Detection of Heat Flux failures in Building using a Soft computing Diagnostic System

Javier Sedano, Emilio Corchado†, Leticia Curiel‡, José Ramón Villar, Enrique de la Cal§*

Abstract: The detection of insulation failures in buildings could potentially conserve energy supplies and improve future designs. Improvements to thermal insulation in buildings include the development of models to assess fabric gain -heat flux through exterior walls in the building- and heating processes. Thermal insulation standards are now contractual obligations in new buildings, and the energy efficiency of buildings constructed prior to these regulations has yet to be determined. The main assumption is that it will be based on heat flux and conductivity measurement. Diagnostic systems to detect thermal insulation failures should recognize anomalous situations in a building that relate to insulation, heating and ventilation. This highly relevant issue in the construction sector today is approached through a novel intelligent procedure that can be programmed according to local building and heating system regulations and the specific features of a given climate zone. It is based on the following phases. Firstly, the dynamic thermal performance of different variables is specifically modeled. Secondly, an exploratory projection pursuit method called Cooperative Maximum-Likelihood Hebbian Learning extracts the relevant features. Finally, a supervised neural model and identification techniques constitute the model for the diagnosis of thermal insulation failures in building due to the heat flux through exterior walls, using relevant features of the data set. The reliability of the proposed method is validated with real data sets from several Spanish cities in winter time.

Key words: *Computational Intelligence, Soft computing, Identification Systems, Artificial Neural Networks, Non-linear Systems, Energetic Efficiency.*

Received: ??
Revised and accepted: ??

*Javier Sedano
Departament of Artificial Intelligence and Applied Electronics. Technological Institute of Castilla and Leon. Poligono Industrial de Villalonquejar , C/Lopez Bravo 70, 09001, Burgos, Spain. `javier.sedano@itcl.es`

†Emilio Corchado
Departamento de Informtica y Automtica. Universidad de Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain. `escorchado@usal.es`

‡Leticia Curiel
Department of Civil Engineering, University of Burgos. EPS Politcnica, Campus Vena, Edificio C. C/ Francisco de Vitoria, s/n, 09001, Burgos, Spain. `lcuriel@ubu.es` Telephone number: +34 947259358.

§José Ramón Villar, Enrique de la Cal
Department of Computer Science, University of Oviedo, Campus de Viesques s/n, 33204 Gijn, Spain `villarjose@uniovi.es, delacal@uniovi.es`. Telephone number: +34 985182597.

1. Introduction

The diagnostic system for identification of thermal insulation failures (TIF) could significantly increase building energy efficiency and substantially contribute to reductions in energy consumption and in the carbon footprints of domestic heating systems. Conventional methods can be greatly improved through the application of learning techniques to detect TIF when a building is in operation through a heat flux model - heat flux through exterior walls in a building-.

Assessing thermal insulation in new buildings is a well-known problem that has not as yet been fully resolved [21, 49]. Several different techniques are proposed in the literature. In [23], thermal insulation leaks are found by measuring thermal resistance and infrared (IR) thermography, while in [2], [37] only IR thermography is used to locate thermal insulation failures. As the main drawback of using IR thermography is the high cost of equipment, alternatives using different technologies are always of interest.

Nevertheless, predicting the thermal dynamics of a building in operation is a complex task. The dynamic thermal performance of a building has mainly been used to estimate its power requirements. As an example, the difficulties of obtaining a black-box model for a generic building are documented in [**?**]. Furthermore, [11] cites examples of the errors associated with different kinds of techniques while providing possible solutions. Local building regulations need to be analyzed in the determination of TIF in order to profile the premises and the legal specifications for their physical parameters.

This interdisciplinary research represents a step forward in the development of techniques to improve dynamic thermal efficiency in existing buildings through a diagnostic system -modeling of heat flux- in the building. Although this may at first appear simple, noise due to occupancy and lighting profiles can introduce distortions and complicate detection. A novel three-step soft computing procedure for testing and validating the model -used in the diagnostic system- is proposed: firstly, the dynamic thermal behavior of a specific configuration is calculated using HTB2 software [29]. The outcome of the HTB2 should then be post-processed to obtain a suitable dataset. Subsequently, the dataset is analyzed using an exploratory projection pursuit (EPP) method [9], [16] called Cooperative Maximum-Likelihood Hebbian Learning (CMLHL) [6], [7], to extract the dataset structure and key relationships between the variables. Finally, a dynamic ANN model is trained and validated with them, which is used for fault diagnosis. This diagnosis dynamic model is responsible for estimating the heat flux through the exterior walls in the building and the results are then compared with the real heat flux. Differences between the estimated and the real measures -above a reference value- are detected which indicate the TIF.

Soft Computing represents a set of several technologies that aim to solve inexact problems [50]. It investigates, simulates, and analyzes very complex issues and phenomena in order to solve real-world problems [40]. Soft Computing has been successfully applied in feature selection, and plenty of algorithms are reported in the literature [4], [5], Principal Component Analysis (PCA) among others [30]. In this study, an extension of a neural PCA version [17] and other extensions are used to select the most relevant input features in a data set as well as to study its internal structure.

This paper is organized as follows. Following this introduction, Section 2 describes the problem. Section 3 introduces the unsupervised connectionist techniques for analyzing the datasets in order to extract their relevant internal structures. Section 4 deals with classical identification techniques used in the diagnostic system –modeling system-. Section 5 describes a real case study in detail and the multi-step procedure. Section 6 describes the experiments and results obtained and finally, the conclusions are set out and comments are made on future lines of work.

2. Spanish regulations and the problem description

Several national regulations on buildings and their construction were approved in Spain, 2007. The minimum pre-requisites for energy efficiency with which buildings must comply are given in the European Directive 2002/91/CE [13]. Project specifications, construction conditions and the basic requirements in Spain are specified in the CTE (*Código Técnico de Edificación* [Building Regulations]) [36]. One of the basic requirements is document HE1 that specifies the energy consumption limitation in buildings [36] and its revised updates.

3. Analysis of the Internal Structure of the data set

In general, to obtain an efficient diagnostic system it is necessary to model it with a good dataset. Often, the systems are modeled using all the variables collected. This is not a proper way as some of them influence in the dynamic of the system and its inclusion only adds complexity to the model process degrading the effectiveness of the final model. For this reason, in this research, it is proposed a previous analysis of the dataset using statistical methods and neural models, as Principal Component Analysis (PCA) [27, 35] and the Cooperative Maximum Likelihood Hebbian Learning model (CMHL) [**?**, 8], respectively, in order to know if the dataset is informative enough and to extract the most relevant variables in order to model it using only the main variables.

Local regulations will be analyzed to extract the minimum requirements and parameters for heating systems and thermal comfort, and the certification procedure for energy efficiency. In Spain, energy efficiency is calculated as the ratio of combustible consumption needed to satisfy the energy demand of the building. Energy efficiency in the case of buildings constructed before the CTE approval is still an open issue, and the assumption is that it will be based on heat flux and conductivity measurement.

In these conditions, it could be interesting to model the heat flux in order to detect the isolation failures in buildings in operation. It is interesting that such model could distinguish the climate zone to analyze, the specific building geometry and orientation, etc. For this modeling task, a novel procedure is proposed. This procedure includes several steps: the thermal dynamics simulation, the feature selection, the heat flux identification using neural networks models and the detection of failures.

3.1 Component Analysis

Principal Component Analysis (PCA) originated in work by Pearson [35], and independently by Hotelling [[27]>] describing multivariate data set variations in terms of uncorrelated variables, each of which is a linear combination of the original variables. Its main goal is to derive new variables, in decreasing order of importance, which are linear combinations of the original variables and are uncorrelated with each other.

3.2 A Neural Implementation of Exploratory Projection Pursuit

The standard statistical method of EPP [**?**, 16], provides a linear projection of a data set, but it projects the data onto a set of basic vectors which best reveal the interesting structure in data; interestingness is usually defined in terms of how far the distribution is from the Gaussian distribution [42].

One neural implementation of EPP is Maximum Likelihood Hebbian Learning (MLHL) [9, 18]. It identifies interestingness by maximizing the probability of the residuals under specific probability density functions that are non-Gaussian.

An extended version of this model is the Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [6] model. CMLHL is based on MLHL [9, 18] adding lateral connections [6, 8], which have been derived from the Rectified Gaussian Distribution [42]. The resultant net can find the independent factors of a data set but does so in a way that captures some type of global ordering in the data set.

Considering an N-dimensional input vector (x), and an M-dimensional output vector (y), with W_{ij} being the weight (linking input i to output j), then CMLHL can be expressed [8, 18] as:

1. Feed-forward step:
$$y_i = \sum_{j=1}^{N} W_{ij} x_j, \forall i \tag{1}$$

2. Lateral activation passing:
$$y_i(t+1) = [y_i(t) + \tau(b - Ay)]^{+} \tag{2}$$

3. Feedback step:
$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i, \forall j \tag{3}$$

4. Weight change:
$$\Delta W_{ij} = \eta . y_i . sign(e_j) |e_j|^{p-1} \tag{4}$$

Where: η is the learning rate, $[\ ]^{+}$ is necessary to ensure that the y-values remain within the positive quadrant, τ is the "strength" of the lateral connections, b the bias parameter, p a parameter related to the energy function [9, 8, 18] and A the symmetric matrix used to modify the response to the data [6]. The effect of this matrix is based on the relation between the distances separating the output neurons.

4. Diagnostic System Using Identification Algorithms

Among the different methods for the detection and diagnosis of faults are: checking limits or thresholds, physical redundancy, deterministic methods -mathematical models-, methods based on knowledge and so on. Some examples are found in the literature [1, 28, 44], [45, 47].

In this context, System identification [31] is concerned with obtaining a model that best suits a given process behavior [31]. Firstly, several measurements are sampled from the process. The data gathered is then analyzed to obtain a model that estimates the desired process behavior. The model is then used to optimize the process output. Finally, the process is modified in order to enhance its outcome. If more adjustments are needed the cycle is repeated.

The system identification procedure includes the experiment design, the data visualization and analysis, the model learning and testing, and the model validation [31, 32, 33, 38, 39, 46, 48].

The experimental design determines the signals to be measured, the sensors to be used and their placement, the sample rate, and the generation of the data sets. Expertise is required as the experimental design decisions are problem dependant. Moreover, it is not always feasible in real world applications to gather data from the most relevant variables and, in most cases, the data is limited by the locations of the sensors that are installed. In other cases, portable instrumentation can be employed to measure some extra process variables. Nevertheless, the human-expert who designs the experiment always has a priori theory and knowledge about the relationships between the variables.

When the data set is gathered, several tasks should be carried out: eliminating missing data and outliers [3, 12, 14, 19, 20] scaling and normalizing the data [43], etc. Whenever the data gathering is expensive and little data is available, it is usual to partition the data generating several train and test data sets. Standardized partitioning schemas are the k-fold cross validation and the 5x2 cross validation. This is all included in the data pre-processing and analysis step.

The selection of the model structure, their training and validation represents the core of the system identification. The classic theory includes a vast amount of model structures and training methods [31, 41]. Well-known functions are also used to rank the goodness of the models. These functions are used as the criteria in the optimization problem of training the model. In the next subsection several criteria functions are introduced and the use of Artificial Neural Networks in system identification is outlined.

4.1 The system identification criteria

According to [31], several measures have been proposed in the literature to evaluate the goodness of a model:

- The representation percentage of the estimated model in relation to the true system, that is, the numeric value of the normalized mean error. There are several typical estimation models used in the literature such as the one-step ahead prediction error (FIT1), the ten-step ahead prediction error (FIT10) and the simulation error (FIT). Equations (5) to (10) are used to calculate the FIT1 and FIT indexes. The FIT10 index can be derived in a similar manner as FIT1. In these equations, $u(t)$ is the input, $y(t)$ is the output, $\hat{y}_1(t|m)$ is the one-step ahead prediction, $\hat{y}_\infty(t|m)$ is the simulated output of the model,$\hat{G}(q)$ is the estimated transfer function from $u(t)$ to $y(t)$, $\hat{H}(q)$ is the estimated transfer function from $e(t)$ to $y(t)$ and q is the forward shift operator. The term $e(t)$ represents the white noise signal and it is included in the modeling errors. The term $e(t)$ is associated with a series of random variables of mean null value and variance λ.

$$\hat{y}_1(t|m) = \hat{H}^{-1}(q)\hat{G}(q)u(t) + (1 - \hat{H}^{-1}(q))y(t) \tag{5}$$

$$J_1(m) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - \hat{y}_1(t|m)|^2 \tag{6}$$

$$FIT1(\%) = (1 - \frac{\sqrt{J_1(m)}}{\sqrt{\frac{1}{N}\sum t = 1N|y(t)|^2}})100 \tag{7}$$

$$\hat{y}_\infty(t|m) = \hat{G}(q)u(t) \tag{8}$$

$$J_\infty(m) = \frac{1}{N}\sum t = 1N|y(t) - \hat{y}(t|m)|^2 \tag{9}$$

$$FIT(\%) = (1 - \frac{\sqrt{J_\infty(m)}}{\sqrt{\frac{1}{N}\sum t = 1N|y(t)|^2}})100 \tag{10}$$

- The loss or error function (V): the numeric value of the mean square error (MSE) that is calculated from the estimation data set by means of Eq. (6).
- The generalization error value: the numeric value of the normalized sum of squared errors (NSSE) that is computed with the validation data set by means of Eq. (6).
- The average generalization error value: the numeric value of the final prediction error (FPE), which is a criterion that is calculated from the estimation data set. Eq. (11) is used to calculated the FPE value, where d_M is the dimension of θ -the estimated parametrical vector- and N is the number of samples of the estimation data set.
$$FPE = \overline{J}(m) \approx J_1(m) + \frac{J_1(m)}{1 - \frac{d_M}{N}}\frac{2d_M}{N} \tag{11}$$
- The graphical representations of true system output and both the one-step ahead prediction $\hat{y}_1(t|m)$, the ten-step ahead prediction $\hat{y}_{10}(t|m)$, and the model simulation $\hat{y}_\infty(t|m)$.

4.2 The ANN in the identification process

The use of ANN in the process of identification requires the selection of several parameters: the number of layers, the number of neurons per layer and the activation functions. The methods by which the parameters are set up are fully documented in the literature. It was found that ANN with two layers using non-linear functions in the hidden layer are universal approximators or predictors [10, 26].

The number of neurons per layer is also a relevant design parameter, and it should be analyzed in order to avoid over fitting [22, 24]. Each algorithm introduces some restrictions in the weight matrix. The most widely used training algorithms in system identification are the Lenvenberg-Marquardt method [15], the recursive Gauss-Newton method [31] and the batch and recursive versions of the back-propagation algorithm [25].

When using ANN, the purpose of an identification process is to determine the weight matrix based on the observations Z^t , so as to obtain the relationships between the nodes in the network. The weight matrix is usually referred as w, W or θ.

The supervised learning algorithm is then applied to find the estimator θ, so as to obtain the identification criterion [40]. Several well-known model structures are used when merging system identification with ANN. If the AutoRegressive with eXternal input model

(ARX) is used as the regression vector, the model structure is called a Neural Network for ARX model (NNARX). Likewise, the Neural Network for Finite Impulse Response model (NNFIR), the Neural Network for Autoregressive Moving Average with eXternal input model (NNARMAX), and the Neural Network for Output Error model (NNOE), are also extensively used [40]. In the same way, it is possible to use an estimator for the one-step ahead prediction of the output $\hat{y}_1(t|m)$, where the polynomial degree values -n_a, n_b, n_c, n_d, n_f and n_k- are given as parameters.

5. A multi-step method for modeling heat flux in buildings

The novel three-step Soft computing method proposed to diagnose insulation failures, for the detection of heat flux through exterior walls in the building incorporates a diagnostic system that integrates different methodologies, to obtain a parametric model which performs the diagnosis.

Firstly, the building is parameterized and its dynamic thermal performance in normal operation is obtained by means of simulation. Then, the data gathered is processed using CMLHL as a dimensionality reduction technique to choose the most relevant features in order to determine the heat flux. The second step outcome is a data set, which is finally used to train and validate the heat flux nonparametric model that was used in the diagnostic system.

Fig. 1 shows the diagnostic system in a global manner. It indicates how training data are acquired from a theoretical model -HTB2-, which incorporates all the dynamic characteristics of thermal system. After data are preprocessed, using feature selection techniques and attributes. A dynamic ANN model is trained and validated with them, which is used for fault diagnosis. The actual data -real data set- of the building's thermal system will be evaluated in the model that is generated by assessing two indexes: the representation percentage of the estimated output in relation to the true output (FIT1), Eq. (7) and the numeric value of the error, which is the difference between the responses of the heat flux -$y_1(t)$- in the building and the estimated heat flux -$y_1(t|m)$- in the diagnostic system.

When the error exceeds a certain value -threshold value- or the FIT1 is less than a reference value, then the diagnostic system determines a failure.

5.1 Thermal dynamics data gathering by means of simulation

The following variables and data sets should be gathered in order to simulate the thermal behavior of a building: building topology; climate zone according to the specific regulations; building materials that comply with local regulations for the chosen climate zone; meteorological data for the climate zone and the simulated time period: such as solar radiation, outdoor temperature, wind speed, etc., and realistic profiles for heating, lighting, small power devices, occupancy and ventilation.

In this study, the system is applied is Spain where the regulations establish five winter/summer zones, from E1 (a more severe climate zone) to A3 (a gentler climate zone).

Having defined and/or gathered these data sets, then the chosen simulation tool is applied to obtain the output data. In our case, the simulation software used is HTB2 [29]. The typical values that each variable could take for an E winter climate zone of maximum severity in Spain -i.e. the cities of Leon, Burgos or Soria among others- are shown in Table I.

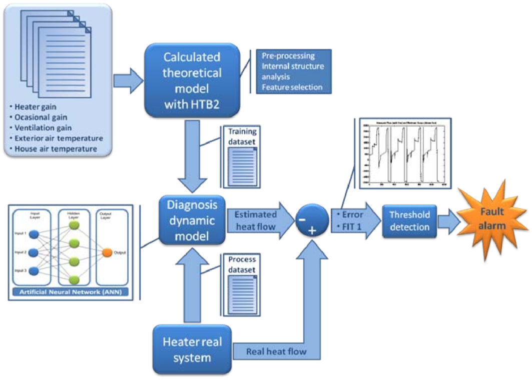

Fig. 1 *The diagnostic system: the data are obtained from a theoretical model - through HTB2-. They are then processed and a better data set is found. The data set is used to train the dynamic ANN model. Actual data - from a thermal system in operation- will be evaluated on the model, identifying errors that will determine the failure.*

5.2 Selection of the relevant features

As detailed in Section 2, PCA (Fig. 2.a) and CMLHL (Fig. 2.b), which were both applied to this real-life problem, are instrumental in identifying the internal structure of the data. In this procedure, the data set gathered in the previous step is analyzed. The objective is to find the relationships between the input variables with respect to the heat flux. CMLHL (Fig. 2.b) allows to detect the relations of dependence and to choose the most relevant features. The outcome of this step is a new data set with the features for which a relationship with the heat flux is found.

5.3 System identification applied to model normal building operation

Once the relevant variables and their transformations have been extracted from the thermal dynamics data, then a model to fit the normal building operation should be obtained in order to identify bias in the heat flux through exterior walls in the building. The heating process exhibits nonlinear behavior between output and inputs, due to which the linear modeling techniques do not behave properly except in the linear behavior zones of the process. Consequently, the heating process has been modeled using soft computing techniques, specifically an ANN.

The different learning methods used in this study were implemented in Matlab© [34]. The experiment followed the identification procedure detailed in Section 4: the model structures were analyzed in order to obtain the models that best suited the dataset. The Akaike Information Criterion (AIC) is used to obtain the best degree of the model and its delay for each model structure. A total of thirty four different combinations of model structures and optimization techniques were considered -such as the Levenberg-Marquardt method and the recursive Gauss-Newton method for the NNARX, NNFIR, NNARMAX and NNOE models [31, 34].

Three different residual analyses based on cross correlation were performed: residual analysis between the residual $\hat{R}_{\varepsilon}^{N}(\tau)$, between the residual and the input $\hat{R}_{\varepsilon u}^{N}(\tau)$ and the

Variable (Units)	**Range of values**	**Transmittance level (W/m^{2}K)**
Fabric gain -heat flux- (w), y_1(t).	0 to -7,100	-External cavity wall: 0.54
Heater gain (W), u_1(t).	0 to 4,500	-Double glazing: 2.90
Occasional gain small power, occupancy and lighting gain- (W), u_2(t).	0 to 5,500	-Floor/ceiling: 1.96
Ventilation gain (w), u_3(t).	0 to -5,500	-Floor/ceiling: 1.96
Exterior air temperature in February (C), u_4(t).	1 to 7	-Others party wall: 1.05
Air temperature of the house (C), u_5(t).	14 to 24	-Internal partition: 2.57

Tab. I *Typical values of each variable in an E winter climate zone city in Spain.*

non-linear residual correlation $\hat{R}^N_{\varepsilon^2 u^2}(\tau)$.

6. Experimentation and results

The theoretical model has been generated from realistic situations. The model used in this study was implemented in HTB2 [29] and used to gather the initial data set. The main output of a HTB2 simulation is the heater gain –the power requirements in the modeled building-, but also the Fabric gain -heat flux- the temperature and other variables in table I.

The realistic materials in the construction, the volumetric measures of each room, the neighbourhood of the rooms, the orientation and geographical earth zone, the solar radiation profile, the environment data, the heating subsystems, the occupancy profile, the temperature-time profile for each heating subsystem, the small power devices and the light ON profileswere considered, among others, to validate the proposal. A building in the E winter zone, in the city of Avila is used as the actual building location. Different sample periods and the length of the simulations have been fixed too.

This initial data set has been analyzed, then, in order to select the features that best describe the relationships with the heat flux. As may be seen in Fig. 2, PCA (Fig. 2.a) and CMLHL (Fig. 2.b), both methods have identified the occasional gain as the most relevant variable but more structured clusters than in the PCA projections may be noted in the CMLHL projections (Fig. 2.b).

Having analyzed the results obtained with the CMLHL model (Fig. 2.b) it can be concluded that CMLHL has identified four relevant variables and seven clusters ordered by occasional gain. Inside each cluster there are further classifications according to heater gain, ventilation gain and, to a lesser degree, exterior air temperature. Accordingly, it may be said that the heat flux and the dataset have an interesting internal structure. When the dataset is considered sufficiently informative, then the third step of the process begins. This step performs an accurate and efficient optimization of the heating system model to detect the heat flux model in the building, through the application of several conventional modeling systems.

Thus, an ANN was used to monitor the thermal dynamics of the building. The objective was to find the best suite of polynomial model orders [n_a, n_{b1}, n_{b2}, n_{b3}, n_{b4}, n_c,

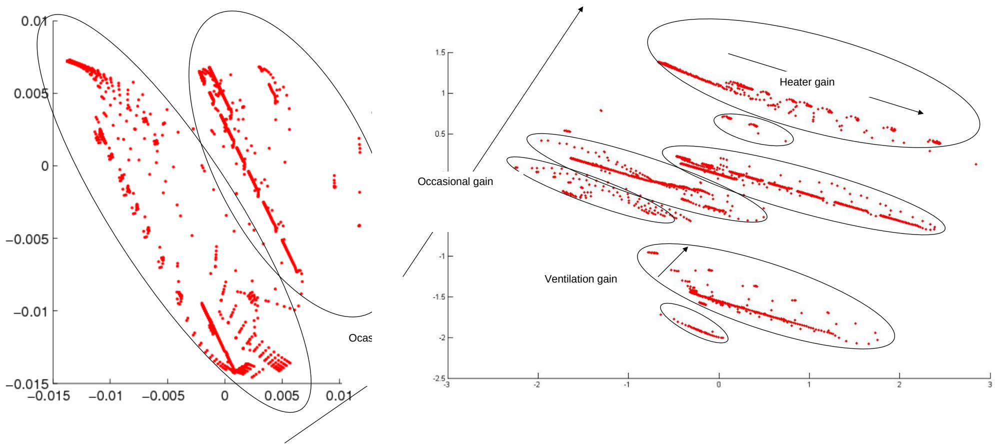

Fig. 2 *PCA projections in left figure (Fig. 2.a) and CMLHL projection in the right figure (Fig. 2.b) after 20000 iterations using a learning rate of 0.05, 3 output neurons p=0.3 and τ=0.3.*

n_d, n_f, n_{k1}, n_{k2}, n_{k3}, n_{k4}]. Using the data set from the previous stage and the Optimal Brain Surgeon (OBS) [22, 24] network pruning strategy to remove superfluous weights, the best suite model was found from the residual analysis. Table II shows the estimation and prediction characteristics and qualities of the chosen ANN, along with their indexes.

Fig. 3 shows the time responses of the heat flux -$y_1(t)$- and of the estimated heat flux -$\hat{y}_1(t|m)$- for the NNARX model [40]. The x-axis shows the number of samples used in the estimation and validation of the model and the y-axis represents the normalized output variable range: which is the normalized heat flux of the house. The estimation and validation data sets include 2000 and 1126 samples, respectively, and have a sampling rate of 1 sample/minute. Fig. 4 indicates the final neural network structure chosen for modeling heat flux, both of which are polynomial model orders. These orders specify the inputs to the ANN –four for a full connected–and the indices of the orders represent each of the thermal system inputs.

From, Fig. 4 it can be concluded that the pruned network of the NNARX model is able to simulate and predict the behavior of the heat flux through exterior walls in the building as a consequence of the heating process- and it is capable of modeling more than 91.4% of the actual measurements. This model does not only present a lower loss function (V) and error values (NSSE and FPE), but also a higher system representation index value (FIT1).

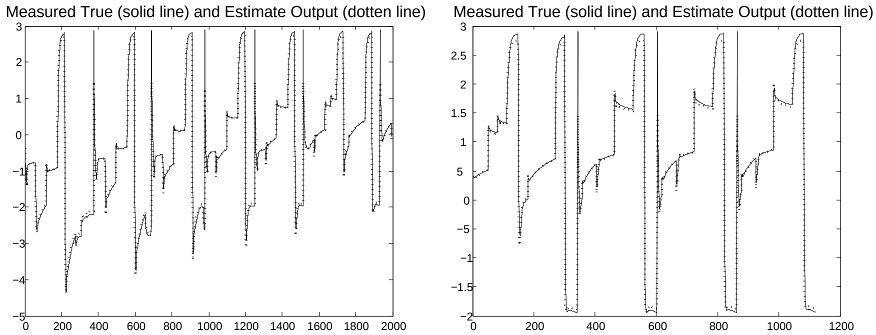

Fig. 3 *Output response of NNARX model: the actual output (solid line) is graphically presented with one-step-ahead prediction (dotted line). In Fig. 3.a (left) the real measure can be compared with the estimated data, while in Fig. 3.b (right) the real measure is compared with the validation data.*

7. Conclusions and future work

Effective thermal insulation is an essential component of energy efficient heating systems in buildings. Thus, the possibility of improving the detection of thermal insulation failures represents a fresh challenge for building energy management.

The new methodology proposed in this study to diagnose insulation failures from the heat flux through exterior walls in the building can be used to determine the normal operating conditions of thermal insulation in buildings in Spain, which has recently became a mandatory test in the evaluation of building insulation.

The novel soft computing diagnostic system as presented here improves fault detection with respect to detection systems that rely on isolated signals -used in the industrial processes-. The detection is based in the analysis of the numeric value of the error -difference between the responses of the real heat flux and the estimated heat flux in the building- and the representation percentage of the estimated output in relation to the true output. This analysis presents a low dependency respect to the input signals.

Future work will create a standard of theoretical failures -data set- in the normal conditions of heating, lighting, small power devices, occupancy and ventilation, so that the diagnostic system in the building –thermal system- can incorporate a global fault classifier. Moreover, automation of the diagnostic system will further improve its performance.

8. Acknowledgments

We would like to extend our thanks to Phd. Magnus Nørgaard for his freeware version of Matlab Neural Network Based System Identification Toolbox. This research has been partially supported through projects of the Junta of Castilla and León (JCyL): [BU006A08], TIN2010-21272-C02-01 from the Spanish Ministry of Science and Innovation, the project of Spanish Ministry and Innovation [PID 560300-2009-11], the project of the Spanish Ministry of Education and Innovation [CIT-020000-2008-2] and [CIT-020000-2009-12], the project of Spanish Ministry of Science and Technology [TIN2008-06681-C06-04] and Grupo Antolin Ingenieria, S.A., within the framework of project MAGNO2008 - 1028.-CENIT also funded by the same Government Ministry.

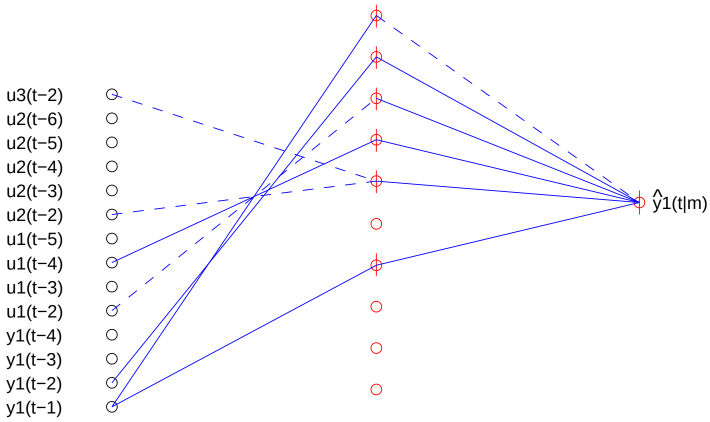

Fig. 4 *Optimal architecture of the NNARX model, with the pruned network, for the heat flux through the exterior walls of the building -output* $\hat{y}_1(t|m)$ *-. Positive weights are represented in solid lines, while a dashed line represents a negative weight. A vertical line through the neuron represents a bias.*

References

[1] M. Alexandru. Building on-line diagnosis system upon artificial neural network techniques. *International Journal on Non-standard Computing and Artificial Intelligence. Neural Network World*, 10(6):523–534, 2000.

[2] C. A. Balaras and A. A. Argiriou. Infrared Thermography for Building Diagnostics. *Future Generation Computer Systems*, 34(2):171–183, 2002.

[3] G.E.A.P.A. Batista and M.C. Monard. An Analysis Of Four Missing Data Treatment Methods For Supervised Learning. *Applied Artificial Intelligence*, 17(5):519–533, 2003.

[4] J. Casillas, O. Cordón, F. Herrera, and P. Villar. A hybrid learning process for the knowledge base of a fuzzy rule-based system. In *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU 2004*, pages 2189–2196, Perugia, Italy, 2004.

[5] T. W. S. Chow, P. Wang, and E. W. M. Ma. A New Feature Selection Scheme Using a Data Distribution Factor for Unsupervised Nominal Data. *IEEE Transactions on Systems, Man and Cybernetics - PART B: Cybernetics*, 38(2):499–509, 2008.

[6] E. Corchado and C. Fyfe. Connectionist Techniques for the Identification and Suppression of Interfering Underlying Factors. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(8):1447–1466, 2003.

[7] E. Corchado and C. Fyfe. Orientation Selection Using Maximum Likelihood Hebbian Learning. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 7(2):11–30, 2003.

[8] E. Corchado, Y. Han, and C. Fyfe. Structuring global responses of local filters using lateral connections. *Journal of Experimental & Theoretical Artificial Intelligence*, 15(4):473–487, 2003.

[9] E. Corchado, D. MacDonald, and C. Fyfe. Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit. *Data Mining and Knowledge Discovery*, 8(3):203–225, 2004.

Model	Indexes
ANN model for the heating process, NNARX regressor, the order of the polynomials of the initial fully connected structure are n_a=4, n_{b1}=4, n_{b2}=5, n_{b3}=1, n_{b4}=4, n_{k1}=2, n_{k2}=2, n_{k3}=2, n_{k4}=2, [4 4 5 1 4 2 2 2 2]. The model was obtained using the regularized criterion. This model was optimised by CMLHL analysis, residual analysis and the pruned network, using OBS. The model structure has 10 hidden hyperbolic tangent units and 1 linear output unit. The network is estimated using the Levenberg-Marquardt method, and the model order is decided on the basis of the best AIC criterion of the ARX model.	FIT1:91.4% V: 0.0068 FPE: 0.12 NSSE:0.0049

Tab. II *The value of the quality indexes obtained for the proposed model. FIT1, V, NSSE and FPE stand for the graphical representation percentage, the loss function error, the normalised sum of squared error and the final prediction error.*

[10] G. Cybenko. Aproximation by Superpositions of Sigmoidal Function. *Math. Control, Sygnals and System*, 2(4):303–314, 1989.

[11] E. de la Cal, J.R. Villar, and J. Sedano. A thermodynamical model study for an energy saving algorithm. In *Lecture Notes in Artificial Intelligence:Hybrid Artificial Intelligence Systems*, volume 5572, pages 384–390. Springer, 2009.

[12] J. Deogun, W. Spaulding, B. Shuart, and D. Li. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. In *Lecture Notes in Computer Science: Proceedings of the 4th International Conference of Rough Sets and Current Trends in Computing (RSCTC'04)*, volume 3066, pages 573–579, Uppsala (Sweden), 2004. Springer.

[13] Directive 2002/91/CE of the European Parliament and the Council of 16 December 2002 on the energy performance of buildings, Oficial Journal of the European Community, 2003.

[14] H.A.B. Feng, G.C. Chen, C.D. Yin, B.B. Yang, and Y.E. Chen. A SVM regression based approach to filling in Missing Values. In *Lecture Notes in Computer Science: 9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES2005)*, volume 3683, pages 581–587, Melbourne (Australia), 2005. Springer.

[15] R. Fletcher. *Practical Methods of Optimization.* Wiley & Sons, Chichester, UK, 2nd edition, 1987.

[16] J.H. Friedman and J.W. Tukey. Projection Pursuit Algorithm for Exploratory Data-Analysis. *IEEE Transactions on Computers*, 23(9):881–890, 1974.

[17] C. Fyfe, R. Baddeley, and D.R. McGregor. *Exploratory Projection Pursuit: An Artificial Neural Network Approach.* Research Report/94/160. University of Strathclyde, 1994.

[18] C. Fyfe and E. Corchado. Maximum Likelihood Hebbian Rules. In *Proceedings of the 10th European Symposium on Artificial Neural Networks(ESANN 2002)*, pages 143–148, Bruges, Belgium, 2002. D-side Publishers.

[19] P.A. Gourraud, E. Ginin, and A. Cambon-Thomsen. Handling Missing Values In Population Data: Consequences For Maximum Likelihood Estimation Of Haplotype Frequencies. *European Journal of Human Genetics*, 12(10):805–812, 2004.

[20] J.W. Grzymala-Busse, L.K. Goodwin, W.J. Grzymala-Busse, and X. Zheng. Handling Missing Attribute Values in Preterm Birth Data Sets. In *Lecture Notes in Computer Science: Proceedings of the 10th International Conference of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'05)*, volume 3642, pages 342–351, Regina (Canada), 2005. Springer.

[21] J. Han, L. Lua, and H. Yang. Investigation on the thermal performance of different lightweight roofing structures and its effect on space cooling load. *Applied Thermal Engineering*, 29(11-12):2491–2499, 2009.

[22] L.K. Hansen and M.W. Pedersen. Controlled Growth of Cascade Correlation Nets. In *International Conference on Artificial Neural Networks (ICANN1994)*, pages 797–800, Italy, 1994. Eds. M. Marinaro and P.G. Morasso.

[23] D.A. Haralambopoulos and G.F. Paparsenos. Assessing the thermal insulation of old buildings. The need for in situ spot measurements of thermal resistance and planar infrared thermography. *Energy Conversion and Management*, 39(1-2):65 – 79, 1998.

[24] B. Hassibi and D.G. Stork. Second Order Derivatives for Network Pruning: Optimal Brain Surgeon. In *Advances in neural Information Processing System 5*, pages 164–171, San Mateo, CA, USA, 1993. Eds. S.J. Hanson et al.

[25] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation*, volume 1. Addison-Wesley, Santa Fe Institute Studies In The Sciences Of Complexity, 1991.

[26] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Aproximators. *Neural Networks*, 2(5):359–366, 1989.

[27] H. Hotelling. Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Education Psychology*, 24:417–444, 1933.

[28] D. Kolokotsa, A. Pouliezos, and G. Stavrakakis. Sensor fault detection in building energy management systems. *Dynamics of Continuous Discrete and Impulsive Systems-Series B-Applications & Algorithms*, 13(6):221–225, 2006.

[29] P.T. Lewis and P.K. Alexander. A Flexible Model for Dynamic Building Simulation: HTB2. *Building and Environment*, 1:7–16, 1990.

[30] D.T. Lin and D.C. Pan. Integrating a Mixed-Feature Model and Multiclass Support Vector Machine for Facial Expression Recognition. *Integrated Computer-Aided Engineering*, 16(1):61–74, 2009.

[31] L. Ljung. *System Identification. Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., USA, 2nd edition, 1999.

[32] G.M. Mendez, L. Leduc-Lezama, R. Colas, G. Murillo-Pérez, J. Ramírez-Cuellar, and J.J. López. Application of Interval Type-2 Fuzzy Logic Systems for Control of the Coiling Entry Temperature in a Hot Strip Mill. In *Lecture Notes in Computer Science: Proceedings of the Hybrid Artificial Intelligent Systems*, volume 5572, pages 352–359, 2009.

[33] O. Nelles. *Nonlinear System Identification. From Classical Approaches to Neural Networks and Fuzzy Models*. Springer, Berlin, Germany, 2001.

[34] M. N¨rgaard. Neural network Based System Identification Toolbox. Technical Report 00-E-891, Department of Automation, Technical University of Denmark, 2000.

[35] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6):559–572, 1901.

[36] Real Decreto 314/2006, de 17 de Marzo, por el que se aprueba el Código Técnico de la Edificación. BOE num 74, España, 2006.

[37] S. Ribarić, D. Marŏetić, and D. S. Vedrina. A knowledge-based system for the non-destructive diagnostics of façade isolation using the information fusion of visual and IR images. *Expert Systems with Applications*, 36(2):3812–3823, 2009.

[38] F. Rodriguez, J.L. Guzman, M. Berenguel, and M. Ruiz. Adaptive Hierarchical Control of Greenhouse Crop Production. *International Journal of Adaptive Control and Signal Processing*, 22(2):180–197, 2007.

[39] J. Schoukens, Y. Rolain, and R. Pintelon. Improved Approximate identification of Nonlinear Systems. In *21st IEEE Instrumentation and Measurement Technology Conference*, pages 2183–2186, Como, Italy, 2004.

[40] J. Sedano, E. Cal, L. Curiel, J.R. Villar, and E. Corchado. Soft Computing for detecting Thermal Insulation Failures in Buildings. In *Proceedings of 9th International Conference on Computational and Mathematical Methods in Science and Engineering*, Gijón, Spain, June.

[41] J. Sedano, L. Curiel, E. Corchado, E. de la Cal, and J.R. Villar. A Soft Computing Based Method for Detecting Lifetime Building Thermal Insulation Failures. *Integrated Computer-Aided Engineering, IOS Press*, 17(2):103–115, 2010.

[42] H.S. Seung, N.D. Socci, and D. Lee. The Rectified Gaussian Distribution. *Advances in Neural Information Processing Systems*, 10:350–356, 1998.

[43] L.A. Shalabi, Z. Shaaban, and B. Kasasbeh. Data Mining: A Preprocessing Engine. *Journal of Computer Science*, 2(9):735–739, 2006.

[44] S. Simani. Fault diagnosis of a simulated industrial gas turbine via identification approach. *International Journal of Adaptive Control and Signal Processing*, 21(4):326–353, 2007.

[45] S. Simani, C. Fantuzzi, and S. Beghelli. Diagnosis techniques for sensor faults of industrial processes. *IEEE Transactions on Control Systems Technology*, 8(5):848–855, 2000.

[46] T. Söderström and P. Stoica. *Nonlinear System Identification. From Classical Approaches to Neural Networks and Fuzzy Models.* Prentice Hall International, London,UK, 1989.

[47] G. Xian and B. Zeng. An effective and novel fault diagnosis technique based on EMD and SVM. *International Journal on Non-standard Computing and Artificial Intelligence. Neural Network World*, 20(4):427–439, 2010.

[48] P.C. Young, C.J. Taylor, W. Tych, D.J. Pedregal, and P.G. McKenna. *The Captain Toolbox*. Centre for Research on Environmental Systems and Statistics, Lancaster University, 2004.

[49] J. Yu, C. Yang, L. Tian, and D. Liao. Evaluation on energy and thermal performance for residential envelopes in hot summer and cold winter zone of China. *Applied Energy*, 86(10):1970–1985, 2009.

[50] L.A. Zadeh. Soft computing and fuzzy logic. *IEEE Software*, 11(6):48–56, 1994.