FILLING MISSING TEMPERATURE VALUES IN WEATHER DATA BANKS

S. Kotsiantis[1], A. Kostoulas[2], S. Lykoudis[3,4], A. Argiriou[3], K. Menagias[5]

[1] *Educational Software Development Laboratory, University of Patras, Greece*
[2] *Mechanical Engineer T.E, M.Sc*
[3] *University of Patras, Department of Physics, Section of Applied Physics, GR-26500 Patras, Greece*
[4] *National Observatory of Athens, Institute for Environmental Research and Sustainable Developement, GR-15236 Palia Pendeli, Greece*
[5] *Mechanical Engineer T.E*

ABSTRACT

Meteorological data (wind speed, wind direction, rainfall, temperature etc) are an essential parameter for energy applications studies and development. Weather data is subject to different types of errors. The most commonly observed problems in temperature data embrace missing observations, unreasonable readings, spurious zeroes, and so on. Therefore, the data must be cleaned – that is, the errors and omissions must be corrected. In this research, the methodology adopted is to discard certain observed values and treat them as 'missing data'. We then examine and analyse the imputation accuracy of different interpolation techniques and filling methods for missing historical records of temperature data. The performance of these techniques as predictors of the missing values is evaluated using standard statistical indicator, such as Correlation Coefficient, Root Mean Squared Error, etc.

Keywords: regression algorithms, data cleaning, missing values.

INTRODUCTION

Weather data are generally classified as either synoptic data or climate data. Synoptic data is the real time data provided for use in aviation safety and forecast modelling. Climate data is the official data record, usually provided after some quality control is performed on it. Special networks also exist in many countries that may be used in some cases to provide supplementary climate data.

The reliability of the automated weather stations has been improved greatly. Failures of weather stations still happen, however, and techniques are necessary to fill gaps in data sequences in order to use them as input in weather or energy models. Several approaches have been pursued by researchers. The naïve hypothesis retained is to impute the missing temperature observation with the same day value of the previous year.

The reconstruction of missing weather data is different from weather forecasting because both data collected before the gap and data collected after the gap can be used. Still, the dependencies between the missing and available weather variables are expected to be complex, and advanced data analysis tools are needed to find and to express these dependencies with sufficient accuracy.

In this study, we examine and analyse the imputation accuracy of different interpolation techniques and filling methods for missing historical records of temperature data. The performance of these techniques as predictors of the missing values is evaluated using standard statistical indicators, such as Correlation Coefficient, Root Mean Squared Error, etc.

The following section describes the applications of meteorological data. Section 3 attempts a brief literature review of the techniques for eliminating noise instance and for filling missing data. Section 4 presents the data set of our study and the standard procedure for daily temperature estimation. Section 5 presents the experimental results for the representative regression algorithms used for filling missing data in our data sets. Finally, section 6 discusses the conclusions and some future research directions.

APPLICATIONS OF THE METEOROLOGICAL DATA

Knowledge of meteorological data in a site is essential for meteorological, pollution and energy applications studies and development. Especially temperature data is used to determine thermal behaviour (thermal and cooling loads, heat losses and gains) of buildings (1). It is also an explicit requirement for sizing studies of thermal (2) and/or PV systems (3),(4). Another major sector where temperature data is fundamental is the estimation of biometeorological parameters in a site (5). In advanced energy system designs the profile of any meteorological parameter is a prerequisite for systems operating management on daily and/or hourly basis. Also, simulations of long-term performance of energy plants require detailed and accurate meteorological data as input. This knowledge may be obtained, either by the elaboration of data banks, or by the use of estimation methodologies and techniques, where no detailed data are available. As nowadays "smart buildings" have became a reality, artificial techniques must be embedded is building management systems (BMS), in order energy profile (loads, gains etc) of a following time period (next hour, next day) to be predetermined. That will lead to a more effective energy management of the building or the energy plant. Weather data from

automated weather stations have also become an important component for prediction and decision making in agriculture and forestry. The data collected from such stations are used in predictions of insect and disease damage in crops, orchards, turfgrasses, and forests (6); in deciding on crop-management actions such as irrigation (7); in estimating the probability of occurrence of forest fires (8); and in many other applications.

Errors in weather data are frequently caused by poorly calibrated instrumentation, but also may be caused by errors in recording the data or while digitising older hard-copy records. In either case, this data must be "cleaned" before accurate valuation analyses can be performed. Weather data cleaning consists of two processes: the replacement of missing values and the replacement of erroneous values. These processes should be performed simultaneously to obtain the best result.

ELIMINATION OF NOISE INSTANCES AND FILLING MISSING DATA

The elimination of noise instances is a difficult problem. Usually the removed instances have excessively deviating instances that have too many null feature values. These excessively deviating features are also referred to as outliers. Variable-by-variable data cleaning is straightforward filter approach (those values that are suspicious due to their relationship to a specific probability distribution, say a normal distribution with a mean of 20, a standard deviation of 3, and a suspicious value of 50). Table 1 shows examples of how this metadata can help on detecting a number of possible data quality problems.

Impossible values should be checked for by the data handling software, ideally at the point of input so that they can be re-entered. These errors are generally straightforward like negative values when positive ones are expected. If correct values cannot be entered, the observation needs to be moved up the hierarchy to the missing value category.

TABLE 1 - Examples for the use of variable-by variable data cleaning

Problems	*Metadata*	*Examples/Heuristics*
Illegal values	Max, min	max, min should not be outside of permissible range
	variance, deviation	variance, deviation of statistical values should not be higher than a threshold
Misspellings	feature values	sorting on values often brings misspelled values next to correct values

Missing Data

Incomplete data is an unavoidable problem in dealing with most of the real world data sources. The topic has been discussed and analyzed by several researchers (9). Depending on the case, the expert has to choose from a number of methods for handling missing data (10):

- *Method of Ignoring Instances with Unknown Feature Values:* This method is the simplest: just ignore the instances, which have at least one unknown feature value.
- *Most Common Feature Value:* The value of the feature that occurs most often is selected to be the value for all the unknown values of the feature.
- *Concept Most Common Feature Value:* This time the value of the feature, which occurs the most common within the same class is selected to be the value for all the unknown values of the feature.
- *Mean substitution:* Substitute a feature's mean value computed from available cases to fill in missing data values on the remaining cases. A smarter solution than using the "general" feature mean is to use the feature mean for all samples belonging to the same class to fill in the missing value
- *Regression or classification methods:* Develop a regression or classification model based on complete case data for a given feature, treating it as the outcome and using all other relevant features as predictors.
- *Hot deck imputation:* Identify the most similar case to the case with a missing value and substitute the most similar case's Y value for the missing case's Y value.
- *Method of Treating Missing Feature Values as Special Values:* treating "unknown" itself as a new value for the features that contain missing values.

We used the most sophisticated technique in our study. We used regression methods for predicting missing values.

DESCRIPTION OF DATASET USED AND THE SIMPLE PROCEDURE FOR DAILY TEMPERATURE ESTIMATION

The values of temperature data used in this paper were obtained from the meteorological station of the Laboratory of Energy and Environmental Physics of the Department of Physics of University of Patras. Collected data cover a four years period (2002-2005). This station records temperature, relative humidity and rainfall data on hourly basis (8760 measurements per year). For the needs of this work mean daily temperature values for the city of Patras were calculated, from the elaboration of the data bank of that station. The mean daily temperature values were inserted in a new data bank with reference to the day of the year (D) (1-365). The data were also elaborated per

month. In that case mean daily temperatures were registered with reference to the number of the month (1-12), the number of the day of the month (1-30) and finally to the day of the year (D) (1-365). Missing hourly temperature data represent 11% of the total and were randomly dispersed over the four-year period (2002-2005). For the validation of the different methodologies for filling missing data, the mean daily temperature values of the year 2005 were considered to be missing.

Simple procedures for meteorological data gap filling can be derived from the elaboration and analysis of past time period measurements using procedures of meteorological data estimation-prediction. Many methods have been proposed so far worldwide for the estimation-prediction of monthly, daily or even hourly values of different meteorological parameters (11), (12), (13), (14), based mainly on past time data analysis. Such a simple method is the one proposed by Kouremenos and Antonopulos (15). This method is the result of the elaboration of temperature measurements made by the Hellenic National Meteorological Service (HNMS) in different sites of Greece. The analysis of this data shows that the yearly variation of the mean, maximum and minimum values of daily temperature can be expressed by the following equation (15):

$$T(D) = A + B\sin(\frac{360}{365}D - f) \qquad (1)$$

where D is the day of the year (1-365), A is the mean yearly temperature in °C, B is the width of the yearly temperature variation in °C and f is the phase shift expressed in degrees or days. These variables are typical and have constant value depending on the site of the country. Their values have been calculated for a number of Greek cities using the least square method. As far as Patras is concerned their values for the calculation of mean daily temperature are given in the table below (elaboration of temperature data of the period 1960-1974).

TABLE 2 - Values of A,B and f for the city of Patra

	Based on 1960-1974 data	Based on 2002-2005 data
A	17,339	18,351
B	-7,47	-8,65
f	-59,691	-62,908
Correlation coefficient	0.8872	0.8881

The values of average daily temperature estimated using the method proposed by Kouremenos were plotted against the measured data of year 2005 and are shown in the Figure 1. As the parameters of eq(1) have been estimated using past time data (1960-1974) we re-estimated them using the 2002-2005 data.

The values of average daily temperature estimated using the method proposed by Kouremenos were plotted against the measured data of year 2005 and are shown in the Figure 1. As the parameters of eq(1) have been estimated using past time data (1960-1974) we re-estimated them using the 2002-2005 data.

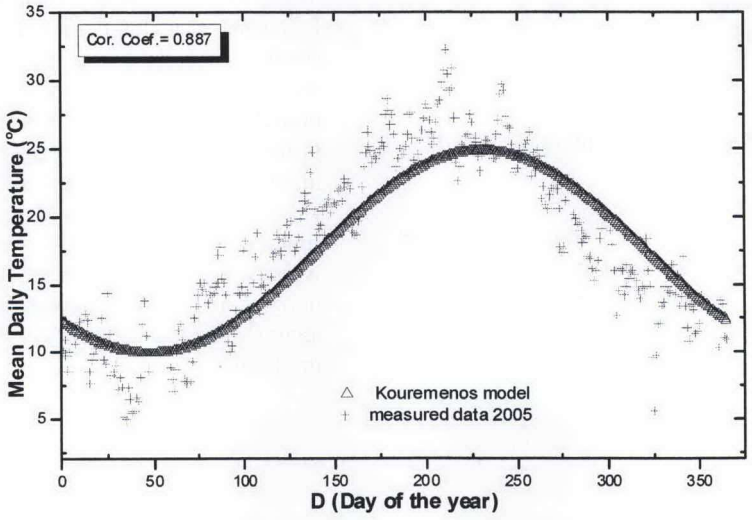

Figure 1: Scatter diagram for the measured values (2005) and the values estimated using the method proposed by Kouremenos based on the 1960-1974 data.

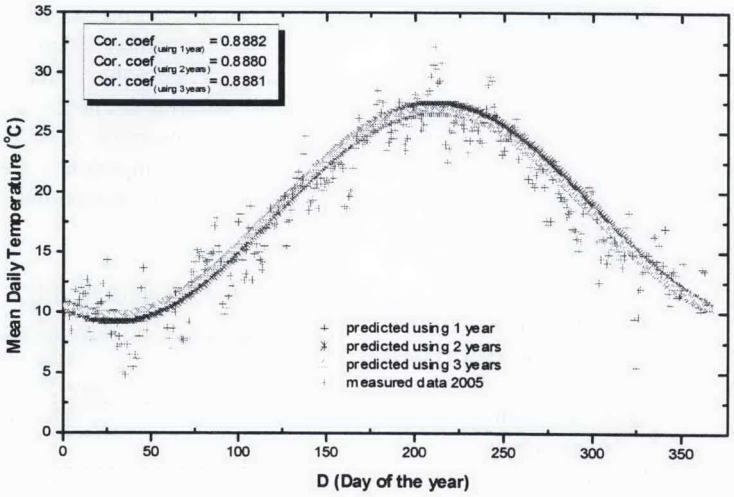

Figure 2: Scatter diagram for the measured values (2005) and the values estimated using the method proposed by Kouremenos based on the 2002-2005 data.

For the modification of the method proposed by Kouremenos the measured data were used to create a various data sets of temperatures. These sets were created from all the possible combination of the available years of data (2004, 2003-2004, 2002-2003-2004). As already mentioned the mean daily temperature values of the year 2005 were considered to be missing in order to validate the results of each methodology. The new values of the parameters A, B, f are given in Table 2 while in Figure 2 the estimated values of average daily temperatures using as input a) the previous year data (2004) b) the last two years (2003,2004) and c) the three last years (2002,2003,2004). Despite the fact that it is not correct, from the statistical point of view, to provide the coefficients A,B and f of eq(1) with three decimals, we kept this format to be in accordance with (15).

REGRESSION METHODS USED IN OUR STUDY

The problem of regression consists in obtaining a functional model that relates the value of a target continuous variable y with the values of variables x_1, x_2, ..., x_n (the predictors). This model is obtained using samples of the unknown regression function. These samples describe different mappings between the predictor and the target variables.

For the propose of our comparison the most common regression techniques namely Model Trees and Rules (16), instance based learners (17) and additive regression (20) are used. In the following we will briefly describe these regression techniques and their results in our dataset.

Model Trees

Model trees are the counterparts of decision trees for regression tasks. Model trees are trees that classify instances by sorting them based on attribute values. Instances are classified starting at the root node and sorting them based on their attribute values. The most well known model tree inducer is the M5′ (16). A model tree is generated in two stages. The first builds an ordinary decision tree, using as splitting criterion the maximization of the intra-subset variation of the target value (16). The second prunes this tree back by replacing subtrees with linear regression functions wherever this seems appropriate. The values of average daily temperature estimated using the M5' were plotted against the measured data of year 2005 and are shown in the Figure 3.

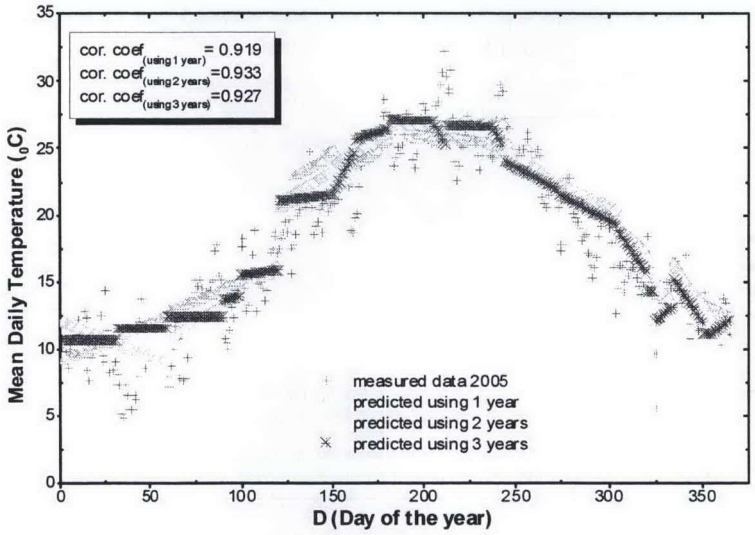

Figure 3: Scatter diagram for the measured values (2005) and the values estimated using the M5' algorithm based on the 2002-2005 data

Regression Rules

M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5'Model trees (18). The algorithm is able to deal with both continuous and nominal variables, and obtains a piecewise linear model of the data. The values of average daily temperature estimated using the M5rule algorithm were plotted against the measured data of year 2005 and are shown in the Figure 4.

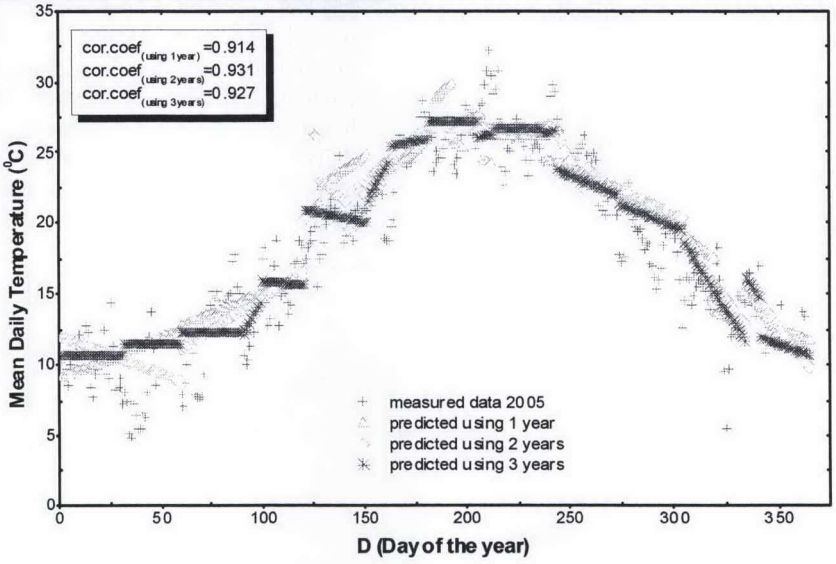

Figure 4: Scatter diagram for the measured values (2005) and the values estimated using the M5rules algorithm based on the 2002-2005 data

Instance Based Learning

Locally weighted linear regression (LWR) is a combination of instance-based learning and linear regression (17). Instead of performing a linear regression on the full, unweighted dataset, it performs a weighted linear regression, weighting the training instances according to their distance to the test instance at hand. This means that a linear regression has to be performed for each new test instance, which makes the method computationally quite expensive. However, it also makes it highly flexible, and enables it to approximate non-linear target functions. K* is a well known technique for instance based learning. The values of average daily temperature estimated using the K* algorithm were plotted against the measured data of year 2005 and are shown in the Figure 5.

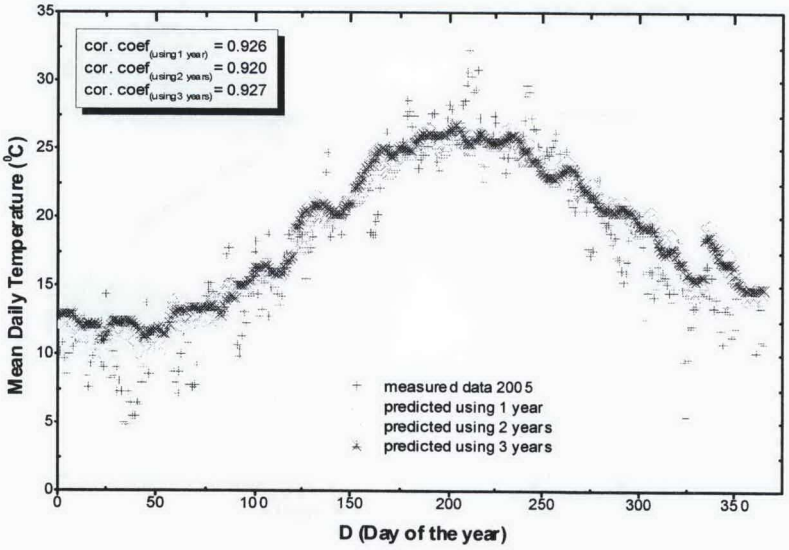

Figure 5: Scatter diagram for the measured values (2005) and the values estimated using the K* algorithm based on the 2002-2005 data

Additive Regression

Combining models is not a really new concept for the statistical pattern recognition, machine learning, or engineering communities, though in recent years there has been an explosion of research exploring creative new ways to combine models. Currently, there are two main approaches to model combination. The first is to create a set of learned models by applying an algorithm repeatedly to different training sample data; the second applies various learning algorithms to the same sample data. The predictions of the models are then combined according to a voting scheme.

A method that uses different subset of training data with a single learning method is the boosting approach (19). The boosting approach uses the base models in sequential collaboration, where each new model concentrates more on the examples where the previous models had high error. Although boosting for regression has not received nearly as much attention as boosting for classification, there is some work examining gradient descent boosting algorithms in the regression context. Additive regression (20) is a well known boosting method for regression. The values of average daily temperature estimated using the additive regression algorithm were plotted against the measured data of year 2005 and are shown in the Figure 6.

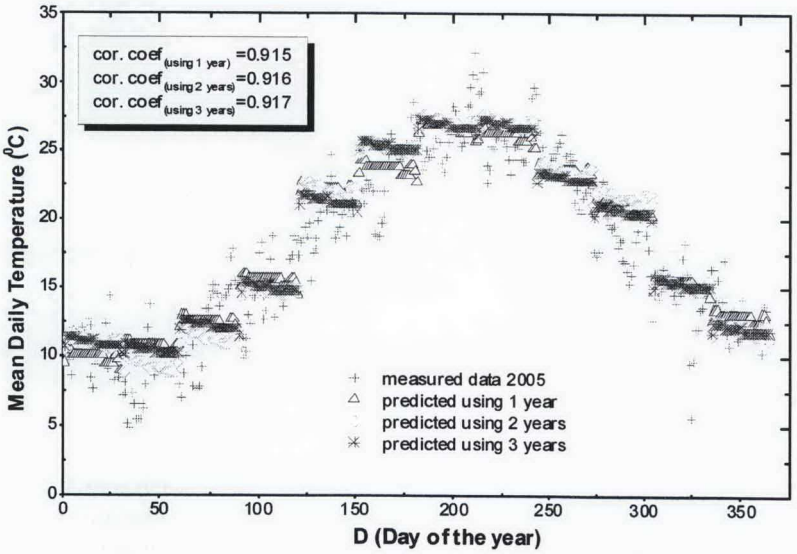

Figure 6: Scatter diagram for the measured values (2005) and the values estimated using the additive regression algorithm based on the 2002-2005 data

Discussion

For the regression methods, there isn't only one regressor's criterion. Table 3 represents the most well known. Fortunately, it turns out for in most practical situations the best regression method is still the best no matter which error measure is used.

TABLE 3 - Regressor criteria (p: predicted values, a: actual values)

Mean absolute error	$\frac{\lvert p_1 - a_1 \rvert + \ldots + \lvert p_n - a_n \rvert}{n}$
Root mean squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_1 - a_1)^2}{n}}$

In order to calculate the models' regressor criteria for our experiments, we used the free available source code for most of the algorithms by (18) for our experiments. In the following tables we present the models' regressor criteria using as input a) the previous year data (2004) b) the last two years (2003,2004) and c) the three last years (2002,2003,2004).

TABLE 4 - Using the previous year data (2004)

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (S.P.)
Correlation coefficient	0.9189	0.914	0.9153	0.9256	0.8882
Root mean Squared error (°C)	2.6454	2.7308	2.6465	2.6977	2.2044

TABLE 5 -Using the last two years data (2003-2004)

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (S.P.)
Correlation coefficient	0.9331	0.931	0.9156	0.9202	0.8880
Root mean Squared error (°C)	2.4635	2.4994	2.7352	2.8006	2.267

TABLE 6 - Using the last three years data (2002-2004)

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (S.P.)
Correlation coefficient	0.927	0.9275	0.9178	0.9275	0.8881
Root mean Squared error (°C)	2.483	2.4376	2.6266	2.8234	2.2233

As a result, it was found that the regression algorithms could enable experts to fill missing values with satisfying accuracy using as input the temperatures of the previous years. The experts are in the position using the temperatures of previous years, to fill the missing values of the examined year with sufficient precision, which reaches 91% correlation coefficient in the initial forecasts (using the data of the previous of the examined year) and exceeds the 93% using the data of the last three years before the examined year. However, because of the small size of training set it is not useful to run a statistical test (t-test) in order to compare these algorithms. The resulting differences between algorithms are not statistically significant. In a following study, we will include more data for this proposes.

CONCLUSION

Ideally, the market needs timely and accurate weather data. In order to achieve this, data should be continuously recorded from stations that are properly identified, manned by trained staff or automated with regular maintenance, in good working order and secure from tampering. The stations should also have a long history and not be prone to relocation. The collection and archiving of weather data is important because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the weather risk market.

Weather data released often must be "cleaned", ie, replacing missing and erroneous data before the weather market can use it. In this paper, we set out to examine and analyse the use of alternative methods when confronted with missing data, a common problem when not enough historical data or clean historical data exist. It was found that the regression algorithms could enable experts to fill missing values with satisfying accuracy using as input the temperatures of the previous years.

The next phase of this work is the implementation and validation of the techniques analyzed and validated here, using minimum and maximum daily temperatures data. The methods used in this work, for the case of Patras, should be tested and in other regions with different climatic profile. Also, other methodologies (like Neural Networks, fuzzy logic techniques etc) have to be validated in many regions of the country covering

its climatic spectrum, including not only temperature data (on any time basis) but other meteorological parameters as well (wind speed, solar radiation etc).

REFERENCES

1. ASHRAE, Handbook of Fundamentals, American Society of Heating, Refrigerating and Air Conditioning Engineers, New York: 1993
2. Klein, S.A, W.A Beckman and J.A. Duffie. 1985. 'A Design Procedure for Solar Heating systems.' Solar Energy 18: 113-127.
3. S. Rahman and B.H. Chowdhury, "Simulation of Photovoltaic power systems and their performance prediction". IEEE Transactions on Energy Conversion 3,440-446 (1988)
4. Duffie, J.A., and W.A Beckman. 1991. Solar Engineering of thermal processes. New York: John Wiley and Sons
5. Matzarakis, A. 1995. Human-biometeorological assessment of the climate of Greece. Ph.D. Dissertation, University of Thessaloniki.
6. Dinelli, D., 1995: What weather stations can do. Landscape Manage., 34 (3), 6G.
7. M. C. Acock, Ya. A. Pachepsky, Estimating Missing Weather Data for Agricultural Simulations Using Group Method of Data Handling, Journal of Applied Meteorology: Vol. 39, No. 7, pp. 1176–1184, 2000.
8. Fujioka, F. M., 1995: High resolution fire weather models. *Fire Manage. Notes,* **57,** 22–25.
9. Jerzy W. Grzymala-Busse and Ming Hu, A Comparison of Several Approaches to Missing Attribute Values in Data Mining, LNAI 2005, pp. 378–385, 2001.
10. Lakshminarayan K., S. Harp & T. Samad, Imputation of Missing Data in Industrial Databases, Applied Intelligence 11, 259–275 (1999).
11. Gelegenis, J.J. 1999. 'Estimation of hourly temperature data from their month average values: case study of Greece.' *Renewable Energy* 18, nos 1: 49-60
12. I.J. Hall, Generation of a Typical Meteorological Year, Proceedings of the 1978 annual meeting af AS of ISES, Denver USA, 1979
13. P.C. Jain, Comparison of techniques for the estimation of daily global irradiation and a new model for the estimation of hourly global irradiation. *Solar and Wind Technology* 1, nos. 2, 1984, pp.123-134
14. K.M. Knight, S.A Klein and J.A. Duffie, A methodology for the synthesis of hourly weather data. *Solar Energy* 46, nos 2, 1991, pp.109-120.
15. Kouremenos D.A, Antonopoulos K.A, Temperature data for 35 Greek cities. In Greek. Athens 1993 – Second Edition.
16. Wang, Y. & Witten, I. H., Induction of model trees for predicting continuous classes, In Proc. of the Poster Papers of the European Conference on ML, Prague (pp. 128–137). Prague: University of Economics, Faculty of Informatics and Statistics (1997).
17. Atkeson, C. G., Moore, A.W., & Schaal, S., Locally weighted learning. Artificial Intelligence Review, 11, (1997) 11–73.
18. Witten, I.H., Frank, E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA, (2000).
19. Duffy, N. Helmbold, D., Boosting Methods for Regression, Machine Learning, 47, (2002) 153–200.
20. Friedman J. (2002). "Stochastic Gradient Boosting," Computational Statistics and Data Analysis 38(4):367-378.