OPTIMIZATION METHODS FOR TRAINING SETS SELECTION IN THE REPRESENTATION OF SOLAR COLLECTOR VIA ARTIFICIAL NEURAL NETWORKS

Luis E. Zárate*, Elizabeth Marques Duarte Pereira,**
Daniel Alencar Soares*, João Paulo D. Silva*, Renato Vimieiro*,
Antonia Sonia Cardoso Diniz***
*Applied Computational Intelligence Laboratory (LICAP)
**Energy Researches Group (GREEN)
***Energy Company of Minas Gerais (CEMIG)
Pontifical Catholic University of Minas Gerais (PUC)
Av. Dom José Gaspar, 500, Coração Eucarístico
Belo Horizonte, MG, Brasil, 30535-610
zarate@pucminas.br, green@pucminas.br

Abstract. *Due to the necessity of new ways of energy producing, solar collector systems have been widely used around the world. The efficiency of this kind of systems is calculated through measurement of process parameters. There are mathematical models that represent these systems. However these models involve several parameters that may lead to nonlinear equations of the process. Multi-layer Artificial Neural Networks (ANN), with supervised learning, have been proposed in this work as an alternative of these models. However, a better modeling of the process by means of ANN depends on a representative training set. To better define the training set, statistical ways and clustering techniques have been proposed and compared in this work. Results of both techniques have been discussed too in this work.*

Keywords. *Neural Networks, Solar Energy, Thermosiphon, Training Set Optimization.*

1. INTRODUCTION

In a reality where natural resources have been scarce, associated with the population increasing, the traditional ways of energy producing (by means of hydroelectric power plants) may not be sufficient. Therefore alternative ways with the purpose of energy producing have been proposed and, among these ways, solar energy systems are an alternative.

Solar energy systems, specifically water heaters, have considerable importance as substitutes of traditional electrical systems. In Figure (1), an example of water heater, called thermosiphon system, is schematically represented. The solar collector (collector plate) is the most important component in a thermosiphon system. The performance of thermosiphon systems has been investigated, both analytically and experimentally, by numerous researches (Morrison & Ranatunga 1980; Huang 1984; Kudish, Santaura & Beaufort 1985). The formula used to calculate the solar collector efficiency is the following:

$$\eta = \frac{\dot{m}c_p(T_{out} - T_{in})}{GA_{extern}} \qquad (1)$$

where η is the thermal efficiency, $\dot{m}$, the flow rate, c_p, the heat capacity of water, T_{out} and T_{in} are, respectively, output and input water temperatures, G, the solar irradiance and A_{extern}, the area of the collector.

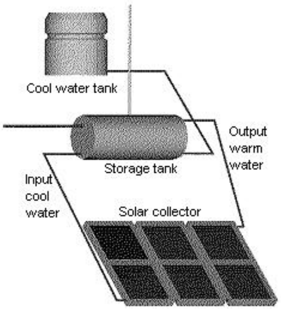

Figure 1. Schematic diagram of thermosiphon system.

Mathematical models (Kudish, Santaura & Beaufort 1985) have already been presented as a way of calculating the efficiency of solar collectors; however the non-linearity nature of those models makes their application discouraging. Linear regression (Kudish, Santaura & Beaufort 1985) has been proposed as a way of modeling solar collectors, instead of those complex mathematical models. But linear regression may introduce significant errors when used with that purpose, due to its limitation of working better only with linear correlated values.

In the last years, ANN (i.e. Artificial Neural Networks) have been proposed as a powerful computational tool. Some researches (Kalogirou 2000, Kalogirou, Panteliou and Dentsoras 1999, Zárate et al. 2003a, Zárate et al. 2003b and Zárate et al 2003c) have already discussed their use in the representation of thermosiphon systems. ANN have several advantages on other techniques, including their performance when dealing with nonlinear problems, their capacity of learning and generalizing, the low time of processing that can be reached when trained nets are in operation etc.

In Zárate et al. 2003a, a net trained with 601 data has been presented, however the time spent to train it is not satisfactory. Trying to obtain a lower training time, Zárate et al. 2003b presents statistical ways, and Zárate et al. 2003c shows clustering techniques, to define better training sets. A better-defined training set has its size decreased and is also better representative of the process. In Moreira and Roisenberg 2003, an alternative way, based in genetic algorithm, to define the training set is presented. This technique has not been considered a satisfactory solution in training set reduction, due to needed time to obtain the optimal training set. In this work, techniques which purpose is the reduction of the training set are discussed and compared, specifically, the statistical analysis and the widely known k-means clustering algorithm.

This paper is organized in six sections. In the second one, solar collectors are physically described. In the third section, collecting data from the solar collector is presented. In the fourth section, techniques to better define training sets are presented and compared. In the fifth section, the representation by means of ANN is discussed. Finally, conclusions are presented.

2. PHYSICAL DESCRIPTION OF THE SOLAR COLLECTOR

The working principles of a thermosiphon system are based on thermodynamic laws (Duffie & Beckman 1999). In those systems water circulates through the solar collector due to the natural density difference between cooler water in the storage tank and warmer water in the collector. Although they demand larger cares in their installation, thermosiphon systems are of extreme reliability and lower maintenance. Their application is restricted to residential installations and to small commercial and industrial installations. Thermosiphon system is presented in Figure (1).

Solar irradiance reaches the collectors, which heat up water inside them decreasing the density of heated up water. Thus cooler and denser water forces warm water to the storage tank. Since this is a constant process, the water flow happens between the storage tank and the collector, resulting in a natural circulation called "thermosiphon effect".

3. COLLECTING DATA FROM THE SOLAR COLLECTOR

Collected data refer to a typical solar collector and have been obtained by means of experiments in different ambient situations, under ASHRAE Standards (ASHRAE 93-86 RA 91). During three days of a characteristic period of the year for those experiments, measurements have been realized several times per day. Figure (2) shows a graphic where the relation between output temperature of water (T_{out}) and the hours during the day (*hours*) can be observed. Notice that the collected data are representative for different operating points and output temperatures.

In order to verify the non-linearity of collected data, the graphics in Figures (3), (4) and (5) have been built. However there is linearity in data sets. Although linear regression gives a valid approach, this work tries to demonstrate Artificial Neural Networks are capable of modeling data and estimating output with more precision.

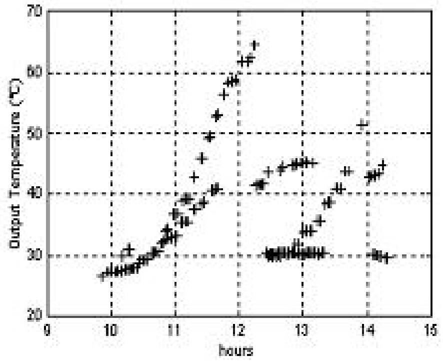

Figure 2. Output water temperatures X Hours.

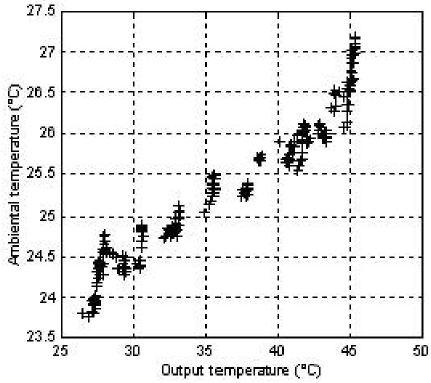

Figure 3. Ambient and output temperatures.

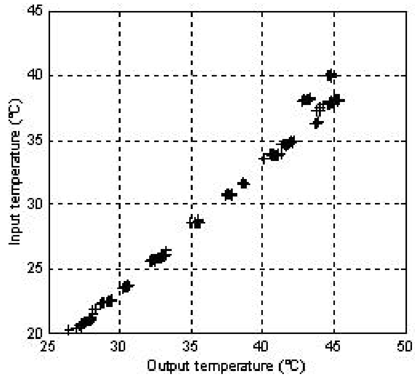

Figure 4. Input and output temperatures.

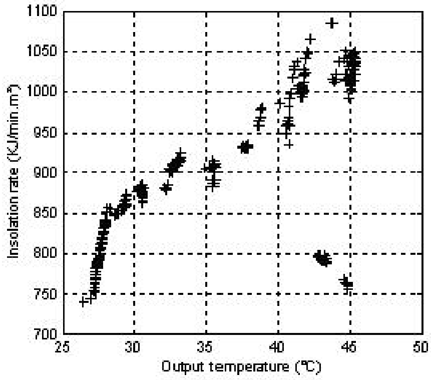

Figure 5. Solar irradiance X Output temperature.

The total number of collected data equal 631; those data include values of solar irradiance (G), ambient (T_{amb}), input (T_{in}) and output (T_{out}) temperatures. Table (I.1) (see append) shows a sample of those collected data. A subset composed by 30 data has been extracted from the original set in order to be used as validation set which is used later. Thus the new training set contains 601 data.

A reduced and better-defined training set must continue representing the problem, maintaining the capacity of generalization of the net, tolerable errors and permitting the reduction of time spent in the training process. In Zárate et al. 2003b, statistical analysis has been used to reduce the training set, resulting in 84 data. The clustering technique called k-means has been used in this work to reduce the training set, maintaining its capacity of represent the problem. In Zárate et al. 2003b and Zárate et al 2003c, statistical analysis and clustering techniques have been respectively used to select the training set, resulting in 84 sets (statistical analysis) and 20 sets (clustering technique). The results of these two techniques are discussed in this work.

4. TECHNIQUES OF TRAINING SET SELECTION

In order to get a better-trained ANN, increasing its capacity of generalization, the training set must characterize the problem and it must also cover all the possible situations that may happen in the problem or, at least, a major part of them. Considering this, the training set should be composed by a great number of data, however training sets formed by greater number of elements leads to a greater training time.

The genetic algorithm technique has been presented in Moreira and Roisenberg 2003. However that technique has not been considered a satisfactory solution, due to, as previously mentioned, the time needed to obtain the optimal training set, besides the considerable computational cost.

The techniques presented in Zárate et al. 2003b and Zárate et al 2003c, which are, respectively, statistical and clustering techniques, will be briefly described in the next subsections.

4.1. Clustering Technique (K-Means)

The k-means algorithm is one of the several techniques of clustering. It divides *n* data into *k* clusters, where *k* is a constant not defined by the algorithm. The result of this algorithm is a frame where all the objects present in a cluster have considerable similarity among them and a great dissimilarity to objects present in other clusters. Each cluster has a center point, which has the principal characteristics of the group. In the center point, the sum of distances of all objects in that cluster is minimized.

In order to build a representative training set, the k-means algorithm has been used in the data set composed by 601 data. As the number *k* of clusters must be explicitly given to the algorithm, *k* value has been varied from 10 to 100. For each test with a different number of clusters, the distance between each point in data set to each cluster center point has been calculated. Figure (6) shows average distances between all points of each cluster and the center points of neighbor clusters, for all the tested quantities of clusters.

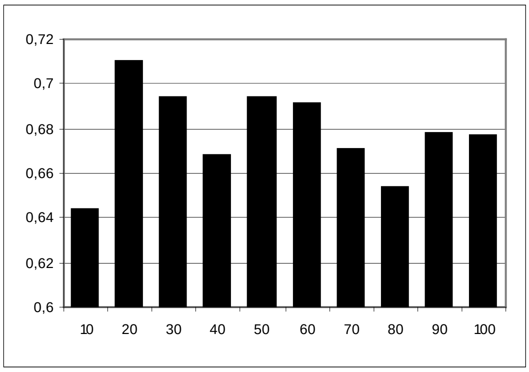

Figure 6. Average distances between the points of each cluster and the neighbor clusters.

Higher average distances between the points of each cluster and the center points of neighbor clusters characterize better-defined clusters. Considering this, the set of 20 clusters has been chosen.

After determining the optimal number *k* of clusters, a technique to select data present in the clusters has been applied. Although most representative characteristics are present in the center point of each cluster, this center point may not correspond to a real point in the data set. Thus, for each cluster, the point closest to the center point has been chosen resulting in 20 sets. Figure (7) shows, graphically, the data set divided in 20 clusters.

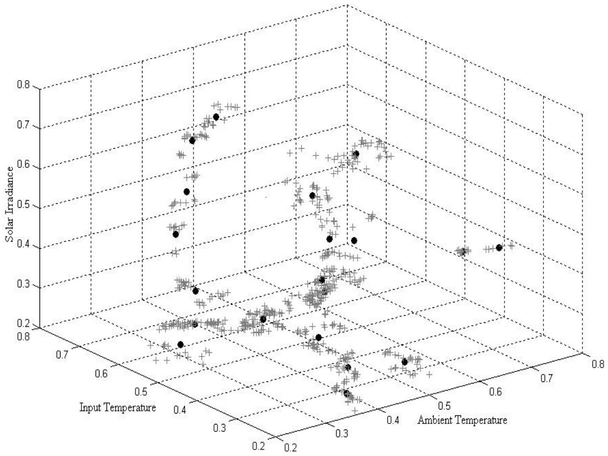

Figure 7. Clusters centers (●) and data sets (+).

4.2. Statistical Analysis Technique

To calculate the number of sets by means of statistical analysis, Equation (2) has been used.

$$n = \left(\frac{z}{e}\right)^2 * (f * (1 - f)) \tag{2}$$

where n is the size of the set, z, the reliance level, e, the error around average and f is the population proportion.

In order to verify the size of the training set, the error in Equation (2), around output water temperature average, has been varied from 0.1 °C to 0.04 °C, with reliance level fixed in 90 % (z = 1.645) and population proportion in 0.5. Table (1) shows the sizes of data sets with the assumed error. It is of extreme importance to note that the error values considered have no relation to the ANN maximum errors. The errors values determine the tolerable error for the statistical process of selecting data in a population, as shown in Equation (2).

Table 1. Calculated set sizes

Error	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Set size	423	271	188	139	106	84	68

With the sizes of data set calculated, a method must be chosen to select, among the 601 data, the necessary number of elements for the training set. The chosen method, presented in Zárate et al. 2003b, proposes three steps to define the training set:

1. For each variable of the net, maximum and minimum values found in the set have been selected.
2. The operation point of the problem (i.e. the most representative parameter in the set) has been chosen and added to the new training set.
3. The remaining elements have been randomly selected, with the purpose of reaching the calculated size for each training set.

To satisfy the first step the two major and the two minor values of each parameter have been chosen, resulting in a number of 12 entries (there are 3 parameters: T_{in}, T_{amb} and G). To satisfy the second step, the average temperature of output temperature has been chosen (which coincides with the critical time for measurements in a typical day – around midday). The rest of entries have been chosen randomly without repetition.

5. NEURAL REPRESENTATION OF SOLAR COLLECTOR

Multi-layer Artificial Neural Networks have been used in this work. Based in cognitive theories, they try to work similarly to human brain. Processing neurons are present in the ANN layers. The input layer receives the external entries while the output layer is responsible of producing an answer to the proposed problem. Choosing the most suitable values of some parameters of ANN, like the number of their neurons, is still a non-solved problem, although there are approaches. The suggested number of neurons in the hidden layer is *2n+1* where *n* is the number of entries of the net (Kovács 1996); the determined number of neurons in the output layer equals the number of wanted output answers from the net.

Input water temperature (T_{in}), solar irradiance (G) and ambient temperature (T_{amb}) are variables used as entries to the ANN. And the output water temperature (T_{out}) is the wanted output from the net. In this work, ANN represents the thermosiphon system according to the follow formula

$$f(T_{in}, T_{amb}, G) \xrightarrow{ANN} T_{out} \tag{3}$$

The structure of the net in this work may be schematically represented as shown in the Figure 8. The net contains seven hidden neurons (i.e. *2*entries+1*) and one neuron in the output layer from which the output water temperature is obtained.

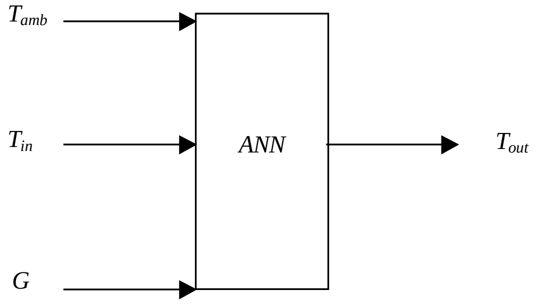

Figure 8. Schematic diagram of ANN

Nonlinear sigmoid function (4) has been chosen in this work as the axon transfer function for being the most consistent with the biophysics of the biological neuron. One of the parameters of this function is the considered sum of products of entries and weights associated with the connections of the neuron.

$$f = \frac{1}{1 + \exp^{-\sum Entries \text{ x Weigths}}} \tag{4}$$

Backpropagation has been chosen as the training algorithm; it uses a set during the training of ANN, from where inputs and wanted outputs may be extracted. For each parameter of the chosen training set, the weights of the net are adjusted in order to minimize errors obtained in the outputs values.

5.1. Preparing Data For Training

Pre-processing input data is a process of considerable importance for the performance of ANN. In this work, the following procedure has been applied to collected data, before the use of them in the net structure:

1. Data values have been normalized in order to be within the interval [0.2, 0.8].

2. Data have been normalized by means of the following formulas

$$f^{a}(\mathrm{L}o) = \mathrm{L}n = (\mathrm{L}o - \mathrm{L}m\acute{\imath}n) / (\mathrm{Lmax} - \mathrm{Lmin}) \tag{5a}$$

$$f^{b}(\mathrm{L}n) = \mathrm{L}o = \mathrm{L}n * \mathrm{Lmax} + (1 - \mathrm{L}n) * \mathrm{L}m\acute{\imath}n \tag{5b}$$

The formulas above must be applied to each variable of the training set (e.g. T_{amb}, T_{in}, G), normalizing all their values.

3. L_{min} and L_{max} have been computed as follows:

$$L_{\mathrm{min}} = L_{\mathrm{sup}} - (N_s / (N_i - N_s)) * (L_{\mathrm{inf}} - L_{\mathrm{sup}}) \tag{6a}$$

$$L_{\mathrm{max}} = ((L_{\mathrm{inf}} - L_{\mathrm{sup}}) / (N_i - N_s)) + L_{\mathrm{min}} \tag{6b}$$

where L_{sup} is the maximum value of that variable, L_{inf} is its minimum value, N_i and N_s are the limits for the normalization (in this case, N_i = 0,2 and N_s = 0.8).

5.2. The Training Process

For the training process, random values (between –1 and 1) have been attributed to the connections weights. As explained in sections (4.1) and (4.2), 20 and 84 of 601 data, respectively through k-means and statistical analysis, have been chosen for the training process. With a learning rate equivalent to 0.08, the error value established in 0.016 (1°C), two neural networks have been trained. After 80800 and 412800 iterations, the nets trained with 20 and 84 sets had, respectively, their training processes concluded. Table (2) shows minimum and maximum errors obtained in the training process.

Table 2 Training results.

	Statistical analysis	K-means
Minimum error(°C)	0.000039	0.017174
Maximum error (°C)	1.021237	0.92959
Error average (°C)	0.244534167	0.33199015

5.3. The Validation Process

Table (I.2) (see append) shows samples of the data sets used to validate the ANN; 30 data have been removed from the original data set to validate the ANN and have not been seen during the training process. Table (I.2) also shows samples of the output of the ANN and the errors obtained, compared to the real output temperatures. Table (3) shows the errors obtained in validation process.

Table 3. Errors from validation process

	Statistical analysis	K-Means
Minimum error (°C)	0.0432	0.0302
Maximum error (°C)	1.4752	1.3599
Error average (°C)	0.6255	0.4585

5.4. Verification Of Results

For the analysis by means of the linear regression, Equation (7) has been used:

$$\eta = F_{R}(\tau\alpha)_{e} - F_{R}U_{L}\frac{(T_{in} - T_{amb})}{G} \tag{7}$$

$F_R(\tau\alpha)_e$ equals 61.1 and F_RU_L, 570.65. F_R corresponds to collector heat removal factor, $(\tau\alpha)_e$, to transmittance absorptance product and U_L, to collector overall loss coefficient.

With the values of the output temperature of the water, the efficiency (*Eff*) of the solar collector can be calculated. Table (4a) e (4b) show a comparison of efficiency calculated through ANN (clustering and statistical methods) and linear regression. Some efficiency error values (i.e. minimum, maximum and average) of the linear regression are lower than the errors of the ANN. However, an advantage of ANN is that an equation to calculate efficiency does not have to be reformulated for new values of entries.

Note that in Tables (4a) and (4b), minimum and maximum errors, error average and standard deviation have been calculated for training sets of different sizes. For the first one (4a), 20 data have been used and for the second one (4b), 84 data have been used. Despite the errors of linear regression increase with the set size, the errors of ANN are maintained similar. This situation shows that linear regression may be less representative when the number of data increases.

Table 4a. Comparison between errors (clustering technique).

	Eff (real) – Eff (ANN) (%)	Eff (real) – Eff (LR) (%)
Average:	3.124178769	1.864363541
Minimum:	0.14650623	0.08587937
Maximum	8.110471201	7.215990429
Standard deviation:	2.672893417	1.816157438

Table 4b. Comparison between errors (statistical technique).

	Eff (real) – Eff (ANN) (%)	Eff (real) – Eff (LR) (%)
Average:	2.2710	2.1041
Minimum:	0.0003	0.0671
Maximum	8.9414	10.5667
Standard deviation:	2.0374	2.0823

6. CONCLUSIONS

In this work, a possible use of ANN to model a solar collector has been presented. It has been also presented techniques to build more representative training sets. These techniques are the widely used k-means clustering method and a statistical analysis. Training sets composed by 20 and 84 data (k-means and statistics, respectively) could be used as results of the application of those methods.

The average errors (Table (2)) obtained for output temperature equal 0.244534167 (statistical analysis) and 0.33199015 (k-means). The maximum and minimum errors of both techniques are, respectively, 1.021237, 0.000039 (statistical analysis) and 0.92959, 0.017174 (k-means). Those results show the optimal approach of ANN, since the error recommended by INMETRO (National Institute of Metrology and Industrial Quality – Brazil) is 1°C.

The efficiency values, calculated via ANN and linear regression, are presented in Table (4b). Although the errors obtained via linear regression are lower, ANN present some advantages on linear regression (e.g. For new situations with unusual values of entries, the equation formulated by means of linear regression may increase the actual errors values, while a trained net may use its capacity of generalization in order to maintain the errors values).

Comparing the results of training and validation processes of nets trained with all data (Zárate et al. 2003a), with data selected by means of statistical analysis (Zárate et al. 2003b) and with data selected by means of k-means, it can be concluded that a better-defined training set may decrease the time spent in training and may also maintain the capacity of generalization of the net, not significantly increasing its original errors. Table (5) presents a comparison among the training results, while Table (6) presents the validation results of the nets discussed above.

Table 5. Comparison of training results

	None technique	Statistical analysis	k-means clustering
Minimum Error(°C)	0.000035	0.000039	0.017174
Maximum Error(°C)	1.19	1.021237	0.92959
Average Error(°C)	0.15	0.244534167	0.33199015
Number of iterations spent in training	7700000	412800	80800

Table 6. Comparison of validation results

	None technique	Statistical analysis	k-means clustering
Minimum Error(°C)	0.02185	0.043265	0.030246
Maximum Error(°C)	0.70706	1.475292	1.359952
Average Error(°C)	0.27365	0.625548767	0.458544167

Analyzing the results of training and validation processes, it can be conclude that, despite the greater number of iterations of statistical analysis, it represents a better solution to the problem of training set optimization, when compared with k-means technique.

7. ACKNOWLEDGEMENTS

This project is financially supported by CEMIG (Energy Company of Minas Gerais - Brazil).

8. REFERENCES

Ashrae 93-86 Ra 91. Methods of testing to determine the thermal performance of solar collectors. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc., Atlanta (1986).

Duffie, J.A., & Beckman, W. A. 1999. *Solar engineering of thermal processes*. 2nd ed. U.S.A.: John Wyley & Sons, Inc.

Huang, B. J. 1984. 'Similarity theory of solar water heater with natural circulation', *Solar Energy*, vol. 25, p. 105.

Kalogirou, S. A., Panteliou S., & Dentsoras A. 1999. 'Modeling solar domestic water heating systems using ANN', *Solar Energy*, vol. 68, no. 6, pp. 335-342.

Kalogirou, S. A. 2000. 'Thermosiphon solar domestic water heating systems: long term performance prediction using ANN', *Solar Energy*, vol. 69, no. 2, pp. 167-174.

Kovács, Z. L. 1996. *Redes neurais artificiais*, São Paulo, Brasil: Edição acadêmica, pp. 75-76.

Kudish, A. I., Santaura, P., & Beaufort, P. 1985. 'Direct measurement and analysis of thermosiphon flow', *Solar Energy*, vol. 35, no. 2, pp. 167-173.

Moreira, F., & Roisenberg, M. 2003. Evolutionary optimization of neural network's training set: application in the lymphocytes' nuclei classification. *In* Hamza, M. H. ed. *International Conference On Artificial Intelligence And Applications*, Benalmádena, Spain, 8-11 September 2003. IASTED: ACTA Press, pp. 358-362.

Morrison, G. L., & Ranatunga, D. B. J. 1980. 'Transient response of thermosiphon solar collectors', *Solar Energy*, vol. 24, p. 191.

Zárate, L. E., Pereira, E. M., Silva, J. P., Vimieiro R., Diniz, A. S., & Pires, S. 2003a. Representation of a solar collector via artificial neural networks. *In* Hamza, M. H. ed.

International Conference On Artificial Intelligence And Applications, Benalmádena, Spain, 8-11 September 2003. IASTED: ACTA Press, pp. 517-522.

Zárate, L. E., Pereira, E. M., Silva, J. P., Vimieiro, R., & Diniz, A. S. 2003b. Neural representation of a solar collector with optimization of training sets. (Unpublished).

Zárate, L. E., Pereira, E. M., Silva, J. P., Vimieiro, R., & Diniz, A. S. 2003c. Optimization of neural network's training sets via clustering: application in solar collector representation (Unpublished).

9. APPEND

Table I.1 Training set sample.

Ambient temperature	Input water temperature	Solar irradiance	Output water temperature
25.05	27.17	908.42	33.97
25.91	34.7	1005.68	41.61
23.51	43.42	967.31	49.43
26.26	39.98	761.83	44.73
22.61	25.31	905.41	32.02
23.12	32.82	922.13	39.23
23.75	57.89	958.19	62.21
24.71	38.32	833.93	43.76
25.66	31.65	958.24	38.58
24.49	22.65	872.67	29.46
24.22	23.01	933.09	30.4
23.53	22.83	958.29	30.41
23.96	20.76	768.96	27.28
23.36	39.89	962.33	45.79
25.99	38.11	794.92	43.15

Table I.2. Validation data sets.

Ambient temperature	Input water temperature	Solar irradiance	Output water temperature	Output water temperature (ANN)	Error (T_{out} (real) – T_{out} (ANN))
23.83	20.66	755.1	27.17	27.630451	0.460451
24.43	20.97	819.75	27.74	28.14392	0.40392
24.61	21.47	850.02	28.07	28.63377	0.56377
24.44	22.5	860.06	29.27	29.388565	0.118565
24.87	23.72	869.47	30.55	30.365038	0.184962
24.81	25.96	912.59	32.85	32.46125	0.38875
25.31	30.81	932.79	37.52	37.219883	0.300117
25.66	31.65	958.24	38.58	38.304413	0.275587
25.82	33.75	993.54	40.78	40.810246	0.030246
25.85	33.81	996.78	40.86	40.902008	0.042008
25.96	34.65	993.69	41.6	41.772133	0.172133
26.03	34.79	1024.01	41.88	42.176178	0.296178
26.45	37.9	1022.66	44.63	45.445923	0.815923
26.66	38.04	1022.82	45.12	45.592113	0.472113
26.98	38.16	1041	45.27	45.85676	0.58676