Imputation of missing test day milk records and its use in genetic evaluation for milk yield in Mehsana buffaloes

Sujit Saha, Swapnil Gajjar, Nilesh Nayee, A. Sudhakar, G. Kishore and R.O. Gupta

National Dairy Development Board, Anand-388001, Gujarat, India

Abstract

Under the smallholder dairying, the establishment of robust performance recording infrastructure for genetic evaluation programmes is quite challenging. The nationwide lockdown due to COVID-19 pandemic hampered the field performance recording activities to a great extent.The restriction in mobility of the field functionaries leads to the accumulation of considerable volume of missing TestDay Records (TDR). As the number of TDR is important for reliable breeding value (BV) estimates, the accumulation of missing values under any genetic evaluation program would be a concern for the implementers. Under such circumstances, imputation of TDR could provide an effective solution.The present study was carried out to assess the accuracy of the imputation of the missing TDR using Linear and Cubic spline interpolation in Mehsana Buffaloes and the effect of using this imputed TDR for BV estimation.The results of this study indicated a high correlation (0.89) between actual and imputed TDR. The BV estimates and their reliabilities obtained using a combination of actual and imputed TDR were found to be at par with that estimated using 100 per cent actual TDR.There was no impact on the ranking of Mehsana bulls. The study revealed that imputation of missing record through interpolation holds the potential of alternate performance recording system with bi-monthly/quarterly interval without compromising with the accuracy. Besides, it could also facilitate to cover more animals under recording with the limited available fund and create a reference population with a wide genetic base for successful implementation of genomic selection using recorded females.

Keywords: Mehsana buffaloes, missing values, interpolation, breeding values.

Introduction

Low capital investment, short operating cycle, steady returns etc. made dairying a preferred supplementary livelihood option for rural households in India. It has been contributing to the farmers in many ways like regular income from milk and milk products, insurance against drought, emergency cash requirements, household nutrition, fuel for cooking, manure for crops, draught power for farming etc. (Rath, 2019). In India, ownership of bovines is fragmented, with a large number of small and marginal farmers, each raising a few animals (with the average herd size of 5 or below) for draught (animal traction) or dairy purposes. Hence, implementation of long term and scientific genetic evaluation programmes through systematic performance recording is a quite challenging and costly proposition. It is estimated that, to implement a progeny testing programme, based on a young sire model, involving small herds of the dairy farmers, around $ 0.57 million (INR 40 million) per annum would be required to carry out Test Artificial Insemination of about 20 young bulls and other related field activities like Nominated AI, milk recording, measuring type traits, growth monitoring of daughters at six-monthly intervals and also for overall monitoring and supervision activities. It is

also estimated that out of the total budget, around 30-40% of the total fund is being utilized for the test day milk recording and milk component analysis at a monthly interval.

The COVID-19 pandemic, which has emerged as one of the biggest pandemics in and around the globe has a devastating effect not only on health but also the global economy (Pathak, 2020). Due to nationwide lockdown, like in many industries, many routine activities under ongoing field-based genetic improvement programs were also got affected. To prevent the spread of infection, farmers even refused to allow AI technicians or milk recorders to enter their premises. Thus activities like AI delivery, monthly milk recording, milk sampling etc. were either stopped or carried out at the bare minimum level. It resulted into a considerable amount of data loss, till the activities resumed at its normal pace. Similar situation arises even during other unavoidable circumstances like natural calamities.

The possible adverse impact of missing records in Breeding value (BV) estimates of the animals and it's reliability, triggered a thought process to adopt a suitable alternative, which enables estimation of breeding values of the animals having less number of test day (TD) records without compromising the reliability of estimates.

With this background, the present study was designed to assess the efficiency of different interpolation approaches to impute missing TD records (TDR) and impact of using such imputed records in estimating BV for milk yield.

Materials and methods

About the "Mehsana" buffaloes

Mehsana buffaloes are one of the best milk breeds of buffalo in India (Gupta, 1997) and are spread in the northern part of the Gujarat State mainly Mehsana, Banaskantha and Sabarkantha districts (Prajapati *et al.*, 2018).

The name of Mehsana buffalo was derived from the town "Mehsana" in the North Gujarat State. The Mehsana Buffalo breed was evolved by crossing Surti and Murrah buffaloes by the farmers to meet the need for higher milk and also for adaptation to the semi arid climate of Gujarat (Pundir *et al.*, 2000).

Mehsana buffalo is recognized as a persistent milker, regular breeder and most economical to adverse climatic condition. According to the National Bureau of Animal Genetic Resources (NBAGR), the average lactation yield of Mehsana buffaloes is about 1988 kg, with maximum yield reported to be 3597 kg (www.nbagr.res.in).

Brief about genetic improvement program on Mehsana buffaloes

For Genetic improvement of Mehsana Buffaloes, NDDB initiated its first field-level progeny testing project (PT) in the year 1987 in association with The Mehsana District Cooperative Milk Producers' Union Limited, Mehsana, Gujarat under Dairy Herd Improvement Programme Actions (DIPA). Under this project, a robust infrastructure was created for milk recording and genetic evaluation (Trivedi, 1997). Trained milk recorders were engaged to record test day milk yield at the monthly interval and collect milk samples for analysis of various milk components till 2012, the project tested about 231 Mehsana bulls. Since 2012-13 to 2018-19, the PT activities were carried out under the supervision of NDDB in National Dairy Plan Phase–I, a World Bank funded central sector scheme of Govt. of India.

Source of data

The present study was carried out using 119066 test day milk records obtained from 11962 Mehsana buffaloes recorded during September 1988 to September 2019 under Meshana Buffalo PT project implemented by The Mehsana Milk Cooperative Union Limited, Mehsana, Gujarat. The TDR for milk were retrieved from NDDB's INAPH database. For the current study, only those Mehsana buffaloes having 3 or more test day records have been considered. Besides this, observations (with respect to various traits) falling outside the defined physiological range (outliers) were removed. Criteria for accepting records for analysis is given in Table 1.

Imputation of missing records through interpolation

Under this study to mimic the actual field scenarios pertaining to the missing test day records, actual test day records of the buffaloes were masked following four different approaches namely ST-1, ST-2, ST-3, ST-4 as shown in Figure 1. The number of records masked under different strategies is elaborated in Table 2.

The individual animal-wise masked records were imputed using two different interpolation methods namely Linear interpolation and Cubic spline using zoo package (Zeileis *et.al*, 2020) of R software. The degree of association between actual test-day records and imputed test day records was assessed through correlation estimate. The imputation accuracy of linear interpolation and cubic spline interpolation was expressed in terms of Mean Absolute Error (MAE) and Root Mean Sum Square Error (RMSE).

Breeding value estimation

BV along with it's reliability of the recorded she buffaloes and buffalo bulls were estimated using DMU software (Madsen and Jensen, 2014) applying random regression (with legendry polynomials) test day model as mentioned below:

$$y_{thijmnkl} = Village_h + YS_i + Age_j + HYMR_m + Owner_n + \sum_{l=0}^{nf} \phi_{ktl}\beta_l + \sum_{l=0}^{nr} \phi_{ktl}u_{kl} + \sum_{l=0}^{nr} \phi_{ktl}pe_{kl} + e_{thijmnkl}$$

where,

$y_{thijmnkl}$ is the test-day milk yield of animal *k* recorded on day *t* within fixed village subclass *h*, fixed YS (year of calving x season of calving) subclass *i,* fixed Age at calving (6 months grouped) subclass *j,* random HYMR (herd x year of recording x month of recording) subclass m and random owner (at the time of first milk recording) subclass n;

β_l are fixed regression coefficients;

u_{kl} and pe_{kl} are the l^{th} random regression for animal additive genetic and permanent environmental effects, respectively, for animal *k;*

Φ_{ktl} is the l^{th} legendre polynomial for the test day record of cow *k* made on t^{th} day in milk;

nf is the order of polynomials fitted as fixed regressions (0 to 2);

nr is the order of polynomials for *u* and *pe* effects (0 to 2);

$e_{thijmnkl}$ is the random residual effect.

Table 1. Acceptable Physiological range for accepting records for analysis.

SN	Traits	Acceptable Range
1.	Days in Milk	5-330 days
2.	Test Day milk yield	1-40 kg
3.	Age at first calving	20-80 months

Table 2. Strategies for Masking of Test day records for different groups of animal.

Strategy	Animals considered	No. of eligible buffaloes in the subset	Total TD records	No. of TD records masked & subsequently imputed	Remarks
ST-1	Having 10 TD records	2915	29150	11660	Alternate TD records masked (40% of the total TD records)
ST-2	Having 10 TD records	2915	29150	17490	alternate two TD records masked (60% of the total TD records)
ST-3	Having 3 to 10 TD records	6120	54495	29545	54% of the total TD records masked
ST-4	Having 3 to 11 TD records	11916	118252	64321	54% of the total TD records masked

Table 3. Correlation coefficient between true and imputed test day milk records.

	Imputation method	
Strategy	Linear Interpolation	Cubic spline
ST-1	0.899*	0.897*
ST-2	0.890*	0.888*
ST-3	0.885*	0.867*
ST-4	0.883*	0.876*

For analysis of each subset of animals, two types of data files were constructed. One with 100% actual test day record TDR and another with the combination of actual and imputed TDR.

The degree of association of BVs (for both the she buffaloes and breeding bulls, respectively) predicted using the combination of imputed and actual TDR with the BV estimates obtained using 100% actual TDR was studied through Pearson's correlation coefficient (Snedecor and Cochran, 1989). In addition to that ranking of these animals based on BV was also compared using Spearman's rank correlation coefficient (Steel and Torrie, 1960).

The schematic diagram of the design of the experiment is elaborated in Figure 2.

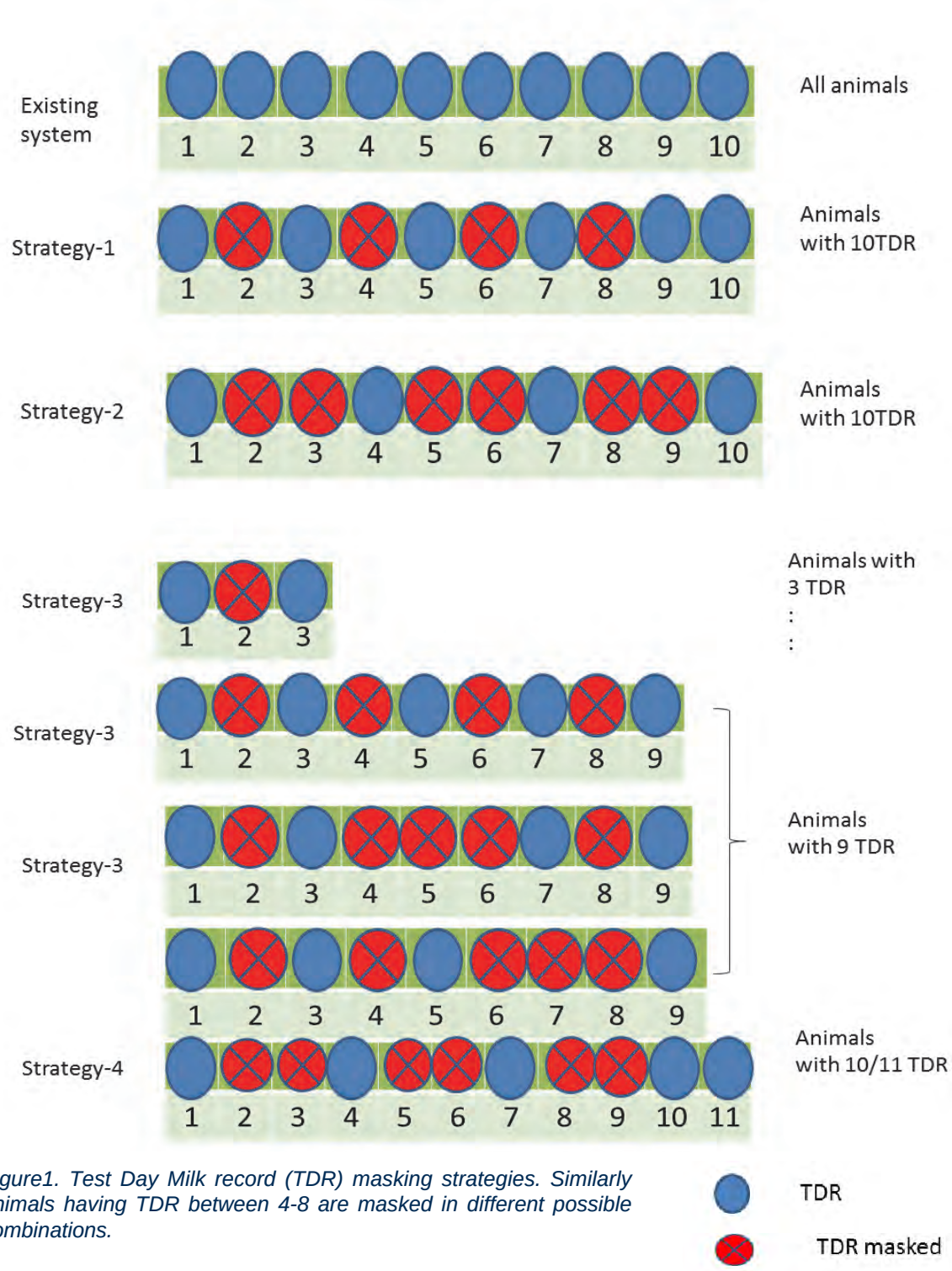

Figure1. Test Day Milk record (TDR) masking strategies. Similarly animals having TDR between 4-8 are masked in different possible combinations.

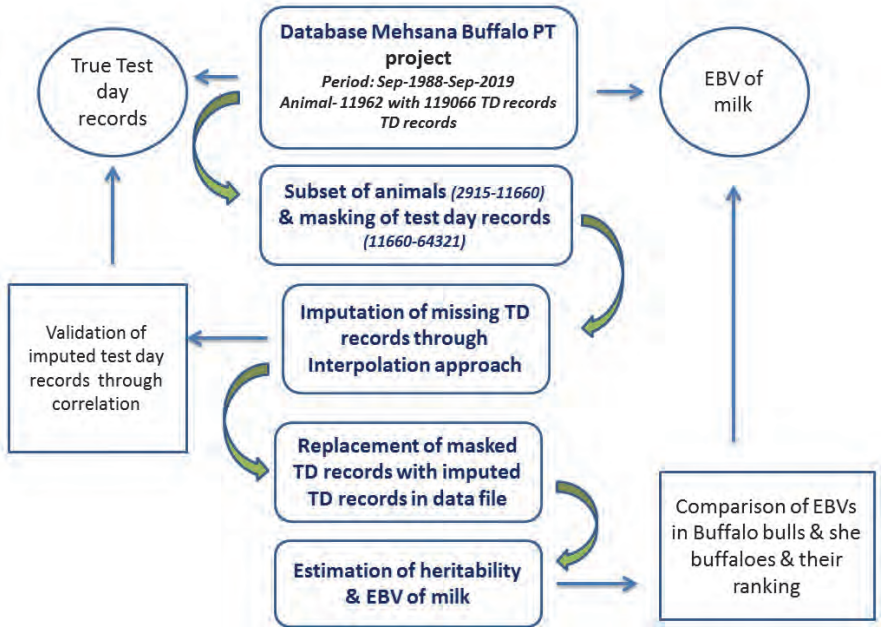

Figure 2. Design of Experiment- a schematic diagram.

Results and discussion

The correlation between imputed TDR (obtained using Liner interpolation and cubic spline method) and the actual test day records are presented to Table 3.

The obtained result indicated a statistically significant and higher degree of association (correlation coefficient >0.86) between imputed TDR with that of actual TDR in Mehsana buffaloes. However, the correlation was observed to decline when the animals having <10 TDR (i.e. 3 to 9 TD records) were masked for imputation.

Comparison of the Linear and Cubic spline interpolation approach expressed in terms of the correlation between actual TDR and imputed TDR and it was found to be marginally higher in linear interpolation than cubic spline interpolation.

The imputation accuracy of Linear interpolation and Cubic spline interpolation expressed in terms of MAE and RMSE was found to be very close (Table 4). However, the MAE and RMSE values were found to be marginally lesser in case of Cubic spline approach.

The BV of the recorded she buffaloes, as well as breeding bulls, was estimated using actual TDR and the combination of imputed and actual TDR, separately. Heritability estimates of test day milk yield obtained under different masking strategies were given below in Table 5.

As presented in Table 5, the heritability estimate of milk yield using 100% actual TDR was found to be 0.26. However, with the use of a combination of imputed missing data along with actual test day records, heritability estimates were observed to reduce marginally (0.24-0.25), which may be due to loss of variance in the test day records due to imputation using interpolation. Heritability estimate thus obtained in this study was found to be slightly higher than the estimate reported by Galsar *et al.* (2016) and Prajapati *et al.* (2018) in Mehsana Buffaloes, whereas, Breda *et al*. (2010) reported TD milk yield heritability between 0.19 -0.31 in Murrah buffaloes.

The correlation coefficient between the BV estimates obtained using 100% actual TDR and a combination of true test day cum imputed TDR are presented in Table 6 (for she buffaloes) and Table 7 (for breeding bulls), respectively.

The results indicated that the replacement of a part of actual TDR with imputed TDR did not have any adverse impact on BV estimation. In case of she buffaloes, the correlation was observed to vary between 0.97 to 0.99 under various strategies for both linear interpolated as well as cube spline interpolated data.

While in the case of breeding bulls, the correlation coefficient for all the cases were observed to be 0.99. All the correlation coefficient estimates were found to be statistically significant. The correlation between the rankings of the animals as estimated by spearman's rank correlation were found to be very high (around 0.98).

Ranking of the top 10 Mehsana breeding bulls based on BV calculated using actual TDR and combination of actual and imputed TDR are presented in Table 8 and Table 9, respectively for linear interpolation as well as spline cube interpolation. The results revealed that there was no marked change in the BV estimates as well as the ranking of the breeding bulls due to incorporation of imputed TDR for data analysis.

Panchal *et al*. (2019), studied the impact of different milk recording strategies on sire evaluation in HF crossbred cattle and found monthly test day milk recording as an optimum strategy for milk recording in genetic evaluation program and concluded that a bimonthly recording can be considered with caution under the conditions wherein deploying monthly test day milk recording is not feasible. The rank correlation between sire breeding values for 305 days milk yield was reported to be 0.79 between monthly milk recording and bimonthly recording strategies, which was lower than the estimates observed in the current study (>0.97). It indicated that in case of presence of the missing TDR, imputation of the missing values with interpolation and its use in breeding value estimation was quite effective in ensuring greater reliability of estimate rather than excluding the same from the analysis.

Conclusion

Based on the results obtained in the present study, it can be concluded that under adverse situations like COVID-19 pandemic or other natural calamities, when regular monthly milk recording activities in the field level are not feasible, imputation of the missing records can be carried out through a suitable interpolation approach, and such imputed data can be effectively utilized in combination of available actual records for genetic evaluation of animals without any adverse impact on breeding value estimation and its reliability.

Under smallholders' dairy production systems, where dairy animals of important breeds are spread across the country under varied environmental conditions, covering a maximum number of animals under performance recording is quite challenging and a costly affair. In such conditions, a performance recording system with the frequency of recording at bi-monthly or quarterly interval holds the potential to include of maximum possible number of animals of a particular breed under milk recording program with the available funds and also to study the influence of Genetic x Environment interaction, In such scenarios, the missing monthly records could be imputed through interpolation and used for genetic evaluation purpose. Further, bringing the maximum number of animals under systematic milk recording program would also enable the creation of a large reference population with a wide genetic base for successful implementation of Genomic selection using recorded females.

Table 4. MAE and RMSE estimate for Linear and Cubic Spline Interpolation.

Interpolation approach	MAE	RMSE
Linear Interpolation	0.713	0.996
Cubic Spline	0.706	0.995

Table 5. Heritability estimates (h^2) of test day milk yield with 100 per cent actual test day record vis a vis combination of actual & imputed test day records.

Strategy	Type of TDR used	h^2	
		Linear Interpolation	Cubic spline
ST-1	Actual + Imputed	0.25	0.25
ST-2		0.25	0.24
ST-3		0.24	0.24
ST-4		0.26	0.25
All animals	100% actual	0.26	

Table 6. Correlation coefficient between EBVs obtained using actual and imputed Test day records in Mehsana buffaloes.

Strategy	Pearson correlation		Rank correlation	
	Linear Interpolation	Cubic spline	Linear Interpolation	Cubic spline
ST-1	0.997*	0.997*	0.996*	0.996*
ST-2	0.993*	0.994*	0.992*	0.993*
ST-3	0.977*	0.971*	0.973*	0.968*
ST-4	0.980*	0.982*	0.977*	0.979*

(*P<0.05)

Table 7. Correlation coefficient between EBVs obtained using actual and imputed Test day records in Mehsana Breeding Bulls.

Strategy	Pearson correlation		Rank correlation	
	Linear Interpolation	Cubic spline	Linear Interpolation	Cubic spline
ST-1	0.999*	0.999*	0.998*	0.999*
ST-2	0.996*	0.997*	0.995*	0.996*
ST-3	0.991*	0.991*	0.989*	0.990*
ST-4	0.982*	0.986*	0.977*	0.982*

(*P<0.05)

Table 8. Breeding value estimates of top ten ranked Mehsana breeding bulls under different strategies using Linear interpolation of missing records.

	Based on Actual Test day records			ST-1			ST-2			ST-3			ST-4		
Bullid	BV	Accuracy	Rank	BV	Accuracy	Rank	BV	Accuracy	Rank	BV	Accuracy	Rank	BV	Accuracy	Rank
Bull-1	467.84	78.03	1	465.18	78.27	1	462.36	78.54	1	472.32	78.55	1	466.83	79.38	1
Bull-2	416.12	85.93	2	406.14	86.15	2	393.36	86.40	2	386.74	86.48	2	389.59	86.98	2
Bull-3	395.71	84.49	3	380.31	84.60	3	374.50	84.73	4	371.52	84.64	4	365.67	85.21	3
Bull-4	383.78	79.52	4	376.85	79.77	4	376.98	80.06	3	373.26	80.11	3	362.71	80.87	4
Bull-5	367.67	88.04	5	371.16	88.13	5	366.32	88.25	5	369.40	88.18	5	348.29	88.62	6
Bull-6	342.92	54.34	6	338.81	55.80	7	329.18	57.22	9	291.56	58.70	10	290.86	60.32	11
Bull-7	342.92	88.02	7	341.52	88.15	6	332.85	88.30	8	332.48	88.30	7	315.34	88.73	9
Bull-8	341.86	81.37	8	334.42	81.68	9	341.26	82.01	6	334.39	82.17	6	353.13	82.85	5
Bull-9	339.45	88.81	9	336.81	88.86	8	341.05	88.93	7	331.31	88.82	8	324.22	89.22	8
Bull-10	333.67	79.68	10	318.14	79.89	10	320.67	80.15	10	324.31	80.16	9	346.26	80.92	7

Table 9. Breeding value estimates of top ten ranked Mehsana breeding bulls under different strategies using Cubic spline interpolation of missing records.

	Based on actual test day records			ST-1			ST-2			ST-3			ST-4		
Bullid	BV	Accurcy.	Rank	BV	Accurcy.	Rank	BV	Accurcy.	Rank	BV	Accurcy.	Rank	BV	Accurcy.	Rank
Bull-1	467.84	78.03	1	463.44	78.21	1	459.88	78.35	1	473.15	78.45	1	479.57	78.36	1
Bull-2	416.12	85.93	2	404.97	86.14	2	392.31	86.30	2	397.17	86.43	2	404.78	86.37	2
Bull-3	395.71	84.49	3	381.06	84.56	3	383.58	84.61	3	376.56	84.61	4	371.73	84.56	3
Bull-4	383.78	79.52	4	375.45	79.73	4	375.77	79.88	4	380.65	80.01	3	369.51	79.93	4
Bull-5	367.67	88.04	5	368.79	88.11	5	362.76	88.17	5	362.78	88.19	5	345.30	88.14	6
Bull-6	342.92	54.34	6	338.26	55.68	8	329.23	56.76	9	287.20	57.95	11	281.13	57.70	14
Bull-7	342.92	88.02	7	340.40	88.13	6	334.37	88.22	8	327.48	88.27	8	311.56	88.22	9
Bull-8	341.86	81.37	8	332.19	81.64	9	338.45	81.86	7	326.23	82.06	9	357.60	81.97	5
Bull-9	339.45	88.81	9	338.33	88.84	7	347.43	88.85	6	340.50	88.82	6	334.10	88.79	8
Bull-10	333.67	79.68	10	319.27	79.85	10	327.42	79.98	10	328.80	80.08	7	344.97	80.00	7

Acknowledgement

The authors acknowledge the contribution of Mehsana PT project officials and Field staff for their sincere efforts in capturing performance data of Mehsala buffaloes in the field.

References

Breda F. C., Albuquerque L. G., Euclydes R. F., Bignardi A. B., Baldi F., Torres R. A., Barbosa L., and Tonhati H. 2010. Estimation of genetic parameters for milk yield in Murrah buffaloes by Bayesian inference . Journal of Dairy Science. TBC.1-8. doi:10.3168/jds.2009-2230

Galsar N. S. , Shah R. R., Gupta J. P. , Pandey D. P. , Prajapati K. B. and Patel J. B. 2016. Analysis of first production and reproduction traits of Mehsana buffaloes maintained at tropical and semi-arid region of Gujarat, India. Life Sciences Leaflets, 77, 65 to 75. Retrieved from http://petsd.org/ojs/index.php/lifesciencesleaflets/article/view/1076

Gupta P. R. (Ed.)1997. Dairy India, Fifth Edition, Statistics, 153-195.

Madsen P., Jensen J., Labouriau R., Cristensen O.F. and Sahana G. 2014. DMU- a package for analyzing multivariate mixed models in quantitative genetics and genomics. In: Proc. 10th World Congress of Genetics Applied to Livestock Produciton, Vancouver, Canada.

Rath D. 2019. Keynote address at 47th Dairy Industry Conference held at Patna during Feb 7-9, 2019: Indian Dairyman.73(3):16-21.

Panchal D., Kakati P., Joshi R. S., Patel A. C., Dholariya S. A., Shah D. B., Patel S. B., Patel C.T., Gajjar S. G., Kishore G. and Rank D. N. 2019.Effect of Alternative Milk Recording Strategies on Genetic Evaluation of Sires of Holstein Friesian Crossbred Cattle. International Journal of Livestock Research.9(8):203-213.

Pathak K. M. L. 2020. Impact of COVID-19 on health of dairy animals. Indian Dairyman.72(5):50-52.

Prajapati B M, Gupta J P, Chaudhari J D, Parmar G A, Panchasara H H, Chauhan H D, Ankuya K J and Prajapati M N. 2018. First lactation production performance of Mehsana buffaloes under field progeny testing programme in semi-arid region of Gujarat. Indian J of Dairy Science. 71(4): 404-408.

Pundir R., Sahana G., Navani N., Jain P., Singh D., Kumar S. and Dave A. 2000. Characterization of Mehsana Buffaloes in India. Animal Genetic Resources Information. 28: 53-62. Doi:10.1017/S101423390000136X.

Steel R.G.D andTorrie J.H. 1960.Principles and Procedure of Statistics with Special Reference to the Biological Sciences. McGrew Hill Book Company Inc. New York. 550

Trivedi K.R. 1997. A case study of Buffalo recording systems under the Dairy Cooperative organizations in India. International Workshop on Animal Recording for smallholders in developing countries. Anand, India 20-23 October,1997.

Snedecor G. W. and Cochran W. G. 1989. Statistical Methods. Eighth Edition.Iowa State University Press, Ames, Iwoa 50010.

Zeileis A., Grothendieck G., Ryan J. A., Ulrich J. M., Andrews F.2020. Package 'Zoo'.version 1.8-8. (URL: http://zoo.R-Forge.R-project.org).