RESEARCH ARTICLE

# Prediction of Lifetime Milk Yield using Principal Component Analysis in Gir Cattle

Nikhil S. Dangar[1]*, Pravin H. Vataliya[2]

## Abstract

The objective of the research was to investigate the relationship among production traits i.e., lactation milk yield, lactation length and lactation peak milk yield of the first three lactations using principal component analysis and formulation of prediction equation to predict lifetime milk production in Gir cattle. Data were from multiparous dairy cows of the University farm. Principal component analysis with correlation matrix was used to find the relationship among lactation milk yield, lactation length and lactation peak milk yield of first three lactation and other fixed effects, including the year of calving, season and parity with random effect of sire. The principal components were fitted to identify the best-fitted model for predicting lifetime milk yield using all principal components as a predictor in different combinations. The first six principal components (first lactation milk yield, lactation length and peak milk yield, second lactation milk yield, lactation length and peak milk yield), explained 98% variation in the estimated values with adjusted $R^2$= 59.85% variation in the estimated values. The curve estimation analysis revealed that the first six principal components as the predictor was the most fitting model for predicting lifetime milk yield. The prediction equation found most fitted will be useful for the selection of Gir cattle at an early stage of lactation.

**Keywords**: Gir cattle, Lifetime milk yield, , Multiple linear regression, Prediction, Principal component.

*Ind J Vet Sci and Biotech* (2022): 10.48165/ijvsbt.18.4.19

## Introduction

High production efficiency in livestock is an economically desirable attribute that ultimately targets genetic up-gradation. The economy of the dairy industry mainly relies upon the performance parameters of dairy animals; therefore, it becomes more relevant to tackle the means for ameliorating the performance efficiencies by developing certain guidelines for selection (Dangar *et al.,* 2015). The aim of an animal breeder is to maximize the genetic gain per unit of time for various traits of economic importance in a breed improvement programme. Dairy cattle breeding implies maximizing genetic gain mainly for milk yield and production efficiency traits. This calls for evaluating a breeding program to assess change in the genetic constitution and environmental (managemental) conditions over time in organized herds of a particular breed. The magnitude and direction of production trends in a herd indicate the effectiveness of the breeding programme and help in developing or modifying appropriate strategies for further improvement. Therefore, the prediction of important production traits of an animal at an early age is of prime importance nowadays (Dangar *et al.,* 2017). In this era of genomics, various scientists are working to predict economic traits at an early age through genomics. Some advanced breeding methodologies also provide some assumptions to do the same by using various analytical methods such as multiple linear regression and principal component analysis. Milk recording is one of the essential criteria for efficient herd management, selecting animals with higher genetic potential, and culling low producing animals (Chaudhary *et al.,* 2022). Gir cattle has huge production potential and sustains productivity in harsh climates (Parikh *et al.,* 2022). Looking at these facts present study was designed to predict the production potential of an animal at an early age using part production records.

[1]Department of Animal Genetics and Breeding, College of Veterinary Science and Animal Husbandry, Navsari Agricultural University, Navsari, Gujarat, India.

[2]Directorate of Extension Education, Kamdhenu University, Gandhinagar, Gujrat, India.

**Corresponding Author:** Nikhil S. Dangar, Department of Animal Genetics and Breeding, College of Veterinary Science and Animal Husbandry, Navsari Agricultural University, Navsari, Gujarat, India., e-mail: drnik2487@gmail.com.

**How to cite this article:** Dangar, N.S., Vataliya, P.H., Prediction of Lifetime Milk Yield using Principal Component Analysis in Gir Cattle (2022). Ind J Vet Sci and Biotech. 18(4), 92-96.

**Source of support:** Nil

**Conflict of interest:** None.

**Submitted:** 15/04/2022 **Accepted:** 20/08/2022 **Published:** 10/09/2022

## MATERIALS AND METHODS

The data pertinent to production traits on 680 Gir cows calving from 1987 to 2010, and progeny of 52 sires maintained at Cattle Breeding Farm, Junagadh, Gujarat, India were considered. The duration of 24 years was divided into 6

periods of four years each. The three seasons were delineated as winter (November-February), summer (March- June) and monsoon (July-October) based on geo-climatic conditions prevailing in the region. The parity was considered up to 12th lactation. First lactation milk yield (TLY1), first lactation length (LL1), first lactation peak milk yield (PMY1), second lactation milk yield (TLY2), second lactation length (LL2), second lactation peak milk yield (PMY2), third lactation milk yield (TLY3), third lactation length (LL3), third lactation peak milk yield (PMY3), were recorded with lifetime milk production of the same cow for the principal component analysis. Records of cows with some specific or non-specific diseases, reproductive disorders and physical injury were excluded from the present investigation. A forward selection strategy was used to find the explanatory variables for the highest determination of coefficients that worked as new explanatory variables added to the model to achieve maximum determination of coefficients. The model equation is given as

$$Y_{ijkmn} = \mu + P_i + C_j + L_k + S_m + e_{ijkmn}$$

Where, $Y_{ijkmn}$-Performance trait of the individual animal (n), calved in $(i)^{th}$ period and $(j)^{th}$ season, of the $(k)^{th}$ parity, born to $(m)^{th}$ sire; μ-overall population mean, $P_i$ fixed effect of period of calving ( i = 1 to 6), $C_j$-fixed effect of season of calving ( j = 1 to 3), $L_k$-fixed effect of parity ( k = 1 to 12), $S_m$-random effect of sire ( m = 1 to 52), $e_{ijkmn}$-random error with mean zero and variance $\sigma^2E$.

The principal component analysis is a method for transforming the variables in a multivariate data set, x1, x2, ...,xp, into new variables, y1, y2, ..., yp, which are uncorrelated with each other and account for decreasing proportions of the total variance of the original variables defined as

$$y1 = a11x1 + a12x2 + ... + a1pxp$$
$$y2 = a21x1 + a22x2 + ... + a2pxp$$
$$yp = ap1x1 + ap2x2 + ... + appxp$$

Table 1: Number of records, N, means, standard deviations, and minimum, maximum for milk yield, lactation length and peak milk yield for selected lactation.

| Variable | Mean | S. D. | N | Min | Max |
|---|---|---|---|---|---|
| First Lactation Milk Yield | 2076.6 | 1037.96 | 88 | 242.5 | 4384.6 |
| First Lactation Length | 376.1 | 131.90 | 88 | 124.0 | 703.0 |
| First Lactation Peak Milk Yield | 8.283 | 2.97 | 88 | 2.670 | 17.6 |
| Second Lactation Milk Yield | 2443.5 | 910.16 | 88 | 493.1 | 4525.3 |
| Second Lactation Length | 368.8 | 119.03 | 88 | 132.0 | 622.0 |
| Second Lactation Peak Milk Yield | 10.559 | 2.57 | 88 | 4.430 | 15.0 |
| Third Lactation Milk Yield | 2398 | 1028.04 | 88 | 312 | 5634 |
| Third Lactation Length | 335.7 | 105.79 | 88 | 105.0 | 645.0 |
| Third Lactation Peak Milk Yield | 11.813 | 3.51 | 88 | 4.400 | 22.0 |

Table 2: Eigenvalues and proportion of the variance of principal components (PC) of the correlation matrix of original variables

| Principal component | Eigen value | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 2.240 | 1.134 | 0.558 | 0.558 |
| 2 | 1.106 | 0.077 | 0.136 | 0.693 |
| 3 | 1.029 | 0.193 | 0.118 | 0.811 |
| 4 | 0.836 | 0.142 | 0.078 | 0.889 |
| 5 | 0.694 | 0.113 | 0.053 | 0.942 |
| 6 | 0.581 | 0.279 | 0.038 | 0.980 |
| 7 | 0.303 | 0.056 | 0.010 | 0.990 |
| 8 | 0.246 | 0.073 | 0.007 | 0.997 |
| 9 | 0.173 | 0.173 | 0.003 | 1.000 |

Table 3: Comparison of determination coefficients using explanatory variable (R square)

| Variable | $R^2$ value |
|---|---|
| First Lactation Milk Yield | 0.94 |
| First Lactation Length | 0.86 |
| First Lactation Peak Milk Yield | 0.78 |
| Second Lactation Milk Yield | 0.91 |
| Second Lactation Length | 0.81 |
| Second Lactation Peak Milk Yield | 0.72 |
| Third Lactation Milk Yield | 0.90 |
| Third Lactation Length | 0.80 |
| Third Lactation Peak Milk Yield | 0.76 |

With the coefficients being chosen so that y1,y2,...,yp account for decreasing proportions of the total variance of the original variables, x1,x2,...,xp. Principal component analysis with correlation matrix was used to find the relationship among TLY1, LL1, PMY1, TLY2, LL2, PMY2, TLY3, LL3 and PMY3 and other fixed effects, including breed, year at calving, season, and parity. Since scales of measurements of the production traits were different; a correlation matrix was used instead of the covariance matrix. According to cumulative explanatory proportions number of principal components was chosen and corresponding scores were estimated. Then, regression analyses were done again; coefficients of determination based on explanatory variables regression and principal component regression scores were compared.

Eigen values and scree plot methods were used to identify the principal components to be retained as predictors. Residual and diagnostics plots were examined to find out appropriate fitted model. Curve estimation analysis was undertaken to determine the models' appropriateness

Table 4: Significance of coefficients of fitted models and respective adjusted $R^2$ values

| Predictor | B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | Adj $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| First PC | 3930.15* (1767.34) | 5.70*** (0.76) | | | | | | | | | 0.39 |
| First Two PC | 7604.99** (2460.93) | 8.47*** (1.52) | -25.09* (11.93) | | | | | | | | 0.41 |
| First Three PC | 14511.46*** (3863.33) | 12.99*** (2.47) | -44.11** (14.32) | -1102.41* (482.96) | | | | | | | 0.44 |
| First Four PC | 9687.58* (3879.03) | 11.07*** (2.38) | -44.89** (13.44) | -1020.14* (454.03) | 3.45*** (0.98) | | | | | | 0.51 |
| First Five PC | 12136.03** 3978.45) | 10.68*** (2.34) | -44.12** 13.18) | -1027.69* (445.08) | 5.61*** (1.41) | -19.39* 9.27) | | | | | 0.52 |
| First Six PC | 12428.35* (4864.46) | 10.56*** (2.65) | -43.66** (13.98) | -1004.05* (500.51) | 5.81* (2.40) | -20.29 (12.59) | -53.90 (509.68) | | | | 0.60 |
| All PC | 11302.74*** (5689.68) | 11.49** (2.67) | -45.50* (13.81) | -1169.22 (489.23) | 3.54 (2.56) | -5.00 (13.05) | -97.05 (497.05) | 1.46 (2.12) | -22.11 (14.33) | 498.80 (398.28) | 0.55 |

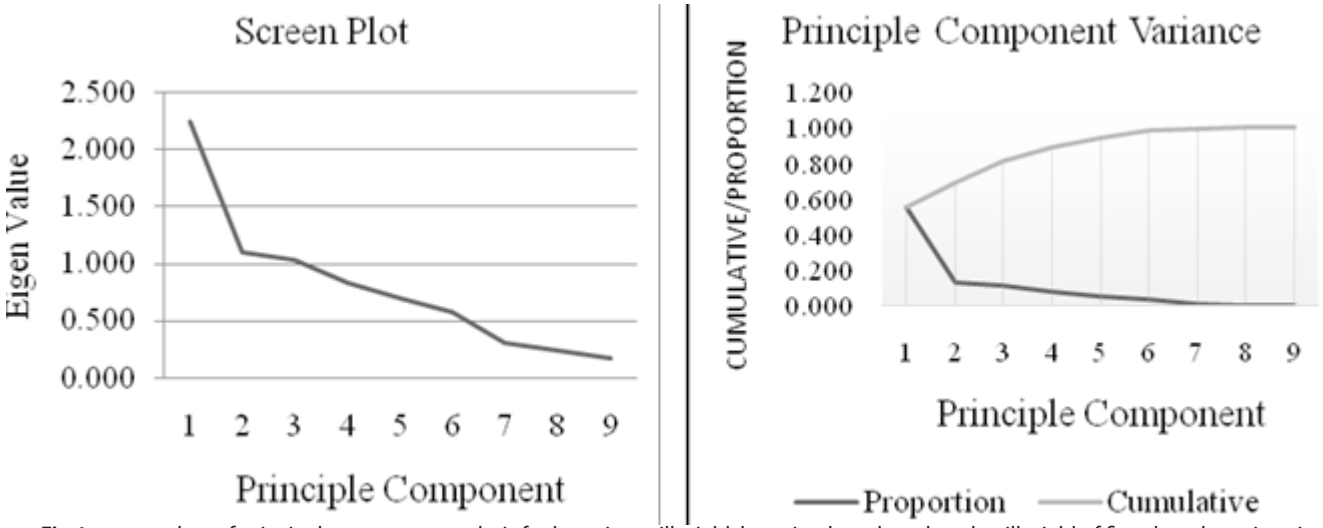


**Fig.1:** screen plots of principal component analysis for lactation milk yield, lactation length and peak milk yield of first three lactations in combinations

(lifetime milk yields with the first principal as predictor). The adequacy of the best-fitted model was adjudged based on the adjusted $R^2$-value and significance of coefficients.

## RESULTS AND DISCUSSION

The statistical method and concept like principal component analysis applies to large volumes of data collected by research organizations and provides small data after analysis to draw any conclusion. Although the limitation in data size, some of the exploratory analyses was statistically significant and will be useful. Table 1 shows the descriptive analysis of data under the present study.

Principal component analysis was performed using the explanatory variables based on the model. Investigation of screen plots (Fig. 1) shows that most variations are explained by the first six principal components (i.e. of lactation first and second). Eigen values and proportion of the variance of principal components (PC) of the correlation matrix of original variables (Table 2) and comparison of determination coefficients using an explanatory variable (R square; Table 3) show the results of the principal component analysis with explanatory variables, R Square. However, using principal components instead of explanatory variables reduced the dimension from 9 explanatory variables to 3 principal components and broke the co-linearity, hence variance inflation factors found more than one for the first three traits on principal component regression. Hence, using principal components instead of explanatory variables, gained both reductions in the explanatory data set and broke the co-linearity. A comparison of predictions of observations using principal components and actual measurements is shown in Figure 2 for first second and third lactation. Visual inspection of Figure 2 showed that predictions were reasonably accurate for all the traits since most points were lying around a straight line with a slope.

The principal component analysis could retain the following first six components, explaining 98% variation of the original variables. The first principal component showed

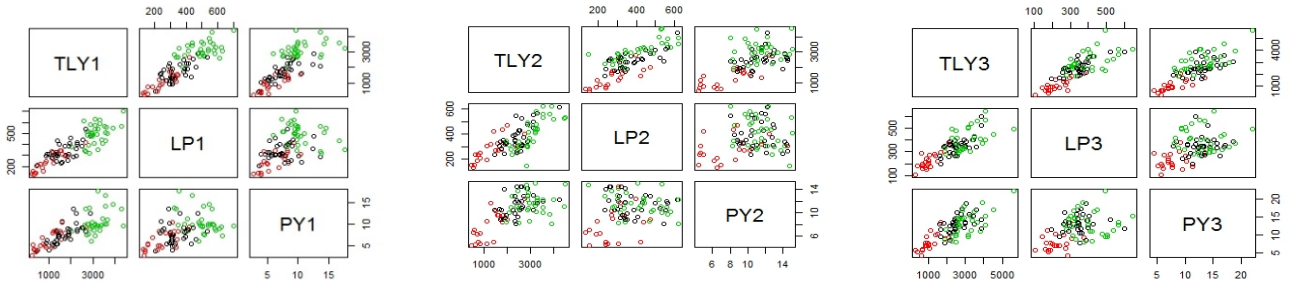


**Fig. 2:** Observations and predictions based on the loadings of the principal components analysis with type traits for first second and third lactation milk yield, lactation length and peak milk yield. Straight middle line as base to check the prediction abilities.

55.8% variation followed by the second 13.6%, third 11.8%, fourth 7.8%, the fifth 5.3% and the sixth principal component 3.8% (Table 2; Fig. 1). Here first three principal components had Eigen values of more than one, but the screen plot showed the appropriateness of the first six components (clearly as curve bend at principal component six). So, the first six principal components were retained to be used as predictors.

The prediction formula was formulated using various principal component combinations (Table 4), and the result revealed that as the number of the principal component increases from 1 to 6 the fitness of the formula also increases. Formula constructed using six principal components having the highest $R^2$ as 0.5985 can be used for predicting lifetime milk production in Gir cattle.

The study was to show the use of principal components regression analysis as principal components are orthogonal contrasts, free from the problem of multi-co-linearity. Since the expression of animal traits on growth, production and reproduction is very complex and correlated, the information generated out of all the traits studied can be included as a null hypothesis as their contribution to lifetime production. The principal components can be used to reduce the data (into components) with the variability explained in the original set of observed traits (variables), as discussed by many workers (Hotelling, 1933; Rugoor *et al.,* 2000 Chapman *et al.,* 2001; Khan *et al.,* 2013).

A 40.32% variation in estimated lifetime yields (total of first 4 lactations- LTMY4) with initial growth, reproduction, and part lactation records with a step-wise procedure of regression analysis in Vrindavani cattle has been reported (Khan *et al.,* 2012). Whereas principal components based on only part lactation records could estimate 54.46% variation in estimation in the same crossbred strain. Bhatacharya and Gandhi (2005) have compared multiple regression analysis and principal components analysis to predict lifetime milk production and found that total variance was lower from the model having PCs as compared to original variables in the regression model. This showed the importance of principal component regression analysis (PCRA) in estimating lifetime production traits.

## CONCLUSIONS

Based on this study, it is concluded that first lactation milk yield, lactation length and peak milk yield, second lactation milk yield, lactation length and peak milk yield contribute 98% of the variance for lifetime production of milk yield. Hence, based on these six records, prediction of an animal's production potential may give a better base of selection at an early age.

## ACKNOWLEDGEMENT

The authors are thankful to the authorities of Cattle Breeding Farm, Junagadh Agricultural University, Junagadh for the permission to utilize milk production records and for this research purpose.

## REFERENCE

Bhatacharya, T.K., & Gandhi, R.S. (2005). Principal components versus multiple regression analysis to predict lifetime production of Karan Fries cattle. *Indian Journal of Animal Sciences*, 75(11), 1317-1320.

Chaudhary, P.N., Kapadiya, P.S., Gadariya, M.R., Gamit, P.M. & Savaliya, B.D. (2022). Test-day and other milk recording options for prediction of lactation milk yield on Gir cows. *Indian Journal of Veterinary Science and Biotechnology*, 18(2), 85-89.

Chapman, K.W., Lawless, H.T., & Boor, K.J. (2001). Quantitative descriptive analysis and principal component analysis for sensory characterization of ultra-pasteurized milk. *Journal of Dairy Science*, 84, 12-20.

Dangar, N., & Vataliya, P. (2017). Genetic, Phenotypic and Environmental Trend for Milk Yield and Production Efficiency Traits in Gir Cattle. *International Journal of Livestock Research*, 7(9), 36-42.

Dangar, N., & Vataliya, P. (2015). Factors affecting lactation milk yield in Gir Cattle. *Indian Veterinary Journal*, 92(7), 71-73.

Hotelling, H. (1933). Analysis of the complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441, 498-502.

Khan, T.A., Tomar, A.K. S., & Dutt, T. (2012). Prediction of lifetime milk production in synthetic crossbred cattle strain Vrindavani of North India. *Indian Journal of Animal Sciences*, 82, 1367-1371.

Khan, T.A., Tomar, A.K.S., Dutt, T., & Bhushan, B. (2013). Principal component regression analysis in lifetime milk yield

prediction of crossbred cattle strain Vrindavani of North India. *Indian Journal of Animal Sciences*, 83, 1288-1291.

Parikh, S.S., Savaliya, B.D., Patbandha, T. K. & Makwana, R.B. (2022). Calf, dam and sire factors affect body weight of Gir calve. *Indian Journal of Veterinary Science and Biotechnology*, 18(2), 49-53.

Rougoor, C.W., Sundaram, R., & Van Arendonk, J.A.M. (2000). The relation between breeding management and 305–day milk production determined via principal components regression and partial least squares. *Livestock Production Science*, 66, 71-83.