ISSN 2635-3296

Journal of Forest Science Environment.
Volume 8 (2023): 40 – 46

Application of Clustering and Principal Component Analysis for Modeling Tree Volume

Clement S A • Kuje E D • Ndagi H I

Website: www.jfseunimaid.com.ng

Journal of Forest Science & Environment

J. For. Sci. Env. Vol. 8 (2023): 40 – 46
ISSN 2635-3296
Available at www.jfseunimaid.com.ng

APPLICATION OF CLUSTERING AND PRINCIPAL COMPONENT ANALYSIS FOR MODELING TREE VOLUME

CLEMENT S A * ● KUJE E D ● NDAGI H I

Faculty of Agriculture, Nasarawa State University, Keffi

*Corresponding authors' email: clementsegun2016@gmail.com Phone: +2348104713317

ORCID iD: https://orcid.org/0000-0002-6732-3221

ABSTRACT: The study evaluated the application of clustering and principal component analysis for modeling tree volume in Shasha forest reserve. Information on the present status of the resource is required for sustainable forest resource management. The data set used in this study was collected from 15 temporary sample plots of 30 m x 30 m dimensional square plots were established. Clustering Analysis was applied to distribute the observations into four different groups using k-means method. Four distance measures and methods were used to group the data from Shasha rain forest. The results of model selection indices on each cluster revealed that in cluster one, model three had the lowest model selection indices of RMSE (0.1404), MAE (0.0952), BIC (-10.2377) and AIC (-16.1279) when compared to models one, two and four (4). In cluster two, model three with RMSE (0.1932), MAE (0.1604), BIC (2.3226), AIC (-5.3092) was selected over models one, two and four as it had the smallest residuals of RMSE (0.1404) and MAE (0.0952). In cluster three, model four had the lowest model selection indices with RMSE (0.2642), MAE (0.203), BIC (14.0543) and AIC (10.3977) when compared to models one, two and three. In cluster four, model 3 had the lowest model selection indices with RMSE (0.2014), MAE (0.1440), BIC (6.2121) and AIC (8.1649) when compared to models one, two and four. The results of the model selection indices on the fifteen (15) sample plots showed that model three had the lowest model selection indices with RMSE (0.1992), MAE (0.1602), AIC (-23.8572) and BIC (-11.5277) when compared to models one, two and four. Hence, model three is most suitable for tree volume predictions in the study area. The application of this management tools in forest modeling aids in prescribing sustainable management strategies for shasha forest reserve.

Keywords: Cluster, Principal Component, Modeling, Volume, Forest Reserve

1. INTRODUCTION

A forest is a complex ecosystem which is predominantly composed of trees, shrubs and is usually with a closed canopy. Forests are storehouses of a large variety of life forms such as plants, mammals, birds, insects and reptiles. Depending on the physical, geographical, climatic and ecological factors, there are different types of forest like evergreen and deciduous. The rainforests of West Africa have been earmarked as one of the world's hotspots of biodiversity (Myers *et al.*, 2000). One of the typical features of tropical rain forests is high species diversity with many of them represented by only a few trees. These biodiversity hotspots have become safe

harbor for slow-growing indigenous species considering the high demands for timber and pressure in favor of establishment of plantations with fast-growing exotic species. Unfortunately, unlike natural forests, plantations are often of same or few species and do not support a variety of natural biodiversity.

Forest management includes the management of land, its total vegetation cover and the surrounding environment and microclimate as well as the people living in such environment. The key to efficient and effective management of the forest is to understand how the forest ecosystems operate and how this knowledge can be used to address the needs of the society, their expectations and values. Forest modelling accurately describes the dynamic nature of the forests. Successful forest modelling usually has a way of improving its management through precise representation of all sections and sectors. Sustainable forest resources management needs a huge amount of supporting information on the present status of resources, estimates the maximum sustainable harvest, and forecast the nature and timing of future harvests.

Cluster analysis is often used for classification and discrimination of dendrometric data in native forests. These techniques generate similarity classifications that exclude the subjective aspects existing in other sorting methods. In forestry, grouping application for data analysis are found in the fitting of stem volume models (Akindele and LeMay, 2006), growth and yield studies (Phillips *et al*. 2002), production stratification (Souza and Souza, 2006), phytogeographic studies (Oliveira-Filho and Fontes, 2000) and tropical species on distribution patterns (Plotkin *et al*., 2002). The general goal of clustering is to partition a set of data such that data points within the same cluster are "similar" while those from different clusters are "dissimilar". At present, there are many clustering algorithms which are categorized based on their cluster model.

Volume models are mathematical expressions which relate tree volume to tree's measurable attributes such as diameter at breast height and/or height. They are used to estimate the average content for standing trees of various sizes and species (Avery and Burkhart, 2002). The volume of the stem of a tree is considered a function of the independent variables, diameter, height, and form (Clutter *et al.*, 1983; Husch, *et al.*, 2003). Foresters need to know every detail about the forest they are managing in terms of location, size, quantity and quality of forest resources available and how these resources change over time.

Therefore, the objective of this study was to apply clustering and principal component analysis for volume modeling of shasha natural forest reserve with a view to improving the management of the forest and evaluate clustering tendency, assign individual observations to clusters using K-means clustering method, reduce the number of variables using principal component analysis method and model volume of trees in Shasha forest reserve.

2. METHODOLOGY

2.1. Study Area

This study was carried out in Shasha Forest Reserve established and managed by Forestry Research Institute of Nigeria (FRIN). The 100 ha Shasha Forest Reserve is located in Ife South Local Government Area in Osun State of Nigeria. The reserve lies between latitudes 4° 18' E and 4° 30' E and longitudes 7° 4' N and 7° 20' N, at an altitude of 122 m above sea level with a mean annual rainfall of 142 mm. Soil type is ferruginous tropical soils on crystalline acid rock. The topography is gently undulating to undulating plain. The vegetation is mainly of the high forest type (Chenge and Osho, 2016).

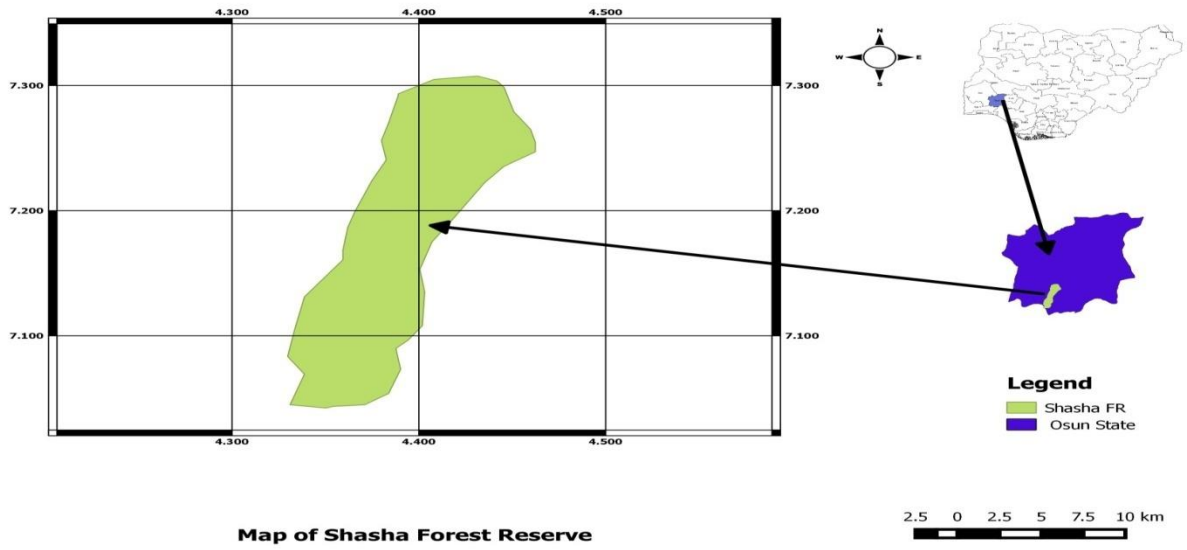

Fig. 1: Map of the Study Area in Shasha Forest Reserve

2.2. Data Collection

The data used in the study was collected from 15 temporary sample plots established in the natural stand at Shasha forest reserve, Nigeria. 30 x 30 m dimension square plots were established (900 m^2 plot size). In each sample plot, all living trees with DBH ≥ 10 cm were enumerated and assessed for diameter at breast height (1.3 meters from the ground) and their corresponding height using a diameter tape and a Spiegel Relakcope while other 48 variables were derived, respectively. A total of 320 observations were available for the analysis.

2.3. Data Analysis

Statistical analysis of the data involved cluster analysis, correlation analysis and regression analysis.

2.4. Cluster Analysis

Clustering was used in this research to group the 50 variables into classes. Although, this analytical tool does not reduce the number of variables measured it gives clear variability that exist within the forest. Correlation analyses were further used to reduce the number of variables for easy handling of regression equations.

2.5. Development of Volume equations

Due to few numbers of trees representing individual species as characteristic of a natural forest, it is impossible to develop species-specific volume equations. Therefore, the volume model was applied to each cluster.

Four volume equations are considered in this study and the equation with the most satisfactory mathematical properties for each cluster was chosen.

The first equation followed the Schumacher-Hall volume model (Schumacher and Hall, 1933) as used by Aigbe *et al*. (2013). The model is expressed as:

$$V_i = b_0 + b_1 D_i^{b_2} H_i^{b_3} + \varepsilon_i \quad \text{.....(1)}$$

Where:
V = tree volume (m^3);
D = diameter at breast height/ stump diameter (cm);
H = merchantable height (m);

b_0, b_1, b_2 and b_3 are the regression parameters, while ε_i is the error term.

This equation is referred to as the combined variable volume function. The model indicates that tree volume increases by certain powers of D and H. Even though Edminster *et al*. (1980), Abayomi (1983) and Akindele (2005) all agreed that H was fixed to the power of 1, the original equation we decided to retain the original equation and changed height to merchantable height.

Schumacher-Hall equation (Malata *et al*., 2017; Aigbe *et al.*, 2013)

$$V = b_0 + Dbh^{b_1} MHt^2 \text{....} (2)$$

Schumacher I equation (Mason, 2011)

$V = exp^{(a-b/d)}$ …… (3)

Weighted quadratic models adopted from Aigbe *et al*., (2013)

$V = b_0 + b_1 Dbh^2 + MHt + Dbh^2 MHt$ …… (4)

$V = b_0 + b_1 Dbh^2 MHt$ ….. (5)

2.6. Goodness of Fit Indices

The volume models were evaluated with the view of recommending those with good fit for further uses. The following statistical criteria were used

I. Akaike Information Criterion (AIC)

This is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model relative to other models. Therefore, AIC provides a means for model selection. The formula is given as;

AIC = n x Ln (*L*) + 2k …… (6)

Where *L* = likelihood function (*L* = RSS/n (RSS = residual or error sums of squares), n = sample size), k = number of regression parameters and Ln = natural logarithm.

II. Regression Mean Square Error (RMSE)

This is also referred to as the standard deviation or residual of the error variance of the estimate. It measures the spread of data and is a good indicator of precision. The value must be small.

RMSE = $\sqrt{MSE}$ …… (7)

Where: RMSE = Regression Mean Square Error and MS $_{Total}$ = Mean square total.

3. RESULTS

Clustering Analysis was used to classify the observations into four different groups using k-means method. Each of the group represents a class with similar characteristics but not the same species. The clusters obtained from the analysis were further subjected to principal component correlation analysis which was used to classify the 50 measured and derived variables from shasha forest reserve into 7 different groups. This was successfully carried out in order to have fewer variables to work with. The principal component analysis of the variables enable smooth application of regression analysis on the four different volume functions.

Table 1: The Principal Component Analysis of the Variables

Variables	Dm	Db	Dbh	Dt	BQ	THt	MHt
Dm	1.00	0.41	0.52	0.75	0.07	-0.05	-0.05
Db	0.41	1.00	0.84	0.31	-0.09	-0.29	-0.29
Dbh	0.52	0.84	1.00	0.42	-0.10	-0.37	-0.37
Dst	0.41	1.00	0.84	0.31	-0.09	-0.29	-0.29
Dt	0.75	0.31	0.42	1.00	0.05	-0.07	-0.07
BQ	0.07	-0.09	-0.10	0.05	1.00	0.15	0.15
THt	-0.05	-0.29	-0.37	-0.07	0.15	1.00	1.00
MHt	-0.05	-0.29	-0.37	-0.07	0.15	1.00	1.00

Dm;diameter at the middle, Db; diameter at the base, Dbh; diameter at breast height, Dt; diameter at the top, BQ; quarter bole height, THt; total height, MHt; merchatable height

The result of evaluation of the model selection indices in cluster one revealed that model three had the lowest model selection indices of RMSE (0.1404), MAE (0.0952), BIC (-10.2377), AIC (-16.1279) with better model parameters of a (-0.9079), b(0.0193), c (0.0088), d (0.0005) when compared to model one, two, and four.

Table 2: Evaluation of Models Selection Indices in Cluster One

Equation	A	B	C	d	AIC	BIC	MAE	RMSE
1	-2.3289	0.24728	0.45796	----	-14.397	-9.6852	0.1087	0.1517
2	0.4376	35.3892	----	----	7.3157	10.8499	0.1904	0.2487
3	-0.9079	0.0193	0.0088	0.0005	-16.1279	-10.2377	0.0952	0.1404
4	-0.0369	0.0007	----	----	-13.2069	-9.6727	0.1141	0.1622

Where: AIC=Akaike information criterion; BIC=Bayesian information criterion; RMSE=Root Mean Square Error and a= intercept while b, c and d are slopes of individual variable (model parameters).

The result of evaluation of the model selection indices in cluster two showed that model three also had the lowest model selection indices of RMSE (0.1932), MAE (0.1604), BIC (2.3226), AIC (-5.3092) with better model parameters of a(0.6024), b(-0.0143), c(-0.0592), d(0.0021) when compared to model one, two, and four.

Table 3: Evaluation of Model Selection Indices in Cluster Two

Equation	A	B	C	d	AIC	BIC	MAE	RMSE
1	-2.74167	0.29494	0.41069		-6.1953	-0.0899	0.1621	0.1964
2	0.9361	56.6418			27.2829	31.8620	0.2473	0.3309
3	0.6024	-0.0143	-0.0592	0.0021	-5.3092	2.3226	0.1604	0.1932
4	-0.1983	0.00102			-7.2474	-2.6683	0.1559	0.1991

Where: AIC=Akaike information criterion; BIC=Bayesian information criterion; RMSE=Root Mean Square Error and a= intercept while b, c and d are slopes of individual variable (model parameters).

The result of evaluation of the model selection indices in cluster three showed that model four had the lowest model selection indices of RMSE (0.2642), MAE (0.203), BIC(14.0543), AIC (10.3977) with better model parameters of a (0.1706), b (0.0011) when compared to model one, two, and three.

Table 4: Evaluation of Models Selection Indices in Cluster Three

Equation	A	B	C	D	AIC	BIC	MAE	RMSE
1	-2.9454	0.3099	0.4054		11.7518	16.6273	0.2076	0.2608
2	0.0896	0.3226			34.6088	38.2655	0.3197	0.4288
3	-2.6830	0.0395	0.1043	-0.00056	13.703	19.79738	0.2060178	0.2605714
4	0.1706	0.0011			10.3977	14.0543	0.203	0.2642

Where: AIC=Akaike information criterion; BIC=Bayesian information criterion; RMSE=Root Mean Square Error and a= intercept while b, c and d are slopes of individual variable (model parameters).

The result of evaluation of the model selection indices in cluster four revealed that model three had the lowest model selection indices of RMSE (0.2014), MAE (0.1440), BIC (6.2121) and AIC (8.1649) with better model parameters of a (-3.459), b (0.0494), c (0.1804), d (-0.0015) when compared to model one, two, and four.

Table 5: Evaluation of Models Selection Indices in Cluster Four

Equation	A	B	C	D	AIC	BIC	MAE	RMSE
1	-2.8551	0.3188	0.3928		9.4609	7.8986	0.2646	0.28
2	3.403	220.359			16.5186	15.3469	0.5845	0.6928
3	-3.459	0.0494	0.1804	-0.0015	8.1649	6.2121	0.1440	0.2014
4	-0.1308	0.0012			12.9555	11.7838	0.4065	0.4851

Where: AIC=Akaike information criterion; BIC=Bayesian information criterion; RMSE=Root Mean Square Error and a= intercept while b, c and d are slopes of individual variable (model parameters).

The result of evaluation of the model selection indices in 15 sample plots showed that model three had the lowest model selection indices of RMSE (0.1992), MAE (0.1602), AIC (-23.8572) and BIC (-11.5277) with better model parameters of a (-0.8393), b (0.0122), c (0.0011), d (0.00099) when compared to model one, two, and four.

Table 6: Evaluation of Models Selection Indices in 15 Sample Plots

Equation	A	B	c	D	AIC	BIC	MAE	RMSE
1	-3.0837	0.2829	0.4537		-9.4319	0.4318	0.1697	0.2189
2	1.310	75.896			86.7967	94.1944	0.2585	0.385
3	-0.8393	0.0122	0.0011	0.00099	-23.8572	-11.5277	0.1602	0.1992
4	-0.2964	0.0011			-4.1858	3.2119	0.186	0.2282

Where: AIC=Akaike information criterion; BIC=Bayesian information criterion; RMSE=Root Mean Square Error and a= intercept while b, c and d are slopes of individual variable (model parameters).

4. DISCUSSION

Clustering analysis was applied to classify the observations of mixed tree species in the study area into four different groups using k-means method. This is in agreement with Jain *et al.* (2000) who stated that data clustering is the process of identifying natural groupings or clusters within multidimensional data based on some similarity measure. Bischof *et al.* (1999) proposed an algorithm based on K-means which uses MDL (conceptually similar to MML). The algorithm starts with a large value for K and proceeds to remove centroids when this removal results in a reduction of the description length. K-means is used between the steps that reduce K. This allowed the identification of different classes of trees with similar characteristics. K-means is generally used as a quick technique for estimation of a clustering, or as an initial clustering for other more expensive methods (Zien and Candela, 2005). The principal component analysis was further applied as variable reduction tools for the application of volume models. It also, enabled the application of volume models on each class in order to estimate the current, future worth and sustainably manage of the forest reserve for wood production, carbon stock, biodiversity among other benefits.

Model selection indices indicate the predictive utility of models to estimates. The lower the model indices, the better the model performance. The result of evaluation of model selection indices in cluster one revealed that model three had the lowest model selection indices of RMSE (0.1404), MAE (0.0952), BIC (-10.2377) and AIC (-16.1279) when compared to model one, two, and four. It implied that model three is the most suitable to estimate tree volume in this class. A forest is generally made up of trees of varying sizes, so a size-class model divides each stand into multiple groups of similar-sized individuals, which are projected through time. Some of the most common size-class models are stand-table projections (Trincado et al., 2003), matrix-based (Picard et al., 2002), diameter-distribution models (Qin et al., 2007) and clustering based modeling with diameter at breast height as an integral variable to predict tree heights. The application of the latter unveils growth characteristics such as tree heights, diameter at breast heights and tree volume in this class.

The result of evaluation of the model selection indices in cluster two showed that model three with RMSE (0.1932), MAE (0.1604), BIC (2.3226), AIC (-5.3092) was performed better than model one, two and four (4) because it has the lowest predicted residuals of RMSE (0.1404), MAE (0.0952). Despite the observed difference between cluster one and two, model three still thrived well in tree volume predictions. The performance of model three might be due to the structure of the model and similarity in the factors of tree growth in the study area. Hence, similarity measures are fundamental components in most clustering algorithms (Jain *et al.*, 1999).

The result of evaluation of the model selection indices in cluster three showed that model four has the lowest model selection indices with RMSE (0.2642), MAE (0.203), BIC (14.0543) and AIC (10.3977) when compared to model one, two, and three. Though, model four has two parameters yet thrived better than three parameter model in this cluster. Its performance could be attributed to the similarities in the factors of tree growth among other classes. Santiago *et al.* (2017) mentioned that the use of growth models is required to generate the most important information for management decision making in time and space.

The result of evaluation of the model selection indices in cluster four revealed that model three had the lowest model selection indices with indices RMSE (0.2014), MAE (0.1440), BIC (6.2121) and AIC (8.1649) when compared to model one, two, and four. It showed that model three is the most suitable to estimate tree volume in this class.

The result of evaluation of the model selection indices in 15 sample plots revealed that model three had the lowest model selection indices with RMSE (0.1992), MAE (0.1602), AIC (-23.8572) and BIC (-11.5277) when compared to model one, two and four. The results further confirmed the size classes' outcomes which had model three as the best performing model based on the model

selection indices. Therefore, model three is the most suitable model to predict tree volume in Shasha forest reserve.

5. CONCLUSION

1. The application of clustering and principal component analytical tools was apt in grouping the diverse tree species into tree classes which serve as an integral process of model applications.
2. The model selection indices showed that model three performed better than the other three models as shown in table six above. Hence, suitable for tree volume predictions in shasha forest reserve.
3. Application of this management tools in forest modeling aids in prescribing management strategies for shasha forest reserve.

REFERENCES

Abayomi JA (1983). Volume tables for Nauclea diderrichii in Omo forest reserve, Nigeria. Nigerian Journal of Forestry, 1983, 13 (1–2): 46–52.

Aigbe HI, Akindele SO, Onyekwelu JC (2013). Volume equation for Oban tropical forest reserve, Cross River, Nigeria. Nigerian Journal of Forestry, 2013, Vol 43 Nos 1&2

Akindele SO, LeMay VM (2006). Development of Tree Volume Equations for Common Timber Species in the Tropical Rain Forest Area of Nigeria. Forest. Ecology and Management, 2006, 226: 41-48.

Avery ET, Burkhart HE (2002). Forest measurements Mc Graw Hill New York U.S.A 5th Edition. 2002, 456 pp

Bischof AL and Selb A (1999). MDL Principle for Robust Vector Quantization. Pattern Analysis and Applications, vol. 2, 59-72

Clutter JL, Fortson JC, Piennnar LV, Brister, GH, Bailey RL (1983). Timber management; A Quantitative Approach. John Wiley and Sons, New York, USA, 1983, 333pp.

Edminster CB, Beeson RT, Metcalf GE (1980). Volume tables and point-sampling factors for ponderosa pine in the Front Range of Colorado. U. S. Forest Service, Rocky Mountains Forest and Range Experimental Station, Research Paper RM-218, 1980, 14Pp

Husch B, Beers TW, Kershaw JA Jr (2003). Forest Mensuration. 4th Edn., John Wiley and Sons Inc., New Jersey, USA., 2003, 949 Pp

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000). Biodiversity hotspots for conservation priorities. Nature, 2000, 403:853–858

Oliveira-Filho AT, Fontes MA (2000). Patterns of floristic differentiation among Atlantic forests in Southeastern Brazil and the influence of climate. Biotropica, 2000, 32: 793–810

Osho JSA (2016). Matrix model for tree population projection in a tropical rain forest of south-western Nigeria. Ecological Modelling, 2016, 59 (3/ 4), 247–255.

Phillips P, yasmani B, Van Gardingen Pr (2002). Grouping tree species for analysis of forest data in Kalimantan (Indonesia Borneo) Forest Ecology and Management, 2002, 157: 205 – 216Pp

Picard N, Bar-Hen A, Franc A (2002). Modelling forest dynamics with a combined maxtrix/individual-based model. Forest Science, 2002, 48: 643–52Pp

Plotkin Jb, Chave J, Ashton PS (2002). Cluster analysis of spatial patterns in Malaysian tree species. The American Naturalist, 2000, 160: 629–644Pp

Qin J, Cao QV, Blouin DC (2007). Projection of a diameter distribution through time. Canadian Journal of Forest Research, 2007, 37, 188–94.Pp

Jain MM and Flynn P (1999). Data Clustering: A Review. ACM Computing Surveys, vol. 31 (3): 264-323

Jain R and Duin JM (2000). Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 (1): 4-37

Schumacher FX, Hall FS (1933). Logarithmic expression of timber-tree volume. Journal of Agricultural Research, 1993, 47 (9): 719-734.Pp

Souza and Souza dr (2006). Análise multivariada para estratificação volumétrica de uma floresta ombrófila densa de terra firme, Amazônia Oriental. Revista Árvore, 2006, 30: 49–54Pp

Trincado GV, Quezada PR, von Gadow K (2003). A comparison of two stand table projection methods for young Eucalyptus nitens (Maiden) plantations in Chile. Forest Ecology and Management, 2003, 180: 443– 51Pp

Zien A, Qui˜nonero J (2005). Candela. Large margin non-linear embedding. In Luc De Raedt and Stefan Wrobel, editors, Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005, volume 119 of ACM International Conference Pro-ceeding Series, pages 1060–1067. ACM, 2005.