# Case-Based Support for Forestry Decisions: How to See the Wood from the Trees

Conor Nugent[1⋆], Derek Bridge[2], Glen Murphy[3], Bernt-Håvard Øyen[4]
nugentc@gmail.com, d.bridge@cs.ucc.ie,
glen.murphy@oregonstate.edu, and
bernt-havard.oyen@skogoglandskap.no

[1] Idiro Technologies,
Ireland.
[2] Department of Computer Science,
University College Cork, Ireland.
[3] Forest Engineering, Resources and Management Department,
Oregon State University, Corvallis, Oregon, USA.
[4] Norwegian Forest and Landscape Institute,
Bergen, Norway.

**Abstract.** In forestry, it is important to be able to accurately determine the volume of timber in a harvesting site and the products that could potentially be produced from that timber. We describe new terrestrial scanning technology that can produce a greater volume of higher quality data about individual trees. We show, however, that scanner data still often produces an incomplete profile of the individual trees. We describe Cabar, a case-based reasoning system that can interpolate missing sections in the scanner data and extrapolate to the upper reaches of the tree. Central to Cabar's operation is a new asymmetric distance function, which we define in the paper. We report some preliminary experimental results that compare Cabar with a traditional approach used in Ireland. The results indicate that Cabar has the potential to better predict the market value of the products.

## 1 Introduction

Forest planners are responsible for deciding how the set of commercially-cultivated forests that are under their control should be developed and eventually harvested. At any given time, a number of different forests are available, and planning how best to utilize them can be a difficult task. For example, forest planners must combine information that comes from processing plants (e.g. sawmills) with information about their forests to decide which forests to use and which trees within those forests to fell.

⋆ This work was carried out while the first author was a member of the Cork Constraint Computation Centre (4C) at University College Cork. The project was an Innovation Partnership (IP/2006/370/), jointly funded by Enterprise Ireland and TreeMetrics, a company that provides forestry measurement systems.

Poor decisions in this planning process are commonplace. But they are not necessarily due to poor judgement. They are often caused by inadequate information at the root of the supply chain, the forest. Forests exhibit high inherent spatial variability (e.g. two trees growing side by side may exhibit different characteristics due to differences in genetics and micro-climate) and temporal variability (e.g. trees continue to grow after being measured). In the vast majority of cases the characteristics of a forest are only vaguely known. This means that resources that have been cultivated over periods of thirty to more than a hundred years are often underutilized based on ill-informed decisions made quickly at the end of their life cycles [1–4]. These poor decisions and the lack of quality forest information naturally have knock-on effects right through the supply chain.

In this paper, we focus on three forest planning tasks:

**Tree taper estimation:** The task here is to estimate the diameter of a tree stem at different heights. Diameters tend to taper, i.e. they decrease with height, and this is why this is known as taper estimation.

**Stem volume estimation:** The task here is to estimate the volume of timber that a tree will produce. This may be based on estimates of tree taper.

**Product breakout estimation:** The task here is to estimate what products can be produced from a tree, e.g. size and quality of planks, amount of wood-chip, and so on. This may be based on estimates of stem volume.

The rest of this paper is structured as follows. Section 2 presents state-of-the-art methods for the tasks listed above; it explains the role of new scanner technologies; it motivates the use of Case-Based Reasoning (CBR) to exploit the scanner data; and it describes related work in CBR. Section 3 presents Cabar, our case-based reasoning system for the tasks listed above. The focus in the section is on case and query representation, along with a new asymmetric distance function that we have defined. Section 4 describes experiments we have conducted and presents our preliminary results.

## 2 The State of the Art: Motivating the Use of CBR

### 2.1 Current Practices

Much of current practice revolves around the prediction of the expected volume of a forest, i.e. the amount of timber a forest is expected to yield. This aids the selection of which of a set of forests to harvest.

In Ireland, a common approach is to take a set of measurements about a forest and use them to access a simple set of look-up tables that translate these measurements into volume figures. The forest manager conducts a survey of the forest, which records the diameter at breast height (DBH) and, perhaps, the height of a number of sample trees.[5]. From these measurements, and perhaps also

[5] In some parts of the world, the survey may also include assessments of stem shape, curvature, and quality (e.g. size of branches, scarring, rot, wood density, etc.) [5].

the forest age and the thinning strategy, the manager can then read-off predictions of the expected volume and, sometimes, the typical dimensions of saw-logs that the forest can yield. The look-up tables are compiled from extensive field measurements and mathematical models. The disadvantages of this approach include: the tables that are available to the manager may not adequately reflect local conditions (soil, weather, tree species, species mixture, etc.); the predictions are made from only a sample of trees in the forest and from only one or two measurements about each tree; and the prediction is only a crude estimate of overall forest volume, and not individual tree volume.

An alternative is to predict volumes on a tree by tree basis. This is usually done by predicting the diameter of the tree stem at different heights along the stem, from which the volume of the tree can be calculated. The equations for predicting diameters are known as taper equations. For the most part, the input parameters are the DBH and the height of the tree [6, 7]. In many cases, the height is not measured; rather, it is estimated from the DBH using a height model [8, 9]. Many different taper equations exist, each making different assumptions about tree shape, and hence using different geometrical principles and mathematical functions. It is necessary to choose the right equation for the species of tree and to calibrate the equation based on local conditions and historical data. The disadvantages of this approach include: the equations that are available to the manager may not adequately reflect local conditions; and the equations use only small amounts of data about the tree (sometimes just the DBH).

From full taper and volume predictions, it is possible to estimate the product breakout, i.e. the products that might be produced from a tree [10]. Some forest managers use software to do this. The software simulates the algorithms that are used in the field by harvesting machines, as explained below.

Trees are rarely transported from the forest as complete units. They are usually first cross-cut into smaller units (logs). The harvesting machine's onboard software decides how to cross-cut a tree. Obviously, its decisions have a major bearing on what products the sawmill will ultimately be able to produce. The harvester is pre-loaded with data about the products to be cut and their priorities (in the form of a set of weights). The harvester begins by taking hold of the base of a tree; it then both measures and infers the dimensions of the tree; and it uses a priority cutting list or a mathematical programming technique (e.g. dynamic programming [11], branch-and-bound [12], network analysis [13]) to determine the optimal or near-optimal way to cross-cut the tree.

By simulating the harvesting machine's decisions in advance on taper and volume predictions, a forest manager can decide which trees to harvest. But the disadvantages include that poor predictions of taper and volume may render estimates of product breakout too unreliable to be useful in practice.

### 2.2 New Technologies

New technologies offer the potential to overcome the lack of information about forests. They may enable us to obtain a greater volume of higher quality data, and to do so at low cost.

In our research, for example, we have been working with a company called TreeMetrics, who provide forestry measurement systems (`www.treemetrics.com`). TreeMetrics has developed a portable 3-dimensional terrestrial laser scanning technology [14, 15]. Their technology makes it possible to capture 3D data about standing trees in a forest prior to harvesting. For each individual tree in a scanned plot the scanner can record hundreds of laser readings. TreeMetrics' software takes a set of readings, tries to work out which tree each reading belongs to, and calculates tree diameters at fixed intervals along the length of the tree. For each diameter, the centre point is also calculated and so curvature information about the tree is also obtained. Such information makes it possible to accurately determine the volume and a quality attribute (from the curvature data) of trees in a forest in advance of harvesting.

### 2.3 A Role for Case-Based Reasoning

Although TreeMetrics' technology provides vastly more tree information than some more traditional approaches, it is not guaranteed to provide a complete picture of each tree. Readings for some sections of a tree may be missing due to occlusion by branches or other trees. This becomes increasingly common the further up the tree the readings are sought due to the effects of branching and the limitations of the laser at increased distances.

This leaves us with a tree profile prediction task, both in terms of interpolating the missing sections and also in terms of extrapolating to the upper reaches of a tree. There is also the task of smoothing or replacing diameter readings which contain noise. In this paper we outline a Case-Based Reasoning (CBR) approach which accomplishes these tasks.

We expect various advantages to accrue from the combined use of better data in greater volumes and the use of CBR in place of pre-compiled look-up tables and equations. These advantages include:

**Flexibility:** As described in Section 2, in Ireland the traditional approach is to make predictions from measurements taken at fixed points, such as the diameter at breast height. Even when more measurements are available, such approaches cannot capitalize on the extra data. Equally, they cannot be used to make predictions if the data they need is not available. The CBR system that we propose can make predictions based on whatever readings are available. In particular, the new similarity measure that we define (Section 3.2) handles any number of readings, and accommodates information about the certainty of those readings.

**Localized predictions:** In traditional approaches, models (e.g. systems of equations) can be calibrated to local circumstances. However, this is complex and the costs of doing it are typically high. Hence it is common to generalize over broad regions, which means that the models often fail to capture finer-grained local variations. CBR offers an approach that can be readily localized. The case base used for Sitka Spruce forests in southwest Ireland, for example, need not be the same as the case base used for Norway Spruce

forests in southeastern Norway. Case bases can be built from harvester data or field measurements that come from the particular area in which the case base will be used, thus implicitly capturing the local characteristics of that area. This is not cost-free, but it is simpler than model calibration.

**Immediate use of data:** With CBR, there is no need for model calibration. In some sense, calibration is implicit: the particular cases in the case base calibrate the system to its local circumstances. Equally, since CBR is a strategy for both problem-solving and learning, by judicious case base update, case bases can be tailored to local conditions over time.

### 2.4 Related Work

Applications of CBR in forestry, while few in number, have a long history. In 1997, Goodenough et al. [16] described the SEIDAM system that tries to keep forest inventories up-to-date through the integration of images and digital maps. Their use of CBR is quite different from ours: they use it to form plans of map update operations. In 1998, Kelly and Cunningham [17] investigated an algorithm for selecting an initial case base from a database of Irish forestry data. Our data is about individual trees, whereas the records in their database provide information about forest 'sub-compartments'. Their CBR system is judged by how well it predicts the proportion of a sub-compartment that should be planted with a particular species. There are also a number of examples of problems within forestry, including the problem of estimating taper on unmeasured portions of a felled tree stem while the stem is being harvested, for which solutions based on *k-NN* have been suggested [18–20, 4].

## 3 Cabar: A CBR Forestry System

Cabar is the name of the CBR system that we have developed.[6] Cabar is designed to deal with the particular characteristics of tree stem data, especially the kind of data that we can obtain from TreeMetrics' laser scanner technology. The emphasis is on taper prediction, which we then use for stem volume prediction, which we use in its turn for product breakout estimation. Cabar's design and operation are explained in the next three subsections.

### 3.1 Cases and Queries

Each case in our case base represents one tree and contains a sequence of real values, which denote the diameter of the tree, usually at 10 cm intervals along its entire length. Sometimes cases come from manual field measurements. But they are also readily available from harvester machines. The on-board software records an overbark profile of each tree that the harvester cuts. There is no solution part to the cases.

[6] "Cabar" is the Gaelic word, often spelled "caber" in English, for a wooden pole.

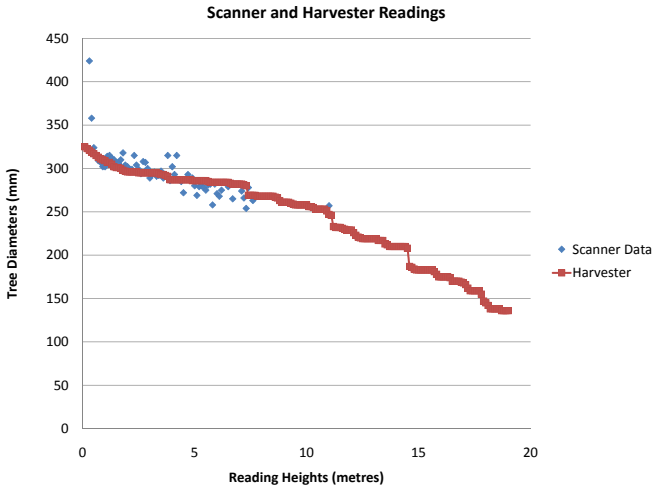


**Fig. 1.** Scanner readings and harvester measurements of the diameter of a tree

Queries in our system are also described by sequences of real values, denoting the diameter of the tree at different heights. Whereas cases may be standing trees that we have measured manually or trees that a harvester has previously felled and measured, queries usually describe standing trees whose taper and volume we wish to predict. They have been scanned, and we have a set of readings from the laser scanner. Hence, we say that cases contain *measurements* and queries contain *readings*.

However, as mentioned above, due to occlusion and the limitations of the scanning technology, a query is often a fairly incomplete impression of the tree profile. This is illustrated in Figure 1, which shows the profile of a tree as recorded by TreeMetrics' scanning device (diamonds) and the profile of the same tree as recorded by a harvester after felling (squares).

The figure shows that the scanner data is noisy (due, e.g., to nodal swelling, dead branch stubs, dead needles, partially hidden stem, etc.), and there are sections of the tree for which the scanner has no readings. Taper prediction involves interpolation of missing sections, extrapolation to the upper reaches of the tree, and smoothing of noise.

In fact, for each reading in a query, we also have a confidence measure. This is a measure of the confidence the TreeMetrics software has in the accuracy of the reading. The software uses 3D image recognition techniques: the 3D coordinates recorded by the scanner are assigned to different trees. The height at which a reading is taken can be determined to a high level of accuracy, but diameters are less reliable due, e.g., to occlusion by parts of neighbouring trees. (There is no

equivalent to these confidence measures in cases because, as we described above, cases describe trees that have been properly measured, either manually or by a harvester. We assume that all readings in cases are ones we can be confident in, although this can be achieved only under carefully controlled circumstances.)

Formally, a case $c$ is a set of pairs, $c = \{\langle i_1, x_1\rangle, \ldots, \langle i_m, x_m\rangle\}$, where $x_1, \ldots, x_m$ are measurements, usually stem diameters, and $i_1, \ldots, i_m$ are the heights at which the measurements were made, e.g. $\langle 5, 300\rangle$ means that at a height of 5 m from the base, the tree's diameter is 300 mm. For each case, it is expected that $i_1, \ldots, i_m$ will be consecutive heights.

A query $q$ is a set of triples, $q = \{\langle i_1, x_1, w_1\rangle, \ldots, \langle i_n, x_n, w_n\rangle\}$, where $x_1, \ldots, x_n$ are stem diameters calculated from scanner readings, $i_1, \ldots, i_n$ are the heights at which the readings were taken, and $w_1, \ldots, w_n$ measure confidence in the readings. For queries, it is typically not the case that heights $i_1, \ldots, i_n$ are consecutive.

### 3.2 Similarity

We need to be able to compute the similarity between cases and queries. In fact, we define a distance function, rather than a similarity measure, whose design is informed in part by the following observations about cases (usually harvester data) and queries (scanner data):

**Sequence data:** Cases and queries both have the characteristics of sequence data. Each data point $x_j$ has a definite relationship with those either side of it, $x_{j-1}$ and $x_{j+1}$. The data is effectively the description of a shape. There is an analogy here between our cases and temporal cases, where values are recorded at different points in time.

**Varying lengths:** Stems vary in height and the raw series data reflect this. Since measurements and readings are taken at fixed intervals, the length of a sequence varies from stem to stem. Cases need not be the same length; queries need not be the same length; and queries need not be the same length as cases. This means that any vector-based similarity measures that assume fixed-length vectors cannot be used directly.

**Partially incomplete:** Both the cases (harvester data) and queries (scanner data) can be incomplete sequences, in the sense that there may not be measurements or readings at certain heights. However, they are incomplete in distinct ways. The harvester data will have a measurement at every interval up to a certain height but may not completely record the final taper of the stem. The point at which the sequence terminates varies with each individual file. The scanner data in contrast contains many missing sections of data due to occlusion from branches and other trees. These effects become more prominent further up the stem. As a result there is typically more information about the sequence at the base of the tree; readings are fewer and more sparsely distributed further up the stem.

We investigated a number of variations of an existing shape-based similarity measure [21], but without great success. We believe that this is because this

measure, the variants we tried, and others like it tend to assume that cases and queries are quite homogeneous and symmetric. In our forestry system, however, we have seen that cases and queries are quite different from each other.

For this reason, we have defined ASES, our own asymmetrical sequence-based Euclidean distance function, which we believe is well-suited to the task at hand:

$$ASES(q,c) = \sqrt{\frac{\sum_{\langle i,x,w\rangle \in q} w \times \mathit{diff}(i,x,c)}{\sum_{\langle i,x,w\rangle \in q} w}} \tag{1}$$

where

$$\mathit{diff}(i,x,c) = \begin{cases} x^2 & \text{if } i \notin \{i' | \langle i', x'\rangle \in c\} \\ (x-x')^2 \text{ such that } \langle i, x'\rangle \in c & \text{otherwise} \end{cases} \tag{2}$$

In essence, this global distance measure is a weighted sum of local distances, where the weights are the confidence measures from the query. For each reading in the query, a local distance is computed. If the query contains a reading $\langle i, x, w\rangle$ and the case contains a measurement that was taken at the same height $i$ (i.e. if it is the case that $\langle i, x'\rangle \in c$), then the local distance is the square of the difference between the query reading and the case measurement, $(x-x')^2$. If the case does not contain any measurement taken at height $i$ (i.e. $i \notin \{i' | \langle i', x'\rangle \in c\}$), then the local distance is the square of the whole amount of the reading, $x^2$. This has the effect of penalizing cases that are too short to match all the readings in the query.

### 3.3 Retrieval and Reuse

Of the four phases in the CBR cycle the two most critical in Cabar are the retrieval and reuse phases. At present, we have experimental results only for quite simple versions of these phases. We have tried more sophisticated techniques, but we will not describe them here because they have not been verified experimentally. We explain how we carry out tree taper prediction, stem volume prediction, and product breakout estimation.

**Tree Taper Prediction** Given a query $q$, we retrieve its $k$ nearest neighbours using the ASES distance measure. In the simple approach for which we have experimental results, we use only $k = 1$. We use this nearest neighbour $c$ for query completion. In fact, the only approach for which we have experimental results is the very simplest one: we take the whole of $c$ unchanged (i.e. without adaptation) in place of $q$.

**Stem Volume Prediction** Stem volume prediction is trivial once the query has been completed. It involves no more than computing the volume based on the inferred diameters.

**Product Breakout Estimation** Product breakout estimation is the most complex task that we carry out. Basically this involves giving a prospective processing plant such as a sawmill a sense of the likely products that can be produced from the stems captured by the scanning device.

As explained in Section 2.1, given a specification of the products that a processing plant is interested in, we can use a cross-cutting algorithm to simulate the actions that a harvester machine would carry out in the forest. Such cross-cutting algorithms are commonly used in forestry and we have designed and developed an adaptation of one which can utilize the extra 3D information captured by the TreeMetrics scanning device. The extra information gives the algorithm knowledge of tree curvature. Our algorithm therefore gives more accurate estimates of the breakout because it better takes problems of curvature into account.

The cross-cutting algorithm we use is an adaptation of the branch-and-bound approach proposed by Bobrowski, which is proven to produce the optimal solution [12]. As each possible solution path in its search tree is evaluated, our variant of Bobrowski's algorithm ensures that any restrictions the products place on acceptable curvature levels are taken into account.

## 4 An Experimental Evaluation of Cabar

Although being able to examine Cabar in real world situations is the ultimate proof of concept, such studies often contain multiple sources of error. Such errors, especially when unquantifiable, make it difficult to properly assess the performance of a new technology and impossible to isolate different areas of failure. In this paper we wish to examine and quantify the performance of Cabar over a range of different scenarios, in particular where there are varying quantities of data and varying levels of noise. To achieve this we developed a testbed to simulate these situations.

The benchmark against which we compare Cabar's performance is similar to many used in forestry and much the same as those described in Section 2.1. It predicts tree taper and then stem volume using a taper equation whose parameter is a DBH measurement, and it estimates product breakout using a harvester simulation model.

We first describe our experimental data; then we describe our benchmark system; finally we present the results of our experiments.

### 4.1 Experimental Data

In our experiments we use harvester data, which we assume to be accurate. Obviously, this is the source of our case data. But it is also the source of our query data. We applied ablation and noise functions to harvester data in order to simulate scanner data of varying quality. The reason we create queries from harvester data is that we need to have a known 'ground truth' against which we can compare Cabar's predictions. What we need in future are readings produced by laser scanning a stand of trees paired with harvester measurements on the

same trees after felling. Lacking enough data of this kind, we were obliged to use harvester data alone. We expect this to be remedied in the future.

In particular, the harvester files used in our experiments are Sitka Spruce tree files from Ireland. We removed from this set of harvester files any which did not have a complete set of readings up to the 70 mm diameter point and also any files in which substantial discontinuities occurred. This left a set of 389 tree stems. We then split the remaining data into two separate, equal-sized data sets. One set was used to generate queries of simulated scanner data; the other formed a case base. We applied ablation and noise functions to the measurements in the query data in order to simulate scanner data, as described in detail below. In our experiments, confidence levels in the query data are all set to 1.

### 4.2 Ablation and Noise Functions

To decide whether to retain or delete a measurement $\langle i, x \rangle$ in a query, we sample a random distribution. If a sampled value $r$ is greater than probability of retention $P(\alpha, \beta, i)$, then $\langle i, x \rangle$ is retained, and otherwise it is deleted. Retention/ablation probabilities are based on the generalized exponential distribution [22]:

$$P\left(\alpha, \beta, i\right) = e^{-\frac{(i-\alpha)}{\beta}} \tag{3}$$

$\alpha$ defines the shape of the distribution, and $\beta$ is a scaling factor. We can adjust the extent to which the readings that are higher in the tree are retained by changing the value of $\beta$. For the remainder of this document we will refer to $\beta$ as the ablation factor.

Figure 2 shows an example harvester tree stem to which we have applied our ablation function. The points remaining after ablation are shown as squares. In this figure the effects of adding noise can also be seen (stars). The noise we added was normally distributed with a variance set to be a percentage of the original diameter. Our Treemetrics expert informally confirmed that this simulated scanner data strongly resembles real scanner data.

We applied varying degrees of ablation and noise to the queries. We altered the noise levels from no noise at all up to a noise level of 20%.[7] The ablation factor took on the values 1, 1.5 and 2.

### 4.3 The Malone Kozak Benchmark System

In order to benchmark the performance of the CBR system against a realistic alternative, we developed a system similar to one of the approaches described in Section 2.1. One important element of this estimate mechanism is the taper equation used. We used a taper equation called the Malone Kozak. The Malone Kozak has been especially calibrated for Sitka Spruce in Ireland [10]. This taper

[7] Noise levels of up to 20% seemed reasonable at the time these experiments were run. Data we have recently acquired for South Australian stems shows that we may need in future to use a slightly larger variance.

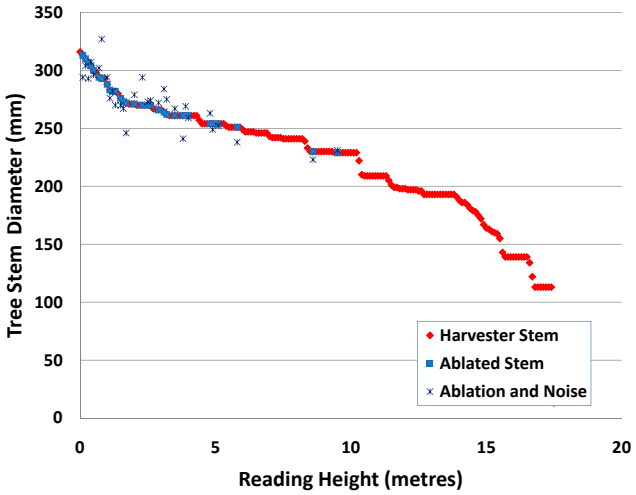


**Fig. 2.** A example of a query and the harvester stem data from which it was created

equation uses DBH and height measurements as its inputs. But since height measurements are often not available, the Malone Kozak comes with a height model that predicts height based on DBH. Since our data-sets did not give us height measurements, we used the height model in our experiments. In generating queries from harvester data (above), we were careful to ensure that we did not ablate the DBH of each of the query trees and that we did not apply noise to the DBH. The Malone Kozak would be particularly susceptible to such noise and would be unusable without a DBH measurement.

For product breakout estimation, the benchmark system uses the same cross-cutting algorithm as the one we developed for Cabar (Section 3.3).

### 4.4 Experimental Methodology

We took the queries and applied a particular level of ablation and noise. We then used Cabar to predict tree taper, stem volume, and product breakout market value. We repeated this 20 times using the same noise and ablation factors, and averaged the results. We then changed the noise and ablation factors and repeated this process for each different level of noise and ablation.

The Malone Kozak system, on the other hand, was run only once on the query set. Only one run is needed because we ensure that the noise and ablation factors do not alter the true DBH, which is the only input in the Malone Kozak equations that we use (Section 4.3).

**Table 1.** A description of the products harvested from an Irish forest

| Product | Length (m) | SED (mm) | Market Value |
|---|---|---|---|
| 4.9 Saw log | 4.9 | 200 | 39 |
| Pallet | 3.1 | 140 | 25 |
| Stake | 1.6 | 70 | 19 |

To perform product breakout estimation, we need an example of product demand. For this, we chose a small set of products which are typical of those harvested in Ireland. A description of these products can be seen in Table 1. The single most important feature of each product is its length. However, each product is further described by other characteristics that restrict whether certain sections of a given tree can produce such lengths. We use the small end diameter (SED), which describes the minimum diameter that the upper or smaller diameter of a cut section is allowed to be. These restrictions are taken into account by our cross-cutting algorithm.

The cross-cutting algorithm also needs priorities, denoting the importance of each type of product. For our experiments, we used indicative market values as the set of weights describing the priorities.

The final outcome of the cross-cutting algorithm is the estimated overall value of the whole query set using the market values in Table 1. We compare these estimates from both systems to the 'ground truth', i.e. the market value of the products that can be cross-cut from the original, noise-free, unablated query tree data. We report the percentage error.

### 4.5 Results

Figure 3 shows the percentage error in the estimates of total product breakout market values for the query set. Two systems are compared: Cabar and Malone Kozak (with a generic height model). In the case of Cabar, there is a separate trend line for each of the three ablation factors that we used. The graph plots the error against different levels of noise. The Malone Kozak system is insensitive to this, as explained in Section 4.4, because we use only the DBH and we ensured that this was noise-free and never ablated. It can be seen that Cabar's error levels are far below the Malone Kozak levels. Of course, if the Malone Kozak system had been calibrated not just for Sitka Spruce in Ireland, but for the particular stand of trees used in the data-set, then its performance would be more competitive. This reinforces the point about the importance (and cost) of model calibration in approaches like this one. Similarly, supplying tree height readings if they had been available instead of using the height model would have made the Malone Kozak system more competitive.

The source of Cabar's greater accuracy becomes clearer when we look at the product breakout predictions for particular noise and ablation factors. Figure 4, for example, shows the product breakout volumes by product type when the

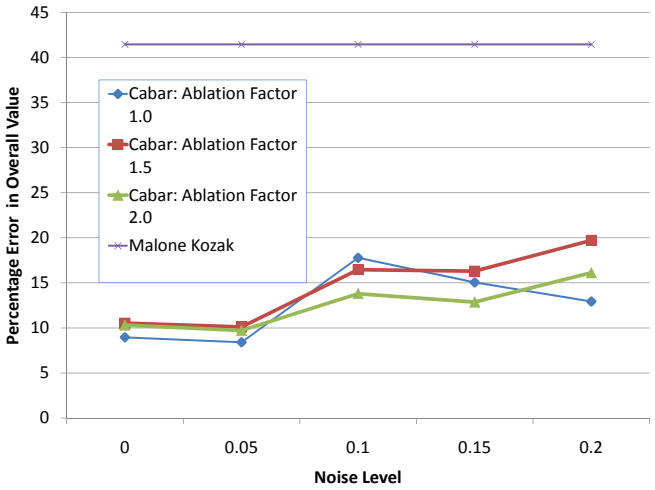


**Fig. 3.** Product breakout market value error results

noise and ablation values are 0.15 and 1.5 respectively. Cabar's predictions are generally reasonably close to the figures produced by the cross-cutting algorithm on the actual stems (first two bars in each group of three). Malone Kozak, by contrast, tends to greatly overestimate the sizes of the trees and hence the product breakout volumes (third bar in each group). This indicates that the trees used in the harvester data-set were shorter and had greater taper than predicted by the Malone Kozak equations and height model.

## 5 Conclusions and Future Work

Improvements in terrestrial scanner technology mean that it is now possible to collect far more information about a forest in advance of it being harvested. However, this data does not offer complete profiles of the trees in the forest. Systems that can 'fill in the gaps' are needed. Traditional approaches to these estimation tasks were not designed with such rich data sources in mind and are unable to exploit them. In this paper we presented Cabar, a Case-Based Reasoning approach, which is better suited to dealing with the abundance of scanner data but also the challenges such data poses. Our preliminary results demonstrate that Cabar provides a viable alternative solution to some of the challenges currently hindering the adoption of terrestrial scanning in forestry.

Although the work we have presented demonstrates that a CBR approach to this problem is promising, there are several possible ways in which it might be

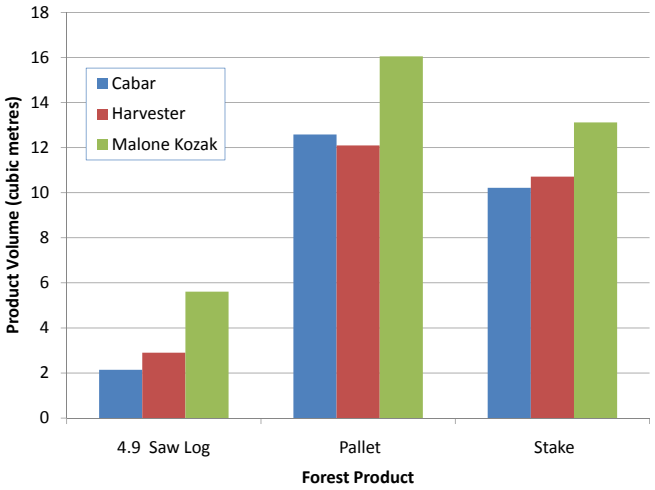


**Fig. 4.** Product breakout volumes (noise factor 0.15; ablation factor 1.5)

improved. The first step would be to extend our system to operate on the $k$ nearest neighbours for $k > 1$ and to develop a more sophisticated approach to case completion from the nearest neighbours. We are also contemplating alternative distance functions that take account of area rather than diameter, which would tend to give greater weight to differences nearer the base of the tree. We are also considering how to make more use of the curvature data that the scanning technology gives us. We use it in our cross-cutting algorithm, but it is not used in the distance function. Using it in the distance function raises methodological problems because no harvester measures this attribute, hence 'ground truth' figures would not be readily available. More empirical evaluation is also called for. We are collecting further data sets on which experiments can be run. In particular, it is likely that we will have data that contains scanner readings and harvester measurements for the same trees. With this, we will not need to use simulated queries, and we can investigate the role of the confidence measures.

## References

1. Boston, K., Murphy, G.: Value recovery from two mechanized bucking operations in the Southeastern United States. Southern Journal of Applied Forestry **27**(4) (2003) 259–263
2. Murphy, G.: Mechanization and value recovery: worldwide experiences. In: Forest Engineering Conference: Forest Engineering Solutions for Achieving Sustainable Forest Resource Management – An International Perspective. (2002) 23—32

3. Kivinen, V.P.: Design and testing of stand-specific bucking instructions for use on modern cut-to-length harvesters. PhD thesis, University of Helsinki (2007)
4. Marshall, H.D.: An Investigation of Factors Affecting the Optimal Log Output Distribution from Mechanical Harvesting and Processing Systems. PhD thesis, Oregon State University (2005)
5. Gordon, A., Wakelin, S., Threadgill, J.: Using measured and modelled wood quality information to optimise harvest scheduling and log allocation decisions. New Zealand Journal of Forestry Science **36**(2/3) (2006) 198–215
6. Newnham, R.M.: Variable-form taper functions for Alberta tree species. Canadian Journal of Forest Research **22** (1992) 210–223
7. Lappi, J.: A multivariate, nonparameteric stem-curve prediction method. Canadian Journal of Forest Research **36**(4) (2006) 1017–1027
8. Uusitalo, J.: Pre-harvest meaurement of pine stands for sawlog production planning. Department of Forest Resuorce Management Pubications, University of Helsinki, Finland (1995)
9. Curtis, R.O.: Height-diameter and height-diameter-age equations for second growth Douglas fir. Forest Science **13**(4) (1967) 365–375
10. Nieunwenhuis, M.: The development and validation of pre-harvest inventory methodologies for timber procurement in Ireland. Silva Fennica **36**(2) (2002) 535–547
11. Nasberg, M.: Mathematical programming models for optimal log bucking. PhD thesis, Linköping University (1985)
12. Bobrowski, P.M.: Branch-and-bound strategies for the log bucking problem. Decision Sciences **21**(4) (1990) 1–13
13. Sessions, J., Layton, R., Guangda, L.: Improving tree bucking decisions: a network approach. The Compiler **6**(1) (1988) 5–9
14. Bienert, A., Scheller, S., Keane, E., Mohan, F., Nugent, C.: Tree detection and diameter estimations by analysis of forest terrestrial laserscanner point clouds. In: ISPRS Workshop on Laser Scanning 2007. (2007) 50–55
15. Bienert, A., Scheller, A., Keane, E., Mulloly, G., Mohan, F.: Application of terrestrial laser scanners for the determination of forest inventory parameters. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **36**(5) (2006)
16. Goodenough, D.G., Charlebois, D., Bhogal, A.S., Matwin, S., Daley, N.: Automated forestry inventory update with SEIDAM. In: Procs. of the IEEE Geoscience and Remote Sensing Symposium. (1997) 670–673
17. Kelly, M., Cunningham, P.: Building competent compact case-bases: A case study. In: Procs. of the Ninth Irish Conference in Artificial Intelligence & Cognitive Science. (1998) 177–185
18. Amishev, D., Murphy, G.: Implementing resonance-based acoustic technology on mechanical harvesters/processors for real-time wood stiffness assessment: Opportunities and considerations. International Journal of Forest Engineering **19**(2) (2008) 49–57
19. Nummi, T.: Prediction of stem characteristics for pinus sylvestris. Scandinavian Journal of Forest Research **14** (1999) 270–275
20. Liski, E., Nummi, T.: Prediction of tree stems to improve efficiency in automatized harvesting of forests. Scandinavian Journal of Statistics **22** (1995) 255–269
21. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**(2) (2000) 99–121
22. Balakrishnan, N., Basu, A.P.: The Exponential Distribution: Theory, Methods, and Applications. Gordon and Breach, New York (1996)