Data Cleaning and Anomaly Detection for an Intelligent Greenhouse

Peter Eredics and Tadeusz 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. Such control requires a good model of the greenhouse. For a large variety of industrial or recreational greenhouses the derivation of an analytical model is not feasible therefore black-box modeling has to be applied. Identification of black-box models requires large amount of data from real greenhouse environments. Measurement errors or missing values are common and must be eliminated to use the collected data efficiently as training samples. Rare weather conditions can temporally lead to unusual thermal behavior around and within the greenhouse. Anomaly detection run on the measurement data can identify such unusual samples and by excluding those from the model building better models and higher validation accuracy can be achieved. This chapter discusses problems of cleaning the measurement data collected in a well instrumented greenhouse, and introduces solutions for various kinds of missing data and anomaly detection problems.

1 The Concept and the Requirements of Intelligent Control

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 simple automated control [2]. Such traditional control systems are based on setting operating levels which are 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:

- 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.
- The control is reactive. It means that the actuators are operated only after the set limit is reached. Unwanted situations cannot be avoided in advance, even if they could be predicted, or the limits must be set much more conservatively.
- The actuators are not synchronized. All actuators work independently based on their rules yielding suboptimal at the best total operation of the greenhouse.

The concept of an intelligent greenhouse is conceived to overcome these limitations [3]. In place of simple operating rules, the greenhouse owner specifies goals for the control (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 [4]. This novel approach is expected to help to overcome the limitations present in current greenhouse control systems.

The prospects of intelligent greenhouse control depend strongly on the accuracy of the modeling. Considering that greenhouses come in different sizes and are designed for different purposes, analytical models are usually not applicable (feasible) to this problem [5]. 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 and tune the model [6]. The training samples must be derived from the training data recorded as time series characteristic to the evolution of the greenhouse where the whole control system is installed.

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 option. To build high precision models several measurement locations have to be set up at strategically selected locations within the greenhouse.

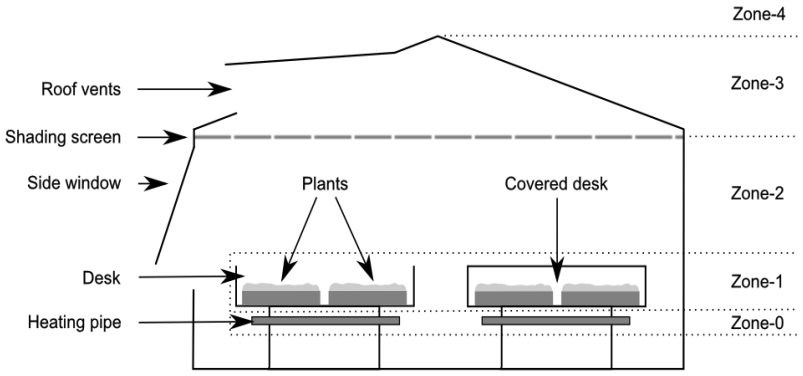
Data recorded in a real environment can be disturbed with several external influences: faulty operation of the data acquisition hardware (missing values or outliers) or rare, local weather phenomena can produce invalid data. The reparation or exclusion of such invalid data is essential to provide a reliable control for the greenhouse, and this is the responsibility of the data cleaning process: data records have to be classified as correct (can be stored without processing), as repairable (the data cleaning process is able to repair it) or as invalid (the record cannot be repaired, and has to be dropped). The problem of data cleaning was dealt with in the conference paper [1]. The main result of the present chapter is the management of the anomalous records. Records after the data cleaning process can all be classified as valid, but their usefulness for the model building can vary. High quality data records can speed up the model building and validation while low quality records should be omitted from the modeling process. The quality of data is determined by the anomaly detection process.

The structure of the chapter is as follows. Section 2 reviews the problem of the black-box modeling of the greenhouse. Section 3 provides information about the data acquisition system and the character of the measured data. Sections 4, 5, and 6 treat the problem of repairing the data coming from various sources. Section 7 introduces the problem of anomaly detection, presents interested cases and the proposed solution. Finally the experience is summarized and further research delineated.

2 Modeling the Experimental Greenhouse

A measurement and traditional control systems have been installed in a 100 m^2 greenhouse to collect real world measurement data. 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 under the shading screen; Zone-3 means the air above the shading screen while Zone-4 represents the environment immediately outside the



Fig, 1. Simplified thermal zone structure of an industrial greenhouse

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.

3 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 1. The recorded attributes are grouped into physical quantities (global zone measurements; desk measurements; data from online sources) and actuator states. 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.

4 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

Table 1. The Structure of Output Data Records of the Greenhouse Data Acquisition System

Field Name	*Zone*	*Unit*
Heating pipe temp.	0	C
Desk 1 temperature	1	C
…		
Desk 18 temperature		
Under shading temp. 1	2	C
Under shading temp. 2		
Under shading radiation		%
Above shading temp.	3	C

Field Name	*Zone*	*Unit*
External local temp.	4	C
External local radiation		%
External online temp.		C
Heating state	-	code
Shading state	-	
Upper windows state	-	
Side windows state	-	
Misting usage time	-	sec

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.

Copy method: Due to the slow process dynamics copying the previous valid recording (if it is not older than 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 cannot be afforded, and in this rare case it is better to drop the whole record.

Regression method: 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.

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

In (1) $T_{regional}$ 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.

Spatial interpolation: Temperature values recorded at the desks holding the plants call for a special handling too. 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 daily recalculated. 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.

5 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.

6 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. As local measurement data is more likely to be available in the system, the regional forecast restoration process can rely on them. A time-series mining method is implemented to produce local weather forecasts. Thus (1) can be used (with the weight factors recalculated if necessary) to present the missing regional prediction from this local forecast. 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).

7 Anomaly Detection

Anomaly detection is the last step of the data preparation process. Data cleaning methods introduced so far were intended to make a crisp decision about the quality of data (whether it is trustworthy of not), and replace the attributes qualified as unacceptable in one way or the other. This way data cleaning ensures a constant flow of acceptable data for the higher levels of the system, but lacks the ability to classify this data any further. In the case of the intelligent greenhouse the concept of anomaly detection refers to the process of examining the output data of the data cleaning process for its further usefulness.

Data quality measures the usefulness of data in the greenhouse modeling process. The data is high quality if it represents well the natural thermal dynamics of the greenhouse and is suitable to run modeling methods efficiently. Data quality will be low, if the data comes from some exceptional phenomena of the greenhouse or if it is disturbed with notable measurement noise. Running model building algorithms on high quality data can be much more efficient, as the low quality records will only introduce additional noise components into the learning process.

Data recorded in the greenhouse can be low quality for many reasons. A malfunctioning actuator can generate unwanted effects inside the house, e.g. if the windows would be stacked closed the whole day, the internal thermal processes would develop differently than on normal days. The weather is also strongly affecting the greenhouse, and rapid weather changes or fluctuations in the solar radiation can have unexpected effects inside the greenhouse. The data quality can be transformed in practice into the qualification as ordinary and extraordinary situations. Ordinary situations are good examples of the normal life of the greenhouse, while extraordinary situations are rarely occurring special events.

With this practical consideration about the data quality the test results of the greenhouse control can be analyzed with more insight as well. The greenhouse control can execute bad decisions and produce suboptimal operation in case of extraordinary circumstances, as these situations are rare and preparing for them is hard. On the other hand the greenhouse control, trained on high quality data, intended to cover the majority of ordinary situations well, has to be able to produce good control performance in case of ordinary circumstances.

7.1 Extending Data Records

The goal of anomaly detection is to identify data records with unusual internal structure (the attributes are messed up by some external disturbance) or unusual dynamics (the thermal processes of the greenhouse evolve abnormally). The first case requires a detailed analysis of the interrelationship between the attributes, while the second case can be handled by using previous data records to identify

tendencies. The method of extending data records detailed here is aimed to gather all necessary attributes into a single extended data record to be analyzed. The data record extender is a function with all temperature and radiation and control inputs listed in Table 1. This means 32 real value inputs along with some administrational data, such as record IDs and timestamps.

Different data types are processed in different ways. Each temperature record is extended into 2+24 values. The first value is the mean of the attribute over the last 2 measurements while the second value is the difference of the attribute and its last measured value. The next 24 values are generated as the difference of the given value and all other 24 temperature measurements of the system.

Each radiation input is extended to 2 values. The first value is the mean of the last 2 measurements while the second value is the difference exactly as in case of temperature data. The low number of radiation measurement (only 2) and the strong effect of the shading screen control onto their relationship makes it hard to express their connection in any simple way, thus radiation relations are not handled by this method.

The actuator control signals are handled universally as if all actuators would have 3 possible states. This assumption is true for the upper windows and the shading screen, while the heating and the side windows use only 2 of the 3 available states. For each actuator the time distribution among the states is calculated based on the last 12 measurement records (60 minutes). Along with the state distributions the number of state changes is also given for each actuator separately as the number of actuator commands issued also holds important information about the state of the control system.

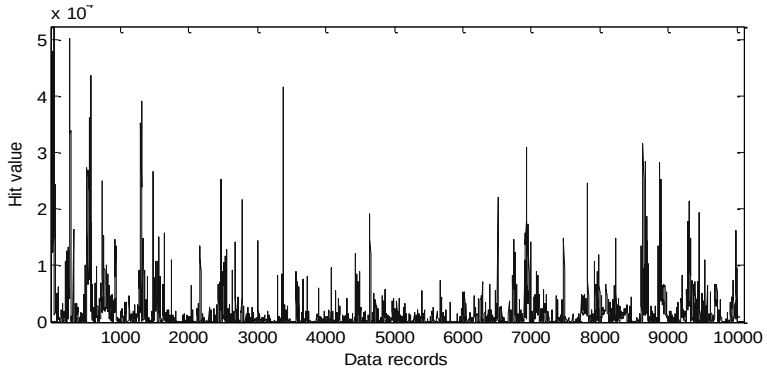
After extending all input parameters, the extended data record contains 644 numeric values. Most of these values represent the inter relation between the different temperature values, and this data is somewhat redundant, but the extended data record describes both the dynamics of the system and the relationships of the attributes together in a simple form.

7.2 Detecting Anomalies in Cleaned Data

A simple model is built from the extended data records at the beginning of the anomaly detection operation, and this model is updated with every new data record analyzed. The model consists of 24 averaged data records for each hour of the day, and 24 data records representing the standard deviations of each element in the model at the given hour. The first 5000 data records (approx. 17 days of measurement) are used to build this model and to calculate the standard deviations. Other data records can be tested against the model. After processing the new records the model is iteratively updated, taking into account the new records with the 0.01 discount factor, and standard deviation values are also updated. This way the

system model can keep up with the structural changes of the greenhouse, while its structure is kept simple.

After building the initial system model data records can be run through the anomaly detection. As the first step the data record has to be extended using the data record extending function. The second step is to select the appropriate hour from the system model based on the measurement time of the input record. After that the two extended data records (the one from the system model, built from several previous averaged measurements; and the one generated from the input record) can be compared attribute by attribute. Every time when the two records differ on any value more than 2 times the standard deviation, the input record gets 1 hit point. These hit points are summed after calculating all differences, and the sum is squared, producing a relative measure of how extraordinary the given input record is, called a hit value. The higher the hit value is, the more unusual the given data record is. Fig. 2 shows an example of this measure on tests run on 10 000 data records from the summer of 2008.



Fig, 2. Anomaly detection run on 10000 data records from the summer of 2008.

7.3 Anomalies Detected

The anomaly detection method produces a hit value to all input data records. The majority of data records have a uniformly low hit value as most of the time the thermal system of the greenhouse is operating normally. The hit value function shows occasional spikes, and this is where the unusual situations take place. After careful investigation of the spikes the following three types of anomalies can be detected in the data records.

At sample 3372 in Fig. 2 the sensors placed on desks 11-13 were malfunctioning, and produced a temperature drop of 5 degrees. The next measurement was correct, and this error was not corrected by the data cleaning system, as similar temperature changes can happen under normal circumstances. The anomaly detec-

tion gave a high hit value to this data record as several temperature values were behaving strange compared to their previous values and to the other measurements around them. Identifying such temporal and rare sensor malfunctioning makes it possible to eliminate such measurements from the model building and validating process yielding higher precision thermal models for the greenhouse.

Around sample 6923 several spikes can be seen in Fig. 2. These measurements were made on 07-07-2008, a summer day with a high cloud activity. The radiation measurements revealed that the solar radiation was rapidly changing during the day. This lead to unusual thermal oscillations inside the greenhouse and this special situation (caused indirectly by the clouds) was indicated by the anomaly detection. Such days should be eliminated from model building and testing as the cloud coverage prediction is out of scope for any complexity of greenhouse models, thus these changing circumstances are impossible to forecast.

Sample 8629 was recorded on 14-07-2008 at 11:01 am. This was the time, when a strong storm reached the environment of the greenhouse causing a sudden end to the usual morning warming process and dropping temperature with more than 5 degrees for several hours. At the middle of the day the parameters of the greenhouse were changing just if it was late evening, and this disturbance was indicated by the anomaly detection. Measurement data from such special weather phenomena are very useful for testing control systems in extreme weather conditions when quick reaction and safe operation has the priority.

Many other spikes were also investigated in Fig. 2, and a lot of these anomalies could be classified in one of the three events detailed before, namely the malfunctioning sensor, the unpredictable weather or the rare (unusual) weather phenomena. Unfortunately it also has to be noted, that the number of false alarms is significant, but the large measurement database lowers the importance of this problem.

8 Results

The data cleaning system has been 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 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.

The data cleaning 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). The number of multiple step training time series examples (especially important for training and testing the models prediction capabilities) has also been notably increased.

In conventional modeling problems measurement records can be used for modeling right after data cleaning. The case of the greenhouse is special, as rare external influences can drive the physical system into unusual states. In such states the collected measurements represent only the temporal state of the greenhouse well, therefore omitting them from modeling seems beneficial. For this aim after cleaning the data anomaly detection took place. This processing was run to identify rare weather phenomena or special sensor malfunctions and to indicate them in the database. During the anomaly detection process the data records were greatly extended to represent all relationships between the measured values. These extended records were compared than element by element to an hourly model of the system. Data records notably different from the system model got a high hit value indicating that these measurements should be handled with care later in the thermal modeling and validation process.

9 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 chapter 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. Data cleaning presented valid data vectors, but did not say anything about the usefulness of these records. That is why the data went through anomaly detection to measure how useful single records are for model building. In this process the data records got a hit value indicating their similarity to the normal operation of the greenhouse. High hit value records must be handled separately, as these high values indicate unusual situations inside the greenhouse.

Acknowledgement

The authors gratefully acknowledge the support of the Hungarian Fund for Scientific Research (OTKA), Grant #73496.

This work is supported by the grant TÁMOP - 4.2.2.B-10/1--2010-0009 and is partly supported by National Office of Research and Technology (NKTH), NAP-1-2005-0010 (BelAmI_H) project.

References

[1] Eredics P, Dobrowiecki TP (2011) Data cleaning for an intelligent greenhouse. In: Proceedings of 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, 293-297

[2] Zazueta FS, Bucklin R, Jones PH, Smajstrla AG (2008) Basic Concepts in Environmental Computer Control of Agricultural Systems. Technical report, Agricultural and Biological Engineering Dept, Institute of Food and Agricultural Sciences, University of Florida

[3] Drummond M, Bresina J (1990) Planning for control. In: Proceedings of the 5th IEEE International Symposium on Intelligent Control, Philadelphia, PA, USA, 657-662

[4] Blasco X, Martineza M, Herreroa JM, Ramosa C, Sanchisa J.: Model-based predictive control of greenhouse climate for reducing energy and water consumption. Computers and Electronics in Agriculture, 55, 49–70 (2007)

[5] Bot GPA (1991) Physical modeling of greenhouse climate. In: IFAC/ISHS workshop held in Matsuyama, Japan.

[6] Cunha JB (2003) Greenhouse Climate Models: An Overview. In: European Federation for Information Technology in Agriculture Conference, Debrecen, Hungary