Bayesian Restoration of Greenhouse Desk Temperature Measurements

P. Eredics*, K. Gáti*, T. P. Dobrowiecki* and G. Horváth*

* Budapest University of Technology and Economics, Department of Measurement and Information Systems, Budapest, Hungary

eredics@mit.bme.hu, gatikr@mit.bme.hu, dobrowiecki@mit.bme.hu, horvath@mit.bme.hu

***Abstract*—Modern greenhouse control solutions usually rely on temperature measurements collected in the greenhouse with high spatial resolution. The most important measurement locations are the close vicinity of the plants, namely on the plant supporting desks. Large number of desk sensors can ensure high spatial resolution, but increased volume of low-cost instrumentation results in a degraded reliability. Several methods have been elaborated in the past to handle the problem of missing or false measurements. This paper proposes a Bayesian approach for the substitution of missing data. The use of Bayesian networks has advantages as not only a single estimated value is returned but a distribution over the possible values as well. Furthermore the architecture of these networks is especially well suited for missing value replacement. The results show that the best overall performance may be obtained by the application of this Bayesian method, however in some other cases inserting the last known good value is also beneficial.**

I. INTRODUCTION

Greenhouses have transparent walls and roofs and are widely used for vegetable production and growing flowers. Sun radiation is essential for photosynthesis of the plants, and also to keep the inner temperature within an acceptable range. In the cold season a heating system may also be necessary. Contrary in hot weather other actuators, like roof vents, shading systems, exhaust fans or evaporative cooling can be used to avoid overheating.

The large number of the actuators in a well-equipped greenhouse makes it a challenging control problem [1]. As a basis of the control measurements have to be collected in the greenhouse [2]. In the recent years affordable solutions of (wired or wireless) sensor networks became available, making it possible to collect greenhouse temperature data of high spatial resolution [3]. Relying on these new, high precision greenhouse control methods can be developed resulting in better control performance and notable savings for the owner [4].

In most greenhouses plants are kept on desks for easy accessibility. Temperature measured on these desks is measured in the closest proximity to the plants, thus it is very important for the control. The large number of desks equipped with temperature sensors can provide valuable high spatial resolution data for the control system, but it also introduces reliability problems as only low-cost sensors can be applied in such a large number [5].

In all data acquisition systems data cleaning plays a key role to eliminate measurement errors and to reduce noise on the collected data. In case of greenhouse control data cleaning of desk measurements is one of the most important tasks of the data cleaning system because of the large number and the importance of these data.

This paper proposes a new method for efficient data restoration of desk measurements. The paper is organized as follows. Section II introduces the experimental data acquisition and greenhouse control system, providing data to experiment with the proposed methods. In Section III methods published and applied in the past are discussed. Section IV proposes an efficient Bayesian approach to the problem. The last two sections summarize the results and draw conclusions.

II. THE EXPERIMENTAL GREENHOUSE

The measurement data used in this paper is originated from an experimental greenhouse located in Western-Hungary. The greenhouse spans an area of 100 square meters, and houses young ornamental plants. Because of the special needs of such plants, the house is equipped with several actuators to keep the inner conditions appropriate. Several temperature sensors are built in as well. The layout of the 18 desks holding the plants and the environment of the greenhouse is illustrated in Fig. 1, with the position of the principal surrounding objects around the greenhouse also depicted. The shading effect of these objects causes different micro-climate on different desks, therefore temperature differences can be observed between desks.

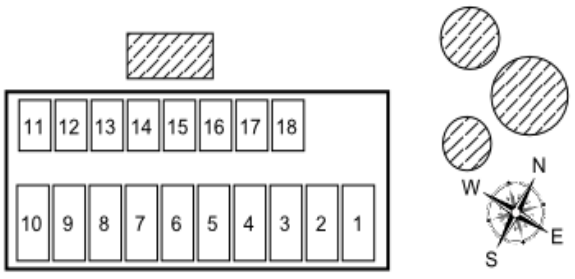

Figure 1. The layout of the experimental greenhouse with the desks and surrounding outside objects shading the house (striped circle = tree; striped rectangle = water tank)

The measurement system is an embedded distributed data acquisition system (detailed in [2]). Desk sensors are connected for every four desks to a desk sensor manager unit. This unit is connected to the main controller via a serial bus. The main controller is running the control algorithm and has a PC interface for computer control as well. Unfortunately the low-cost components integrated into the system result in reliability problems, sometimes

yielding errors and in consequence the measurement data cannot always reach the PC.

A. Measurement Errors

The measurement errors can be classified based on the number of affected sensors as follows:

a) An error affects only one desk sensor, in case of faulty sensor operation.

b) If a desk sensor manager reveals some temporal error, or the bus communication is distorted, all 4 desk measurements can be missing.

c) If the whole bus communication is down for some reason all 18 desk measurements can be absent, but data from other measurement locations in the greenhouse can be still present.

d) In case of the malfunctioning of the main controller unit, all measurement data about the greenhouse is missing.

Measurements are collected every 5 minutes (yielding a total of 288 measurements a day). Some errors are temporal (only affecting one measurement), while other errors are present for a prolonged period of time affecting several sequential measurements.

The different types of errors from a) to d) and the various length of error sequences call for different treatment and algorithms.

B. Test and Validation Data

Data cleaning methods presented in the next sections were developed and validated on real measurement data from the experimental greenhouse. A total number of 397711 measurements were recorded in the years 2009-2011. The desk sensors on desk 7, 10 and 17 have been almost permanently malfunctioning at that time; therefore these desks were excluded from the samples. With the remaining desks the number of correct data vectors was 284349 (71.5% of all data vectors were correct). Based on the incorrect data vectors the statistics of errors could be calculated. Fig. 2, shows the error sequence length distribution for all variables in the incorrect vectors.

The data of the 15 functioning desk sensors and the house global internal and external temperature measurements were used in all tests as inputs to the algorithms. These samples were used for testing the methods with artificially injected errors. The distribution of errors exactly followed the distribution in Fig. 2, but only errors shorter than 1 hour (i.e. 12 measurements in a sequence) were injected, because the effect of longer sequences would make some algorithms discussed later impossible to compare.

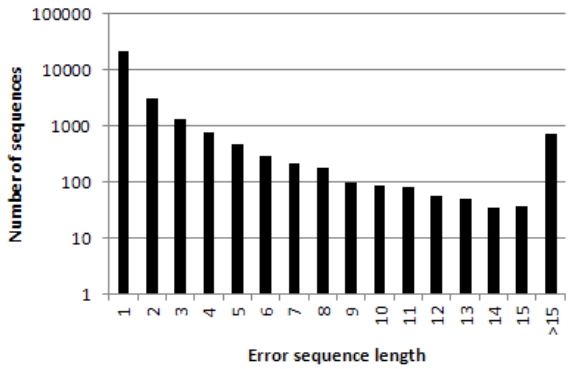

Figure 2. Histogram of the error sequence lengths

III. Earlier Methods

Several data cleaning methods were implemented and tested for the problem already in the past [6]. The two most effective methods are discussed in this section.

A. Copy Method

The copy method is the easiest way to restore missing data by simply substituting incorrect variables with their last known values. The method is very simple to implement, and also very robust as it can handle all types of measurement errors from the simplest case a) to hardest problem of d). It has good accuracy for temporal errors, but is absolutely unusable for permanent errors (error sequences longer than 12 steps). Unfortunately as the examples of desk 7, 10 and 17 show the case of permanent errors is impossible to neglect, therefore more complex solutions are also needed.

B. Spatial Interpolation Method

The main idea of spatial interpolation is to use the known measurements from similar desks in the data restoration process. Desks can be similar if their temperatures are running close to each other the whole day. As a basic assumption desks closely located to each other are likely to be similar. Unfortunately this similarity cannot be taken for sure, as the utilization of the desks, the different types of plants grown or the foil cover applied to protect young plants can cause very different thermal behavior even on desks close to each other.

Therefore simple spatial interpolation of the closest desk measurements is not applicable. The method has to be extended with the preliminary identification of the similar desks to serve as the basis of interpolation. The learning process of the extended method is as follows:

Step 1: Collecting 864 correct (or previously corrected) measurements from all desks and the internal sensor and the external sensor (3 days data). This means 15+2 vectors with a length of 864.

Step 2: Calculating the MSE between all 17 vectors, by calculating squared error between all corresponding elements of these vectors. This step results in a metric of similarity between any two sensors.

Step 3: Identifying for each sensor the 3 most similar other sensors, by finding the ones with the lowest calculated MSE. These related sensors are remembered, and a weighting factor (based on the inverse MSE values scaled to be summed to 1) is also stored.

The learning process is repeated after every 288 new measurements (practically every day). The recall phase is very simple: when a missing value has to be restored, the weighted average of the previously selected 3 most similar values are used. If some of these values are also missing, only the remaining ones are used with an appropriate weighting. In the rare case when all similar values are missing, the record is considered unrecoverable.

The main advantage of this solution compared to the copy method is its independence from the previous measurements. The missing values can always be restored (if similar data is present), thus a several days long sensor malfunction can still be tolerated. On the other hand the

problem of bulk missing values cannot always be solved, as case c) and d) cannot be handled this way.

IV. BAYESIAN METHOD

Bayesian networks are graphical representations of probabilistic distributions. They are represented as DAGs (directed acyclic graphs), where the nodes of the graph are random variables and an edge between two variables may be interpreted as direct causal relationship. To every node in the graph corresponds a conditional probability distribution which specifies the probability of a node given the values of its parents. Parents are those nodes which have a directed edge pointing at the specified node.

These probability distributions may be discrete or continuous depending on the data, or the combination of these, since there may be some nodes with discrete set of values, and some with a continuous domain.

The point of creating such networks is to decompose the joint probability distribution function (pdf) into smaller pdfs, which may be described with much less parameters. For example if the joint probability of 20 binary random variables has to be specified then there are 2^{20}~10^6 combinations of the variables, and for every such combination one must assign its probability value. However if a Bayesian network is constructed with the maximum number of parents limited to 3, only $20 \cdot 2^3$=160 values have to be estimated, which is a remarkable reduction in the number of required parameters.

In our approach we can utilize the architecture of Bayesian networks, and their capability to learn inductively both the structure and the parameters of the network from the pure dataset available. The same dataset may be used for both training the structure and the parameters.

There are two further properties of Bayesian networks which can be exploited. First its ability to represent arbitrary probability distributions, and second that expert knowledge may be applied directly to both the structure of the network and its parameters. For the preceding case the expert may directly specify the existence of a given edge in the graph or forbid it. For the latter case, the expert may estimate, based on the knowledge of historical data, the probability of some variable given its parents, filling the parameters of the conditional probability.

The Environment R [7] was used for executing the calculations. Bayesian networks were used from the *bnlearn* [8] package of R. In this approach only a minimal amount of expert knowledge was uesed, and almost every aspect of the network was trained with the *bnlearn* package using the available 3 year of data.

The measured data is quantized to 0.5 C°, so it was straightforward to apply a discrete Bayesian network first. In the discrete Bayesian network every random variable has a discrete domain. The structure of the network was trained using the Max-Min Hill Climbing algorithm [9].

This algorithm is a hybrid-structure learning algorithm, meaning that it is combining the two main approaches in structure learning, the constraint-based and the score-based training algorithms.

The result of training was not satisfactory both in the learned structure and in the parameters. The structure of the network contained very few edges. This is understandable, since pairs of neighboring tables were connected, but a global structure of the network could not be learned by the algorithm. Furthermore the trained probability distribution was not plausible, as visible in Fig. 3. The **P**(*Table16*|*Table15*) conditional probability distribution can be seen in Fig. 3, when the temperature of Table15 was 2.5 °C. There are two peaks in the figure, one centered at 3 °C suggesting that if on Table15 we measure 2.5 °C, most probably there is 3 °C on Table16. However there is another peak at approximately 8 °C, which is not easy to interpret, as this may be the result of overfitting of the network.

Because of such problems a continuous network was applied instead. An additional problem with the network could have been the quite high number of values of the random variables. For the training of continuous Bayesian network structures a different method called the Grow-Shrink method [10] was used. This method, contrary to Max-Min Hill-Climbing is a constraint-based algorithm.

One result of the trained network can be seen in Fig. 4. Beyond the temperature measurements additional variables were added to the network, like the month and hour of the measurement. As it can be seen these variables are independent from other measurements, as there is only one edge between the measurement of hour and the outside temperature.

The expert knowledge was also introduced here, because the edges between the inside and outside temperatures and the temperature of the tables were specified to be included in the structure. The reason for this was that these two temperatures have a well known and notable effect on the temperatures of the tables.

Please note that in the network structure the spatial relationship between the tables may be discovered. Connections like Table1 and Table2, Table5 and Table6, Table12 and Table13 or Table3 and Table18, all of them

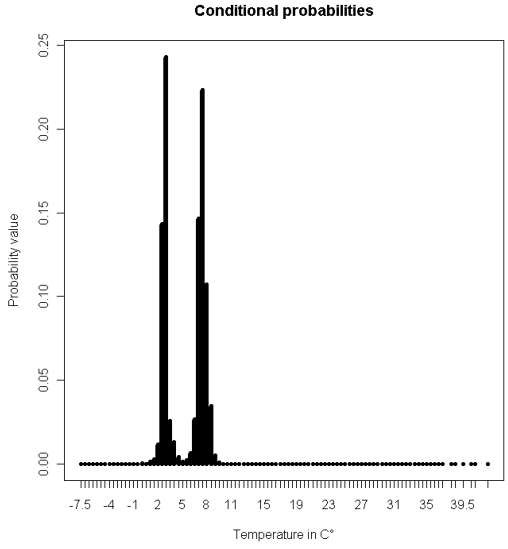

Figure 3. Conditional probability distribution function

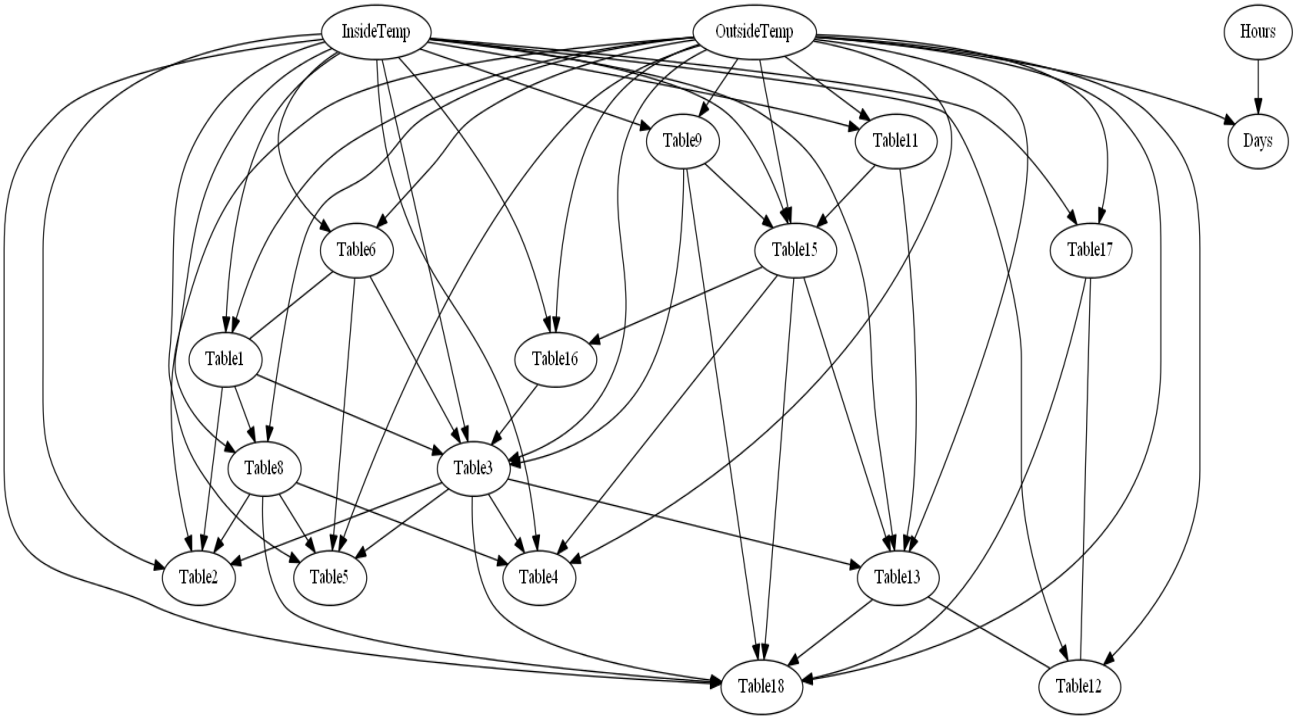

Figure 4. Structure of the trained Bayesian network, where Days means the day of the year, when the measurement was executed, Hours is the hour of the day, Inside and Outside temperature are temperatures measured as "global" temperature values inside and outside the greenhouse. Nodes denoted with TableX are the measured temperatures of table X, where X means the number of table based on Fig. 1.

show that the network could learn these connections from the data, which further means that there is some correlation between the measured values of these tables. There are also edges in the network, which cannot be explained by such spatial relation, e.g. the edge between Table9 and Table18. Here further investigation would be required to decide with other methods like correlation analysis, if these connections may be present in the data.

If these edges are just due to overfitting of the network, other structure training algorithms should be applied, or if the data contains this relationship but the expert knowledge deny the existence of this edge then the whole structure should be based on expert knowledge, or based on the earlier mentioned spatial positions of the tables.

One last assumption not taken into account was the non-stationary character of the measured data. The relationship among the tables and between inside, outside temperatures could vary based on the time of the year, or other circumstances. For this to handle we have to retrain the network for every month, so every time only a small portion of the data was used for the training of the network.

V. RESULTS

To compare the methods 10 different input sets were generated from each year between 2009 and 2011 based on the statistical properties of the measurement data. This way the methods were compared on 30 datasets, each containing 17 attributes and ca. 90000 records. The number of missing values was somewhat above 10500 in each case, and the distribution of error sequence lengths strictly followed the one displayed in Fig 2.

Fig. 5 shows the precision of the methods by depicting the mean absolute errors for each year separately. Fig. 6 displays the average precision of the methods for all test datasets along with the variance.

The Bayesian approach presents the best results in case of 2009, however worse result were obtained for the next two years. For that reason a deeper analysis of the data is needed. For all 3 years the worst results were obtained by the Spatial Interpolation, but the smallest error for a given year can be obtained by the Bayesian approach.

VI. CONCLUSIONS

The problem of data cleaning and missing value handling is a key element of modern greenhouse control solutions requiring measurements with high spatial resolution. The three methods detailed in this paper have

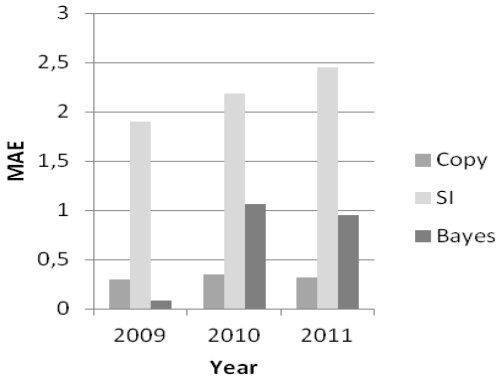

Figure 5. Mean Absolute Errors of the methods on data collected in the experimental greenhouse in 2009-2011

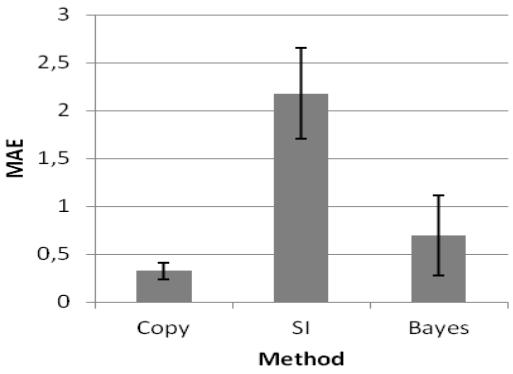

Figure 6. The overall performance (MAE and variance) of the methods tested on all generated input samples

different capabilities in handling missing values. The Copy Method is a computationally effective solution. With short missing data sequences this is the ideal solution. If the sequences are longer the Spatial Interpolation Method has a moderate precision with quite low computational requirements. On the other hand the accuracy of the method is not optimal.

The Bayesian method proposed in this paper has some advantages as handling expert knowledge explicitly, providing a probabilistic approach and the capability of learning from the data, so it can be adapted to the ever changing environment. This method can yield better results than the simple Copy Method, as it can be seen in Fig. 5. Further investigation is required how to select the structure of the network more efficiently and more accurately for the available information. First of all more expert knowledge should be included in the model, mainly at the structure learning phase, but further improvement at the parameter learning phase would be handy.

ACKNOWLEDGMENT

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

This work is supported by National Office of Research and Technology (NKTH), NAP-1-2005-0010 (BelAmI_H) project.

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).

This work was partially supported by the ARTEMIS JU and the Hungarian National Development Agency (NFÜ) in frame of the R3-COP (Robust & Safe Mobile Co-operative Systems) project.

REFERENCES

[1] X. Blasco, M. Martineza, J.M. Herreroa, C. Ramosa, J. Sanchisa: 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.

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

[3] Ling-ling Li, Shi-feng Yang, Li-yan Wang, Xiang-ming Gao. The greenhouse environment monitoring system based on wireless sensor network technology. 2011 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 265-268, Kunming, China, 2011.

[4] 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.

[5] T. Ahonen, R. Virrankoski, M. Elmusrati. Greenhouse Monitoring with Wireless Sensor Network. IEEE/ASME International Conference on Mechtronic and Embedded Systems and Applications, pp.403-408, Beijing, China, 2008.

[6] P. Eredics, T.P. Dobrowiecki. Data Cleaning for an Intelligent Greenhouse, 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, pp 293-297, 2011

[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0.

[8] Marco Scutari (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1-22. URL http://www.jstatsoft.org/v35/i03/.

[9] Tsamardinos I, Brown LE, Aliferis CF (2006). "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm". Machine Learning, 65(1), 31-78.

[10] Margaritis D (2003). Learning Bayesian Network Model Structure From Data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. Available as Technical Report CMU-CS-03-153.