African Journal of Agricultural Research Vol. 5(12), pp. 1509-1514, 18 June, 2010
Available online at http://www.academicjournals.org/AJAR
DOI: 10.5897/AJPP10.053
ISSN 1991-637X ©2010 Academic Journals

Full Length Research Paper

Locational classification of walnut (*Juglans regia* L.) genotypes collected from Lake Van basin by using mixture modeling

Abdullah Yeşilova[1]*, Koray Özrenk[2], Bariş Kaki[1], M. Nuri Almali[3] and Fikri Balta[2]

[1]Department of Animal Science, Biometry and Genetic Unit, Faculty of Agricultural, Yuzuncu Yil University, Van, Turkey.
[2]Department of Horticulture, Faculty of Agricultural, Yuzuncu Yil University, Van, Turkey.
[3]Department of Electrics and Electronics Engineering Unit, Faculty of Engineering and Architecture, Yuzuncu Yil University, Van, Turkey.

Accepted 28 April, 2010

In mixture modeling, it is assumed that the data set shows a heterogeneous structure. This heterogeneity is defined as unobservable heterogeneity. The data set's heterogeneity produces serious deviations in the parameter estimates and the standard deviations. Heterogeneity is overcome when the data set divides itself into homogeneous sub-populations. Thus, while homogeneity is attained for sub-populations, the heterogeneity between the sub-populations is tried to be put forward. Akaike's information criteria (AIC), Bayesian information criteria (BIC), and Entropy classification criteria are used to determine the number of sub-populations. After the number of sub-populations is determined, the model determines the probability that each observation will fall within a particular sub-population. In this study, the classification of districts based on fruit traits is achieved by applying mixture modeling to walnut fruits collected from eight districts. According to the AIC, BIC, and entropy criteria, a model with five sub-populations was chosen where the data set is the most distributed. Therefore, it was determined that each district does not form a different population according to the studied walnut fruit traits, but are distributed into five sub-populations. The fourth sub-population had the most desirable traits for walnut improvement, and the highest proportion of these traits came from the naturally grown populations of Adilcevaz and Ahlat districts.

Key words: Mixture model, classification, EM algorithm, Latent class analysis, Walnut.

INTRODUCTION

Data set can be obtained from a population composed of different sub-populations. In other words, the data set can show a heterogeneous structure obtained not only from a single population but also from a multitude of populations. (Wang et al., 1996; Okut et al., 2002; Dalrymple et al., 2003; Yeşilova, 2003; Leisch, 2004). In mixture modeling, the objective is to determine the sub-populations by assuming that the observations in the data set may belong to unobserved sub-populations, to be subsequently determined for each observation with which probability is included under each sub-population (Wang et al., 1996; Arminger and Stein, 1997; Jedidi et al., 1997; Muthén and Muthén, 2002; Martinez et al., 2009). In mixture modeling, parameter estimations are obtained via the maximum likelihood (ML) method, using an EM algorithm (Dempster et al., 1977; Jansen, 1993; Wang et al., 1996; Wang and Putterman, 1998; Han, 2009). The EM approach is an algorithm composed of E and M steps. In the E step, the conditional expected values are used on the observed values and the missing observations are estimated. Here, the missing observations are thought to be latent classes (Roeder et al., 1999). In the M step, by maximizing the log-likelihood function, ML parameter estimations are obtained. Whatever the distribution of the data, classification can be made by using a multinomial logit model (Okut et al., 2002). In the mixture model, the multinomial logit model is used to determine which observation belongs to which sub-population. In other

*Corresponding author. E-mail: yesilova@yyu.edu.tr. Tel: +90 4322251795. Fax: +90 4322251104.

words, after determining the appropriate sub-population number in step E, step M estimates which observation will be included in which sub-population using the Multinomial Logit model. Akaike's information criteria (AIC), Bayesian information criteria (BIC), and entropy correction classification criteria are used in selecting the suitable model (Wang et al., 1998; Muthén and Muthén, 2002).

Juglans regia L. is an edible walnut species that occurs in Turkey and is ideally grown in Gevaş, Ahlat, and Adilceviz districts around the Van Lake Basin. The Lake Van Basin has large genetic variability of *Juglans regia.* The high yield, good nut and kernel characteristics, high lateral bud fruitfulness, late bud breaking, late flowering, winter hardiness and tolerance to diseases are among the most important improvement criteria for walnuts (Yarilgac, 1997; Sen et al., 2001; Muradoglu, 2005). In addition, walnut kernels contain the unsaturated fatty acids that are valuable for human health. When the oil contents and oil acid compositions of walnuts grown in the Van Lake Basin is examined, it is seen that they are especially rich in unsaturated fat acids (Yarilgac, 1997). Compared to walnuts grown in other regions of Turkey, those grown in the Van Lake Basin offer commercial advantages, with high fruit weights and high kernel percentages. Previous studies on regional classification of walnut fruit characteristics generally used cluster, factor and discriminant multivariate analysis methods. Cluster and factor analyses also classify the data set. However, mixture modeling has two major advantages when compared with cluster and factor analyses (Arminger and Stein, 1997; Muthén and Muthén, 2002, Jones et al., 2001; Jones and Nagin, 2007). The first advantage is that, for each observation, the method determines the probability that it will be included within a particular sub-population; the second advantage is that parameter estimations are obtained for each sub-population (Wang et al., 1998; Muthén and Muthén, 2002; R, 2007; Martinez et al., 2009).

In this study, the fruit traits of walnuts obtained from the Van Lake Basin are defined and classified according to the districts of the basin. This study aimed only to classify the data set using mixture modeling. Information about the data set is given. As well as the multinomial logit model and the mixture models are discussed. The model selection, the distribution of observations into sub-populations, and the estimated means of walnut fruit traits for each sub-population are given. The discussion and conclusion are given in the last part of the paper.

DATA SET

This study was conducted in 2006, using seedling walnut (*Juglans regia L.*) trees growing in eight districts (central Van, Edremit, Gevaş, Çatak, Erciş, Tatvan, Adilcevaz, and Ahlat) located in the Van Lake Basin (Van and Bitlis provinces), in eastern Turkey. Following surveys in the study area, ten walnut genotypes representing each district were marked. Walnut fruits were collected from the trees at the harvest time. A sample of fifteen fruits was taken randomly from each tree and their important nut characteristics for variety-breeding objectives were evaluated. After the green hulls of the collected fruits were removed, the fruits were first stored at room temperature for several days and then dried in an oven at 30°C for 24 h (Şen et al., 2001).

The dried samples were analyzed for pomological traits, including: fruit weight, kernel weight, kernel ratio, fruit width, fruit length, fruit height, and shell thickness, which are considered to be important fruit characteristics for walnut breeding. The fruit weight and kernel weight were measured in gram using a scale with a sensitivity of 0.0001. Each sample was weighed with its shell and then the shell was broken and weighed in full to determine the kernel weight. The fruit width, length, height, and shell thicknesses were measured in millimeter using a digital caliper with a sensitivity of 0.01. Accordingly; fruit height was measured along the axis bisecting the fruit; the fruit length was measured as the length that cuts this axis perpendicularly; and the fruit width was measured as the cheek length of both sides; all measurements were in millimeter. To measure the shell thickness, each cheek was broken from the middle and the thickness was measured in millimeter using a digital caliper, from the point equidistant from both ends.

METHODS

Multinomial logit model

Let $\pi_{ik} = P(c_i = k)$ denotes the probability that the i^{th} response falls in the k^{th} sub-population (Okut et al., 2002). In multinomial distribution, the probability distribution of count C_{ik} given the total N is,

$$p\{C_{i1} = c_{i1}, C_{i2} = c_{i2}, ..., C_{ik} = c_{iK}\} = \binom{N}{c_{i1}, ..., c_{iK}} \pi_{i1}^{c_{i1}}, ..., \pi_{iK}^{c_{iK}} \quad (1)$$

The model for the multinomial logit is,

$$\pi(c_i) = \frac{\exp\{\nabla_{ik}\}}{\sum_{k=1}^{K} \exp\{\nabla_{ik}\}} \quad (2)$$

$$\nabla_{ik} = \log \frac{\pi_{ik}}{\pi_{iK}}$$

Where

Let u represent a binary outcome of the Latent class analysis model (LCA) and c represent the categorical latent variable with K classes. The marginal probability density of u_i is,

$$p(u_i = 1) = \sum_{k=1}^{K} P(c = k)\, P(u_i = 1 \mid c = k) = \sum_{k=1}^{K} \pi_k f(y, \theta_k) \quad (3)$$

The conditional function of y is,

$$f(y/x, \theta_k) = \sum_{k=1}^{K} \pi_k f(y/\theta_k) \quad (4)$$

Where, θ_k k-dimensional unknown parameter vector, π_k the probability of sub-populations k, and y is a dependent variable. The posterior probability that observation y belongs to class j is (Grün and Leisch, 2009),

$$P(j \mid y, \theta_k) = \frac{\pi_j f(y \mid \theta_j)}{\sum_{k=1}^{K} \pi_k f(y \mid \theta_k)} \quad (5)$$

The log likelihood function for the complete data is,

$$\log L = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f(y_i \mid \theta_k) \right) \quad (6)$$

EM algorithm finds the maximum likelihood estimates using an iterative procedure consisting of two steps: an E-step and a M-step. At the E-step, $\hat{P}_{nk}$ posterior class probability for each observation can be given as following (Leisch, 2004),

$$\hat{P}_{ik} = P(k \mid y_i, \theta_k)$$

Using equation 5 and the ML estimators of the sub-populations (the prior class probabilities) are calculated as the sample averages of the estimated weights (Leisch, 2004),

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} \hat{P}_{ik}}{N}$$

At the M-step, After maximizing log-likelihood function in equation 6, ML estimations of unknown parameter vector (θ_k) are obtained.

Model selection

Akaike's information criteria (AIC), Bayesian information criteria (BIC) are two widely used criteria for mixture model. The model having the lowest AIC and BIC values and the highest entropy classification probability is accepted as the best model. In addition, entropy is used to make classification of each observation correctly. AIC and BIC are (Wang et al., 1996; Wang and Putterman, 1998),

$$AIC = -LogL + 2p \quad (7)$$

$$BIC = -LogL + p \ln(n) \quad (8)$$

Where p denotes number of parameters. After the number of classes is determined, the entropy criterion determines the probability of all the individuals being in all the classes. Entropy criteria is,

$$E_c = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik}) \ln \hat{p}_{ik}}{n \ln K} \quad (9)$$

Entropy value ranges between the ranges of $0 \leq E_C \leq 1$. As the E_C value approaches 1, it is understood that the probability for the individual being distributed correctly is high (Okut et al., 2002).

RESULTS

Statistical analyses were made using Mplus software. Initially, the data was analyzed to determine the number of different sub-populations of walnut fruit traits within the eight districts. If there are similarities in the fruit traits between districts, the number of sub-populations within these eight districts must be determined. The model selection criteria for the mixture model used to determine the numbers of sub-populations are given in Table 1. The model selection criteria for a model with up to seven sub-populations are given in Table 1.

Since the AIC and BIC criteria continue to increase after the model with five sub-populations, the models containing more than seven sub-populations were not given. It is generally accepted that the model with the smallest AIC and BIC model selection criteria best describes the data set. The AIC and BIC selection criteria for the model with five sub-populations were smaller than those obtained for the models with other sub-populations. Therefore, the model with five sub-populations best described the data distribution.

Thus, it was determined that the eight districts are separated into five sub-populations according to the fruit traits. The entropy criteria given in Table 1 shows the extent to which the models with different sub-populations are correctly classified. The highest entropy criterion was obtained for the model with five sub-populations. This result shows that the model with five sub-populations was 91.7% accurate classifying the districts according to the fruit traits.

Since the model with five sub-populations is the best for the districts according to the fruit traits, the results were interpreted according to the model with five sub-populations. Table 2 shows the distribution of the 1200 observations obtained from eight districts according to the model with five sub-populations. For example, the

Table 1. Model selection criteria for different sub-populations.

Sub-populations	Selection criteria		
	AIC	BIC	Entropy (%)
Model with one sub-population	4566.433	4637.694	-
Model with two Sub-populations	4193.171	4376.413	86.9
Model with three Sub-populations	3784.653	4079.878	90.9
Model with four Sub-populations	3624.859	4032.069	86.6
Model with five Sub-populations	3289.583	3808.771	91.7
Model with six Sub-populations	3304.610	3935.779	88.9
Model with seven Sub-populations	3319.039	3963.039	87.9

Table 2. Distribution of data set to model with five sub-populations.

Sub-populations	N	Ratio (%)
Sub-population 1	279	23.25
Sub-population 2	292	24.33
Sub-population 3	133	11.08
Sub-population 4	241	20.08
Sub-population 5	255	21.25

Table 3. Rates of correct classification for model with five sub-populations.

	Sub-populations				
	1	2	3	4	5
1	0.941	0.045	0.001	0.007	0.006
2	0.025	0.940	0.017	0.009	0.015
3	0.006	0.070	0.898	0.010	0.015
4	0.006	0.015	0.010	0.965	0.004
5	0.002	0.014	0.008	0.004	0.972

model allocated 279 (23.25%) observations to the first sub-population. As can be seen from Table 2, the second sub-population had the highest number of observations and the third sub-population had the fewest observations. The correct classification ratios for observations that fall into each sub-population were given in Table 3. The correct classification ratio for the first sub-population was obtained to be 94.1%, for the second sub-population, it was 94.0%, for the third sub-population, it was 89.8 %, for the fourth sub-population, it was 96.5%, and for the fifth sub-population, the correct classification ratio was obtained to be 97.2%. As it can be seen in Table 3, in the model with five sub-populations, the correct classification ratio for each sub-population was very high.

The mean values of the fruit traits for each of the five sub-populations are given in Table 4. For instance, the means of fruit weight was 8.221 g for the first sub-population, 8.575 g for the second sub-population, 10.585 g for the third sub-population, 13.357 g for the fourth sub-population, and 14.068 g for the fifth sub-population.

The distributions of the 8 districts studied with respect to the fruit traits are given in Table 5. For instance, the walnuts taken from Gevaş district, 31 (20.67%) were included into the second sub-population, 10 (6.67%) into the fourth sub-population, and 109 (72.27%) into the fifth sub-population. Of the walnuts taken from Tatvan district 14 (9.33%) were included into the second sub-population, 37 (24.67%) into the third sub-population, 3 (2%) into the fourth sub-population, and 96 (64%) to the fifth sub-population.

Conclusion

The AIC and BIC were found to be lowest for the model with five sub-populations. For five sub-populations, the determined entropy criterion of 91.7% showed how accurately the observation values were classified. In

Table 4. Estimated means of variables for model with five sub-populations.

Sub-populations	Variables					
	Fruit weight (g)	Kernel weight (g)	Fruit height (mm)	Fruit width (mm)	Fruit length (mm)	Shell thickness (mm)
1	8.221	1.228	33.537	27.877	29.532	1.550
2	8.575	3.848	32.986	29.225	28.857	1.365
3	10.585	4.881	35.207	32.440	31.758	1.287
4	13.357	6.299	36.815	33.510	33.632	1.206
5	14.068	5.487	42.684	34.906	33.952	1.496

Table 5. Distribution rates and counts of disctricts into model with five sub-populations.

Sub-populations	Districts								
	Gevaş (%)	Tatvan (%)	Adilcevaz (%)	Ahlat (%)	Edremit	Erciş	Çatak (%)	Van-central (%)	Total (%)
1	-	-	-	7 (4.67)	12 (8)	150 (100%)	-	110 (73.33)	279 (23.25)
2	31(20.67)	14 (9.33)	-	4 (2.67)	114(76%)	-	119 (79.33)	10(6.67)	292 (24.33)
3	-	37 (24.67)	23 (15.33)	70 (46.66)	-	-	-	3 (22)	133 (11.08)
4	10 (6.67)	3 (2)	123 (82)	69 (46)	-	-	16 (10.67)	20 (13.33)	241(20.08)
5	109 (72.27)	96 (64)	4 (2.67)	-	24(16%)	-	15 (10)	7 (4.67)	255 (21.25)
Total	150	150	150	150	150	150	150	150	1200

parallel to this, it was determined that there was a very high probability that the sub-populations given in Table 3 were classified correctly. The modeling results in Table 4 show that the fruit weight, kernel weight, fruit width, fruit length, and fruit height of the walnut fruit increase from the first sub-population to the fifth sub-population. In all districts within the first sub-population in Table 4, the lowest average values were obtained for all fruit traits apart from the shell thickness. The thinnest shell value, but nonetheless the highest average values for the kernel weight and the kernel ratio were within the fourth sub-population. In walnut improvement, high kernel weight and thin shell are desired fruit traits (McGrananhan and Leslie, 1991; Germain, 1997). We think that it is no coincidence that these characteristics, which are important for walnut improvement, were included in the fourth sub-population and that the highest addition percentages to these traits occur particularly within the Adilcevaz and Ahlat districts.

Previous walnut selection studies conducted in the Van Lake Basin reported that walnuts with the highest improvement value and those that were the most promising were selected, and that the Van Lake Basin area showed wide genetic diversity between walnut traits (Şen et al., 2001; Muradoğlu, 2005). In addition, it was also not considered to be a coincidence that walnuts within the fifth sub-population have the second highest value for kernel weight and that the biggest addition to this value came from Gevaş and Tatvan districts. On the other hand, the lowest kernel weight, which is an undesirable trait for improvement (Yarilgaç, 1997), occurred in the walnuts in the first and second sub-populations. The highest addition to these undesired traits came from walnuts growing naturally in the districts of Edremit, Erciş, Çatak, and Van center. This result can be attributed to the low quality of the natural walnut population in these districts, the difference in climatic and ecological conditions, or to insufficient technical and cultural practices. In addition, the results of this study suggest that mixture modeling can be used to statistically analyze genetic resources on the basis of population and location by taking into account the fruit

improvement traits.

ACKNOWLEDGEMENT

Financial support from Y.Y.U. Scientific Research Projects Project no: 2006-FBE-B05 is gratefully acknowledged.

REFERENCES

Arminger G, Stein P, (1997). Finite mixtures covariance structure models with regressors. Sociol. Method. Res., 26: 148-182.

Dalrymple Hudson IL, Ford RPK (2003). Finite Mixture, Zero-Inflated Poisson and Hurdle Models with Application to SIDS. University of Canterbury, Christchurch, New Zealand. p. 19.

Dempster AP, Laird NM, Rubin DB (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Royal. Stat. Soc, 39: 1-18.

Germain E (1997). Genetic improvement of the persian walnut (Juglans regia L.). Acta Hortic., 442: 21-31.

Han J (2009). Initial classification of joint data in EM estimation of latent class joint model. J. Multivariate. Anal, 100(10): 2313-2323.

Jansen RC (1993). Maximum Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm. Biometrics, 49(1): 227-231.

Jedidi K, Jagpal HS, Desarbo WS (1997). Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. Mark. Sci., 16: 39-59.

Jones B, Nagin SD, Roeder K (2001). A SAS Procedure Based on Mixture for Estimating Developmental Trajectories. Sociol. Methods.Res., 29(3): 374-393.

Jones B, Nagin SD (2007). Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. Sociol. Methods. Res., 35(4): 542-571.

Grün B, Leisch F (2008). Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects. J. Classification, 25: 225-244.

Leisch F (2004). A General Framework for Finite Mixture Models and Latent Class Regression in R. J. Stat. Software, 11(8): 1-18.

McGrananhan GH, Leslie C (1991). Walnuts (*in*). (Ed. HJ Brooks) Genetic Resources of Temperate Fruits and Nut Crops, ISHS, Wageningen, The Nederlands, 2: 907-951.

Martinez MJ, Lavergne C, Trottier C (2009). A mixture model-based approach to the clustering of exponential repeated data. J. Multivariate. Anal., 100(9): 1938-1951.

Muradoglu F (2005). Selection of promising genotypes in native walnut (Juglans regia l.) populations of Ahlat (Bitlis) and central Hakkari districts, and genetic diversity. Y.Y.U. J. Grad. Sch. Sci. (unpublished doctorate thesis), Van (in Turkish).

Muthén L.K, Muthén B (2002). Mplus: User's guide. Los Angeles, CA: Muthén and Muthén

Okut H, Duncan ET, Duncan CS, Strycker AL (2002). Latent Variable Mixture Modeling: Analyzing Mixture and the Structural Portion of Model. Joint Satatistical Meetings (JSM), August, New York City. R

Development Core Team R (2007): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. pp. 11-15.

Roeder K, Lynch GK, Nagin SD (1999). Modeling Uncertainty in Latent Class Membership: a Case Study in Criminology. J. Am. Stat. Assoc., 447: 766-776.

Şen SM, Balta F, Koyuncu MA., Koyuncu F, Yarılgaç T, Kazankaya A (2001). Lateral fruitfulness on Turkish standard walnut cultivars and promising selections. Acta Hortic., 317: 171-174.

Wang P, Puterman ML, Cockburn IM, Le N (1996). Mixed Poisson Regression Models With Covariate Dependent Rates. Biometrics, 52: 381-400.

Wang P, Cockburn IM, Puterman ML, Cockburn IM (1998). Analysis of Patent Data-Mixed Poisson Regression Approach .J. Bus. Econ. Stat., 16(1): 27-41.

Wang P, Putterman ML (1998). Mixed Logistic Regression Models. J. Agric. Biol. Env. Stat., 3(2): 175-200.

Yarılgaç T (1997). Studies on breeding by selection of walnuts on Gevaş district. Y.Y.U. J. Grad. Sch. Sci. (unpublished doctorate thesis), Van, Turkey.

Yeşilova A (2003). The Use of Mixed Poisson Regression Models for Categorical Data in Biology. Ph. D. Dissertation, University of Yüzüncü Yıl, Van, Turkey.