Accounting for heterogeneity of data in implementing genomic selection models applicable to crossbred cattle in India

Ana Guintard, Laurianne Bourguignon, Gwendal Restoux, Vincent Ducrocq

Accounting for heterogeneity of data in implementing genomic selection models applicable to crossbred cattle in India

A. Guintard[1*], L. Bourguignon[2], G. Restoux[1], V. Ducrocq[1]

[1]Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France;
[2] Institut Agro (Agrocampus Ouest), 65 rue de Saint Brieuc, 35 042 Rennes, France ;
*ana.guintard@inrae.fr

Abstract

Indian dairy production takes place in very diverse environments. The objective of this study was to identify genotype by environment interactions in the context of establishment of a reference population for genomic selection. Data from 9,265 cows in six states were analysed. The high variability in the distribution of breed types and milk yield across India, as well as the wide range of climatic zones existing in the country are highlighted. This heterogeneity must be taken into account, defining, for example, groups of districts (administrative regions within states), rather than considering India as one homogeneous country, in order to better predict the genomic merit of individuals in their local environment.

Introduction

The dairy sector plays a major role in India, contributing 5.3% of the gross domestic product, and representing an essential source of animal protein in a country where one-third of the population is vegetarian (Belhekar and Dash, 2016). Improving productivity at the animal level is necessary to meet the demand and to ensure self-sufficiency (Chawla, 2009). Genomic selection is a way to increase the milk production in Indian smallholder farming systems (Al Kalaldeh et al., 2021). To implement it, BAIF Development Research Foundation (https://baif.org.in/) has been collecting performances and genotypes for the past four years to set up a reference population. However, the heterogeneity of collected data reflects a huge diversity of production systems over India (across states, climatic zones, herd size, breeds, etc.). Thus, the aim of this study was to analyse phenotypic data from BAIF's database in order to characterise the heterogeneity of performances to account for in the establishment of a reference population for genomic selection.

Material and Methods

Data used for this study originates from the BAIF database and encompass phenotypic data recorded in the states of Andhra Pradesh, Bihar, Gujarat, Jharkhand, Maharashtra, Odisha, Punjab, Rajasthan and Uttar Pradesh between August 2016 and November 2021. Animals were crossbred cattle ranging from pure indigenous *Bos indicus* to pure exotic *Bos taurus* (Holstein or Jersey) cows. The database, contains information on daily milk yield, day in milk, lactation number, last calving date, breed type, the type of recording (by an enumerator or by the farmer himself) the farm, the CDC (Cattle Development Centre, which is a local centre providing services such as artificial insemination for eight to ten villages), the district and the state.

In order to make the dataset suitable for analysis, we filtered the data as follows: (1) Andhra Pradesh and Gujarat states were removed because data were available only on pure zebus

(breeding tracks), and were not representative of the diversity of breeds in these states, (2) data from cows with fewer than four records per lactation and with a lactation lasting less than 200 days were removed, (3) data were retained only if they were recorded between eight and 340 days after calving, (4) data from cows with a coefficient of variation in daily milk yield greater than 0.8 on a given lactation or a daily milk yield greater than 25kg were removed, being more than three times greater than the average daily milk yield (Al Kalaldeh et al., 2021), (5) data from cows with a lactation rank greater than seven were deleted, and (6) for each CDC and breed, data were deleted if they were recorded on fewer than ten cows.

Weather data (from worldweatheronline.com), temperature, humidity and precipitation were also considered at different levels of season which were defined as: winter (January to February), pre-monsoon (March to May), monsoon (June to September) and post-monsoon (October to December).

After data filtering and integration of weather data (not available for all the districts), 211,545 records from 9,265 cows in six states (Bihar, Jharkhand, Maharashtra, Odisha, Punjab and Uttar Pradesh), 26 districts, 75 CDC and 4,168 farms were retained for the analysis. The number of cows per herd ranged from one to 47 with an average of 2.2. Overall, 44% of the herds were composed of a single cow while 85% of the herds had three cows or less. The number of cows per CDC varied from ten to 456.

The daily milk yield was transformed into a 305-day production and corrected, using a linear mixed model including the effects of breed type, calving season, lactation rank and type of recording, all treated as fixed.

A recursive partitioning classification using decision trees allowing for clustering of individuals in homogenous sub-populations according to explanatory variables (Breiman et al., 1984), was performed on corrected milk production to identify regions and subsequent characteristics with contrasted production levels.

Statistical analyses were done using RStudio (version 3.5.0) and the ggplot2 package. Recursive partitioning classification was performed using the rpart package.

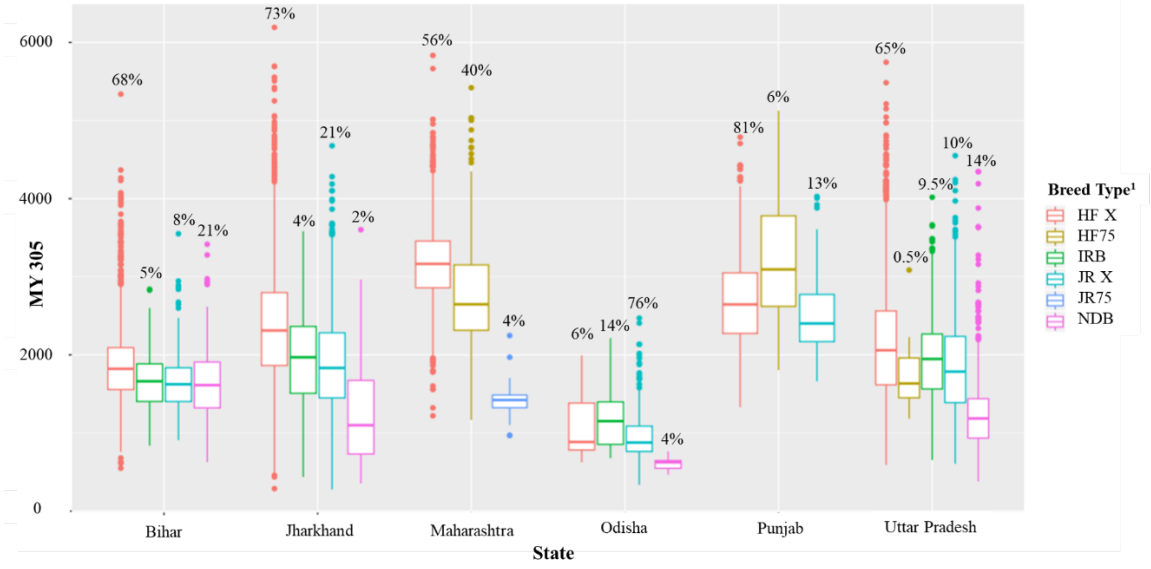
Results and discussion

Breed distribution.

To understand the variability of milk yield, we first looked at the distribution of uncorrected milk production according to breed types between and within states (Figure 1). The percentage of each breed between states varies a lot, ranging from places where indigenous zebus and non-descriptive cows are an important part of the herd (Bihar, Odisha and Uttar Pradesh), to states that orientated their breeding towards high levels of exotic breeds (Maharashtra, Jharkhand).

Holstein crosses usually show the highest milk production in most of the states. However, we can identify some scaling effect (a significantly higher production in Maharashtra than in other states) and re-ranking effect (in Odisha, Holstein crosses have, on average, a lower milk yield than indigenous breeds).

Figure 1 shows a high variability in the distribution of breed types across the six states. Corrected milk yield was significantly variable among districts (ANOVA, $F=170.5$, $df=25$, $P<0.05$)



[1] HF 75: Crosses with more than 75% of Holstein origin, HF X: Crosses with less than 75% of Holstein origin, JR75: Crosses with more than 75% of Jersey origin, JR X: Crosses with less than 75% of Jersey origin, IRB: Indian Recognised Breed, NDB: Non-Descriptive Breed.

Figure 1: Milk Yield production (MY 305) and percentage of each breed type within each state.

Milk production under diverse climatic conditions.

The recursive portioning classification separated districts into four climatic clusters (figure 2), with the following predicted average corrected milk productions: 2,999 kg cluster A: Ahmednagar, sd= 649.4 kg, cv=0.2), 2,750 kg (cluster B: Faizabad, Hazaribag, Khunti, Lohardaga, Pune, Ranchi, Ropar and Saharanpur, sd=767.3 kg, cv=0.3), 2,459 kg (cluster C: Ara, Bahraich, Chhapra, Gonda, Hardoi, Hoshiarpur, Kannauj and Meerut, sd=759.4 kg, cv=0.4), and 2,237 kg (cluster D: Allahabad, Bargarh, Bolangir, Buxar, Etawah, Lucknow, Sambalpur, Unnao and Varanasi, sd=701.8 kg, cv=0.4).

Maharashtra shows the highest corrected milk production, especially the district of Ahmednagar (cluster A). This district, situated in a rain shadow region, has a tough climate, with very low rainfall, even during monsoon. The higher average milk production in Maharashtra, is more likely due to a longer technical breeding support of BAIF to local farmers in this state, where BAIF influence started more than 30 years ago. The higher milk production results from years of breeding with exotic breeds, and particularly Holstein cattle. This illustrates that genotype by environment interactions do not only depend on the state, and that beyond climatic conditions, other factors greatly impact production. Particular attention should be paid to the farming system, and further analyses of agro-economic data (major crops, size of the farm, family income, economic status and feeding system of the cows) should help to identify significant determinants of milk yield in India.

Figure 2 also shows that homogeneous groups of weather are not always located within the same state. Indeed, all dots of the same cluster share similar climatic conditions, and we can see that districts of the same cluster can be found in different locations. Mountains and rivers (especially the Ganges river) play an important role in the local weather, resulting in various sub-groups of climates, influencing crops and feedstuff in the area. Considering these regional

variations by classifying districts into homogenous climatic zones could help to account for the wide heterogeneity of weathers across India in a genomic evaluation.

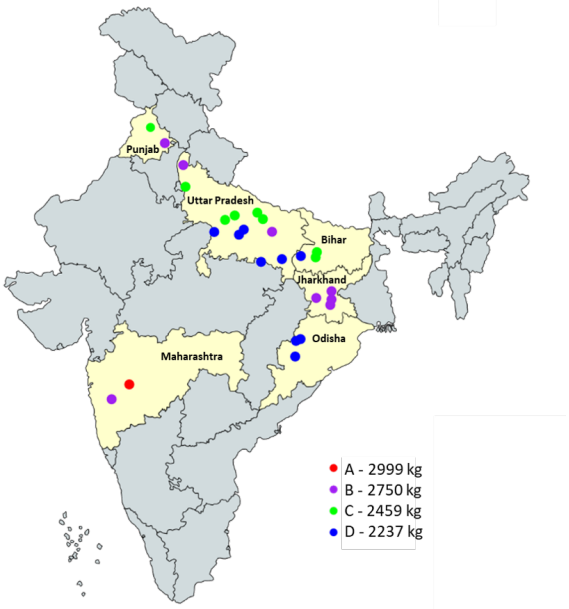


Figure 2 Map of Districts partioned by corrected milk production using weather information, each dot corresponding to a homogenous climatic area.

Conclusion

India, given its size and its particular milk production, mostly ensured by small and marginal farmers, shows extreme variability in terms of production, breeds and climates, resulting in strong genotype by environment interactions. It, thus, appears crucial to account for them, along with agro-economic characteristics, when establishing a reference population in order to get an appropriate estimation of the genomic breeding value of individuals. In particular, defining a few groups of homogeneous environments across India, rather than considering it as a uniform country, could help to conduct more accurate genomic evaluation in the field.

References

Al Kalaldeh M., Swaminathan M., Gaundare Y., Joshi J., Aliloo1 H., et al. (2021) Genet. Sel. Evol. 53:73 https://doi.org/10.1186/s12711-021-00667-6.

Belhekar S., and Dash S. (2016) Paripex - Indian Journal of Research 5(11): 509 – 510 https://doi.org/10.36106/paripex.

Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. (1987) Cytometry, 8(5): 534-535 https://doi.org/10.1002/cyto.990080516.

Chawla A. (2009) Milk and Dairy Products in India – Production, Consumption and Exports. Available at: https://www.hindustanstudies.com/files/dairysept09report.pdf.