

USING PARTIAL LEAST SQUARES REGRESSION TO DETERMINE VARIABLE IMPORTANCE IN EXPECTED GENETIC GAIN IN OPEN NUCLEUS BREEDING SCHEME TO IMPROVE MILK PRODUCTION IN EGYPTIAN BUFFALO

S.A.M. Abdel-Salam¹, M.A.M. Ibrahim¹, Manal El-Sayed² and S. Abou-Bakr¹

1- Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt, 2- Animal Production Department, Faculty of Agriculture, Ain Shams University, Cairo, Egypt

SUMMARY

The purpose of this paper was to study variable importance in the projection (VIP) determining the expected genetic gain when an applying open nucleus breeding scheme (ONBs) to improve milk production of Egyptian buffaloes.

Populations of buffaloes with one milk record for each cow were generated with assumed mean (0) and variance (1). Four different population sizes (z): 10000, 25000, 50000 and 100000 cows were generated. Four generations of progeny were obtained by selection of sires and dams of the next generations. The structure of ONBs was applied for different population size. Genetic gain per generation was calculated for the whole system of ONBs (GGS), base population (GGB) and nucleus population (GGN).

A high correlation was observed between GGS and generation (G), mating ratio (MR) and fraction of base sires born in nucleus (w). The high correlation between explanatory variables was observed between fraction of total population in nucleus (p) and fraction of base dams born in nucleus (y). G , MR and w had high influence in the model ($VIP > 1$). The goodness of fit coefficients, the analysis of the model corresponding to GGS allow to conclude that the model is well fitted ($R^2 = 0.77$). The G had the highest weight in the model of GGS. Under cross-validation, the model of predicted GGS was fit.

MR had the highest influence on GGB and that reflect the importance of high selection intensity in maximizing the genetic gain in base population. The G and MR had high influence in the model ($VIP > 1$) of GGN. While, w had a moderate ($0.8 < VIP < 1$) influence on GGN.

Keywords: Open nucleus breeding scheme, genetic gain, simulation, PLS, Egyptian.

INTRODUCTION

According to FAOSTAT (2009), the Egyptian buffalo represents about 2.7% and 8.4% of the world buffalo's milk and meat, respectively. The proportion of number of Egyptian buffaloes to the world buffalo milk production decreases year after year. The increase of total milk production of Egyptian buffaloes was due only to increasing buffalo population. Lack of effective sustainable breeding programs for local breeds in developing countries is a reason that such breeds lose their

Issued by The Egyptian Society of Animal Production

competitive advantage, especially where production systems or external conditions are subjected to change (Hiemstra *et al.*, 2007).

The Egyptian sustainable agriculture development strategy (SADS) towards 2030 published by the Ministry of Agriculture and Land Reclamation (MALR 2009) focused on improving agriculture productivity and on the importance of Buffalo in agricultural production in Egypt. Which contains, one of the main components of a major program to develop animal production was adopting a long term national program for the genetic improvement of buffaloes, through establishing nucleus herds at the level of governorates.

The open nucleus breeding scheme offers a simpler procedure for producing and disseminating breeding stock of known value (Hinks, 1978; Cunningham, 1979 and 1987 and Jasiorowski, 1991). Bondoc and Smith (1993) recommended the establishment of two-tier open nucleus breeding system to maximize genetic improvement, and to reduce the inbreeding rate and the total cost of recording in smallholder system. Several studies indicated the significance of using open nucleus breeding scheme to improve milk production of buffalo and increase the rate of genetic gain (Dixit and Sadana, 1999, Abdel-Salam *et al.*, 2004 and Nigm *et al.*, 2005).

Abdel-Salam *et al.* (2010) suggested that the open nucleus breeding scheme offers a suitable practical procedure for producing and disseminating buffalo bulls of known breeding values. The authors found that applying ONBS for many generations of selection could accelerate the rate of genetic gain of milk production in buffalo and increase the average milk yield by 15% in first generation (*G1*) to 26% in fourth generation (*G4*). They recommended taking into account the mating ratios and the combinations of population size (*z*) and fraction of total population in nucleus (*p*) which affect proportion of all males born used as sires (*a*) and fraction of base sires born in nucleus (*w*).

The purpose of this paper was to use the partial least squares regression (PLSR) to study variables importance in the projection (VIP) that determines the expected genetic gain of applying ONBs to improve milk production of Egyptian buffaloes.

MATERIALS AND METHODS

Basic parameters

Simulation technique was used to generate population of buffaloes with one milk record for each buffalo. A total number of 3526 lactation records (from first to sixth lactation) of 2179 buffaloes recorded by Cattle Information Systems/Egypt (CISE, 2007) of the Faculty of Agriculture, Cairo University during the period from 1990 to 2006 were used to estimate various parameters (Table 1) which were used to generate the simulated populations.

Table 1. The estimates used in generating the simulated populations

Parameter	Estimate	Reference
Average total milk yield (TMY)	1884, kg	CISE, 2007
Phenotypic standard deviation for TMY	438, kg	CISE, 2007
Average generation interval	5.78, yr	Mourad, 1990
Heritability	0.17	CISE, 2007

CISE: Cattle Information Systems/Egypt

Generation of the simulated population of buffaloes

Populations were generated using stochastic Monte Carlo simulation procedure of SAS (2004) with assumed mean (0) and variance (1). Simulation was replicated four times to generate four different sizes of populations: 10000, 25000, 50000 and 100000 animals. Four generations were generated by selection of progenies with the highest breeding values for milk production to be sires and dams of the next generation. Selected bulls and heifers were mated at random. Abdel-Salam *et al.*, 2010 showed the structure of designed open nucleus breeding scheme and gene migration from nucleus to base and vice versa.

Genetic gains per generation in ONB system i.e. the whole breeding system (GGS), base population (GGB) and nucleus population (GGN) were estimated as the average genetic selection differentials of animals used in nucleus and base (James, 1977). The rate of progress as response to selection is modified by migration of genes between parts of the population with different mean breeding values, i.e. from the nucleus to the base and vice versa. The best selected males are kept for breeding in the nucleus while other selected males are given to the base herds for breeding. The best selected females are maintained in the nucleus while females disseminated to the base herds are surplus to the nucleus replacement. By this means, improvements are quickly spread throughout the population. The nucleus remains open to animals from the base herds, the best selected females in the base are transferred to the nucleus. Thus, the ultimate genetic gain was estimated by a weighted average of the genetic selection differentials in the nucleus and base, the weights being the gene migration rates from each part of the population to the other. Both additive genetic and residual effects were assumed to be distributed normally. This procedure was carried out for four generations. It was assumed that all buffalo females were mated naturally or artificially to bulls coming from the nucleus or base. Mating ratio (male: females) were 1:25, 1:50 and 1:100 in natural mating (NM), and 1:1000, 1:2500 and 1:5000 in artificial insemination (AI).

The genotypes of animals were simulated by the following formula (1) according to (Meuwissen, 1991):

$$g_i = \frac{1}{2}g_s + \frac{1}{2}g_d + a_i\sqrt{0.5h^2\sigma_p^2} \quad (1)$$

Where g_i , g_s , and g_d are the additive genetic values of individual i , its sire (s) and its dam (d), respectively, h^2 is the heritability in population ($h^2 = 0.17$), σ_p^2 is the phenotypic variance, a_i is the random number from the distribution $N(0,1)$ and

$\sqrt{0.5h^2\sigma_p^2}$ is the Mendelian sampling term.

The record of buffalo i was simulated by the following formula (2) according to (Meuwissen, 1991):

$$Y_i = \mu + g_i + E_i \quad (2)$$

Where Y_i is the lactation yield of the i^{th} buffalo, μ is the mean total milk yield, g_i is the additive genetic value of individual i , E_i is the environmental effect of the i^{th} buffalo assumed to equal $\sigma_e a_i$, σ_e is the square root of the error variance and a_i is the random number from the distribution $N(0,1)$.

Estimation of breeding values

Breeding values were estimated using the Derivative – Free Restricted Maximum Likelihood (DF-REML) procedure (Meyer, 1998).

The linear animal model (3) used for analyzing simulated records was as follows:

$$C = D\beta + Z_a a + e \quad (3)$$

Where C is the vector of observations, D is the incidence matrix for assumed fixed effect required by software to run distributed to the data at random, β is vector of an overall mean and fixed effect (2 levels) distributed to the data at random, Z is the incidence matrix for random effects, a is the vector of direct genetic effects of buffalo and e is vector of random errors normally and independently distributed with zero mean and variance $I\sigma_e^2$.

Calculation of the expected genetic gain in milk yield

Seven variables, namely, number of generations (G), population size (z), nucleus size (p), mating ratio (male: females) (MR), proportion of males born used as sires (a), fraction of base sires born in nucleus (w) and fraction of base dams born in nucleus (y) were studied.

The parameters and values of variables assumed for estimating the genetic gain of milk yield of the simulated population applying open nucleus breeding scheme and levels of variables are shown in table 2. Genetic gain per generation and annual genetic gain of milk yield were calculated according to James (1977).

Table 2. Input variables for calculating genetic gain

Item Description	Symbol	Value
Heritability	h^2	0.17
Replacement rate	R	0.20
Fraction of total population in nucleus	P	0.05, 0.10, 0.20
Fraction of nucleus dams born in base	X	0.50
Fraction of base dams born in nucleus	Y	0.08, 0.14, 0.17
Fraction of nucleus sires born in base	V	0.00
Fraction of base sires born in nucleus	w	0.10, 0.20, 0.30, 0.60, 1.00
Proportion of all males born used as sires	a	0.05, 0.10, 0.20
Proportion of all females born used as dams	b	0.80
Simulated population size (number of breedable females)	z	10000, 25000, 50000, 100000
Generations	G	1, 2, 3, 4

Partial Least Squares Regression (PLSR)

PLSR analysis finds a linear model that relates the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), for explanatory purposes. The PLS model is based on principal components on both the independent data (seven variables), X, and the dependent variable (expected genetic gain), Y.

Gaspar *et al.*, 2008 illustrated that the PLS regression is a multivariate calibration method, by which two sets of data, X and Y are related by means of regression. The purpose of PLSR is to establish a linear model, which enables the predictions of Y from the measured X. The contribution of each variable to predict the genetic gain in this study was evaluated using the regression coefficients obtained for the

standardized variables. These coefficients allow the selection of those variables that most contributed to the prediction.

Hajiha, 2009 reported that PLS provides various types of plots that can be used to assess the quality of a model. Correlations are seen using coefficient plots where X is centered and scaled, while Y is scaled but not centered. The coefficient plot shows how strongly Y is correlated to each input variable, X, since the data are scaled and centered, the coefficients are comparable. A bar in the coefficient plot shows the correlation of each input with that specific output, where bars above zero represent positive correlations, and bars below zero represent negative correlations. If the error bar (the bar passing through the correlation bar) passes through the zero line, a conclusive correlation has not been found for that variable. Variable Importance in the Projection (VIP) summarizes the importance of X-variables both for the X and Y models. VIP values are calculated for each X by summing the squares of the PLS weights, taking into account the amount of Y-variance explained in each dimension. There is always only one VIP for a given model and problem. Observe that VIP values are all positive as it is a squared function of the PLS weights. The VIP values may be used for variable selection since variables larger than 1 indicate "important" X variables, and values smaller than 0.8 indicate "unimportant" X variables.

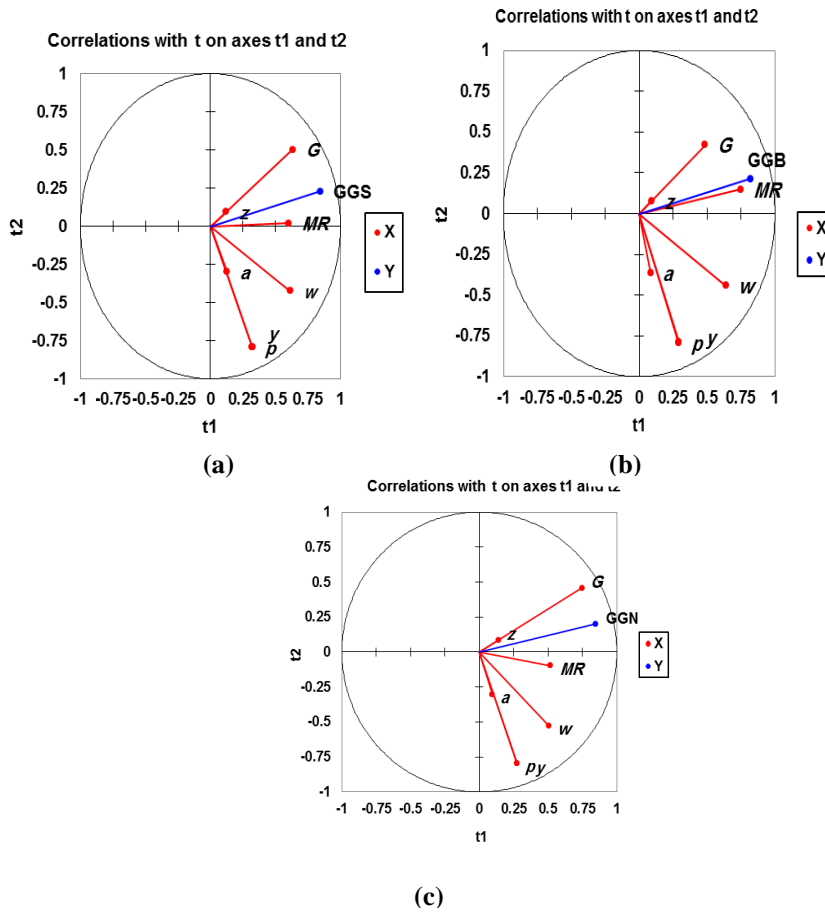
The response variables (Y-variables) analyzed by PLS in the present study are the expected genetic gain in ONBs (GGS), the expected genetic gain in Base population (GGB) and the expected genetic gain in nucleus population (GGN) of milk production in kg. The X-variables consisted of the 7 variables (number of generations (G), population size (z), nucleus size (p), mating ratio (male: females) (MR), proportion of males born used as sires (a), fraction of base sires born in nucleus (w) and fraction of base dams born in nucleus (y)). PLS regression analyses in this study were done using PLSR procedure for XLSTAT (2010).

RESULTS AND DISCUSSIONS

Figure 1.a shows the correlations map between variables of the determined genetic gain in ONBs (independent variables) and the genetic gain in ONBs; GGS; (dependent variable). We can see that for some variables displayed at the center of the map, the correlations are low. The high correlation was observed between GGS and G , MR and w . The high correlation between independent variables was observed between p and y . The same trend was observed in the correlations map between variables of determined genetic gain in base and the components and the genetic gain in base, GGB, (Figure 1.b) and the correlations map between variables of determined genetic gain in base and the components and the genetic gain in nucleus; GGN; (Figure 1.c).

Figure 2 shows the VIPs for each variable (explanatory variables). This allows to quickly identifying which are the explanatory variables that contribute the most to the model. The G , MR and w had high influence in the model ($VIP > 1$) which means that these three variables had a high influence in the expected genetic gain of milk yield when applying ONBs (figure 2.a). G had the highest influence on GGS due to the cumulative genetic improvement by generations. Mueller and James (1983) reported that the cumulated gain after 10 generations of selection, in a likely sheep or cattle system, would be overestimated by about 20%. This may be important in evaluations

of nucleus schemes. *MR*, also, had a high influence on GGS and that reflects the importance of selection intensity in maximizing the genetic gain.



* $t = [t_1, \dots, t_c]$ is the matrix of latent components (PLS components)

Figure 1. Correlation circle between explanatory variables (X: G: generations, MR: mating ratio, w: fraction of base sires born in nucleus, z: population size, a: proportion of males born used as sires, p: nucleus size and y: fraction of base dams born in nucleus) and dependent variable (Y, GGS: genetic gain in ONBs, GGB: Genetic Gain in Base and GGN: Genetic Gain in Nucleus).

As shown in figure 2.a, the top variables in the VIP plot have higher correlation with the genetic gain. On the average, variables with magnitude greater than one in the plot are more relevant in predicting genetic gain. Neglecting these variables will definitely result in a low expected genetic gain by applying ONBs. At the same time, variables with low correlation with genetic gain must not be ignored. The goodness

of fit coefficients, the analysis of the model corresponding to GGS, allows to conclude that the model is well fitted ($R^2=0.77$).

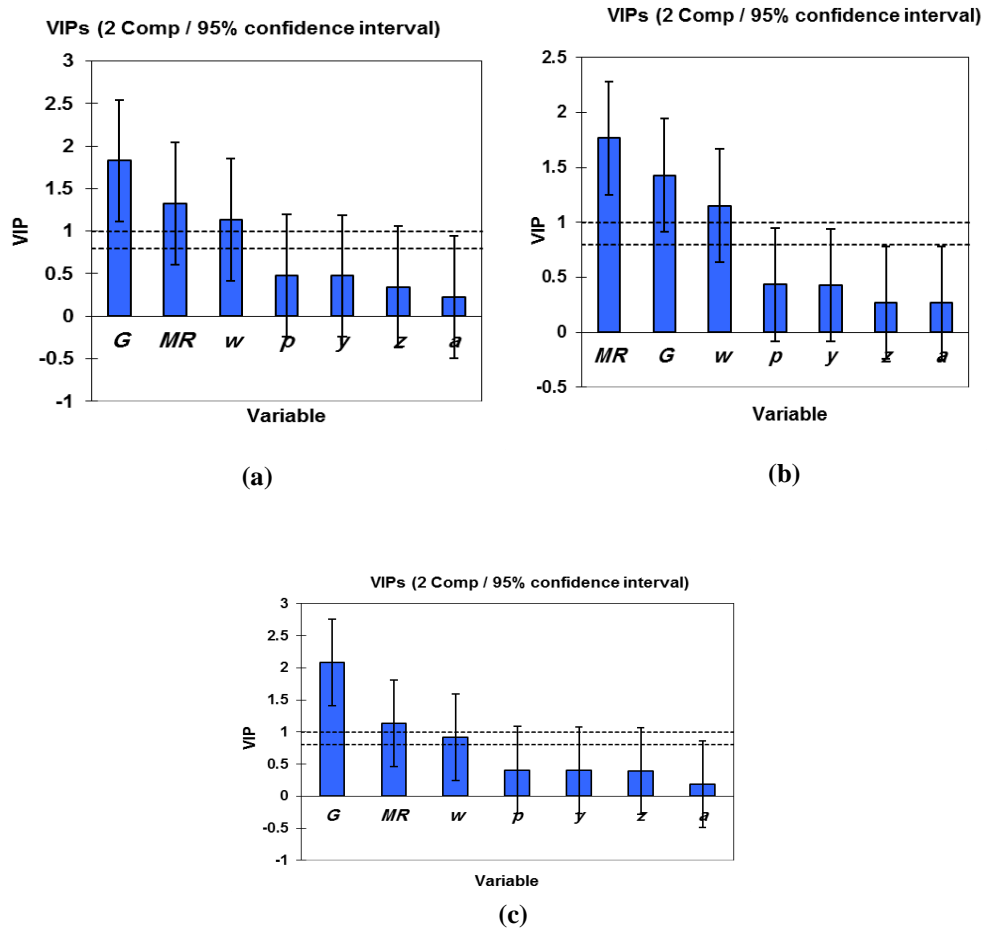


Figure 2. Variable importance in the projection (VIP)

The *MR*, *G* and *w* had high influence the model ($VIP > 1$) on expected genetic gain in base population of milk yield when applying ONBs (figure 2.b). *MR* had the highest influence on GGB reflecting the importance of high selection intensity in maximizing the genetic gain in base population. Also, *G* had a high influence on GGB. Fraction of base sires born in nucleus (*w*) had a high influence on GGB and that interprets the importance of the migration of the sire from the nucleus to the base. The goodness of fit coefficients, the analysis of the model corresponding to GGB allows to conclude that the model is well fitted ($R^2=0.72$). The *G* and *MR* had high influence in the model ($VIP > 1$) which means that these two variables had a high

influence on expected genetic gain in nucleus population (figure 2.c). *G* had the highest influence on GGN and *MR* had a high influence on GGN, while, *w* had a moderately ($0.8 < VIP < 1$) influence on GGN. The goodness of fit coefficients, the analysis of the model corresponding to GGN allows to conclude that the model is well fitted ($R^2=0.76$).

The standardized coefficients of the variables that determine the expected genetic gain when applying ONBs is shown in figure 3. The standardized coefficients allow comparing the relative weight of the variables in the model. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. In general, *G* and *MR* had the highest weight in the model.

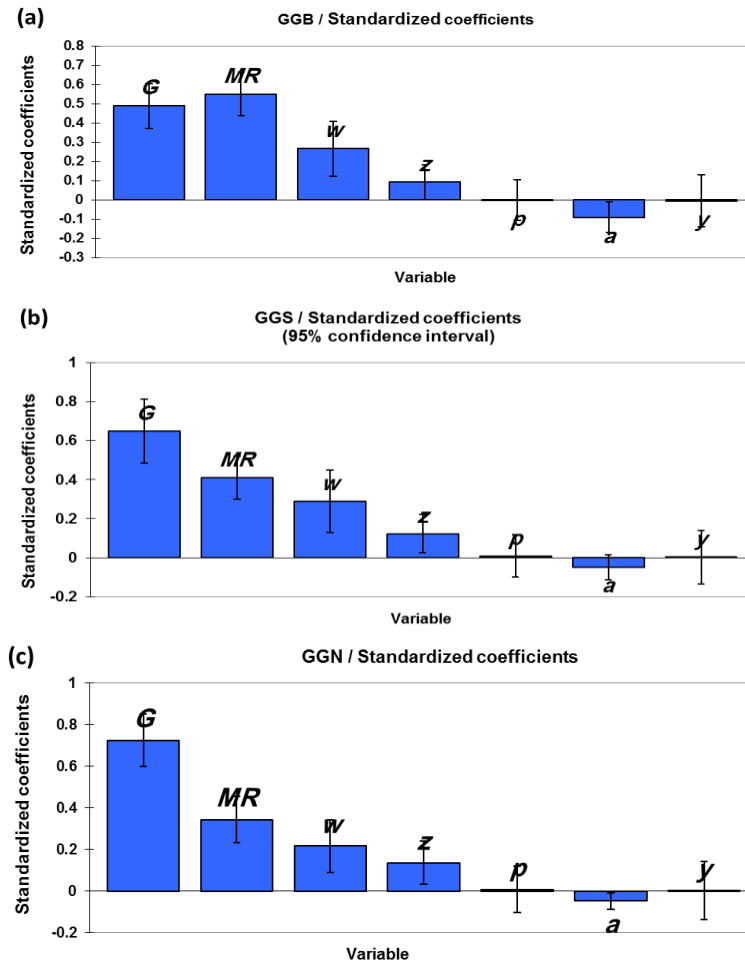


Figure 3. Standardized coefficients of the variables that determine the expected genetic gain.

Cross validation was used in variable selection. Figure (4) shows the distance between the predicted and the observed values. However, under cross-validation, the model of the predicted GGS was fitted. All the points were on the bisecting line reflecting the goodness of fit of the model of the predicted genetic gain. The same trend was observed under cross-validation of the models of predicted GGB and GGN were fitted.

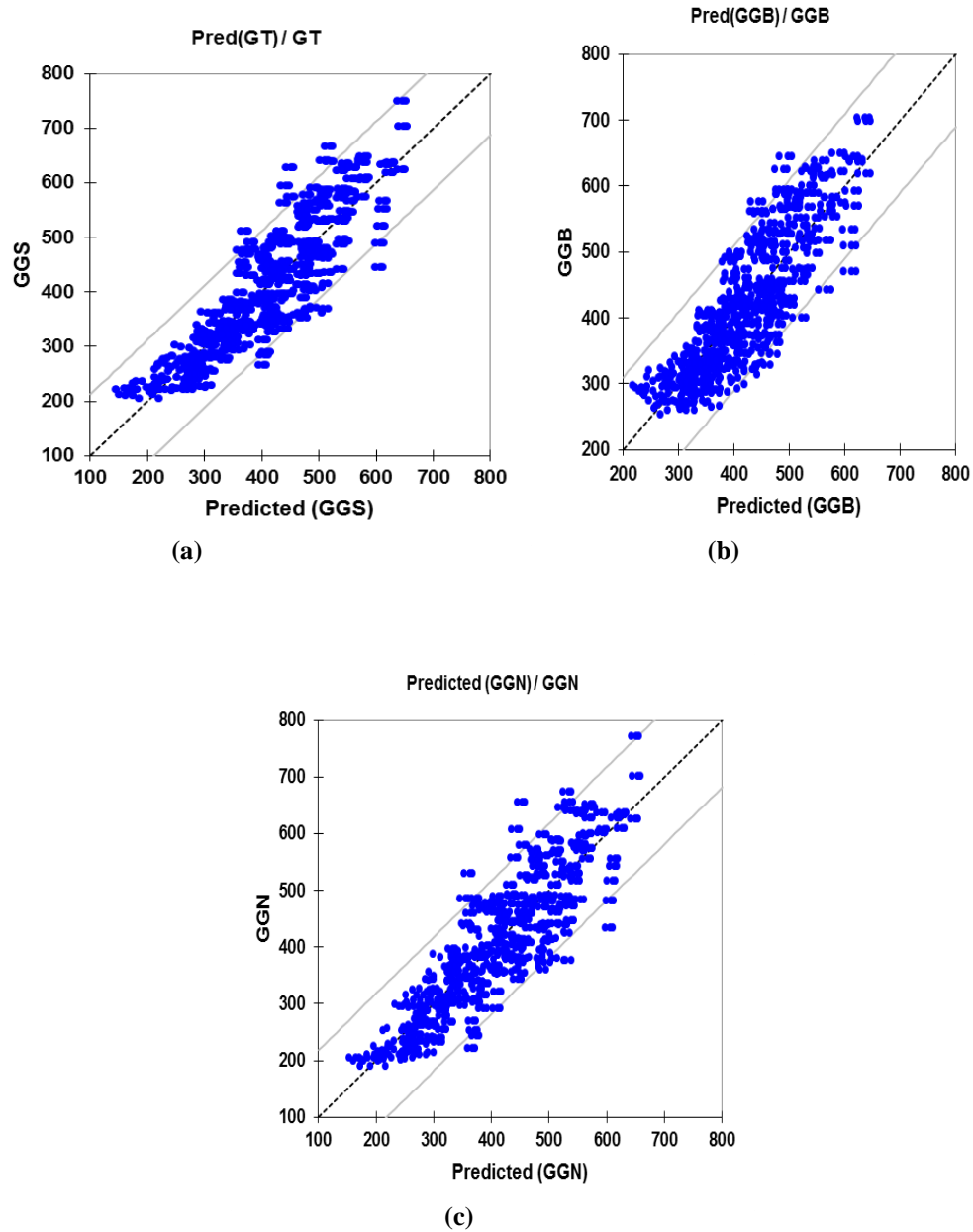


Figure 4. Cross-validation of the model to predict genetic gain

CONCLUSION

Using PLS regression illustrated the importance of *G* and *MR* when applying ONBs. The *G*, *MR* and *w* had a high influence on expected genetic gain of milk yield when applying ONBs. In base population, it is important to consider *MR* and *w*

REFERENCES

- Abdel-Salam, S.A., A.I. Sayed, Manal Elsayed and S. Abou-Bakr, 2010. Genetic gain in open nucleus breeding scheme to improve milk production in Egyptian Buffalo. *Livest. Sci.* 131: 162-167.
- Abdel-Salam, S.A.M., A.A Nigm, Manal Elsayed, R.R. Sadek and A.S. Abdel-Aziz, 2004. Genetic gain of milk production in Egyptian buffalo as affected by population size, nucleus size and selection intensity in open nucleus breeding scheme. *Egypt. J. Anim. Prod.* 42:33-42.
- Bondoc, O.L. and C. Smith, 1993. Deterministic genetic analysis of open nucleus breeding schemes for dairy cattle in developing countries. *J. Anim. Breed. Genet.* 110: 194-208.
- CISE, 2007. Buffalo herdbook 1st edition. Cattle Information Systems/Egypt (CISE), Faculty of Agriculture, Cairo University, 12613 Giza, Egypt.
- Cunningham, E.P., 1979. The importance of continuous genetic progress in adapted breeds. Report of the FAO Expert Consultation on Dairy Cattle Breeding in the Humid Tropics, pp. 35-41. FAO, Rome.
- Cunningham, E.P., 1987. Crossbreeding *Bos Taurus* and *Bos Indicus* for milk production in the tropics. *Animal Production and Health Paper* 68: 15. FAO, Rome.
- Dixit, S.P. and D.K. Sadana, 1999. Response of single trait selection in open nucleus schemes for buffalo breeding. *Indian J. Dairy Sci.* 52: 17-20.
- Eriksson, L., E. Johansson, N. Kettaneh-Wold and S. Wold, 2001. Multi-and Megavariate Data Analysis, Umetrics Academy.
- FAOSTAT, 2009. FAO Statistics Division., FAO, Rome, Italy, www.fao.org.
- Gaspar, P., F.J. Mesías b, M. Escribano a and F. Pulido, 2008. Gross output and livestock sales modelling in Spanish extensive farms using PLSR. The 107th EAAE Seminar "Modelling of Agricultural and Rural Development Policies". Sevilla, Spain, January 29th -February 1st, 2008.
- Hajiha, H., 2009. Multivariate analysis of variables affecting thermal performance of black liquor evaporators. M.Sc. Thesis, Faculty of Applied Science and Engineering, University of Toronto, Canada.
- Hiemstra, S.J., A.G. Drucker, M.W. Tvedt, N. Louwaars, J.K. Oldenbroek, K. Awgichew, S. Abegaz Kebede, P.N. Bhat, and A. da Silva Mariante, 2007. What's on the menu? Options for strengthening the policy and regulatory framework for the exchange, use and conservation of animal genetic resources. *Anim. Genet. Resour. Inf. Bul.* 4, 65-74.
- Hinks, C.J.M., 1978. The use of centralized breeding schemes in dairy cattle improvement. *Animal breeding Abstract* 46: 291.
- James, J.W., 1977. Open nucleus breeding systems. *Anim. Prod.* 24: 287.

- Jasiorowski, H.A., 1991. European animal husbandry: a model to adopt or reject by developing countries ?. on the eve of the 3rd millennium, the European challenge for animal production, EAAP Publication 48: 127. (CAB Abst.).
- MALR., 2009. Sustainable agriculture development strategy (SADS) towards 2030. Ministry of Agriculture and Land Reclamation, Dokki, Giza, Egypt.
- Meuwissen, T.H.E., 1991. The use of increased female reproductive rates in dairy cattle breeding schemes. *Anim. Prod.* 52: 21.
- Meyer, K., 1998. DFREML (Derivative-Free Restricted Maximum Likelihood) package. Version 3.0 α . 6th World Cong. Genetic. Appl. Livest. Prod., Armidal.
- Mourad, K.A., 1990. Buffaloes in Egypt. Proceedings of FAO Workshop on Buffalo Open Nucleus Breeding Schemes (ONBS), Shuman, Bulgaria.
- Mueller, J.P. and J.W. James, 1983. Effect of reduced variance due to selection in open nucleus breeding systems. *Aust. J. Agric. Res.* 34: 53-62.
- Nigm, A.A., S.A.M. Abdel-Salam, Manal Elsayed, R.R. Sadek and A.S. Abdel-Aziz, 2005. Preliminary results on use of the open nucleus breeding scheme for improving milk production of Egyptian buffalo. *Egypt. J. Anim. Prod.* 42:1-9.
- Palermo, G., P. Piraino and H. Zucht, 2009. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry.* 2: 57-70.
- SAS Procedures Guide, Version 9. 2004. SAS Inst., Inc., Cary, NC.
- XLSTAT. 2010. Data analysis and statistical solution for MS Excel. Addinsoft

إستخدام طريقة الانحدار الجزئي للحد الأدنى للمربعات لتحديد المتغير الهام في التحسين الوراثي المتوقع لقطعان النواة المفتوحة لتحسين إنتاج اللبن في الجاموس المصري

سامح عبد الفتاح محمد^١، محمد عبد العزيز إبراهيم^١، منال السيد^٢ وسامى أبو بكر^١

١ - قسم الإنتاج الحواني، كلية الزراعة، جامعة القاهرة، الجيزة، جمهورية مصر العربية، ٢ - قسم الإنتاج الحواني، كلية الزراعة، جامعة عين شمس، القاهرة، جمهورية مصر العربية

الغرض من البحث هو دراسة أهمية المتغيرات التي تحدد كمية التحسين الوراثي عند تطبيق نظام القطعان المفتوحة لتحسين اللبن في الجاموس المصري. باستخدام المحاكاه تم تكوين عشائر من الجاموس لها سجل لبن واحد لكل جاموسة بفرض أن المتوسط يساوي صفر والتباين يساوي واحد. تم تكوين اربع احجام مختلفة من العشائر وهي ١٠٠٠٠٠، ٢٥٠٠٠، ٥٠٠٠٠، ١٠٠٠٠٠٠ جاموسة. تم التطبيق الهيكلي لقطعان النواة المفتوحة. حسبت كميات التحسين الوراثي المتوقع لكل نظام قطعان نواة مفتوحة وحسبت على مستوى عشيرة القاعدة وعشيرة النواة. لوحظ ارتباط عالي بين كمية التحسين الوراثي المتوقع لتطبيق نظام النواه المفتوحة على كل العشيرة وبين رقم الجيل، ونسبة التزاوج ونسبة الذكور المستخدمة في عشيرة القاعدة والمولوده في النواة. كما لوحظ ارتباط قوي بين المتغيرات التفسيرية وبين حجم النواة ونسبة الأمهات في القاعدة المولودة في النواة. رقم الجيل ونسبة التزاوج ونسبة الذكور المهاجرة من النواة إلى القاعدة كان لها تأثير عالي في النموذج المفسر لأهميه المتغيرات. مقياس صلاحية المعاملات المستخدمه في النموذج، حيث أظهر تحليل النموذج المسؤل عن كمية التحسين الوراثي المتوقع أنه جيد وبدرجة ثقة ٠.٧٧. وأظهرت أن رقم الجيل له أعلى وزن في النموذج. تحت اختبار مدى صحه النموذج للنتبأ بكمية التحسين الوراثي المتوقع ظهر انه جيد. وكان لنسبة التزاوج أكبر تأثير على كمية التحسين الوراثي المتوقع لعشيرة القاعدة وهذا يعكس أهمية شدة الانتخاب في تعظيم كمية التحسين الوراثي المتوقع لعشيرة القاعدة. أما بالنسبة لعشيرة النواة فأظهر رقم الجيل ونسبة التزاوج أعلى تأثير في نموذج حساب كمية التحسين الوراثي المتوقع لعشيرة النواة. بينما كمية الذكور المستخدمة في القاعدة والمولودة في النواة لها تأثير متوسط.