Data Cleaning for an Intelligent Greenhouse

P. Eredics* and T.P. Dobrowiecki*
* Budapest University of Technology and Economics, Department of Measurement and Information Systems, Budapest, Hungary
eredics@mit.bme.hu, dobrowiecki@mit.bme.hu

Abstract— **The effectiveness of greenhouse control can be improved by the application of model based intelligent control. However for this a good model of a greenhouse is needed. For a large variety of industrial or recreational greenhouses the derivation of a fully blown analytical model is not feasible and simplified models serve no practical purpose. Thus black-box modeling has to be applied. Identification (learning) of black-box models requires large amount of data from real greenhouse environments. After recording long time series of greenhouse measurements to serve its purpose the data has to be checked for validity. Measurement errors or missing values are common and must be eliminated to use the collected data efficiently as training samples for the greenhouse model. This paper discusses problems of cleaning the measurement data collected in a well instrumented greenhouse, and introduces solutions for various kinds of missing data problems.**

I. Introduction

Greenhouses are built in various sizes and types all around the world to house plants needing special environmental conditions. Greenhouses are widely used both for vegetable and ornamental plant production.

The basic operation of a greenhouse is as follows: The transparent walls and roofs allow the solar radiation to pass through, but the warmed up air is kept inside. To prevent extreme high or low temperatures inside the house several actuators can be utilized. Shading curtains, automatic windows and active cooling systems can slow down the temperature rise in the summer while heating appliances are used in the cold season.

Most greenhouses are provided with some kind of automated control [1]. Such traditional control systems are based on operating levels decided by the owner, e.g. the owner has to set a window opening temperature limit along with a window closing temperature limit. Such rules have to be created for all different actuators and control situations. Main advantages of this control scheme are its simplicity (yielding high reliability) and simple working logic (the owner of the greenhouse always knows exactly why particular actions are happening). On the other hand traditional control solutions have some major disadvantages:

a) The owner has to adjust operating rules relying only on his or her expertise about the facility, and is not supported by the control system to do it optimally or even efficiently.

b) The control is reactive: it means that operations are executed only after the set limit is reached. Unwanted situations can not be avoided in advance or the limits must be set much more conservatively.

c) The actuators are not synchronized: all actuators work independently based on their rules yielding suboptimal of the most total operation of the greenhouse.

II. The Concept and the Requirements of Intelligent Control

The concept of an intelligent greenhouse is conceived to overcome the limitations of traditional control mentioned above [2]. In place of simple operating rules, the greenhouse owner specifies goals for the control system (e.g. in the form of target parameter zones). The system should then build a model of the greenhouse and predict its future states to avoid unwanted circumstances in advance. Using the predictions from the greenhouse model AI planning could be used to create plans for all actuators jointly. This novel approach is expected to help to overcome the limitations present in current greenhouse control systems.

The success of intelligent greenhouse control depends strongly on the accuracy of the modeling. Considering that the greenhouses come in different sizes and are designed for different purposes, analytical models are not applicable (feasible) to this problem [3]. Solely a black box-model might be able to adapt to any greenhouse it is installed in. The main drawback of black-box modeling, however, is the large number of training samples needed to construct the model [4]. The training samples must be derived from the training data recorded as time series characteristic to the evolution of the greenhouse the whole control system is installed in.

The accuracy of greenhouse models depends also on the time and space resolution of the measurements. While measuring e.g. every 5 minutes seems to be acceptable (due to the slow dynamics of the thermal processes in the greenhouse), the usual single location measurement used in traditional control systems is much too limited. To build high precision models several measurement locations have to be set up at strategically selected locations within the greenhouse.

III. Modeling the Experimental Greenhouse

The measurement and traditional control systems have been installed in a 100 m^2 greenhouse to collect real world measurement data [5]. The greenhouse has 18 desks holding most of the time very young and sensitive ornamental plants.

The measurement system records temperatures from all desks and also from thermally quasi-homogeneous larger parts of the greenhouse called zones. Fig. 1 shows the zone partitioning of the greenhouse: Zone-0 is the heating pipe; Zone-1 contains the desks (some covered with foil for humidity protection); Zone-2 means the interior air

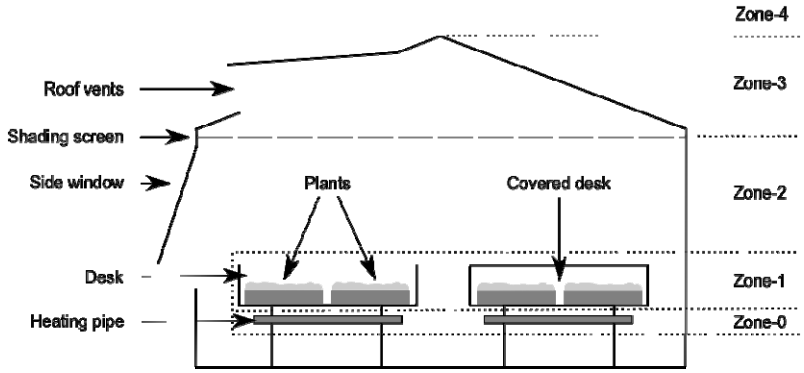

Figure 1. Simplified thermal zone structure of an industrial greenhouse

under the shading screen; Zone-3 means the air above the shading screen while Zone-4 represents the environment immediately outside the house. Temperature data is collected every 5 minutes from all zones with 0.5 degree accuracy. In addition online weather data for the region of the greenhouse is read with hourly resolution. Regional weather data and forecasts are also recorded.

The goal of data acquisition is to design the thermal model of the greenhouse. With so many measurement locations a monolithic model would contain too many parameters; its tuning would be possible only with an unacceptably high number of training data. To overcome this problem the decomposition proposed in [6] divides the greenhouse model into 6 modules, shown in Fig. 2. This functional decomposition still contains many black-box components (Module C, D and F), therefore although the number of training data is reduced, it still remains an important factor.

IV. THE DATA RECORDS

The data acquisition system is already installed and running since early 2008. Since then 297037 records were stored in the central database representing 23250 hours of measurement. Unfortunately the recording was not without breaks (the system was sometimes turned off for maintenance and power outages occurred also), but this amount of rough data is what is available for the modeling, nothing else. The structure of the records is presented in Table I. The recorded attributes are grouped as follows:

- Physical quantities
  - Global zone measurements
  - Desk measurements
  - Data from online sources
- Actuator states

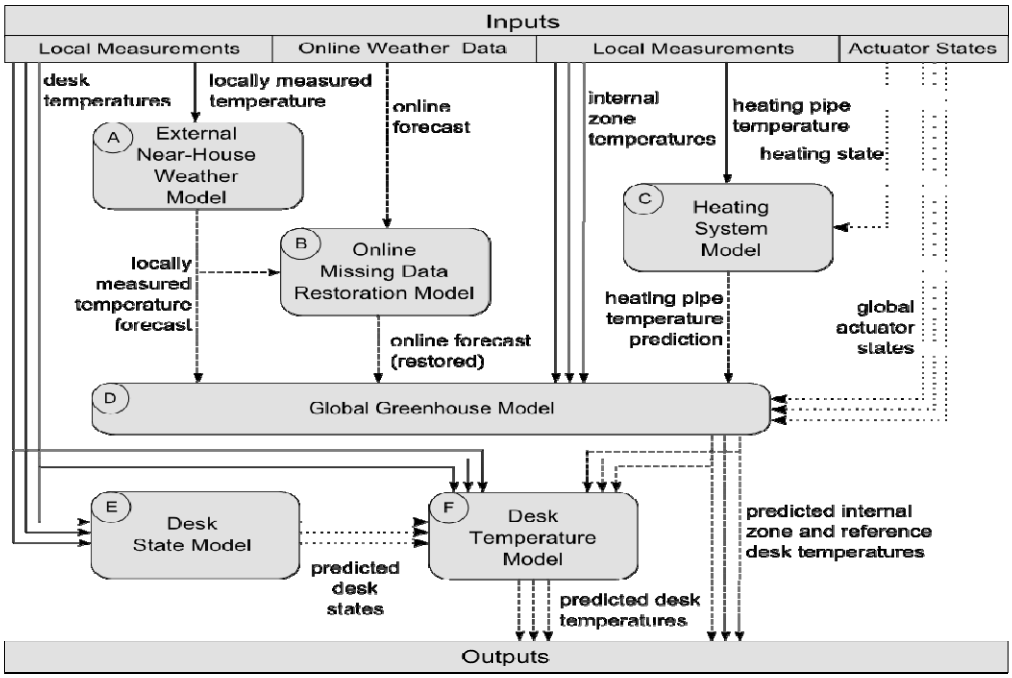

Figure 2. The proposed decomposition of the global greenhouse model into 6 modules related to the thermal substructures of the greenhouse (solid lines: measurements; dotted lines: states; dashed lines: predictions) [6]

TABLE I.
THE STRUCTURE OF OUTPUT DATA RECORDS OF THE GREENHOUSE DATA ACQUISITION SYSTEM

Field Name	Zone	Unit	Precision
Recorded	-	date and time	-
Heating pipe temperature	0	°C	0.5
Desk 1 temperature	1	°C	0.5
…			
Desk 18 temperature			
Under shading temperature 1	2	°C	0.5
Under shading temperature 2			
Under shading radiation		%	5
Above shading temperature	3	°C	0.5
External local temperature	4	°C	0.5
External local radiation		%	5
External online temperature		°C	1
Heating state	-	state code	-
Shading state	-		-
Upper windows state	-		-
Side windows state	-		-
Automatic misting usage	-	sec	5

These different groups of collected data might contain different types of errors and consequently call for different handling. The system also records regional temperature forecasts from an online source for 6 hours ahead forming a 3rd group of data available to the system.

V. PROCESSING PHYSICAL QUANTITIES

Records of physical quantities can contain three kinds of invalid values. The data acquisition system might enter an error code instead of the valid data (in case of errors at the lower system levels, e.g. error in the communication with the sensors). These error codes are represented by numbers out of measurement range for a given sensor, thus they can be processed together with the out of range measurements caused sometimes by sensor malfunctioning. Both cases can be differentiated by checking the recorded values against the upper and lower bounds of the operating range of the given sensor.

Some physical processes, driving particular quantities can exhibit large time constants, thus the change rate of these values will be limited. Comparing the actual value to the one read in the last recording can indicate such measurement errors with extreme high dynamics. Unfortunately identifying problematic values in the data alone is not enough to support the modeling process. Most black-box methods used for model building cannot handle missing input values and such attributes must be repaired in some way.

A. Repair by Copy

Due to the slow process dynamics copying the previous valid recording (if it is not older then 10 minutes) is an acceptable solution in case of most physical quantities. If the last available record is too old, time-series prediction methods might be considered, but in case of many values missing this would be too resource intensive. Taking into account the high computational needs of the intelligent control such expensive data restoration methods can not be afforded, and in this rare case it is better to drop the whole record

B. Repair by Regression

There is a specific physical parameter, namely the regional temperature value recorded from the internet that can be handled differently. This value is special in two main aspects: it is updated only once in an hour (thus copying the former value is acceptable in an hour range) and it contains somewhat redundant information as its value cannot be independent from the local temperature measurements. Considering it a simple regression model can be built based on local measurements to approximate the regional temperature value in case of internet black-out. This model uses the local temperature and radiation measurements as regressor values. A simple model of (1) was chosen for approximation as its accuracy is close to the sensitivity of the approximated data.

$$\widetilde{T}_{regional} = a_0 + a_1 * T_{local} + a_2 * R_{local} + a_3 * R_{local}^2 \quad (1)$$

where $T_{regional}$ is the approximated regional temperature value, T_{local} is the local temperature measurement while R_{local} is the local radiation data.

Higher order regression models were also tested but their accuracy did not prove to be significantly better. The weighting parameters a_i must be calculated periodically to follow the changes in the relationship between the two measurement locations (i.e. the weather station and the greenhouse), but this calculation (using least squares method) is necessary only once a week to keep the approximation accurate.

C. Repair by Spatial Interpolation

Temperature values recorded at the desks holding the plants call also for a special handling. Desks close to each other in position have closely related temperatures, thus restoring a missing desk value can be accomplished by a spatial interpolation.

Unfortunately some of the desks used to be covered with transparent foil to keep the humidity high while other desks are left uncovered based on the plants' actual needs. This fact makes it impossible to simply use the data from the closest desks for the reconstruction process. The desks for the interpolation must be selected dynamically for every restoring operation. On the other hand the state of the desks (i.e. foil covering) does not change frequently therefore these relationships can be cached and calculated only once a day.

In the current implementation the desks are sorted by the similarity (calculated based on the mean square difference of temperature values measured on the desks) of the previous 3 days. The weighted average of the 3 most similar desks is used as a replacement. The weighting factor is the normalized similarity value and the factors are recalculated only once a day. This method makes the system able to operate with some permanently

malfunctioning desk sensor, which is very important considering the increasing unreliability of the large number of desk sensor used in the greenhouse.

VI. Processing Actuator States

Two kind of actuator states are recorded in the given greenhouse: some actuators are characterized by their states (such as the windows with open/half-open/closed) while other actuators by the operation times (e.g. the operating time of the misting system) associated with them. The first group can be considered as a simpler case of the physical quantities as only error codes and out of the range state codes have to be identified and replaced. In the current implementation the only replace procedure available here is the copy method – later on with the greenhouse model available the system might be able to approximate the unknown state of the actuators based on the thermal behavior of the house.

For the second group of state descriptors, the only possible way to replace the missing data is to assume that the actuator was turned on, as it is the case most of the time.

VII. Processing Regional Temperature Forecasts

The regional temperature forecasts are obtained from a free online source. Sometimes this source becomes inaccessible (because of site maintenance or other reasons) resulting in missing predictions. In Fig. 2 Module-A and B are related to the temperature forecast and are implemented as part of the data cleaning system: Module-A generates a local forecast for the close proximity of the greenhouse based on local measurements. This module implements a time-series mining method detailed in [7] for the local temperature forecast problem. The local radiation forecast is a much simpler problem, as the radiation value follows the current weather phenomena.

As local measurement data is more likely to be available in the system, the regional forecast restoration process of Module-B can rely on them. Thus (1) can be used (with the weight factors recalculated if necessary) to present the missing regional forecast from the available local data. Of course if there is a former regional forecast at hand it might be used in the first few hours (e.g. a 4 hours old, 6 hours long forecast has yet valid data for the first 2 hours, thus only the last 4 hours must be calculated).

VIII. Results

Fig. 3 shows the data cleaning system implemented in the experimental greenhouse. All measurements (both historical and new measurements are handled the same way) are stored in the raw measurement data table in the central database. The data cleaner application reads this data along with the metadata (holding information about the sensors validity range, the actuators type and legal state count, etc.). If the data record can be repaired then it is processed accordingly in this application. The regional weather data is the only exception, as it is repaired in a separate module shared with the regional forecast related modules. All data vectors are stored in the output data table along with indication of valid measurement values or repaired approximations.

The weather forecasts are recorded in a raw weather forecast data table. The forecast cleaner application looks for raw weather data in every hour. If the forecast is present it is simply copied to the output table. In case of

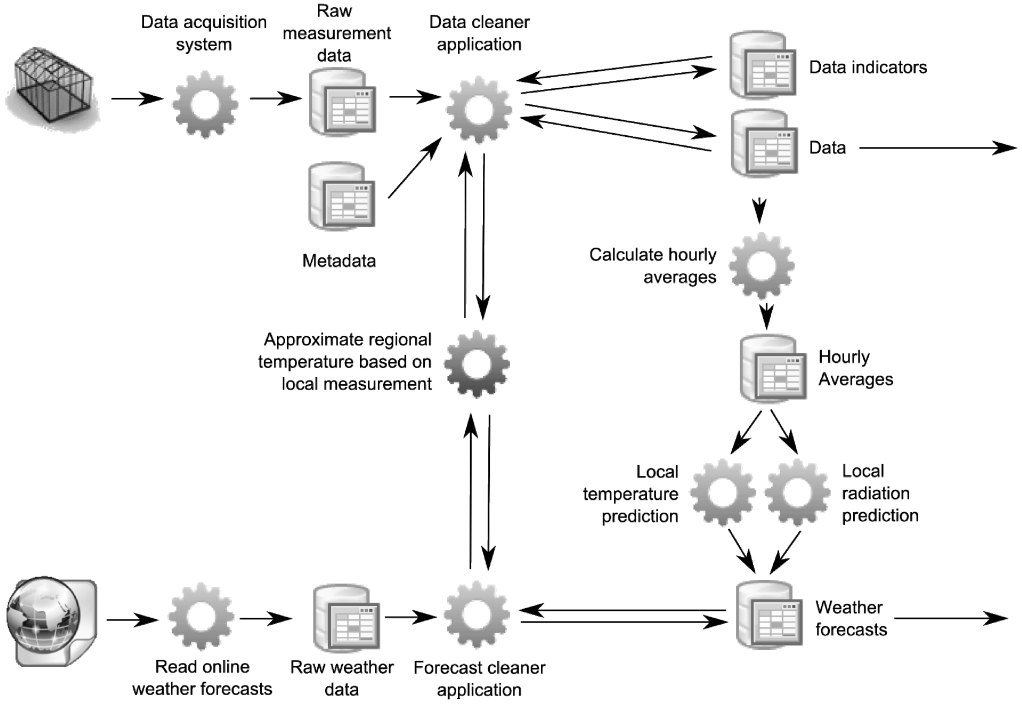

Figure 3. Architecture of the data cleaning system

missing forecast the local weather forecast is needed: due to the 1 hour time resolution of the weather forecasts, hourly averages are calculated for all important attributes. These values are stored in a separate data table, because computing them every time would mean useless overhead for the system. Based on hourly averages the local temperature and radiation forecasts are generated. These forecasts are used by the shared regional temperature approximation module and are stored in the output weather data table.

This system was used to process all 297037 data records (containing more than 9.8 million measurements) collected in the greenhouse. The copy method was used 32109 times to restore missing values. The spatial interpolation between desks has been executed 781150 times – this large number is caused by some permanently malfunctioning desk sensors. During spatial interpolation the weighting factors have been recalculated 535 times. The regional temperature approximation module was used 50362 times to restore actual regional temperature and 11412 times to rebuild regional forecast values based on locally generated forecasts. This way the system was able to raise the number of full data vectors from 184809 to 276760 available for the model building (50% gain). This way the number of multiple step training time series examples (especially important for training and testing the models prediction capabilities) has also been notably increased.

IX. Conclusion

The efficiency of greenhouse control systems can be improved by implementing model based intelligent control methods. Thermal modeling applicable to a whole variety of the greenhouses is only feasible with black-box models because of their ability to adapt to any greenhouse the system is installed in. Black box modeling requires large amount of measurement data from the greenhouse with high resolution both in time and space. A sophisticated measurement and control system working round the clock and for a long time, might have malfunctioning sensors and any kind of internal errors causing incomplete measurement records. These records have to be corrected before using them with black box modeling techniques.

This paper introduced various recovery methods for different types of measurement data collected in the greenhouse. Randomly missing attributes were always replaced by their pervious know value. In case of consequently missing attributes either spatial interpolation or regression methods were used. This way the number of useable data records has been increased by 50%. This means that the time needed to collect data before training the black box models can be decreased by 1/3.

Acknowledgment

The authors gratefully acknowledge the support of the Hungarian Fund for Scientific Research (OTKA).

This work is connected to the scientific program of the "Development of quality-oriented and cooperative R+D+I strategy and functional model at BME" project. This project is supported by the New Hungary Development Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

[1] Zazueta, F.S., Bucklin, R., Jones, P.H., Smajstrla, A.G.: Basic Concepts in Environmental Computer Control of Agricultural Systems. Agricultural and Biological Engineering Dept, Institute of Food and Agricultural Sciences, University of Florida, 2008.

[2] Blasco, X., Martineza, M., Herreroa, J.M., Ramosa, C., Sanchisa, J.: Model-based predictive control of greenhouse climate for reducing energy and water consumption. Computers and Electronics in Agriculture, pp. 49–70, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, 2007.

[3] G.P.A. Bot: Physical modeling of greenhouse climate. In: IFAC/ISHS workshop held in Matsuyama, Japan, 1991.

[4] Cunha J.B. Greenhouse Climate Models: An Overview. In: European Federation for Information Technology in Agriculture Conference, Debrecen, Hungary, 2003.

[5] P. Eredics: Measurement for Intelligent Control in Greenhouses, 7th International Conference on Measurement, pp 178-181, Smolenice Castle, Slovakia, 2009.

[6] P. Eredics, T.P. Dobrowiecki. Hybrid Knowledge Modeling for an Intelligent Greenhouse, 8th IEEE International Symposium on Intelligent Systems and Informatics, Subotica, Serbia, pp 459-463, 2010.

[7] P. Eredics: Short-Term External Air Temperature Prediction for Intelligent Greenhouse by Mining Climatic Time Series. In: 6th IEEE International Symposium on Intelligent Signal Processing, pp. 317–322, Budapest, Hungary,2009.