Indian Journal of Animal Sciences **90** (8): 1186–1190, August 2020/Article

Artificial insemination for milk production in India: A statistical insight

AMIT SAHA[1] and SANGEETA BHATTACHARYYA[2]✉

Central Silk Board, Ministry of Textiles, Government of India

Received: 19 November 2019; Accepted: 1 January 2020

ABSTRACT

Though India is a global leader in milk production, on the flip side, about 80% cattle belonging to indigenous and non-descript breeds are low yielders whose productivity needs to be improved by adopting appropriate breeding techniques and Artificial Insemination (AI) comes to this rescue. AI plays a vital role in improving the productivity of bovines by upgrading their genetic potential thereby enhancing the milk production and productivity in the country. Though milk production is influenced by a number of factors, the authors analyzed one of the revolutionary innovations in Indian dairy sector, the artificial insemination (AI) in bovines which was introduced in India in 1951–56. Hence a statistical approach to inspect the influence of artificial insemination as a factor behind the growth in milk production in India was undertaken. In this study, Linear Regression (LR) and Support Vector Regression (SVR) were utilized. LR was used to establish the linear relationship between variables and determine the role of AI in that relation. SVR is an eminent machine learning technique which works on the structural risk minimization principle to minimize the generalization error which leads to better prediction accuracy, whereas LR provides the value by which one can estimate that to what extent AI can show its' impact on milk yield. Empirical results noticeably reveal the positive impact of AI on milk yield using LR and better prediction accuracy of SVR as compared to LR for both in training and testing dataset.

Keywords: Artificial insemination, Linear regression, Milk production, Support vector regression

India has a livestock population of 535.78 million as per the 20th Livestock Census Report of 2019. It is a 4.6% increase over the population calculated in the 19th Livestock Census of 2012 (PIB, 2019a). The quinquennial review of the country's animal husbandry and dairying scenario is a truckload of statistics which speak volumes of the actual growth of the booming dairy sector in India. The total bovine population (Cattle, Buffalo, Mithun and Yak) registered an increase of 1% over the previous census. The total milch animals (in-milk and dry) in cows and buffaloes are 125.34 million which is an increase of 6.0% over the previous census (PIB 2019). As the bulk of milk produced in India is from the bovine population, hence the spurt in population when compared to the increase in estimated milk production statistics is quite less. The estimated milk production of India in 2013–14 was 137.7 million tonnes whereas it is 188.1 million tonnes in 2018–19 which is an increase of 36.6% (PIB 2019b). In 2017–18 it was 176.3 million tonnes which meant a per capita availability of 375 grams per day (NDDB 2019). So an intrusion into more statistical data and field surveys brought into the limelight that the number of AIs performed across India over the years has also shot up. Government of India had introduced AI officially into farms in the first five-year plan (1951–56) through 150 key village centers to improve the genetic quality of cattle and buffaloes in this country (Buffalopedia 2019). As per available data from only Government Veterinary Hospitals or AI centers, it has been seen that 2013–14 registered 58,839,000 AIs performed while in 2017–18 it was 70,062,000 which is an increase of 24.19% (NDDB 2019). Not only in India but Sapkota *et al.* (2016) had reported that AI had led to increase in milk yield in livestock of Nepal. It was quite a matter of inquisition as to whether AI had a role in influencing the milk production in India and if yes then as to what extent. Another important thing was to predict the future production of milk based on the annual number of artificial inseminations performed, so that, government can take more steps to increase the number of artificial insemination to maintain the increasing trend of the milk production even if the number of bovine reduces in near future as seen in US where in 1942 the cattle population was more (25 million) but the milk production was less (118 billion pounds) as compared to 2007 when the cattle population decreased (9.1 million) but the milk production (185 billion pounds) registered high growth due to AI (Memon 2018).

In our study, LR and SVR have been used to fit the model and predict the milk production on the basis of artificial insemination. SVM is one of the most important supervised

Present address: [1]Central Silk Board, Ministry of Textiles, Government of India. [2]ICAR-Central Citrus Research Institute, Nagpur, Maharashtra. ✉Corresponding author email: sangeeta.bhattacharyya2012@gmail.com.

machine learning technique which is a part of artificial intelligence. Machine learning is a technique which allows the machine/computer to learn by itself. Cortex and Vapnik (1995) developed SVM technique for problems of classification which was based on Vapnik-Chervanenkis theory. Vapnik *et al.* (1997) successfully extended SVM to regression problems, and it is called as SVR. A good review on SVR has been discussed in many research papers (Gunn, 1998, Dibike *et al.* 2001, Kecman 2001, Basak, *et al.* 2007, Cherkassky and Ma, 2004; Yu *et al.* 2006, Raghavendra and Deka, 2014). SVR has been used in water demand prediction (Msiza *et al.* 2008), on-line health monitoring (Zhang *et al.* 2008), river stage prediction (Wu *et al.* 2008), response modeling (Kim *et al.* 2008), long-term monthly flow discharge (Lin *et al.* 2006) and tourism demand forecasting (Chen and Wang 2007).

The significant thing in SVR is the selection of kernel function which plays an important role on the performance of the SVR model. So, proper care should be taken during the selection of kernel function. The power of prediction mostly depends on the proper selection of this function. There are various types of kernel function, viz., Linear SVM, Polynomial SVM, Radial Basis function (RBF) and Multi-Layer Perceptron (MLP).

Hsieh *et al.* (2011) applied Least square support vector machine (LS-SVM) to calibrate the prediction model for adulteration ratio. It was found out that the adulteration ratio above 10% clearly differs from 0% samples. Alonso *et al.* (2013) revealed that it can be possible to predict carcass weights 150 days before the slaughter day using SVR. Mammadova and Keskin (2013) used SVM technique to detect mastitis detection. SVM showed its classification ability to ascertain the presence of subclinical and clinical mastitis in dairy cows (Holstein cows). Shine *et al.* (2019) applied SVM in predicting and analysing the consumption of annual dairy farm electricity to improve the sustainability of the projected expansion of milk production in Ireland.

So the authors have applied the SVR model to a longitudinal data base of milk production and AI in India and compared the prediction accuracy with LR model. The in sample forecast was done and matched with the available data on record. Also a case of out sample forecast was performed to forecast milk production for two years when only data on AI is available. On the other hand,LR has provided the value by which one can estimate that to what extent AI can show its' impact on milk yield.

MATERIALS AND METHODS

Data description: Time series data on Milk Production ('000 tonnes) and Artificial Inseminations Performed ('000 Nos.) of India from 2001 to 2017 were taken from the website of National Dairy Development Board. The data from 2001 to 2015 have been utilized for model building purpose and the data of Artificial Inseminations Performed ('000 Nos.)in 2016 and 2017 are used to predict the milk production for the validation purpose.

Linear regression: Linear regression method estimates the relationship between the response variable and explanatory variable. After establishing the relationship, one can predict the response based on the value of explanatory variable. Linear regression method is known as simple linear regression if there is a single explanatory variable, otherwise it is multiple regression in case of two or more number of explanatory variables.

In notation Simple linear regression model is:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \ldots (1)$$

where, X, Explanatory variable; Y, Response variable. β_0 and β_1 are the parameters to be estimated and ε is the error term.

Ordinary least square method (OLS) is the most commonly used regression method to estimate the parameters which is based on the minimization of square of residuals of the above model. The solution will be

$$\hat{\beta} = (X'X)^{-1}(X'Y) \quad \ldots (2)$$

After determination of explicit form of regression equation by OLS method, the aim is to forecast the response variable for a given value of explanatory variable.

Support vector regression (SVR): Cortex and Vapnik (1995) developed SVM technique for problems of classification which was based on Vapnik-Chervanenkis theory. SVM is not only popular for the classification but also for its modelling and prediction performance. A tremendous advantage of SVM is that it is not model dependent as well as independent of linearity. The training of the data driven prediction process SVM is done by a function which is estimated utilizing the observed data.

The prediction function for linear regression is defined as:

$$f(y) = (w.y) + c \quad \ldots (3)$$

whereas, for non linear regression, it will be:

$$f(y) = (w.\varnothing(y)) + c \quad \ldots (4)$$

where, w dentoes the weights, c, represents threshold value, $\varnothing(y)$ is known as kernel function.

If the observed data is linear, then equation (3) will be used. But, for non-linear data, the mapping of y(t) is done to the higher dimension 'feature' space through some function which is denoted as and eventually it is transformed into the linear process. Afer that, a linear regression will carry out in that feature space.

The significance of kernel function in non-linear support vector machine (NLSVR) is very much important for mapping the data into higher dimension feature space in which the data becomes linear. Generally notation for kernel function is given as;

$$k(y,y') = \langle \varnothing(y), \varnothing(y') \rangle$$

Kernel function are used for the transformation of the given data into the required form. Kernel function is actually a mathematical function. RBF is mostly used kernel function.

Radial Basis function

$$k(x,y) = \exp\left(\frac{\|x-y\|^2}{2\sigma^2}\right)$$

or $$k(x, y) = \exp(-\alpha\|x-y\|^2) \quad \ldots (5)$$

where, shape of hyperplane is controlled by.

A detailed description of SVR has been discussed by Vapnik *et al*. (1997).

RESULTS AND DISCUSSION

Linear regression: Linear regression method has been used to establish the relationship between the response and explanatory variable. In our study, response variable and explanatory variable are the Milk Production ('000 tonnes) and explanatory variable is Artificial Inseminations Performed ('000 Nos.) respectively. Parameter estimation through OLS method is depicted in Table 1.

Table 1. Parameter estimation using OLS for LR

Coefficients	Estimate	Standard Error	t-value	P-Value
Intercept (β_0)	54,025.923	2,314.554	23.34	<0.0001
X (β_1)	1.431	0.052	27.53	<0.0001

Parameter estimation of SVR: Parameter selection is the most important part to obtain the better model. So, best parameters have to be selected to improve the performance of the model.

Grid search method has been used which is the standard way to search the parameters in order to get best model by training various models with different combination of cost and epsilon and next step is to select the best one after comparing their performance in terms of root mean square error (RMSE) which is the indicator to measure the performance.

The analysis has been done using "e1701" package (David, 2017) in R software. This process of finding and choosing these parameters by grid search method is known as hyper parameter optimization. R software has been used to analyze the support vector regression. The package 'e1071' (David 2017) has been utilized for this study. The performances of 1100 trained models are presented in Fig. 1 which also depicts the value of RMSE in the right side.

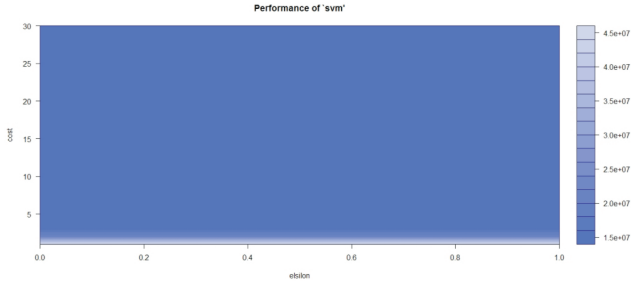

Fig. 1. Plot to find out the best parameters of the SVR model

Table 2. Parameter estimation of SVR

Sampling method	10-fold cross validation
Epsilon (best parameter)	0.1
Cost (best parameter)	9.0
Gamma (best parameter)	1.0
Number of support vectors	5.0
SVM-type	eps-regression
SVM-kernel	Radial

Table 2 shows the estimated best parameters of SVR after sufficient tuning of SVR model and these best parameters have been utilized to build the SVR model. It has been seen that the best SVM-kernel function is Radial basis function for SVR.

In Fig. 2, Blue line represents the original data whereas red and black line denote the fitted line of Linear Regression and support vector regression respectively. After the perpendicular line, the three lines show the forecasted (out of sample) value for the year 2016 and 2017 for the purpose of validation of the models. It has been seen that MAPE of LR and SVR model are 8.29 and 5.18 respectively. So, the result shows the better generalization ability of SVR over the LR model.

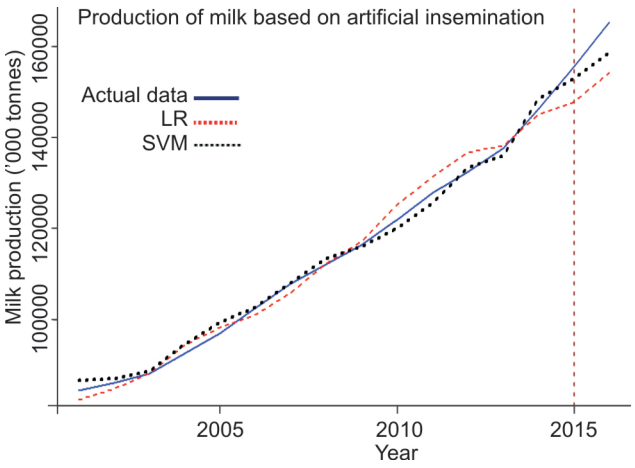

Fig. 2. Graphical representation of the performance of the models.

The Mean square error (MSE), Mean absolute error (MAE) and Mean absolute percentage error (MAPE) values of LR and SVR are given in Table 4.

The intercept and slope parameters of LR model are significant at 1% level of significance (Table 1). Further, the model is also significant at 1% level of significance as the P-value of model is <0.0001. The value of is 0.98 which tells that 98% of variability of milk production is explained by the Artificial Inseminations Performed. The value of β_1 is 1.431 which reveals that one unit increment in AI tends to 1.431 unit increment in annual mil production. Fig. 2 depicts the performance of 1,100 trained SVR models where the aim was to find out the best model based on minimum RMSE. Various kernels have been tried but RBF kernel had provided the best result. The best values of Epsilon, cost and gamma

Table 3. Actual and predicted value using LR and SVR

Year	In sample forecast			
	Actual production ('000 tonnes)	Artificial inseminations performed ('000 nos.)	Linear regression (LR)	Support vector regression (SVR)
2001	84,406	19,766	82,319.94	86,663.38
2002	86,159	21,519	84,829.27	86,936.56
2003	88,082	23,835	88,144.50	88,511.64
2004	92,484	28,225	94,428.56	94,512.51
2005	97,066	30,940	98,314.94	99,322.52
2006	102,580	32,868	101,074.80	102,813.57
2007	107,934	36,205	105,851.50	108,360.40
2008	112,183	40,706	112,294.50	113,924.41
2009	116,425	44,037	117,062.60	116,582.35
2010	121,904	49,821	125,342.10	120,486.39
2011	127,904	54,063	131,414.30	125,649.01
2012	132,431	57,787	136,745.10	133,213.96
2013	137,685	58,839	138,250.90	135,823.40
2014	146,314	63,619	145,093.30	148,585.41
2015	155,491	65,567	147,881.70	153,223.86
	Out of sample forecast			
2016	165,404	70,062	154,284.60	158738.90
2017	176,347	73,369	158,942.70	165165.00

are 0.1, 8, and 1 respectively (Table 2). After that, SVR model was fitted with these best values of parameters. It is observed from Table 3 that overall performance of SVR is better than the LR model in case of in sample forecast. In out of sample forecast, the forecasted value of SVR is much closer to the actual data compared to the LR model. Graphically, the superiority of SVR over the LR has been observed from the Fig. 2 as the forecasted values of SVR are closer to the actual values than the forecasted values of LR. It can also observe from Table 4 that the MSE, MAE and MAPE of SVR models are smaller than the LR model. From the above discussions, it can be said that the SVR has provided better fitting model over the LR.

Table 4. Model performance

Method	MSE	MAE	MAPE
Linear regression	8,063,486	2,111.13	1.76
Support vector regression	2,657,182	1,410.93	1.25

The aim of this study was to establish a relationship between AI and milk production and to know how AI data can help in forecasting of milk production of the country and thus help the nation to achieve its target of not only food and nutritional security but also livelihood security.

LR and SVR have been utilized in this study. The results have revealed that SVR model can perform better than LR model in terms of forecasting accuracy. It can be seen that the prediction for 2016 and 2017 using SVR is closer to the actual milk production compared to the LR based on the value of artificial insemination. As SVR deals with the simultaneous minimization of both empirical risk and the confidence interval, hence minimum error has been achieved. However, as a machine learning technique, SVR cannot able to expose that up to what extent AI influences milk production which was answered by LR. Finally, it can be concluded that AI has a positive impact on increasing milk yield in India and prediction can be done using SVR. So, this study is not only to show the better forecasting performance of SVR over the LR but also to estimate the role of AI in influencing the milk production in India which can't be found out by SVR. In future, this study can also be extended by using other machine learning techniques.

REFERENCES

Alonso A, Rodríguez A and Antonio C B. 2013. Support vector regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and Electronics in Agriculture* **91**: 116–20.

Basak D, Pal S and Patranabis D. 2007. Support vector regression. *Neural Information Processing – Letters and Reviews* 11.

Buffalopedia. 2019. Indian Council of Agricultural Research, Department of Agricultural research and Education, Ministry of Agriculture and Farmers' Welfare, Government of India.

Chen K Y and Wang C H. 2007. Support vector regression with genetic algorithm in forecasting tourism demand. *Tourism Management* **28**: 215–26.

Cherkassky V and Ma Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* **17**(1): 113–26.

Cortes C and Vapnik V. 1995. Support-vector network. *Machine Learning* **20**: 1–25.

David M. 2017. E1071: Misc Functions of the Department of Statistics. *Probability Theory Group R package version* **1**: 6–8.

Dibike Y B, Velickov S, Solomatine D and Abbott M B. 2001. Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering* **15**(3): 208–16.

Department of Animal Husbandry and Dairying. 2019. Ministry of Fisheries, Animal Husbandry and Dairying. Government of India. http: //dahd.nic.in/about-us/divisions/cattle-and-dairy-development.

Gunn S R. 1998. Support vector machines for classification and regression. *Image, Speech and Intelligent Systems Tech. Rep.* University of Southampton, UK.

Hsieh C L, Hung C Y and Kuo C Y. 2011. Quantization of Adulteration Ratio of Raw Cow Milk by Least Squares Support Vector Machines (LS-SVM) and Visible/Near Infrared Spectroscopy. Iliadis L and Jayne C (eds). Engineering Applications of Neural Networks. EANN 2011, AIAI 2011. IFIP Advances in Information and Communication Technology, 363. Springer, Berlin, Heidelberg.

Kecman V. 2001. Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press, Cambridge, Massachusetts.

Kim D, Lee H J and Cho S. 2008. Response modeling with support vector regression. *Expert Systems with Applications* **34**(2): 1102–08.

Lin J Y, Cheng C T and Chau K W. 2006. Using support vector machines for long-term discharge prediction. *Hydrological Sciences–Journal* **51**(4): 599–611.

Memon, M. 2018. Could Artificial Insemination be the Solution to Increase Milk Production and Ensure Food Security? Agrilinks. July 10, 2018. https: //www.agrilinks.org/post/could-artificial-insemination-be-solution-increase-milk-production-and-ensure-food-security.

Msiza I S, Nelwamondo F V and Marwala T. 2008. Water demand prediction using artificial neural networks and support vector regression. *Journal of Computers* **3**(11): 1–8.

National Dairy Development Board. 2019a. Milk Production in India, https: //www.nddb.coop/information/stats/milkprodindia.

National Dairy Development Board. 2019b. Artificial Insemination Performed by States, https: //www.nddb.coop/information/stats/insemination.

Nazire M and Ismail K. 2013. Application of the Support Vector Machine to Predict Subclinical Mastitis in Dairy Cattle. *Scientific World Journal*.

Press Information Bureau. 2019a. Department of Animal Husbandry and Dairying releases 20th Livestock Census, Total Livestock population increases 4.6% over Census–2012, Increases to 535.78 million, Ministry of Fisheries, Animal Husbandry and Dairying, Government of India. New Delhi, India. https: //pib.gov.in/PressReleasePage.aspx?PRID=1588304

Press Information Bureau. 2019b. Growth of Dairy Sector, Ministry of Fisheries, Animal Husbandry and Dairying, Government of India, New Delhi, India. https: //pib.gov.in/newsite/PrintRelease.aspx?relid=191790.

Raghavendra N S and Deka P C. 2014. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing* **19**: 372–86.

Sapkota S, Gairhe S, Kolakshyapati M and Upadhyay N. 2016. Adoption of artificial Insemination technology in dairy animals and impact on milk production: a case study in Nawalparasi and Chitwan districts of Nepal. *Nepalese Journal of Agricultural Sciences* **14**: 79–84.

Shinea P, Scullyb T, Uptonc J and Murphy M D. 2019. Annual electricity consumption prediction and future expansion analysis on dairy farms using a support vector machine. *Applied Energy* **250**: 1110–19.

Vapnik V, Golowich S and Smola A. 1997. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems. Mozer M, Jordan M and Petsche T (Eds). MIT Press Cambridge, MA **9**: 281–87.

Valergakis G, Arsenos G and Banos G. 2007. Comparison of artificial insemination and natural service cost effectiveness in dairy cattle. *Animal* **1**: 293–300.

Wu C L, Chau K W and Li Y S. 2008. River stage prediction based on a distributed support vector regression. *Journal of Hydrology* **358**(1–2): 96–111.

Yu P S, Chen S T and Chang I F. 2006. Support vector regression for real-time floodstage forecasting. *Journal of Hydrology* **328**: 704–16.

Zhang J, Sato T, Iai S and Hutchinson H. 2008. A pattern recognition technique for structural identification using observed vibration signals: Nonlinear case studies. *Engineering Structures* **30**: 1417–23.