Onimit Data preparation procedure

This update is about work package 3 that includes the data filteration and preparation procedure. The aim of this step was that SEGES gets hands on experience with the data filtration process and produce the criteria needed to be implemented in the pipeline for data preparation.

Data filtration procedure:

This procedure is based on the manuscript from Peter L. of AU entitled "**A data driven approach to processing sniffer-based gas emission data from dairy cattle**". Because sniffer system is an add on to existing AMS or feeder machines, therefore, they require further data streaming programs to convert the raw data to useable information in the breeding system. Sniffer data contain the methane and CO2 measurements on second basis that are collected on farm. AMS data is the milk recording information for each visit to the milking robot machine installed on the farms.

The steps needed to be taken are: 1) alignment of sniffer data with the AMS or feed data to obtain the animal ID; 2) cleaning procedure to remove outliers; 3) creating a phenotype that can be used by farmers as indicators and breeding organizations as selection tool.

The synchronization of sniffer data to AMS machines are currently run by GEDA program. The program is modified and updated recently in September 2023.

Here, the steps to remove the outliers and create the phenotype will take place.

Cleaning procedure to remove outliers:

1. Obtaining baseline values from idle periods: idle periods are the time between each milking until the AMS machine gets occupied again. This can be from seconds to minutes. The baseline value is Idle periods vary in duration from a few seconds to several minutes, as also AMS cleaning cycles are considered idle periods. At the beginning of an idle period the sneezer valve will block the airstream (disturb measured values) and at the end of an idle event if time synchronization is not perfect, the last part of the idle event may be disturbed by breath from the new cow. These are called "edge effects".

In this study, the edge effects were removed, and the rest of the data was condensed and mean and SD was used to obtain the "true" background concentrations. However, simply removing the data can create additional issues. So different scenarios were investigated to obtain the best possible approach. moving average (MA) methods were used which is based on simple exponential smoothing functions.

Handling idle periods helps to avoid the slip over effects of each cow record on the next cow data. The resulting baseline concentrations were then subjected to further analysis for systematic and random effects.

2. Factors affecting baseline concentrations: here the systematic and random effects such as time of the day, AMS unit, and random effects of date, as well as drift over calendar days, or even the duration of the period are handled. Here a linear mixed model was used to obtain the effects of each factor. Fitted values from moving averages for CH4, CO2 and their STD's were estimated here.

3. Emission data obtained when cows were in the AMS: the cow data is the data that is assigned to a cow during the milking period. This data also needs to be filtered as the idle period. The data could be affected by multiple factors such as carry over and measurement noises. Milking duration can be from few minutes to several minutes. The concentration of gas emissions are highest during the early milking recording period. Pruning of cow records to remove "carry over" and noise at long milkings was investigated similar to the data from idle periods. In addition, the behaviour of the cows affects the data, the main behaviour is "head lifting". The gas concentration is highest when the cow puts her head into feeder as the sensors are in the feeder bin. However, they lift their head out of the bin when the bin is empty. This behaviour affects the data. a "head lifting" threshold can be imposed, so that records with concentrations below baseline plus a latitude above it should be omitted.

4. Summary variables for each visit: here the response variable is defined. For this study, the simple mean and median per visit and per date was obtained. In addition, the ratio between excess CH4 and CO2 concentrations were calculated. Then the variables were analysed with a linear mixed model equation that included the effects of AMS machine and time of day as fixed effects, and Date, Cow and Residuals were random effects.

Mahmoud Shirali

SEGES Innovation, Specialkonsulent

SEGES INNOVATION

SEGES Innovation
Agro Food Park 15, 8200 Aarhus N

T: +45 8740 5000 - F: +45 8740 5010 - E: info@seges.dk