Jul 1st, 12:00 AM

Applying Data Mining Methods for Forest Planning Data Validation

A. Mäkinen

A. Kangas

T. Tokola

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Applying Data Mining Methods for Forest Planning Data Validation

A.Mäkinen[a], A.Kangas[a] and T.Tokola[b]

[a]*Department of Forest Resource Management, University of Helsinki, Latokartanonkaari 11 00014 University of Helsinki, Finland (antti.makinen@helsinki.fi)*

[b]*Faculty of Forestry, University of Joensuu (timo.tokola@joensuu.fi)*

Abstract: Decision making in forest planning is based mostly on simulated forest management scenarios. A fundamental tool in creating these scenarios is the forest planning system, which utilizes a set of models for projecting the future development of forests and assessing the effects of alternative management tasks, such as timber harvests [Burkhart 2003]. Input data for forest planning is obtained from several different sources, such as remote sensing and field measurements and visual assessments. All data collection systems include errors, both due to human and technical sources, which eventually affect the quality of forest plans. Part of the errors can be considered as next to impossible to detect but part of the errors are outliers and can be separated from the data. Statistical outlier detection methods have been used in data processing, although the statistical methods do not work well for multi-dimensional data. Data mining offers some interesting possibilities for the outlier detection task. The different data mining schemes for outlier detection include for example distance-, density- and clustering –based algorithms that have been proven to work with multi-dimensional data. In the field of forest research, data mining methods have not been studied almost at all. In this study we compared three different outlier detection schemes for finding the outliers in a large forest inventory data. The tested algorithms were Nested-Loop distance-based outlier detection [Knorr & Ng 1998], Simple-Pruning distance-based outlier detection [Bay & Schwabacher 2003] and Outlier Removal Clustering [Hautamäki et al. 2005]. The data included a total of 5090 field measured sample plots on 578 forest stands with a number of stand-level aggregate attributes representing different characteristics of the growing stock, the forest site and the surrounding region. Each of the examined methods has a number of parameters having a strong effect on the outlier detection result. Also the selection of the attributes which were used in the outlier detection strongly affected the results. None of the three methods proved to be superior compared to the others in finding the outliers. The large natural variation in the forest attribute values made the task of separating the outliers difficult. However, the examined data mining methods showed very promising results in finding outliers.

Keywords: *Forest Planning; Data Mining; Outlier Detection*

1. INTRODUCTION

The purpose of forest planning is to offer support for the decision maker [Buongiorno & Gilless 2003]. Generally the goal in forest planning is to find the best way to manage forest treatment units, forest stands. Forest planning systems are computer decision support tools that are used for evaluating alternative forest treatment scenarios. The planning systems include models for predicting the future development of forests, eg. forest growth models and models for predicting the possible outcomes from alternative forest treatment scenarios, eg. yield models [Burkhart 2003]. Usually some kind of optimization technique is applied for finding an optimal forest management scenario which will maximize the benefit of the decision maker. Errors in the inputs of a forest growth model will result in errors the predictions about the future. This in turn might lead to sub-optimal forest management, causing for example economical losses or losses in the social or ecological

values of the forest. This is why the quality and reliability of the forest planning data has a major effect on the quality of forest planning.

Most forest planning systems require some basic information about the forest site properties, the growing stock and the geographical area as inputs. Attributes that contain information about the forest site can include for example soil type and growth potential. Information about the growing stock is usually given as a list of trees and tree properties or some aggregated stand-level mean attributes such as mean diameter, mean height, volume of trees and main species.

Data for forest planning in Finland is collected mainly through partly subjective visual estimation of stand-level aggregate attributes. The method, though cost-effective, is highly subjective and prone to human errors [Haara 2003, Kangas et al. 2004]. In this method the expertise and experience of the forest planner has a strong effect on the quality of the data. The quality of the data can rarely be verified, except by labour intensive and costly reference measurements in the field. As the data for forest planning is acquired from multiple sources and in many cases without good estimates on the data quality, forest planning would benefit from tools that could be used to identify at least the coarsest errors in a data set.

The quality of the forest planning data varies between different forest types and inventory methods. In stand-wise visual estimation, the errors in the data can be caused by sampling, assessment and classification errors [Haara 2003]. According to different studies, the relative mean square errors (RMSE) for estimation accuracies of different aggregate attributes in stand-wise subjective field estimation range from 10 to 80 percent, with large variation between different attributes. For example the RMSE for mean height of trees has been reported to range from 12 to 17 percent whereas the RMSE for number of trees per hectare has been around 80 percent. [Poso 1983, Laasasenaho & Päivinen 1986, Kangas 2004]. The correlation of estimation error between different variables or the estimation error for categorical variables, such as soil type, has not been reported.

Recent developments in various forest inventory technologies, mostly in remote sensing, provide different types of data that varies in accuracy and reliability. With novel remote sensing techniques, such as airborne laser scanning, the RMSE's have been reported to range from around 5 to 50 percent. With laser scanning, the RMSE for mean height of trees has been reported to range from 3 to 15 percent and the RMSE for number of trees per hectare has been between 13 and 74 percent according to different studies [Naesset 2002, Maltamo et al. 2004]. The different remote sensing techniques have shown very promising results in forest inventory accuracy, but still at the moment the prevailing method for forest planning data collection in Finland is the stand-wise subjective field sampling.

Coarse errors in forest planning data might be detected by searching for outliers as the individual observations with large estimation errors might stand out from the rest of the data. Some of the problems with traditional statistical or distribution-based outlier detection methods are that they require a lot of manual work, demand prior information about data distributions and cannot handle multi-dimensional data [Barnett & Lewis 1994]. Detection of outliers has been studied also in the data mining research field and is considered to be an important data mining task [Knorr 2000]. Outlier detection is one of the many applications of data mining and it has been successfully utilized in many applications, such as fraud detection and computer network intruder detection [Hodge & Austin 2004, Han & Kamber 2006]. As the data sets in forest planning can be large and usually multi-dimensional, the data mining –based outlier detection methods could possibly provide good tools for finding errors in forest planning data.

This paper describes how different data mining –based outlier detection algorithms could be used to detect coarse errors from a typical forest planning data. A method that could automatically identify a significant proportion of the coarse errors in the input data would greatly improve the quality and reliability of forest planning.

2. OUTLIER DETECTION WITH DATA MINING

Data Mining (DM) is an umbrella term which covers a number of different methods for finding frequent patterns end previously unknown information from large amounts of data automatically. Data mining is considered to be a part of Knowledge-Discovery in Databases (KDD) process. Data mining methodologies are usually related to different scientific core fields such as computer science, statistics and machine learning. Common DM tasks include classifying, clustering, association rule mining and outlier detection [Piatetsky-Shapiro 1991, Han & Kamber 2006]. Different types of DM tasks can be divided into four main categories according to Knorr & Ng [1998]: 1. dependency detection, 2. class identification, 3. class description and 4. exception/outlier detection. In the first three categories the focus is on the majority of the objects in the data, where as the fourth category concentrates on a small portion of the whole data.

One definition for outliers is that an outlier is *an observation which appears to be inconsistent with the remainder of that set of data* [Barnett & Lewis 1994]. According to Papadimitriou [2002] different approaches for identifying outliers in a data set can be classified into distribution based, depth-based, clustering-based, distance-based and density-based techniques. The distribution-based techniques have their background in statistics and they utilize a distribution model fitted to the data and single out objects deviating from the distribution as outliers [Hawkins 1980, Barnett & Lewis 1994]. The distribution-based methods are usually univariate and need a priori assumptions about the data distributions. Depth-based techniques utilize computational geometry to compute layers of *k-dimensional* convex hulls and use the layers to divide the objects into outliers and normal objects [Johnson et al. 1998]. The depth-based methods suffer from the curse of dimensionality and cannot be used with high-dimensional data.

Usually the exceptions or outliers are considered to be mere annoyances, which the different data mining algorithms try to ignore. However, in many applications the outliers are actually the most interesting objects as they can lead to the discovery of new and meaningful knowledge [Knorr 2000]. Number of different approaches for identifying outliers in large data sets have been proposed in recent years [Hodge & Austin 2004].

Distance-based outlier detection introduced by Knorr & Ng [1998] utilizes a distance measure between objects to identify the outliers. The distance between two objects is usually computed as euclidean or mahalanobis distance in a k-dimensional feature space. The outlyingness is considered to be a binary property so that an object either is or is not an outlier. Distance-based outlier detection algorithms normally view the data objects at global scale and compare an object to all other objects in the data, which makes them computationally heavy for large data sets. A lot of research has been done in order to improve the efficiency of these algorithms [Ramaswamy 2000, Bay & Schwabacher 2003].

Density-based outlier definition offers a bit different approach for identifying the outliers. Algorithms for density-based outlier detection compute a continuous Outlier Factor (OF) locally for each object which can be used to divide objects into outliers and normal objects by setting a threshold value. The outlier factors are computed locally for example as a mean or maximum distance from object O to the object's k nearest neighbours. In the proposed density-based outlier detection algorithms, the outlier factor has been defined as Local Outlier Factor (LOF) [Breunig et al. 2000] or Connectivity Outlier Factor (COF) [Tang et al. 2002]. Many of the proposed distance- and density-based outlier detection algorithms handle well both *large* and *high-dimensional* ($n > 5$) datasets [Hodge & Austin 2004].

Clustering is a group of unsupervised classification methods used extensively for exploratory data analysis by many disciplines. Some of the most common clustering methods include hierarchical single- and complete-link clustering and partitional k-means and k-medoids clustering [Jain et al. 1999]. Many of the clustering algorithms, such as DBSCAN and it's many variations use a density-based notion of clusters to classify the objects in the data set [Ester et al. 1996, Brecheisen et al. 2006]. Most clustering algorithms consider outliers as noise that should be dealt with by identifying and removing the outlying objects. Because of this, some clustering algorithms provide outlier detection as a side product of clustering or are designed entirely for the purpose of identifying the outliers [Tang & Khosgoftaar 2004, Hautamäki et al. 2005].

3. MATERIAL AND METHODS

3.1. Stand-wise estimation data and reference measurements

The data in the study consisted of stand-wise visual estimates conducted from a helicopter by an experienced forest planning professional, and reference sample plot measurements conducted in the field. The total number of forest stands in the data was 578 and the number of reference sample plots was 5090, an average of 8.8 sample plots per stand. For each stand, a number of attributes describing the average properties of the growing stock, site quality and surrounding environment, were estimated. The sample plots were located inside the stands in a grid formation and the same set of attributes as for the stands were measured by field groups. In this setting the stand-level estimates for the different attributes represents the normal accuracy of stand-wise subjective sampling. The reference measurements in the sample plots can be considered very precise and they represent the true values for the different attributes.

Altogether six attributes, four of them numerical and two categorical, were selected for this study. These particular attributes were chosen because their values affect strongly the behaviour of the growth models in a forest planning system. Coarse errors in the values of these attributes will lead to notable errors in the growth and yield predictions. The selected attributes and their measurement units, abbreviations and some mean statistics are listed in Table 1.

Table 1. Stand-level mean attributes and abbreviations used in the analysis

attribute	abbreviation	mean	median	stdev
mean diameter [cm]	D_{MEAN}	16.1	-	8.0
mean height [m]	H_{MEAN}	13.6	-	6.4
basal area per hectare [m^2/ha]	BA	16.7	-	7.6
mean age [years]	A_{MEAN}	61.1	-	27.6
development class	DC	-	3	-
forest site class	SC	-	4	-

As the sample plot measurements were accurate, we could compute the true values for the stand-level attributes for each stand. This was done by calculating means for the attributes from the sample plots that were located inside a given stand. After we had the stand-wise, possibly erroneous, estimate x_{ij} and the true value y_{ij} for each attribute, the relative to true error δ_{ij} of the stand-wise estimation was calculated for each stand i and each attribute j. The error computed for the categorical attributes was a binary value; *TRUE* if the stand-wise estimate was correct and *FALSE* if not.

After calculating the relative estimation errors for each stand and numerical attribute, the errors were classified into four classes by the absolute magnitude of the error $|\delta_{ij}|$, so that in class I $|\delta_{ij}| < 0.25$, class II $0.25 < |\delta_{ij}| < 0.50$, class III $0.50 < |\delta_{ij}| < 0.75$ and class IV $|\delta_{ij}| > 0.75$. The percentage of observations in each of the four error classes for each numerical attribute is shown in Figure 1. Only a small proportion, under five percent of the stands had relative estimation errors in error class IV. The proportion of observations relative error greater than 0.25, classes II to IV, ranged from 27.6 percent to 45.4 percent depending on the attribute in question.

In addition to the four numerical attributes, D_{MEAN}, H_{MEAN}, BA and A_{MEAN} Figure 1 has also column *MAX,* where each stand was classified into one of the four error classes by the maximum error in any of their four numerical attributes. Most of the stands, around 60 percent, had at least one attribute in error class II and only in 13 percent of the stands the maximum error class was I.

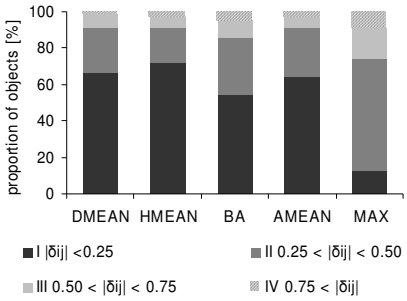

Figure 1. Proportion of objects in the four error classes for each of the four numerical attributes.

3.2. Outlier detection schemes

In this study, three different outlier detection algorithms were tested for identifying the coarse estimation errors in the visually estimated stand-wise data set. The applied algorithms were Nested-Loop distance-based outlier detection proposed by Knorr & Ng [1998], Simple Pruning distance-based outlier detection proposed by Bay & Schwabacher [2003] and Outlier Removal Clustering by Hautamäki et al. [2005]. These algorithms were selected because they are not univariate and because they do not require a priori information about data distributions. Forest stand data is multivariate and we do not necessarily have good a priori information about the data distributions.

The Nested-Loop distance-based outlier detection algorithm, *DB(p, D)* had parameters *p* and *D*, where an object *O* is considered to be an outlier if a proportion *p* of all objects are further than distance *D* away from *O*. The complexity of the Nested-Loop algorithm was $O(kN^2)$, where *k* is the number of attributes and *N* is the number of objects.

The Simple-Pruning distance-based algorithm was based on the Nested-Loop design, but randomization and a simple pruning rule were added for increased performance. The Simple-Pruning algorithm has showed nearly linear behaviour with large and multi-dimensional data sets. The parameters of the Simple-Pruning algorithm were *n* and *k*. The outliers were defined as the top *n* objects whose average distance to *k* nearest neighbours was greatest.

The Outlier Removal Clustering (ORC) consisted of two stages with were repeated multiple times. In the first stage, a k-means clustering was performed until convergence and in the second stage each object was assigned an outlyingness factor, which depended on the object's distance from the cluster centroid. The algorithm searched for the object with the maximum distance to the cluster centroid d_{MAX} and after that assigned the outlyingness factor o_i for each object x_i (1). Outlier Removal Clustering algorithm had parameters *k* which was number of cluster centroids and *T* which was the outlier threshold for o_i between 0..1.

$$o_i = \left\| x_i - c_{p_i} \right\| / d_{MAX} \tag{1}$$

The distance d_{xy} between two objects *x* and *y* were calculated as normalized euclidean distances (2) for the continuous attributes and as Hamming's distance between categorical attributes.

$$d_{xy} = \sqrt{\sum_{i=1}^{P} (x_i - y_i)^2 / \sigma^2} \tag{2}$$

The outlier detection algorithms were run multiple times with changing parameter values for the visually estimated stand-wise data. During each run, a number of stands were classified as outliers and the list of outliers was stored for later analysis. As the aim of this study was to test if the outlier detection algorithms could identify the stands with coarse estimation errors, the outlier lists from the different runs were examined against the classified errors. With this kind of cross-examination we could determine what proportion of the stands with coarse estimation errors, ie. errors in classes II-IV, had been defined as outliers and what proportion of the outliers did not have any coarse errors.

4. RESULTS

We executed several runs with different parameter values for all three outlier detection methods in order to see how the parameters affected which stands were classified as outliers. With Nested-Loop algorithm, increasing the value of parameter *p* from 0.9 to 0.995 decreased the proportion of outliers in all objects from 96.2 to 4.8 percent. Also the value of the parameter *D* had a strong effect on the proportion of outliers so that when the value of *p* was fixed to 0.95, increasing the value of *D* from 0.5 to 1.5 decreased the proportion of outliers from 74.7 to 6.2 percent. All three algorithms were affected strongly by the chosen parameter values.

The proportion of outliers in different error classes depended on the different parameter values in all algorithms. This was the most interesting part as the main idea of the tests was to find out if the stands with coarse errors have a higher probability to be classified as an outlier. A notable trend in the results was that the percentage of outliers was clearly higher in the stands where at least one variable had an estimation error in class III or IV. This behaviour can be seen in Figure 2, where the proportion of outliers in error class IV increases sharply as the proportion of outliers in all stands increases. Very similar behaviour was found in the performance of all three algorithms. Individual attributes seemed to affect the results to some extent. In all of the tests, the coarse, class III and IV, errors in the variables D_{MEAN} and *BA* increased the stand's probability to be classified as an outlier.

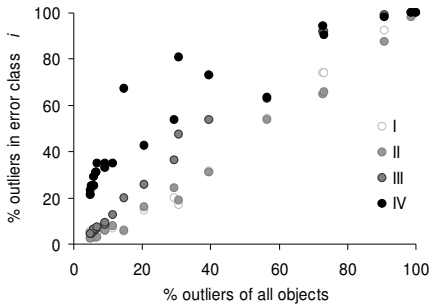

Figure 2. Relation between the percentage of outliers in all objects and percentage of outliers in each error class. Nested-Loop algorithm, 6 attributes.

We ran the outlier detection algorithms with two attribute combinations, where in the other combination the four continuous attributes were included in the outlier detection and in the other all six attributes were included. The difference in the results between the two attribute combinations was minimal as the proportion of outliers was almost exactly the same with both attribute combinations.

5. CONCLUSIONS

The objective of this paper was to test if data mining –based outlier detection algorithms could be used to identify coarse assessment errors from a typical forest planning data set. The errors were classified into magnitude classes in order to examine how the magnitude of the error affected the probability of a stand to be classified as an outlier. The probability for a stand to be an outlier was notably higher when at least one of the attributes had a coarse estimation error in class IV. This suggests that coarse individual estimation errors will be likely to stand out from the rest of the data as outliers.

All of the three algorithms had different parameters that had a strong effect on what proportion of the whole data was classified as an outlier. This is quite obvious as the parameters were simply defining what an outlier is. The higher the proportion of the outliers was, the more stands with no coarse estimation errors were classified as outliers. Also, when only a small proportion of all stands were classified as outliers, then the number of outliers in the stands with coarse estimation errors was also smaller. In this kind of a trade-off situation, it was difficult to find a balance so that the proportion of outliers in the stands with no coarse errors would be minimized and the number of outliers in the stands with coarse errors maximized. Also, ranking the different algorithms in some supremacy order was difficult as the results were very similar. We used only a single data set in this study and it would be valuable to test these methods with different forest stand data sets to find out if the optimal parameter values depend on the data set as suggested by Hautamäki et al. [2005].

Normally, different forest attributes correlate with each other to a certain degree. Tall trees tend to have larger diameter than short trees and in a forest with a lot of tall and broad trees, the total volume of timber is normally high due to so called allometric correlations. Because of these correlations, observations in a forest stand data set tend to bunch together into groups or clusters. When we are measuring forest attributes and we do a severe error in the measurement of one of the attributes, this observation might stand out from the rest of the data if it does not fit these correlations. This would be the case for example when mean diameter and height would be measured correctly and total volume would be coarsely underestimated. These kind of outlying observations can possibly be detected with outlier

detection methods even though the methods are not based on analysing the correlations. However, if we systematically overestimate the values of most of the attributes, the correlation between the different attributes might fit the normal correlation and that observation would not necessarily stand out as an outlier. Because of this, all estimation errors can not be considered to be outliers or vice versa. This has a lot to do with the actual data collection method. When forest planners are estimating the forest properties with a highly subjective method, human errors are possible. Classification errors or data input errors could produce individual coarse errors which could be fairly easy to identify as outliers. As in future the forest data may originate from very many sources, from old data to new measurements, the possibility of independency between errors increases and the usefulness of datamining probably increases.

The size of the data was fairly small compared to many data mining –related studies. Yet, the dimensionality of the data is quite high. The problem was that as precise reference measurements require a lot of manual work and expertise, the costs for acquiring this kind of data are high. However, as this study was the first step in applying data mining methodology in a forestry setting, we decided that this kind of fairly small data set should be sufficient.

In spite of the problems with the tested methods, the outlook for applying data mining –based methods for identifying coarse estimation errors in forest planning data is promising. If the forest planning system could identify the outliers, which might be due to coarse estimation errors, this information could be used to provide some estimate about the likely quality of the results. It is possible that the detected possible errors could be used for instance as a basis for planning field checking [Store 2007], for detecting illogical treatment recommendations made in field or even for imputing more logical data to replace the most susceptible observations. However, this is all left to future work.

The next step in studying data mining –based outlier detection methods for identifying coarse estimation errors in forest planning data should be testing of different types of approaches, such as density –based outlier detection. Also, the algorithms should be tested with larger data sets. Another interesting research question would be to study what are the optimal parameter values for the algorithms and how they depend on the particular data set.

ACKNOWLEDGEMENTS

The authors would like to thank the Foundation for Research of Natural Resources in Finland for funding this research project. Initial steps in this research topic were taken as a part of SIMO (SIMulation and Optimization for next generation forest planning) project (http://www.simo-project.org)

REFERENCES

Barnett, V. & Lewis, T. Outliers in statistical data. John Wiley & Sons, Chichester, 3rd edition, 584 pp. 1994.

Bay, S.D. & Schwabacher, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. SIGKDD ’03, August 2427, Washington, DC, USA. 2003.

Brecheisen S., Kriegel H.-P. & Pfeifle M. Parallel density-based clustering of complex objects, Proceedings of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'06), Singapore, 2006, in: Lecture Notes in Artificial Intelligence (LNAI), Springer, Vol. 3918, 179-188. 2006.

Breunig, M.M., Kriegel, H-P., Ng, R.T. & Sander, J. LOF: Identifying density-based local outliers. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.

Buongiorno, J. and Gilles, J.K. Decision methods for forest resource management. Academic Press. 439 p, 2003.

Burkhart, H. Suggestions for choosing an appropriate level for modeling forest stands. In: Amaro, A., Reed, D. & Soares, P. (eds.) Modelling forest systems. CABI Publishing, 398 p. 2003.

Ester, M., Kriegel, H-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, 1996, pp. 226-231. 1996.

Haara, A. Comparing simulation methods for modelling the errors of stand inventory data. *Silva Fennica 37*(4): 477–491. 2003.

Han, J. & Kamber, M. Data mining: concepts and techniques, 2nd ed. Morgan Kaufmann Publishers, 2006.

Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T. & Fränti, P. Improving k-means by outlier removal. In H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 978–987, 2005.

Hawkins, D. Identification of outliers. Chapman and Hall, London. 1980.

Hodge, V. J. & Austin, J. A Survey of outlier detection methodologies. *Artificial Intelligence Review 22*: 85–126, 2004

Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. *ACM Computing Surveys 31*(3) (1999) 265–323. 1999.

Johnson, T., Kwok, I. & Ng., R. Fast computation of 2-dimensional depth contours. Proc. KDD, pp. 224-228. 1998.

Kangas, A., Heikkilä, E. & Maltamo, M. Accuracy of partially visually assessed stand characteristics: a case study of Finnish forest inventory by compartments. *Canadian Journal of Forest Research 34*, 916–930, 2004.

Knorr, E.M. & Ng, R.T. Algorithm for mining distance-based outliers in large datasets. VLDB Conference Proceedings, September 1998

Knorr, E.M., Ng, R.T. & Tucakov, V. Distance-based outliers: algorithms and applications. *The VLDB Journal 8,* 237–253, 2000.

Laasasenaho, J., and Päivinen, R. On the checking of inventory by compartments. *Folia Forestalia 664*, 1986. (In Finnish)

Papadimitriou, S., Kitagawa, H., Gibbons, P.B. & Faloutsos. C. LOCI: Fast outlier detection using the local correlation integral. Proceedings of the 19th International Conference on Data Engineering (ICDE), p. 315-326, 2003.

Maltamo, M., Eerikäinen, K., Pitkänen, J., Hyyppä, J. & Vehmas, M. Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sensing of Environment 90*, 319–330, 2004.

Næsset, E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment 80*, 88– 99, 2002.

Piatetsky-Shapiro, G. & Frawley, W. J. Knowledge discovery in databases. AAAI/MIT Press. 1991.

Poso, S. Basic features of inventory by compartments. *Silva Fennica 17*, 313–349, 1983. (In Finnish)

Ramaswamy, S., Rastogi, S, R. & Shim, K. Efficient Algorithms for mining outliers from large data sets. Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, 427-438, 2000.

Store, R. Erilaisten tavoitteiden ja maastotyötarpeiden huomioon ottaminen metsäsuunnittelun maastoinventoinnin suunnittelussa. In: Tikkanen, J. & Lappalainen, S. (toim.). Asiakaslähtöisyys metsäsuunnittelun kehittämishaasteena. Metlan työraportteja 65. (http://www.metla.fi/julkaisut/workingpapers/2007/mpw065.htm) 125 s, 2007.

Tang, J., Chen, Z., Fu, A.W. & Cheung, D. A robust outlier detection scheme for large data sets. PAKDD. 2002.

Tang, W. & Khosgotaar, T.M. Noise Identification with the k-means Algorithm. Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04) - Volume 00, 2004.