Thermodynamic Identification of Buildings using Wireless Sensor Networks

E. Toffoli, G. Baldan, G. Albertin L. Schenato * **A. Chiuso** ** **A. Beghi** *

* *Department of Information Engineering, University of Padova*
emails: {toffolie—gbaldani—galberti—schenato—beghi}@dei.unipd.it
** *Department of Management and Engineering, University of Padova, Vicenza, Italy, (email: chiuso@dei.unipd.it)*

Abstract:
In this paper we study different strategies for identifying thermodynamic models of buildings using experimental data collected from large scale wireless sensor networks. Wireless sensor networks can easily provide temperature, humidity, and solar radiation measurements from tens to hundreds of sensors, thus potentially providing a fine grain spatial-temporal resolution. In order to cope with such a large number of inputs and outputs, we tested sub-space identification algorithms which are particularly suitable for identifying large scale MIMO systems. The identified model can be used to evaluate the thermodynamic efficiency of the building. We also explore different sensor selection strategies in order to choose among all sensors the most informative ones. The use of a small set of sensors not only greatly reduces the computational burden in the identification algorithms, but also allows to predict with high accuracy the measurements of the other sensors using Kalman filtering techniques. The adopted identification and Kalman filtering algorithms, as well as the sensor selection strategies have been tested and analyzed using experimental data collected from 65 sensors deployed in a $80m^2 - 200m^3$ building over an 11 day period.

Keywords: Subspace methods; Hybrid and distributed system identification

1. INTRODUCTION

There is an ever growing attention on energy conservation policies and technologies that can improve energy efficiency with low negative environmental effects, due to the steadily increasing prices of energy resources and environmental concerns about climate changes. In particular, energy expenditure for temperature control in buildings account for up to 30% of total budget and it is doomed to increase with the proliferation of air conditioning systems. Therefore, there is a need to use energy more efficiently and in a cleaner manner in both new and old buildings, as testified also by a recent European Community Directive 2002/91/EC which imposes several actions in these directions. Therefore, there is a strong need to develop both technologies and tools that can provide:

- *a-posteriori* evaluation of the thermal efficiency of a building, i.e. energy labelling based on experimental data;
- thermal monitoring and comfort control systems especially in large building
- energy saving quantification after remodeling and energy-specific retrofitting of existing buildings
- automatic fault-detection and monitoring of Heating Ventilation and Air Conditioning (HVAC) systems

Wireless sensor networks (WSNs) seem to be a particularly useful and suitable technology in this prospect. In fact, a WSN is a network of small devices, called motes, provided with sensors (such as temperature, humidity and solar radiation sensors), a microcontroller, some memory and I/O ports, and a wireless antenna which allow them to communicate with their neighbors. WSNs are easy to deploy since they are battery powered, they do not need to be placed in specific locations since the network is self-configurable and adaptive, they are non-intrusive since each device is smaller than a cigarette packet, and finally they are quite inexpensive. As a consequence a WSN, by avoiding the need of cabling, can be rapidly installed also in existing buildings with minor costs and intrusion, and collect measurements from hundreds of locations for long periods of time. Such measurements thus provide an unprecedented quantity of information that can be used to identify a fine grain model of the building and to certificate its thermal efficiency. Moreover, it is possible to envision the use of WSNs not only for thermal efficiency certification, but also for collecting data for realtime thermal monitoring and regulation systems especially in large buildings. However, the effective use of WSNs for thermodynamic identification requires the development of novel mathematical tools that can cope with such a large number of sensors. So far ther-

⋆ This work was partially supported by EU fund MIRC-6-CT-2005-014815 "SENSNET" and by the national project *New techniques and applications of identification and adaptive control* funded by MIUR.

modynamic identification of buildings have been developed based on data collected from a small number of sensors, mainly due to the fact that measurement collection is expensive and time-consuming. The most popular tools adopted for thermodynamics identifications of buildings are based on ARX, ARMAX and Neural Networks models (see Dodier and Henze (2004)); these two latter model classes turn to be particularly difficult to handle when the number of inputs and outputs grow very large, leading also to ill-conditioned estimation problems when the inputs and outputs are highly correlated, as it is the case for measurements collected from sensors which are closely located. Moreover, the choice of the location of sensors from which data for identification is collected, is generally based on experience and rule of thumbs, and little has been done to experimentally evaluate which are the most informative locations where to place the sensors. With respect to this, WSN can prove to be an extremely valuable resource. In fact, it is easy to place a large number of sensors, therefore systematic analysis can be performed and possibly experiment-based sensor placement guidelines can be derived. Finally, it is of great interest also to be able to select among all sensors in the WSNs initially adopted to identify a model, a small set from which we can still predict the measurement of all the others with high accuracy. These sensors, for example, could be used for collecting real-time measurements for thermoregulation systems, which have been shown to grant an increase of the energy performance when good predictive models of the building thermodynamical evolution are available, see AA. VV. (1994).

In this work, we propose to use WSNs for collecting data for thermodynamic identification of building, since they can be used to rapidly collect measurements from a large number of sensors. In order to cope with the large number of available measurements we adopted recently developed sub-space identification tools that, when compared to the traditional methods mentioned above, have the advantage to be numerically efficient also for large scale MIMO systems. To our knowledge, this is the first attempt to apply sub-space methods for identification of thermodynamical models of buildings. Then we deal with the problem of optimal sensor selection in terms of extracting the most informative sensors from an identification perspective. With this respect we propose some selection heuristics that seem to provide good performance. Finally, we also show how a small number of sensors, if appropriately chosen, can predict with very high accuracy the readings of all the other sensors by using Kalman filtering techniques, thus providing a useful tool that can be used to close the loop around a thermoregulation systems. The proposed methodologies were tested using experimental data collected from 65 sensors deployed in a $80m^2 - 200m^3$ building over an 11 day period. The main limitation of these experiments is that data where collected in open-loop thermodynamical conditions of the building, i.e. the temperature inside the building was not regulated by any heating/cooling system therefore the state of the building is affected by the external temperature, humidity and solar insulation only. Although the very goal of identification of building thermodynamic models is the energy efficiency of the building under closed-loop conditions, i.e. when the heating/cooling regulation is in place, we believe that the methodologies proposed can be readily extended by simply including as inputs also the external loads associated with the heating/cooling system.

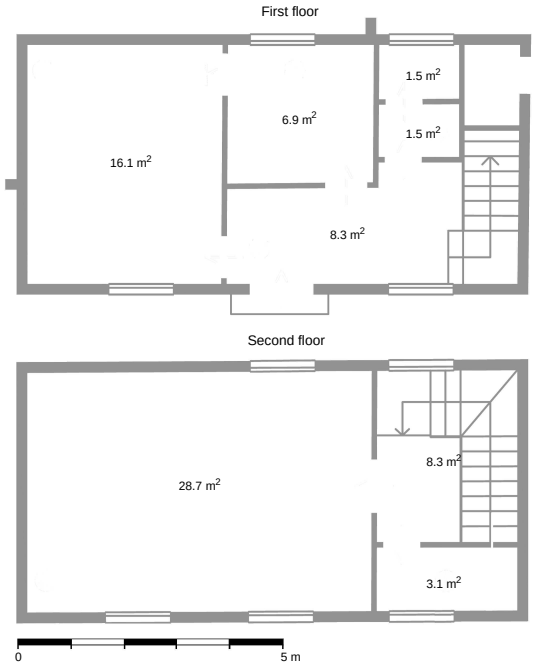

Fig. 1. Map of the building used as test case.

Fig. 2. Front face of the building.

2. EXPERIMENTAL TESTBED AND DATA COLLECTION

As mentioned in the previous section, we tested the identification techniques based on experimental data collected from a real building. The edifice we took into exam is a small, two-story residential building of about 80 m^2 and 200 m^3 (see Figs. 1 and 2.) The building is located in Padova (Italy) at a latitude of 45.41 °N. The experimental data was collected through a WSN made of 65 *Tmote-Sky* nodes produced by Moteiv Inc. Each Tmote-Sky [1] is provided with a temperature sensor, a humidity sensor, and

[1] Datasheet available at `http://www.moteiv.com/products/docs/tmote-sky-datasheet.pdf`

a total solar radiation photoreceptor (visible + infrared). Measurements from the humidity sensors were not used in the identification schemes.

The data were collected under ideal conditions. In particular, during the data collection period, the building was not inhabited, and all external windows and doors were closed. This prevented the natural thermal dynamics to be disturbed by nonlinear and unpredictable phenomena due to air exchange with the external environment. Also, the building thermoregulation system was not in service. As a consequence, the system dynamics were subject to external, climatic inputs only. On the other hand, we have not been able to derive a thermodynamic model of the building in closed-loop with the thermoregulation system, so that it is not possible to analyze the energy efficiency of the building. Also, it has not been possible to use the building heating/cooling system to produce strongly exciting inputs to be used for system identification.

As far as the initial placement of the 65 available sensor nodes is considered, one sensor measuring temperature and one sensor measuring total solar radiation were placed on each wall on the outer surface of the building at an height of about 4.5 m. The remaining 57 sensors, used as temperature sensor only, were positioned inside the building. In particular, they were uniformly distributed in space so that the resulting model could describe precisely the temperature in each part of the building, and some of them were placed in the proximity of windows and doors where most part of the heat exchange takes place.

A single experimental measurement session has been performed, covering a period of 11 days (from June 15th, 2007 to June 26th, 2007.) We used a sampling time of 10 minutes. As mentioned above, data have been collected in poor model input excitation conditions (constant weather conditions with sun shining every day.) The external high temperature was always around 31 °C while the low temperature was about 24 °C. The mean internal temperature of the building increased every day passing from about 25 °C on the first day to about 28 °C on the last day.

3. SUB-SPACE MODEL IDENTIFICATION WITH INPUT SELECTION

The goal of this section is to model the building thermal dynamics in terms of a discrete time, linear, time invariant model in state space form with both exogenous input and stochastic zero-mean additive noise. Model Inputs $u(t) \in \mathbb{R}^m$ are a subset chosen from one or more classes of sensors such as

- 4 external temperature sensors
- 4 external total solar radiation (visible and infrared spectrum)
- 4 internal temperature sensors placed on the ground floor

while outputs $y(t) \in \mathbb{R}^l$ are the (remaining) internal temperature measurements in the various rooms of the building. All inputs have been properly scaled in order to avoid numerical issues. The rationale behind the use of temperature sensors as inputs is that they are indirectly related to the heat exchange between the build and the air and the building and the ground.

Assuming we have N measurements for each sensor, we used the linear model in forward innovation form

$$\begin{cases} x(t+1) = A\,x(t) + B\,u(t) + K\,e(t) \\ y(t) \quad\quad = C\,x(t) + e(t) \end{cases} \tag{1}$$

where $t = 1, 2, \ldots, N$, $e(t)$ is the zero-mean white noise, i.e. $E[e(t)] = 0$ and $E[e(t)e^T(t)]$ is known. We also assumed that there is at least a one-step time delay between input and output, i.e. the output $y(t)$ is not directly affected by $u(t)$. The dimension of the state space is denoted by n, i.e. $x(t) \in \mathbb{R}^n$, and the model matrices are sized accordingly.

The number of inputs and outputs is large and hence we decided to use subspace identification techniques. These methods are based on robust, non iterative and numerically efficient linear algebra tools which, contrary to other methods based on the optimization of some cost function (e.g. Prediction Error Methods, see Ljung (1997)) do not require performing costly iterative minimization thus also avoiding the risk of getting stuck in local minima, see e.g. Van Overschee and De Moor (1996),Chiuso (2007). In particular, we employed the *N4SIS* algorithm [2] available in the MATLAB® System Identification Toolbox, and the recursive version of the $PBSID_{opt}$ algorithm in Chiuso (2007).

As mentioned in the introduction, using a large number of inputs and outputs data can potentially provide a great wealth of information to obtain a detailed model for the building thermodynamics. However, if the inputs and the outputs of the model are very similar, i.e. they are highly correlated, as is the case when the sensors are closely positioned, we may incur in severe numerical problems due to collinearity. This means that the estimated models will be very sensitive to the available data (i.e. the estimators' variance will be large). Indeed, some preliminary identification tests confirmed that the model identified with the full set of inputs provided lower prediction performance in fitting the validation data, as compared to models identified with only a subset of total inputs. This is a classic problem in model identification where there is a tradeoff between bias and variance. To cope with this problem the model can forced to be "simple" by adding some regularization terms or by imposing that the model has only a small number of parameters. Another strategy is to select only a fraction of all possible inputs and outputs. Here we follow this latter approach, it being more suitable for subspace identification algorithms.

Note that, if the linear model were known (as is clearly not the case in system identification problems), one could obtain a measure of relative importance of each single input (see the survey by van de Wal and de Jager (2001) for a presentation of various methodologies). Methods for avoiding the collinearity problems include extensions of Principal Component Regression (see e.g. Greenberg (1975)), of PLS (see Wold (1966)) and its dynamic extensions (see e.g. Qin (1998) and references therein).

In this preliminary work we take a simple route which we describe next, leaving to future work the analysis and development of more sophisticated techniques. The

[2] The current implementation in Matlab is actually a mixture of the most well known methods, see Van Overschee and De Moor (1994), Verhaegen (1994) and Larimore (1990).

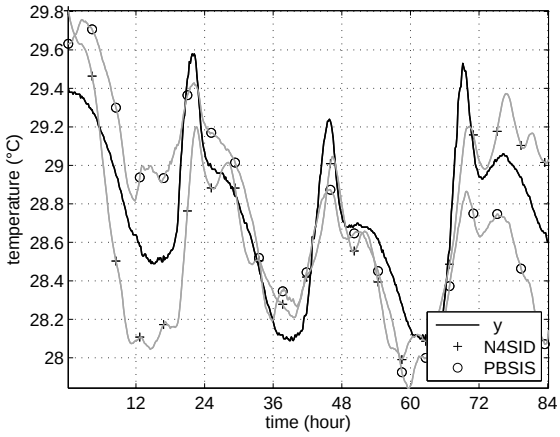

Fig. 3. Temperature measured by one internal sensor, i.e. one entry of output vector y, and simulated temperature y_u using the models identified by *N4SID* and $PBSID_{opt}$.

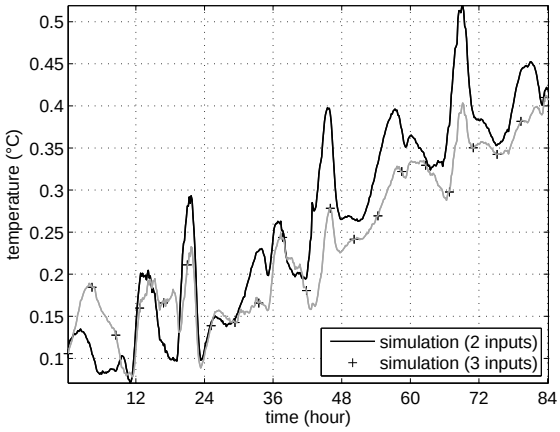

Fig. 4. Simulation error fit(M) based on the model identified by $PBSID_{opt}$ using 2 and 3 inputs.

input selection has been achieved by constructing, for each candidate subset, a linear state space model and by choosing the best fitting one with respect to a performance index, that in the situation at hand is the scalar variance of the simulation error, that is, for model M

$$\text{fit}(M) = \text{trace}\left(E[e_u(t)e_u^T(t)]\right). \quad (2)$$

Of course we approximate $E[e_u(t)e_u^T(t)]$ with the sample variance of the fitting error, i.e. $e_u(t) = y(t) - y_u(t) = y(t) - C\,x_u(t)$, where $x_u(t)$ is the state obtained by setting $e(t) = 0$ and by using the identified initial condition $\hat{x}(0)$ in the model dynamics given by 1.

Clearly, an extensive research of every combination of inputs is not conceivable because the number of instances is very large and, even assuming a fixed cost for the identification algorithm, it can not be performed in reasonable time. So we adopted an iterative greedy approach to the selection problem. First of all, we divided our candidate inputs in three classes: external temperatures, external solar radiation, and internal temperatures. We start with an empty input set to be used for the identification. For each class of candidate inputs we choose the best input in term of the performance measure proposed above. Then, such input is removed from the candidate input set and placed into the input set selected for identification. Then, we either repeat the process to select another input from the remaining inputs in the same class or we pass to another class if we have already reached the desired number of inputs from that class. During this process we use the inputs which have been previously selected in the intermediate identification process. This selection algorithm is heuristic and the ordering for the class selection as well as the number of inputs per each class can be chosen by the user. Experimental evidences showed that the chosen subset is mostly independent from the class ordering. The results based on the previous identification methodologies on validation data set are shown in Figure 3 and Figure 4. In particular, in Figure 3, we compared the predictive performance of the two models identified with *N4SID* and $PBSID_{opt}$ algorithms by using the best input for each of the three classes. In order to have acceptable performance from *N4SID* implementation we had to force both the use of an high order model (namely at least as big as output length) and the matrix A to be asymptotically stable. Differently, the model identified by $PBSID_{opt}$ was always asymptotically stable and showed limited output error even with a low order model (namely about one fifth of output dimension). Besides these numerical issues, the $PBSID_{opt}$ identified a model that performed consistently better than the *N4SID* algorithm under different testing conditions, therefore we focus only on $PBSID_{opt}$ in the following experimental analysis. In Figure 4 we show the predictive performance of $PBSID_{opt}$ identified model as the number of inputs is increased. In the 2-input model we used the most informative external temperature sensor and internal ground temperature sensor, while in the 3-input model we added the most informative radiation sensor. It is interesting to mention that the input selection algorithm found that most informative external temperature sensor was the one placed on the south side and not on the north side as commonly suggested (see e.g. AA. VV. (1994)). This might be the result of the particular data-set we used, however most common sensor placement strategies are based more on experience rather than mathematical analysis, so that this aspect deserves more investigation.

4. PREDICTION ENHANCEMENT VIA KALMAN FILTERING

One limitation of the linear model previously identified is that it has unsatisfactory prediction performance. In fact, small environmental changes in the training data strongly affect the outcome of the identification algorithm, and even if the initial conditions can be perfectly estimated, the predicted output of the model becomes unreliable within few days.

One way to overcome this problem is to use of more sensors in the identification phase and longer training data sets. Another possible solution is to use few sensors to improve the prediction of the output for all the other sensors using a filtering approach. That is, we use the linear model previously identified to design a Kalman estimator with a small subset of the output vector $y(t)$. This filter estimates the state vector with a given confidence and then we obtain the full set of outputs using the original linear relation

$y(t) = C\,\hat{x}(t)$. In particular, we use a time varying Kalman filter, Kalman (1960), since it provides the best output estimate at any time-step.

Let us denote with $\hat{l} \in \{1, 2, \ldots, l\}$ the subset of outputs which will be used to recover the state in the Kalman filter. Let $M_{[r,c]}$ be the matrix obtained by choosing the rows in the set $r \in \{1, 2, \ldots, p\}$ and columns in the set $c \in \{1, 2, \ldots, q\}$ from matrix $M \in \mathbb{R}^{p \times q}$. Similarly, V_r will be the vector obtained by choosing entries in set r of the column vector V. The dot inside a square bracket, as in $M_{[r,\cdot]}$, indicates that all columns of the original matrix have been retained; a similar notation holds for the rows. Based on Equation 1, we build the reduced linear system

$$\begin{cases} x(t+1) &= A\,x(t) + B\,u(t) + K\,e(t) \\ y_{[\hat{l}]}(t) &= C_{[\hat{l},\cdot]}\,x(t) + e_{[\hat{l}]}(t) \end{cases} \tag{3}$$

with model error covariance matrix $Z_{[\hat{l},\hat{l}]}$, $Z = E[e(t)e^T(t)]$. The filter requires some initial tuning, in particular it is necessary to define the initial state estimate and error covariance. The former is obtained in a rather arbitrary way by using the Moore-Penrose matrix pseudo-inverse of C, $C^\dagger$, and the mean of the effective internal sensor readings at zero-time $y_{[\hat{l}]}(0)$. That is, if $\hat{n} = \dim(\hat{l})$ we have

$$\bar{y}_0 = \frac{1}{\hat{n}} \sum_{i \in \hat{l}} y_i(0)$$
$$\mu_0 = C^\dagger \left[\bar{y}_0 \ \ldots \ \bar{y}_0\right]^T.$$

The initial state covariance matrix P_0 was instead chosen by a trial and error procedure. Further improvements could be achieved by applying some whiteness tests. However, these tuning are only necessary to improve the estimate performance during the initial transient period, since the filter always converges to the same steady-state error independently of the initial conditions.

Again we face the problem of selecting a small subset of outputs from a large set, yet providing good predictive performance. That is, we have to select $\hat{l}$ in such a way that the global error on internal temperature estimates is reduced. First of all, we need to define a metric in order to compare different sensors choices. We adopted two natural metrics to evaluate the quality of sensors subsets. The first one is based on the discrete algebraic Riccati equation (DARE): starting from the solution of DARE we calculated the theoretical variance of the outputs

$$P(\hat{l}) = \bar{P} - \bar{P}\,C^T_{[\hat{l},\cdot]}\,(C_{[\hat{l},\cdot]}\,P\,C^T_{[\hat{l},\cdot]} + R)^{-1}\,C_{[\hat{l},\cdot]}\,\bar{P}$$
$$\Phi(\hat{l}) = C\,P(\hat{l})\,C'.$$

The cost function to minimize is then

$$J_d(\hat{l}) = \mathrm{tr}(\Phi(\hat{l})). \tag{4}$$

The second metric is based on the empirical error and the cost function is the sum of square deviation on each sensor at each time

$$J_f(\bar{l}) = \sum_t ||\hat{y}(t) - y(t)||_2. \tag{5}$$

Finding the optimal set of output sensors which minimizes one of the previous metrics is a combinatorial problem, therefore some efficient suboptimal strategy is required. In this work we explored three different strategies:

Greedy Search (GrS) This strategy splits the sensors into two sets: the selected sensor set $\mathcal{S}$ and the remaining ones $\mathcal{R}$. It starts with an empty set $\mathcal{S}$ and then sequentially finds among all the sensors in $\mathcal{R}$ the one that provides best performance when added to the set $\mathcal{S}$. Once this sensor is found, it is removed from $\mathcal{R}$ and placed in $\mathcal{S}$. The algorithm iteratively proceeds as above by finding the next best sensor from the set $\mathcal{R}$ till $\mathcal{S}$ has reached the desired sensor number.

Local Search (LS) This strategy, similarly to the previous one, splits the sensors into two sets: the selected sensor set $\mathcal{S}$ and the remaining ones $\mathcal{R}$. The algorithms start with a candidate sensor sets $\mathcal{S}$ chosen at random from $\mathcal{R}$ or obtained by running the GeS algorithm. Then it sequentially swaps one sensor between the two sets $\mathcal{R}$ and $\mathcal{S}$ and finds among all possible swapping combinations the one that leads to the best performance improvement. Then the best swapping is actually performed between the two sets and the process proceeds similarly avoiding to search previous swapping combination. This procedure is guaranteed to improve performance at every step and it stops when a local minimum is reached and a certain number of iterations have been performed.

Genetic Search (GeS) The previous strategy based on LS is likely to end up in local minima. To reduce this risk we adopted a genetic algorithm, see e.g. Goldberg (1989), to find good sensors swapping between the sets $\mathcal{R}$ and $\mathcal{S}$. In particular, it starts with several candidate sets $\mathcal{S}$, called populations, and then it swaps sensors among them (breeding and mutation) and the only the best performing new populations are likely to survive. This process is continued till no major improvements are observed or a certain number of iterations have been performed. Although these algorithms are based on heuristics and are not guaranteed to find the global minimum, they often lead to good performance.

We applied the three algorithms in sequence, i.e. we used the solution of GrS to initialize LS, and the solution of LS to initialize GeS. The optimal solution has been used to compare the performance between simulation based only on inputs $u(t)$ and prediction using the Kalman filter based on the outputs $y_{[\hat{l}]}$. The dynamical model used for both the simulation and the Kalman predictor was obtained using the $PBSID_{opt}$ algorithm with three inputs chosen based on the greedy algorithm described in the previous section. Figure 5 shows the real temperature of a sensor not included in the 5 outputs $y_{\hat{l}}$ used by the Kalman filter and the corresponding y_u and $\hat{y}$ given by the open-loop simulation based only on the input u and the kalman predictor, respectively. The improvements given by the Kalman filter are evident, in fact the use of only 5 sensors is sufficient to reconstruct the temperature of all 57 sensors with high precision.

Figures 6 (single sensor output) and 7 (mean square error) show that the prediction error improves as the number of sensors increases, in particular during the transient period. However, even with only two sensors the mean square error is smaller than half a degree during transient

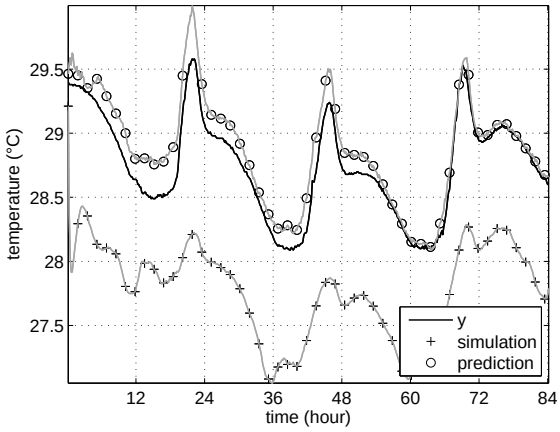

Fig. 5. Comparison between the true temperature y measured by one internal sensor not included in the output set $\hat{l}$, the corresponding value y_u simulated using the PBSID model, and $\hat{y}$ predicted by the Kalman filter using 5 outputs $\hat{l}$.

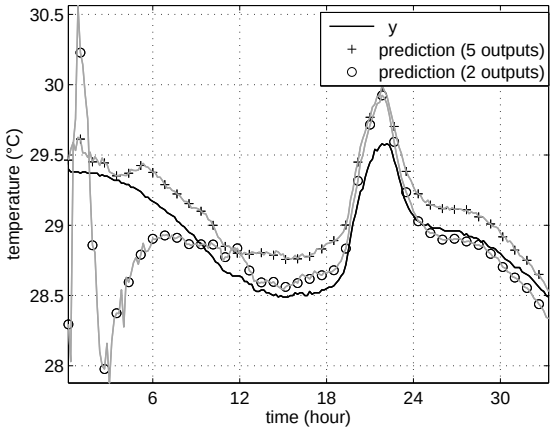

Fig. 6. Comparison of the true output y of a single sensor not included in the set $\hat{l}$, and the predicted output $\hat{y}$ obtained from Kalman filter based on 2 and 5 outputs $\hat{l}$.

period and smaller than a tenth of degree at steady-state. Of course, such a small error is also a consequence of the specific experimental conditions, i.e. high sensor density and unpopulated building. However, it suggests that the linear model identified by the $PBSID_{opt}$ algorithm is rather effective to describe the behavior of building thermodynamics, in particular when paired with Kalman filtering.

Finally, Figure 8 show the mean square error using the 3 best output sensors selected by the three strategies described above. The LS algorithm always offers great enhancements while the GeS presents some problems connected to parameters calibration. We also noticed that sometimes the application of a random starting solution to the local search produced better results if compared with the ones given by the solution of greedy algorithm.

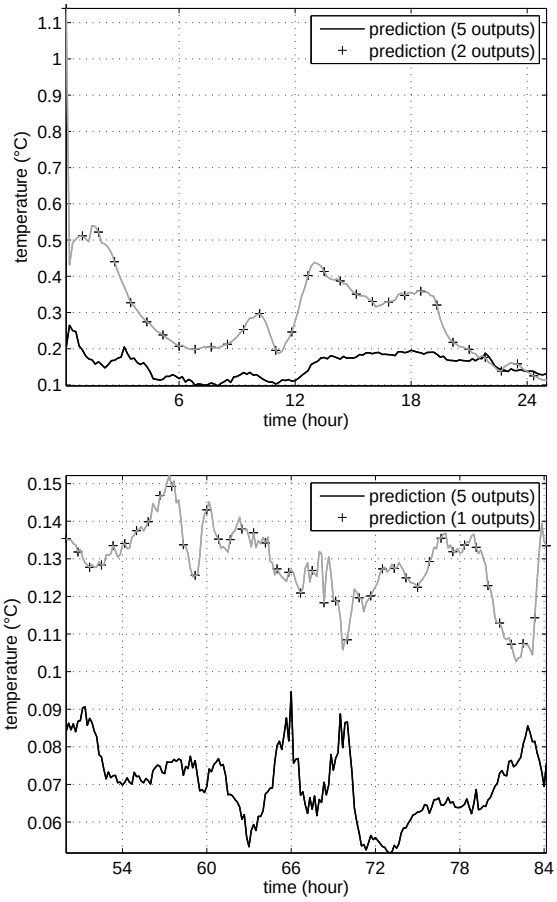

Fig. 7. Comparison of average prediction error for all sensors not included in the set $\hat{l}$ during transient period (*top panel*) and at steady-state (*bottom panel*) between the true output $y(t)$ and the predicted output $\hat{y}(t)$ using the Kalman-filter with 2 and 5 outputs.

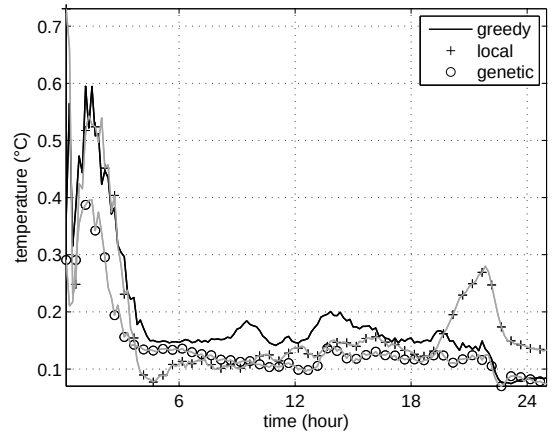

Fig. 8. Comparison on a single sensor of the output obtained from the models improved with Kalman filter with subsets selected by different heuristic algorithms.

The previous results are based on the metric $J_d(\hat{l})$, however we observed that there was substantial agreement with the empirical cost $J_f(\hat{l})$.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed the use of WSNs as a useful technology for identification of building thermodynamics. In fact, they provide the means to rapidly and inexpensively collect measurements from hundreds of temperature, humidity and light radiation sensors for long period of times. Although such a great wealth of data potentially provides fine grain information about building thermodynamics, it also poses novel challenging problems, in particular in terms of model identification. In fact, the number of inputs and outputs provided by WSNs are at least an order of magnitude larger than the number that can be handled by traditional identification tools for building thermodynamics (AA. VV. (1994)). In this paper we proposed to address this problem by adopting subspace identification tools which have been recently developed for identifying large scale MIMO systems. Indeed, we believe that identification of building thermodynamics can be a very useful testbed to evaluate and possibly improve subspace identification algorithms. In this work we tested a standard subspace method (the Matlab *N4SID*) and a recently developed subspace method (PBSID$_{opt}$, Chiuso (2007)) and we observed that the latter systematically outperformed the former, however a more detailed investigation is still required. Also we found that these tools do not perform well when the number of inputs and outputs are very large, mainly due to the fact that inputs and outputs are highly correlated. Therefore we proposed some heuristics for input and output selection for identification purposes, however we believe that more systematic and mathematically sound tools are necessary. We feel that systematic extension of principal component regression (PCR) and Partial Least Squares (PLS) can give significant improvements and hence will be subject of future research.

The simulation and prediction performance obtained by using only a properly chosen subset of inputs and outputs are remarkable. Although this is also a result of the particular experimental conditions, we believe that subspace identification techniques provide a viable and effective solution, and we are currently performing more realistic experiments.

REFERENCES

AA. VV. *System Identification applied to building performance data*. European Community, Luxembourg, 1994.

A. Chiuso. The role of Vector AutoRegressive modeling in subspace identification. *Automatica*, 43(6):1034–1048, June 2007.

R.H. Dodier and G.P. Henze. Statistical analysis of neural networks as applied to building energy prediction. *J. of Solar Energy Eng.*, 126(1):77–83, 2004.

David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

Edward Greenberg. Minimum variance properties of principal component regression. *Journal of the American Statistical Association*, 70(349):194–197, 1975.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 (1):35–45, 1960.

W.E. Larimore. Canonical variate analysis in identification, filtering, and adaptive control. In *Proc. 29th IEEE Conf. Decision & Control*, pages 596–604, Honolulu, 1990.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1997.

J.S. Qin. Partial least squares regression for recursive system identification. In *Proc. of IEEE CDC*, S. Antonio, Texas (USA), 1998.

M. van de Wal and B. de Jager. A review of methods for input/output selection. *Automatica*, 37(4):487–510, 2001.

P. Van Overschee and B. De Moor. N4SID: Subspace algorithms for the identification of combined deterministic–stochastic systems. *Automatica*, 30:75–93, 1994.

P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems*. Kluwer Academic Publications, 1996.

M. Verhaegen. Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. *Automatica*, 30:61–74, 1994.

H. Wold. *Research Papers in Statistics*, chapter Nonlinear estimation by iterative least squares procedures. Wiley, New York, 1966.