Generalized Profile Function Model Based on Neural Networks*

Pero Radonja¹, Srdjan Stankovic²

Abstract: Generalized profile function model (GPFM) provides approximations of the individual models (individual stem profile models) of the objects using only two basic measurements. In this paper it is shown that this GPFM can be successfully derived by using artificial computational intelligence, that is, neural networks. GPFM is obtained as a mean value of all the available normalized individual models. Generation of GPFM is performed by using the basic dataset, and verification is done by using the validation data set. Statistical properties of the original, measured data and estimated data based on the generalized model are presented and compared. Testing of the obtained GPFM is performed also by the regression analysis. The obtained correlation coefficients between the real data and the estimated data are very high, 0.9946 for the basic data set, and 0.9933 for the validation dataset.

Keywords: Generalized profile function model, Individual stem profile model, Neural networks, Histogram, Scatter plot.

1 Introduction

In this paper a derivation of the generalized profile function model (GPFM) based on neural networks is described. GPFM provides approximations of individual models (individual stem profile models) of any object in region using only two basic measurements. Development of the nonlinear generalized models is given in [1,2] and linear in [3]. In this paper GPFM is obtained using all available normalized individual models as described in Section 4. A region can contain thousands of individual objects (spruce trees) with their own individual profile model. Since the measurements on all objects are practically impossible, we shall try to find GPFM that enables obtaining any individual model without detailed measurements on the very subject. In other words, we are trying to obtain a generalized model, in order to be able to get an approximation of any individual model, only by using the basic measured

¹Institute of Forestry, 11030 Kneza Viseslava 3, Belgrade, Serbia; E-mail: radonjap@eunet.rs
²Faculty of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Serbia; 11000; E-mail: stankovic@etf.bg.ac.rs
*Research is supported by the Ministry of Science and Environmental Protection of Serbia, as a part of the project EE - 273015.B.
values D and H, that is, the sets of values, data pairs, (1.3,D) and (H,0), where D is the diameter of a tree at breast height, 1.3m, H is the total height of a tree, and, generally, the values of data pairs (h,r) denote the tree radius r at the height h.

It is known that neural networks (NN) can be used very successfully in modeling of many nonlinear stochastic processes. The applications of adaptive neuro-fuzzy inference system (ANFIS) in forestry is given in [4] and in fuzzy controller in [5]. The modeling of a highly nonlinear biological process is given in [6,7] and modeling in forestry in [8,9].

By using the individual models we can get very accurate volumes of all the objects [3]. On the other hand, by using the obtained approximations of the individual models we can compute only approximate values of these volumes. Obviously, it is important that the errors in volume computing are unbiased. In this instance, the sum of the volumes of all objects from a region will be computed with minimal error. The overall volume of all the objects from whole region is important because of the sustainable management of the considered ecological system. Modeling of nonlinear relationships in ecology and complex environmental data are given in [10,11], and management and organizational topics are analyzed in [12,13].

2 Input Data Specification

For a problem statement as is defined in the Introduction, the total data set contains 260 data pairs, that is, 20×13 measured data pairs, (h,r), which are divided into two datasets [14]. The basic data set, which contains 182, (14×13) data pairs, will be used for obtaining the generalized model and the validation data set, which contains 78, (6×13) data pairs, will be used for verification of the obtained generalized model. In this instance we have used 13 data sets (h, r) for every object. In fact, in the considered case 14 objects will be used in order to get the generalized model and verification will be performed on 6 new objects.

The basic statistics of the measured r, average values, medians, standard deviations, SD, min. and max values for all three input data sets, are summarized in Table 1. Histogram of the measured values r for the total dataset is shown in Fig. 1. Numbers of data in fixed bands, frequencies (frequency - common statistical term), are given in histograms, Figs. 1 to 3. In the Fig. 1 the first band ranges from –2 to 2, second from 2 to 6, etc., whereas the last one ranges from 26 to 30cm.

Histograms of the measured values of the basic and validation dataset are shown in Figs. 2 and 3.
Fig. 1 – Histogram of the measured values of the total dataset.

Fig. 2 – Histogram of the measured values of the basic dataset.

Fig. 3 – Histogram of the measured values of the validation dataset.
Table 1
Summary statistics of the measured $r$.

<table>
<thead>
<tr>
<th></th>
<th>Numb. of data</th>
<th>Aver. [cm]</th>
<th>Med. [cm]</th>
<th>SD  [cm]</th>
<th>Min. [cm]</th>
<th>Max. [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Dataset</strong></td>
<td>260</td>
<td>8.1008</td>
<td>7.2000</td>
<td>5.9418</td>
<td>0.0</td>
<td>29.50</td>
</tr>
<tr>
<td><strong>Basic Dataset</strong></td>
<td>182</td>
<td>7.4975</td>
<td>6.3250</td>
<td>5.9019</td>
<td>0.0</td>
<td>29.50</td>
</tr>
<tr>
<td><strong>Validation Dataset</strong></td>
<td>78</td>
<td>9.5083</td>
<td>10.4250</td>
<td>5.8315</td>
<td>0.0</td>
<td>23.75</td>
</tr>
</tbody>
</table>

The scatter plots and time sequence plots of the measured values of the total, base and validation datasets are presented in the figures in Appendices A, B and C, respectively.

3 Individual Models Based on Neural Networks

Artificial neural networks, in short, neural networks (NN), represent a very efficient and powerful tool of computational artificial intelligence. In modeling many different biological process NN ensure smaller modeling error than the classical methods based on polynomials or exponential approximation functions [6,7,8].

A typical structure of a NN is the three-layer feed forward NN. The commonly used activation functions of neural networks include linear functions for output neurons, logistic sigmoid functions for hidden neurons, and identity functions for input neurons.

It is known that the best model for the observed data in the regard of the generalization is obtained in the case of a minimal number of neural network layers and neurons (for a general discussion, see [15,16]). In this instance, network complexity is determined simply by the number of free model parameters, that is, by the number of tansig neurons in the second (hidden) layer. The tansig neurons have logistic sigmoid tangent hyperbolic transfer function. In this instance, we found the best network structure by decreasing, step by step, the number of tansig neurons in the hidden layer [17].

In this study, a three layered feed forward NN with the back propagation algorithm and with 2 tansing neurons in the hidden layer was used.

As the illustration, we shall consider two very different individual models. The process of NN learning (or training) for the first model is shown in Fig. 4a. As a result of the good convergence properties, the Levenberg-Marquardt’s algorithm is used in the process of training [17]. In Fig. 4b, the first obtained (calculated) individual model is presented. Errors of fitting or modeling for the first model are shown in Fig. 4c. It is clear that these errors represent deviation model from measured values.
Fig. 4 – (a) The process of training for the first model; (b) The first obtained individual model; (c) Errors of modeling for the first model.
Fig. 5 – (a) The process of training for the second model; (b) The second individual model; (c) Errors of modeling for the second model.
The process of training for the second model is shown in Fig. 5a.

The second obtained individual model is presented in Fig. 5b. Errors of modeling for the second model are shown in Fig. 5c.

4 Generalized Profile Function Model

In order to get GPFM, it is necessary, in the first step, to calculate the normalized individual models of all the available objects. Normalization is performed by using the largest values on the $x$ and $y$ axes. In fact, the normalization of $x$ axis is performed by using $H$, and of $y$ axis by using $r(0)$.

GPFM is obtained as a mean value of all the available normalized individual models. The obtained generalized model is presented in Fig. 6.

![Figure 6](image)

**Fig. 6 – The generalized profile function model.**

Now the generalized model presented in Fig. 6 can be used for generating approximations of all the individual models. Renormalization is performed so that each individual (renormalized) model passes through the characteristic points $(x_0, y_0)$, $(x_0$ breast height, $1.3m$, $y_0 = D/2$ radius at breast height) and the final point $(H, 0)$. In accordance with this, renormalization per $x$ axis is performed using $H$, and per $y$ axis using $y(0)$ where:

$$y(0) = \left[\frac{(D/2)}{[(1.3/H)]}\right]. \tag{1}$$

The value $y(1.3/H)$ is obtained by using the generalized model, Fig. 6.

Quality and performance of the obtained GPFM will be analyzed in the subsequent sections.
5 Basic Statistics of the Estimated Data

In applications of artificial intelligence (neural networks) and verifications of obtained results, it is typical to use two datasets, the basic (model) and the validation dataset. The validation dataset can have the same dimension as the basic dataset, but more frequently it is smaller. In this instance, the validation datasets was about 40% of the basic dataset.

Testing the model accuracy by using the validation dataset we get the real model accuracy. Evidently, it is better to use more validation datasets. If we validate the model with the same data used during model derivation, we get the initial model accuracy. Typically, the initial model accuracy is higher than the real one.

Summary statistics of the estimated data for the validation and basic datasets are given in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Num. of data</th>
<th>Aver. [cm]</th>
<th>Med. [cm]</th>
<th>SD [cm]</th>
<th>Min. [cm]</th>
<th>Max. [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation</strong></td>
<td>78</td>
<td>9.2297</td>
<td>10.0040</td>
<td>5.6736</td>
<td>0.0671</td>
<td>22.6973</td>
</tr>
<tr>
<td><strong>Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Basic</strong></td>
<td>182</td>
<td>7.4744</td>
<td>6.3828</td>
<td>5.8483</td>
<td>0.0204</td>
<td>29.6145</td>
</tr>
</tbody>
</table>

As it was expected, a better agreement between the measured and estimated data, from the point of view of the average value, exists for the basic dataset, 7.4975 and 7.4744, than for the validation dataset, 9.5083 and 9.2297, Tables 1 and 2. (GPFM is generated by using the basic dataset).

6 Testing of the Obtained GPFM by the Regression Analysis

Usability of the obtained generalized model will be tested by comparing the measured data and the corresponding estimated data. Testing of the accuracy of the generalized model is practically performed by applying the regression analysis technique. The comparison of the measured and the corresponding estimated data is performed for both available datasets. The result of testing for the validation dataset is presented on Fig. 7a. Residuals which determine the standard estimation error for the validation data set are presented in Fig. 7b. The residuals represent deviation of the presented data in Figs. 7a and 8a from regression lines. In fact, the residuals are very close related with standard error of estimation $S$. 
Fig. 7 – (a) Statistical performance in radius estimation (validation data set); (b) Residuals (validation data set).

The first pair of results of data comparison represents statistical performance in data estimation. This pair contains the regression parameters: standard error of the estimate \( S \) and the correlation coefficient \( R \). The values of the regression parameters \( S \) and \( R \) (0.66236 and 0.99325) are presented in Fig. 7a. The second pair of the results contains the slope of the regression line and the intercept on the \( y \) axis. Evidently, the angle of the slope of the
regression line, obtained by comparing the estimated and the measured data, ideally must be 450 and has to start from the origin. In other words, the value of the parameter $b$ must be near 1.0, and the value of the parameter $a$, that is, the intercept on the $y$ axis, must be near 0. In the considered case the values of the parameters $b$ and $a$ (0.9664 and 0.0412), are presented in Table 3.

**Table 3**

Values of the statistics of data comparison by the regression analysis.

<table>
<thead>
<tr>
<th>Statistics of the $r$ comparison</th>
<th>Interc. $a$</th>
<th>Slope $b$</th>
<th>Std.Err. $S$</th>
<th>Corr. coeff. $R$</th>
<th>$R^2$ [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation Dataset</strong></td>
<td>0.0412</td>
<td>0.9664</td>
<td>0.6624</td>
<td>0.9933</td>
<td>98.65</td>
</tr>
<tr>
<td><strong>Basic Dataset</strong></td>
<td>0.0849</td>
<td>0.9856</td>
<td>0.6070</td>
<td>0.9946</td>
<td>98.93</td>
</tr>
</tbody>
</table>

The comparison between the measured and the estimated data for basic data set is done, as well. The result of testing is presented in Fig. 8a. Residuals which determine the standard estimation error in the case of the basic data set are presented in Fig. 8b.

![Graph of regression line](image)

**Fig. 8a** – Statistical performances in radius estimation (basic data set).

Values of the statistics of data comparison by the regression analysis, for the both datasets, are presented in Table 3.

As it was expected, the parameter $b$, the slope of the regression line, is closer to 1.0 in the case of data from the basic data set, 0.9856, than in the case of the validation dataset, 0.9664. Also, the obtained standard error 0.6070 is lower in
comparison with 0.6624, and the correlation coefficient is higher (0.9946 compared to 0.9933). However, the intercept on y axis is lower for the validation dataset, 0.0412, in comparison with 0.0849 for the basic dataset. By using the same dataset for both modeling and validation, Korol and Gadow [2] have obtained a lower standard error, 0.526, in comparison with 0.607 in this instance.

Fig. 8b – Residuals (basic data set).

7 Conclusion

The estimated data based on the proposed GPFM are compared with real, measured data, using the basic statistics and the regression analysis. The obtained statistics for the measured and the corresponding estimated data are very similar. This conclusion is valid for both the validation and the basic dataset. Also, correlation coefficients between the measured and the estimated data are very high, over 0.99, and the standard error of estimation very low, less than 0.7, for both datasets. It can be seen that the obtained generalized profile function model can be very successfully used in the process of estimating the individual profile models. Practically it implies that the errors in volume computing are small ($S$ small) and unbiased ($a$ close to 0 and $b$ close to 1). Consequently, the sum of the volumes of all objects from a region will be computed with very small error. The overall volume of all objects from whole region is important because of the sustainable management of the considered ecological system.

8 References


P. Radonja, S. Stankovic


Appendix A

Fig. A1 – Scatterplot of the measured values of the total dataset.

Fig A2 – Time sequence plot of the measured values of the total dataset.

Appendix B

Fig. B1 – Scatterplot of the measured values of the basic dataset.
Fig. B2 – *Time Sequence Plot of the measured values of the basic dataset.*

Appendix C

Fig. C1 – *Scatterplot of the measured values of the validation dataset.*

Fig. C2 – *Time Sequence Plot of the measured values of the validation dataset.*