Supplementary materials for this article are available at https://doi.org/10.1007/s13253-024-00622-0.

Using Density and Fuzzy Clustering for Data Cleaning and Segmental Description of Livestock Data

Torgunn Aslaug SKJERVE, Gunnar KLEMETSDAL, Bente Aspeholen ÅBY, Jon Kristian SOMMERSETH, Ulf Geir INDAHL, and Hanne Fjerdingby OLSEN

The cluster algorithms density-based clustering with noise and fuzzy c-means were used to edit and group a large, noisy data set from a livestock herd recording scheme consisting of slaughter records on 73,743 bulls. Density-based clustering with noise was used for data selection with $\epsilon = 0.06$ and minPts = 8. The remaining data ($n = 65{,}446$) was exposed to a fuzzy c-means analysis to partition data based on three variables: Age at slaughter, carcass weight, and average daily carcass gain. Appropriate number of clusters was chosen by the maximum value of the modified partition coefficient ($k = 3$ clusters). Cluster validation for both hard cluster assignment and cluster membership was performed with linear models and a permutation test. The three clusters had centroid values for slaughter age and carcass weight interpreted as: Part of production systems characterized by high bull turnover (Cluster 1), production systems aiming for heavy slaughter weights (Cluster 2), and a less intensive system with higher roughage proportions (Cluster 3). The results show that the approach can be successfully combined to segment meaningful groups from large, noisy industry data, exemplified by the description of slaughter performance records.

Supplementary materials accompanying this paper appear online.

Key Words: Beef production; DBSCAN; Fuzzy c-means; Norwegian red.

T. A. Skjerve (✉) · G. Klemetsdal · B. A. Åby · H. F. Olsen, Department of Animal and Aquacultural Sciences, Norwegian University of Life sciences, Oluf Thesens vei 6, 1430 Ås, Norway
(E-mail: *torgunn.aslaug.skjerve@nmbu.no*)
G. Klemetsdal (E-mail: *gunnar.klemetsdal@nmbu.no*)
B. A. Åby (E-mail: *bente.aby@nmbu.no*)
H. F. Olsen (E-mail: *hanne.fjerdingby.olsen@nmbu.no*).
J. K. Sommerseth, Farm Advisory Services - R&D Department, TINE SA, BTB-NMBU, 5003, 1430 Ås, Norway
(E-mail: *jon.kristian.sommerseth@tine.no*).
U. G. Indahl, Faculty of Science and Technology, Norwegian University of Life Sciences, Drøbakveien 31, 1430 Ås, Norway (E-mail: *ulf.indahl@nmbu.no*).

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-024-00622-0

Published online: 14 May 2024

1. INTRODUCTION

Large-scale field data collection in livestock production has great potential as a resource, both in farm management and research. However, the degree to which it is utilized is often hindered by technical challenges such as incompatible structure and data quality. Thus, there is a need for efficient methods to explore, analyse, and transform data into compiled summaries that can be used in communication and dissemination. To address this issue, it is necessary to build efficient approaches to data processing (Eastwood et al. 2019; White et al. 2021; Cravero et al. 2022).

Traditionally, data filtering has hinged on a detailed knowledge of how data are typically distributed within a given space, an approach that relies heavily on previous experience. There are however methods available that bypass the need for such a priori assumptions. Implementing machine learning techniques has been proposed as a possible step towards improving the data process. By leveraging machine learning algorithms, the industry and researchers can effectively analyse the vast amounts of data collected, enabling them to unlock insights that would be difficult to discern otherwise (Hudson et al. 2018). One of the most popular and effective machine learning clustering techniques used for data analysis is the k-means algorithm. It is however, sensitive to noise in the data, which can result in sub-optimal clustering results (Ahmed et al. 2020). Being able to efficiently interpret clustered observations would highly improve the data exploration process in livestock farming, but the presence of noise is inevitable. Thus, although k-means clustering can be a useful candidate tool for analysing livestock farming data, the method's sensitivity to noise can pose a significant challenge.

In livestock production, dairy farms especially have a long-standing tradition of data-driven management, with the widespread adoption of milk recording systems dating back to the late nineteenth century (Armitage 2007; Hudson et al. 2018). The availability of these large generational data sets to scientists and farm advisory services has been instrumental in the modern development of dairy production. It is, however, recognized that this data is prone to recording error and reporter bias (e.g. Espetvedt et al. 2012; Koeck et al. 2012). As a consequence, editing and processing represent a significant and time-consuming aspect of working with these type of databases (Hudson et al. 2018).

The density-based clustering algorithm, Density-based Spatial Clustering of Application with Noise (DBSCAN), was initially developed for use on spatial data. As it is based on density, it requires no other assumptions on distribution and relies solely on the parameters' region radius (ϵ) and a minimum number of points (minPts) to cluster and identify outliers (Ester et al. 1996). This feature makes the algorithm appealing for both data editing and data selection tasks, particularly when it is necessary to define sub-groups within the data. A few studies applying DBSCAN on agricultural sensor farm data have proven the efficiency of the algorithm (Ismail et al. 2019; Miao et al. 2021), but to the best of the authors' knowledge has yet to be employed on larger-scale population data sets, such as national herd scheme data.

A further challenge of applying k-means to livestock farming data is the abrupt way the observations are assigned a classification. Although there may be cases where this approach is appropriate, much of the data is likely to overlap. This can be addressed by fuzzy set theory,

which introduces the concept of uncertainty in membership through a membership function (Zadeh 1965). The clustering algorithm fuzzy c-means, as proposed by Bezdek (1981), is the fuzzy counterpart of k-means and has been applied in livestock farming research for tasks such as recognizing body attributes and monitoring animal welfare (Zhang et al. 2018; Ojo et al. 2022). It assigns membership degrees rather than membership akin to assigning a probability of each observation belonging to a particular cluster.

One example highlighting the challenges posed by livestock farming data is the data collection process on the bulls of the Norwegian red breed (NR). In Norway, a large proportion of domestic beef is produced by combined dairy- and beef enterprises. Most dairy farms participate in the Norwegian Dairy Herd Recording System (NDHRS), which focuses on data of importance to dairy production, used by advisory services, breeding organizations, and research. Finisher bull rearing strategy, production goals of age at slaughter and carcass weight are important both for the environmental and economic outcomes (Nguyen et al. 2010; Bonesmo and Randby 2011). However, this information is usually unavailable in the large NDHRS data sets, making detailed production analysis of NR finisher bulls challenging. The NDHRS slaughter performance records feature a high volume of records, but low degree of details and significant levels of noise, and discerning finisher rearing bull strategy in these data sets are therefore well-suited as a test case for new data exploration approaches.

In this study, we aimed to assess the combined approach of utilizing DBSCAN and fuzzy c-means clustering algorithms to clean and categorize a large data set on livestock production, using slaughter performance data on NR bulls as an example.

2. DATA

Yearly domestic beef production in Norway is made up of 86 000 tons of slaughter approved for human consumption (Statistics Norway 2017–2021). Approximately 300 000 animals are slaughtered each year, of which 46% are of the slaughter category "young bull" (bulls between 301 and 730 days of age.) Most bulls come from combined dairy-beef operations (Animalia 2017–2021; Tine Rådgivning 2017–2021; Statistics Norway 2017–2021). Rearing strategies on finisher bulls in combined dairy- and beef production are usually based on varying amounts of silage and concentrate. The choice of rearing strategy, the goal for slaughter weight, slaughter age, and concentrate levels depends on the availability and price of feed resources (Bonesmo and Randby 2011).

The NHRDS covers approximately 97% of all dairy herds in the country. Through its large proportion of NR, with 88% of the roughly 500 000 animals active in the recording system (Rådgivning 2022), it is also the largest source of data on the NR bulls. The system stores on-farm data from all participating herds, in addition to data from advisory services, slaughterhouses, milk laboratories, the breeding company Geno, accountancy data, and Dyrehelseportalen, a nationwide database for animal health records used by veterinarians and farmers. NDHRS is owned by the company Mimiro, which manages a large proportion of large-scale field data recordings in Norwegian agriculture. The data collected are used for various purposes, such as research, breeding evaluations, and governmental reporting. It is organized into several tables based on the source of the reporter and the theme of the

record. These tables are linked using unique identity keys for the farm, owner, or animal, or a combination of all three, as well as specific table identifiers.

In this study, a data set of slaughter performance records on 76 293 bulls slaughtered in the year 2017 was used. The sources for this information are farmers and slaughterhouses. The data were edited based on the three variables, namely slaughter age (Age), carcass weight (CW) and average daily carcass gain (ADCG), with initial corrections for recorded slaughter category and gender. In finisher bull rearing, average daily gain is an important factor, as age and average daily gain are the two main determinants of the animal's nutritional requirements. Average daily gain is defined as (Live weight at slaughter – birthweight)/(Age at start of period – Age at end of period). It is composed of growth in all tissues, including bone and organs. ADCG is based on carcass weight and can be used as approximation when live weight is not recorded as is the case data set used for this study.

3. METHODS

3.1. Data Compilation and Pre-processing of Raw Data

Slaughter records are reported by the slaughterhouse and stored by NDHRS, including date of slaughter, CW, EUROP carcass and fat classification, and slaughter category. In addition, to determine the animals' age, registered breed, and gender, birth information was obtained from the Birth Information table from the NDHRS, which contains data reported by producers. Information regarding the primary cause of culling and the farm at the time of slaughter was extracted from the Animal History table, which is a chronological record of all the times and reasons a farmer changed an animal's status in the NHDRS registration systems. The date when the animal was first recorded in NDHRS is also recorded. In most cases, this date corresponds with the birth date given in NDHRS, but there are a few instances where it differs significantly. However, this was not the case for any of the slaughter records used in this analysis.

All data editing of the raw data was done using DATA STEP and PROC SORT in SAS®9.4 (SAS Institute Inc 2013). Age at slaughter (Age) was calculated as the difference between date of birth and date of slaughter. ADCG was estimated as (CW-(birth weight/2)/Age). As birth weight is not recorded in NDHRS, it was put at 43.5 kg, which is the five-year average for NR in the Norwegian Beef Cattle Herd Recording System for the period 2015–2019 (Animalia, 2015–2019; $n = 749$). Only records with the event categorized as "Out to Slaughter" for animals slaughtered in 2017 were kept ($n = 201{,}162$). Of these, 77,955 were recorded with gender "male" in their birth record, where of 106 were classified in a slaughter category for female animals ("Cow", "Young Cow" or "Heifer"). The latter group was most likely erroneously reported to NDHRS and was therefore excluded. The slaughter category "Steer" was also excluded ($n = 1234$). Finally, animals dead for other reasons than slaughter were eliminated by the recorded reason for the update in the database ($n = 294$), and by primary reason for culling indicating disease ($n = 28$). Thus, the data used in the DBSCAN analysis consisted of 76,293 records from 5424 farms, with 2 550 being reported as "bull", 5943 as "calf" and 67,800 as "young bull". Plots were generated using the ggplot2-package in RStudio (Wickham 2016).

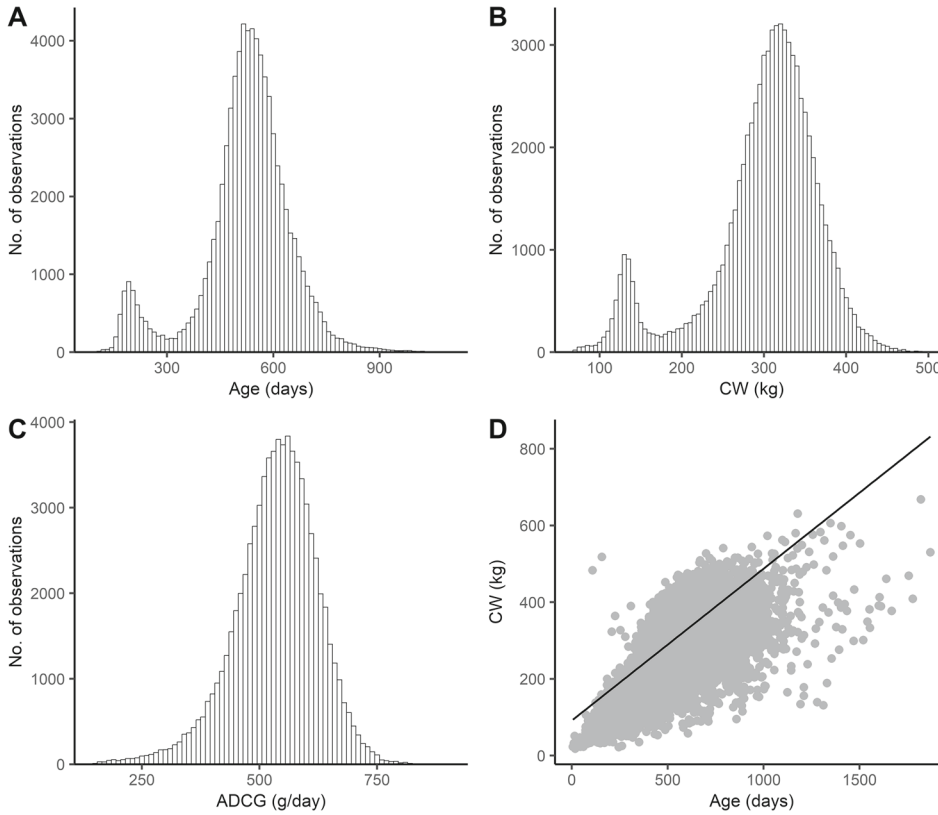

Figure 1. Raw data distribution of the variables **A** age in number of days at slaughter (Age), **B** carcass weight (CW) in kg, **C** average daily carcass gain (ADCG) in g/day, and **D** a scatterplot of Age and CW, with, respectively, 216, 324, 144 and 4 extreme observations removed for visibility.

Age and CW were bimodally distributed with a smaller peak for the period corresponding to the age and weight of rose veal slaughter (Age <301 days) and a higher peak corresponding to the age and weight of bull slaughter (Fig. 1A, B). The modality shifted at approximately 300 days of age and approximately 180 kg carcass weight. Figure 1C displays the distribution of the ADCG. The scatterplot of Age and CW (Fig. 1D) indicates non-collinearity as the variation around the regression line, which is a close proximate of ADCG, is considerable. Nevertheless, it is important to note that Age and CW exhibited a strong positive correlation with a coefficient of 0.74. The correlation between Age and ADCG ($r = -0.25$) was weaker, whereas CW and ADCG had a moderately strong correlation ($r = 0.41$).

The mean value of Age, CW, and ADCG was 525 days, 299 kg, and 534 g/day, respectively, with minimum and maximum values of 6 and 5505 days, 18 and 668 kg (CW), and −208 and 4271 g/day (ADCG).

3.2. NOISE REMOVAL, DATA SELECTION AND GROUPING BY CLUSTERING

DBSCAN was used for noise removal and to separate the lowest and the highest peak in the distribution of Age and CW, using the package dbscan in RStudio. Data were scaled

using the scale() function prior to clustering (Hahsler et al. 2019; RStudio Team, 2022). The DBSCAN algorithm relies on density estimation, which involves counting the number of data points within a specified radius (ϵ) from each data point. A region is designated as dense if it meets the minimum number of points requirement (minPts) (Schubert et al. 2017). Based on the structure of the raw data and purpose of the analysis, the following priors were used for setting the sensitivity of the DBSCAN parameters:

1. The bimodality of the distribution of Age and CW will create two distinct density clusters. The density cluster formed around the higher density peaks in Fig. 1 (upper panels) relates to bull slaughter, the density cluster formed around the lower relates to rose veal slaughter.

2. For further analysis, the two clusters must be clearly separated.

In Sander et al. (1998), the authors conclude that the choice of minPts is not very crucial for the DBSCAN algorithm and propose a default value of 2*dimensions -1. The ϵ-value was therefore set by examining the separation of the two density clusters around the Age and CW density peaks with minPts = 5 (Fig. 2A). MinPts was then adjusted to ensure a clear separation and avoid classifying noise as natural variation, following the recommendations by Schubert et al. (2017) for noisy data sets. The largest DBSCAN cluster ($n = 65{,}438$) and a smaller cluster nested within the largest ($n = 8$) were retained for further consideration and subject to the subsequent fuzzy c-means cluster analysis. For more details on DBSCAN, see Supplementary material 1.

The selected data set consisted in total of 65 446 records on bulls slaughtered from 326 to 809 days of age, collected from 4 600 farms with a total of 13 063 distinct slaughter groups (bulls slaughtered from the same farm on the same day). Fuzzy c-means cluster modelling with the variables Age, CW, and ADCG was executed using the fclust package in R (Ferraro et al. 2019). The Modified Partition Coefficient (MPC) as suggested by Davé (1996) was used to decide the number of clusters ($k = 3$). The fuzzifier value of the algorithm was taken to be 2, as recommended by Bezdek (1981) for fuzzy clustering.

3.3. Cluster Validation

Effects of hard cluster assignment were tested with linear regression. The highest membership value was used as criterion for hard cluster assignment for each individual bull. Effect on EUROP carcass conformation class, EUROP carcass fatness class and carcass value (NOK per carcass) was estimated using PROC GLM in SAS® 9.4 (SAS Institute Inc 2013) according to the model

$$y_{ij} = \mu + cl_i + e_{ij} \tag{1}$$

where y_{ij} is EUROP carcass confirmation, EUROP carcass fatness classification or carcass value of the individual bull j in cluster i, cl_i is the fixed effect of the individual cluster assignment of the bull, and e_{ij} is the random residual term for the jth bull in cluster $i \sim N(0, \sigma^2)$. Prior to analysis, EUROP classifications were transformed to a 15-point numerical scale as done in Hickey et al. (2007).

For carcass value, the explanatory effect of hard cluster assignment of bulls from the same farm slaughtered at same date (slaughter group) was compared to a model only including the random effect of slaughter group by the means of log-likelihood ratio testing. In these analyses, slaughter groups with less than 4 slaughtered bulls were excluded (n = 5 389 slaughter groups). The number of slaughter groups retained, described and analysed was 7 684 from 3 378 different farms, with 55 818 bulls. The following three models were run:

$$y_i = \mu + e_i \tag{2}$$

$$y_{ij} = \mu + cl_i + e_{ij} \tag{3}$$

$$y_{ij} = \mu + sl_i + e_{ij} \tag{4}$$

where y is the EUROP carcass confirmation, EUROP carcass fatness classification or carcass value for bull i, cl_i is the random effect of the individual cluster assignment with $\sim N(0, \sigma^2_{cl})$, sl_i is the random effect of slaughter group $\sim N(0, \sigma^2_{sl})$, and e is the random residual term for the jth bull in cluster $i \sim N(0, \sigma^2_e)$. Significance of the random effects was tested with a log-likelihood ratio test, using the following test statistics: D = -2[log-likelihood full model – log-likelihood of the simple model] with one degree of freedom. The log-likelihood values were obtained by PROC MIXED in SAS® 9.4 (SAS Institute Inc 2013).

Finally, cluster membership was explored using a permutation test. Effect of country region on cluster membership was explored using a permutation test of 5000 repetitions on regional means, executed in R Studio using package dplyr (Wickham et al. 2023) and base R (R Core Team 2023). Regions based on geographic location and production conditions, as defined by the Survey of Account Statistics for Agriculture and Forestry (the Norwegian Farm accountancy data network), were used. The regions are geographically distributed throughout the country, and certain production conditions are indicated. They are named as follows: “Southwest Marginal”, “Southwest”, “East marginal”, “East Lowlands”, “North”, “Central Marginal”, “Central Lowlands” and “West”. Lowlands and the region “Southwest” are areas with relatively favourable production conditions, whereas “marginal” indicates the opposite. A map of the regions can be found as Fig. S1.1 in Supplementary material 1. For each permutation, region identity was shuffled randomly between the 65 446 bulls in the original data set, and average cluster membership for each region was calculated, resulting in a data set of 5000 permutated region means. Euclidean distance between the permutated and realized region means towards the average mean cluster memberships of the data set was calculated and tested with the H_0-hypothesis: Euclidean distance between regional average cluster membership and overall data set average cluster memberships is equal to 0. Detailed description and programs for the permutation test can be found in Supplementary material 1.

The descriptive statistics were generated using Microsoft® Excel. Plots were made using R base plot (R Core Team 2023) and ggplot2(Wickham 2016).

4. RESULTS

4.1. CLUSTERING

The DBSCAN parameters were set by testing the separation of two larger clusters based on two peaks in the density distribution of Age and CW, where the largest of the two was assumed to correspond to bull slaughter and the smaller rose veal slaughter. The default value suggested by Sander et al. (1998) for minPts (2*dimensions -1) was used to define the ϵ-region. To allow for the greatest possible age span within the clusters, ϵ was set to 0.06 which is the largest value at which the clusters separate in the 3D-space. The minPts parameter was then adjusted to avoid noise classified as natural variation by the more generous ϵ-value (minPts = 8). Figure 2 depicts the changes in number of clusters, number of noise points, number of observations in the largest cluster, and number of observations in the smaller cluster at the default minPts = 5 (Panel A) and the adjusted minPts = 8 (Panel B). Notably, increasing minPts to 8 increases the value of ϵ where the two clusters re-merge to 0.07, indicating a very low density in the region between the two. The lower number of clusters forming also indicates that less nonsensical clusters are formed, avoiding the misclassification of noise in the data.

The DBSCAN resulted in a total of 106 clusters and 5 017 noise points. Among these clusters, only the two largest clusters had more than 45 observations, with the smaller cluster containing 4 372 observations and the larger cluster containing 65 438 observations. To visualize the clustering results, scatter plots of Age versus CW and Age versus ADCG were generated and coloured by cluster, as shown in Fig. 3. Based on these plots, the largest cluster and a smaller one nested within it were selected for further analysis. Notably, correlations between the three variables changed, where the correlation between Age and CW decreased ($r = 0.54$), whereas the correlations between ADCG and Age ($r = -0.48$) and ADCG and CW ($r = 0.48$) increased. The decrease in correlation between Age and CW and the increase between Age and ADCG indicate that the data selection was successful, as the variable ADCG is the factor that will directly differentiate the feed rations.

The modified partition coefficient (MPC) for $k = 1$ to $k = 10$ was calculated, and its highest value was reached at $k = 3$ (value = 0.37). For cluster 1, the average membership for observations in the data set was 0.328, while for cluster 2 it was 0.367, and for cluster 3, it was 0.305. Overall, 8 880 of the observations had an unclear cluster assignment, determined as membership degree <0.5, with the highest proportion in cluster 3 (0.15) and the lowest in cluster 2 (0.12). The scatter plot in Fig. 4 shows the relationship between Age and CW, coloured by cluster membership with cluster centroids shown as yellow diamonds.

4.2. CLUSTER VALIDATION

To evaluate the performance of the approach, descriptive statistics were examined to identify relevant patterns, and effects were tested in several statistical models, both for hard clustering assignment and cluster membership patterns. Based on the descriptive statistics in Table 1, the three clusters can be described as follows: Cluster 1 (red): “Early Age and low CW”, Cluster 2 (blue): “Medium Age and high CW”, and Cluster 3 (green): “Late Age

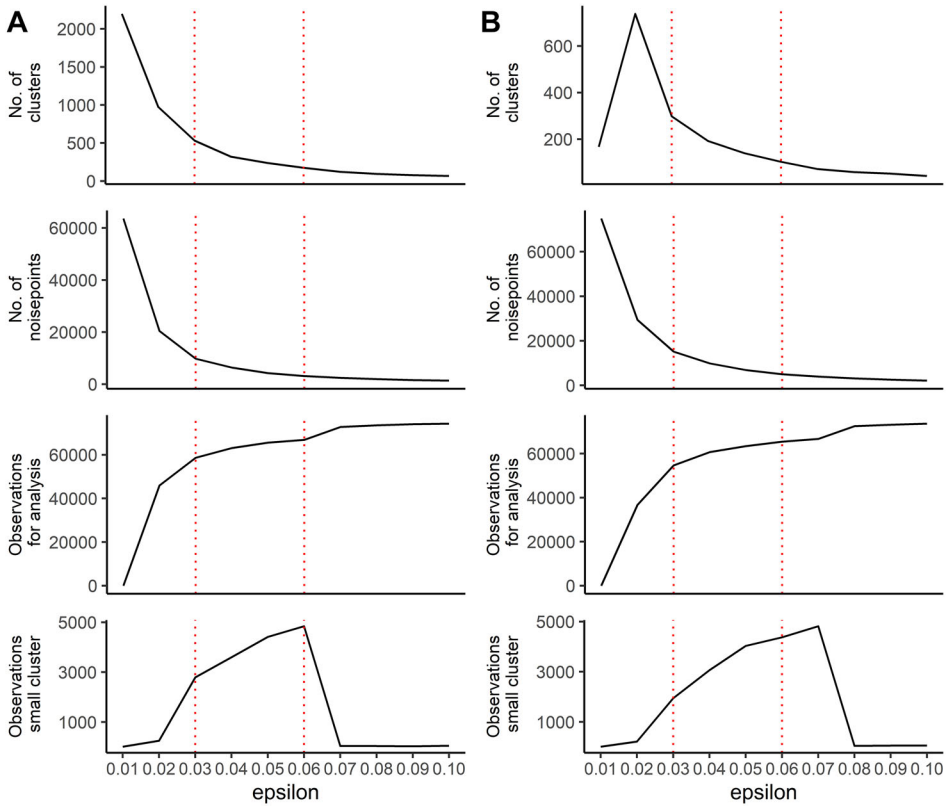

Figure 2. Number of formed clusters, number of noise points and number of observations in the largest and the smallest cluster for ε-values between 0.01 and 0.1, at **A** minPts = (2*dimensions—1) and **B** minPts = 8. The red dotted lines indicate the ε-values where the two larger clusters first separate and remerge at the default minPts-value suggested by Sander et al. (1998) (Color figure online).

and Medium CW". Bulls with the highest cluster membership to Cluster 1 falls within the interval 326 to 622 days with CW ranging between 177 and 315 kg. Likewise, bulls with the highest membership to Cluster 2 falls within the interval 370 to 690 days, with CW ranging between 276 and 448 kg, and bulls with the highest membership to Cluster 3 finally ranging between 552 to 809 days, with CW ranging between 232 and 442 kg.

The distribution of clusters within farm slaughter groups was explored with hard cluster assignment. Of the slaughter groups, 22.7 % had bulls all within one cluster, 57.6 % had all bulls within two clusters, and the remaining had bulls spread across all clusters. Slaughter groups with individuals assigned to Clusters 1 and 2 were the most prevalent, making up 32.5 % of the slaughter groups. Average membership for individual bulls in single-cluster slaughter groups ranged between 0.70 and 0.71, for two-cluster slaughter groups the range was between 0.41 and 0.47 for the two clusters in question, and three-cluster groups between 0.30 – 0.32.

The effect of the clusters on slaughter return, EUROP carcass confirmation classification, and EUROP carcass fatness classification was found to be significant, with the most prominent effect on slaughter returns with R^2 at 0.45 (Table 2). When the effect of the cluster was

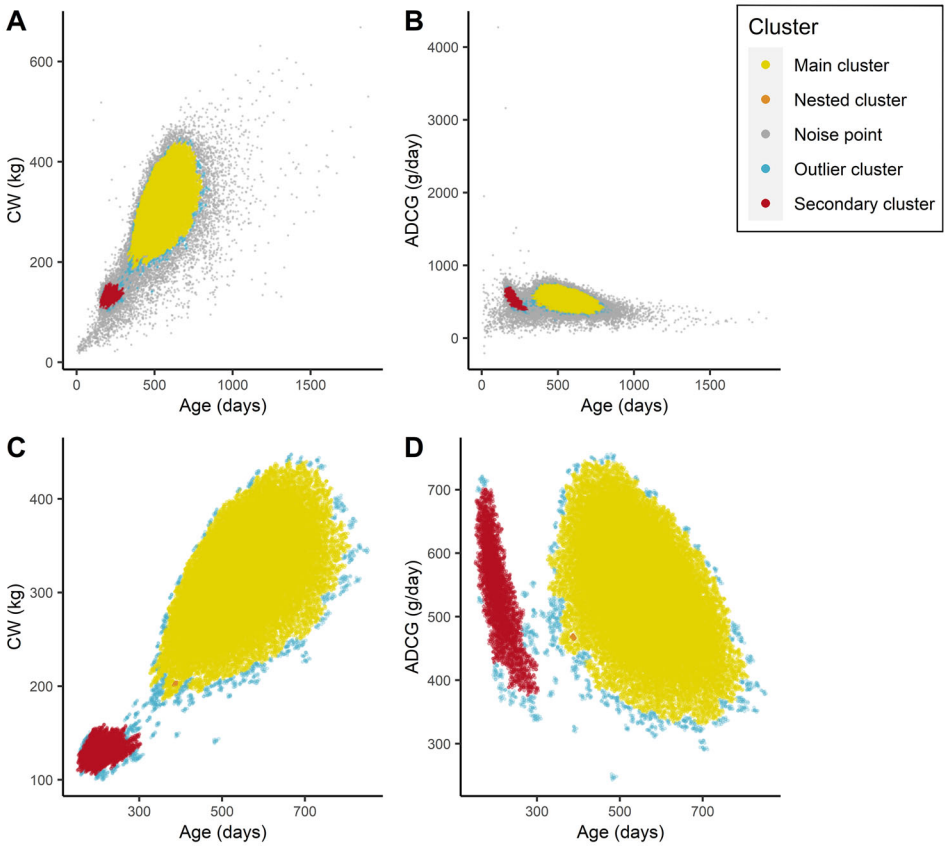

Figure 3. Scatter plots of age at slaughter (Age) and Carcass weight (CW) (**A** and **C**) and Age and Average daily carcass gain (ADCG) (**B** and **D**), with noise points (**A** and **B**, $n = 76{,}287$) and without noise points (**C** and **D**, $n = 71{,}276$), showing the main cluster (yellow; $n = 65{,}438$), lateral cluster (red; $n = 4372$), a small cluster nested within the main cluster (orange; $n = 8$), noise clusters (blue; $n = 1458$), and outliers (grey, $n = 5017$) (Color figure online).

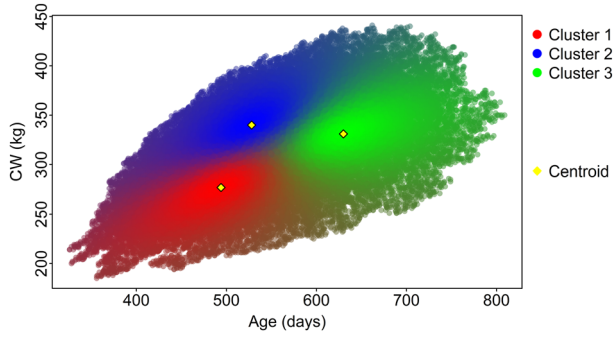

Figure 4. Scatter plot of age at slaughter (Age) and carcass weight (CW), where the colours indicate the observations' belonging to clusters 1–3. Cluster centroids are marked with yellow diamonds (Color figure online).

Table 1. Cluster information and cluster central values for age at slaughter (Age), carcass weight (CW), average daily carcass gain (ADCG) and average carcass quality measures for Clusters 1 to 3

Variables	Cluster 1	Cluster 2	Cluster 3
Cluster size[1]			
N observations	21,521	24,401	19,524
Average membership degree	0.69	0.71	0.69
Centroid values			
CW (kg)	277	341	331
ADCG (g/day)	520	606	494
Mean economic measures[1]			
EUROP carcass conformation class[2]	4.89	5.72	5.36
EUROP carcass fat classification[2]	6.23	6.62	7.14
Carcass value (NOK)	13,211	16,678	16,229

[1]Hard cluster assignment based on highest membership degree

[2]EUROP class transformed to a 15-point numerical, linear scale according to Hickey et al. (2007)

Table 2. Estimated test statistics (F- and P-values) as well as coefficient of determination (R^2) of individual cluster assignment on EUROP carcass classification, EUROP fatness classification and carcass value; likelihood ratio test statistics (D) of the random effect of cluster and slaughter group on carcass value in NOK

	F	P	R^2	D
Fixed model				
EUROP carcass conformation class[1]	6 864.3	< 0.001	0.17	
EUROP carcass fatness class[1]	3 971.9	< 0.001	0.11	
Carcass value (NOK)	27 315	< 0.001	0.45	
Random model				
Carcass value (NOK) as effect of cluster				33 097.7
Carcass value (NOK) as effect of slaughter group				29 070.6

[1]EUROP class transformed to a 15-point numerical, linear scale as in Hickey et al. (2007)

Hard cluster assignment is according to highest membership degree

compared with effect of slaughter group as random effects, the log-likelihood test statistics were in favour of the cluster model (Table 2). An example of how the scatter plot in Fig. 4 can be used in a real-life sense, as graphic feedback to the farmer on the performance of the slaughter group compared to the rest of the population can be found in Supplementary material.

The permutation test results are visualized in Fig. 5. For all but two regions, the realized Euclidean distance for the region mean from the overall mean in the data set (shown in the figure as a dotted red line) fell significantly above any of the permutated values. For the remaining two regions, the North and the Central Lowlands, the permutated means were 19.6 % and 45.2 % higher, respectively. The largest distance was found for the regions Southwest and East Lowlands, where membership to the clusters with heavy slaughter weights (Cluster 2 and 3) was higher, and West where membership leaned towards Cluster 1 (early slaughter) (Table S1.1 in Supplementary material 1).

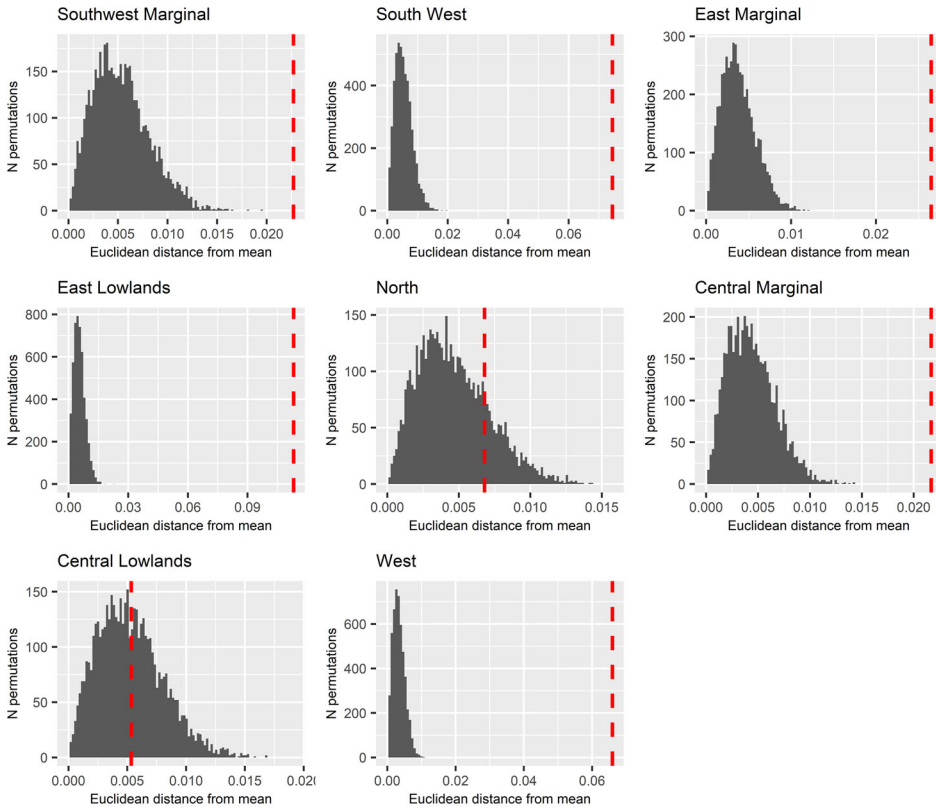

Figure 5. Histogram of regional means for 8 agricultural regions from the permutation test of 5000 permutations with random shuffling of region identity of 65,466 bulls from the data set. The red dotted line is the realized regional mean for each separate region (Color figure online).

5. DISCUSSION

Looking at the results, the approach of sequential DBSCAN and fuzzy c-means on the test case data set performed well. Cluster validation also showed that the resulting fuzzy clusters can be useful in a real-life context. The DBSCAN algorithm was developed to efficiently handle large databases where information in and about the data is limited (Ester et al. 1996). While used in a variety of fields, including precision farming, auto-detection of crop quality, and plant development (Ismail et al. 2019; Miao et al. 2021), it has to the authors' knowledge never been applied to large-scale field data in livestock production. There may be several advantages to using DBSCAN compared to heuristic frameworks based on preconceptions or apparent data distributions. For one, it allows for an effective definition of the typical space of data points, exemplified by the largest cluster representing bull slaughter. Furthermore, DBSCAN identifies outliers in n-dimensional space, rather than from a one-dimensional variable distribution. Consequently, single variable values that would be within the expected limits on their own, but combined with other variable values, make little sense and can be easily detected. DBSCAN was also tested on a 30-year older data set from 1993 consisting of more heterogeneous data than the one used in this study. After the required

adjustment of the minPts and ϵ parameters, the algorithm effectively distinguished between a calf and a bull slaughter cluster. This demonstrates that the algorithm is effective also with older data produced under different conditions. A perceivable disadvantage is the arbitrary nature of deciding on the ϵ and minPts parameters. For both tested data sets, we used the density distribution of two variables (Age and CW) with the goal of isolating two modalities, already being aware of the presence of two distinct periods for slaughter in Norwegian bull slaughter production. Where this kind of a priori knowledge is lacking, guidance to minPts and ϵ can be difficult. However, the algorithm is under constant development, and even today some versions exist that can be used if the data itself does not give a clear indication of parameter choice (Karami and Johansson 2014).

To extract classification clusters from this particular data set, a fuzzy logic approach was chosen over clear-cut algorithms that assign absolute identities to each observation. Fuzzy logic is frequently utilized for agricultural data analysis, for the same reason as it was used here: the absolute classification is most likely unsuitable or too complex, but exploring the relative relationships can give nuanced insight and meaningful support in decision-making (e.g. Dutta et al. 2015; Mota et al. 2018; Heiß et al. 2021). The fuzzy c-means analysis on the slaughter records yielded a pattern that makes sense for real-life finisher bull rearing strategies, with rationales based either on bull turn-over (Cluster 1; Age 494 days and CW 281 kg), slaughter weight (Cluster 2; Age 529 days and CW 342 kg) or a less intensive production with higher roughage proportion (Cluster 3; Age 639 days and CW 332 kg). A similar cluster pattern was also identified in the 1993 data set used for testing the applicability of DBSCAN, demonstrating that the approach can be used when data are collected under different production conditions with similar results.

The degree of separation between the clusters was not very large, which is a result of the fuzzy logic approach. If the aim was to define strongly contrasted groups, a clear-cut cluster assignment obtained from the regular k-means algorithm would have done a better job. However, in this case, the goal was to discern farming practices that commonly overlap. The likelihood of multiple practices being present in the same space is also an issue. Thus, the emphasis was put towards similarity to ground truth, rather than contrasting and clear classification. This approach to data analysis is more appropriate for complex issues often encountered in agriculture and agricultural research, as overlapping classifications can often provide a more accurate representation than a rigid class assignment. Several studies have demonstrated the efficacy of fuzzy classification for a range of agricultural issues. For instance, Ji and Wu (2022) proposed an application that monitors the severity of black measles infestation in grapes. Similarly, an approach that enables the monitoring of irrigation requirements for alfalfa was demonstrated by Li et al. (2019).

As examples of how the proposed approach can be applied, clusters were evaluated on both clear-cut cluster assignments and cluster memberships. In a log-likelihood comparison, the effect of the main cluster assignment explained as much as 45% of the variation in carcass value, and when compared with the explanatory value of slaughter groups (bulls slaughtered from the same farm on the same day), the latter was outcompeted (Table 2). These verification analyses indicate that the clusters can absorb considerable information about production intensity, as well as a potential for advisory services at the farm level. To explore its validity on a larger scale, the difference in cluster membership of agricultural

regions was examined by the measure of Euclidean distance. As the assumption is that a farm adjusts its strategy to the given production conditions, it was expected to find these differences reflected in the cluster memberships. Using a permutation test on the distance of region means towards the overall mean, overlap was found for only two of the eight regions, showing that the realized region differences were indeed significant. Furthermore, membership to Cluster 1, with the assumed high bull turnover strategy, was found relatively higher in regions with restricted access to roughage, while membership to Clusters 2 and 3 was found relatively higher in regions where feed is presumably less of a concern.

Using herd recording scheme-type data is a quick way of obtaining population information, but there is a limit to how much information it is reasonable to ask contributors to include. As an alternative, this study demonstrates that partitioning clustering techniques such as fuzzy c-means can improve the interpretability of data without any additional contributions from reporters. It is, however, important to keep in mind that one specific analysis does not give the final answer to all questions. Several other clustering techniques than the one used here could provide interesting insights into the data, although not necessarily useful for the question in mind (Mota et al. 2018; Rodriguez et al. 2019).

6. CONCLUSION

In summary, using DBSCAN and partitioning clustering algorithms has the potential to improve and expand the usability of large-scale field recordings from the industry and other databases. This study shows how the approach can be applied to slaughter performance data to form real-life applicable groups. However, this is not the only purpose for which this type of processing is relevant. In research, the approach can be a cost-effective way of addressing detail concerns with large data sets, and potentially improve different modelling tasks, such as modelling of farm emissions and economy.

SUPPLEMENTARY INFORMATION

Additional analysis and R code are included in Supplementary material 1. Examples of intended use of analysis in advisory services can be found in Supplementary material 2.

ACKNOWLEDGEMENTS

The authors want to thank TINE SA for providing data from the Norwegian Dairy Cattle Recording System.

Funding Open access funding provided by Norwegian University of Life Sciences. This study was funded by the Norwegian University of Life Sciences and the Research Council of Norway through the LIVESTOCK project (295189).

Declarations
Data availability statement The data used in the study are the property of Tine Rådgivning, used under license for this study, and not publicly available.
Conflict of interest The authors declare no conflict of interest.
Ethics approval Not applicable.

[Received September 2023. Revised March 2024. Accepted April 2024.]

REFERENCES

Ahmed M, Seraj R, Islam SMS (2020) The k-means algorithm: a comprehensive survey and performance evaluation. Electronics 9(8):1295. https://doi.org/10.3390/electronics9081295

Animalia (2017–2021). Annual report: Kjøttets tilstand. Animalia, Oslo, Norway

Armitage F (2007) Milk recording: its role, past, present and future. Animal production and animal science worldwide: WAAP book of the year 2007(4):169. https://doi.org/10.3920/978-90-8686-656-4

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Springer, New York. https://doi.org/10.1007/978-1-4757-0450-1

Bonesmo H, Randby ÅT (2011) The effect of silage energy concentration and price on finishing decisions for young dairy bulls. Grass Forage Sci 66:78–87. https://doi.org/10.1111/j.1365-2494.2010.00765.x

Cravero A, Pardo S, Sepúlveda S, Muñoz L (2022) Challenges to use machine learning in agricultural big data: a systematic literature review. J. Agron 12(3):748. https://doi.org/10.3390/agronomy12030748

Davé RN (1996) Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognit Lett 176:613–623. https://doi.org/10.1016/0167-8655(96)00026-8

Dutta R, Smith D, Rawnsley R, Bishop-Hurley G, Hills J, Timms G, Henry D (2015) Dynamic cattle behavioural classification using supervised ensemble classifiers. Comput Electron Agric 111:18–28. https://doi.org/10.1016/j.compag.2014.12.002

Eastwood C, Avre M, Nettle R, Rue BD (2019) Making sense in the cloud: farm advisory services in a smart farming future. Njas-Wagen J Life Sci 90:100298

Espetvedt MN, Wolff C, Rintakoski S, Lind A, Østerås O (2012) Completeness of metabolic disease recordings in Nordic national databases for dairy cows. Prev Vet Med 105:25–37. https://doi.org/10.1016/j.prevetmed.2012.02.011

Ester M, Kriegel J, Sander X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining, 2–4 August 1996, Portland, USA, 226–231

Ferraro MB, Giordani P, Serafini A (2019). fclust: an R Package for Fuzzy Clustering. R J 11. Available at: https://journal.r-project.org/archive/2019/RJ-2019-017/RJ-2019-017.pdf

Hahsler M, Piekenbrock M, Doran D (2019) dbscan: Fast density-based clustering with R. J Stat Softw 9:1–30. https://doi.org/10.18637/jss.v091.i01

Heiß A, Paraforos DS, Sharipov GM, Griepentrog HW (2021) Modelling and simulation of a multi-parametric fuzzy expert system for variable rate nitrogen application. Comput Electron Agric 182:106008. https://doi.org/10.1016/j.compag.2021.106008

Hickey JM, Keane MG, Kenny DA, Cromie AR, Veerkamp RF (2007) Genetic parameters for EUROP carcass traits within different groups of cattle in Ireland. JAS 85(2):314–321. https://doi.org/10.2527/JAS.2006-263

Hudson C, Kaler J, Down P (2018) Using big data in cattle practice. In Pract 40(9):396–410. https://doi.org/10.1136/inp.k4328

Ismail ZH, Chun AKK, Razak MIS (2019) Efficient herd–outlier detection in livestock monitoring system based on density-based spatial clustering. IEEE Access 7:175062–175070. https://doi.org/10.1109/ACCESS.2019.2952912

Ji M, Wu Z (2022) Automatic detection and severity analysis of grape black measles disease based on deep learning and fuzzy logic. Comput Electron Agric 193:106718. https://doi.org/10.1016/j.compag.2022.106718

Karami A, Johansson R (2014) Choosing DBSCAN parameters automatically using differential evolution. Int J Comput Appl 91:1–11. https://doi.org/10.5120/15890-5059

Koeck A, Miglior F, Kelton DF, Schenkel FS (2012) Health recording in Canadian Holsteins: data and genetic parameters. J Dairy Sci 95:4099–4108. https://doi.org/10.3168/JDS.2011-5127

Li M, Sui R, Meng Y, Yan H (2019) A real-time fuzzy decision support system for alfalfa irrigation. Comput Electron Agric 163:104870. https://doi.org/10.1016/j.compag.2019.104870

Miao T, Zhu C, Xu T, Yang T, Li N, Zhou Y, Deng H (2021) Automatic stem-leaf segmentation of maize shoots using three-dimensional point cloud. Comput Electron Agric 187:106310. https://doi.org/10.1016/J.COMPAG.2021.106310

Mota VC, Damasceno FA, Leite DF (2018) Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in agricultural engineering. Comput Electron Agric 150:118–124. https://doi.org/10.1016/j.compag.2018.04.011

Nguyen TLT, Hermansen JE, Mogens L (2010) Environmental consequences of different beef production systems in the EU. J Clean Prod 18(8):756–766. https://doi.org/10.1016/j.jclepro.2009.12.023

Ojo RO, Ajayi AO, Owolabi HA, Oyedele LO, Akanbi LA (2022) Internet of Things and machine learning techniques in poultry health and welfare management: a systematic literature review. Comput Electron Agric 200:107266. https://doi.org/10.1016/j.compag.2022.107266

R Core Team (2023) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, da F. Costa L, Rodrigues FA (2019) Clustering algorithms: a comparative approach. PloS One 14(1):e0210236. https://doi.org/10.1371/journal.pone.0210236

Sander J, Ester M, Kriegel H-P, Xiaowei X (1998) Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min Knowl Discov 2:169–194. https://doi.org/10.1023/A:1009745219419

SAS Institute Inc (2013) SAS® 9.4 programmer's guide: essentials. SAS Institute Inc, New York

Schubert E, Sander J, Ester M, Kriegel H, Xu K (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst 42(3):1–21. https://doi.org/10.1145/3068335

Statistics Norway (2017–2021). Table 04181: Public meat inspection. Carcasses approved for human consumption (tonnes) (C) 2001H1–2022H1

Tine Rådgivning (2017–2021). Annual report: Statistikksamling for ku- og geitekontrollen. Tine Rådgivning, Ås, Norway

Rådgivning Tine (2022) Annual report: Statistikksamling for ku- og geitekontrollen. Tine Rådgivning, Ås, Norway

White EL, Thomasson JA, Auvermann B, Kitchen NR, Pierson LS, Porter D, Werner F (2021) Report from the conference, 'identifying obstacles to applying big data in agriculture'. Preci Agric 22:306–315. https://doi.org/10.1007/s11119-020-09738-y

Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer, New York

Wickham H, Francois R, Henry L, Muller K (2023). _dplyr: A grammar of data manipulation_. R package version 1.0.10, https://CRAN.R-project.org/package=dplyr

Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zhang AL, Wu BP, Wuyun CT, Jiang DX, Xuan EC, Ma FY (2018) Algorithm of sheep body dimension measurement and its applications based on image analysis. Comput Electron Agric 153:33–45. https://doi.org/10.1016/j.compag.2018.07.033