# Estimating the Characteristics of a Marked Stand Using k-Nearest-Neighbour Regression

Mikko Tommola[1]
Mika Tynkkynen[2]
Jussi Lemmetty[3]
Pertti Harstela[4]
Lauri Sikanen[5]
*Joensuu, Finland*

## ABSTRACT

The purpose of this study was to develop the k-nearest-neighbour method as a wood procurement planning tool. Traditionally, sampling measurement of standing trees has been used to obtain advance information on marked stands. In this study, key figures such as sawtimber/pulpwood ratio in pine and spruce stands, diameter and height distribution in spruce stands, diameter and quality distribution in pine stands, and quality distribution by diameter classes in pine stands were estimated using k-nearest-neighbour regression. The material consisted of 716 stands. Stands were located in the eastern Finland. Information regarding every stand was collected from the information system of one large Finnish timber-procurement organization. The accuracy of the k-nearest-neighbour method was compared with the traditional planning inventory method and stand inventory method. The created model was found to be a useful tool in the planning of wood procurement.

Keywords: *Non-parametric estimation, planning inven to y, production planning, stand characteristics, wood procure ment.*

## INTRODUCTION

In a shortwood method, which is used mainly in Scandinavia, stems to be removed were marked by a forest officer in the old days. Nowadays, a harvester operator has to cope with operational planning of work, selection of strip roads, selection of removed trees in thinnings (=marking), bucking decisions and grading of lumber. However, the term "marked stand" is still valid and in use in Finland. A stand marked for cutting simply means same as the cutting area.

Timber procurement is among the costliest operations of the Finnish forest industries [8]. For example, as much as 60-70 per cent of a pulp mill's budget can be consumed in covering raw material costs (information provided by engineer Petri Lassila, Enocell Ltd.). The planning of wood procurement is an extremely important function and a great deal of costs can be saved by applying a proper planning system. Different kinds of information systems and models have been developed to enhance the efficiency of wood procurement. The question that remains is that of how to obtain accurate data required when using these tools.

Measuring a stand marked for cutting by some pre-logging method is one way to obtain accurate stand data. Generally, some sample trees are measured prior to felling. The sample size and methods vary between different organizations. These data are then used in planning wood-purchasing operations and in controlling the flow of wood from the stump to the mill process [5]. However, a variety of pm-logging measurement methods have turned out to be too expensive or their reliability has not been sufficient [7,9].

An information system formerly used by a large Finnish wood-processing enterprise differed from its wood- procurement system in that it was inflexible and not fast enough to be able to cope with a continually changing environment. because of this, it has been quite difficult to realise the basic idea of logistics: The right material to the right place at the right time. In *ad hoc* situations, e.g. when a sawmill gets a huge order for sawtimber needing to be delivered within the space of just a couple of weeks, an inflexible information system and insufficient data on marked stands can result in great losses throughout the production chain. The amount of non-marketable raw material lying in the storage would decrease were accurate data on marked stands readily available. Control over value-based cross-cutting also calls for advance data on marked stands.

Optimization of transportation activities, length and quality classes, and planning the output at sawmills require prediction of the diameter distribution and the sawtimber/pulpwood ratio of the incoming timber flow as accurately as possible. Up until now, production planning has been based on the planner's experience. This is followed by the marketing section selling the product beforehand in the market-

The authors are respectively [1]*Planning Officer, Tieto Corporation,* [2]*Researcher, Forest Technology University* of *Joensuu, Faculty* of *Foresty,* [3]*Customer Service Manager, Stora Enso Oyj,* [4]*Professor, Forest Technology, University of Joensuu, Faculty* of *Forestry, and* [5]*Lecturer, Forest Technology, University of Joensuu, Faculty of Forest y.*

place. Better and more accurate information of reserves of marked stands would result in more optimal decisions in production planning. It would be easier to share the products yielded by marked stands among sawmills so that each mill's raw material requirements would be fulfilled as well as possible.

The purpose of the present study was to introduce and test a new alternative for planning timber-procurement activities. The idea was to utilize the existing data available in a company's database. This procedure meant that the company's Woods Department's costs did not increase. The goal of the study was to create a non-parametric regression model for predicting the following:

- Sawtimber/pulpwood ratio in pine and spruce stands.
- Quality- and diameter distributions of pine stands.
- Diameter- and height distributions of spruce stands.
- Quality distribution of pine stands (classification by diameter classes).

At the outset, it was realised that the to-be-created model would have to be flexible and able to be applied throughout all areas of the company's operations. The estimastes of stand characteristics yielded by the model would also be compared with the figures yielded by other planning methods.

## MATERIAL

All the data used in this study were initially gathered for practical purposes, not for research. Consequently, some relevant data are missing and there is some degree of inaccuracy in the data. However, the k-nearest-neighbour method can be used despite these shortcomings, as long as some of the data are accurate. On the other hand, because this method is going to be part of a real information system, it has to work with real data in order to yield results of maximum validity.

The data consisted of 716 stands marked for cutting (located in the eastern Finland), which was recognised as being enough to test the method. For example Moeur and Stage had the data of 236 stands [6]. All the stems in every stand were measured by means of log-measurement instruments installed at the various sawmills. Common data on the stands and contracts (Table 1) were obtained from the information system.

Table 1. Data entered into the information system in contract-of-sale situations.

| Variable name | Unit of measure | Example |
|---|---|---|
| Batch number | Key figure | 1514001 |
| Area | Key figure | 24 (Karjala) |
| District | Key figure | 01 (Joensuu) |
| Foreman | Key figure | 23 (Huovinen, Reijo) |
| Team | Key figure | 01 (Team of Kontiolahti) |
| Municipality | Key figure | 214 (Kontiolahti) |
| Position | Geographical position | 123456, 123456 |
| Date of contract | Dd/mm/yy | 11.12.93 |
| Stand area | Are | 34 |
| Validity of logging | Key figure | 1 (Winter) |
| Logging method | Key figure | H1 (First thinning) |
| Age class | Number | 04 |
| Amount of logged wood | Cubic meters by species | Pulpwood, pine $45 m^3$ |
| Height above ground of first dead branch | Decimetres | 50 dm |
| Mean stem volume of trees | Litres | 1201 |
| Mean diameter by species | Centimetres | Pine 16 cm |
| Age | Years | 40 a |

Position is a very important variable in creating the model, because geographical position has a powerful impact on tree structure. For example, the tree-size structure of stands can vary a lot in the different districts. All the figures are quite reliable, except for the area of the marked stand, which is more or less an estimation. The volumes of logged wood are mainly based on the forestry workplans. In addition to the variables mentioned above, some additional variables were used when creating the model. Some of these variables were measured accurately by means of the aforementioned log-measurement instrument installed at sawmills. All the selected independent variables are presented in Table 2.

## METHODS

Previously, the non-parametric k-nearest-neighbour method has been used in forestry for forest-inventory purposes. In Finland, for example, it has been used in estimating the basal-area diameter distribution and in generalizing sample-tree data (diameter- and height distribution) for the stand [2,4]. The results obtained have been encouraging.

The main advantage of the non-parametric model (also referred to as local location estimator [1]) compared to the parametric regression model is that the former is more realistic: unrealistic values are avoided when extrapolating because all the values are estimated on a real-life basis. The non-parametric model is also easy to maintain and update. K-nearest-neighbour regression is a powerful data-analytical tool when used both as a stand-alone technique and as a supplement to parametric analyses [1]. Parametric models tend to decrease the deviation of the predicted values. Real variation is, however, very important when predicting the characteristics of individual marked stands.

The first phase in creating a model is to decide on the type of distance function to be used when seeking out the most similar reference plots [2]. Then one is required to decide as to the size of the neighbourhood and finally as to the weights of the reference plots. This reference-plot idea is quite close to the idea involved in this study, where the reference plots are simply replaced with the marked stands. The fundamental idea in this study was to predict the key figures for marked stands with inadequate advance information but having, nevertheless, some accurate stand data.

In the present study, the first step was to replace absolute numbers with relative ones so that stand area would not affect the choice of nearest neighbours. Following this, the form was standardized to avoid the influence of different scales on the variables [2]. The mean of every variable in standardized material is 0 and the variance is 1.

Following this, the nearest neighbows for each stand in regard to certain variable, i.e. the distance between the similarity of the reference stand (from the data base) and the target stand, were determined by calculating the Euclidean distances in regard to each variable for each stand (Eq. 1).

Equation 1. Euclidean distance.

$$e = \sqrt{\sum_{i=l}^{k} \left(W * (y_{ij} - y_{il})^2\right)}$$

in which

$e$ = Euclidean distance

$k$ = Number of variables

$i$ = Considered variable

$y_{ij}$ = Value of the considered variable in the reference stand j, j=1..n

W = Weight of the variable $x_i$

$y_{il}$ = Value of the considered variable in the target stand 1, l=1..n

The root mean square error (RMSE) is a common tool to evaluate the estimations given by the k-nearest-neighbour model (Eq. 2). In the present study, RMSE was used to evaluate the estimations of the sawtimber/pulpwood ratio.

Equation 2. Root mean square error.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}\left(y_{ij} - \hat{y}_{ij}\right)^2}{N-1}}$$

in which

RMSE = Root mean square error

N = Number of observations

$y_{ij}$ = The real value of the variable i in stand j

$\hat{y}_{ij}$ = The estimated value of the variable i in stand j

The estimations of height, diameter and quality classes were evaluated by the probability of what percentage of logs in the stand were located in the correct class.

The calculations did not include all the possible alternatives. Haara et al had ended up in using 10 neighbours to determine the independent variables [2]. Thus, the search for independent variables was carried out with 10 nearest neighbours and without weights. All alternative combinations of variables were tested. Second, the optimum weights were searched for heuristically by minimizing the error of the estimates. Finally, the optimal number of neighbours was determined by maximizing the number of correctly-located logs and minimizing the RMSE.

There are many techniques available for selecting the optimal size of the neighbourhood depending on the smoothness of the estimates required. If one uses a very large number of neighbours, the estimation will be very smooth and close to the mean of the data. On the other hand, with a small number of neighbours, the estimates obtained are almost unbiased and over-fitting. However, increasing the number of neighbours improves the accuracy of estimates [4]. Also some analyses of sensitivity with different weights and neighbourhood sizes have been conducted [4]. No major differences in mean residuals were observed.

The nearest-neighbour estimator can be defined as the mean of a constant number of neighbouring observations [1,3,4]. In the present study, the estimate was weighted by the inverse of the Euclidean distance (Eq. 3).

Equation 3. The nearest-neighbour estimator.

$$\hat{y}_i = \frac{\sum_{j=1}^{n} (\frac{1}{e} y_{ij})}{(\sum_{j=1}^{n} \frac{1}{a})}$$

in which
$y_i$ = The estimated value of the variable i
$y_{ij}$ = The value of the variable i in stand j
n = The number of the nearest neighbours
e = Euclidean distance

Some of the variables were measured accurately and some estimated by employees. The data thus obtained was utilized by creating the model itself using accurately measured values and then the estimated values were used when computing the final results. In the computation phase, one marked stand (the target stand) was extracted from the marked stand population and the estimates were then computed to take into account the thus excluded stand (cross validation system).

## RESULTS

The optimal model was defined by searching for the nearest neighbours with regard to variables presented in Table 2. The computation was done in four phases with different independent variables because using the same variables did not explain all the key figures accurately. Weights and numbers of the neighbours were estimated heuristically by optimizing formulas 2 and 3.

The fact that the dead-branch line and sawmill were not significant enough to be used as independent variables was surprising. One explanation could be that the dead-branch line is not estimated accurately enough and it is used on an accrual basis. On the other hand, other variables probably explain the differences between the sawmills. The sawmill is used as an independent variable when creating diameter and height distributions for spruce, because cross-cutting is controlled by the sawmill. Some sensitivity analyses were done with the number of neighbours. The size of the neighbourhood did not seem to have a powerful influence on the results when there are between 5 to 10 neighbours (e.g. Figure 1). However, when the size of the neighbourhood exceeds 20 or so, the error-percent begins to rise again.

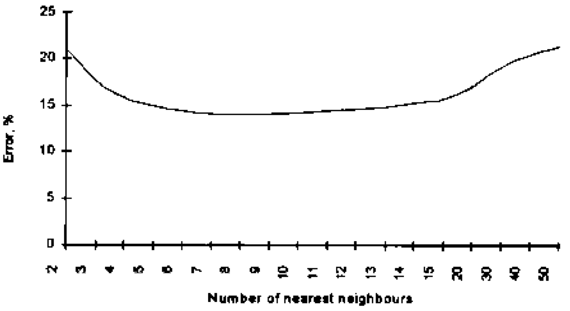

Figure 1. RMSE (%) of sawtimber/pulpwood ratio estimate as a function of the number of nearest neighbours.

Table 2. Independent variables (vertical axis) for tbe various dependent variables (horizontal axis). The numbers of nearest neigbbours and weights (in parentheses).

| | Quality distribution of pine stems by diameter and quality class | Diameter and height| distribution of spruce stems | Sawtimber/l pulpwood ratio of pine stems | Sawtimber/ pulpwood ratio of spruce stems |
|---|---|---|---|---|
| X co-ordinate | *x (0.5)* | *x(0.5)* | x (0.5) | X(0.5) |
| Y co-ordinate | x (0.5) | x (0.5) | x(0.5) | X(0.5) |
| Mode of logging | x(1.0) | x(1.0) | x(1.0) | X(1.0) |
| Size of pine logs | x (2.0) | | x(1.0) | |
| %l of pine | x(1.0) | x(1.0) | x(1.0) | X(1.0) |
| Sawmill | | x(1.0) | | |
| Age class | | x(1.0) | | |
| Stand density | | x(1.0) | | X(1.0) |
| Size of spruce logs | | x(1.0) | | X(1.0) |
| Number of neigbbours | 14 | 19 | 8 | 12l |

When estimating the **sawtimber/pulpwood** ratio for individual stands, the average error in pine stands was 14.7% and to spruce stands 11.3%. This is small considering that the entire material error was not as large. For example, when considering pine stands, the error in the **sawtimber/pulpwood|** estimate is less than 10% in 75% of all the pine stands (Figure 2).

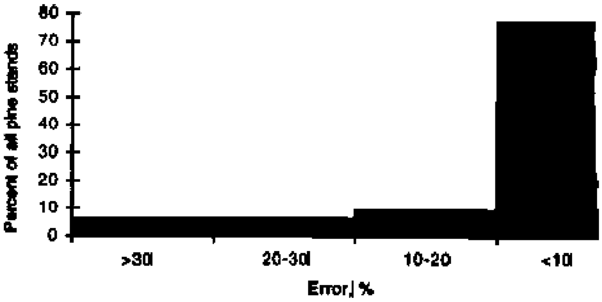

Figure 2. Error **insawtimber/pulpwood|** estimates of pines when considering the entire material.

When estimating the diameter distributions, the model tends to yield results that are closer to the average (e.g. Figure 3). At its best, the model was accurate up to 86.7% in a pine stand and 80.8% in a spruce stand.

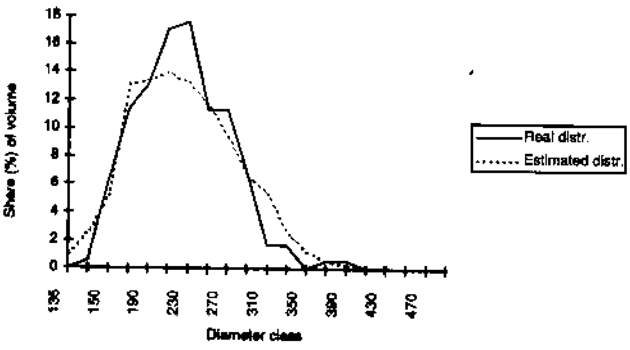

Figure 3. A real and an estimated diameter distribution of pine logs in one stand. This is a typical example of the results of the present study.

Estimating the height distribution of spruce was not so accurate: on average, only 69.9% of all trees were located in the correct class. Changes in cross-cutting may have **beenl** the cause of this result. The quality distribution can be estimated quite reliably for individual stands: on average, 87.6% of the logs were located in the correct quality class. When the quality and diameter distributions are handled together, the average is 74.4 percent.

If more stands are considered for analysis, errors decrease: errors at the stand level compensate each other and the error approaches zero. It becomes easier to plan wood-procurement operations and sawmill production when accurate information is available on incoming raw material. In the present study, the error of every estimated parameter approaches zero.

## DISCUSSION

According to the results obtained in the present study, the nearest-neighbour method can even be used to estimate quality classes as a function to diameter classes when the examination includes several stands.

The results are encouraging: sawtimber/pulpwood ratio estimation was not as accurate as in preharvest measurement-studies, but still more accurate than the results of studies concerning the inventory of stands [5]. Estimating quality classes is about as accurate with the nearest-neighbour method as are the results of a forestry planning inventory. The results of diameter- and height distribution estimates of the present study were not able to compare to the results of other pre-logging measuring methods because there is no scientific experiments of the diameter- and height distribution estimates obtained by traditional pre-logging measuring methods.

Systematic errors have been among the biggest problems associated with planning inventory methods. The nearest-neighbour method does not include this kind of error, but it does involve some significant weaknesses compared to planning inventories: the results yielded by the nearest-neighbour method are strongly contingent on size and quality requirements and control of cross-cutting. In addition, the nearest-neighbour method can not be used to estimate the amount of timber; just the relative values.

The material of the present study contained a lot of data on final-felling stands, which has to take into consideration when surveying the results. More thinning stands in the material could have led to more inaccuracies. The model itself could be improved by adding new, dependent variables, e.g. stand density, site-quality class, and the relative position of the trees.

When this method is used as a part of an information system, it improves the advance information on stands marked for cutting without increasing the foremens' workload as well as making it relatively easy to update the system. The k-nearest neighbour regression method works quite well and is suitable for wood-procurement planning.

When bucking to value and demand approach is used in optimisation of bucking done by harvester, pm-logging information is crucial if we want to control harvester group instead of single harvesters. Control of harvester group is one of the most important areas of research in further development of customised timber procurement.

Artificial intelligence and neural network techniques and methods based on statistical classification could also be utilized to solve this kind of problem. Traditional planning inventory and the nearest-neighbour method together with specific remote-sensing techniques could also be one solution in the endeavour to obtain even more accurate advance information on marked stands and to improve the productivity of wood-procurement operations.

## REFERENCES

[1] Altman, N. S. 1992. An introduction to kernel and nearest-neighbour nonparametric regression. Am. Stat. 46: 175-185.

[2] Haara, A, Maltamo, M. and Tokola, T. 1997. The k-nearest neighbour method for estimating basal-area diameter distribution. Scandinavian Journal of Forest Research 12: 200-208.

[3] Härdle, W. 1989. Applied nonparametric regression, Cambridge university, Cambridge, 323 pp. ISBN O-521-38248-3.

[4] Korhonen, K. and Kangas, A. 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scandinavian Journal of Forest Research 12: 97-101.

[5] Lemmetty, J. and Mäkelä, M. 1992. Suunnittehunittauksen perusteet ja toteutus. Summary: Methods for measurement of a stand for harvest planning. Metsätehon katsaus 11.4 pp.

[6] Moeur, M. and Stage, A. 1995. Most similar neighbour. An improved sampling Inference procedure for Natural Resource planning. Forest Science 41 (2): 337-359.

[7] Suunnittelumittaus. 1996. Enso Ltd. Loppuraportti. 10 pp.

[8] Tolvanen-Sikanen, T. 1994. Tekoäly| metsäteknologiassa¦ University of Joensuu, Fac- ulty| of forestry. Research notes 23.33 pp.

[9] Uusitalo, J. 1995. Pm-harvest measurement of pine stands for sawing production planning.

From Volume 10, No. 1, the following is the missing figure from the paper entitled "The Barycentric Coordinates Solution to the Optimal Road Junction Problem" by Francis E. Greulich.

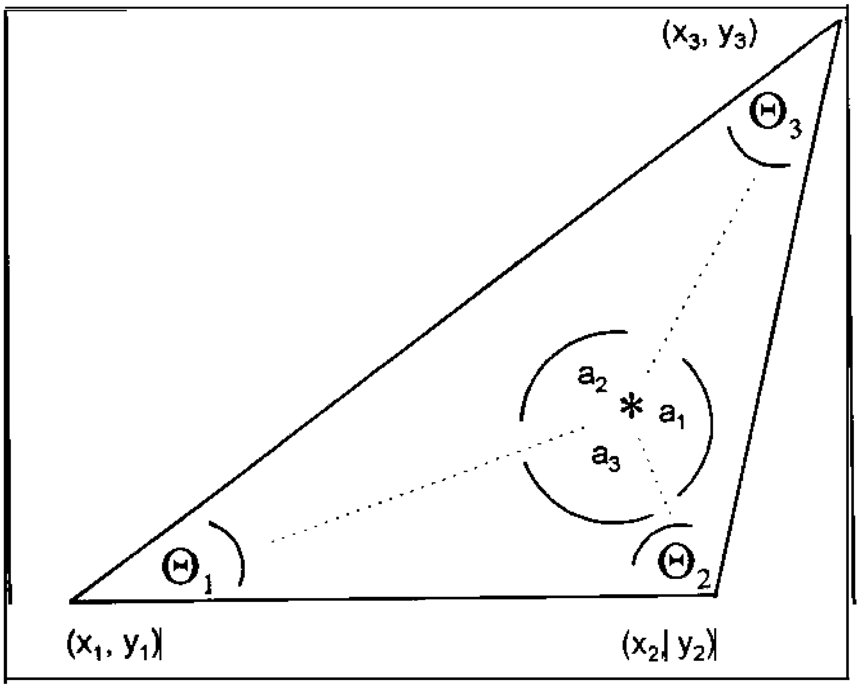

Figure 1. A location triangle with counter-clockwise control point indexing. The point to be resected is indicated with an asterisk.