The role of data sample size and dimensionality in neural network based forecasting of building heating related variables

Martin Macas, Fabio Moretti, Alessandro Fonti, Andrea Giantomassi, Gabriele Comodi, Mauro Annunziato, Stefano Pizzuti, Alfredo Capra

Abstract

Energy consumed in buildings represents a challenge in the context of reduction of greenhouse gases emission. For this reason and due to the growing interest in operative costs reduction the energy used by buildings (tertiary and privates) for heating, ventilating, and air conditioning (HVAC) is even more investigated. Due to the nature of the energy consumption profile a predictive optimization method is one of the solution the scientific literature spreads even more. However optimization techniques need a good and reliable prediction of the variables of interest over a time horizon. This work focuses on methods to obtain a robust and reliable predictor based on Artificial Neural Networks. For the optimization purposes the neural model predicts total heating energy consumption (gas), internal air temperature and aggregated thermal discomfort 12 hours ahead. Training and testing data are simulated using a simulator based on Heat, Air and Moisture model for Building and Systems Evaluation (HAMBASE), by which a real office building was modeled. Influence of training data sample size and selection of predictor inputs is examined. Several combinations of early stopping condition and network complexity are tested for different training sample sizes. It is observed that the early stopping mechanism is crucial especially but not only for small training data, because it reliably overcomes overfitting problems. Surprisingly, relatively small networks were sufficient or performed best, although examined range of training sample covered up to five heating seasons. The use of a model tuning is thus supported by the results. Further, two strategies of selection of suitable input variables are demonstrated. While the input selection does not degrade the prediction performance, it is able to reduce the dimensionality and thus to save computational, communication, time, and data acquisition demands. The importance of inputs selection in HVAC modeling is thus pointed out and demonstrated.

Keywords: Heating, power consumption, temperature, thermal comfort, neural networks, prediction, feature selection, input selection

1. Introduction

Heating, ventilation, and air conditioning (HVAC) in buildings are the most important consumers of energy. In developed countries, buildings account for 20 – 40% of the total final energy consumption [1]. HVAC consumption accounts for half the energy use in buildings and one fifth of the total national energy use [1]. In Europe, buildings accounts for 40% of total energy use and 36% of total CO2 emissions [2]. Moreover, 46% of energy is used for heating and cooling, 33% is used for mobility reasons, and 21% is used for electrical power; also in the United States, HVAC systems constitute even over 50% of the building energy. Within the commercial sector, office and retail buildings are together the biggest energy consumers. All those examples suggest that the energy efficiency in building heating should be one of world’s priorities.

This article follows this goal by focusing on predictive models for one-step ahead forecasting of total gas consumption, internal temperatures, and thermal discomfort in one particular gas–heated office building. The modeling and prediction of heating related variables has been intensively studied for a long time. A very comprehensive review of modeling methods for HVAC systems can be found in [19]. The predictive models can be further used for an optimization of heating operation [20, 5]. In such scenario, a building management system can find the inputs of those predictive models that lead to proper values of their outputs. Such inputs can be future set points for indoor air temperature and supply water temperature. The optimization task can be further defined as constrained optimization or multi-objective optimization. An example of the former is the minimization of the predicted gas consumption subject to a pre-defined maximum level of the predicted thermal discomfort [9]. An example of the latter is the simultaneous minimization of predicted gas consumption and thermal discomfort and selection of one of multiple pareto-optimal solutions in terms of multi-objective optimization [21]. Another example can be a direct control of thermal discomfort by minimization of the difference between its predicted value and a reference [22]. In those approaches, it is crucial to reach a good prediction accuracy of consumption, thermal discomfort and temperature, because the control errors cannot be reduced under the limit im-
Table 1: Summary of selected studies that use a non-linear predictor. The abbreviations are: Feedforward neural network (FFN), Radial Basis Function (RBF), General Regression Neural network (GRNN), Non-linear Auto-regressive Network with Exogenous Inputs (NARX), Adaptive Neuro-Fuzzy Inference System (ANFIS), Echo State Network (ECHO) and Support Vector Regression (SVR).

| Reference          | Problem                     | Model  | Train/ 
|--------------------|-----------------------------|--------| test | Input selection | Model tuning | Early stopping |
| Dombayci 2010, [3] | Heating consumption         | FFN    | yes  | no   | no            | no            |
| Congradac 2012, [5]| Chiller states              | FFN    | no   | no   | yes           | no            |
| Yuce 2014, [7]     | El.&heat. consumption/Comfort | FFN   | yes  | no   | yes           | no            |
| Yokohama 2009, [8] | Electrical consumption      | FFN    | no   | yes  | no            | no            |
| Ferreira 2012, [9] | Temperature/Humidity        | RBF    | yes  | no   | yes           | yes           |
| Huang 2012, [10]   | Temperature                 | NARX   | yes  | yes  | yes           | yes           |
| Mustafaraj 2010, [12]| Temperature          | NARX   | yes  | yes  | yes           | no            |
| Deihimi 2013, [14]| El. consumption/temperature | ECHO  | yes  | no   | yes           | no            |
| Qiong 2009, [15]   | Cooling consumption         | RBF/FFN/GRNN | yes  | no   | yes           | no            |
| Jovanovic 2015, [16]| Heating consumption       | FFN/RBF/ANFIS | yes  | yes  | yes           | no            |
| Jain 2014, [17]    | Electrical consumption      | SVR    | yes  | no   | yes           | x             |
| Protic 2015, [18]  | Heating consumption         | SVR    | yes  | yes  | yes           | yes           |

posed by the prediction performance of the model the controller is based on.

An important novelty of this article is that the prediction of three different heating related variables in multiple zones is ex- amined (although some studies predict both the consumption and air temperature [4, 14]). In HVAC literature, the main fo- cus is naturally devoted to variables related to consumption (see [23] or [6] for literature survey). In some cases, however, it can be also required to predict internal temperature. Since the current internal temperature is one of the inputs of consumption and discomfort models, in case of multiple steps ahead predic- tion, the temperature must be predicted and fed back as an input to predict the next step. In another scenario, temperature model can be used for estimation of the time needed to pre-heat the building to some also important for many heating man- agement systems. Finally, thermal comfort is also forecasted here. In HVAC studies and artificially ventilated buildings, it is often assessed in terms of a Predicted Mean Vote (PMV) model [25]. A forecasting of PMV values is not so common, because it is difficult and expensive to measure training data.

Here, this expensiveness and lack of the data is overcome by the use of simulation. In real case, the thermal comfort can be assessed by questionnaire or some simplified measurements of thermal comfort [26]. It should be remarked that there are other variables that can be worth to predict. Examples of such vari- ables are air quality, relative humidity, electrical consump- tion [5], or occupancy and behavior of people [28, 29]. Those variables are however not focused in this article.

Three main types of HVAC predictive models exist - physi- cal–semi–physical and data–driven [23]. In this article, purely data–driven black-box models are considered. The focus on data–based statistical modeling technique can be motivated by some practical features of the heating management problem. Physical or semi–physical models of HVAC related variables can give precise predictions, but are highly parameterized and computationally expensive for control applications. Moreover, measured data are necessary also for those physical models and their calibration and validation. Therefore, the data-based pre- dictive models of HVAC related variables in buildings are of great importance for many HVAC control and energy manage- ment problems. A very representative example is the model predictive control, which optimizes forecasts of target variables over a certain time horizon and applies only the first time step of this horizon. With some exceptions [9], linear models are mostly used that make it possible to solve the optimization problem using fast quadratic programming methods. The pa- rameters of such model are usually estimated by system identi- fication methods (e.g. identification of parameters of state space equations). On the other hand, modeling of the energy variables using ANNs is a very popular approach especially in last few years, because of their ability to represent the non-linear dependencies. A review on the use of ANN in energy applications in buildings can be found in [30], where one can observe that common feed-forward networks (FFN) are the most popular mod- els. This has been also supported by Predictor Shootout con- tests organised by ASHRAE, where feed forward ANNs trained by the back propagation algorithm were the most efficient and effective models for energy prediction [31]. This trend is also followed here, only feedforward neural networks (multi–layer perceptrons) are considered. Another very common method is the non-linear auto regression with exogenous inputs [11], which can be understood as a recurrent neural network. The recurrent neural networks, although used less often than FFNs, are also popular (e.g. [32, 33, 34]). Other much less frequent approaches are Artificial Neural Fuzzy Interface System [35], combination of multiple ANNs [36], pseudo-dynamic transi- tional characteristics used as one of inputs for a feed-forward
However, the non-linear character of those models implies a need of more complex, often population-based optimization heuristics [38, 5]. Those optimization methods repeatedly run the model in a loop, which causes high computational demands that can make the real-world implementation impossible. This is a very strong motivation to make the model as simple and fast as possible, this goal can be partly reached by using techniques like input selection or model tuning. The reduction of complexity is however not the only potential benefit of those approaches, those techniques can also improve the generalization abilities and performance of the final predictive models. The model tuning is related to the fact that an optimal predictive model depends on the available training data, especially the sample size (number of data instances) and dimensionality (number of model inputs) are easily observable properties. While the first can be hardly changed without measuring more data and implies a need for model tuning, the second can be reduced by input selection methods.

To motivate our research by a literature evidence, we summarized some examples of recent studies dealing with prediction of HVAC related variables in Table 1. Because of the limited space, the table does not comprehensively summarize all the research, but only those similar to our approach with iteratively trained predictors. The table examines if a division of the data into test/train set is described (column 4), if a selection of input variables was performed (column 5), if the predictor was tuned by some means (column 6) and if an early stopping was used (column 7).

Although a vast majority of studies clearly describes a division of available data into training and testing set and time periods spanned by those data sets (column 4 of Table 1), most of those studies avoid discussing of the satisfactory sample size and mostly satisfy with momentary accessible data. The size of the training and testing data thus spans different values from a couple of days [10, 6], months [17, 18] to several years [3, 14]. We found only one study that examines an impact of the sample size on forecasting of heating related variables in [9]. Therefore, the influence of sample size on the prediction performance is focused in this article. Moreover, this article shows a strong dependence between data sample size and feasible model complexity (number of hidden neurons). A direct consequence of this is that a model tuning should be always performed to reach a good predictive performance. As can be seen from Table 1, some studies do not perform (or do not describe) model tuning and simply guess a suitable model complexity. Moreover, the methodology of model tuning is often not correct or sufficiently described. Some models are tuned on the testing data, which usually gives biased estimate of the final performance. This occurs in [3] or [4], where different numbers of neurons are compared on testing data, but the results of comparison are not further validated on independent data. In [5], five different model architectures are compared, but it is not clearly explained what is the testing set.

From survey described in [30, 6] it appears that the most common inputs of heating consumption prediction studies are internal and external temperature, solar radiation, relative humidity, wind speed, occupancy and current time. Considering that also lagged versions of those variables are used and that for multiple zones some of those variables are vectors, the number of inputs can become too high with respect to available sample size, which can cause overfitting. Therefore we believe that an input selection can bring significant benefits. Column 5 of Table 1 shows that there is a small number of studies that deal with a good choice of input variables. In [6], a very simple selection of inputs is performed in terms of comparison of several ANNs with different inputs on validation data. In [39], statistical tests are used to decide, which inputs can be removed. In [16], well-known sequential forward selection procedure is adopted. In [11], differential entropy was used to select lags for particular inputs, but the inputs themselves were not selected. In [12], an advanced optimal brain surgeon pruning was used for selection of inputs and pruning of hidden neurons.

For all those reasons, the role of the training data sample size (number of instances), input dimensionality of the model (number of input variables) and early stopping procedure is examined with some important conclusions. Namely, it is shown that suitable model complexity is crucial for good predictive performance and its choice strongly depends on the available sample size of the training data. Model tuning should be therefore always used to avoid unnecessary complexity. Moreover, it is shown that the model complexity can be also cheaply but significantly reduced by input selection strategies without a serious loss of performance. The results should motivate the HVAC researchers to devote higher effort to those two aspects.

The rest of this article is organized as follows. In section 2, the examined building and its simulation are described. Section 3 summarizes all the variables and their notation. Section 4 is focused on ANN model, its training procedure, parameters, and also on methods of input selection and early stopping. All empirical results are summarized in section 5. The article is concluded in section 6 by a comprehensive discussion and an important future work summary is provided in section 7.
2. F40 building simulation

A real office building located at Casaccia Research Centre of Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA) was considered as a case study. The building is depicted in Figure 1. Its structure is composed of three floors and a thermal plant in the basement. There are 41 office rooms of different size with a floor area ranging from 14 to 36 m², two electronic data processing rooms each of about 20 m², four laboratories, one control room and two meeting rooms. Each office room has from one up to two occupants. For each room and laboratory, as thermal exchangers, there are fan-coils with on-off fan speed controlled by a proper thermostat with hysteresis equal to 1°C. A detailed plan of its ground floor with dispositions of rooms, fan-coils and some dimension lines can be found on Figure 2.

The thermal plant is composed of a natural gas boiler (winter) and of three electric compressor chillers (summer). This study is related only to heating during winter season. Furthermore, the building is equipped with an advanced monitoring system collecting data from sensors of environmental conditions and electrical and thermal energy consumption. Despite the fact that some real building data are already available, simulated data are used here that enable us to reliably estimate the true prediction error using a testing data. The use of simulated data, preferred in many studies [5, 10, 20], is not unusual if a study is more oriented on methods rather than on their application.

To obtain sufficient training and testing data, a Matlab Simulink simulator based on Heat, Air and Moisture model for Building and Systems Evaluation (HAMBASE) ([40], [41]) was developed. In particular, the building was divided into 15 different zones according to different thermal behavior depending on solar radiation exposure. A zone consists of a group of rooms with similar climatic conditions and the same climate control policy. Figure 3 shows the division of the floors into zones. As consequence of this assumption, thermostat behaviour in all rooms within one zone was assumed equal. Although there are 15 zones at all, five of them are not considered. Three zones (white) correspond to corridors and do not have sensors and two zones (gray) are meeting halls that do not have sensors and remotely controllable fan-coils.

In the simulation, the gas consumption is derived by proper integration of the natural gas mass flow which depends directly from the discharge and return water temperature at the thermal plant and from the thermal plant efficiency. The fan-coils are modeled by the ε-Number of Transfer Units (ε-NTU) method.
[42] which allows to derive the heat injected in the zones and the outlet water temperatures from known zone air temperatures, fan-coil inlet water flows and fan speeds. The inputs of the simulation are indoor temperature set points and external meteorological data. Based on those inputs, the simulator is used to generate output data (radiant temperature, air temperature, relative humidity, thermal consumption) from which the thermal comfort can be evaluated. The casual heat gains due to people’s presence and activity (mostly work on personal computer) were modeled as static daily pattern obtained by averaging available real daily occupancy profiles.

To evaluate thermal comfort inside particular zones, we used Predicted Mean Vote model (PMV). The model was developed by P. O. Fanger [25] using heat balance equations and empirical studies about skin temperature to define comfort. The PMV model has seven inputs summarized in Table 3 and two output variables. The first output, not used here, is PMV value. It is a continuous variable between −3 and +3. Its ideal value is zero. Positive PMV means too hot environment and the negative PMV corresponds to cold environment. The comfort zone is for PMV between −0.5 and +0.5. An alternative output of the PMV model is Predicted Percentage of Dissatisfied (PPD), which is always greater than 5% and its value 10% corresponds to the comfort zone. The PPD value was used here as the discomfort assessment. In our experiments we used implementation from Technical University of Eindhoven, Netherland [43].

The PMV model uses building simulation outputs as inputs. All the inputs needed for PPD evaluation are summarized in Table 3. For each zone, the building simulation output includes radiant temperature, air temperature and relative humidity. In additional to those taken from the simulation, there are other four variables - metabolism, external work, clothing, and air velocity. Since an office building is considered, those parameters can be understood as more or less constant (people are dressing similarly, do not change the dress during the day and are mostly working on a computer). The constant values were set heuristically by examining the typical behavior of occupants. This simplification is used, because a sufficient dynamic model of occupants and their behavior is not available yet. Its detailed examination is however out of scope of this paper and will be solved in a future work. It should be stressed that some simplifications of classical PMV model also exist. For example, in [44], authors present a model, which, in moderate environment, can estimate PPD only from temperature and relative humidity. Since the real building presented here has no sensors of the radiant temperature, this approach can be very useful also for our real building.

### 3. Variables and problem definition

Further, the variables and prediction tasks are described more formally. The prediction of the variables is performed every 12 hours, because the future control decision will be made in the morning or in the evening for whole daytime (6:00-18:00) or

<table>
<thead>
<tr>
<th>Number</th>
<th>input</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S_A(t+12)$ Air temperature set point in zones [$^\circ C$]</td>
</tr>
<tr>
<td>2</td>
<td>$S_W(t+12)$ Supply water temperature set point [$^\circ C$]</td>
</tr>
<tr>
<td>3</td>
<td>$W_1(t)$ Diffuse solar radiation [Wm$^{-2}$]</td>
</tr>
<tr>
<td>4</td>
<td>$W_2(t)$ Exterior air temperature [$^\circ C$]</td>
</tr>
<tr>
<td>5</td>
<td>$W_3(t)$ Direct solar radiation [Wm$^{-2}$]</td>
</tr>
<tr>
<td>6</td>
<td>$W_4(t)$ Cloud cover (1..8)</td>
</tr>
<tr>
<td>7</td>
<td>$W_5(t)$ Relative humidity outside [%]</td>
</tr>
<tr>
<td>8</td>
<td>$W_6(t)$ Wind velocity [ms$^{-1}$]</td>
</tr>
<tr>
<td>9</td>
<td>$W_7(t)$ Wind direction [degrees from north]</td>
</tr>
<tr>
<td>10</td>
<td>$T_1(t)$ Air temperature in zone 1 [$^\circ C$]</td>
</tr>
<tr>
<td>11</td>
<td>$T_2(t)$ Air temperature in zone 2 [$^\circ C$]</td>
</tr>
<tr>
<td>12</td>
<td>$T_3(t)$ Air temperature in zone 3 [$^\circ C$]</td>
</tr>
<tr>
<td>13</td>
<td>$T_4(t)$ Air temperature in zone 4 [$^\circ C$]</td>
</tr>
<tr>
<td>14</td>
<td>$T_5(t)$ Air temperature in zone 5 [$^\circ C$]</td>
</tr>
<tr>
<td>15</td>
<td>$T_6(t)$ Air temperature in zone 6 [$^\circ C$]</td>
</tr>
<tr>
<td>16</td>
<td>$T_7(t)$ Air temperature in zone 7 [$^\circ C$]</td>
</tr>
<tr>
<td>17</td>
<td>$T_8(t)$ Air temperature in zone 8 [$^\circ C$]</td>
</tr>
<tr>
<td>18</td>
<td>$T_9(t)$ Air temperature in zone 9 [$^\circ C$]</td>
</tr>
<tr>
<td>19</td>
<td>$T_{10}(t)$ Air temperature in zone 10 [$^\circ C$]</td>
</tr>
</tbody>
</table>

### Table 2: The description of input variables

<table>
<thead>
<tr>
<th>Number</th>
<th>input</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Metabolism [Wm$^{-2}$] From simulation</td>
</tr>
<tr>
<td>2</td>
<td>External work [Wm$^{-2}$] From simulation</td>
</tr>
<tr>
<td>3</td>
<td>Radiant temperature [$^\circ C$] From simulation</td>
</tr>
<tr>
<td>4</td>
<td>Air temperature [$^\circ C$] From simulation</td>
</tr>
<tr>
<td>5</td>
<td>Relative humidity [-] From simulation</td>
</tr>
<tr>
<td>6</td>
<td>Clothing [-] 1</td>
</tr>
<tr>
<td>7</td>
<td>Air velocity [ms$^{-1}$] 0.1</td>
</tr>
</tbody>
</table>

Table 3: The description of input variables. Only three input parameters were taken from the simulation output. For simplicity, the other four were set to their typical values.

---

1(see standards ISO EN 7730, CR 1752 or ASHRAE 55)
night time period (18:00-6:00). In this setting, one-step ahead prediction corresponds to 12 hours-ahead prediction. The division of the time axis into nighttime and daytime intervals for one heating season with 68 days is shown in Fig. 4. All the important variables are established in the following:

- Let \( S_W(t) \) be the supply water temperature set point and \( S_A(t) \) be the air temperature set point for all zones that is held constant within all 12 hours preceding to hour \( t \). For simplicity, we consider \( S_A(t) \) to be changed every 12 hours.

- Let \( t_z(t) \) be the air temperature inside the zone \( z \) measured at the end of hour \( t \).

- Let \( p_z(t) \) be the PPD value inside the zone \( z \) at the end of hour \( t \) computed from values of zone air temperature, relative humidity and radiant temperature.

- Let \( c(t) \) be the hourly total consumption measured at the end of hour \( t \).

- Let \( w(t) = [w_1(t), w_2(t), \ldots, w_7(t)] \) be the vector of 7 weather measurements acquired at the end of hour \( t \). The meaning of particular weather variables can be found in Table 2.

- Let

\[
T_i(t) = \frac{\sum_{i=t-12}^{t} t_i(t)}{12} \tag{1}
\]

be the average of the hourly values of air temperatures \( t_i(t) \) over 12 hours preceding to hour \( t \).

- Let

\[
W_j(t) = \frac{\sum_{i=t-12}^{t} w_j(i)}{12} \tag{2}
\]

be the average of the hourly values of weather variable \( w_j(t) \) over 12 hours preceding to hour \( t \).

- Let

\[
C(t) = \sum_{i=t-12}^{t} c(i) \tag{3}
\]

denotes the total gas consumption of the building in last 12 hours preceding to hour \( t \).

- Let

\[
P_z(t) = \sqrt{\frac{1}{12} \sum_{i=t-12}^{t} p_z(i)^2} \tag{4}
\]

be the quadratic mean of PPD values \( p_z(t) \) computed over all hourly values measured within last 12 hours.

The main task is to predict:

- total gas consumption in the following 12 hours long period \( C(t + 12) \),

- air temperature in each zone \( z \) 12 hours ahead \( t_z(t + 12) \),

- quadratic mean \( P_z(t + 12) \) of thermal comfort computed from the following 12 hours long period, given

- the constant set points for the following 12 hours \( S_A(t+12) \) and \( S_w(t + 12) \),

- averages \( T_z(t) \) of last 12 hourly values of air temperature for each zone \( (z = 1 \ldots 10) \)

- averages \( W_j(t) \) of last 12 hourly values of the weather variables \( (j = 1 \ldots 7) \).

4. Neural network model

The inputs and outputs of all models are depicted on Figure 5. In all the underlying experiments, feed-forward neural network with one hidden layer was used. Levenberg-Marquardt training algorithm was adopted, because of its relatively high speed, popularity and because it is highly recommended as a first-choice supervised algorithm, although it requires more memory than other algorithms [45]. The network was simulated in Neural Network Toolbox for Matlab. The hidden and output neurons with sigmoid and linear transfer function, respectively, were used. The mean square error was minimized by training procedure. The training was stopped after the number of training epochs exceeded 300 or if the error gradient reached 10^{-7}. If a "stop-test" data based early stopping was used, one more stopping condition was the achievement of 100 epochs without any improvement of mean square error computed on stop-test dataset. Although the training was based on the mean square error, the performance measure of the experiments was the mean absolute percentage error (MAPE), which is, together with coefficient of determination (R^2), most typical in prediction and forecasting experiments. MAPE is defined as:

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{5}
\]

where \( n \) is the number of data instances, \( y_i \) is the true (actual) output value and \( \hat{y}_i \) is the predicted output value for data instance \( i \).

Stop-test data are commonly called validation data. Here we adopt the term "stop-test" to avoid confusion with validation data that are used to decide about number of inputs.
As mentioned above, the resulting network can be further used as a data-based black box model for optimization of the building heating. Each 12 hours, a building management system finds the air and supply water temperature set points that correspond to some optimal behavior (e.g. lead to a minimum output of the consumption model and a proper level of modeled thermal comfort, keeps a reference value of thermal comfort). Since the cost function is highly nonlinear and multi-modal, population based metaheuristical optimization can be used with advantage. In such a case however, the neural network is used in optimization loop many times. It is thus crucial to keep the network as small and simple as possible. Moreover, if the optimization is performed remotely, one must minimize the amount of data that must be measured and transferred from the building to the optimization agent. All these requirements imply a critical need for a proper selection of inputs, which leads to reasonable data acquisition requirements and proper prediction accuracy.

Because different ANNs can significantly differ in their dynamics, input selection should be adapted for particular network in terms of well known wrapper approach. To select proper inputs tailored for particular network, we decided to use a sensitivity based method developed by Moody [46]. It is called Sensitivity based Pruning (SBP) algorithm. It evaluates a change in training mean square error (MSE) that would be obtained if jth input’s influence was removed from the network. The removal of influence of input is simply modeled by replacing it by its average value. Let \( x_j(t) = (x_{1j}, \ldots, x_{ij}, \ldots, x_{Dj}) \) be the jth of N instances of the input vector (N is the size of the training data set). Let \( x'_j(t) = (x_{1j}, \ldots, \sum_j x_{ij}/N, \ldots, x_{Dj}) \) be jth instance modified at jth position. For each data instance \( j \), partial sensitivity is defined by

\[
S_{ij} = (f(x'_j) - y_i)^2 - (f(x_j) - y_i)^2, \tag{5}
\]

where \( f \) is the neural network function and \( y_i \) is the target value for ith data instance. Further, the sensitivity of the network to variable \( i \) is defined as:

\[
S_i = \frac{\sum_j S_{ij}}{N} \tag{6}
\]

In our implementation of SBP, the algorithm starts with the full set of inputs (\( D = 19 \)). At each step, a target neural network is trained. Further, its sensitivity is computed for particular inputs and the input, for which the sensitivity is the smallest one is removed from the data. Note, that a new neural network is trained at each backward step.

An obvious question is, how many inputs to select. To answer this question, we test the neural networks with different number of inputs on an independent validation data set and select the proper number of inputs according to the validation MAPE. The final errors of the methods are estimated on another independent testing data set, which is not used in any part of the predictor design process. This process of data res-sampling is described in more details in section 5.4.

4.2. Early Stopping

The training procedure of a neural network usually minimizes an error measure computed on a training data. The stopping condition of such training algorithm can be the accomplishment of a predefined threshold of training error, its gradient, maximum number of epochs, iterations or seconds, which does not avoids a potential overfitting. A possible solution of this problem is to split the training data to a real training data set and a special stop-test data set and stop the training if the error on the latter one reaches a minimum and starts to increase. This can prevent a possible over-fitting. A possible disadvantage is the reduction of the size of the original training set by the stop-test data size. However, in practical applications, one must ensure that the neural model will work reasonably and thus, particularly for smaller sample sizes, this early stopping procedure can be necessary. Here, whenever the early stopping procedure was used, the stop-test set consisted of the last 25% of the original predictor design instances. Note that a whole block of data was used as the stop-test rather than randomly selected instances. This is more fair since a random distribution could add an optimistic bias by using temporary close instances both for training and stop-test. The stop-test dataset based early stopping was used in [6]. Here, the role of the early stopping procedure is also examined.

5. Experimental work

This section describes the empirical analysis, whose aim is to demonstrate the importance of predictor tuning and input variable selection. First we define the data sets and experimental
settings. Next, two experiments and their results are described showing the impact of sample size and importance of input selection.

5.1. Data

The behavior of supply water temperature set point was controlled by a simple weather compensation rule. To excite the dynamics of the system in a proper degree, we also added a random component. A similar randomized approach was used for example in [9]. The value of the supply water temperature set point \( S_w(t) \) is Gaussian random number with standard deviation \( 4 \)°C and mean equal to \( 85 - 2T_e \), where \( T_e \) is the mean of previous day external temperature. The behavior of air temperature set points \( S_A(t) \) differs for daytime and nighttime hours. Between 6:00 and 18:00, they are also Gaussian random numbers with mean 21.5° C and standard deviation 1° C, which guarantees an acceptable level of thermal comfort. Moreover, there is a saturation under 20° C and above 25° C. Between 18:00 and 6:00, there is a nighttime regime and air temperature set points are Gaussian random numbers with mean 20° C, standard deviation 1° C and upper saturation level 23° C.

5.2. Experimental settings

To ensure the validity of the results, we simulated 9 heating seasons covering years 2004-2013. The model uses real weather data measured in Rome area during the examined years. Each data instance describes one 12 hours-long time period and consists of 19 input variables (see Table 2 for the description of the inputs) and one output variable (gas consumption, temperature or quadratic mean of PPD) computed using aggregations described above in section 3. As can be seen from Fig. 4, each heating season corresponds to 68 days, which gives 134 data instances. The total number of data instances for all 9 heating seasons is \( 9 \times 134 = 1206 \).

All simulated data were further split into smaller training, testing and validation subsets used for error estimation needed in different parts of experiments. Since this data handling is crucial for validity of the experiments and must be clearly described, it is graphically summarized in Figure 6 for the two particular experiments with training sample size (6a) and experiments with input selection (6b).

As can be seen from the figure, for both types of experiment, the testing data (data set B) covers four heating seasons 2009-2013 corresponding to \( 4 \times 134 = 536 \) samples. The testing data set is used only for evaluation purposes and is never used in the predictor design process.
5.3. Experiments with training sample size

In this section, the dependence between the training sample size and a proper network complexity is shown, supporting the importance of a model tuning. Further, the influence of the early stopping on the testing results is examined and pointed out.

5.3.1. Experiment description

The splitting of data into different subsets for predictor design and testing procedures is depicted in Figure 6a. To obtain several predictor design data of different sample sizes, we created data sets by taking only part of the first 670 samples. In this way we obtained several predictor design data sets A containing 134, 201, 268, 335, 402, 469, 536, 603, 670 instances. Further, we tested each combination of sample size, number of hidden units and early stopping (see Table 4) by repeating 100 runs of the predictor design and testing procedure. For testing, the same data set B containing block of 536 data instances 671-1206 was used in all experiments. The neural network described in section 4 was used.

Since the predictor training is a stochastic procedure, it may happen that some runs lead to a local optimum or overfit unacceptably. Typically, those unacceptable results appear very rarely, but can give very poor testing performance. In practical case, it is important to avoid those extreme situations, because such a network can lead the subsequent predictive optimization to an uncomfortable or energy inefficient control. This has been taken into account in the evaluation of results. First evaluation criterion used here is median of MAPE error. The median is highly robust to extreme values of the failed runs. Therefore, apart of the median MAPE, we also measure the reliability of the network as a percentage of the failed runs with an unacceptably high testing error. The acceptance thresholds on MAPE results was subjectively chosen as 12% for all three predicted variables by visually inspecting some representative results.
5.3.2. Importance of model tuning

All the results are depicted in Figures 7, 8 and 9. The upper subfigures depict the median MAPE and lower subfigures show the percentage of failed runs. The left-hand subfigures correspond to the case when early stopping procedure is not used, i.e. the stop-test set is empty and all predictor design samples are used solely for training. The right hand subfigures correspond to the early stopping procedure using 25% of the data for the testing of stopping condition.

One can observe very similar phenomena in all Figures 7, 8 and 9. First, when the data covers less than one heating season for the building (i.e. 68 days corresponding to 134 instances), it is best to choose only one hidden unit. Such an extremely small network can be a reasonable choice also for larger data sets depending on the predicted variable. However, as the sample size increases, the network becomes too simple to encode some more complex dependencies and is significantly outperformed by bigger networks regardless of whether the early stopping was used or not. This strong dependence of suitable model complexity on the available sample size implies that one should always use a model tuning method.

5.3.3. Importance of early stopping

The influence of early stopping for consumption prediction can be examined by comparing left and right sub-figure of Figure 7. For networks with more than three units and especially for the network with ten units, one can observe a clearly visible and statistically significant impact of the early stopping procedure. Simultaneously, there is relatively small difference between curves for 3, 5 and 10 units if the early stopping is applied. For such case, the training algorithm prevents overfitting, which makes the differences between 3, 5 and 10 units insignificant. If the early stopping is not used, it is unacceptable to use 10 hidden units if we have less than 200 days. This result demonstrates that in the examined scenario, the early stopping procedure is necessary to overcome the over-fitting problems. On the other hand, creation of the stop-test set reduces the training data size, which can increase the median MAPE, as it is for consumption prediction with 3 hidden units and big training data (comparison of blue line with triangle marker on the left-hand and right-hand side on Figure 7). A similar observation can be made from the statistical point of view. From the practical point of view, the difference in medians is much less than 1%, which most probably does not affect significantly the utilization of the network in subsequent optimization.

The results for the prediction of indoor temperature $T(t + 12)$ in zone 7 are depicted in Figure 8 and are different to the consumption prediction. First of all, it seems reasonable to use hidden neurons with early stopping for any size of the data set. The model with 1 hidden unit is not sufficient enough to represent the association between input and output variables even for small sample size. With the early stopping procedure, there is no occurrence of unacceptable models. Both, the median of MAPE error and ratio of failed runs are very small compared to consumption prediction. The temperature prediction seems to be an easier task. This is a very important point, because the internal temperature is an important input of all of our predictors, which is a common case in many heating related articles (e.g. [4] or [14]).

Finally, the prediction of zone 7 discomfort $P_d(t + 12)$ is depicted on Figure 9. One can see that the MAPE is relatively high compared to consumption and temperature, the discomfort prediction seems to be the most difficult task and probably cannot be used in any predictive optimization of building operation. Its improvement will be focused in future work.

5.4. Experiments with input selection

In this section, the results of input selection are presented. First, the experimental settings are described and two strategies of choosing the final input dimensionality are proposed. Next, the results are described and finally summarized.

5.4.1. Experiment description

A detailed description of data handling in this experiment is depicted in Figure 6b. The predictor design data set $A$ is used in a little bit more tricky way, because the input selection is performed before the training procedure. The input selection consists of two parts. First, the predictor design data set $A$ is split into two blocks $A_1$ and $A_2$, which consist of 75% and 25% blocks of $A$. Data set $A_1$ is used by the backward elimination, which sorts the input variable according to the order of the elimination and creates subsets of input variables of size $D, D - 1, \ldots, 1$.

Second, those $D$ subsets of input variables are evaluated by training predictors on $A_1$ and testing on dimensionality decision data set $A_2$, which gives $D$ error values. The final number of selected inputs is decided according to an input dimensionality decision strategy:

- **Strategy 1** is to choose the subset that leads to minimum of $D$ errors computed on dimensionality decision data set $A_2$. This strategy is focused on performance maximization.
- **Strategy 2** is to compute median value of the $D$ errors and choose the smallest subset, whose error is less or equal to this median. This strategy is focused more on dimensionality reduction.

After sorting the inputs and deciding, how many of them to choose, both subsets $A_1$ and $A_2$ are merged back to $A$ and a new predictor with selected inputs is trained on whole predictor design data set $A$ and tested on data set $B$.

Simultaneously, both types of the training procedures - the one inside the input selection (performed on data set $A_1$) and the final training procedure (performed on data set $A$) use the early stopping method (split the data into training and stop-test subsets). Within the input variable selection loop, the use of early stopping is even more important, because the neural network is trained $D$-times and the probabilities of a failed run sum up. Such a failed run can obviously absolutely degrade the input variable selection by eliminating a very important variable.
<table>
<thead>
<tr>
<th></th>
<th>Strategy 1</th>
<th>Strategy 2</th>
<th>Strategy 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>c</td>
<td>6.2</td>
<td>18</td>
<td>7.4</td>
</tr>
<tr>
<td>t1</td>
<td>0.9</td>
<td>9</td>
<td>0.9</td>
</tr>
<tr>
<td>t2</td>
<td>1.0</td>
<td>16</td>
<td>1.0</td>
</tr>
<tr>
<td>t3</td>
<td>1.2</td>
<td>11</td>
<td>1.3</td>
</tr>
<tr>
<td>t4</td>
<td>1.5</td>
<td>9</td>
<td>1.6</td>
</tr>
<tr>
<td>t5</td>
<td>0.9</td>
<td>9</td>
<td>1.0</td>
</tr>
<tr>
<td>t6</td>
<td>0.9</td>
<td>14</td>
<td>0.9</td>
</tr>
<tr>
<td>t7</td>
<td>1.2</td>
<td>11</td>
<td>1.3</td>
</tr>
<tr>
<td>t8</td>
<td>0.9</td>
<td>11</td>
<td>1.0</td>
</tr>
<tr>
<td>t9</td>
<td>0.9</td>
<td>15</td>
<td>1.0</td>
</tr>
<tr>
<td>t10</td>
<td>1.2</td>
<td>6</td>
<td>1.2</td>
</tr>
</tbody>
</table>

(a) 202 samples

<table>
<thead>
<tr>
<th></th>
<th>Strategy 1</th>
<th>Strategy 2</th>
<th>Strategy 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>c</td>
<td>5.05</td>
<td>14</td>
<td>5.54</td>
</tr>
<tr>
<td>t1</td>
<td>0.9</td>
<td>14</td>
<td>0.9</td>
</tr>
<tr>
<td>t2</td>
<td>1.0</td>
<td>16</td>
<td>1.0</td>
</tr>
<tr>
<td>t3</td>
<td>1.2</td>
<td>11</td>
<td>1.3</td>
</tr>
<tr>
<td>t4</td>
<td>1.5</td>
<td>9</td>
<td>1.6</td>
</tr>
<tr>
<td>t5</td>
<td>0.9</td>
<td>14</td>
<td>0.9</td>
</tr>
<tr>
<td>t6</td>
<td>0.9</td>
<td>14</td>
<td>0.9</td>
</tr>
<tr>
<td>t7</td>
<td>1.2</td>
<td>11</td>
<td>1.3</td>
</tr>
<tr>
<td>t8</td>
<td>0.9</td>
<td>11</td>
<td>1.0</td>
</tr>
<tr>
<td>t9</td>
<td>0.9</td>
<td>15</td>
<td>1.0</td>
</tr>
<tr>
<td>t10</td>
<td>1.2</td>
<td>6</td>
<td>1.2</td>
</tr>
</tbody>
</table>

(b) 403 samples.

<table>
<thead>
<tr>
<th></th>
<th>Strategy 1</th>
<th>Strategy 2</th>
<th>Strategy 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>c</td>
<td>220</td>
<td>403</td>
<td>604</td>
</tr>
<tr>
<td>Consumption</td>
<td>-0.1</td>
<td>-0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Discomfort</td>
<td>0.8</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

(c) 604 samples.

Table 5: Influence of input selection on MAPE and dimensionality

Table 6: Impact of input selection on MAPE and dimensionality averaged over zones
Therefore, the early stopping is always used in those input selection experiments.

The results of input subset selection are depicted in Figure 10. On the horizontal axis, the number of selected inputs is depicted. On the vertical axis, median of MAPE error computed over 100 runs of the experiment is shown. The particular number of inputs is decided according to the value of validation error (dotted line) computed on data set \( A_2 \). The value of the testing error (solid line) corresponds to the predictor trained on whole data set \( A \) and tested on data set \( B \). From all sub-figures of Figure 10, one can observe that the curves of testing and validation error are very similar, which proves that the validation error is a reasonable approximation of the testing error.

The vertical shift of those curves is caused by the fact that validation error is computed for models trained on smaller data set \( A_1 \).

An important phenomenon, which appears for all studied variables is high potential of dimensionality reduction. One can see that a visually recognizable disruption of the testing MAPE appears if less than 5 inputs are used for all predicted variables.

As mentioned above, two possible strategies of dimensionality selection (decision how many inputs to choose) are examined. The strategies are compared quantitatively in Tables 5a, 5b and 5c for the size of training data equal to 202, 403 and 604 samples, respectively. The tables report median values of testing MAPEs and chosen dimensionality. For lack of space and better and better clearness, standard deviation are not reported in Table 5.

Strategy 3 represents a trivial strategy of selecting all 19 inputs. The influence of the input selection on the MAPE performance can be expressed in terms of comparison of input selection strategy 1 or 2 to the trivial reference strategy 3 that selects all inputs. The cells that correspond to statistically significant difference from strategy 3 are colored. The input selection strategy 3 leads to statistically significant MAPE increase is depicted in red. The improvement in terms of statistically significant MAPE reduction is depicted in blue color.

### 5.4.2. Results for dimensionality decision strategy 1

The strategy 1 is focused on minimization of the validation error. First, the small-sized training sample case can be examined in Table 5a.

It can be seen that for forecasting of consumption and temperatures, this strategy mostly outperforms strategy 3. This difference is often statistically significant, although from practical point of view, the difference is always less than 1%. This show that the input selection using strategy 1 can bring significant benefits for small training data in terms of both the dimensionality reduction and MAPE performance. The situation is slightly different for discomfort forecasting, where prediction is much inaccurate. The strategy 1 often worsens the prediction, although it leads to a maximum absolute degradation of MAPE equal to 1% for most zones. The only notable exception is the prediction of \( P_t(t + 12) \), where the input selection strategy 1 causes an unacceptable increase of average MAPE by approximately 5% from 7.41% to 12.32% due to excessive dimensionality reduction from 19 to 2 inputs. This is caused by much more complex character of discomfort prediction, a poor correspondence of curves for validation and testing error (positions of their minima) due to the small sample size, and in consequence an incorrect estimate of the best dimensionality.

Regarding the medium and large training data size (Tables 5b and 5c), we observe similar phenomena as in the previous case of small training data. However, the strategy 1 never increases MAPE by more than 1%, which means that input selection does not lead to any critical accuracy reduction.

### 5.4.3. Results for dimensionality decision strategy 2

Strategy 2 is focused more on dimensionality reduction. It tries to avoid a significant degradation of performance by choosing the smallest subset of inputs, whose MAPE is still smaller than MAPE’s median value. From Table 5, one can see that this strategy mostly brings statistically significant performance degradation, which is however smaller than 1% in most cases. Again, an exception is the discomfort prediction with small training data, where only one input is often selected, which leads to very poor prediction accuracy. On the other hand, strategy 2 strongly reduced the input dimensionality by eliminating 13 and more inputs. This can be a good reason to adopt this strategy, especially for consumption and temperatures prediction and can be therefore very useful in some scenarios.

### 5.4.4. Summary of input selection

This section summarizes the results described above. To clearly compare the input selection strategies for temperature and discomfort, we aggregate the MAPE and dimensionality values over all 10 zones. Let \( M_i(z) \) be the value of MAPE for strategy \( i \) and zone \( z \) averaged over 100 runs of the experiment. For strategy 1, the mean absolute MAPE increase computed over all zones for temperature and discomfort is:

\[
\frac{1}{10} \sum_{z=1}^{10} \frac{M_i(z) - M_1(z)}{M_1(z)}. \tag{7}
\]

Similarly, let \( D_i(z) \) be the number of selected inputs and let \( D_3(z) = 19 \) be the original number of inputs. For strategy 1, mean relative reduction of dimensionality averaged over all zones is:

\[
1 - \frac{D_i}{D_3} = 1 - \frac{1}{10} \sum_{z=1}^{10} \frac{D_i(z)}{D_3(z)}. \tag{8}
\]

Analogically for strategy 2, the mean absolute MAPE increase is defined as \( \frac{M_i - M_3}{M_3} \) and mean relative reduction of dimensionality as \( 1 - \frac{D_i}{D_3} \). For total consumption of all zones, this aggregation is not needed. Therefore, the averaging is not used in the above definitions.

Summary of all the measures is reported in Table 6, where one can more clearly observe the real benefits of input selection. In average, Strategy 1 does not degrade the prediction by more than one percent of MAPE. For strategy 2, this is not true for discomfort prediction. Strategy 2 however extremely reduces
the number of predictor inputs for all predicted variables and all sample sizes.

In summary, for small data with 202 samples, strategy 1 leads to a dimensionality reduction (40% in average), and mostly improves the prediction. For large training data, it always slightly worsens the prediction by less than 1%, and eliminates only 28% of inputs in average.

Although it is not the aim of this article, one can interpret the input selection results by observing which inputs were selected in most runs and which were not. For example, in case of the total consumption prediction, strategy 1 selects always inputs $S_A(t + 12)$, $S_W(t + 12)$ and $T_d(t)$. The last one, air temperature in zone 4, is characteristic for its large size and small number of fan-coils and its air temperature is usually not saturated on the set point value. This is why this zone is so important. Another example, for prediction of $t_4(t + 12)$, temperature inside zone 4 in zone 4, is characteristic for its large size and small number of fan-coils and its air temperature is usually not saturated on the set point value.

It is shown that the prediction quality and a suitable choice of input selection is useful, but its further detailed description is out of scope of this article.

Finally, an important point is an evident instability of the input selection. Some inputs, especially temperatures, are selected in approximately half of the runs depending on random training procedure inside the input selection. This apparent strange behavior is again justified by the fact that the higher set point is usually not reached in large zone 4 and there, the discrete set point value is not important for the future temperature prediction. Those two cases demonstrate how the interpretation of input selection results is useful, but its further detailed description is out of scope of this article.

In the first part of the experimental work, some important parts of the prediction system design process were pointed out. It is shown that the prediction quality and a suitable choice of predictor’s complexity (topology) strongly depends on the sample size. The combination of those factors plays a key role. An example is the prediction of the discomfort described above. Here, for the smallest number of training instances and without early stopping, the MAPE can range between 14% for network with one unit and 24% for network with 10 units. Our results thus strongly support an importance of model tuning approaches and their comprehensive description.

A very interesting point is the complexity of the best performing networks. For all three prediction tasks and for our range of examined training sample sizes, the MAPE plots show that it does not make a sense to use network with more than three units. Moreover, a single unit network is the best for some smaller data. This does not correspond to the literature (see Table 1, where most studies use much more than 10 units).

6. Conclusions

In this work, we omitted any consideration of non-stationarity of occupant presence, its influence on the other variables and its prediction [47]. It was involved in terms of internal heating gain different for different weekdays and hours, but not changing throughout the year. For example, it has been shown in [48] that the occupancy can have a huge influence of on the building cooling load. Sun et al. [49] observed that occupancy profile has a significant impact on cooling load in the building during morning. This point must be considered in future works and can lead to additional occupancy based predictor inputs.

For future work, we are going to conduct similar experiments with recurrent neural networks and make a comparison to find out if and when their higher complexity brings benefits in terms of prediction performance. As far as we know, there is no such comparative study. We expect that a proper input selection will be even more important for recurrent models. Although ANNs are very popular soft-computing techniques used in industrial applications, recurrent neural networks are not used so often.
like the feed-forward models. One possible cause is the fact that their training is usually much more difficult and more complex recurrent models are more sensitive to over-fitting. It can be therefore crucial to perform a proper selection of network inputs, which can simplify the training and can lead to a better generalization abilities [50]. A proper input selection was observed to be very important particularly in real-world applications. Moreover, there is also lack of comparisons between ANN approaches and common system identification methods.[49, 97] This topic must be also focused in future.

The next stage of our effort will be devoted to the use of our models in an optimization loop, where the values of set points will be sought that minimize a cost function based on the predicted variables. In future, a per zone approach can be used with different value of air temperature set point for each zone. This increases the number of degrees of freedom and can help to reach higher savings, but simultaneously increases the complexity of the black box model (and also a further optimization process) and training data requirements.

Acknowledgement
The research was supported by the Czech Science Foundation project no. 13-21696P. Feature selection for temporal context aware models of multivariate time series.

References