Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

Nina Salvesen

Digital threshold markers of atypical sheep movement on rangeland pastures by the use of the machine learning models *k*-means and DBSCAN

Master's thesis in Physics and Mathematics
Supervisor: Svein-Olaf Hvasshovd
Co-supervisor: Jon Andreas Støvneng
June 2022

NTNU
Kunnskap for en bedre verden

Nina Salvesen

Digital threshold markers of atypical sheep movement on rangeland pastures by the use of the machine learning models *k*-means and DBSCAN

Master's thesis in Physics and Mathematics
Supervisor: Svein-Olaf Hvasshovd
Co-supervisor: Jon Andreas Støvneng
June 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

NTNU
Norwegian University of
Science and Technology

Abstract

Determining the data driven description of the behaviour in sheep on rangeland grazing pastures is in demand to ensure welfare and sustainability in both Norwegian and worldwide modern pastoral systems. This study investigated the movement pattern of sheep on rangeland pastures in Norway in terms of different types of breed and diurnal and seasonal behavior traits. High resolution digital GPS collar data on free range sheep enables a deeper dimension of research on typical sheep behavior. The research in this thesis was done with the unsupervised machine learning models k-means and DBSCAN and statistical analysis. Data from 391 sheep of six breeds in five different grazing areas from 2012-2016 and 2018-2020 was studied. Variables included in the analysis were time of day, activity levels, altitude, trajectory angle, season, breed, number of lambs, temperature, and age of sheep. Diurnal behavioural traits could be identified, and the k-means model categorized their day into four characteristic activity periods. Digital threshold marker values were calculated from the mean of the behavioural outliers detected by the DBSCAN model, which when triggered indicate possible atypical movement patterns. The analysis found that the sheep velocity threshold marker should be considered split into four different values based on the circumstances surrounding the herd. The threshold marker values may be used as trigger conditions in future collar alerting technology, and thus give more specialized herd welfare warnings to the farmer.

Preface

This report is the master's thesis of Nina Salvesen for the study program Master of Science in Applied Physics and Mathematics, at the Norwegian University of Science and Technology. It is written for the Department of Computer Science under the supervision of professor Svein-Olaf Hvasshovd, with data provided by the Norwegian Institute of Bioeconomy and by the Norwegian Institute of Nature Research. The thesis will study the data driven description of the movement pattern of sheep on rangeland pastures in Norway, in connection with both typical and atypical behavior. The analysis will be completed by the use of unsupervised machine learning, statistical research and ethological theory.

I would first and foremost like to sincerely thank my main supervisor Svein-Olaf Hvasshovd, who not only have guided me and given me many motivating and encouraging words throughout, but who also put his trust in me and gave me the freedom to make this thesis my own. It has been exceptional!
I also want to thank the people at NIBIO and NINA. To Lise Grøva, Inger Hansen, Erling Meisingset and Jennifer Stien, thank you for all the input, feedback and time invested from you, and for making me better.
Lastly I would like to thank professor and co-supervisor Jon Andreas Støvneng and student adviser Brit Wenche Meland at the Institute of Physics, for all the help. You made this thesis possible.

"Not all those who wander are lost."
- *Bilbo Baggins*

Contents

Abstract i

Preface ii

Contents v

List of Figures vi

List of Tables vii

Abbreviations ix

1 Introduction 1
1.1 Motivation . . . 1
1.2 Project description . . . 2
1.3 Stakeholders . . . 3

2 Theory 5
2.1 Ethological theory . . . 5
2.1.1 Breed of sheep and its impact . . . 5
2.1.2 Hierarchy and herd behavior . . . 7
2.1.3 Habits and movement patterns of sheep . . . 7
2.1.4 Diseases and behavior when sick . . . 11
2.1.5 Herd reaction to predators . . . 11
2.2 Calculating with geodata . . . 12
2.3 Machine learning . . . 14
2.3.1 Supervised vs. unsupervised learning . . . 14
2.3.2 Feature engineering . . . 15
2.3.3 Clustering . . . 16
2.4 Statistical significance . . . 19
2.4.1 The central limit theorem . . . 19

3 Methods 21
3.1 Work method . . . 21
3.1.1 CRISP-DM and the agile work process . . . 21
3.1.2 Gantt chart and work progression . . . 23
3.2 Software and libraries used . . . 23

3.3 Exploratory Data Analysis . . . 24
3.3.1 Visualization of data . . . 24
3.4 Data wrangling . . . 26
3.4.1 Handling ungenerated points . . . 26
3.4.2 Handling error in time . . . 26
3.4.3 Selecting universal time ranges . . . 27
3.4.4 Other data cleaning . . . 27
3.5 Feature engineering . . . 29
3.5.1 Generating new features . . . 29
3.5.2 Transforming temporal data . . . 32
3.6 Implementing the machine learning models . . . 34
3.6.1 K-means . . . 34
3.6.2 DBSCAN . . . 35
3.6.3 Other . . . 37
3.7 Two-tailed two sample t-test . . . 37

4 Results **39**
4.1 Statistical results and visualizations . . . 39
4.1.1 Feature correlation . . . 39
4.1.2 Statistics before data selection . . . 42
4.1.3 Statistics after data selection . . . 42
4.1.4 Map of trajectories . . . 46
4.1.5 Activity distribution . . . 47
4.2 Results of the machine learning models . . . 48
4.2.1 K-means . . . 48
4.2.2 DBSCAN . . . 51
4.3 Digital threshold markers . . . 55
4.3.1 Sheep movement against threshold values . . . 56

5 Discussion **59**
5.1 Data wrangling decisions . . . 59
5.2 Feature engineering decisions . . . 60
5.3 Statistical analysis . . . 60
5.3.1 Feature correlation . . . 60
5.3.2 Data sets after wrangling . . . 61
5.3.3 Description of normal sheep behavior . . . 62
5.4 Machine learning models . . . 64
5.4.1 K-means and diurnal behavioral patterns . . . 64
5.4.2 DBSCAN and atypical behavioral patterns . . . 66
5.4.3 Proposed digital threshold markers . . . 67
5.5 Future work . . . 69

6 Conclusions **71**

References **73**

Appendices: **79**

A - Code **80**

B - Specialization Project **82**

C - Statistics before the data cut **83**

D - Feature correlation pairplot matrix **89**

E - K-means results 24-hour clock view **90**

F - K-means mean cluster feature values **91**

List of Figures

2.1.1 Hierarchical position of sheep 8
2.1.2 Diurnal rhythm of sheep, previous study 9
2.1.3 Diurnal sheep activity, previous study 10
2.2.1 Illustration of latitude and longitude 12

3.1.1 Gantt chart of work progress 23
3.5.1 Trajectory angle 30
3.5.2 Time as a cyclic feature 33
3.6.1 K-means elbow method on velocity and time 34
3.6.2 K-means elbow method for all numerical features 35
3.6.3 DBSCAN elbow method for all numerical features 36

4.1.1 Feature correlation heatmap 39
4.1.2 Feature correlation matrices part 1 40
4.1.3 Feature correlation matrix part 2 41
4.1.4 Data set sizes 43
4.1.5 Mean sheep velocity 44
4.1.6 Mean sheep altitude 45
4.1.7 Mean sheep inverse trajectory angle 45
4.1.8 Map of sheep range trajectories 46
4.1.9 Activity distribution 47
4.2.1 K-means velocity clustering 49
4.2.2 Polar line plot for k-means clustering 50
4.2.3 Polar line plot for DBSCAN clustering for all data 52
4.2.4 Polar line plot for DBSCAN noise points 53
4.3.1 Mother-lamb velocity against threshold value 57
4.3.2 Mother-lamb altitude against threshold value 57
4.3.3 Temperature against threshold value 58
4.3.4 Mother-lamb inverse angle against threshold value 58

C.1 Data set sizes 83
C.2 Mean activity per hour 84
C.3 Mean activity per hour outliers 85
C.4 Mean activity per year per date in Fosen 86
C.5 Mean activity per year per date in Tingvoll 87
D.1 Feature correlation pairplot matrix 89

E.1 K-means velocity clustering 90

List of Tables

1.1.1 Lost sheep per county in Norway in 2021 2

3.4.1 Selected time ranges for all data 28
3.6.1 Epsilon values for DBSCAN for all data split configurations 36

4.1.1 Dynamic feature statistics with outliers 42
4.2.1 Feature cluster statistics from k-means 51
4.2.2 Dynamic feature statistics from DBSCAN noise points 53
4.2.3 Dynamic feature statistics from DBSCAN split data noise points . . 54
4.3.1 P-values for the statistical significance of data split differences . . . 55
4.3.2 Digital threshold values 56

C.1 Statistics on sheep velocity 88
F.1 Mean feature cluster statistics from k-means 91

Abbreviations

List of all abbreviations in alphabetic order:

- **API** Application Programming Interface
- **CRISP-DM** CRoss Industry Standard Process for Data Mining
- **EDA** Exploratory Data Analysis
- **GIS** Geographic Information System
- **GNNS** Global Navigation Satellite System
- **IQR** Interquartile range
- **Mamsl** meter above mean sea level
- **NIBIO** Norwegian Institute of Bioeconomy
- **NINA** Norwegian Institute for Nature Research
- **NKS** Norsk Kvit Sau (Norwegian White Sheep)
- **NTNU** Norwegian University of Science and Technology
- **PCA** Principal Component Analysis

Chapter 1

Introduction

1.1 Motivation

The sheep industry in Norway consists of over 13 thousand sheep farms, where every year around two million sheep will be released on rangeland pastures during the summer to graze [1, 2]. While out on the outfield rangelands the sheep herds will move freely over large areas, and be for most of the time unsupervised except for occasional check-ups by the farmers. More than 30 thousand sheep were reported lost due to predators in 2021, resulting in over 44 million NOK in compensation claims [3]. Sheep are also lost due to causes like diseases, accidents, parasites or eating poisonous plants. In table 1.1.1 the reported amount of lost lambs and adult sheep per county in Norway in 2021 is presented, where the areas with the highest loss percentages are Innlandet and Trøndelag. Sheep lost while on rangeland pastures will not only have economical consequences for both the farmer and the government, but there is also an important factor of ensuring animal welfare. Without supervision sheep may be sick or hurt for a long time before the farmer can help, and they may not become aware of the problem at all before it is too late. In recent years modern agricultural technology has been used in the form of electronic collars on a selection of sheep within herds, to track their movement while out on pastures. These collars will send their UTM position, and may have other sensors installed as well to track other conditions about the sheep. Optimizing the collar systems and furthering the technology may substantially help with measuring the behavior of the sheep, and thus in extent also to assess their welfare. Should the sheep movement pattern diverge from the typical and expected behavior, this might give indication that something is wrong. Detecting this as early as possible gives more time to fix the problem, help the sheep in need, ensure a higher standard of animal welfare and give long term economical gain. The sheep loss may thus decrease. Identifying atypical behaviour in sheep on rangeland grazing pastures is therefore in demand to ensure welfare and sustainability in both Norwegian and worldwide pastoral systems. The access to high density movement data from numerous electronic collars facilitates the possibility to collect behavioural information at both individual and herd level. Knowledge of typical behaviour patterns and deviations from these has the potential to provide a tool to detect atypical behaviour in real time, that may be caused by for example predatory attacks or disease.

County	Lost lambs	Lost sheep
Agder	793	134
Innlandet	7648	1508
Møre og Romsdal	1433	156
Nordland	4405	983
Oslo	0	0
Rogaland	95	16
Troms og Finnmark	2453	541
Trøndelag	6578	1491
Vestfold og Telemark	985	167
Vestland	948	259
Viken	1216	298

Table 1.1.1: Lost sheep per county in Norway in 2021, data collected from *Rovbase* [3].

1.2 Project description

This master's thesis was given by the Norwegian University of Science and Technology (NTNU), in collaboration with the Norwegian Institute of Bioeconomy (NIBIO) and the Norwegian Institute of Nature Research (NINA). The data used in the analysis was given by NIBIO and NINA, and consists of coordinate positions on different herds on rangeland pastures in Fosen, Møre og Romsdal and Tingvoll, Trøndelag. The goal of the analysis was to investigate and describe the normal behavior of sheep on rangeland pastures based on digital data, and to examine if it is possible to determine atypical sheep behavior using machine learning and data mining. It is expected behavior for sheep to flee from predators while out on rangeland pastures, so normal behavior embraces both typical habitual and calm demeanor, and more atypical behavior in e.g. extreme movement perhaps because of predators. Atypical behavior is within the normal spectre, but defined as the less common and more irregular movement patterns detected. The project aims to be able to recommend further development within the technology of electronic collars for sheep, by looking into what properties that may be the most important to represent atypical sheep behavior and if it is possible to quantitatively describe them. The analytical methods and strategies to solve the problem were done with guidance and recommendation from the supervisors from NTNU, NIBIO and NINA.

The available data did not have many cases of known target situations where anti-predatory or sick behavior occurred. The machine learning will therefore be unsupervised, and it will not be possible to check the accuracy of the model predictions or verify the results without further research. The analysis still aims

to form a basis on which future studies may build on, and work as a possible starting point in examining the development possibilities of threshold markers in livestock collar technology.

1.3 Stakeholders

A. Sheep farmers: From both an ethical and economical perspective, sheep farmers will have a high motivation to to reduce the amount of lost sheep each year. Both time and resources from the farmer are allocated to ensuring animal welfare for the sheep while on rangeland pastures, locating lost individuals and reporting any deceased sheep. An improved digital solution might simplify and give more control over all tasks.

B. The government: The government has several departments and agencies that handles livestock welfare and losses, and compensation claims when sheep are lost due to predators. There may be a large economical gain in reducing the amount of lost sheep and by lessening the workload for the government employees associated with better sheep welfare.

C. Producers of electronic sheep collars: Providers of electronic collars for livestock will have an interest in bettering their products and service. The solutions on the marked today will sometimes have a type of stress detector from excessive movement, and the research done in this study may further the accuracy and scope of realizable parameters incorporated in the collars.

D. NIBIO and NINA: Both collaborating institutes on this project are analyzing and researching digital solutions and future possibilities within agricultural technology daily. Specialized knowledge and skills from students doing detailed research may help to advance and develop their innovations.

Chapter 2

Theory

The ethological behavior of sheep and the theory of calculating with geographical data were studied in the project thesis. The following sections 2.1 and 2.2 are partially excerpts from the findings stated in the specialization project report [4], with some changes and additions to better fit the analysis in the master thesis. The project thesis can be found linked in appendix B.

2.1 Ethological theory

There are many factors that may influence the behavior of how a sheep acts naturally when out on rangeland pastures, without the security of constant human protection. A special field of interest in this thesis is how sheep behave when there is an indication that something is wrong, compared to their typical, relaxed behavior. Since the end goal is to differentiate these two, and hopefully be able to see tendencies of this in the data, it is very important to understand what to look for, and what affects the data. Parameters that may influence the nature of the sheep behavior include type of breed, hierarchy order, environment and weather, age and gender, available resources and individual personality. The findings on factors influencing the temperament and movement pattern of sheep are presented below.

2.1.1 Breed of sheep and its impact

Observations clearly state that the breed of sheep will influence their behavior, especially on rangeland pastures where they can roam freely over large areas uninterrupted by human interaction [5, 6]. Sheep have been domesticated and bred by humans for at least 7000 years [7], and trait selection has long been a part of breeding the animal. Especially in more modern times, when sheep farming has gradually evolved from being a livelihood within a family, to be an international economical and sometimes industrial occupation. Traits that maximize profits and streamlines operations are often more wanted and bred upon. This includes traits like growing better quality wool, more muscle for more meat per sheep, less flocking, and being more friendly and social towards humans [5]. This will in turn affect their survival instincts on rangeland pastures, and make them more dependent on humans as a species. Research indicates that the more bred upon

a sheep breed is, the more tame and heavier it will be, and the sheep might thus have weakened instincts to help guard themselves against predators [8]. However, breed differences and their antipredatorial behaviors are generally poorly documented.

2.1.1.1 Description of relevant Norwegian breeds

Spæl: Originating from one of the oldest breeds in Norway, the Spæl sheep is still considered to have a close resemblance in manner and physical features to the earliest domesticated Norwegian sheep. It has been somewhat crossbred during the first half of the 20^{th} century creating different kinds of Spæl. The main types are Old Norwegian Spæl, White Spæl and Colored Spæl. Spæl has less meat content and fat than most other breeds, but has stronger mother instincts and milks very well. Despite its slightly smaller and fragile build, it is light on its feet and robust, more vigilant and runs fast. While on pasture they gather in a herd [9, 10].

Old Norwegian: Also called Wild sheep or Stone age sheep, like Spæl it stems from the old original Norwegian breed. It is small but highly cautious and sturdy, with strong mother and herd instincts.

Norwegian Fur: A relatively new breed from the 1960s, and slightly larger and heavier than Spæl. It has good mother instincts, and will roam widely on pastures.

Dala: One of the largest breeds, and has a calm, tame demeanor. It has some difficulties with breeding, and will often scatter more on pastures.

Norwegian White Sheep (NKS): This is a crossbred sheep consistent of different both Norwegian and foreign types. Lately some of the other Norwegian breeds like Dala, Rygja and Steigar are sometimes being classified as NKS instead due to crossbreeding. It is a fast growing and heavy breed, very fertile and social towards humans, making it the most populous breed in Norway. It will often walk separate on pastures and not flock together.

Grey Trønder: A cross between Old Norwegian and Tauter from Trøndelag, with a medium weight. It is alert, frugile and has good herd instincts.

2.1.1.2 Reactions to predators by breed

Research on the antipredatorial reaction in ewes of several typical Norwegian breeds were done in Trøndelag [8]. The breeds were Old Norwegian Sheep, Spæl, Norwegian Fur Sheep, Suffolk, Steigar Sheep, and Dala sheep. Their reaction was recorded against the threats stuffed wolverines, lynx, and bears, humans and dogs, and big unfamiliar objects. This study found that lighter breeds have stronger instincts and reactions against predators than heavier breeds. The breed that had the longest de-reaction time, longest flee distance, most defensive behavior, and flocked the most together was the Old Norwegian Sheep. After that came Spæl and then Fur sheep, and the Fur Sheep was also the most offensive towards the predator types. The rest of the breeds had no to little differences in reactions

between them. Steigar Sheep flocked the least together when threatened. The predator figures that initiated the longest de-reaction times were in decreasing order dog, lynx, wolverine and bear. The ones that stimulated the longest flee distance were in decreasing order wolverine, bear, lynx, and then dog. Human and large objects fell significantly below the other threats in reactions. There were also stronger reactions during the fall experiment than the one in spring time.

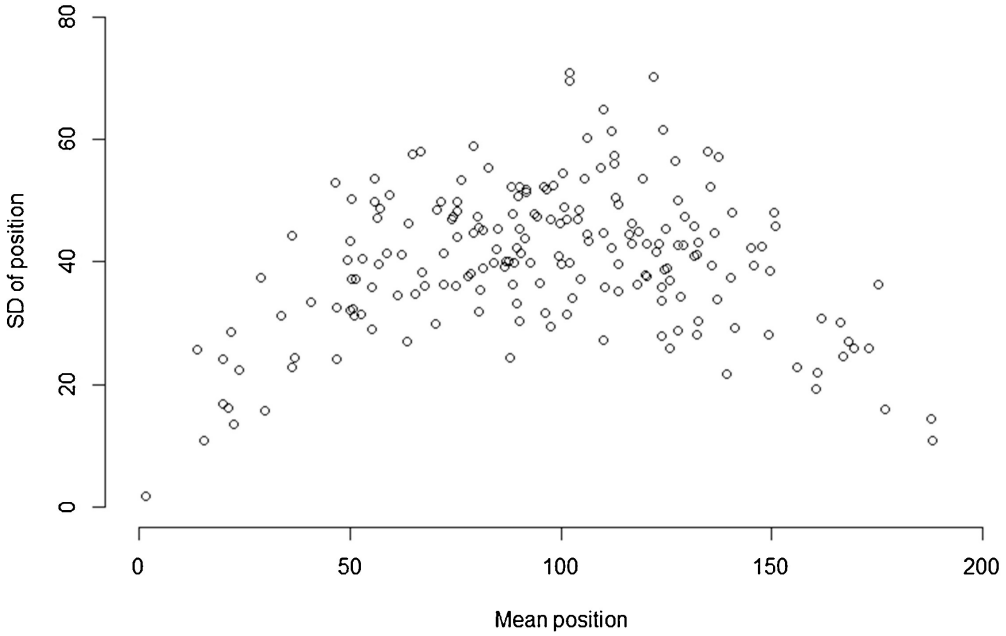
2.1.2 Hierarchy and herd behavior

When referring to a herd in this thesis what is meant is the whole collection of sheep belonging to a farmer, while when referring to a flock it is meant a smaller portion of the herd, like a family group. When on rangeland pastures, the adult sheep are usually only ewes. In a study conducted at the University of New England, Australia, herd position during movement and hierarchical and temperamental differences were examined in adult Merino ewes [11]. The movement was enforced by high value food motivation along a track. Every sheep was given a lameness score each test day, to map their positional variance against their health status and mood. The research show strong correlation in especially preceding runs, and hints towards strong front- and rear hierarchical order. In figure 2.1.1a, the mean position of each sheep relative each other and the standard deviation of their position are given. The curve plainly shows that the further to the front or back in the herd an ewe is located, the more consistent their relative position is. For the sheep in the middle of the herd, the standard deviation is both higher and varies more, which may indicate that hierarchy is mostly important for sheep with either strong leadership or follower traits. The study also found that for days when sheep showed lameness, their positional mean were 20.5 ± 5 % further back than on days they were not deemed lame. Tømmerberg (1985) followed a herd of Dala sheep over two years on rangeland pastures, and recorded their rank and compared it against their age, as seen in figure 2.1.1b [12]. Here there is a correlation between older age giving higher rank in the herd. He also found that once hierarchy was established in spring, this stayed constant for the whole season out on the pastures. Similar results for Border Leicester sheep and other Merinos have also been found, but more extensive research is needed in order to conclude generally about the spatial leadership for sheep [13, 14].

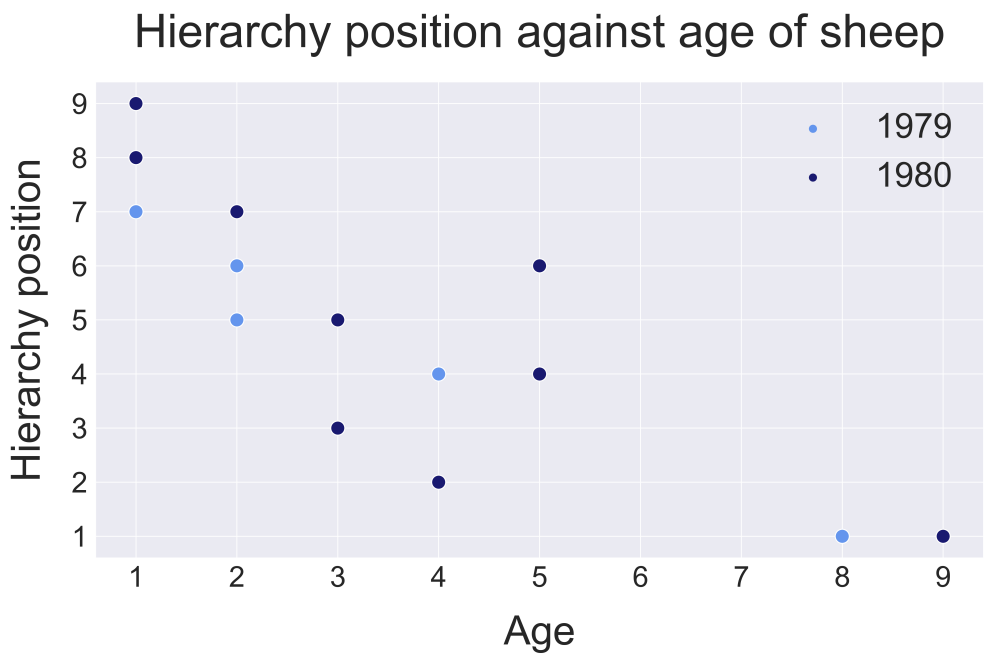
2.1.3 Habits and movement patterns of sheep

2.1.3.1 Home range

Sheep are habitual animals, and when roaming on rangeland pastures they will mostly keep to a fixed area each year called home range [9]. William Burt (1943) defined it as "...the area, usually around a home site, over which the animal normally travels in search of food. Territory is the protected part of the home range, be it the entire home range or only the nest." [15]. The extent of the home range will vary with grazing qualities, topography, predators, and quantity of other grazing animals. Tømmerberg (1985) found that the home range increased in size later in the season, which may be explained by that the lambs are older and need more nutrition, and vegetation grows more slowly towards fall [12]. Each individual sheep or small family group will have their own home range, and will



(a) Hierarchical position of sheep in movement



(b) Hierarchical position by age of sheep

Figure 2.1.1: Figure (a) shows individual mean position of each sheep (n=196) for all runs and their standard deviation (SD) (figure taken from Doughty et al. [11]), and (b) shows hierarchy position of one herd over two years compared to age, where a low position means higher rank (figure taken from Tømmerberg [12].)

also follow more or less the same route every year. This route will pass down from ewe to daughter, and stay approximately consistent throughout the generations [12]. Trying to change pasture may prove difficult, as the sheep often will return to the area they grew up with. If the home range of a sheep is changed it is most likely due to disturbances, like predators or human intervention [16].

2.1.3.2 Reactions to different types of weather

The movement pattern of sheep will be affected by the weather. On sunny days they are more likely to be in higher altitudes, probably to seek more wind and cooling temperatures [9]. Likewise, they keep to lower grounds if the weather is cold or rough. They roam the most on dry, cloudy days, while staying more put if it is either very hot or stormy. Their behavior is not influenced by a normal amount of rain, but should it rain a lot they will usually find shelter in the form of vegetation or natural formations. When there are a lot of insects, typically when sunny, hot and calm winds, they will move onto higher grounds to try to find more wind to get relief from the nuisance. One behavior to note here is that when bothered by insects sheep will regularly shake their heads and fidget more, which may affect the GPS positioning. At very hot days, grazing at night will increase while the sheep instead will relax more during the day. Scott and Sutherland (1981) found for Merino sheep that grazing were at a maximum around 10-15°C, and relaxing and shade group formations increased noticeably at temperatures above this [17].

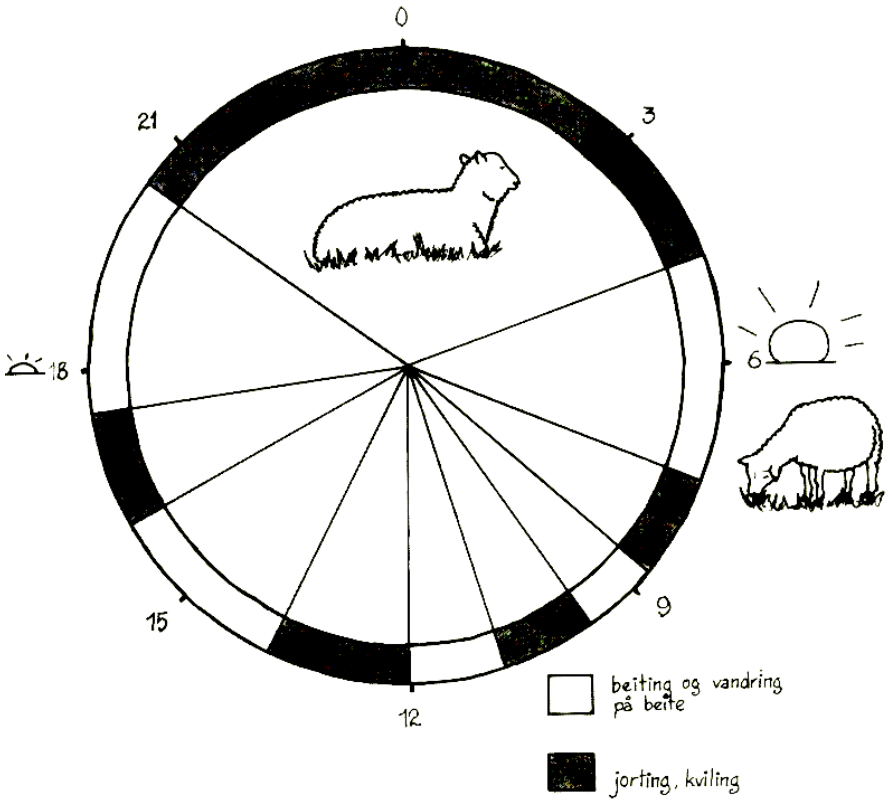


Figure 2.1.2: Schematic breakdown of the diurnal rhythm of sheep while on rangeland pastures, figure retrieved from *Saueboka* [9]. The white segments represent grazing and moving about, while the black segments represent resting and chewing cud.

2.1.3.3 Diurnal routines

The major part of the day for a sheep is spent alternating between grazing and chewing cud, as can be seen in figure 2.1.2. Grazing will occupy up to 7-11 hours

each day, and cud chewing about 5-9 hours [18, 9]. This depends on the quality of the pasture relative the amount of sheep and time of the season. The most active grazing periods will be at dawn and at dusk, while the sheep will rest during the early afternoon when temperatures and sun are peaking [12, 17, 19]. Towards night the herd will usually seek higher up in the terrain, and come down again in the morning.

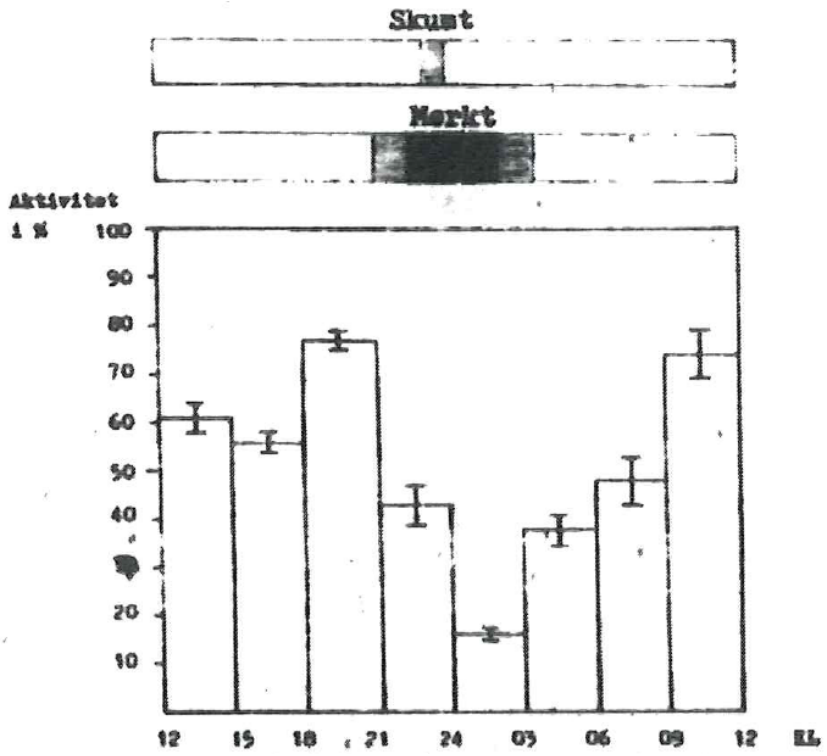


Figure 2.1.3: The diurnal sheep activity measured in percentage, from the study of Tømmerberg [12]. The activity is registered from mid-July to mid-August on adult ewes with lambs.

2.1.3.4 Seasonal habits

Typically the sheep will be released on rangeland pastures around the end of May/beginning of June, and be gathered home again in the middle of September. Depending on the breed, a herd will sometimes split into smaller flocks of family groups out on the pastures. As discussed in 2.1.1.1 the larger and more docile breeds like NKS will split more than e.g. Spæl. When a flock is grazing together they will often move in a wide formation forwards, but if a flock is migrating to another area they will walk in a row behind each other instead [9]. They seek secure places for when resting or chewing cud, either higher up or around vegetation. Their route throughout the season will often stop by mineral blocks if laid out. Migration of a flock is mostly motivated by better vegetation and grazing opportunities elsewhere. At the end of the season a herd will often return homewards by itself, from where it had been grazing. This will also often happen if it is scared by predators [20, 12]. Activity during the night in the late season significantly decreased compared to the early season according to Tømmerberg, which may be explained by darker and longer nights. The home range also increased later in the

season in his study. Around 90 % of this active time were used for grazing, the rest for walking, running, nursing et cetera.

2.1.4 Diseases and behavior when sick

When sheep get sick while on rangeland pastures, they may fall behind the flock, become lethargic, have less energy and generally move more slowly than when healthy. This could be because of several diseases, like scrapie, hypokalemia, coccidiosis, or anaplasmosis [9]. Alveld is a disease that arise by consumption of the plant bog asphodel, and will make the sheep sensitive to light and uneasy, and they will often seek shade. Lambs are typically more susceptible to get sick than adult ewes.

2.1.5 Herd reaction to predators

Tømmerberg observed the herd running upwards in the terrain when frightened by something, and hypothesized that the shift to higher altitudes at night is a measure to better be able to detect approaching predators when dark [12]. This argument is further supported by that this behavior largely followed the length of the day, where the move upwards started earlier in late summer when the sun set earlier. The given feedback from farmers stated that sheep will return either home to the farm if close enough, go to roads or cabins, or seek towards water when predators are near their pasture. They will be noticeably stressed and restless. One report from the Norwegian Institute for Nature Research in 2016 found an indication of correlation between heightened deaths because of golden eagles and deaths because of diseases and accidents on rangeland pastures, even though the data did not provide a basis to conclude [21]. It is worth noting that a prevalence of more diseases in a herd might cause easier prey and hence more predatory deaths, or that the presence of predators might increase the chances of diseases.

2.2 Calculating with geodata

Latitude and longitude are geographical coordinates, and in Geographic Information Systems (GIS) they are often given in decimal degrees, as were the case with the data from Tingvoll and Fosen. Latitude is the degrees in the north-south orientation, and thus represented by horizontal lines called parallels, and goes from 0° at the equator to ±90° at the poles. Longitude is the east-west orientation, and thus represented by vertical lines called meridians, and goes from 0° at the prime meridian in Greenwich to ±180° east/west [22]. Latitude and longitude are illustrated in figure 2.2.1. When moving along latitudinal lines with an unchanged longitude, the trajectory will slice the earth through the center along the meridian, but when moving along longitudinal lines the trajectory will depend on the latitudinal value.

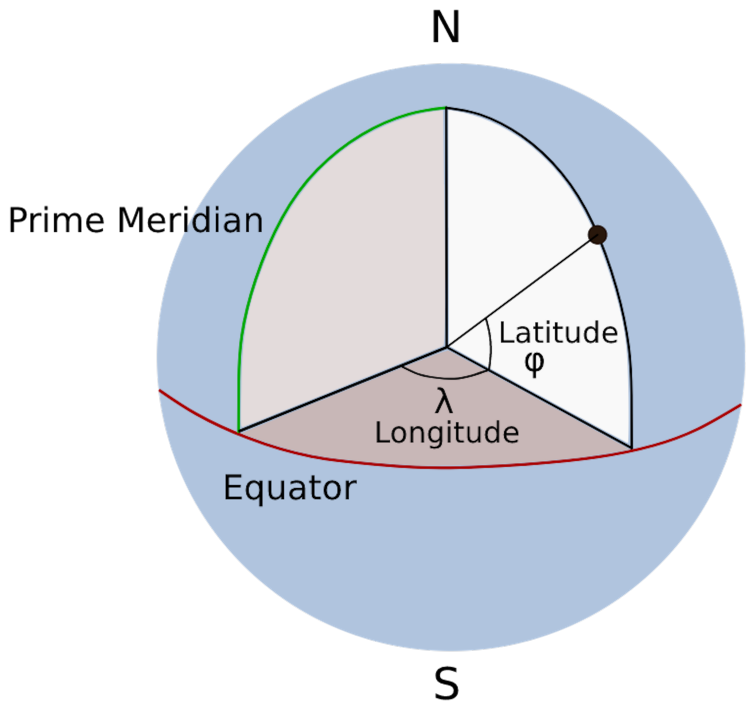


Figure 2.2.1: Illustration of the earth, and how latitudes and longitudes are calculated with respect to the equator and prime meridian.

The earth in itself is approximately spherical, but because the earth is spinning about a fixed axis though its center, the fictitious inertial centrifugal force compresses the middle of the earth outwards, and deforms the earth into an obloid spheroid sometimes called an ellipsoid of rotation [23]. This deformity affects how latitudes and longitudes are computed in a Global Navigation Satellite System (GNNS) and a GIS, which need a reference in approximately the same shape of the earth to calculate values. This reference ellipsoid for a given coordinate system should be specified as in accordance with the ISO 19111:2019 standard in order to be sufficiently accurate [24]. Most world wide services like Google Earth and OpenStreetMap, which are 3D geographical coordinate systems that can chart coordinates, uses the WGS84 geodetic reference frame (datum) to define the ellipsoid reference of earth [25, 26]. The identifier used to chart coordinate values on the

WGS84 ellipsoid is EPSG 4326 [27]. Note that this is in relation to 3D-mapping of coordinates to locations, when applied to the flat 2D plane on projected coordinate systems like maps, the identifier EPSG 3857 for coordinate transformations is used for WGS84 [28]. In Norway and other European countries, the EUREF89 reference frame is also used, which is based on WGS84 and Europe's tectonic position in 1989. Because of continental drift, Europe and America moves a couple of centimeters apart each year, increasing the difference between EUREF89 and WGS84 [29]. For the data from Fosen and Tingvoll, the exact GNNS and reference frames are unknown, which may affect results in especially orthometric height (altitude) calculation. The orthometric height is the vertical height from a given point to the geodetic reference frame, approximating the meters above the mean sea level [30]. For this project it will be assumed that GPS and WGS84 are used, based on probability, but the difference from other standards are not deemed significant. Further on, the major part of this project will depend on positional differences and qualitative analysis, not exact location, and therefore be mostly independent of reference frame as long as the calculations are consistent.

If the distance is less than about 200 km, the earth may be assumed spherical when calculating the length between points [31]. The Haversine formula given in equation (2.1) is a method used to calculate the distance traveled by coordinate values and taking into account the curvature of a spherical surface [32]. The Haversine is defined as $hav(\theta) = \sin^2(\theta/2)$, R is the sphere radius and $\Delta\phi$ and $\Delta\lambda$ are the differences in latitude and longitude, respectively. This formula is an approximation, because of the obloid curvature of the earth the Haversine formula may still have an error up to 0.5 % [33]. Vincenty's formula will take into account the reference frame, most commonly WGS84, but it is computationally heavy, and for a qualitative analysis on relatively small distances not necessary [34].

$$\begin{aligned} a &= \sin^2(\Delta\phi/2) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2(\Delta\lambda/2) \\ d &= 2R \cdot \arcsin(\sqrt{a}) \end{aligned} \tag{2.1}$$

Some error in GNNS values must be assumed, and could be caused by for example signal noise, clock offsets between satellites, atmospheric conditions, or attitude in orbit inclination. But for e.g. the American system GPS, this fault will only account for a couple of meters [35].

2.3 Machine learning

Machine learning is a subcategory of artificial intelligence which applies statistical analysis and mathematical models on data in order to make a computer detect trends, or make decisions or predictions [36]. An algorithm is said to learn if it can make decisions based on empirical data without being explicitly programmed on how to do so. The concept was formally defined by Tom Mitchell as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [37]. A well-posed learning problem therefore must be able to identify the three features (1) the tasks, (2) the measure of performance, and (3) the training experience.

A machine learning program may take into account many different input variables, which are called the features or attributes. The parameters of the model are internally configured variables which are tuned and adjusted as the model learns during training, and hyperparameters are coefficients externally predefined to guide the learning process [38]. The goal is to optimize both parameters and hyperparameters so the model will be specialized enough to discover the pattern but general enough to accurately predict or classify unseen data. If a model is made too complex it may learn to incorporate the noise and errors in the data as part of the structure. This is called overfitting. On the other hand the model will be underfitted if it is not trained enough or too rudimentary to detect a correlation in the data. Tactics against under- and overfitting may be to get more training data, stopping the training process early, feature selection and regularization [38].

2.3.1 Supervised vs. unsupervised learning

In supervised learning the target values to be predicted, called labels, will be known when training a machine learning model. The values of all features of one point x^i, a so-called example of data, will correspond to one output variable label y^i. The algorithm will learn through experience using a data set with known labels, and tune the parameters in the model in order to optimize its predicting performance on new data [39]. Labels for data with unknown output values can then be predicted with no other information available other than the example values and the tuned model parameters. Accuracy for the model may then be tested by splitting the data into train and test sets, where you predict labels for the test set after tuning the model with he train set, and compare the predicted values with the real ones. With unsupervised learning the target labels are not known, and the main objective is rather to find underlying structure and dependencies in the data. The accuracy of the model thus cannot be directly calculated, and other methods must be used to have a metric of the model performance.

For a supervised model the training variance and test error may be calculated. If it is underfit both will be high, while if overfit the variance will be very low while the test error will be high [38]. For unsupervised models there is no way of calculating an error without true labels, and avoiding under- and overfitting is rather reduced into a problem of e.g. deciding the optimal number of clusters.

2.3.2 Feature engineering

Feature engineering involves using domain knowledge, intuition and mathematical computation to transform the data into feature representations better suitable to portray the underlying problem for the model. It is done after the data cleaning process on the raw data, and gives a better quality input to the algorithm as to improve the model accuracy and performance. It is one of the most important steps in a machine learning pipeline, as the selection and form of the input will have a major impact on the model performance, often more than the algorithm itself [38].

2.3.2.1 Feature generation and data transformations

Raw data will not always include all variables relevant to the problem the model aims to solve, and constructing new features will often be necessary. Creating new features may be done by transforming existing data from one or more features, or by aggregating data from outside sources. Sufficient domain knowledge is crucial for feature generation in order to know what type and form of information that will affect the issue, and thus knowing what is most likely needed for the model. This includes for example binning from continuous variables to categorical, by creating discrete ranges to place each value in order to improve the model performance.

How each feature is represented will have a major impact on the results of the algorithm [38]. It must be in a form the machine learning program understands and that make sense computationally, and physically. The data must therefore sometimes be transformed or converted into another, more suitable shape.

2.3.2.2 Scaling

The different input variables in a data set will most likely have a wide range of both values and units depending on what they are describing. For example the age of a group of people versus their income level will be of a quite dissimilar order of magnitude. Algorithms in machine learning will often employ a distance metric between features to calculate similarities in the data, and critically disparate value intervals might give a disproportionate bias and severely affect the outcome. Features with higher values will be emphasized more, even though they might not be more important for the model compared to other features of different ranges and units of measurement. Feature scaling is therefore needed to make divergent attribute values comparable. This is done with normalization (Min-Max Scaling) and standardization (Z-Score Normalization) [38].

Min-Max Scaling, given by equation (2.2a), transforms all the data to a similar scale so that their range becomes bounded by [a, b] [40]. Here a is the minimum and b the maximum, typically [0, 1]. Normalization simply rescales the range, and does not change the relative positioning of any outlier points. Z-score, given by equation (2.2b), shifts the features such that they follow normal distribution rules, which means that the mean is zero and the standard deviation is one [41]. The z-score is calculated by subtracting the mean μ and dividing by the standard deviation σ. This ensures that variables of different types of measurement

are comparable. Normalization and standardization transformations are not always needed, but useful when the features have contrasting characteristics and/or especially when distance metrics are used in the model calculations [38].

$$\text{Min-Max: } x_{i,new} = (b-a)\frac{x_i - x_{min}}{x_{max} - x_{min}} + a, \quad \in [a,b] \tag{2.2a}$$

$$\text{Z-Score: } x_{i,new} = \frac{x_i - \mu_x}{\sigma_x} \tag{2.2b}$$

2.3.3 Clustering

The technique of clustering in machine learning is an unsupervised method and uses the notion that a set of observed and abstract points will be more similar to each other the closer they are [38]. Given a set $S = \{x_1, x_2, ..., x_n\}$ of $n \geq 2$ heterogeneous points, and the pairwise distances between them given by the distance function $d(x_i, x_j)$ where $x_i, x_j \in S$, the points may be grouped together based on the similarity of closeness. The resulting partitions $T_1, T_2, ..., T_k \subset S$ of the grouping function f are called clusters, where the points belonging to the same cluster are closer together than to the points in different clusters [42]. The cluster assumption states that if points are in the same cluster, they are more likely to be in the same class [39].

2.3.3.1 *K*-means

k-means is a method for clustering data into k non-overlapping clusters, where each data point belong to the cluster with the smallest mean distance to the cluster center (centroid) [43, 44]. The goal is thus to minimize the intra-cluster variance, while the inter-cluster variance is maximized [45]. The algorithm for k-means is as follows:

- Specify the number of clusters k.
- Initialize the cluster centroids either by randomly selecting seeds or by the use of an optimization method.
- Calculate the sum of squared distance between all data points and all clusters, and assign each point to the nearest cluster center.
- Compute new centroids by finding the mean point of all the data within each cluster.
- Repeat the two previous steps until convergence or until a stopping condition is met.

The sum of squared distance (SSD) is given by equation (2.3a) using euclidean distance, where μ_q is the centroid for cluster q. The sum of squared distance will give the cluster variance as it computes the square deviation of each example

from the mean, and may also be used to measure the total error since the intra-distance will rate the dissimilarity for each cluster [43]. This is therefore sometimes shortened to SSE (sum of square error) as given in equation (2.3b), where $r = 1, 2, ..., k$ describes which cluster point x_i belongs to and δ_{rq} is the Kronecker delta.

$$SSD = \sum_{i=1}^{n} (\vec{x_i} - \vec{\mu_q})^2 \tag{2.3a}$$

$$SSE = \sum_{q=1}^{k} \delta_{rq} SSD \tag{2.3b}$$

To optimize k-means one may use an improved initialization algorithm, such as k-means++. This chooses the initial centroid values, or seeds, to set a better starting point for the rest of the k-means algorithm [46]. K-means++ works by first randomly selecting the first centroid from the data points available, it then computes the distance between the chosen centroid and all other points, and selects the next centroid as the point with the maximum distance to the previously chosen centroids. The last step is repeated until k centroids have been chosen.

2.3.3.2 The elbow method

To check the model validity and performance of the k-means clustering, the elbow method will be used. By plotting the explained variation (inertia) against the number of clusters, a cut-off cluster value is given at the "elbow" of the curve. The curve will linearly level out at higher numbers of clusters after this point, which means adding an additional class will not necessarily yield a much better model of the data. It is a heuristic and does not guarantee the global optimal solution, but will help get satisfactory results where the model will most likely be accurate enough at an acceptable cost and where k-means will at least terminate at a local optimum [47]. The reasoning behind the concept of the elbow method is that with more classes available the fit will naturally improve and explain more of the variation as more compact clusters can be used, but at some k the model will be overfitted. SSE is used as the measure of explained variance. By implementing the elbow method to choose the number of classes to include in the model, over- and underfitting is hopefully avoided and validity strongly suggested [48].

2.3.3.3 DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise, and will cluster data together while also detecting outliers [49]. It will group points together that are close and with many nearby neighbours, while marking points that are alone in low-density regions as anomalies. The clustering will therefore not be particularly affected by outliers [50]. DBSCAN characterizes all the data into three different types of points, i.e. core points, boundary points and noise

points [51]. This is based on the number of neighbours a point have within a given neighbourhood. The neighbourhood of a point is defined as the distance radius of the point, which is a hyperparameter decided by the user, and the number of points within the circle spanned by the radius is defined as the point neighbours. A core point will have at at least as many neighbouring points as a given minimal value also given by the user, counting itself. A boundary point does not fulfill the criteria of enough neighbours, but will be in the neighbourhood of a core point. Noise points are all the the rest of the data that do not fit into either of the two other categories. Further, a point p is said to be density connected to point q if there exists a chained set of core points, connected through their neighbourhoods, such that p and q are each in the neighbourhood of a core point in the set. The DBSCAN algorithm works as follows:

- Categorize all the data into either core, boundary or noise points by the given hyperparameters.
- Remove all noise points from further consideration.
- Allocate a cluster to a core point, and assign all other density connected points to the same cluster.
- Repeat the last step until all core points have been assigned to a cluster.
- Assign each boundary point to the cluster of the nearest core point.

2.3.3.4 K-Nearest Neighbours

Likewise the elbow method for k-means, an optimization algorithm will be used to optimize and choose the hyperparameter value for the point radius ϵ in DBSCAN. As proposed by the original authors of DBSCAN, Sander et al. 1998 [50], the method of K-Nearest Neighbours (K-NN) is used. It calculates the euclidean distances to the K nearest neighbours of each point, sorts the distances from minimal to maximal value and plots them to find the elbow of the curve. This will give a parameter radius distance ϵ which will separate the majority of the intra-cluster distances from the noise points further apart from the rest. A heuristic to decide the input variable number of neighbours K in the K-NN algorithm was proposed by Sander et al. to be K$= 2 * Dim - 1$, where Dim is the dimensionality of the data, i.e. the number of independent variables (or features) present. Further, the hyperparameter for the minimal amount of neighbours needed to be considered a core point in DBSCAN, $minPts$, needs to be decided by domain knowledge. Generally it should be greater than or equal to the dimension, and be chosen larger for larger and more noisy data sets [50].

2.4 Statistical significance

To validate hypotheses made from the data driven results, the statistical significance has to be computed. It is used as a test to check whether the results obtained are likely due to chance alone or due to an actual global relationship between two or more variables [52]. The method calculates through statistical hypothesis testing the p-value of the results (or the sample), which signifies the probability of observing as extreme values from the data if it is presumed that the results actually were only because of chance [53]. The null hypothesis H_0 represents the notion that there exists no significant relationship between the data and that any observed difference is due to chance, and the alternative hypothesis H_A is some form of opposite to H_0, often the assumed hypothesis made by the researchers that they are studying if is true [54]. Hypothesis testing determines whether to reject the null hypothesis or not, and thus prove that the sample of a situation that were studied can be extrapolated to the entire population and be globally generalized also outside the sample. Most hypothesis tests assumes a normal distribution in the data analysed. A study is considered statistically significant if its p-value is less than the pre-defined significance level α, typically set to 5 %.

2.4.1 The central limit theorem

The central limit theorem demonstrates that when a large enough normalized random sample is taken from a population, the sample mean will incline towards a normal distribution even when the population is not normally distributed [55]. An adequate sample size for the theorem to hold will be around 30-50 samples [56].

Chapter 3

Methods

3.1 Work method

A smooth and strategic workflow is essential for obtaining legitimate and reliable results, and to keep the research process effective, systematic and logical. It is important to utilize a work method appropriate for the problems to be solved and that fits to fulfill the objectives of the study. The process model chosen for this thesis was the CRoss Industry Standard Process for Data Mining (CRISP-DM) [57]. The method is suitable for data mining projects, and is well documented and tested [38]. CRISP-DM provides the workflow and necessary steps to execute an analytical project through iterative phases. The analysis is expected to change direction and requirements as new discoveries are made, so the CRISP-DM model was paired with a flexible and agile project management.

3.1.1 CRISP-DM and the agile work process

A key principle of CRISP-DM involves prioritizing to understand the business of and gain domain knowledge about the field of research. This will be valuable when choosing how to solve the problems, give important insights that might affect the analysis, and help interpret and recognize results. The CRISP-DM method contains six phases in a cycle, where the discoveries made guide the requirements of the data project in an agile and iterative manner. As more knowledge is gained from the analysis, the workflow may need to be adjusted, and the different steps repeated. To use an agile development approach to the workflow implies to be adaptive of changes throughout the whole project, where the model design and implementation are interleaved and under constant development. Formal documentation is kept at a minimum until the end, as the solution is meant to be reassessed and changed often [58]. The six phases of the CRISP-DM method are described below.

1. Business understanding: The initial phase focuses on understanding the environment of the given problem, recognizing the business context, and determining the business goals and objectives. The first phase is crucial for later analysis, such that when defining the business requirements it is known how a data mining process can answer the problem and which success criteria to set. The business

understanding hence includes gaining domain knowledge relevant for the available data, knowing what the existing solutions to the problems are and what are in need of improvement. The first phase should be used to specify and concretize the issues, determining limitations or caveats, and to reduce the scope of the objectives if necessary. A tentative project plan should be made, in accordance with the principles of the six phases of CRISP-DM and the agile methodology. The business in this thesis refers to the industry of sheep farming and sheep welfare technology on rangeland pastures.

2. Data understanding: The second phase involves investigating the available data in detail, and taking the necessary steps to understand all the data before the process of analysing it can begin. This will include retrieving the data, describing all the different features, performing exploratory data analysis (EDA) and visualizing the attributes, and verifying the quality of the data at hand.

3. Data preparation: Data preparation is the major part of the data mining process, and will take up to 60-70 % of the total time of the project [38]. In order to prepare the data for the modelling in the later analysis, the data has to be cleaned, wrangled, transformed and curated to a convenient format. Errors, anomalies and missing values has to be checked upon and handled, new attributes might be constructed (feature generation), and the most important features are selected for further use in the model.

4. Modeling: The modeling phase is the process where the machine learning model is designed and built, evaluated and tuned. The model can then be run with the clean and formatted data to deliver results, and the model performance assessed with respect to the business objectives and success criteria. Previous steps may be revised and iterated until satisfactory results are achieved based on the domain knowledge.

5. Evaluation: The fifth phase involves a detailed assessment of the final models and the given results, evaluating the significance of the findings and analysing their relevance for the problem. An overall evaluation of the process should be done, and further work recommended. Whether the business understanding or the objectives need to be refined should be discussed, and necessary adaptations for any future iterations suggested.

6. Deployment: The final phase is to prepare the models for deployment. This entails to document the project and the results, suggest a plan for future monitoring and maintenance, and discuss how to validate the performance of the proposed solution. It should be clear how and when to update, replace or retire models after they have been deployed.

3.1.2 Gantt chart and work progression

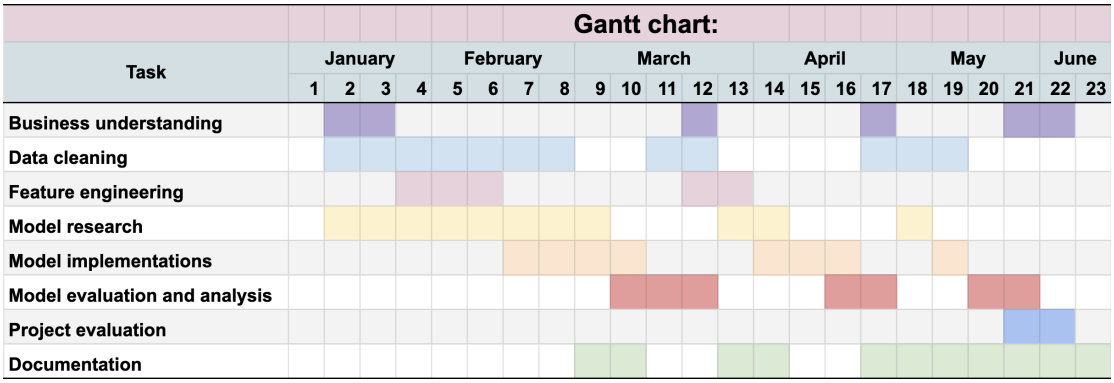


Figure 3.1.1: Gantt chart of the tentative progress plan of the project.

As described by the CRISP-DM phase one, a tentative project plan of the timeline was made as a Gantt chart shown in figure 3.1.1. The different tasks belonging to the different phases of CRISP-DM are planned in iterative sprints, as the model design and requirements are updated in accordance with new knowledge and revised business understanding. The plan is flexible and will change as needed, but will give a framework to keep the workflow effective and systematic.

3.2 Software and libraries used

All data on the sheep were given as comma separated files. The programming language *Python* was used for all coding, and the library *Pandas* was used for handling the data. *Pandas* is a software library especially well suited for data manipulation and data analysis, where data is represented as a *DataFrame* object for operating all information [59]. All data plots included in the thesis were made with the data visualization libraries *Matplotlib*, *Seaborn* and *Plotly*. Illustrations were made with *Inkscape* and *Adobe Photoshop*. The machine learning models were built with the open source library *scikit-learn* [60]. All the code files for the data cleaning, EDA, feature engineering and machine learning are collected in a Github repository, which can be found in appendix A.

3.3 Exploratory Data Analysis

An important step in data driven science, and in accordance with the CRISP-DM method, is understanding the data at hand. The data and its properties needs to be explored and visualized, to help understand what is going on in the data. It will also be useful when interpreting the results and to identify the possibilities in the later analysis. Examples of what to look into are how many attributes the data contains and what are their statistical properties, what do the different attributes look like and how is the data described.

Fosen:
The data from Fosen had 10 attributes in total, where the *datetime*-object, the sheep identification number, and the latitude and longitude values were used further. A sensor in the electronic collar of the sheep registered the temperature in real time, but an external and more reliable source was used instead for this information. There were a total of 583858 rows of data, distributed over 346 individual sheep trajectories, three separate farms and three years from 2018-2020. After data wrangling the final amount of individual sheep trajectories were 309, where 145 were of adult sheep and 164 were of lambs. The distribution of the different breeds of sheep were 193 NKS, 90 White Spæl, 1 Old Norwegian Spæl, and 25 Old Norwegian sheep. There were a total of 13 individual sheep who wore an electronic collar for two seasons, and three wore one thrice. Three sheep were confirmed killed by predators, one by golden eagle and the other two by unknown type.

Tingvoll:
The raw data retrieved from Tingvoll had 18 attributes in total, whereas only date, time, latitude and longitude continued to be used in the analysis. The attributes altitude and temperature, which were generated directly in the collar, were not of good enough quality to be reliable. These were thus later collected from external sources. The rest of the included attributes were not of interest in the analysis of sheep behavior, like battery power or collar status. The data from Tingvoll had a total of 444234 rows of data, distributed over 89 individual sheep trajectories, two separate farms and five years from 2012-2016. There were a final amount of 82 individual sheep trajectories after data wrangling, where all were adult sheep. The distribution of the different sheep breeds were 40 Grey Trønder sheep, 32 Spæl, 7 NKS and 3 not specified. Also for the Tingvoll data there were 13 sheep who wore an electronic collar for two seasons, and one wore it three times. There were no confirmed kills by predators in the data from Tingvoll.

For both locations there were given files with extra information on each sheep, including individual identification number, electronic collar id, age, number of lambs, breed, farm, and birth year. Some of the information on these files were used to supplement the data used in the final model.

3.3.1 Visualization of data

In the directory *EDA* in the Github repository lies all code files for visualizing and exploring the data. The activity of the sheep in the form of movement velocity be-

tween points were plotted in various ways, the size of all individual trajectory sets checked, the different home ranges of all farms explored against maps, a heatmap were made of the correlation between features and the inter-feature dependencies further investigated with pairplots. Additionally, since the start and end dates for the individual trajectory sets were somewhat dispersed they were plotted in a histogram to visualize and inspect the possible trends. The available information on all individual trajectory sets were also summarized and stored in csv-files.

3.4 Data wrangling

Before any analysis and visualizations can take place, the raw data has to be pre-processed and prepared properly. Data wrangling is the task of cleaning and transforming the material into more convenient formats. Instrumental data generated in real time retrieved from sensor systems will most likely have intermittent erroneous point generations. The format in which the data is produced may also not be suitable or appropriate for the later studies, and needs to be converted into a shape understandable for the machine learning models. Cleaning and processing the data will as aforementioned demand about 60-70 % of the total time spent on data analysis problems, and the wrangling process will iterate over time as new issues arise. Real life big data will seldom be perfect or without any faulty information included, the goal is to minimize the amount of defect data so that it is still useful and representative for the analysis. It is not always evident or obvious whether a point is faulty or just an extreme true observation. Since part of this study involves identifying atypical sheep behavior, outliers of extreme observations are key to recognize deviant movement patterns and it will be important to preserve this data in order to discover any would-be relationships between the variables at these instances. It is therefore essential to be careful, slightly restrictive and deliberate when deleting, imputing or altering any anomalous points.

3.4.1 Handling ungenerated points

In the data from Fosen there were occasional time stamps where no coordinates had managed to generate, but were set to zero instead. These faulty points were often at either the beginning or the end of each sheep trajectory. In the script *PointClean.py*, ungenerated points were handled. Since there seemed to be a connection with the first few points before and after switching the power of the electronic collars, the first and last five points of each trajectory set were deleted. Afterwards, all additional zeros at the end or the beginning of each set were deleted if there existed more over the five points just removed. Lastly, the faulty points in the middle of functional data were imputed with new values based on the mean of the two closest adjacent errorless points. If the consecutive sequence of faulty points were an odd number, the middle index were imputed with the mean of the existing outer points, and the upper and lower voids recursively filled with the mean of the newly imputed value and the existing upper and lower points, respectively. If the consecutive sequence were an even number the mean of the outer points were imputed into the lower of the two middle indices, and the rest filled recursively as with the odd sequence.

3.4.2 Handling error in time

For both data sets there were erroneous time stamps in the data, in various forms. All trajectory sets were for one season and should be valid only for one specific year, however sometimes the year in the time stamp attribute were different than the rest in the same set. The change of months should only increment the month in the time stamps upwards by one, and the day before the month changes should

also be either the 30^{th} or 31^{st} since the data is set to the summer months only. The days should likewise just increment upwards by one, except at the change of months, and the last clock time before changing a day should be close to 23:00. The time range between point generations were either an hour for the Fosen data or 30 minutes for the Tingvoll data. If the clock time difference between two adjacent points varied by more than a user set threshold, while the time difference with the next point over was less than the given threshold, then the clock time will be faulty for the middle point. Clock time also only goes forwards. In the script *Timeclean.py* all these instances of errors in time were handled. The user set threshold for clock time difference were set to four times the time interval between points. The faulty time stamps were also checked against the date and time two points over, in case it was a case of a long time off for the electronic collar and not an isolated time generation error. In the data from Tingvoll there were additionally cases of the same point being copied and stored twice in a row, so that one had to be deleted.

3.4.3 Selecting universal time ranges

The first and last days of the sheep trajectory sets in the data consisted of the sheep either being driven off towards outfield rangeland from the farm, or being collected down to the farm again in the fall. This data is not representative of normal free range sheep behavior and needs to be cut off. The mean activity per year per date for the data from both Fosen and Tingvoll are attached in appendix C4 and C5, respectively. Based on the slightly heightened activity in the beginning and end for all years, and the mean start and end date for all the individual trajectory sets, every trajectory set cut off the first and last few recorded days in the script *TimeInterval.py*. In addition, when comparing behavioral patterns it is advantageous that the time ranges are approximately equal, so that the data is comparable across years on the seasonal differences. Since the trajectory set time ranges were somewhat varied also within the same year, the range cut decision was based on trying to preserve most of the data while keeping the dates as universal across the years as possible. The selected time ranges for all data is presented in table 3.4.1.

Further a cut-off value of the individual trajectory set sizes were set at 10 % of the average size in the data from Fosen and Tingvoll separately. This was done to ensure a minimal amount of data for a set to be considered valid enough to be included in the analysis, and to remove faulty sets. Trajectory sets that were only slightly above the cut-off value were manually checked, and deleted if they looked defective.

3.4.4 Other data cleaning

The data from Tingvoll had twice the resolution as the data from Fosen, as a point were generated each half hour where in Fosen the time interval were only every hour. Sheep will change direction as they move, and may also turn back from where they came, and with a higher point resolution the values of the e.g. movement velocity will be affected and be more accurate. To be able to compare

Fosen	**Start date**	**End date**
2018	03.06	29.06
2019	03.06	03.07 or 31.08
2020	03.06	05.09
Tingvoll	**Start date (farm 1/2)**	**End date**
2012	09.06	07.09
2013	23.06 / 15.06	25.08
2014	05.06 / 25.06	10.09
2015	13.06 / 03.07	06.09
2016	17.06	22.07

Table 3.4.1: Selected time ranges for all individual trajectory sets in the data.

the two sets, they have to have the same resolution so that the features based on distance calculations are made with the same basis. Therefore the data resolution from Tingvoll had to be reduced by deleting every other point where the time interval difference were less than an hour. There were also 32 trajectory sets of data from Fosen that lacked extra information on them that had to be deleted as analysis on breed, age, and number of lambs thus could not be done. The last data wrangling done were to update the format of the data by deleting extraneous attributes and concatenating the two data sets together.

3.5 Feature engineering

3.5.1 Generating new features

The first new feature that was generated was the movement velocity of the sheep. In the script *Haversine.py*, the Haversine distance represented in meters was calculated between all points, given by equation (2.1). The distance was then divided by the time difference between the points represented in hours, to give the velocity in m/h. A threshold value of the maximal distance a sheep could run over the course of an hour had to be set, to filter out any unreasonable coordinates. It is important to not set this threshold too small, as the extreme cases of behavior are what this study aims to find and describe. It was assumed that a sheep might run a maximum of about 10 km/h, and to give some room to preserve any extreme cases the threshold was set at 15 km/h. This value was also compared to the range of unmodified velocity values, where the maximum was over 17 km/h. The data wrangling process thus started a new iteration, by cleaning the erroneous coordinates that resulted in a too high velocity. At the points where the maximum velocity threshold were exceeded, the coordinates were changed to the mean of the closest viable adjacent points and a new velocity were calculated.

Another feature created was the inverse trajectory angle. Two vectors were defined between three coordinate points A, B and C as $\vec{BA}$ between point A and B and $\vec{BC}$ between point B and C. The angle was then calculated using the dot product between the vectors, $\theta = arccos((\vec{BA} \cdot \vec{BC})/(|\vec{BA}| \cdot |\vec{BC}|))$. Thus the movement pattern of the sheep is included as a feature, giving either an acute and small trajectory angle as seen in figure 3.5.1 for the movement ABC" when changing direction, or an obtuse angle close to 180 degrees as seen for the movement ABC' when continuing straight. To make the extreme cases of directional change the highest in value, that is when the direction is changed towards the complete opposite, the inverse of the angle was made as the feature.

The altitude in the form of orthometric height for each coordinate pair had to be collected from an external source. The public government service for maps *Kartverket* were used, where all latitude and longitude coordinates and the corresponding altitude are stored in a database in extensible markup language trees [61]. An application programming interface (API) call had to be made for each coordinate pair to retrieve the altitude for each point, which resulted in several hundred thousand API calls. This is highly time consuming, so multithreading were implemented. Multithreading is the property of handling several threads of executions concurrently within the cores of a central processing unit [62]. For a multi-core processor many threads can be run at perceivably the same time. With tasks that include wait time, as with making an API request to retrieve information from *Kartverket*, other threads may use the processing resources as another thread is waiting for the returned answer. The limiting restriction on the number of threads to be used was then how many simultaneous calls the server of *Kartverket* could handle, which turned out to be three. The API calls for several coordinate points could then be made in parallel, cutting down the execution time of retrieving the altitude values by a third. The altitude values were given in the

unit meters above mean sea level (*mamsl*). After retrieving the altitude for all points, the data again had to go through a cleaning iteration. Altitude points below zero means that the coordinates are over a body of water, where the sheep have not been. These points are therefore erroneous. New coordinates were then instead imputed by taking the mean of the two closest adjacent points, and a new altitude value retrieved and a new inverse trajectory angle calculated.

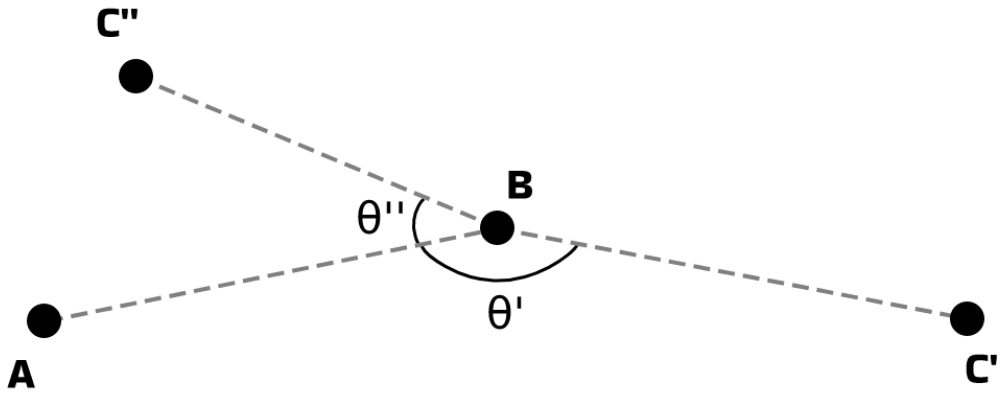


Figure 3.5.1: The trajectory angle found between three points A, B and C. The two different C-points show the angle gotten for relatively unchanged directional trajectory with C', and opposite directional trajectory with C".

The temperature also had to be collected from an outside source, and was downloaded from *Norsk Klimaservicesenter* delivered by *Meteorologisk Institutt* [63]. The weather stations used were Rissa III (23 *mamsl*) for Fosen and Sunndalsøra III (6 *mamsl*) for Tingvoll. The weather observation data had measured the temperature every hour, so each date and hour of the weather data were matched with the date and hour of the sheep data, and the temperature inserted for each point. Not all hours had a registered temperature in the weather data, so if no value were given in the temperature feature for a point, the mean of the two adjacent points imputed the value instead. The weather stations are not exactly at the same locations as where the herds resided, but they were the closest found with observed data for the time periods in question. The temperature used in the machine learning models is therefore an approximation, but it is assumed the data will qualitatively be in the same order of magnitude as the temperature the sheep actually experienced.

New features were also generated from the files of extra information on each sheep. These features were the age, the number of lambs, the breed, the farm, and identification number. Lambs of age zero would naturally not have any lambs themselves. The number of lambs are based on the amount born by each ewe, which might have changed during the grazing season if any were e.g. lost, killed or slaughtered. Additionally, *Miljødirektoratet* was contacted in order to obtain data on observed

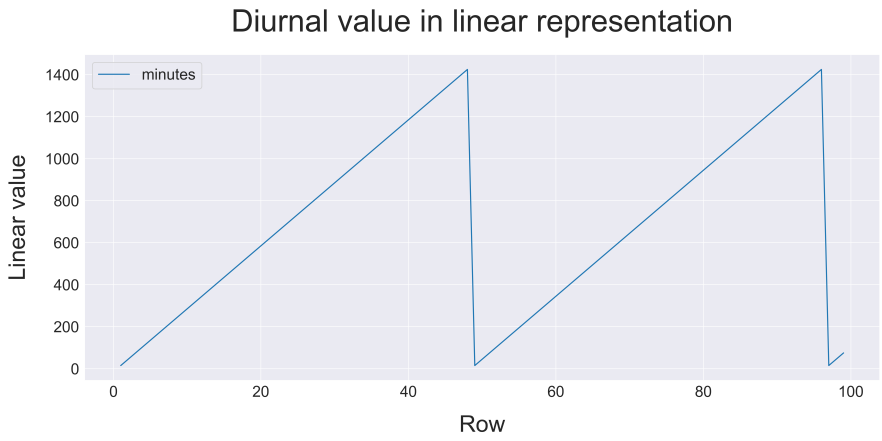
locations of predators in Fosen and Tingvoll. The data was based on the approximate coordinates an observer noticed some sort of trace of a predator, and the time stamp of when the predator most likely had been there based on how fresh the trace seemed. The traces could be footprints, fur, feces, carrion or live observations. However, the observations were irregular and very few, and relatively imprecise. This data was hence not used in the analysis.

3.5.2 Transforming temporal data

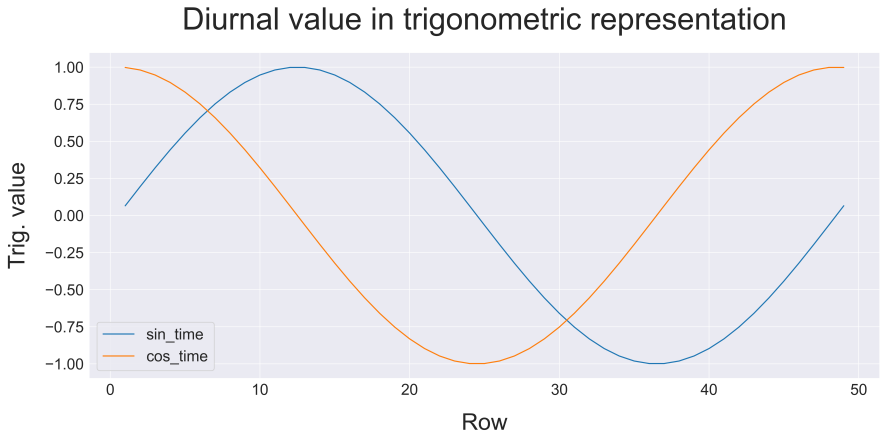
The time feature in the data is represented as *datetime* objects in *Pandas*, which is not in a numerical format the model will understand and can compute with. A new feature had to therefore be generated by transforming the *datetime* variable. When examining the movement pattern of sheep on rangeland pastures the interest in this study will mainly be in their diurnal cyclic habits. It is expected to see some variations in behavior across dates and throughout the season, but this can be dealt with by splitting the data into categories of longer time periods and cross-examining the differences. The times of the day follow a cyclic process of 24 hours however, and the time difference across the 24-hour mark must be equal to other time differences on the clock. If e.g. only the hour is extracted to consider the diurnal behavior, the computer will interpret 23:00 and 01:00 to be 22 hours apart, instead of actually only being two hours apart. These points in time will therefore be far apart from each other as input in the model and will most likely not be clustered together, even though it is expected of the sheep to behave approximately the same at these times. This is illustrated in figure 3.5.2a, where the daily cycle will drastically jump at 00:00 when using linear time. The solution is to transform the 24-hour clock to cyclic values using trigonometric representation instead [64]. The *datetime* objects in the data were generated either at the minutes :00, :15 or :45, with some variations to the exact time the point was fixed. The resolution of the time was therefore set to minutes, and the 24-hour clock represented as minutes after midnight. Since both sine and cosine will intercept the same value twice in their individual cycle, in the same way an analogue clock will hit an hour twice every day, each time must be represented by a sine-cosine pair given by equation (3.1). Sine and cosine are out-of-phase with each other, breaking the symmetry, and thus ensuring that every sine-cosine pair represents a unique value, as illustrated in figure 3.5.2b. When plotting the time pair representations against each other for the first 100 rows of data, time has become a cyclic feature as seen in figure 3.5.2c, represented now as a 24-hour analogue clock. Midnight is hence at the top of the clock, every quadrant measures six hours, and midday is at the bottom of the clock. This transformation also ensures that the time feature is normalized and standardized, since trigonometric values will always be bounded by [-1, 1].

$$\begin{aligned} \text{Sine time} &= sin(2\pi \cdot \text{minutes after midnight}/(24 \cdot 60)) \\ \text{Cosine time} &= cos(2\pi \cdot \text{minutes after midnight}/(24 \cdot 60)) \end{aligned} \tag{3.1}$$

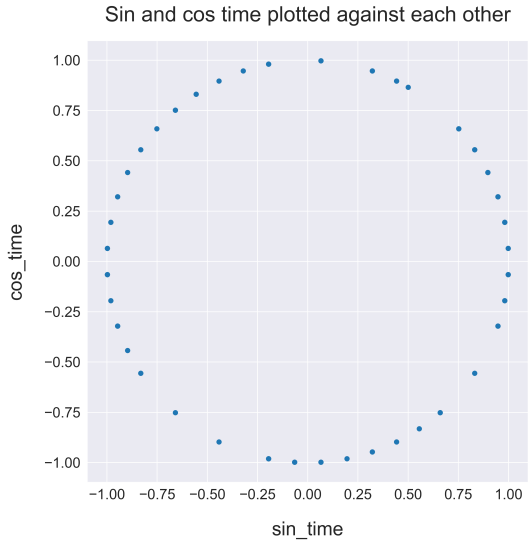
Thus the final features in the data after the feature engineering are latitude, longitude, datetime, velocity, identification number, sine time, cosine time, age, number of lambs, breed, farm, temperature, altitude and inverse trajectory angle. The features that will not be used directly in the model are latitude and longitude (preserved in velocity and altitude), datetime (preserved in trigonometric time), identification number, farm, and breed.



(a) Time as a linear variable.



(b) Time represented as pairs of sine and cosine.



(c) Time as a cyclic feature.

Figure 3.5.2: Figure (a) shows the time variable in the data for the first 100 rows using linear time as minutes past midnight, figure (b) shows the trigonometric time representation for one cycle where each time point has a unique sine and cosine-pair value, and figure (c) shows time as a cyclic feature with the trigonometric pairs.

3.6 Implementing the machine learning models

Before running both k-means and DBSCAN, the features of interest were selected, and the rest deleted. The remaining features were standardized and normalized, and the hyperparameters optimized. With unsupervised learning there is not a direct way of checking the accuracy of the model, since the predicted target values cannot be compared to any true labels as with supervised learning. Optimizing the model must therefore be done by other means, like making sure the hyperparameters are as optimized as possible to reduce the variance and error in the model.

3.6.1 *K*-means

The first model k-means were optimized using the elbow method by comparing the explained variance against the number of clusters. The model were run for two different set ups, one where only the velocity of the sheep and the time of day were included as features, and one where all features were included. By only checking how the velocity of the sheep changes throughout the day, the clusters generated will describe the activity periods sheep will go through based on the simplest example of data. K-means is best suited for describing the normal behavior of the sheep, as all points will be assigned to a cluster, outliers included. The outliers may skew the clustering results, but this will also be useful in determining the usual activity habits of the sheep as it will take into account when the atypical behavior usually occurs. The results of the elbow method for the velocity and time is shown in figure 3.6.1, where the determined elbow point is at four clusters. The k-means model were implemented by setting up the algorithm with the found optimized hyperparameter, fitting the model on the data, and labeling all points to their respective clusters.

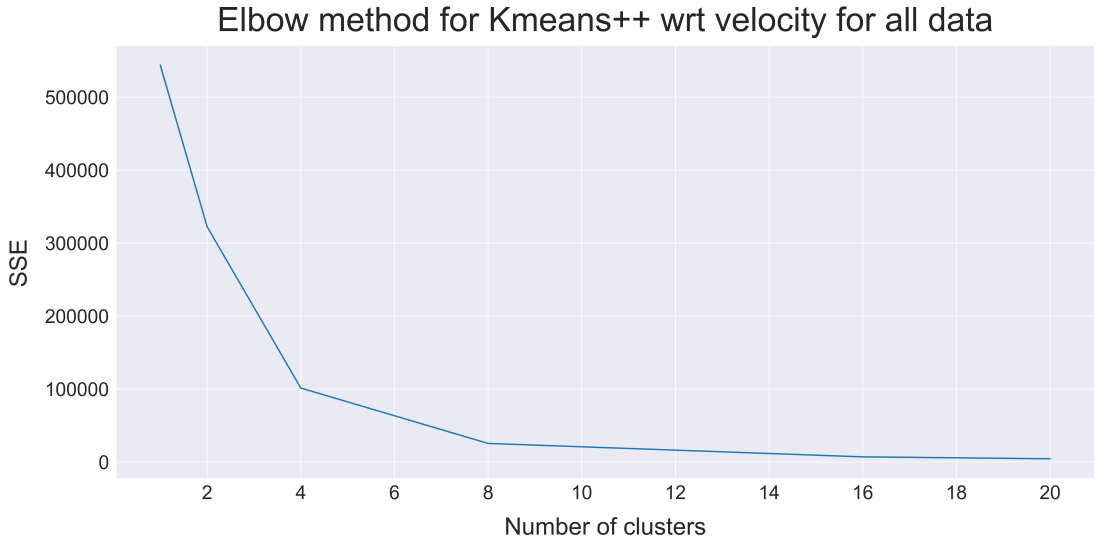


Figure 3.6.1: Explained variance as SSE for the data plotted against number of clusters included, with respect to the velocity for all sheep.

The second run were done with all the features included. The cluster number optimization had to be run again, where the results are shown in figure 3.6.2. The optimal amount of clusters for all features is also here four.

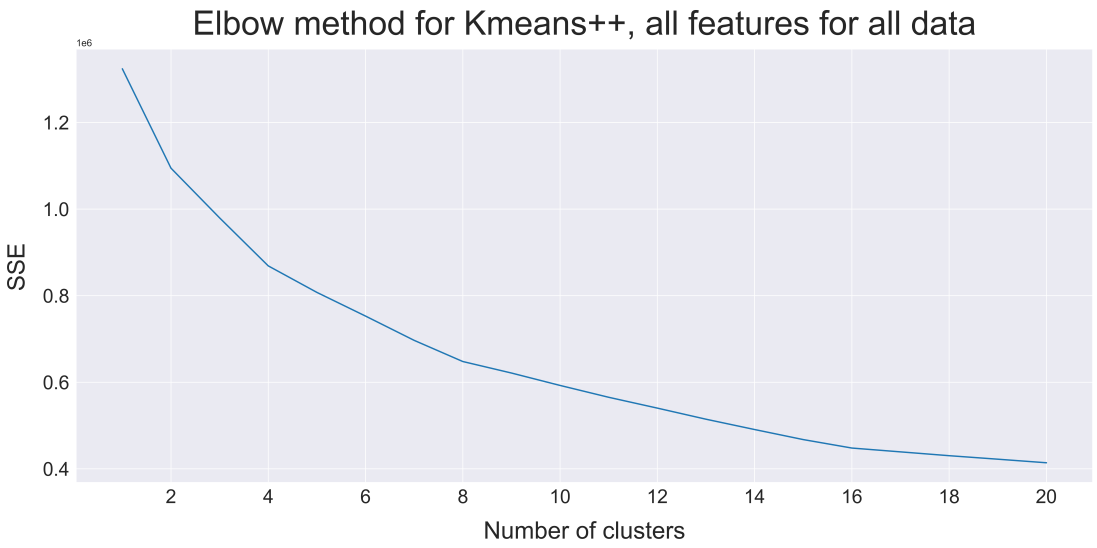


Figure 3.6.2: Explained variance as SSE for the data plotted against number of clusters included, with respect to all features.

3.6.2 DBSCAN

The optimization of DBSCAN will be done by using the elbow method on radius ϵ-plots calculated by the K-Nearest Neighbours algorithm. DBSCAN will be used mainly to describe the atypical behavior of the sheep. As the outliers in DBSCAN will be set aside when the algorithm describes the final clusters, some critical sheep behavior will not affect the clustering outcome. Atypical behavior, such as flight from predators, are still within the normal span of how sheep behave even though it is not the predominant behavior. Therefore when describing the overall behavior of sheep k-means were used where the outliers were included. The main advantage with DBSCAN is that it can separate the outliers such that the atypical behavior alone can be analyzed and described. The model were first run with all data and all features, where the resulting ϵ-plot for hyperparameter optimization is shown in figure 3.6.3. The epsilon radius, which is the maximum distance between two data points for them to be considered neighbours, was determined to be $\epsilon_{all} = 0.37$. Three different data split configurations were run with DBSCAN, with all features included for all. The first data split configuration were done on different types of sheep breed, where the lighter short tailed breeds of Grey Trønder, Old Norwegian and all types of Spæl were separated into one data split, and the heavier long tailed breed of NKS were the other. The second data split were done on early versus late season, and the last data split were to look into the differences in behavior between day and night. The resulting epsilon values for all the split data sets are given in table 3.6.1.

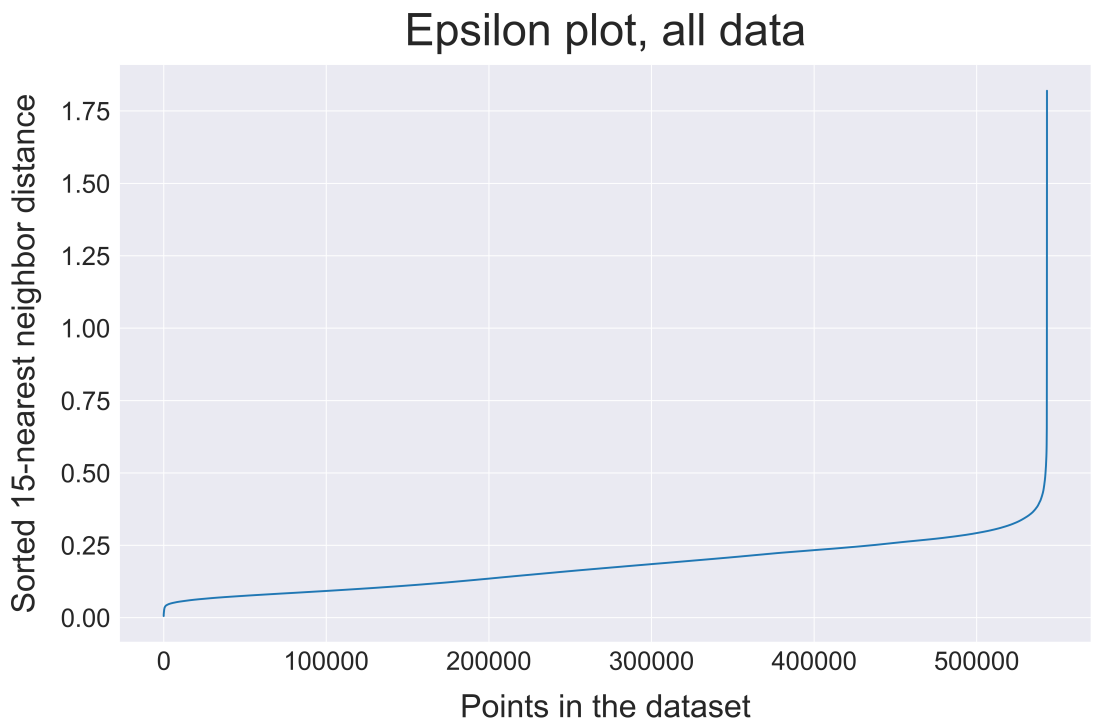


Figure 3.6.3: Sorted K-NN distance for all data points, using all numerical features. The elbow to the right is used to choose the value for the epsilon hyperparameter in DBSCAN.

Data split	**Epsilon**
All data	0.37
Heavier breed	0.39
Lighter breed	0.39
Early season	0.42
Late season	0.39
Day	0.37
Night	0.39

Table 3.6.1: Epsilon radius values for the DBSCAN hyperparameter optimalization for all data set configurations.

The DBSCAN model was implemented by setting up the algorithm with the optimized hyperparameters, fitting the model on the data, and assigning each point to a cluster. The outliers, or noise points, could then be separated from the rest and inspected. Threshold values of atypical sheep behavior were then calculated based on the mean of the outlier points for the different dynamical features, by reversing the standardization and normalization of the data values.

3.6.3 Other

Principal Component Analysis (PCA) were implemented and tested for k-means, but not used for further analysis. The reason is that PCA will assume a linear relationship between the features, and works best when the features have a strong correlation [65], which were not quite applicable to the data in this study. PCA is also sensitive to outliers. The code for the PCA model is however still included in the script *Kmeans.py*.

3.7 Two-tailed two sample *t*-test

The hypothesis test chosen to check the statistical significance of the data split threshold values was the two-tailed two sample t-test. The t-test is suited to compare and examine any difference between two samples of the same variable but with underlying differences for each set [66]. Here the different sample configurations are the noise data on lighter short tailed sheep breeds versus the heavier long tailed sheep breed NKS, the early versus late season, and lastly the daytime hours versus nighttime hours. The test will be two-tailed since the quality to be examined is if the sample mean difference is not equal to zero, and hence that the threshold marker values for the feature should be differentiated on the data split configurations.

$$H_0 \quad : \mu_1 - \mu_2 = 0 \tag{3.2a}$$

$$H_A \quad : \mu_1 - \mu_2 \neq 0 \tag{3.2b}$$

To find the p-value, the standard error (SE) and test statistic t_{stat} are calculated by equations (3.3a) and (3.3b). Here σ_i is the sample standard deviation and n_i the sample size. The test statistic is compared to a t-Distribution table against the degrees of freedom in the variables, $dof = n_1 + n_2 - 2$. Since the test is two-tailed, the chosen criterion for the significance level α has to be divided by two, with one segment of significance for each tail on the distribution. The p-value is the α that corresponds to some critical value t_{crit} in the table for two-tailed alpha-values. The null hypothesis can be rejected, and the achieved mean sample difference can thus be declared statistically significant, if the t_{test} value is more than the t_{crit} of the chosen criterion for α.

$$SE \quad = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{3.3a}$$

$$t_{test} \quad = |\frac{\mu_1 - \mu_2}{SE}| \tag{3.3b}$$

Chapter 4

Results

4.1 Statistical results and visualizations

This section shows the results from the EDA. The data from Fosen and Tingvoll are here merged after the finished data wrangling, so the results include all data collected unless indicated otherwise.

4.1.1 Feature correlation

The pairwise correlation between features were calculated, and plotted in a heatmap in figure 4.1.1. The colors show how linked and dependent two features are to each other. In machine learning, the input variables to a model should stay mostly uncorrelated with each other, if two variables are very highly correlated they will cause redundancy of information, decrease efficiency, and reduce the generalisation ability and the accuracy. One of the features should thus be removed. Most of the features have close to zero correlation, except a not too surprising positive correlation between age and number of lambs and a slight negative correlation between temperature and time of day. However, one will not be removed as the dependency is not deemed too high.

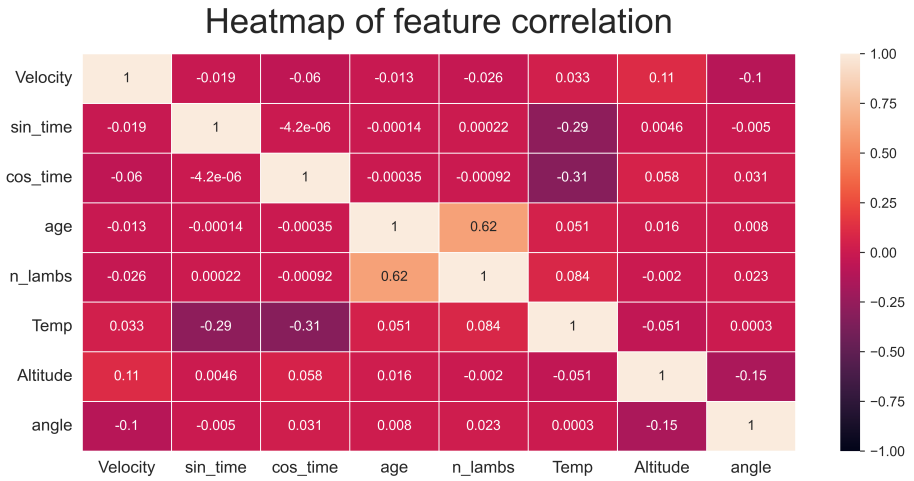
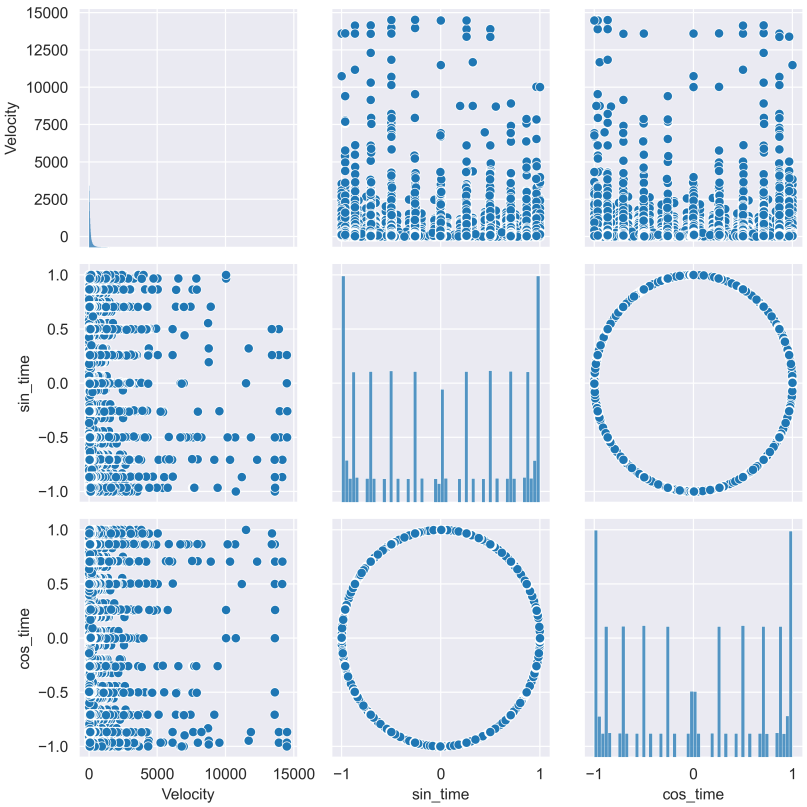
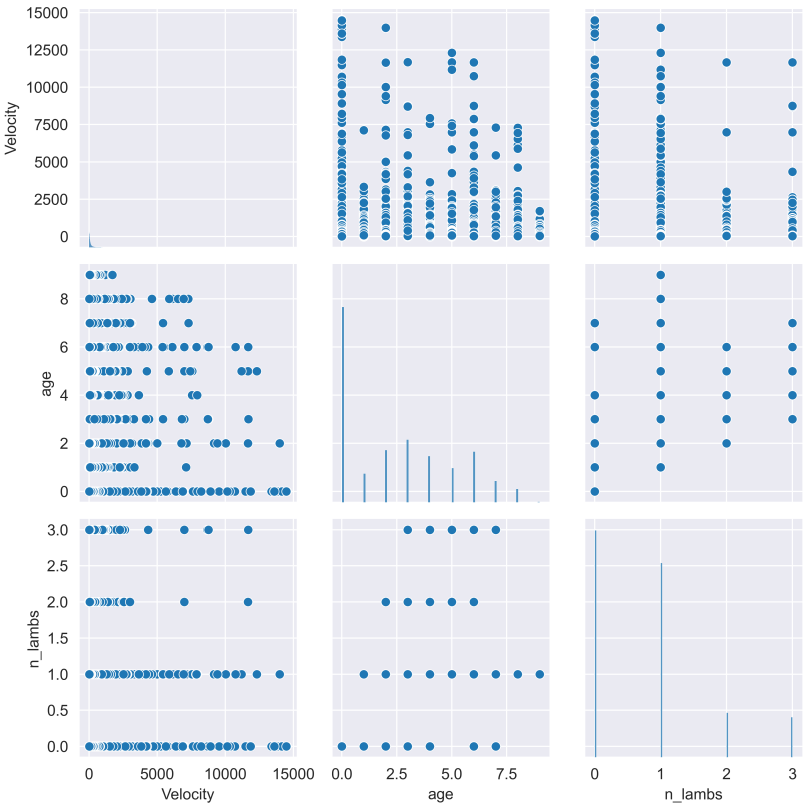


Figure 4.1.1: Heatmap of the correlation between features, where -1 is perfect negative correlation, 0 is no correlation and 1 is perfect positive correlation.



(a) Correlation between velocity and time.



(b) Correlation between velocity, age and number of lambs

Figure 4.1.2: Feature correlation matrices between the features velocity, time, age and number of lambs showing their relationship with each other, and the attribute distributions.

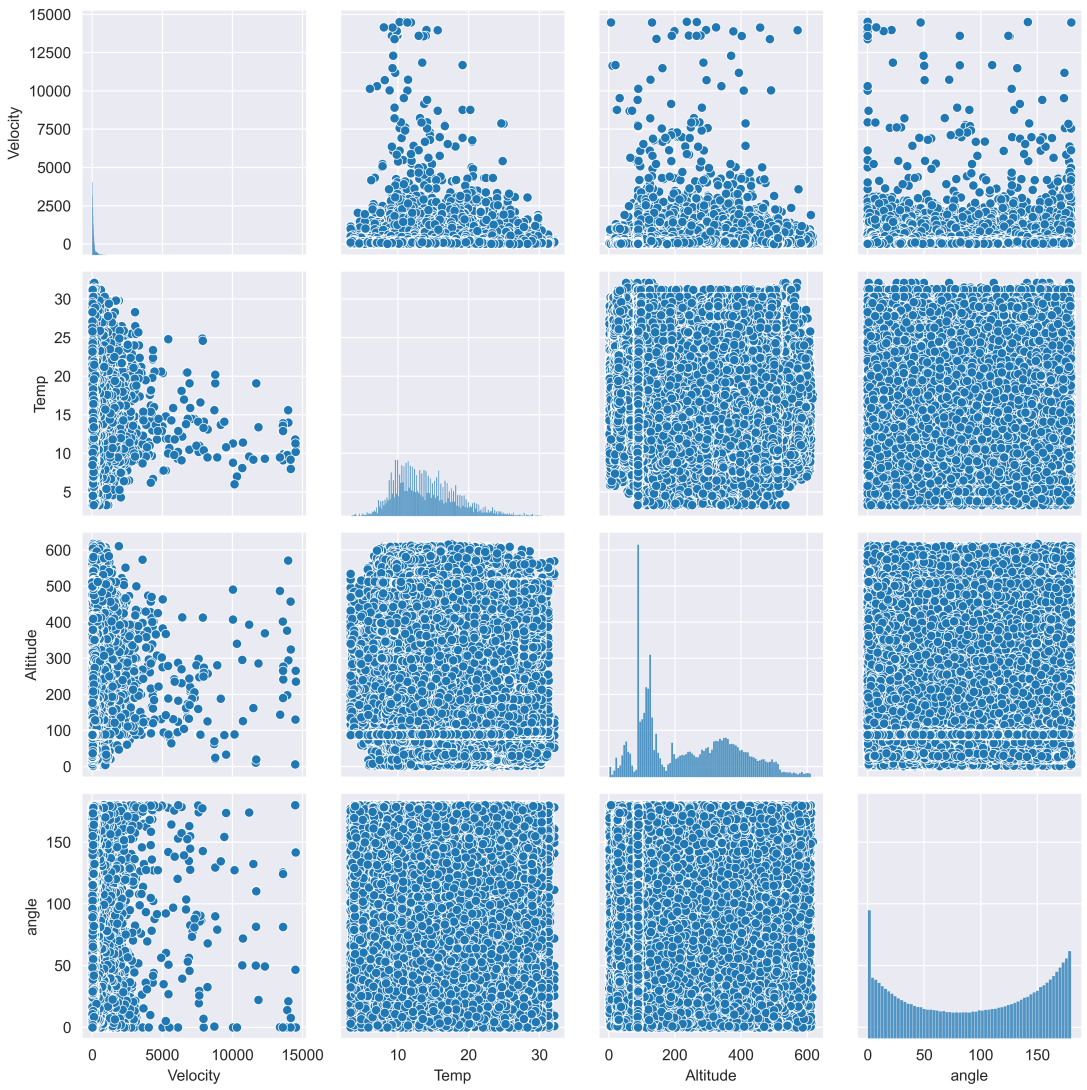


Figure 4.1.3: Feature correlation matrix between the features velocity, temperature, altitude and inverse trajectory angle showing their relationship with each other, and the attribute distributions.

In order to investigate further the inter-feature relationships and their dependencies to each other, scatter plot matrices (pairplots) were made as shown in figures 4.1.2 and 4.1.3. A pairplot will plot the values in all the features two and two against each other, revealing how they correlate. It will also plot the univariate histograms for the individual features on the diagonal, thus showing the distribution of the features, and will not plot the pairplot of a feature against itself because that will just become a linear function. Because of the number of features present, the plot was split, where all plots contain the velocity as this is the most important attribute to compare with when it comes to the sheep activity. Some features were not considered for the pairplot as they are not directly numerically comparable. The attributes left out were datetime, breed, farm and identification number. The split of which attributes to put against each other was determined based on which feature correlations were the most interesting, but a full correlation matrix with all features can be found in appendix D. Interpretations and further explanations can be found in chapter 5.

4.1.2 Statistics before data selection

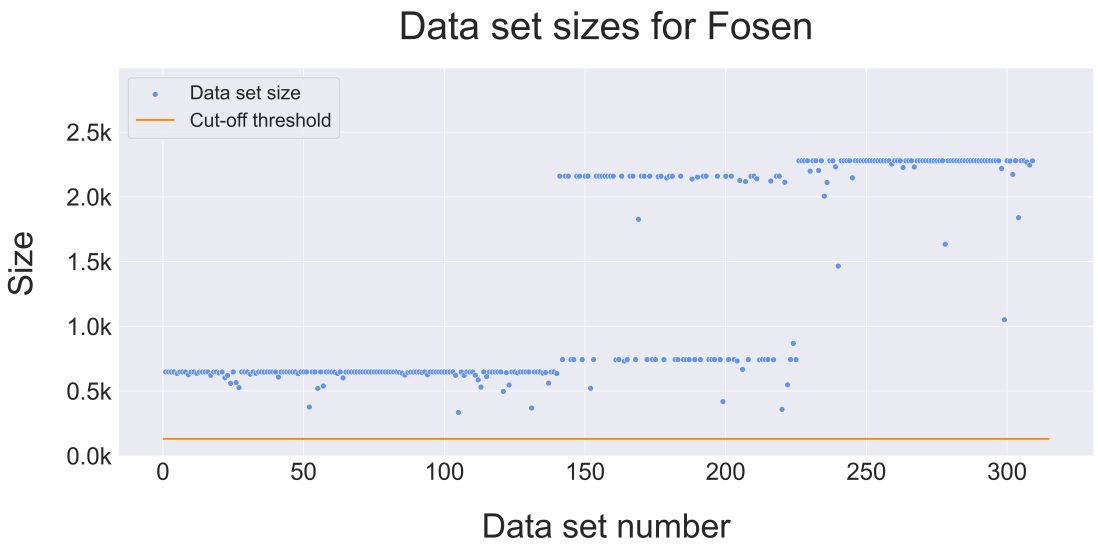
As part of the EDA, before any data were cut away in the data preparation process, preliminary visualizations were made. The data was cleaned of obvious errors, and the velocity of the sheep in terms of meter per time were added as a feature. This attribute was the first collected which could give some initial intuition on the behavioral pattern of the sheep. Plots of the data set sizes and different types of mean registered activity can be found in appendices C1-C5. A table of statistics on the velocity of the sheep can be found in appendix C6. All representations were done before any deletion of data, e.g. when fixed time frames were selected, and further cleaning in the iterative wrangling process as the data was more understood modified some values in the later analysis. Therefore these plots will be somewhat different than the ones presented later in this section. It may still be valuable to have available the unprocessed visualizations, as assumptions were made when threshold values for deletion were chosen, and some useful information may have been lost in this process.

4.1.3 Statistics after data selection

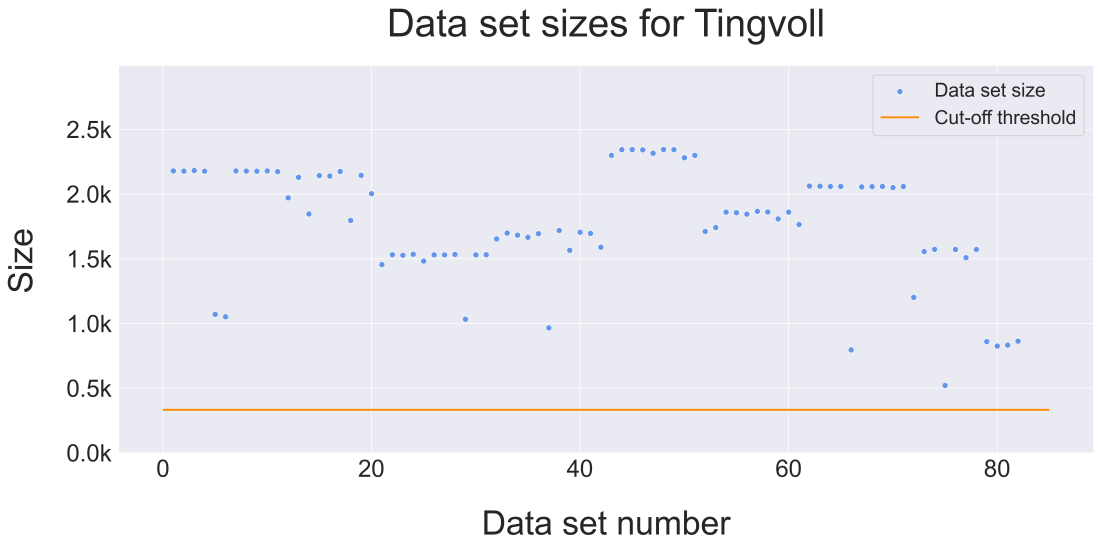
A table of the different statistics on the numerical and dynamic features after the data cut is given in table 4.1.1. This includes outliers that might skew the results upwards. The features considered are the numerical dynamic attributes, excluding time, as these are the features of interest for digital threshold values. Static or categorical attributes like age, breed or number of lambs will mainly not change while on rangeland pastures, and thus not influence a live statistical model.

Statistic	**Velocity**	**Altitude**	**1/Angle**	**Temp.**
Mean	122.68	240.98	93.75	13.95
Std	224.51	145.88	60.39	4.44
Q1	28.00	111.60	34.15	10.60
Median	63.00	223.20	99.59	13.30
Q3	137.00	359.10	151.99	16.70
Min	0.00	1.00	0.00	3.30
Max	14519.00	616.70	180.00	32.10

Table 4.1.1: Table of dynamic feature statistics where outliers are included, for all data points. Velocity is given in *m/h*, the altitude in *mamsl*, the inverse trajectory angle in 1/degrees, and temperature in degrees Celsius.



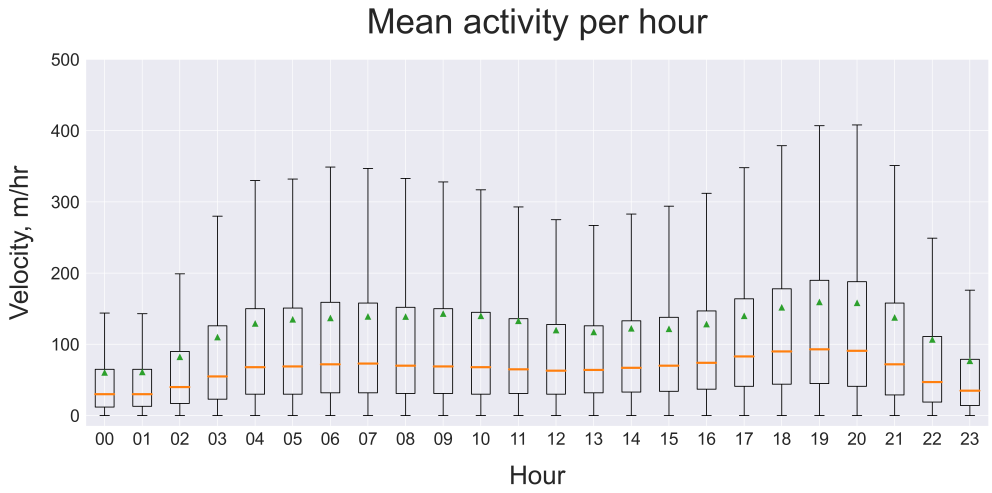
(a) Data set sizes in Fosen after the data cut.



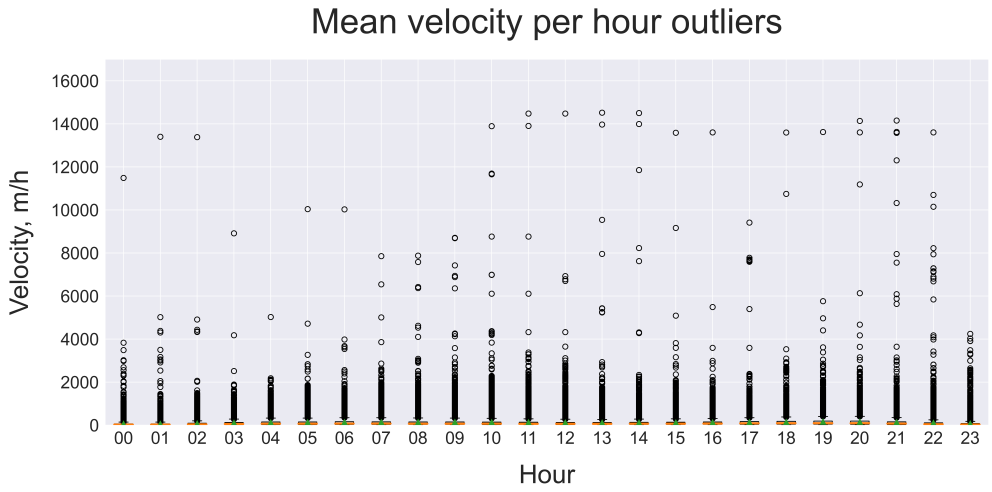
(b) Data set sizes in Tingvoll after the data cut.

Figure 4.1.4: The figures show the trajectory data set sizes of each data set, after the extraneous information were cut off. The orange line indicates the cut-off threshold for the data sets. The trajectory data set numbers are ordered chronologically in time, from (a) 2018-2020 in Fosen and (b) 2012-2016 in Tingvoll.

After the data wrangling and feature engineering were done, the total final data were explored and visualized. In figure 4.1.4 the size of each data set for the individual sheep trajectories are shown, i.e. each scatter point represents the total number of generated attribute points for every sheep. The final threshold cut-off values were 131 data points for Fosen, and 333 for Tingvoll. The values are based on 10 % of the mean set size after the data was cut due to setting a fixed time frame for each year and each place. All data sets in the figures are above this level, the ones below were deleted due to being too small to be considered.



(a) Mean sheep velocity



(b) Outliers of mean sheep velocity

Figure 4.1.5: The figure shows (a) the mean sheep velocity per hour in a boxplot, and (b) the outliers of the boxplot.

Figure 4.1.5 shows boxplot statistics and the outliers of the boxplot of mean sheep velocity, given in meters per hour. The boxes are bounded by the interquartile range (IQR), where the lower boundary is the 25^{th} percentile (Q1, first quartile), the upper boundary is the 75^{th} percentile (Q3, third quartile), and the orange line is the median (50^{th} percentile, Q2). The green triangle is the mean, and the whiskers are delimited by $\mp 1.5 \cdot$ |IQR| of the first and third quartile, respectively. Outliers are here points lying outside the defined whiskers, and the colored marks at the bottom of each hour in figure 4.1.5b are the boxes in the upper plot. There are two clearly defined local optima in figure 4.1.5a around 06 and 19 with a local minimum in between, and the global minimum is approximately at midnight. Outliers can be seen throughout the entirety of the day, but are more heavily concentrated between 10-22.

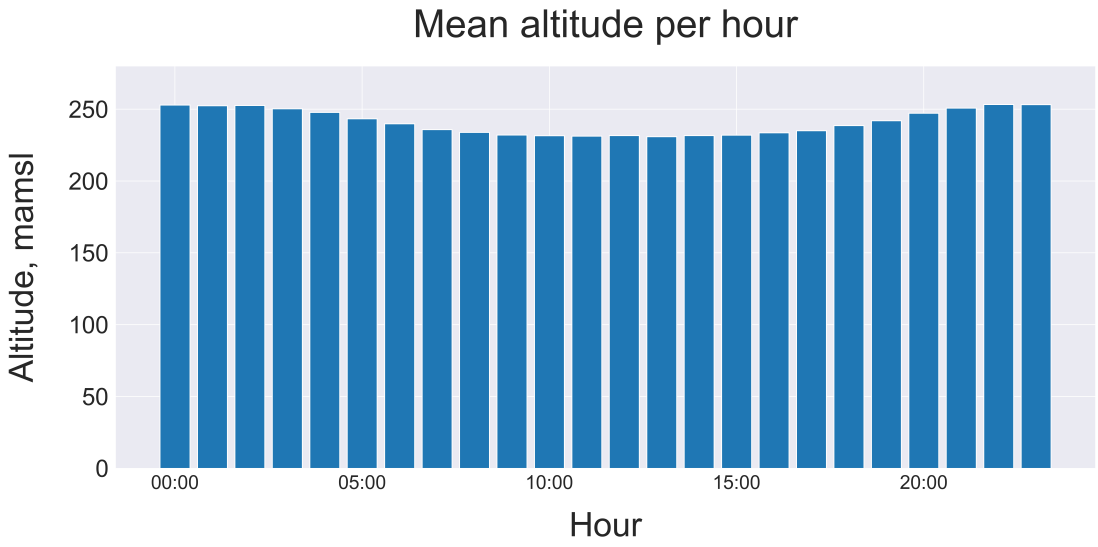


Figure 4.1.6: Mean sheep altitude per hour of the day, given in meter above mean sea level. There were no outliers.

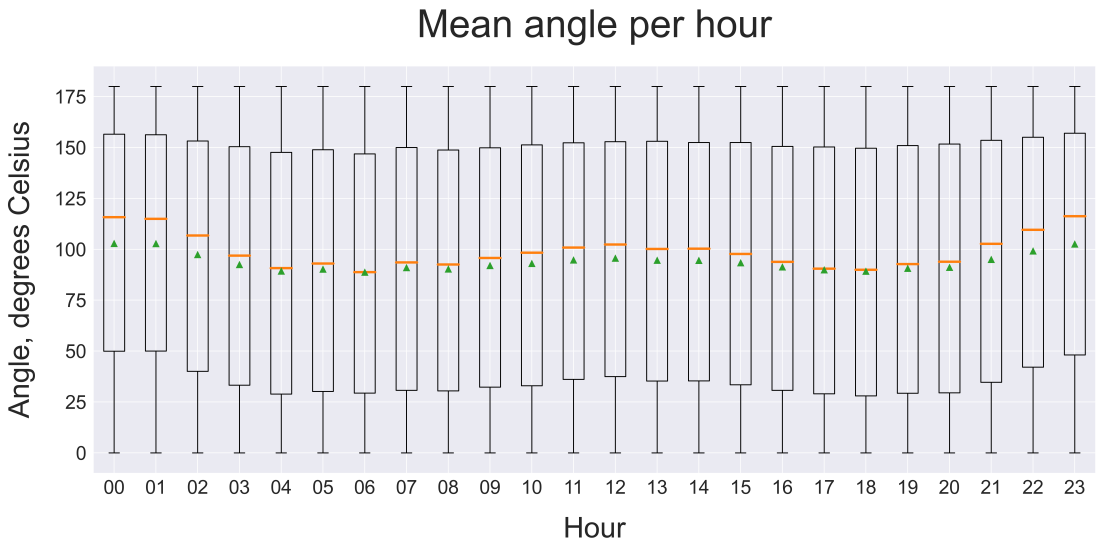
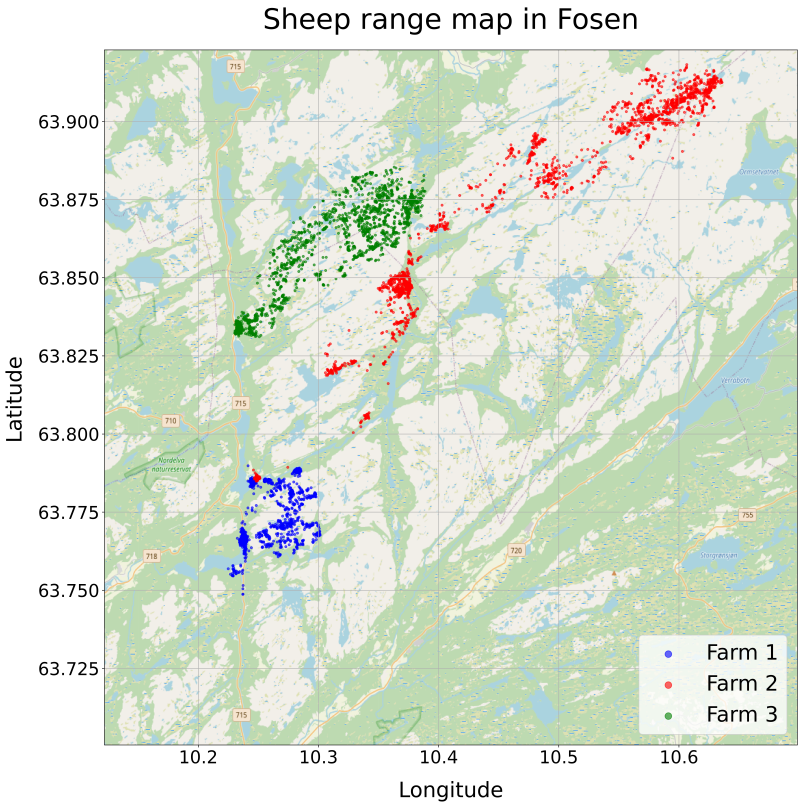


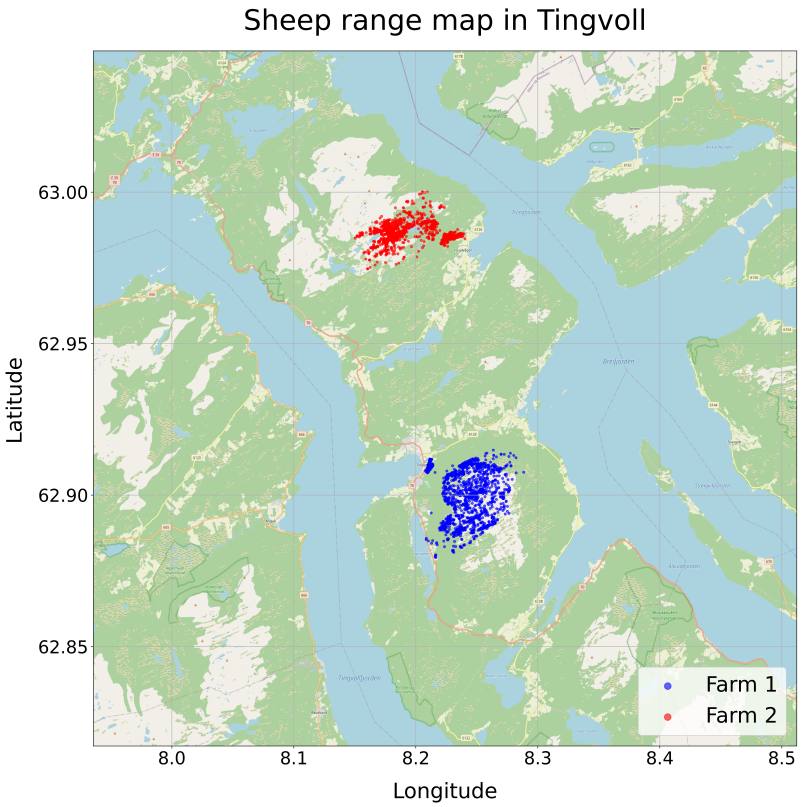
Figure 4.1.7: Mean sheep inverse trajectory angle per hour in a boxplot, given in 1/degrees. There were no outliers.

The mean sheep altitude per hour of the day is shown in figure 4.1.6. There is a small change in altitude between daytime and nighttime, where the sheep move upwards at dusk and come down again at dawn. Figure 4.1.7 gives the inverse trajectory angle per hour, where while the minimum and maximum stay roughly equivalent throughout, the mean and median follow a sinusoidal path with maxima at midnight and midday. The upper third quartile has some slight variation in accordance with the mean but is mostly flat, while the lower first quartile varies more and deviates the most between 22-02.

4.1.4 Map of trajectories



(a) Plot of sheep range trajectory on map, in Fosen.



(b) Plot of sheep range trajectory on map, in Tingvoll.

Figure 4.1.8: Individual sheep trajectories plotted on a map, showing examples of sheep home ranges in (a) Fosen and (b) Tingvoll.

Figure 4.1.8a shows three trajectory examples for each separate farm unit in Fosen, where each color plot represents approximately the different pasture home ranges of the units. The same is shown for the two separate farm units in Tingvoll in 4.1.8b. The maps are retrieved from *Open Street Map* [67].

4.1.5 Activity distribution

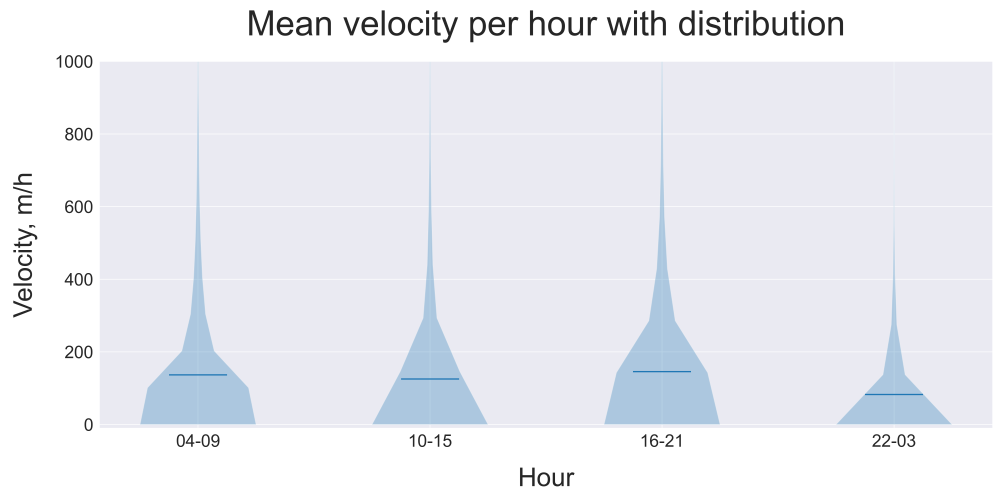


Figure 4.1.9: Sheep activity distribution per hour presented as a violin plot with four bins. The width of the violin plot indicates the frequency of the y-value, while the mean is indicated at the blue horizontal lines.

Based on the activity levels presented in figure 4.1.5, four bins were defined with a duration of six hour intervals, capturing the four extrema of the mean velocity. The frequency distribution of the bins and their means are plotted in figure 4.1.9, using a probability density function which smoothes the four histogram levels. The larger the area the more activity, and a taller spire means more intense activity. The values on the horizontal axis represents all points that are within the hour stated, meaning that e.g. a point at 09:45 will belong to the bin of 09. The distribution is widest above the mean at 16-21, indicating this is the most active time for the sheep with the most high-intensity movement. The distribution is also quite wide at 04-09 below the mean, suggesting that at these hours the sheep will be highly active in general but in a more moderate and relaxed manner than at the afternoon.

4.2 Results of the machine learning models

The task of clustering the data with unsupervised machine learning models, so that it could be described and characterized was done with k-means and DBSCAN. The results of the machine learning is presented below.

4.2.1 *K*-means

Several analyzes were done with k-means, mainly to determine normal behavior as described by data. The first were to investigate the basic case of looking at only velocity in terms of time of day to look into their diurnal activity traits, that is their daily movement pattern. The results are presented in figure 4.2.1, where the sheep activity clusters into four distinct time intervals, thus categorizing their day into four characteristic activity periods. These time periods are 22:30-04:30, 04:30-10:30, 10:30-16:30, 16:30-22:30. This is also highly in accordance with the density distribution bins given in figure 4.1.9. The plot in 4.2.1 is also shown in appendix E1 with a top-down perspective, to show the 24-hour clock view of the activity periods. The more intense velocity points are in higher density during the day, indicating that atypical activity behavior perhaps should have a higher threshold value at day than at night.

K-means was also run with all numerical features included, where the number of clusters given by the elbow method was four also here. Visualizing eighth dimensional data to show the resulting clustering is difficult to do in an understandable and readable way. Therefore the mean and standard deviation for all features in all clusters were calculated and plotted in polar line plots in figures 4.2.2a and 4.2.2b. This shows the identified clusters' different average scopes. The mean and standard deviation of all cluster values are also numerically presented in table 4.2.1, while the feature mean for the individual four clusters are attached in appendix F.

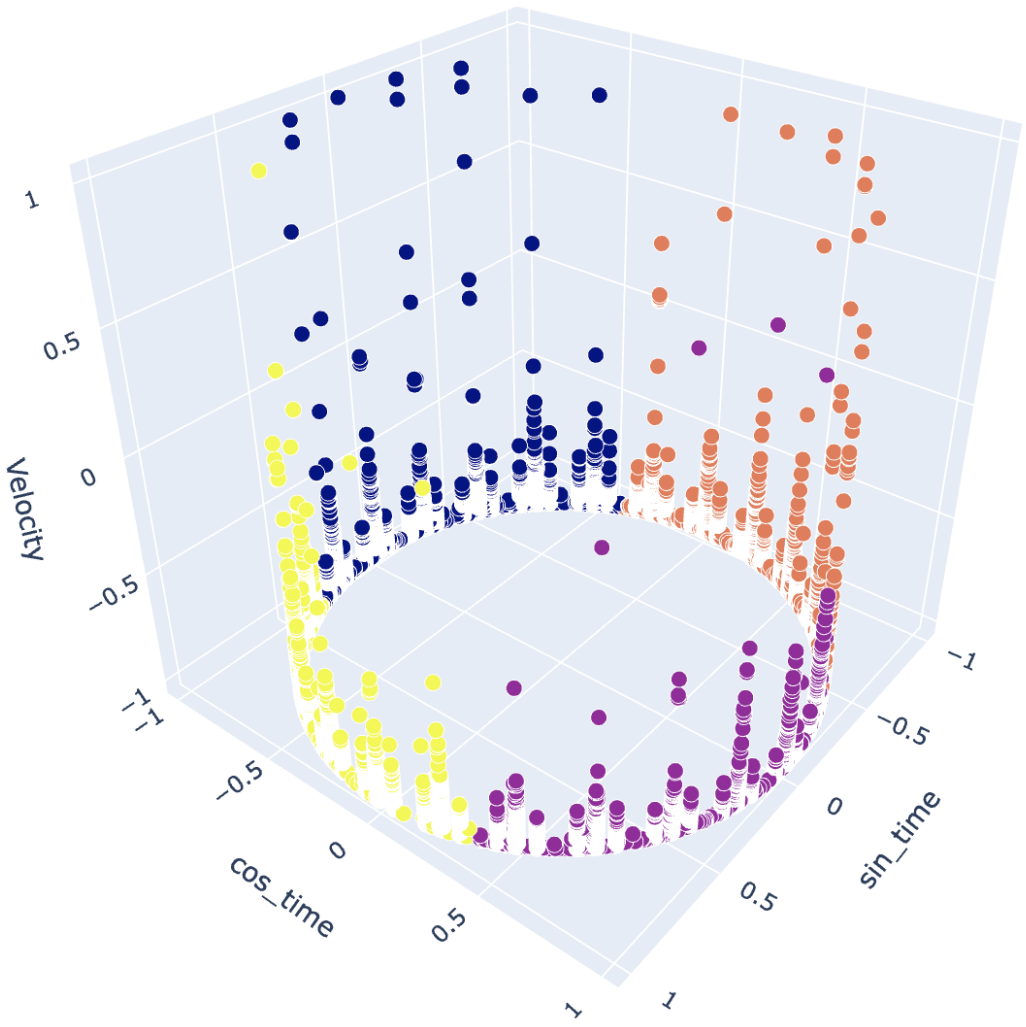
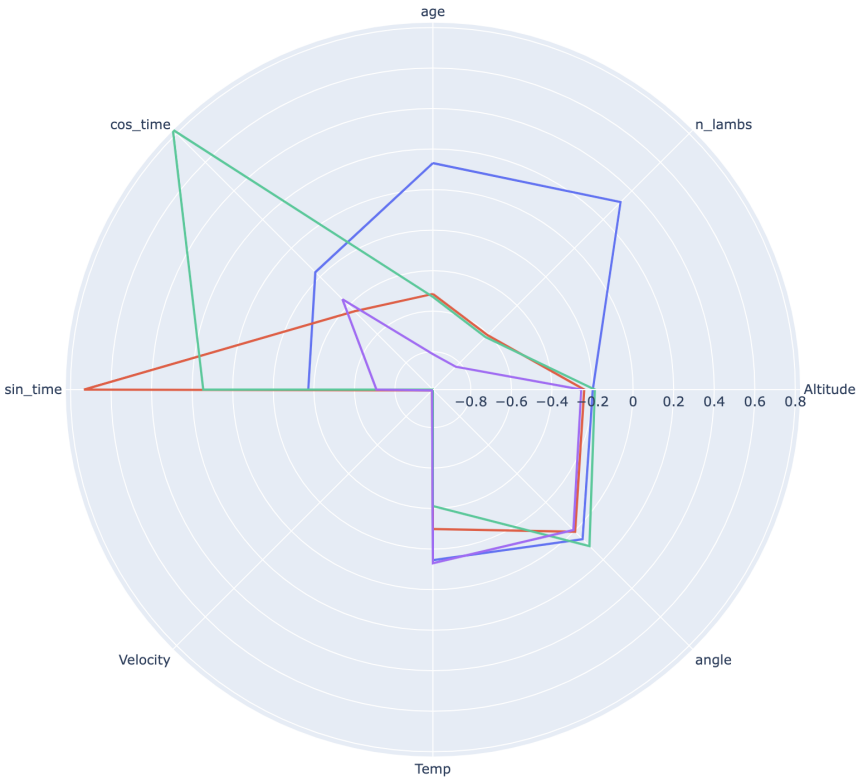
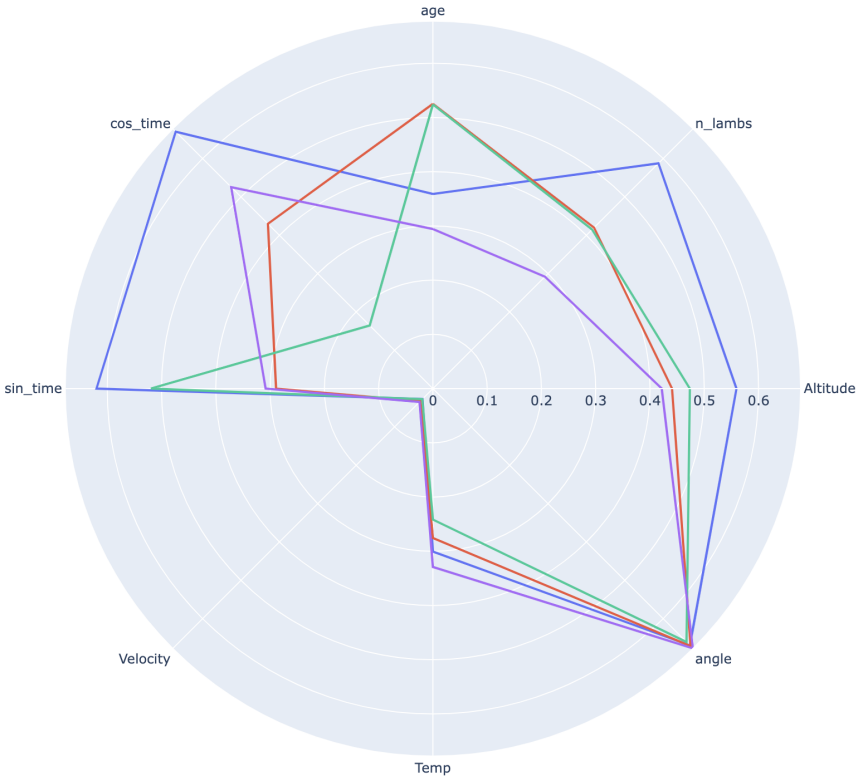


Figure 4.2.1: Clustering with K-means showing their diurnal activity patterns, using the features velocity, sine time and cosine time for all data. The purple cluster at sine time equal to zero corresponds to midnight. The number of clusters $k = 4$ were decided using the elbow method as shown in figure 3.6.1.



(a) Polar line plot of the k-means clustering for all numerical features and all data.



(b) Polar line plot of the standard deviation of the k-means clustering for all features and all data.

Figure 4.2.2: K-means clustering results in mean and standard deviation for all numerical features for all data, with $k = 4$. Each color represents a computed cluster. The features included are velocity, altitude, trajectory angle, sine and cosine time, temperature, age and number of lambs.

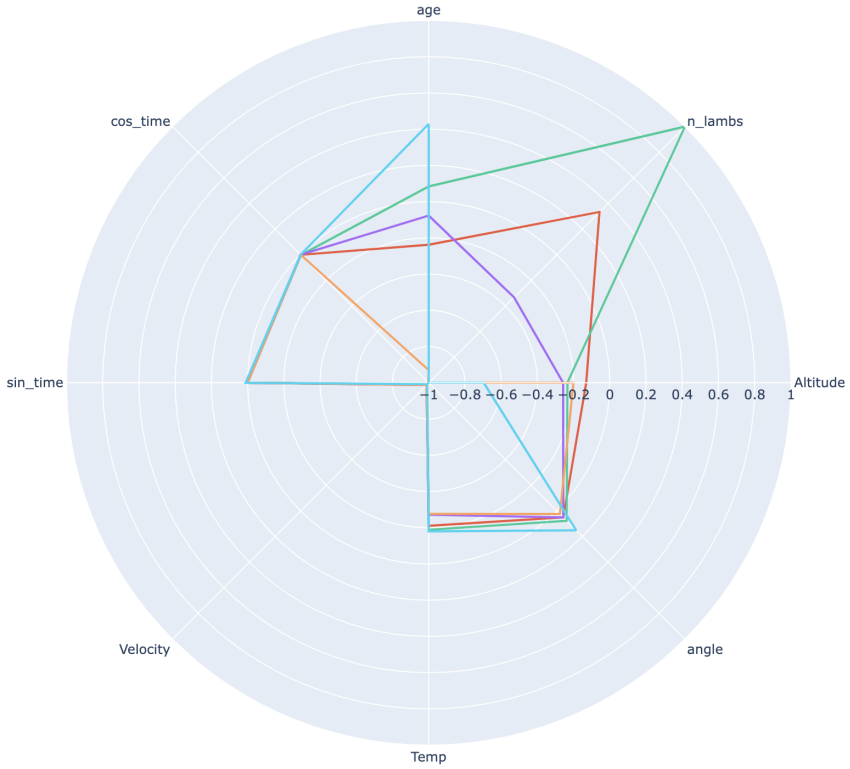
Feature	Mean	Std
Velocity	-0.983	0.003
Altitude	-0.221	0.032
1/Angle	0.040	0.051
Temp.	-0.247	0.134
Age	-0.431	0.399
n_lambs	-0.433	0.513
sin_time	-0.051	0.630
cos_time	-0.035	0.585

Table 4.2.1: Feature cluster statistics from the results of the k-means model for all features and all data, showing the mean and standard deviation across the four identified clusters. All values are standardized and normalized.

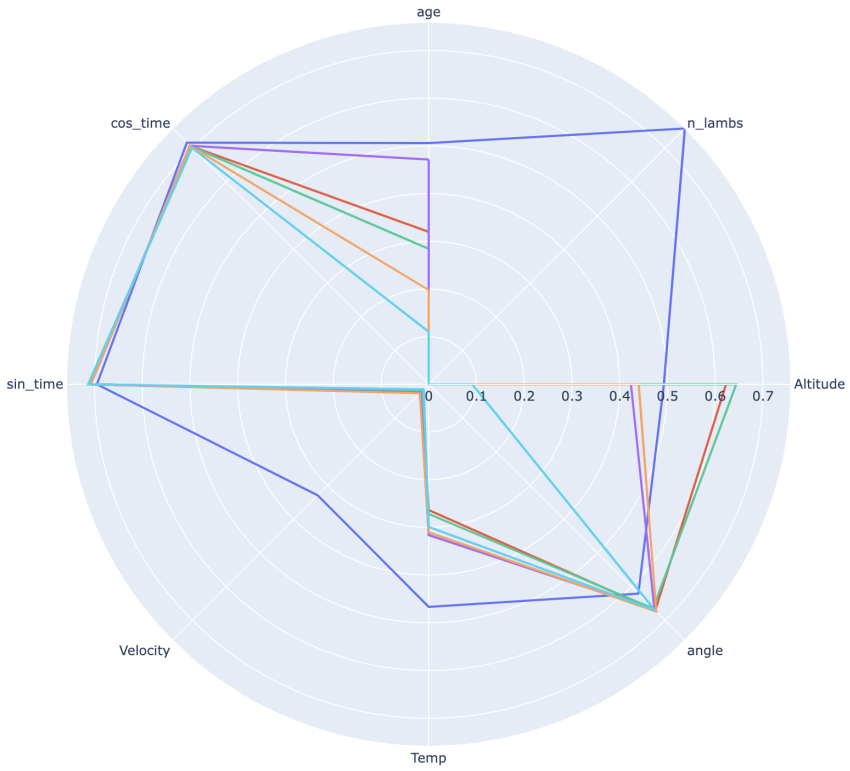
4.2.2 DBSCAN

The domain knowledge acquired about the normal behavior of sheep from the k-means model suggests that a good estimate at the number of clusters for typical behavior should be around four. Likewise, the minimal number of nearest neighbours in DBSCAN should be at least as large as the dimension of the data and larger for bigger and more noisy data sets. Therefore the $minPts$ hyperparameter for DBSCAN was set to a value that gave approximately the equal amount of typical clusters of that of k-means, and that produced a comparably similar cluster plot.

The minimal number of nearest neighbours for all features and all data was determined to be $minPts = 16$, which is twice the dimension of the data. This gave a plot closest related to that of normal behavior in k-means. The model determined there to be five clusters of typical behavior. The resulting polar line plots of the mean cluster values and the standard deviation are shown in figure 4.2.3. The mean of the noise points, i.e. the outliers of the data representing the atypical behavior, is likewise plotted in figure 4.2.4. Examining other values of $minPts$ showed that the outliers had approximately both the same plot shape and values regardless. Statistics from the dynamic features of the noise points are given in table 4.2.2, where the mean values are slightly elevated from the same statistics given in table 4.2.1, except for inverse trajectory angle where it is slightly lower. Only the dynamic features are considered since the goal is to identify digital threshold marker values to imply atypical sheep behavior, and these attributes have to be able to change in order to notice deviant actions in real time. The total amount of outliers determined by the model was 1521.



(a) Polar line plot of the mean of the DBSCAN clustering for all numerical features and all data.



(b) Polar line plot of the standard deviation of the DBSCAN clustering for all numerical features and all data. The blue line is the standard deviation for the noise points.

Figure 4.2.3: DBSCAN clustering results in mean and standard deviation for all numerical features for all data, showing the five computed clusters. The features included are velocity, altitude, trajectory angle, sine and cosine time, temperature, age and number of lambs.

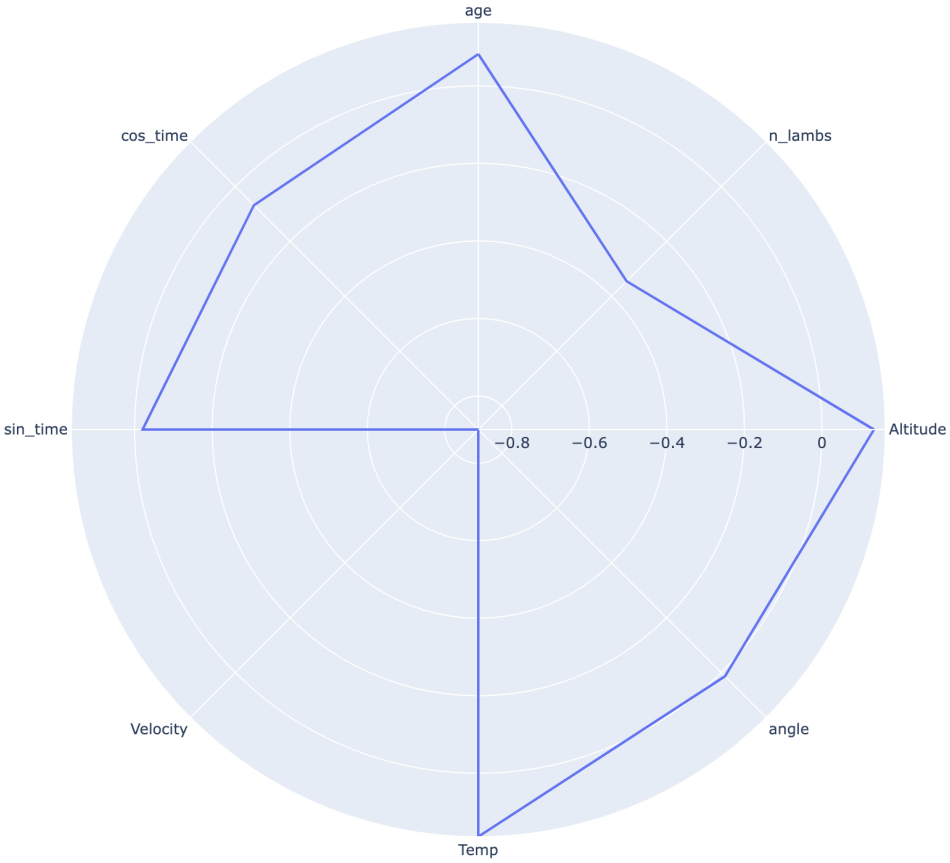


Figure 4.2.4: Polar line plot of the mean for the detected noise points from the DBSCAN clustering results for all features and all data.

Statistic	**Velocity**	**Altitude**	**1/Angle**	**Temp.**
Mean	-0.886	0.134	0.013	0.164
Std	0.329	0.492	0.621	0.466
Q1	-0.998.	-0.202	-0.537	-0.292
Median	-0.991	0.150	0.025	0.278
Q3	-0.975	0.523	0.563	0.521
Min	-1.000	-0.999	-1.000	-1.000
Max	1.000	0.979	1.000	1.000

Table 4.2.2: Table of dynamic feature statistics for the noise points determined by DBSCAN, calculated with all data. All values are standardized and normalized.

The data was split into several versions to look at behavioral differences with different factors included. The first configuration was to look at the breed differences between the heavier long tailed breed NKS, and the lighter short tailed breeds present. The minimal number of nearest neighbours for all model splits converged

at $minPts = 15$. For both types of breed the model produced six clusters apiece. The model detected 1126 outliers for the heavier type, and 1065 for the lighter breeds. The second type of split was to look at early versus late season, where the limit was set at before and after the first of July. Here the model resulted in five clusters, for both sets. The amount of outliers were 810 for the early season, and 1325 for the later season. The final split was to look at daytime values versus night time, since the sheep activity differs substantially during the course of a day, as seen in figure 4.1.5 and 4.1.9. The diurnal hourly split were decided by the hourly sheep velocity, where the minimal low between 22-03 were determined to be the night hours, and 04-21 were set to be the day hours. The amount of clusters also here became five for both versions. The total number of outliers were 1136 for the day hours, and 549 for the night hours.

Statistic	**Velocity**	**Altitude**	**1/Angle**	**Temp.**
Heavier breed				
Mean	-0.908	0.014	0.034	0.029
Std	0.278	0.392	0.645	0.513
Lighter breeds				
Mean	-0.924	0.166	-0.011	0.144
Std	0.279	0.510	0.618	0.474
Early season				
Mean	-0.873	-0.517	-0.118	-0.027
Std	0.381	0.512	0.560	0.376
Late season				
Mean	-0.918	0.181	0.039	0.130
Std	0.248	0.452	0.623	0.482
Day				
Mean	-0.880	0.140	0.014	0.192
Std	0.342	0.501	0.632	0.472
Night				
Mean	-0.922	0.145	-0.051	0.373
Std	0.261	0.505	0.610	0.485

Table 4.2.3: Table of dynamic feature statistics for the noise points determined by the DBSCAN model, for the different data split configurations. All values are standardized and normalized.

All mean and standard deviation values for the noise points of the dynamic features, in all data splits, are listed in table 4.2.3. These values will be used to calculate if there is any statistical significance on the difference between the split data configurations, to test if the factors of type of breed, date, or hour of the day will have any noteworthy impact on any threshold marker values that might be implemented. Since the numerical values are the factor of interest, all plots of clustering results and noise plots are not included in the results, only the table values. They were however of approximately the same shape and size as the results for all data.

4.3 Digital threshold markers

The statistical significance was computed for all data splits and all dynamic features from equation 3.3b with the mean values given in table 4.2.3. All features for all data splits were statistically significant for at least $p<0.05$, except for the difference in velocity and in inverse trajectory angle between breed types, and the difference in altitude between day and night. This indicates that there should be set different threshold marker values implemented into the electronic collars depending on the situation for the sheep. All the calculated p-values are given in table 4.3.1.

Data split	**Velocity**	**Altitude**	**1/Angle**	**Temp.**
Breed type	X	<0.0001	X	<0.0001
Season	<0.005	<0.0001	<0.0001	<0.0001
Daytime	<0.01	X	<0.05	<0.0001

Table 4.3.1: The calculated p-values for the statistical significance of the differences between all data splits for all dynamic features, by the two-tailed two sample t-test.

The suggested threshold marker values for atypical sheep behavior for all the dynamical features are presented in table 4.3.2, calculated from the mean of the outlier noise points. A proposal is also included for the undivided data, thus it is possible to only implement one value for each feature if that is preferable. Both inverse angle and temperature stay roughly around the same values for all splits, even though they have statistical significance on the difference, while velocity and altitude diverges more. The percentiles of the threshold values calculated with all the data are given in the last row of the table. Velocity and temperature both has a high percentile and for altitude it is somewhat lower, but for the inverse angle the percentile is set at slightly less than half of the value distribution.

Data split	**Velocity**	**Altitude**	**1/Angle**	**Temp.**
Heavier breed	X	313	X	18
Lighter breed	X	356	X	20
Early season	919	148	79	17
Late season	478	365	93	20
Daytime	869	X	91	20
Nighttime	531	X	85	19
All data	824	350	91	20
Percentile	98.7	72.6	46.9	89.9

Table 4.3.2: Suggested digital threshold marker values for both all data splits and the undivided data. Velocity is given in m/h, the altitude in $mamsl$, the inverse trajectory angle in 1/degrees, and temperature in degrees Celsius.

4.3.1 Sheep movement against threshold values

When analysing sheep movement it is important to look at the behavioral pattern of the nearby flock, not just individual sheep. One sheep may act atypical as an exception, but if several sheep or the whole herd act atypical it is more reason to believe it might be from actual external provocations. Not all sheep in the herd wore the electronic collars, and the herd will divide further into smaller flocks of close family and friends, so it is not known which sheep are closer together at all times. However, there are known mother ewe and lamb-pairs, which will stay together throughout the season and move approximately the same. In figures 4.3.1-4.3.4 the dynamic features and their suggested threshold values for one such mother-lamb pair were plotted. The data from the pair is from 2019, and they were NKS sheep. The inverse trajectory angle varied greatly, and to make the plot in figure 4.3.4 readable and understandable, a smaller segment of the data were chosen instead. The plot shows the inverse angle for a day of high movement in late july. As the figures show, the general movement of mother and lamb is highly equal, for all features of sheep behavior. For velocity, the sheep only cross the threshold a handful of times, while for altitude, inverse angle and temperature, the sheep stay above the line for longer periods of time.

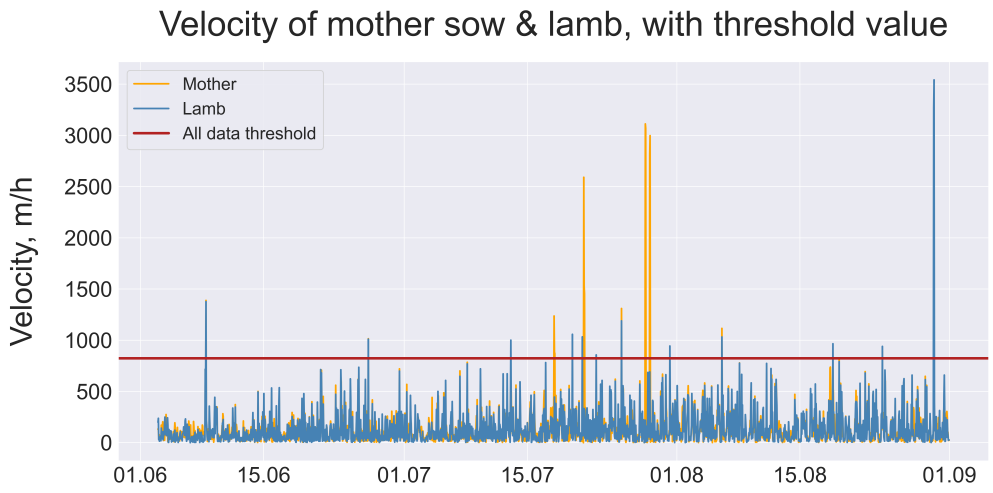


Figure 4.3.1: The velocity of the mother ewe and lamb throughout the season in 2019, against the threshold value calculated for all data.

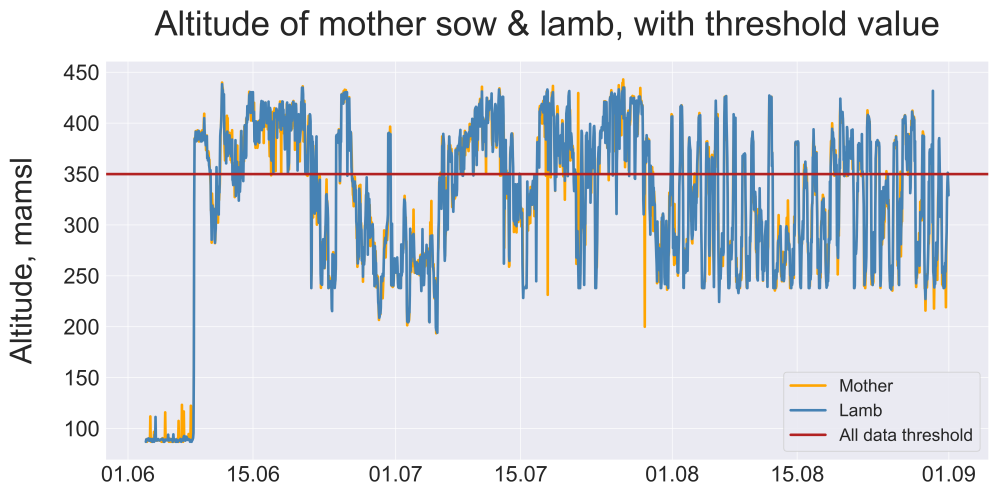


Figure 4.3.2: The altitude of the mother ewe and lamb throughout the season in 2019, against the threshold value calculated for all data.

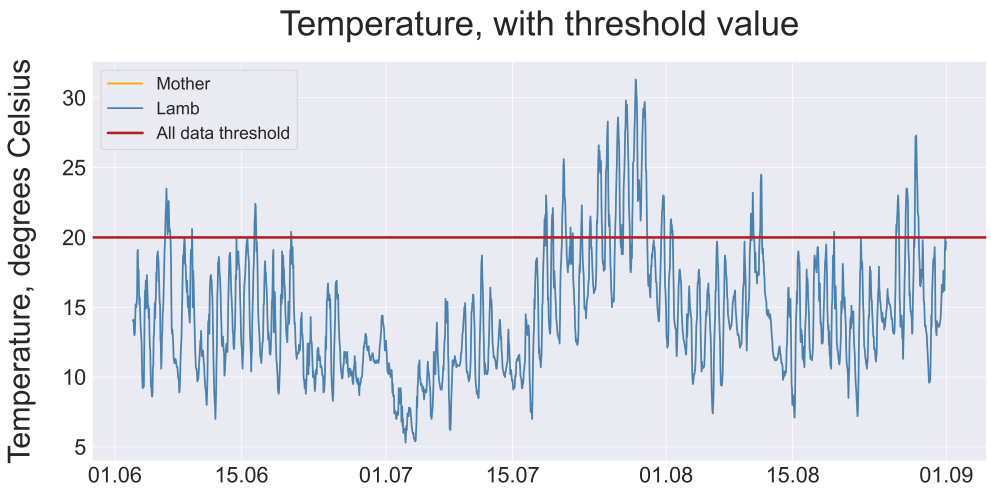


Figure 4.3.3: The registered temperature throughout the season in 2019, against the threshold value calculated for all data.

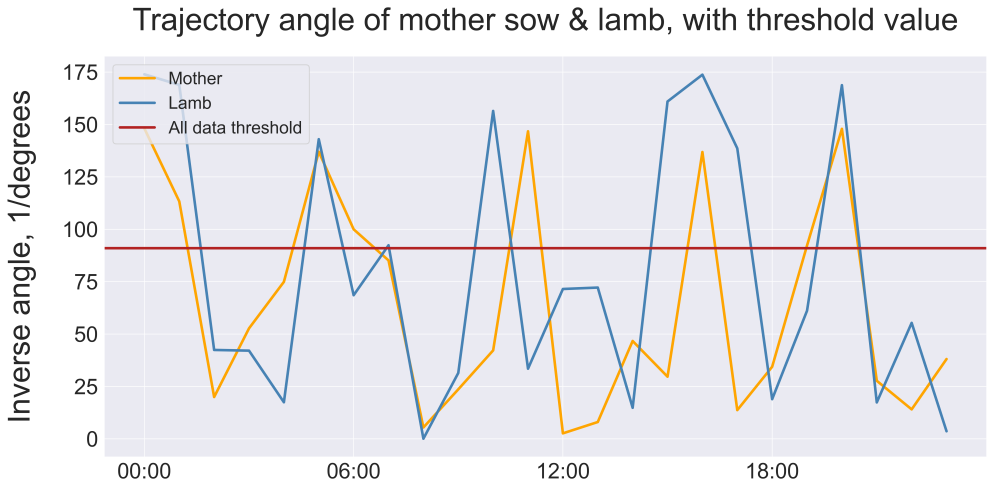


Figure 4.3.4: The inverse trajectory angle of the mother ewe and lamb throughout a day in 2019, against the threshold value calculated for all data.

Chapter 5

Discussion

5.1 Data wrangling decisions

As seen in the table of selected time ranges in table 3.4.1 and illustrated by the data set sizes for Fosen in figure 4.1.4, the trajectory set sizes were smaller in 2018 and partially for 2019 than in 2020 and the rest of 2019. They all started at approximately the same date, but some trajectory sets only lasted about a month before being terminated in the middle of the rangeland grazing season. To preserve most of the data, the larger sets were kept at full length of about two months until the end of the rangeland grazing season while the smaller sets were set at about half as big. Even though not all sets then can contribute with the same information of seasonal behavior, the smaller sets are able to provide data on approximately the full early season so this was deemed unproblematic. The start and end date variation in the data from Tingvoll were more dispersed, and dependent on which farm the different herds belonged to. The selected time ranges differed from either full rangeland grazing season, only late grazing season, or almost full leaning towards late grazing season. The overall data coverage on sheep behavior from both Fosen and Tingvoll is hence relatively balanced.

The threshold cut-off value of 15 km/h travelled by the sheep for the velocity feature might have been too high for a reasonable assumption, and can for future research be considered to be set slightly lower. A relatively tolerant maximum value had to be assumed, and given some lenience to include most target behavioral cases, but in retrospect of the analysis such a high velocity threshold before imputing new values may have contributed to skew the data some in the more extreme direction. This might have influenced the calculated digital threshold marker values to be higher than necessary, and if future research should be done with alterations to the threshold values it might be expected that the velocity threshold will turn out somewhat lower.

5.2 Feature engineering decisions

The sheep movement velocity calculated with the Haversine formula is more an attribute describing the distance travelled over an hour, and not the instantaneous velocity a sheep is moving in. The points were however not consequently generated correctly, so the time difference was not always the same between all points. To make the distance feature independent of time interval errors, the distance travelled were divided by the hourly point difference to give the velocity. For the acquired data on temperature, the closest weather stations were several hundred meters lower in altitude than where the sheep were mostly located. The temperature feature may therefore also be somewhat higher than it should be for the given data points in time and space. The externally retrieved data were however more accurate and reliable than the temperature measured by the sensors in the electronic collars.

There were several possibilities on what to study within the data and how to prioritize the analysis, and thus which feature engineering to carry out. To examine the home range on sheep which wore a collar over several years or the last behavior of the confirmed killed sheep were considered, but since the data on both these cases were very limited and not enough to be of statistical relevance this was not explored further in this study. The time available for a thesis is finite, and the main objective must be narrowed down. The aim was therefore set to look into the general data driven description of the normal behavior of sheep, and to propose digital markers of atypical sheep movement.

5.3 Statistical analysis

5.3.1 Feature correlation

The correlation between features seen in the heat map in figure 4.1.1 was generally very low, with most features very close to zero. Naturally, when the cosine value increases the sine value will decrease, and vice versa, resulting in a negative correlation. There was also some positive correlation between the age of the sheep and the number of lambs, which is further understood by the pairplot given in figure 4.1.2b with an increasing trend to the right connecting the two features. When comparing the velocity against the time of day in pairplot 4.1.2a, the velocity stays relatively steady at all times below 5000 m/h and has less point density above. The sine time is most dense below zero and up to about 0.5, while the cosine time is quite spacious in the middle and more concentrated along the extremes at positive and negative one. These times correspond to 10:00-00:00, with some emphasis on around 18:00 and at midnight. This distribution confirms unsurprisingly that sheep are the most active during the day and that this is reflected in the data, but that their activity will also fluctuate in periods during the hours of the day. The velocity and thus the activity levels are denser and more extreme the younger the age, where the lambs of age zero (born the same year), which also do not have lambs of their own, are clearly the most active. The sheep that only have one lamb is also discernibly more active than compared to the sheep with two or three lambs. This might be explained by having less

freedom and time to e.g. play as an adult the more children the sheep have to watch over. These observations are in accordance with normal expected sheep behavior. That this can be seen by data visualization establishes that the quality of the data is at least to some degree good enough to analyse real life sheep behavior.

For figure 4.1.3, the velocity is distinctly upwards limited at just below 5000 m/h for the closely compacted part of the plot, across all the other three features in the matrix. Above this value, the points spreads more. For the inverse angle, there seems to be no obvious correlation between the more spread points above 5000 and what trajectory the sheep have taken. This might imply that when highly active or when the sheep run far, they do not necessary have a preference on whether to change direction or not. The altitude has the most high velocity points around the middle. This might mean that any disturbances to the herd making them move further usually happens around a mid-height, and that when trekking upwards or further down the interference is subdued or not as much present. From figure 4.1.6 it is apparent that the sheep seek higher altitudes at night, as expected from the theory of their diurnal routines as stated by Garm et al., Tømmerberg, and Tribe [9, 12, 19]. This is also a time when they are less active, which may explain that the higher altitude has less high velocity points. From ethological theory it is already a hypothesis that sheep seek upwards at night to escape predators, get a better view and starting point, and feel safer [9, 12]. The data here confirms the behavior, without any assumptions about the reason why, which might strengthen this theory. This behavior is however of course dependent on the topographic conditions, and their possible limitations on vertical movement and how their environment looks, so this might not be true for all herds. Consequently, any threshold value for altitude will not be universal, but rather be dependent on the individual and local circumstances. For the temperature, the velocity and thus the activity of the sheep is clearly concentrated around a mild climate of about 10-15 degrees Celsius. This is exactly in accordance with the findings of Scott and Sutherland (1981) [17], where very hot or cold days resulted in more docile sheep and less movement. It is therefore expected that high activity when it is also very hot or very cold will be classified as atypical behavior. For very cold temperatures below 5 degrees, the sheep sought a middle height, staying away from the lowlands and high altitudes.

5.3.2 Data sets after wrangling

By comparing the final data set sizes after the data wrangling in figure 4.1.4 with the set sizes of the raw data enclosed in appendix C1, the average is a lot more even and uniform. The data from Fosen has two distinct set size lines, with much less spread around these levels and less deviating variations in between after the data wrangling. The data set numbers are chronologically ordered, which shows how the data from 2018 and partially 2019 are prominently smaller than those from 2020 and the rest of 2019. A cut-off threshold for when a data set was to be considered too small to be a part of the analysis was defined at 10 % of the average, and is indicated by the orange line in the figures. For the pre-wrangling figure in the appendix, several sets had to be cut away, and after cleaning and removing unwanted sets none were below the cut-off value. The data set sizes in

Tingvoll were much more varied and dispersed, and some were also much larger in size compared to the ones in Fosen. One of the reasons for this was because many electronic collars were not turned off after the sheep were herded home at fall. When a set date range were selected for each year and each place, and the data from Tingvoll were reduced to hourly point generation, the set sizes became more uniform across both places.

5.3.3 Description of normal sheep behavior

The mean velocity per hour of the day in figure 4.1.5 has two local maxima at 06 and 19, indicating the sheep's most active periods. The mean is significantly skewed upwards from the median, signifying a wider range of higher outlier values. The median is likewise centered below the middle of the IQR meaning that the lower activity values are more dense than the higher values as half of all data lie below this line. The sheep will alternate between grazing and active play, and relaxing, and the division seen in the figure is comparably equal to that in figure 2.1.3. Although there has been used different units of measurement (velocity versus activity percentage), both graphs show a unit of movement flowing in two maxima and two minima. However, the positions of the extrema are somewhat skewed compared to each other. The sheep studied by Tømmerberg were later active with a preferance towards during the middle of the day, whereas the sheep in this study have been more active at break of day and early morning. This difference is deemed not significant, and the behaviors seem to have the same qualities. The sheep in Fosen and Tingvoll both independent from each other had the same activity pattern with maxima at 04-06 and 18-20, and minima at 12 and 00 as seen in appendix C2. Sheep in Tingvoll had a slightly more extreme movement pattern, with higher velocities in general and being more active during the early hours of the day compared to Fosen which were most active in the evening. This could be due to just herd differences, sheep personalities, or the topographic environment, but might also be because of predatory conditions. If there are more predators hunting in the morning in Tingvoll than in Fosen, or generally just more predators, this could affect the data in this way. However by looking at appendix C3, there are considerably more outliers of the activity in Fosen compared to Tingvoll, which may be used to argue for the opposite. While the outliers for the Fosen data is more dispersed and even throughout the day, the outliers for the activity in Tingvoll are centered at two instances, one during the morning around 05-06 and one in the early afternoon around 16. From table 1.1.1, there are five times more sheep deaths because of predators in Trøndelag compared to Møre og Romsdal, so a higher predator density in Tingvoll may be a reasonable assumption. However, since no labeled data is available it is not practicable to conclude if the differences seen possibly come from predators or some other explanation. Hence, the data from Fosen and Tingvoll were merged to give the machine learning models as much material to work with as possible.

As per the ethological theory of sheep behavior and shown in figure 4.1.6, the sheep will on average seek higher altitudes at night, and come down again at dawn. This behavior is observable and evident in the data, but the difference in mean altitude is however small. The mean difference is of about 30 meters, which is not very pro-

nounced. This might simply be because the topography of the herds are relatively level. The mean inverse trajectory angle plotted in figure 4.1.7 will flow contrary to that of the mean activity per hour, with maxima at midnight and midday, and minima at 04-06 and 18-20. This means the sheep will change directions the least during the times with the most activity, and at times where the altitude changes the most. This could just imply that when moving to change pasture from the nightly relaxation to the daily grazing, they move relatively straight forward and with a purpose. The mean and median stay roughly the same, at approximately the middle of the box plots, indicating that the dispersion of inverse angle values are quite symmetric about the median. The trajectories taken by the sheep will thus roughly be of equal probability of being acute or obtuse. Per table 4.1.1, the total range is naturally of 0-180 degrees, where the inverse median is at 99.59 and the mean at 93.75, both above the right angle 90 degrees. The data is consequently moderately skewed downwards, with a wider and more spacious range at lower values below Q2. The feature is the inverse angle, so that the higher the value, the less the change in trajectory. The behavior of the sheep hence nominally leans towards not making extreme directional changes.

The violin plots of the activity distribution on figure 4.1.9 further confirms the activity behavior pattern of the sheep, with both most intensity and amount happening at dawn and dusk.

5.4 Machine learning models

Machine learning algorithms are often stochastic and use some form of randomness when optimizing and training, e.g. how the starting seeds for each run are chosen. This will often result in different outcomes every run. Even though the model results varied some each round, there were a limited amount of possible outcomes, and the findings were approximately equivalent nevertheless.

5.4.1 *K*-means and diurnal behavioral patterns

The k-means model identified four characteristic activity periods for the normal behavior of sheep throughout the day, that were comparatively consistent with the acquired mean activity flow and the mean activity density distribution. Theory stated that the most active grazing periods will be at dawn and dusk, which are here discovered as two of the distinctive diurnal behavioral stages. The different stages and their characteristics are summarized below, based on the results from k-means, statistics, and established theory.

- **04:30-10:30:** The sheep wakes up, migrate to lower altitudes, and have their first grazing period of the day. This stage is hereby referenced as *the first grazing period*.
- **10:30-16:30:** The sheep will relax more during the peak of the temperature and day, and chew cud from their first grazing period. They may also move about and graze some more, but less intensively. This stage is hereby referenced as *the moderate period*.
- **16:30-22:30:** Grazing intensity and movement activity will increase, and the sheep will seek upwards in the terrain towards higher altitudes again. This stage is hereby referenced as *the second grazing period*.
- **22:30-04:30:** This stage will be calm and restfull, the sheep will mostly sleep, relax and chew cud. This stage is hereby referenced as *the resting period*.

The resulting time intervals of the activity periods are quite similar to that of the violin plots in figure 4.1.9, only shifted half an hour later. From the results of k-means including all numerical features, the different day periods identified by the four output clusters were roughly the same as those stated above, only slightly skewed. If compared, the red cluster in figure 4.2.2 corresponds to the first grazing period, the purple to the moderate period, the blue to the second grazing period, and the green to the resting period. By evaluating the cluster centroids by the mean sine- and cosine-time pairs in appendix F on the k-means clock view in appendix E, and accordingly assigning the points to the closest cluster center, the time shift from including all features in the model can be estimated. The first grazing period is somewhat extended in time in both ends, making it the longest active period. The moderate period is hence shifted clockwise, but remains seemingly the same length. The second grazing period is decreased in both ends, and thus the resting period is correspondingly shifted counterclockwise. The changes

do not appear to be more than about an hour back or forth at each borderline from the original time stamps stated. With more features and more information, it is more likely the model will be more accurate when describing the behavior of the sheep. The time ranges of the behavioral stages may therefore be updated to stage 1: 04:00-11:30, stage 2: 11:30-17:30, stage 3: 17:30-21:30, and stage 4: 21:30-04:00. These limits are however only an estimation and are not discrete, the behavior will flow gradually from one characteristic into the other. By comparing the results with the approximate and schematic diurnal rhythms proposed by Garmo et al. in figure 2.1.2 the activity periods look remarkably alike, strengthening the ethological theory. The diurnal behavior has however here been demonstrated and established by data driven verification.

All periods have a low velocity mean at almost negative one, indicating that normal sheep behavior for all parts of the day is to remain relatively docile. This heightens the importance of high velocity as a trigger for atypical behavior. The temperatures are divided into two groups, appropriate for the nightly and the daily climate. The second grazing and the resting period are placed at higher altitudes than the other two. The resting period has a higher mean of trajectory change, where the first grazing period and the moderate grazing period have the lowest and thus the most linear movement. The standard deviation is relatively high for the inverse angle for all clusters, meaning that the range of values are more dispersed and spread out from the mean. The altitude varies the most for the second grazing period, while the temperatures at day are more diverse than the temperatures at night. The rest of the features are not directly relevant to analyze as only the effects of the dynamic features will be used for marker threshold values. Further, even though the static features like age and number of lambs are a part of the model to determine herd behavior as a whole and therefore impact the given values for the dynamic features, they do not influence the results as much. However, it is worth noting that for the second grazing period, the most prominent sheep in the data were the older and the ones with more lambs, while the younger sheep were the most prominent for the moderate period. It is expected that the more lambs to feed, the more a mother ewe has to eat and therefore graze more actively, which might explain this correlation. On the contrary, the younger sheep do not need to eat as much and might require more sleep, and therefore might be more still and thus influence the moderate period the most which tends towards slightly less activity.

The k-means algorithm does not handle well non-globular, anisotropic data, because k-means tends to select spherical groups. However, this is not problematic as long as the clustering makes sense and is intentional. When observing only the time versus the velocity, the data is cylindrical, and there is a big distance between points that are far from each other in time. These points will therefore not be clustered together with k-means. The objective to only look at time versus velocity was to determine the characteristic activity periods for the sheep during a day, using the most important measure for activity. It is intentional and suitable that the resulting clusters have to be limited by time periods close to each other to fulfill this purpose, instead of having points of the same behavior connected over all clock values. It is therefore interesting that when the points are more

complex in higher dimensions with more features included, k-means still returns approximately the qualitatively equivalent result.

5.4.2 DBSCAN and atypical behavioral patterns

DBSCAN is density-based and calculates the clusters by considering if the next point over is close enough. The cluster distribution will therefore not necessarily be as polarized as with k-means, but more layered and mixed for more complex data. It can also be explained visually by how a lot of the lower velocity points in figure 4.2.1 will be dense enough to be assigned to the same cluster all along the circle of time, while the next cluster would rather most likely be on top of that on a level above where the velocity points starts to diverge more. This principle continues as more dimensions are added and as the distribution becomes more complex. Describing the diurnal typical behavior of sheep is therefore not as straight forward when using DBSCAN. As can be seen in figure 4.2.3a, where both the mean sine and cosine time are about zero for all clusters. This confirms that the clusters made by DBSCAN found similar behavior across all times of day, such that the mean is zero.

K-means with the help of statistics described the normal behavior of the sheep, but to depict the atypical behavior DBSCAN is a better algorithm. The goal of optimizing the algorithm was to make the normal clusters become similar to that of k-means, such that the assumed regular behavioral points are preserved in some typical cluster also in DBSCAN. The typical clusters made from DBSCAN are as explained above more complex to describe and understand, however the main motivation for implementing DBSCAN is rather the binary task of separating the typical behavior from the outlier noise points such that the outliers alone can be described. As with the typical results, both mean sine and cosine time is centered approximately at zero for the noise, probably because there will exist outliers all around the clock. They are nonetheless both skewed a small amount in the negative direction, that is towards the moderate and the second grazing period, indicating that there is some predominance of outliers at these times.

From the statistics in table 4.2.2, all feature means except the inverse angle is elevated from the k-means normal point values in table 4.2.1, confirming that the outliers are more extreme than the ordinary. The outlier standard deviation is considerably higher than for most normal clusters, especially for time and angle. The mean inverse angle is also situated around zero, establishing an equal weight of values above and below. From this it can be concluded that atypical behavior may therefore occur at all times of day, and as previously discussed this also furthers the hypothesis that the sheep may not have a preference on directional changes when stirred. The outlier values for age, temperature and altitude have a slightly above zero mean, and a standard deviation of about 0.5. The ranges are here compressed from -1 to 1, where zero is the relative middle value. The most of the atypical behavior happens by adult sheep of most ages, but more for the above average age and less for the very young lambs or the older ewes. The same principle applies to temperature and altitude, where it is expected most atypical behavior at a slightly above middle height, and on warmer days, but the

standard deviation range mostly only excludes the extreme altitudes and temperatures. This is the same conclusion as expected from the correlation plots.

For the three features altitude, inverse angle and temperature, the median and the mean are comparably similar, while the third quartiles are not disproportionately high compared to the means. A low difference between median and mean dictate unskewed normalized data, and ensures that the mean is a good measure to represent the atypical behavior for that attribute. For velocity however, the mean is substantially higher than the median, and the third quartile is over six times closer to the median than the mean. This reveals how for the velocity there will be a few very high outliers skewing the data. The mean may therefore be higher than it perhaps ought to be to describe when atypical behavior occurs. When suggesting threshold marker values the mean will anyhow be the measure used as a starting point, and more research will instead be needed to test and examine how they should be altered to be more optimal.

The amount of outliers expresses the frequency of deviant sheep actions present in the data. While it will not give information on what type or how extreme, it may work as a measure of expectations for the atypical behavior. For the data split on breeds the amounts were quite similar, while significantly higher for the late season compared to the early season. The late season data was the split with the most outliers of all. It may therefore be expected that there are more alerts of atypical behavior, and thus maybe more trouble with predators, in the late season. The amount of outliers at day versus night were also very different, with the most at day. This is most likely a consequence of the sheep sleeping and staying mostly put at night when it is dark.

5.4.3 Proposed digital threshold markers

The results of the calculated statistical significance lead to that the only data split configurations that did not have statistical significance were the velocity difference and inverse angle between heavier and lighter breeds, and the altitude distance between day and night. This is somewhat unexpected, and diverges from the ethological theory and from the findings of Hansen et al. [8]. It could be the case that there is not a big enough difference in the behavior, but it is also important to note that more experimentation and analysis should be done before any conclusions can be made. The given results could be due to other causes such as the topography, breed composition within the herd, or insufficient data. A difference in altitude between night and day could still be seen in the statistical representation in the EDA, even though it did not here result in statistical significance.

Another aspect to consider is the statistical test in itself. A t-test will usually assume that the samples are normally distributed and randomly sampled from the population. In this case the populations, that is the data split configurations, were not necessarily normally distributed. The samples, that is the different noise data for each data split, were not randomly sampled either as they were selected by the machine learning model based on patterns found in the population data. The size of the noise data were relatively large, and the central limit theorem

states that a sample size of over 30-50 can be assumed approximately normally distributed if randomly sampled. Since the noise data were not randomly selected, the hypothesis testing does not meet all the requirements, and the calculated statistical significance should be evaluated with some skepticism. Other hypothesis tests may be more appropriate, like Z-testing or Mann–Whitney U testing, which may be considered for future examinations.

The threshold values for temperature for all splits were centered not too far from the one given for all data at 20 degrees Celsius. The difference between the temperature furthest from this value (17 degrees) is not deemed contrasting enough to be of consequence. The proposed threshold marker value for atypical behavior is therefore set to that of all the data at 20°C. In its own the temperature does not describe the behavior of the sheep, it is in combination with other characteristics that this becomes important. It has been hypothesised in prevoius research that sheep will generally be more docile above and below 10-15 degrees Celsius [17]. The mean velocity threshold values given by the noise points are all well above the mean for the normal behavior, showing that as expected the atypical behavior the machine learning model alludes to is high activity, and thus especially at higher temperatures. The combination of warm weather and high velocity should be a clear sign of herd irregularities. The altitude values are mostly positioned around 350 *mamsl*, except for the data for the early season. As seen in figure 4.3.2 the altitude is very low at the beginning of the season, either because the sheep have not yet been released for the rangeland pastures from the farm or because they stay close to the farm in the beginning before moving on further. Either way, the danger of predators close to the farm will be relatively small compared to up in the mountains, so the probability of atypical behavior because of predators at this altitude will be low. The altitude later in the season varies around the same levels, suggesting that the low mean value for the early season is skewed by the first few days, and will also gradually converge towards about 350. It might therefore be a safe assumption that a lower threshold value for altitude for the early season will not be necessary. The proposed threshold marker value for altitude is as well set to that of all the data at 350 *mamsl*. Furthermore, the inverse angle points are evenly dispersed, and it is not clear whether the atypical behavior happens above or below the given threshold value. This is heightened by that the threshold value for all the data, which is quite equal to all other data split values, is only at the 46.7^{th} percentile. In figure 4.3.4 it can be seen how the inverse angle varies approximately equally above and below 90 degrees over the course of a day. The trajectory angle is thus proposed dropped as a trigger property.

The velocity threshold values are quite varied for the different data splits, and are proposed kept separated for the early and late season, and between day and night. For velocity and temperature the percentiles for the threshold values for all data are quite high, calculated from data that already consists of outliers only, ensuring that the value most likely will describe atypical behavior. The altitude percentile is somewhat lower, and it could be considered to set the threshold even higher. Further, the local topography will vary greatly, so to generalize it will be best to use the percentile value for the altitude for the approximate local range the herd will move in, instead of the fixed number 350 mamsl that might not apply in dif-

ferent areas. The same may be considered for also the temperature, but this might not vary locally just as much. By comparing the final proposed threshold marker values with the total statistic data in table 4.1.1, they are well above the mean, ensuring that when triggered something irregular most likely have happened.

If these threshold values were to be implemented in electronic sheep collars, there are several ways of executing it. Atypical sheep behavior is not necessarily decided by one characteristic, like for example high activity, but a combination of several contributions. Whether a feature trigger is able to detect irregularities on its own, or needs to be connected to another feature trigger before alerting the farmer, have to be investigated further. Another thing to consider is how complex the alerting system should be, by assessing whether to enforce threshold markers for all features suggested, and whether to split on season and/or daytime. By looking at figures 4.3.1-4.3.4, the only feature that crosses the threshold as an exception and does not have long periods of being above is the velocity. This suggests that perhaps the velocity is able to be implemented as a trigger alone, and emphasizes that the other features need to be in combination with other triggers in order to be functional for atypical behavior.

The proposed threshold values for the atypical behavior are heightened from the normal, indicating they point to instances where the sheep are provoked and able to move freely. They will not detect sick sheep, as their behavior when sick will on the contrary be lethargic and slow. From the interviews of farmers done for the specialization project, they reported that the sheep behavior and reaction pattern to predators will also depend on individual personality and the type of predator. The threshold values may therefore as well vary with different herds and different areas where other predators are more prone.

5.5 Future work

The features temperature and altitude had to be retrieved from external sources as the in-collar generated values were not reliable enough for the analysis. If these features should be used for monitoring live behavior against threshold markers, the firmware and sensors that generate the different attributes should be dependable enough to be able to rely on the alerts of irregularities when they happen. An alternative is to look further into parallel and/or sequential irregularities in behavior, where one point of triggered threshold values is not enough to set off any alarms, but e.g. several in the flock at the same time must by triggered where the collars can communicate with each other, or several points in a row for one sheep must be irregular. The latter is however not recommended if the time interval is one hour, as many hours of predatory danger before making the farmer aware will most likely be too long to be able to save the sheep. Nevertheless, looking at more than a single trigger will make the alerts more reliable for actual atypical behavior worth checking up upon.

An initial analysis has here been done based on unsupervised data, and different threshold values has been proposed for different attributes affecting the sheep be-

havior. The next steps going forwards from this are to implement these values into the electronic collars to give alerts when triggered, and to test their accuracy by observing the sheep behavior both at typical and atypical points. If possible, labeled data of predatory presence to the herd should be collected and analysed. New and more precise threshold values should then be computed, and the electronic collars updated. More features and their thresholds could also be developed, and further research should be done on which sheep attributes and which feature trigger combinations are the most important when it comes to predatory reactions, and which that give the most accurate results for alerting when predators are nearby. It is also recommended to look further into the different data split configurations, and to test their individual threshold values specifically to see if split values make a difference in real life or if all data values work just as well. Deeper levels of data splits could also be made, like investigating the behavioral difference of day and night at early season for the lighter breeds, and so forth. Lastly, it would be interesting to test the threshold values on herds in new areas to see if the conclusions would be the same and how well the results translate to the general case.

To gather labeled data in order to do a supervised analysis should be of interest for the future work on sheep welfare. It is recommended that if new data is to be collected in new herds that these are chosen from more predator prone areas. It is expected that the most vulnerable sheep in the herd are at the highest risk of being taken by predators, hence it might be valuable to install electronic collars on sheep that are older, lonely grazers, low rank or lambs. Further, it would be an advantage if there were data over several years where the sheep wearing the collars were consistently the same individuals each year so the home ranges to a greater extent could be investigated. Sheep are habitual animals, and deviations from their usual home range is an established theory of atypical behavior.

Chapter 6

Conclusions

The main objective with this study was to look at both normal and atypical sheep behavior not only from ethological theory or observations done by the naked eye, but by data driven verification. From the analysis done for this thesis, and in accordance with other published theories, it has been established that sheep on rangeland pastures will have a habitual and regular diurnal routine. Their most intense periods of movement will be at dawn and dusk, and their activity levels will rise and fall throughout the day. The characteristic activity periods, determined by a k-means machine learning model, were termed the first grazing period (04:00-11:30), the moderate period (11:30-17:30), the second grazing period (17:30-21:30) and the resting period (21:30-04:00). The first and second grazing periods are the most active, while the moderate period involves more resting during the peak of day, and the resting period will mostly be for sleeping. They will seek higher altitudes at night, but this may not necessarily be a big difference in height, and for the calculations the altitude difference proved not statistically significant.

Typical sheep behavior is generally to remain relatively docile. They prefer more moderate temperatures, and will be more active at around 10-15 degrees Celsius. At higher or lower temperatures than this they will stay more still, and any intense activity at these times is atypical behavior for the sheep. Detected atypical behavior were the most frequent among adult sheep, and less for young lambs and older ewes. When agitated the sheep did not show a preference on directional changes, and would either modify their trajectory or not with about equal probability.

Digital threshold marker values of atypical behavior were computed, and are suggested to be implemented in the electronic collars on the sheep to alert the farmer of irregularities. For the temperature the threshold were proposed at above 20 degrees Celsius, or at above the 90^{th} percentile of the local summer climate. The altitude threshold were proposed at above 350 $mamsl$, or at above the 72.6^{th} percentile of the local topography. The velocity threshold were proposed at over 824 m/h travelled if only one value for velocity should be used in the collars, or at above the 99^{th} percentile. For several situational splits on the velocity, the proposed trigger values are proposed at above 919 m/h for the early season and above 478 m/h for the late season, and at above 869 m/h during the daytime periods and above 531 m/h during the nighttime period. The trajectory angle

is recommended not being used as a behavioral attribute. Further research and testing are needed to validate and optimize the threshold values, and to analyse what combination of triggered threshold markers before sending an alert gives the most accurate results of e.g. predatory presence to the herd. Should new research be initialized it is recommended to collect data from more predator prone areas, and to put the electronic collars on the same sheep individuals over several years to investigate their home range variations. To increase the chances of obtaining more labeled data it is also preferable that the collars are set on the more high risk and vulnerable sheep in the herd.

References

[1] SSB. *Gardsbruk, jordbruksareal og husdyr*. Feb. 2022. URL: https://www.ssb.no/jord-skog-jakt-og-fiskeri/jordbruk/statistikk/gardsbruk-jordbruksareal-og-husdyr.

[2] SSB. *12660: Husdyr på utmarksbeite (k) 1995-2020*. 2020. URL: https://www.ssb.no/statbank/table/12660/.

[3] Rovbase. *Erstatning for sau*. 2021. URL: https://www.rovbase.no/erstatning/sau.

[4] N. Salvesen. "*Statistical Analysis Of The Movement Pattern Of Sheep And The Occurrence Of Predators*". MA thesis. Trondheim, Norway: Norwegian University of Science and Technology, Dec. 2021.

[5] J. R. E. Johanssen and K. M Sørheim. *Atferd og velferd hos sau*. May 2021. URL: https://www.agropub.no/fagartikler/atferd-og-velferd-hos-sau.

[6] P. Le Neindre et al. "*Influence of breed on reactivity of sheep to humans*". In: *Genetics Selection Evolution* 25 (Nov. 1993). DOI: 10.1186/1297-9686-25-5-447.

[7] Encyclopaedia Britannica. *Sheep*. May 2020. URL: https://www.britannica.com/animal/domesticated-sheep.

[8] I. Hansen, H. S. Hansen, and F. Christiansen. "*Kartlegging av antipredatoratferd hos ulike saueraser*". In: (1998).

[9] T. H. Garmo et al. *Saueboka*. 2nd ed. Landbruksforlaget, 2006.

[10] Norsk Sau og Geit. *Sauerasene i Norge*. URL: https://www.nsg.no/sau/saueraser/.

[11] A. K. Doughty et al. "*The influence of lameness and individuality on movement patterns in sheep*". In: *Elsevier* (Mar. 2018). URL: https://www.sciencedirect.com/science/article/pii/S0376635717302449.

[12] W. O. Tømmerberg. "*Atferd hos frittlevende domestiserte sauer på fjellbeite*". MA thesis. Trondheim, Norway: University of Trondheim, June 1985.

[13] L. A. Syme. "*Social disruption and forced movement orders in sheep*". In: *Elsevier* (1981). URL: https://www.sciencedirect.com/science/article/pii/S0003347281801765.

[14] V. R. Squires and G. T. Daws. "*Leadership and dominance relationships in Merino and Border Leicester sheep*". In: *Elsevier* (1975). URL: https://www.sciencedirect.com/science/article/pii/030437627590019X.

[15] W. H. Burt. "*Territoriality and home range concepts as applied to mammals*". In: *Journal of Mammalogy* (Aug. 1943). DOI: https://doi.org/10.2307/1374834.

[16] P. A. Jewell. "*The concept of home range in mammals*". In: *Symposium of the Zoological Society of London* 18 (1966), pp. 85–109.

[17] D. Scott and B. L. Sutherland. "*Grazing behaviour of merinos on an undeveloped semi-arid tussock grassland block*". In: *New Zealand Journal of Experimental Agriculture* 9 (Jan. 1981). DOI: https://doi.org/10.1080/03015521.1981.10427794.

[18] G. P. Hughes and D. Reid. "*Studies on the behaviour of cattle and sheep in relation to the utilization of grass*". In: *Journal of Agricultural Science* 41 (4 Oct. 1951), pp. 350–366. DOI: https://doi.org/10.1017/S0021859600049534.

[19] D. E. Tribe. "*Some seasonal observations on the grazing habits of sheep*". In: *Empire Journal of Experimental Agriculture* 27 (1949), pp. 105–115.

[20] A. Murie and W. B. Davis. "*The Wolves of Mount McKinley. Fauna of the National Parks of the United States*". In: *Journal of Mammalogy* 26 (1 Feb. 1945), pp. 100–101. DOI: https://doi.org/10.2307/1375039.

[21] A. Stien et al. "*Kongeørn som tapsårsak for sau og lam. Tapsstudier i Rødsjø beiteområde 2014-2015.*" In: (Aug. 2016). URL: https://brage.nina.no/nina-xmlui/handle/11250/2402971.

[22] J. K. Rød. *Geografiske koordinater*. Jan. 2021. URL: https://snl.no/geografiske_koordinater.

[23] L. D. Talley et al. *Descriptive Physical Oceanography (ed. 6). Academic Press*. 2011. DOI: https://doi.org/10.1016/C2009-0-24322-4.

[24] ISO. *Geographic information — Referencing by coordinates*. 2019. URL: https://www.standard.no/en/PDF/FileDownload/?redir=true&filetype=Pdf&preview=true&item=1022546&category=5.

[25] Ø. B. Dick, J. K. Rød, and L. Mæhlum. *Geodetisk datum*. Jan. 2021. URL: https://snl.no/geodetisk_datum.

[26] Ø. B. Dick and J. K. Rød. *WGS84*. Feb. 2021. URL: https://snl.no/WGS84.

[27] European Petroleum Survey Group. *About the EPSG Dataset*. URL: https://epsg.org/home.html.

[28] Spatial Reference.org. *EPSG Projection - WGS 84*. URL: https://spatialreference.org/ref/epsg/wgs-84/.

[29] J. K. Rød and L. Mæhlum. *EUREF89*. Mar. 2021. URL: https://snl.no/EUREF89.

[30] Ø. B. Dick. *Ortometrisk høyde*. Mar. 2020. URL: https://snl.no/ortometrisk_h%C3%B8yde.

[31] J. K. Rød. *Innføring i GIS og Statistikk*. 2nd ed. Fagbokforlaget, 2017, p. 113.

[32] N. R. Chopde and K. N. Mangesh. "*Landmark Based Shortest Path Detection by Using A* and Haversine Formula*". In: *International Journal of Innovative Research in Computer and Communication Engineering* 1 (2 Apr. 2013).

[33] V. Agafonkin. "*Fast geodesic approximations with Cheap Ruler*". In: (May 2016). URL: https://blog.mapbox.com/fast-geodesic-approximations-with-cheap-ruler-106f229ad016.

[34] T. Vincenty. "*Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations*". In: 23 (Apr. 1975).

[35] GPS.gov. *GPS Accuracy*. URL: https://www.gps.gov/systems/gps/performance/accuracy/.

[36] A. Tidemann and A. C. Elster. *Maskinlæring*. June 2019. URL: https://snl.no/maskinl%C3%A6ring.

[37] T. M. Mitchell. *Machine Learning*. McGraw Hill, Mar. 1997, p. 2.

[38] D. Sarkar, R. Bali, and T. Sharma. *Practical Machine Learning with Python*. Apress, 2018. DOI: https://doi.org/10.1007/978-1-4842-3207-1.

[39] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[40] scikit-learn developers. *sklearn.preprocessing.MinMaxScaler*. Computer Software. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[41] scikit-learn developers. *sklearn.preprocessing.StandardScaler*. Computer Software. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[42] J. Kleinberg. "*An Impossibility Theorem for Clustering*". In: *Proceedings of the 15th International Conference on Neural Information Processing Systems*. NIPS'02. Cambridge, MA, USA: MIT Press, 2002, pp. 463–470.

[43] I. Dabbura. *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Sept. 2018. URL: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

[44] scikit-learn developers. *sklearn.cluster.KMeans*. Computer Software. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[45] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[46] P. Fränti and S. Sieranoja. "*How much can k-means be improved by using better initialization and repeats?*" In: *Pattern Recognition* 93 (2019), pp. 95–112. DOI: https://doi.org/10.1016/j.patcog.2019.04.014.

[47] E. Umargono, J. E. Suseno, and S. K. Gunawan. "*K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula*". In: 474 (Jan. 2020). DOI: 10.2991/assehr.k.201010.019.

[48] R. L. Thorndike. "*Who belongs in the family?*" In: *Psykometrica* 18 (4 Dec. 1953), pp. 267–276. DOI: https://doi.org/10.1007/BF02289263.

[49] scikit-learn developers. *sklearn.cluster.DBSCAN*. Computer Software. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

[50] J. Sander et al. "*Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications*". In: *Data Mining and Knowledge Discovery* 2 (1998), pp. 169–194. DOI: https://doi.org/10.1023/A:1009745219419.

[51] N. S. Chauhan. *DBSCAN Clustering Algorithm in Machine Learning*. Apr. 2022. URL: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html.

[52] G. S. Braut. *Statistisk signifikans*. June 2018. URL: https://snl.no/statistisk_signifikans.

[53] G. S. Braut. *P-verdier*. Feb. 2019. URL: https://snl.no/p-verdier.

[54] S. Dahlum and S. Grønmo. *Hypotesetesting*. Mar. 2021. URL: https://snl.no/hypotesetesting.

[55] "*Central Limit Theorem*". In: *The Concise Encyclopedia of Statistics*. Springer, 2008, pp. 66–68. DOI: 10.1007/978-0-387-32833-1_50.

[56] A. Ganti. *Central Limit Theorem (CLT)*. May 2022. URL: https://www.investopedia.com/terms/c/central_limit_theorem.asp#citation-4.

[57] Data Science Process Alliance. *What is CRISP-DM*. URL: https://www.datascience-pm.com/crisp-dm-2/.

[58] I. Sommerville. *Software Engineering*. 10th ed. Pearson Education Limited, 2015.

[59] The pandas development team. *pandas-dev/pandas: Pandas*. Version 1.4.2. Feb. 2020. DOI: 10.5281/zenodo.3509134.

[60] F. Pedregosa et al. "*Scikit-learn: Machine Learning in Python*". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[61] Kartverket. *Brukarrettleiing for høgdeprofil-API*. Sept. 2021. URL: https://www.kartverket.no/api-og-data/friluftsliv/hoydeprofil.

[62] A. C. Elster. *parallellprosessering*. Dec. 2020. URL: https://snl.no/parallellprosessering.

[63] Norsk Klimaservicesenter. *Observasjoner og værstatistikk*. URL: https://seklima.met.no/observations/.

[64] H. Hain. *Time as a Machine Learning Feature*. May 2020. URL: https://henrikhain.io/post/time-as-a-machine-learning-feature/.

[65] Keboola. *A Guide to Principal Component Analysis (PCA) for Machine Learning*. Oct. 2020. URL: https://www.keboola.com/blog/pca-machine-learning.

[66] S. A. Hill. "*Statistics*". In: *Foundations of Anesthesia (Second Edition)*. 2nd ed. Mosby, 2006, pp. 207–217. DOI: https://doi.org/10.1016/B978-0-323-03707-5.50024-3.

[67] OpenStreetMap contributors. *Open Street Map*. 2022. URL: %5Curl%7B%20https://www.openstreetmap.org%20%7D.

Appendices

A - Code

All code scripts used in this thesis are included in the Github repository linked below. Explanations are given in the readme-file. The *Python* files are:

Data cleaning:

- FosenDelete.py
- PointClean.py
- TimeClean.py
- TimeInterval.py
- TingvollReduce.py
- UpdateFormat.py

EDA:

- ActivityPerTime.py
- InfoGenerationFosen.py
- InfoGenerationTingvoll.py
- MapPlotFosen.py
- MapPlotTingvoll.py
- SizeCheck.py
- StartEndDates.py
- Eda.py

Feature engineering:

- Altitude.py
- Angle.py
- Haversine.py
- InfoFeatureGeneration.py
- TimeScale.py
- WeatherFeature.py

Machine Learning:

- DBSCAN.py
- Kmeans.py
- StatSignificance.py

Github repository link

- `https://github.com/ninasalvesen/master_thesis_ninasalv`

B - Specialization Project

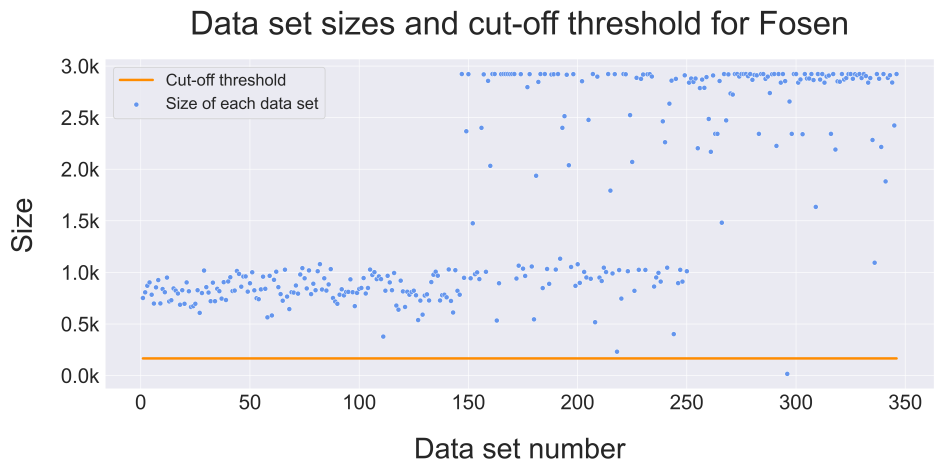
The specialization project were written during the fall of 2021, and the final report is linked below.

Specialization project report link

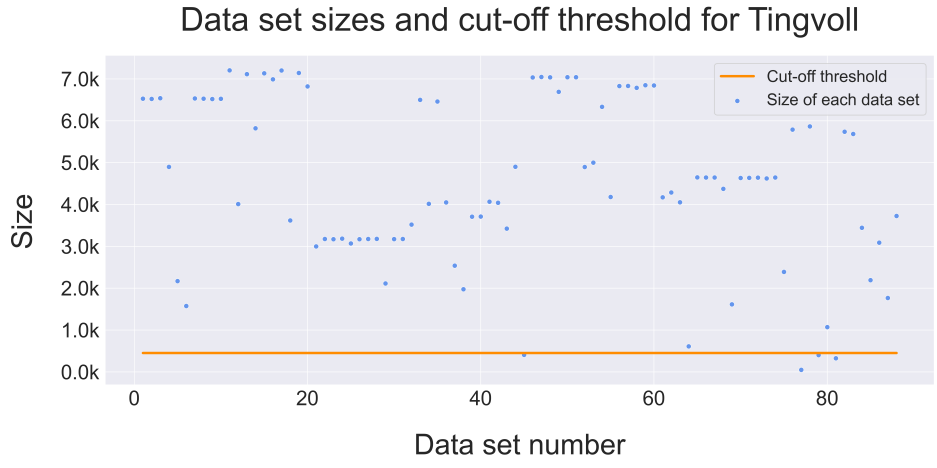
- `https://bit.ly/3NVOO5E`

C - Statistics before the data cut

C1 - Data set size



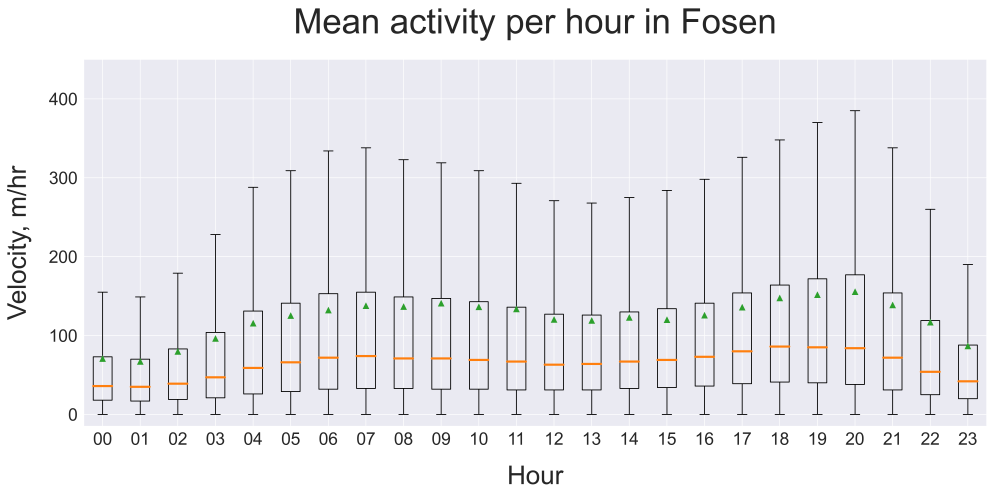
(a) Data set sizes in Fosen before the data cut.



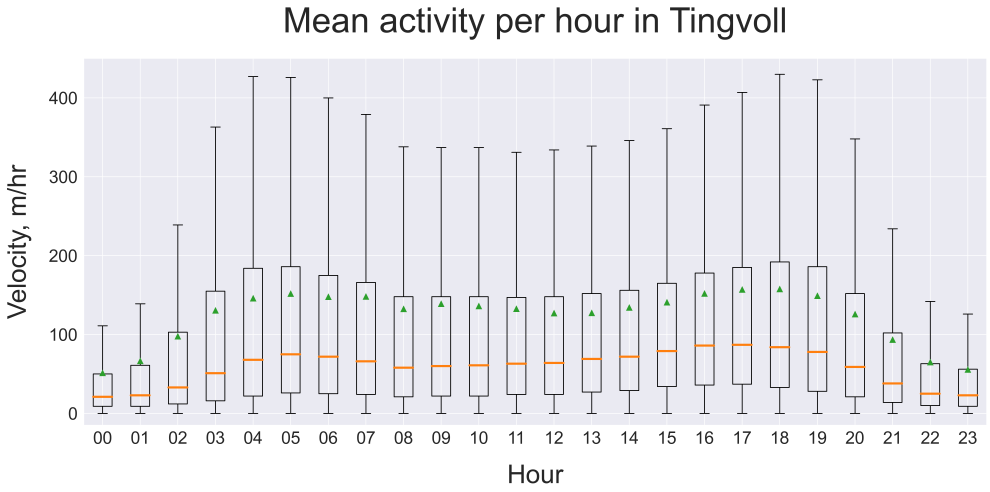
(b) Data set sizes in Tingvoll before the data cut.

Figure C.1: The figures show the trajectory data set sizes of each data set and the proposed cut-off threshold set at 10 % of the mean trajectory set size, before any information were cut off from the cleaned raw data. The data set numbers are ordered chronologically in time, from (a) 2018-2020 in Fosen and (b) 2012-2016 in Tingvoll.

C2 - Mean activity per hour



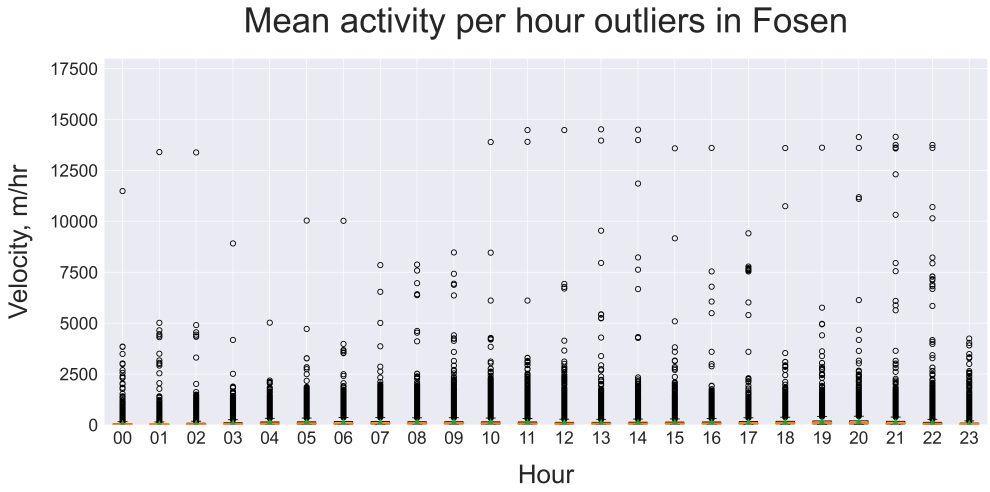
(a) Mean activity per hour in Fosen before the data cut.



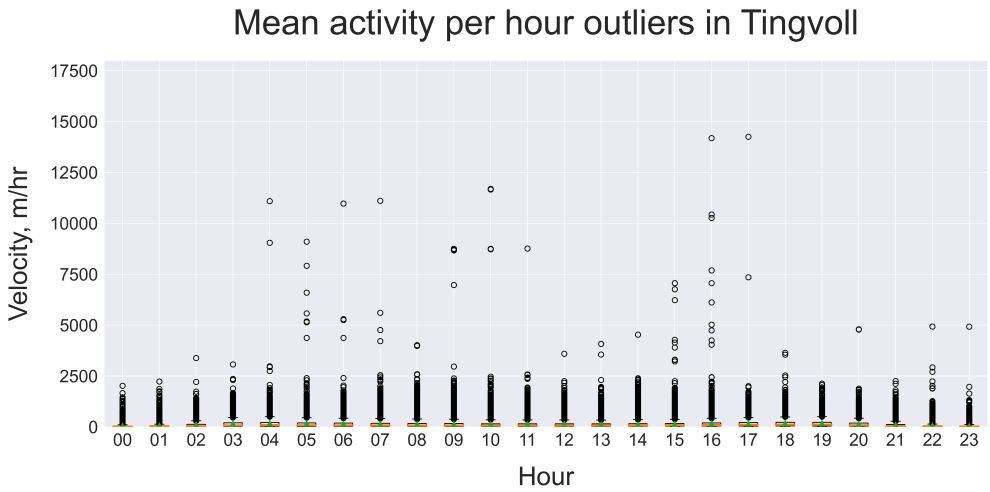
(b) Mean activity per hour in Tingvoll before the data cut.

Figure C.2: Boxplots of the mean activity per hour of the day for all sheep in (a) Fosen and (b) Tingvoll before the data cut.

C3 - Mean activity per hour outliers



(a) Mean activity per hour outliers in Fosen before the data cut.



(b) Mean activity per hour outliers in Tingvoll before the data cut.

Figure C.3: Outliers (fliers) of the mean activity per hour of the day for all sheep shown for (a) Fosen and (b) Tingvoll before the data cut. The colored marks at the bottom corresponds to the boxplots given in appendix C2, and the scatter plot represents the outliers in the data.

C4 - Mean activity per year per date in Fosen

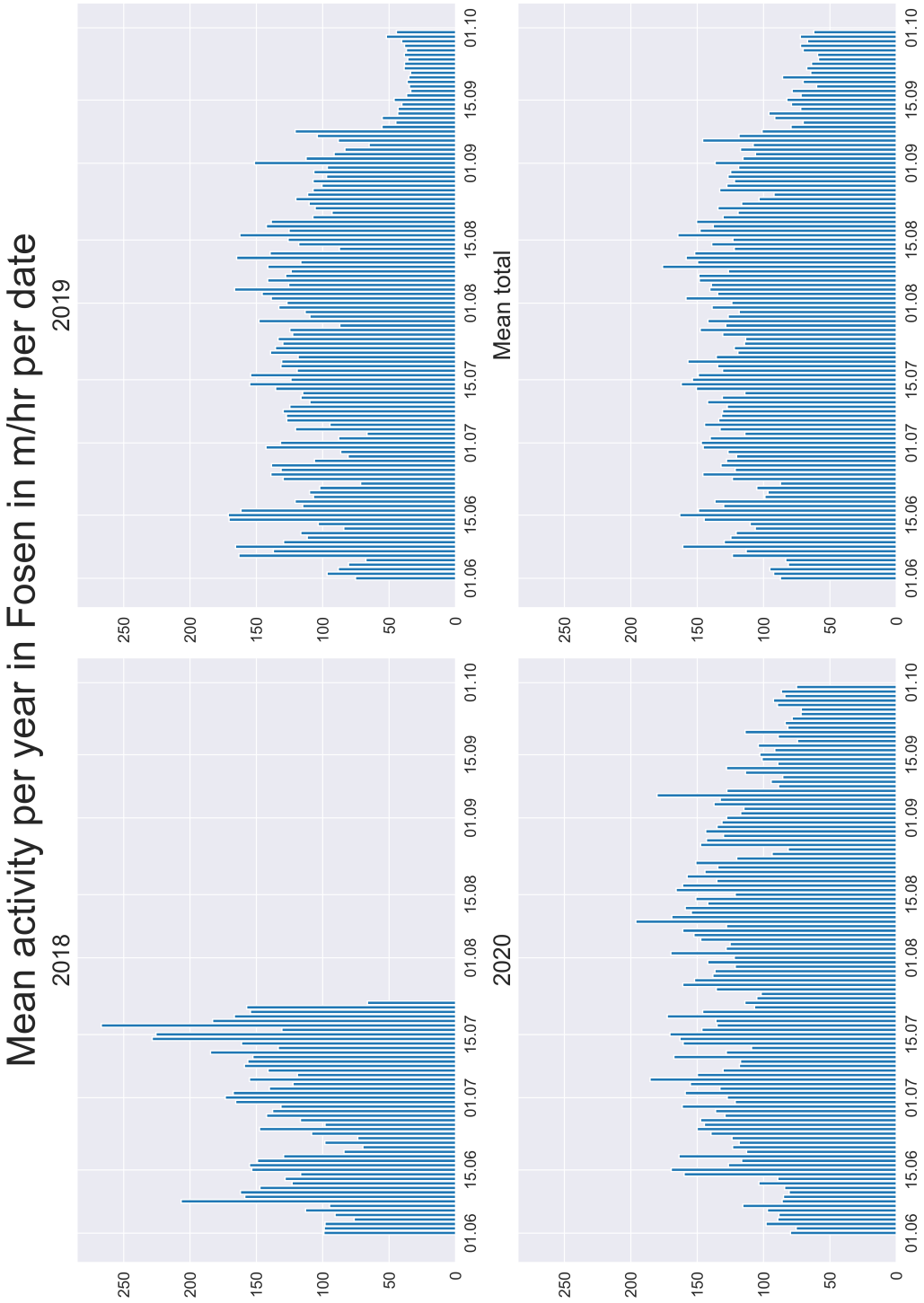


Figure C.4: Mean activity per year per date and the mean total per date for all years in Fosen before the data cut, measured in m/h.

C5 - Mean activity per year per date in Tingvoll

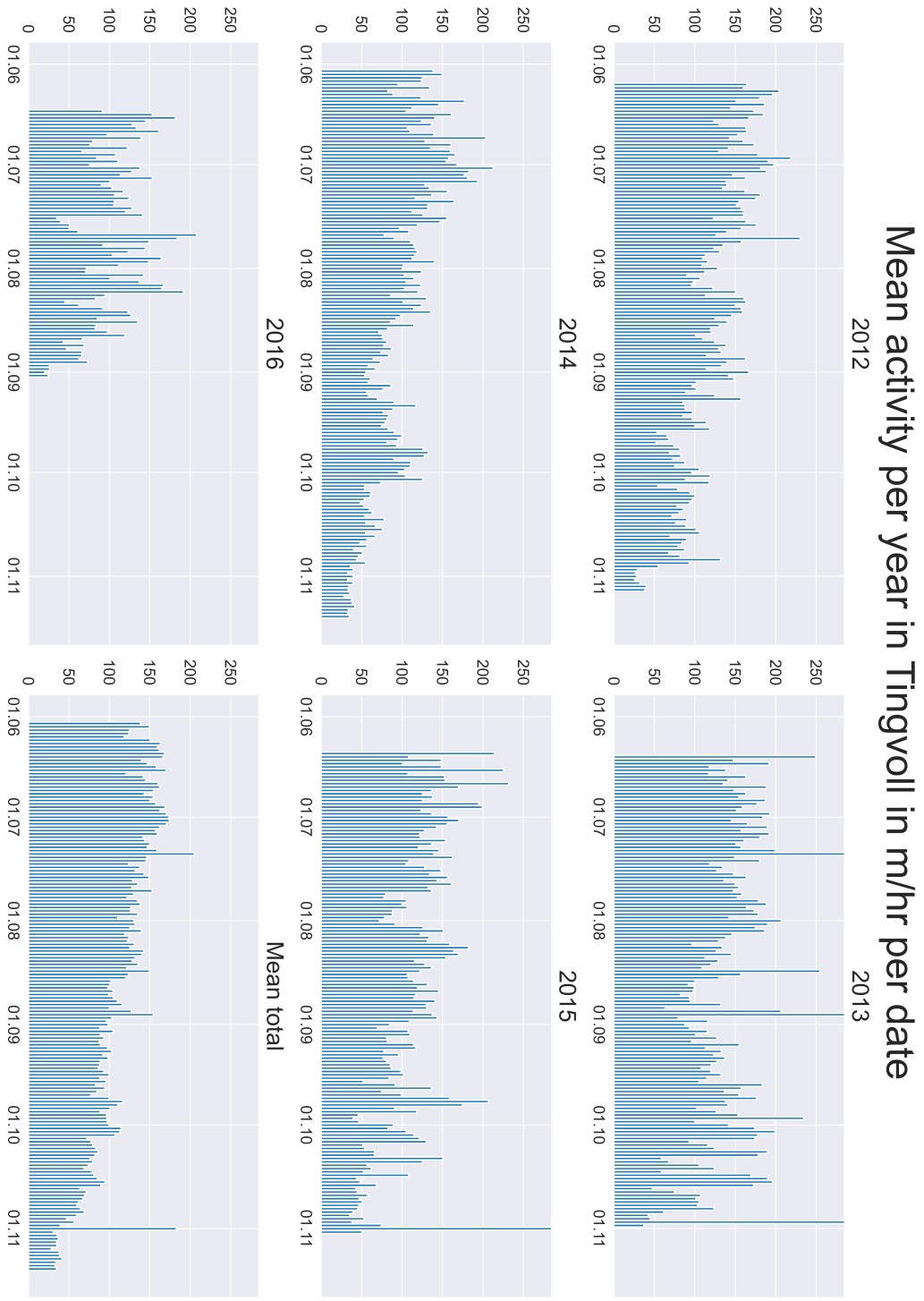


Figure C.5: Mean activity per year per date and the mean total per date for all years for Tingvoll before the data cut, measured in m/h.

C6 - Statistics on sheep velocity

Statistic	**Fosen**	**Tingvoll**
Count	387317	283960
Mean	130.66	134.18
Std	248.09	230.32
Q1	31.00	21.00
Median	67.00	63.00
Q3	142.00	159.00
Min	0.00	0.00
Max	14519.00	14253.00

Table C.1: Table of statistics on sheep velocity in m/h before the data cut.

D - Feature correlation pairplot matrix

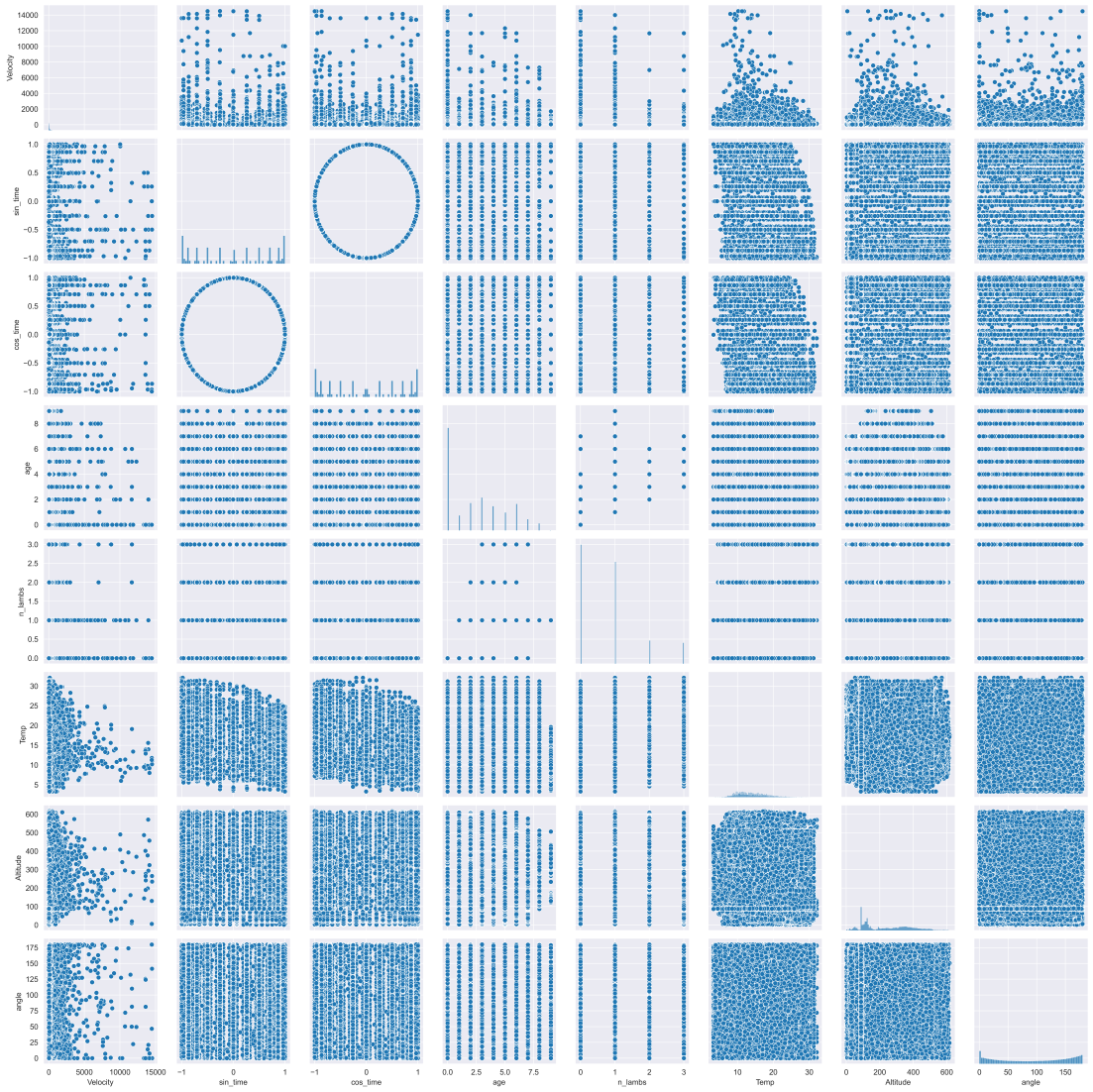


Figure D.1: Full feature correlation matrix in a pairplot, with the feature histogram distribution along the diagonal. The sequence of features from left to right, and from top to bottom, is velocity, sine time, cosine time, age, number of lambs, temperature, altitude and angle.

E - K-means results 24-hour clock view

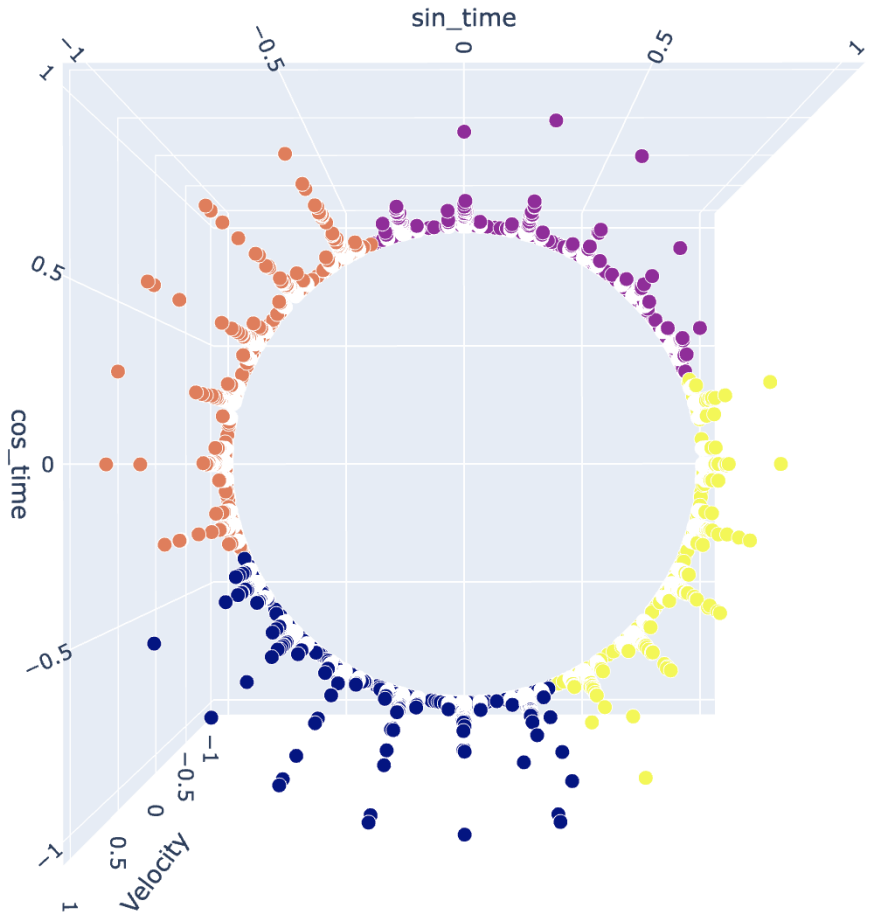


Figure E.1: Results from k-means with time and velocity, for all data, in a top-down perpective to show the activity against the 24-hour clock. The top point [sin(0), cos(1)] is 00:00, and every 90 degrees clockwise corresponds to six hours later.

F - K-means mean cluster feature values

Feature	$Mean_1$	$Mean_2$	$Mean_3$	$Mean_4$
Velocity	-0.981	-0.981	-0.983	-0.987
Altitude	-0.241	-0.256	-0.199	-0.189
1/Angle	0.005	-0.007	0.057	0.105
Temp.	-0.299	-0.130	-0.145	-0.412
Age	-0.515	-0.812	0.131	-0.529
n_lambs	-0.607	-0.827	0.321	-0.621
sin_time	0.733	-0.708	-0.372	0.145
cos_time	-0.440	-0.356	-0.167	0.826

Table F.1: Mean feature cluster statistics from the results of the k-means model for all features and all data, for the four identified clusters. Cluster one is the first grazing period, two the moderate period, three the second grazing period and four the resting period. All values are standardized and normalized.

Salvesen Nina

Master's thesis

NTNU
Kunnskap for en bedre verden