Monday, March 11, 2024

Abstract Submission Form

Title (Mr./Mrs/Dr./Prof.) Mr

Presenting author Ilia Rukin

Institute

Institute/company: JSC Agroplem

Adress: 42 Bolshoy boulevard, office 1466, Skolkovo Innovation center

ZIP/Postal code: 143026

City: Moscow

Country: Россия

Insert all authors and institutions

Rukin I.V. (1), Sudarkina S.I. (1), Krutkina M.S. (1), , Kamaldinov E.V. (2)
(1) JSC «Agroplem», Moscow, Russian Federation
(2) Novosibirsk State Agrarian University, Novosibirsk, Russian Federation

Preferred presentation Oral Poster

Preferred session Session 6: SC Dairy Cattle Milk Recording – Presentation and evaluation of new analytical parameters in herd management for dairy farms

Email of corresponding author irukin@agroplem.ru

Title of your paper Application of machine learning methods to control the milk samples analysis results reliability

Insert ABSTRACT text

Modern animal breeding methods, such as genomic evaluation of breeding values (GEBV), are based on large amounts of phenotypic and genetic data. The reliability of the GEBV results and general selection process depends on the reliability of the primary phenotypic data. Phenotypic data, particularly dairy cattle control milk data are derived from on-farm control milkings and may be incorrect. A decrease in the raw data reliability may occur for several reasons, including violation of the milk TD samples storage and transportation conditions, problems with sampling, as well as intentional data falsification. As a result of errors in the sampling process and intentional data falsification, TD milk samples may actually be dispensed from a sampler or milk tank. We will further call such milk samples dispensed or DS. The aim of our work was to develop and introduce an effective DS identification system. During the study, we were faced with two main tasks: 1) to determine the presence of DS in the TD samples batch, and 2) to identify DS for subsequent data filtering. The milk analysis laboratory already uses a developed method to detect DS in the orders, but the existing method requires milk samples to be enumerated in the sequential order they were collected and is based on the calculation of standard deviation (std) in groups of sequential

samples. It shows high accuracy in the recognition of DS in TD milk samples dispensed sequentially from one tank, however, if samples were dispensed not sequentially but were mixed with unique samples, or if samples were dispensed from multiple milk tanks, the accuracy of this method reduced. To improve the DS recognition, we developed a clustering algorithm which relies on the density-based method of finding clusters in spatial data – OPTICS (Ordering Points To Identify the Clustering Structure). The idea of the algorithm is to find clusters of high-density points in the space of milk sample parameters. Additionally, we have developed a unique clustering algorithm based on searching for areas with the least change in milk sample component values (e.g. fat and protein content) in a set of milk samples ordered by component values. Each clustering method requires a certain set of input parameters. During our research, we learned to recognize whether there were DS in a batch, and from how many milk tanks they were dispensed, and also compared DS recognition methods. Identification of all serial numbers of DS in a batch requires namely a more accurate selection of the algorithm parameters. We were able to determine the range of parameter values needed for this task, but the specific values for each batch are selected manually. Clustering quality estimation carried out on 4 tests, involving 2640 samples in total, 264 of which were DS, resulted in a mean Rand index score close to 0.9 indicating a high proportion of correctly clustered points, and a mean silhouette score close to 0, which is associated with the erroneous identification of some unique samples as dispensed. A comparison of DS recognition methods by std and by clustering showed similar results on batches in which DS were consecutive, and an advantage of clustering methods on batches in which DS were taken from several tanks and/or mixed with unique ones. New DS recognition methods have already found their application in the work of the milk analysis laboratory and will be further developed for use in more accurate data filtering before using this data in genomic evaluation of breeding value systems.

Enter keywords

data quality control, TD milking, milk samples dispensed, dairy cattle, genomic selection reliability, clustering method