Machine learning for dairy cow behaviour classification

Manod Lewis Williams

2019

A thesis submitted at the Institute of Biological, Environmental and Rural Sciences (Aberystwyth University) for the degree of Doctor of Philosophy

Declaration

Word count of thesis: 61,450

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Candidate name …… Manod Williams

Signature……

Date…… 7th April 2020

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signature……..

Date…… 7th April 2020

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signature……..

Date…… 7th April 2020

Acknowledgements

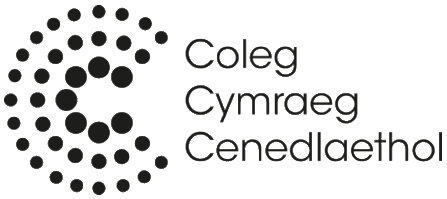
Ni fyddai'r ddoethuriaeth yma wedi bod yn bosib heb gefnogaeth ariannol y Coleg Cymraeg Cenedlaethol. This work would not have been possible had it not been for the financial support of the Coleg Cymraeg Cenedlaethol.

I would like to thank my supervisors, Dr Michael Rose and Mr Penri James for their continued support and guidance over the course of this project both from an academic perspective and also for being great colleagues.

Yn ogystal â hyn diolch i deulu fferm Brysgaga am roi fi ar y trywydd cywir ac i Rowland am awgrymu'r coleg i mi yn y lle cyntaf. Y cyfnod ym Mrysgaga yn sicr yw un o uchafbwyntiau'n fywyd hyd yn hyn. Er bod gwaith academaidd yn medru bod yn heriol, nid oes gwaith mwy caled na chaib a rhaw.

Hoffwn ddiolch i ti Emma am dy help trwy gydol y blynyddoedd diwethaf ac am ein sgyrsiau amryw ar y pwnc yma.

Yn olaf, diolch i ti mam am fod yna i mi bob dydd o'm mywyd. Ni fyddai hyn (nac unrhyw beth) yn bosib hebot ti (a Cai).



Summary

This thesis describes the use of machine learning (ML) techniques applied to data gathered from GPS receivers attached to pasture-based dairy cows for the purpose of automatic behaviour identification. Automatically identifying the behaviour of cattle will allow livestock practitioners to make more informed decisions on their management. Furthermore, daily behaviour data can be utilised for earlier disease diagnosis. For example, if the feeding duration of a cow is below its expected target then managers can intervene. Individual animal data were previously unattainable, with cattle usually managed on a herd basis. This thesis begins with an introduction that summarises the ongoing research in the field of precision livestock farming (PLF) and how farmers are implementing some PLF systems for the management of livestock. The main PLF systems discussed are those that incorporate on-animal sensors for the detection and classification of key behaviours associated with production and health. The main body of the thesis is divided into three experimental chapters. Chapter 1 (published in the Journal of Dairy Science) describes the development of a behavioural model of pasture-based Holstein dairy cows using data collected from GPS receivers and processed using ML techniques. Chapter 2 (published in Computers and Electronics in Agriculture) discusses a further modification to the behavioural model which improves its ability to categorise behaviours. Finally, Chapter 3 describes the use of a data partitioning technique often used for timeseries analysis as an alternative method for the development of behaviour prediction models of dairy cows. Chapter 3 was published in the journal Biosystems Engineering. The thesis concludes with a discussion of each chapter in light of the wider research and highlights some necessary areas for further work.

Statement of candidate's contribution to the work

This thesis, presented in alternative format is structured around three main chapters which have been submitted and accepted for publication in peer-reviewed journals. Below is a description of the contribution of the candidate (Williams, M.L) and co-authors to each chapter.

Chapter 1

A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques
Full reference:

Williams, M.L., Mac Parthaláin, N., Brewer, P., James, W.P.J. and Rose, M.T., 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *Journal of Dairy Science*, *99* (3), pp. 2063-2075.

Contributions
Williams, M. L; Mac Parthaláin, N; James, W. P. J., and Rose, M. T conceived the work and developed the project idea. Williams, M. L constructed the experiments, collected the data, performed the statistical analysis, contributed to developing the behaviour models and co-wrote the manuscript. Mac Parthaláin, N provided machine learning expertise, developed the behaviour models and co-wrote the manuscript. Brewer, P provided RTK-GPS hardware and expertise and undertook data processing for GPS unit calibration. James, W. P. J provided expertise in GIS, coordinate mapping and contributed to editing the manuscript. Rose, M. T undertook data collection, provided advice on statistical analysis, contributed to editing the manuscript and supervised this work.

Declarations

Williams, M. L .....

Mac Parthaláin, N .....

Brewer, P .....

James, W. P .....

Rose, M. T .....

Chapter 2

Fixed-time data segmentation and behavior classification of pasture-based cattle: enhancing performance using a hidden Markov model
Full reference:

Williams, M.L., James, W.P. and Rose, M.T., 2017. Fixed-time data segmentation and behavior classification of pasture-based cattle: Enhancing performance using a hidden Markov model. *Computers and Electronics in Agriculture*, *142*, pp. 585-596.

Contributions

Williams, M. L conceived the work, undertook all data processing, analysis and modelling and wrote the manuscript. James, W. P undertook coordinate mapping and assisted in creating figures for the manuscript. Rose, M. T edited the manuscript and supervised this work.

Declarations

Williams, M. L .....

James, W. P .....

Rose, M. T .....

Chapter 3

Variable segmentation and ensemble classifiers for predicting dairy cow behaviour
Full reference:

Williams, M.L., James, W.P. and Rose, M.T., 2019. Variable segmentation and ensemble classifiers for predicting dairy cow behaviour. *Biosystems Engineering*, *178*, pp. 156-167.

Contributions
Williams, M. L conceived the work, undertook all data processing, analysis and modelling and wrote the manuscript. James, W. P provided feedback on the manuscript. Rose, M. T edited the manuscript and supervised this work.

Declarations

Williams, M. L .....

James, W. P .....

Rose, M. T .....

Table of Contents

List of Figures

Introduction

Chapter 1

Chapter 2

Chapter 3

List of Tables

Introduction

Chapter 1

Chapter 2

Chapter 3

Discussion

Abbreviation

A

acf (autocorrelation function)

AF (Advanced features)

AMS (Automatic milking system)

ANN (Artificial neural network)

ANOVA (Analysis of variance)

AUC (Area under the curve)

B

BCPA (Behavioural changepoint analysis)

BN (Bayesian network)

C

CA (Classification accuracy)

CEP (Circular error probable)

CFS Correlation-based feature selection

CI (Confidence interval)

CLI (Command line interface)

CMS (Conventional milking system)

CNN (Convolutional neural network)

cpm (changepoint model)

CUSUM (Cumulative sum control chart)

D

DBA (Dynamic body acceleration)

DIM (Days in milk)

DM (Dry matter)

DNN (Deep neural network)

E

EGNOS (European geostationary navigation overlay service)

F

FCV ([*n*]-Fold cross-validation)

FM (Fresh matter)

FN (False negative)

FP (False positive)

FTS (Fixed-time segmentation)

FW (Fresh weight)

G

GPS (Global positioning system)

GUI (Graphical user interface)

H

HMM (Hidden Markov model)

HR (Hedgerow)

I

IGER (Institute of grassland and environmental research)

ION (Institute of navigation)

K

KFI (Knowledge flow interface)

L

LMT (Logistic model tree)

LR (Logistic regression)

M

ML (Machine learning)

MO (Movement object)

MP (Multilayer perceptron)

msm (Multi-state model)

N

NB (Naïve Bayes)

NBTree (Naïve Bayes tree)

O

ODBA (Overall dynamic body acceleration)

OF (Open field)

OneR (One rule)

P

PART (Partial decision tree)

PELT (Pruned exact linear time)

PSM (Previously surveyed mark)

R

RIPPER (Repeated incremental pruning to produce error reduction)

RMSE (Root mean squared error)

ROC (Receiver operating characteristic)

RTK-GPS (Real-time kinematic global positioning system)

S

SCAY (Static component of acceleration in Y-axis)

SCC (Somatic cell count)

SD (Standard deviation)

SEM (Standard error mean)

SL (Simple logistic)

SMO (Sequential minimal optimisation)

SMOTE (Synthetic minority over-sampling technique

SO (Segment object)

ST (Static component of acceleration)

SVM (Support vector machine)

T

TMR (Total mixed ration)

TN (True negative)

TP (True positive)

V

VeDBA (Vectorial dynamic body acceleration)

VS (Variable segmentation)

W

WEKA (Waikato Environment for knowledge analysis)

Introduction

1. Precision livestock farming

1.1 A developing livestock industry

Livestock production systems now incorporate a huge level of diverse expertise at all levels of the supply chain. Such has been the rapid development in data capture technology, the livestock industry is developing and augmenting systems that can make efficient use of data for enhanced productivity from fewer resources (Kamilaris *et al*., 2017). For example, pig producers will soon be able to automatically estimate the weight of their animals frequently during the growing phase using image analysis, reducing the need for handling and therefore reducing stress (Kashiha *et al*., 2014). Great progress has been made in many aspects of livestock production by bringing together a number of professional roles including animal practitioners, data analysts, software and hardware engineers, experts in animal behaviour, disease transmission and experts in climate modelling to name a few. Collectively, these roles underpin the concept of precision livestock farming (PLF) which encapsulates the ever-increasing use of data in the livestock industry to assist in managing the health, welfare and performance of animals.

There has been a rapid increase in the number of scientific publications in PLF. A broad literature search using the Web of Science (Thomson Reuters, 2019) database with keywords “precision livestock farming” found 234 publications produced since 2000; 170 of which (73%) were published from 2013 (

Figure 1). The uptake of PLF on farms is gradual however and tends to vary between systems (Wathes *et al.*, 2008; Werkheiser, 2018). Farmers are likely to be motivated by a number of factors when considering a PLF system such as the cost of any intervention (Russell and Bewley, 2013), the longevity of the technology (Van De Gucht *et al.*, 2018), ease of use (Sharma and Kaushik, 2019) and the financial benefits (Rutten *et al.*, 2013). The effect of farmer age on the uptake of PLF technologies on cattle farms tends to vary depending on the technology (Corner-Thomas *et al.*, 2015; Lima *et al.*, 2018; Abeni *et al.*, 2019). However, farms with higher cow-to-staff ratios seem to adopt more technologies than farms with fewer cows per staff member (Abeni *et al.*, 2019).

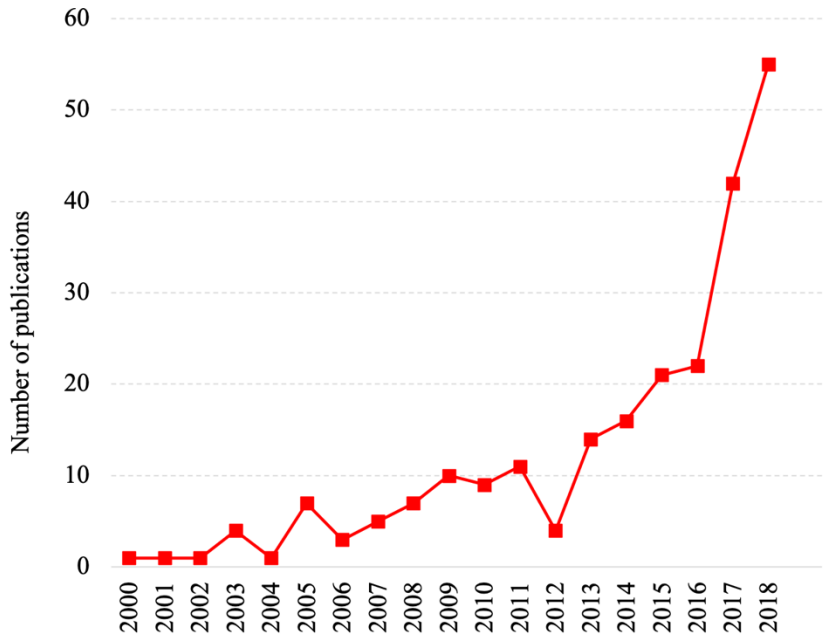


Figure 1. Trend of published research in PLF between 2000-2018 using keywords "precision livestock farming." Chart shows a total search-find of 234 publications produced since 2000 as indexed by Web of Science (Thomson Reuters, 2019).

Contact time between operators and their animals is likely to be highly influenced by system type and species farmed. For example, a dairy operator will directly realise the economic consequences of a case of clinical mastitis where milk

must be discarded to avoid processor penalties. The subsequent use of intramammary and possibly systemic antibiotics, veterinary assistance and labour further increases the cost of production and the losses quickly become tangible. A dairy operator may be more inclined therefore to invest in a tool that can automatically monitor milk-variables associated with mastitis such as milk electrical conductivity for earlier mastitis detection (Borchers and Bewley, 2015). Conversely, the consequences of an outbreak of contagious ovine digital dermatitis on the productivity of ewes may be less obvious. Suboptimal fertility may only manifest itself months later during pregnancy scanning where overall foetal numbers are lower than the desired target (Clements, 2015). Investing in a system to monitor the grazing duration and frequencies of sheep for earlier disease recognition may be less attractive without prior knowledge of the consequences.

1.2 Utility of precision livestock farming systems

A number of PLF systems for monitoring the performance and health of livestock are commercially available, many of which have been developed for use in the dairy industry to support daily management. The majority of systems are on-animal sensors which focus on retrieving information on variables that are both meaningful to farmers and well supported in the literature to be associated with animal health and performance. For example, rumination is a necessary biological component of cattle behaviour and farmers can identify potentially sick animals by their rumination patterns (Calamari *et al*., 2014). Collars capable of detecting rumination events are useful for benchmarking individual animals against their own rumination profiles and also for herd benchmarking (Marchesini *et al*., 2018). Commercially available rumination sensors include the SCR VocalTag (Soriani *et al*., 2013), Moomonitor+ sensor

(Dairymaster Bromsgrove, Worcestershire, United Kingdom) and the Nedap Smarttag sensor (Van Erp-Van der Kooij *et al*., 2016).

It is paramount that the data retrieved from PLF systems are easily interpretable and generate useful action points for consideration by farmers. Unfortunately, the utility of such data is somewhat unrealised in production systems and where necessary, more training is needed for farmers, their employees and their advisors in fully utilising this information (Morgan-Davies *et al*., 2018; Van Hertem *et al*., 2017). Data use and utility can often be complicated by the fact that many PLF systems work independently of one another. For example, milk yield and activity data are most often provided by two separate systems and so the ability of farmers to identify relationships between datasets is difficult. This is further complicated by the fact that current research is only just beginning to identify the complex associations between animal behaviour and performance. These relationships will be important for creating robust and trustworthy models for PLF systems that can alert farmers when necessary. Moving forward, care will be needed in the management of data to ensure that processed data in itself can be accessed at some level by users if required. An age has been entered where dependency on algorithms for decision making is greater than ever before and is likely to continue to aid decision making. In these so-called 'black box' support systems where data are collected and classified output is provided, care will need to be taken to ensure that any hidden biases are accounted for in the decision process and that interpretable logic is provided to aid the decision if at all possible (Pedreschi *et al*., 2018).

1.3 Sensors and information processing

A PLF system can be structured such that farm operators receive frequent feedback from animal or infrastructure-based sensors. This information can be used to make long-term or more frequent management decisions which are implemented

manually or autonomously. Measuring the activity of dairy cattle using pedometers for the purpose of oestrus detection is an example of the partial automation of a subset of a dairy cow system. This has led to greater conception rates and improved farm efficiency (Ferguson and Skidmore, 2013).

Other sectors are benefitting too. Some producers in the poultry sector are making use of precision feeding systems where feed allocation can be tailored to meet the growth targets of individual birds (Zuidhof *et al.*, 2017). In future, there may be possibilities to add further dimensionality to these systems. Data acquired from other sensors or the internet can add context and enrich the management process with real-time information on the behaviour of conspecifics, microclimate, parasite burden, disease transmission as well as market information, all of which will allow operators to respond instantly (Neethirajan, 2017). In fact, a range of tools and techniques are now under development for most species and for a variety of purposes. These include tools for measuring automatically the total feed and water intake of animals (Maselyne *et al.*, 2015), condition score (Spoliansky *et al.*, 2016), conspecific interactions (Handcock *et al.*, 2009), disease (Carpentier *et al.*, 2018), parturition (Menzies *et al.*, 2018), posture (Thompson *et al.*, 2016), temperature (Voss *et al.*, 2016) and vocal sounds (Meen *et al.*, 2015). Sections 2.1-2.3 will discuss some of the key areas where PLF techniques could be used to improve the welfare and performance of livestock.

To capture behavioural information, a bespoke set of sensors and technologies are required for specific tasks, the selection of which will depend on a number of variables. For example, the species in question, the desired metric for measurement and quantification (e.g. animal posture and frequency of postures), the housing system (indoors, free ranging), herd or flock size, farm layout, internet connectivity, staff training opportunities and desired outcomes. Sensors must be environmentally robust,

dependable and integrated with powerful and easy to use support systems that provide reliable and user-friendly information to aid decision making (Borchers and Bewley, 2015; Lima *et al*., 2018). PLF systems should ideally be flexible enough to be multifunctional (e.g. a 3-axis accelerometer fitted to the necks of cattle can be programmed to record rumination, feeding and posture) but some systems will be specific to a particular purpose. For example, an image-based automatic posture classification system will be specifically designed to discriminate between lame and non-lame cattle. While Chapters 1-3 focus on the use of global positioning systems (GPS) for cattle behaviour classification, for comparison, the use and functionality of other PLF tools will be discussed in Section 3.

1.4 Drivers for change towards precision livestock farming

The utility of any PLF system is fundamental to the level of adoption amongst farmers and practitioners. Other influential factors may include government incentives or pressures for the accurate collection of herd or flock data (Lima *et al*., 2018). There may be desires to increase herd or flock size for reasons of profitability or efficiency, or managers may foresee a reduction in available labour and skilled operators. Data collected on sheep farms in England and Wales found that knowledge of IT, the use of smartphones, the time farmers spent managing their sheep and the need to intensify production were all significantly associated with the likelihood of uptake of an electronic identification system (Lima *et al*., 2018).

Farmers may also want to improve performance in a given area, to partially or fully automate a division of the system for greater efficiency or to provide more time to focus on another aspect of the system. Van De Gucht *et al*. (2017) undertook a survey of 135 dairy farmers in Belgium. They found that the decision to invest in an automatic lameness detection system significantly depended on the importance a farmer attached

to lameness (more importance = greater perceived utility of system), the interval between calving and first insemination on their farm (increased interval = increased perceived utility of system) and whether farmers had already adopted an oestrus detection system (already adopted = increased perceived utility of system). Interestingly, once the cost of lameness had been communicated to the farmers in this study, the perceived utility of a lameness detection device significantly increased but this was only true for those farmers that were already using an oestrus detection device (Van De Gucht *et al*., 2017). Clearly, more needs to be done to inform farmers of the economic impact of some production disorders, in particular those problems where there is seemingly no direct financial consequence.

As the pressures increase on producers to improve efficiency, reduce the use of antimicrobials and provide assurances for animal welfare, there are clear opportunities for the development and adoption of PLF technologies by farmers for management support. A clear advantage of using PLF technologies is the labour-saving potential. Indeed, labour reduction is probably one of the most important reasons that farmers may want to invest in a sensor system (Lima *et al*., 2018; Morgan-Davies *et al*., 2018). One of the most common uses of PLF in the dairy industry is in oestrus detection. Without automated oestrus detection, farmers are recommended to observe cows for at least 20 minutes, three times per day for heat events (Firk *et al*., 2002). An automatic system would likely reduce time and labour costs. Furthermore, PLF technologies are objective and when supported by robust prediction algorithms may be able to identify animals that are in need of attention sooner compared to human observers (King *et al*., 2017). In some cases however, more work is needed in system development as diagnostic power has been shown to be less effective than professional opinion (Bicalho *et al*., 2007).

Whatever the driver for adopting or integrating PLF techniques, it is at least in the near-future unlikely that human operators will be fully removed from the various processes that govern livestock production systems. It is equally unlikely that PLF will replace poor management (Wathes *et al.*, 2008). Instead, the future of livestock production systems looks highly likely to be supported by decision support tools that will enhance the quality of livestock products and provide opportunities for systems to be more productive and sustainable (Scholten *et al.*, 2013). The remainder of the discussion will focus on future demands on the dairy industry and how precision technologies can help support dairy farmers.

2. Future demands on dairy production systems

There has been significant growth and intensification of global dairy production systems in the past 50 years, and the demand for dairy products is likely to continue to increase in the foreseeable future (Fuller *et al.*, 2006; Britt *et al.*, 2018). Globally, dairy consumption is expected to rise on average by 27% from 87kg to 119kg per person by 2067 driven largely by population growth and urbanisation (Britt *et al.*, 2018).

For the dairy industry, this growth in production will be largely met through improvements in feed intake and the efficiency of nutrient utilisation driven by genetic selection (Britt *et al.*, 2018; Cole and VanRaden, 2018). For many systems, this has led to intensification, driven by economic pressures and consumer expectations and for the major milk-producing regions, intensively managed systems dominate (Powell *et al.*, 2013). Additional pressures will ensue with these systems which include ensuring sustainability, ethical acceptance and that systems are environmentally benign (Augustin *et al.*, 2013). For example, intensively managed dairy systems usually mean that cows are confined to housing for the majority of their milk-producing lives, usually resulting in greater production potential but also greater consumer awareness for the

welfare of cows and safety of milk products (Drake, 2007). On the other hand, pasture-based systems are usually recognised as being more natural by consumers and welfare considerations may be overlooked (Arnott *et al*., 2017). For pasture-based systems, value can also be added to milk products by highlighting some of the milk quality-associated attributes such as fatty-acid profile (Dewhurst *et al*., 2006). However, these systems face their own challenges such as managing cows in adverse weather conditions, and the potential for greater difficulty in monitoring physiological indicators of health due to reduced contact time with staff (Arnott *et al*., 21017).

With these challenges come opportunities and the dairy industry is rapidly evolving and making use of the technologies available for supporting daily management (Rutten *et al*., 2013). From a production perspective, efforts are now focusing on developing sensors that can support the daily management of individual animals. Most studies have previously focused on fertility management (Mottram, 2016) and locomotion problems (Van Nuffel *et al*., 2015) but increasingly, more attention is being paid to sensors that can measure individual feed intake (Leiber *et al*., 2016) and detect disease (Steensels *et al*., 2017).

2.1 Dairy cattle performance and welfare

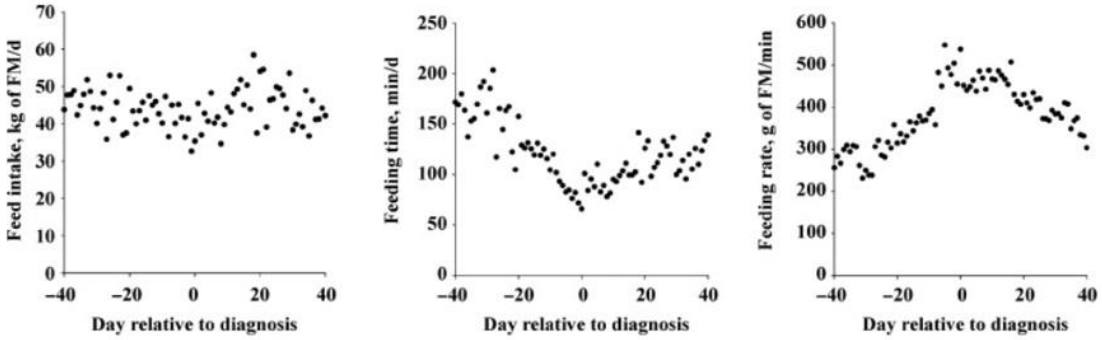
Largely driven by economic pressures, the number of UK dairy herds is decreasing rapidly and as of January 2019 a total of 9,170 producers were recorded in England and Wales; 33 fewer than the previous month and 177 fewer than January 2018 (AHDB Dairy, 2019[a]). With this, the number of cows per herd is increasing. The current average UK herd size (data for 2018) is estimated to be 148 cows which is an increase of two cows per herd on the previous year (2017) and 32 more than in 2008 (AHDB Dairy, 2019[b]). Average UK milk yield per cow for 2018 was estimated at 7,825 litres (AHDB Dairy, 2019[c]). In 2008, this figure was 6,974 litres. Despite increased

production often being associated with the decreased health of dairy cows (Barkema *et al*., 2015), few studies are available on the impact of milk production as a risk factor for disease. One study suggested that higher production was associated with higher incidences of mastitis and lameness (Koeck *et al*., 2014) but others found no relationship between increased milk yield and diseases such as dystocia and metritis (Ingvartsen *et al*., 2003).

To the candidate's knowledge, the current number of staff available per animal on UK dairy farms is unknown but most evidence suggests that the best economically performing farms use less labour per cow, with the top 25% of UK farms allocating 35 h/cow per year compared to 49 h/cow per year in the bottom 25% of herds (AHDB Dairy 2013; AHDB Dairy, 2015). Some evidence suggests that as the number of cows per labour unit increases, so too does the prompt identification of diseases such as lameness (Leach *et al*., 2010). Herd expansion may also be associated with higher somatic cell count (SCC) (Archer *et al*., 2013). However, the relationship between herd size and cow welfare is complex and some have found evidence of decreasing lameness risk in larger herds (Chapinal *et al*., 2014) but an increase in within-herd prevalence of other infectious diseases such as Johne's disease and bovine tuberculosis (Doyle *et al*., 2014; Wolf *et al*., 2014).

PLF technologies offer opportunities to objectively measure daily variations in the behaviour of cows and alert farmers to the need for interventions when required. Feeding behaviour for example can provide a multidimensional outlook on the health and performance of cows. In a retrospective analysis of individual feeding data from computerised feeders, González *et al*. (2008) found that dairy cows subsequently diagnosed with ketosis exhibited significant reductions in daily feed intake (-10.4 kg fresh matter (FM)/d), daily feeding time (-45.5 min/d) and daily feeding rate (-25.2 g

FM/min per d), 3.6 days on average, before diagnosis. Cows diagnosed with locomotor problems had a decreased daily feed intake (-1.57 kg of FM/d), daily feeding time (-19.1 min/d) and a daily increase in feeding rate (+21.6 g FM/min per d) (Figure 2). A number of publications have reported similar patterns for production disorders (Huzzey *et al*., 2007; Sepúlveda-Varas *et al*., 2016; Schirmann *et al*., 2016), and this is now a very strong area of research in the dairy sciences.



(González *et al*., 2008)

Figure 2. Changes in the feed intake, feeding time and feeding rate before and after a cow was diagnosed (day 0) with foul of the foot. Adapted from González *et al*. (2008).

2.2 Sensor systems to support dairy operations

As dairy systems evolve, the opportunities for managerial support using sensor-based technologies are many and farmers are already realising the benefits of certain tools. Indeed, it seems that as the number of cows per farm increases, so too does the level of adoption of precision technologies (Steeneveld and Hogeveen, 2015; Gargiulo *et al*., 2018). A Dutch survey (Steeneveld and Hogeveen, 2015) showed that 39% of farms surveyed (n = 512) used sensor systems and that sensors for mastitis and oestrus detection were the most common. Furthermore, farms with sensor systems had fewer labour hours per cow but productivity per cow measured as milk production per cow, number of cows per hectare and milk production per hectare did not differ between

farms with and without sensor systems. In an Australian study (n = 199 farms), farms with more than 500 cows (16.4%) adopted between two and five times more precision technologies compared to smaller farms (Gargiulo *et al*., 2018) such as electronic cow identification systems, automatic sorting gates and herd management software. Indeed, the adoption of precision technologies on farms with larger herds may be because herd expansion requires investment in improved and more efficient milking systems and often, these systems come integrated with automated cow management tools (Rutten *et al*., 2013). For example, Steenveld and Hogeveen (2015) found that farmers with automatic milking systems (AMS) made no conscious decision to invest in the accompanying sensor systems and that the sensor systems either came as standard or were bought at reduced cost.

It is fundamentally important that farmers are made aware of the potential production responses resulting from the adoption of sensor systems before investments are made. The extent to which any benefits may be realised may vary greatly between milking systems. For example, Steeneveld *et al*. (2015[a]) found that herd average SCC increased (+12,000 cells/mL) on AMS farms in the years after investing in a sensor system for mastitis detection whereas average SCC decreased (-10,000 cells/mL) in the years after investment on farms with conventional milking systems (CMS). The authors concluded that such effects may be explained by farms undergoing other major changes to their systems such as investment in a new milking system. Speculatively, it could be that farmers with CMS make more conscious decisions to invest in sensor systems and are better able to evaluate the impact of a sensor system due to more time being spent with their animals. Care needs to be taken to ensure that informed choices are made before investing in sensor systems such that any improvements in efficiency are not lost in other aspects of the production system (Steeneveld *et al*., 2015[b]). Reducing labour

was one of the most important reasons for investing in sensor systems in that study. The overall economic benefit of investment may come as a result of a reduction in labour cost and not necessarily from improvements in health and performance.

To date, the majority of sensor systems used on farms can be characterised as systems that are either purchased as part of a larger investment (e.g. milk conductivity meters within an AMS) or those that are purchased as part of a specific strategy to improve performance in a particular area (e.g. pedometers for oestrus detection). From a research perspective, a wide range of sensor systems are under development for use in both whole herd and individual cow management. In a review, Rutten *et al*. (2013) categorised 126 publications (2002-2012) describing research undertaken on sensor systems into four levels. More than half of all publications using sensors for capturing information on locomotion (n = 38) and metabolic issues (n = 16) in cattle were undertaken at level one, (something is measured about the cow e.g. rumen pH). The majority of studies concerned with mastitis (n = 31) and fertility (n = 41) were undertaken at level two (captured data is used to make further interpretations on the status of a cow e.g. decreased activity). No publications were found at levels three (sensor information is supplemented with advice from external resources e.g. economic) and four (farmer or sensor system makes a decision based on the information e.g. call a veterinarian). It is envisaged that in the next decade, sensors with the capability of monitoring a breadth of cow activity variables will be in use, allowing farmers' access to information that was previously unattainable. More is needed to ensure that conceptual research is taken to further levels in this hierarchy.

2.3 Conceptual sensor system research in the dairy industry

Reducing labour costs is not the only objective of sensor system development. Sensor systems have the potential to provide access to previously unattainable

information that can not only be leveraged for labour reduction but also allow farmers to make informed choices on the management of individual animals.

A body of research is now developing in the use of various sensors for quantifying dairy cow behaviour for detecting diseases such as ketosis (Rodrigues-Jimenez *et al*., 2018) and lameness (Barker *et al*., 2018) and also variables associated with the onset of calving such as temperature and rumination time (Rutten *et al*., 2017). Despite being in its infancy, the use of behaviour as a proxy for production and welfare management has great potential and is likely to feature highly in future sensor system applications. Currently, more is needed to quantify which behaviours carry the most importance when used as proxies for disease onset and how these behaviours might change over time. Furthermore, the impact of other variables on behaviour need to be fully assessed. For example, in a pasture-based system, the daily variation in pasture quality could impact the amount of time spent grazing by cows, potentially confounding any indicators of the onset of a particular disease.

Selection of the most appropriate sensor for research is also vital. Sensor choice will depend heavily on the research objective (e.g. support fertility, reduce use of antibiotics), management system (e.g. housed cows vs. grazed), the behaviours to be identified (affected by research objective), data sampling interval, sensor data capacity and battery life and the location of the sensor on the animal. From a research perspective, the sensors used to monitor individual behaviours in most cases are designed and constructed for the task at hand and are not typically commercially available products. In cattle research, these sensor systems have, over the last 20 years evolved, both because of improvements in sensor technology and a realisation that the data gathered from sensors that are seemingly unidimensional (e.g. GPS for location) can be leveraged for a breadth of statistical analyses. However, few publications

currently discuss the performance of commercially available products. Those that do are usually based on research paradigms that are well developed and that have had a financial impact at farm level e.g. oestrus detection (Saint-Dizier and Chastant-Maillard, 2012).

One primary research objective over the last few years has been to identify individual behaviours from sensor data using a variety of classification techniques and many have been successful at this (Williams *et al*., 2016; Rahman *et al*., 2018). In these instances, the analyst benefits from directly interpretable behavioural information. However, more recently, attention is being paid to utilising raw sensor data directly. One example includes analysing the space-use patterns of housed dairy cows for behavioural analysis (Diosdado *et al*., 2018). Such techniques do not require complex pre-processing techniques and may be an effective means of gaining more information about livestock movement while maintaining the capability of monitoring long-term health.

The realisation of the possibility of precision individual-animal management for a range of commercial and welfare-important parameters has to some extent erupted in the last 10 years. Before proceeding to a discussion on a series of animal-based sensors currently used in dairy research, it is pertinent to summarise the main evaluation metrics of models derived from sensor data used primarily in classification studies.

2.3.1 Model performance evaluation

2.3.1.1 Introduction

The method of evaluation of classification models derived from sensor data will be largely dictated by the study methodology. For example, if the objective is to identify the grazing behaviour of cattle then an analyst may consider a relatively simple threshold-based classification approach (Section 3.1.3) for data collected from head-mounted accelerometers (Rayas-Amor *et al*., 2017; Arcidiacono *et al*., 2017). Regression analyses can then be undertaken to model the relationship between visually observed and predicted grazing events. A more computationally-intensive modelling procedure based on pattern-recognition and machine learning (ML) is also possible, and studies using these procedures usually follow common model evaluation methods (Caraviello *et al*., 2006; Martiskainen *et al*., 2009; Mansbridge *et al*., 2018). However, despite a breadth of such work appearing in the PLF literature over the last 10 years, there is no standard model evaluation procedure, but each has its own merits and drawbacks. Some evaluation techniques are also more suited to binary tasks (e.g. discriminating between two behaviours) compared to multiclass classification tasks (Sokolova and Lapalme, 2009). Model performance evaluation is a topic strongly discussed in the field of computer science and many evaluation measures are derived from medical and behavioural sciences (Klein *et al*., 1994; Cohen, 2013). As such, this section is not an attempt at defining the best evaluation methods but rather, a short discussion of those most commonly used.

2.3.1.2 Classification accuracy

Classification accuracy (CA) is probably one of the most utilised evaluation metrics of model performance. It can be described as the total number of correct predictions of all predictions made using a classification model (Sokolova and Lapalme, 2009). CA requires knowledge of the number of true-positives (TP) predicted by the classifier, the number of true-negatives (TN), false-positives (FP) and false-negatives (FN) and is expressed as ((TP + TN) / (TP + FN + FP + TN) (Sokolova and Lapalme, 2009). CA has been used previously in several studies of behaviour classification of livestock (Martiskainen *et al*., 2009; Dutta *et al*., 2015; Williams *et al*., 2016; Giovanetti *et al*., 2017; Walton *et al*., 2018) and can be used in both binary and multiclass situations (Sokolova and Lapalme, 2009). However, class imbalances make the use of CA as an evaluation metric problematic because often, the class represented by the most instances can be selected for which can lead to inflation of CA and therefore poor class discrimination (Stąpor, 2018). Because of this, it is sometimes useful to include further evaluation metrics and some of the most common are discussed next.

2.3.1.3 Sensitivity, Specificity, Precision and F-measure

Analysts will often include measures such as sensitivity (TP / (TP + FN)) which is the proportion of target class labels that are correctly classified as the target class and specificity (TN / (FP + TN)), which is the number of non-target classes that are identified as such (Kourou *et al*., 2015; Weng *et al*., 2017). In addition, the precision (TP / (TP + FP)) of a classifier is often used and represents the number of correctly classified instances amongst all classified instances (Asri *et al*., 2016; Higaki *et al*., 2019). Collectively, these measures can provide a better idea of the power of a classifier particularly when no confusion matrix is published to make these estimations. These three measures either in conjunction with CA or as stand-alone metrics are often reported in articles considering livestock behaviour classification.

Some publications, however, do report further evaluation metrics and one of the most common is the F-measure (sometimes referred to as F1) which is an extension to precision and sensitivity expressed as (2TP / (2TP + FP + FN)). Mathematically, F-measure is the harmonic mean of precision and sensitivity and the closer it is to 1, the better the performance of the classifier, whereas a score of 0 indicates very poor performance (Sokolova *et al*., 2006). Ultimately, the rationale is to express a single value that represents the trade-off between the precision and sensitivity of the classifier. For example, a classifier with high precision but low sensitivity will yield highly accurate results but will have misclassified a large proportion of relevant instances (Sokolova *et al*., 2006). Thus, the F-measure is useful in support of either sensitivity or precision alone, but some have also reported all three measures (Alvarez *et al*., 2018; Rodríguez Alvarez *et al*., 2019). However, the F-measure and its component metrics have been criticised for their failure to account for true-negative instances (Powers,

2011) which could be particularly costly in the field of disease diagnostics where training sets are likely to be imbalanced (Qin *et al*., 2010).

2.3.1.4 Receiver operating characteristics

Receiver operating characteristics (ROC) analysis is sometimes documented to demonstrate classifier performance (Ortiz-Pelaez and Pfeiffer, 2008; Williams *et al*., 2016) particularly in the medical sciences (Ball *et al*., 2016). The measure uses both sensitivity and specificity and was traditionally used for binary classification tasks (Fawcett, 2006). A strong argument for its use in measuring classifier performance is because of its lower level of bias, particularly because it accounts for negatively classified instances and it provides a benchmark for chance-level classification (Powers, 2011).

Graphically, ROC curves are used to demonstrate the power of a classifier where the TP rate (sensitivity) is plotted on the Y-axis and the FP rate (specificity) is plotted on the X-axis. Generally, better performance is indicated by curves that occupy the north-west area of the plot (Fawcett, 2006). A degree of caution is needed in this interpretation however because classifiers that produce data that proceeds directly along the Y-axis may be an indication of a model that classifies TPs well in the occurrence of very strong evidence only (Fawcett, 2006) and could therefore be at risk when the evidence for TP instances is weaker. Chance-level classification is indicated by the diagonal line $y = x$ in ROC space and represents classifiers that are randomly assigning classes to instances in the dataset.

The two-dimensional ROC curve can be reduced to a single value that represents the performance of a classifier and this is termed the area under the ROC curve (AUC). The AUC value will always lie between 0 and 1 and a value >0.5 indicates a classifier that is performing better than chance (Bradley, 1997). Both ROC

and AUC analysis can be extended to multiclass classification tasks. One example of such an analysis is the reference class formulation (Fawcett, 2006) also known as the 'one versus all' class strategy (Provost and Domingos, 2000). The issue with such a strategy is that it violates the main attraction of ROC over other classifier evaluation strategies in that it is robust to class imbalance. A method insensitive to class imbalance was put forward by Hand and Till (2001) and this strategy is implemented in the R (R Core Team, 2014) statistical software package 'pROC' (Robin *et al.*, 2011).

2.3.1.5 Other model evaluation measures (Youden's index and Discriminant power)

Despite there being several evaluation measures to select from when assessing model performance, it is argued that some of these measures do not fully meet the demands of classification tasks where each class shares equal importance and where several algorithms are to be compared (Powers, 2011). Youden's index (sensitivity + (1 − specificity)) has been proposed which evaluates the ability of a classifier to avoid failure (Youden, 1950; Khalaf *et al.*, 2017). Youden's index can be used in binary classification tasks and cases where there are more than two classes to discriminate between (Nakas *et al.*, 2010). Discriminant power is another measure that utilises sensitivity and specificity and measures how well a classifier discriminates between positive and negative instances (Sokolova *et al.*, 2006). Discriminant power is written as: $(\frac{\sqrt{3}}{\pi}(\log X + \log Y))$ where X = sensitivity / (1 − sensitivity) and Y = specificity / (1 − specificity). Discriminant power been used previously for feature selection in the medical sciences (Mert *et al.*, 2015). Few publications on livestock behaviour classification use these performance measures (Amrine *et al.*, 2014; Zehner *et al.*, 2019).

2.3.1.6 Summary

This is not an exhaustive list of metrics available to the analyst for classifier performance evaluation, in fact, many more exist. Other mathematical approaches also exist in the literature for livestock behaviour analysis from sensor data. Categorising behaviour based on thresholds in sensor data requires less computational power than techniques where complex pre-processing is needed first. The success rate of such methods can be measured (Arcidiacono *et al.*, 2017) using similar methods to those described above. Some approaches have used raw movement data directly and these movements modelled with the objective of inferring differences in cattle health status (Diosdado *et al.*, 2018). Models are then assessed using specific selection criteria such as the Aikaike Information Criterion (Sakamoto *et al.*, 1986).

In the context of ML, the decision on which set of measures to use may hinge on several elements including whether it is a dichotomous or multiclass classification task and the relative importance of each class to the research question. Furthermore, analysts may be confined to the in-built evaluation metrics provided by the software in use. For example, the Explorer interface of the WEKA (Waikato Environment for Knowledge Analysis) data mining software (Witten *et al.*, 2016) is limited to the main evaluation parameters discussed previously (Sections 2.3.1.3 and 2.3.1.4), but there are options to extend these measures using the Experimenter interface (Section 5.5). Programming languages such as R can provide more flexibility in this regard and allow analysts to tailor their own evaluation measures. Further discussion on classifier evaluation metrics can be found in Sokolova *et al.* (2006) and Tharwat (2018). For a brief discussion on the statistical comparison of model performance measures, see Section 5.5.2.

3. Sensors in precision dairy research

3.1 Accelerometers for behaviour classification

3.1.1 Background

Accelerometers measure g-force and tilt in three axes (X, Y and Z) at pre-programmed intervals. These sensors have been used extensively to monitor human behaviour (Mathie *et al*., 2004; Lowe and ÓLaighin, 2014) and over the last decade, have been shown to be very valuable in behaviour classification studies of both free-ranging (Sakamoto *et al*., 2009) and domestic species (Watanabe *et al*., 2008). For dairy animals, their utility for classifying behaviours associated with performance and welfare is very promising. For example, the objective of some of the earliest uses of accelerometers fitted to ruminants was to identify attributes of feeding behaviour (Chambers *et al*., 1981).

The research of the last 10 years has focused heavily on the lying and standing behaviours of cattle (Ito *et al*., 2010) as well as for general behaviour classification (Borchers *et al*., 2016). By now, it is well recognised that these identifiable behaviours are closely associated with diseases such as lameness (Thorup *et al*., 2015) and mastitis (Stewart *et al*., 2017) as well as overall cow comfort (Ito *et al*., 2010). There is now a strong focus on using accelerometers to identify multiple cow behaviours that are both performance and welfare-associated, some of which are discussed next. Despite not being used in any of the experimental chapters herein, a discussion on the utility of accelerometers is useful as their use in livestock research has increased dramatically since 2009 and indeed, since the beginning of this candidature.

3.1.2 Lying and Standing

Automatically identifying the lying and standing behaviours of dairy animals can be undertaken using a single, leg-mounted accelerometer usually positioned so that the X-axis is parallel to the ground (Ito *et al*., 2010). For behaviour analysis, the Y-axis values are used to confirm lying or standing events and the Z-axis confirms lying laterality (left or right). Identifying the number of lying and standing bouts and bout durations undertaken by cows has been used extensively to assess associations between lameness and leg injuries (Charlton *et al*., 2016; Westin *et al*., 2016) and this information can be useful for herd benchmarking and also to provide farmers with advice on system improvement. For example, Westin *et al*. (2016) found that lame cows lay down for 0.6 h/d longer than non-lame cows and that leg lesions were associated with shorter lying times. Lying and standing frequencies have also been used as indicators of calving in dairy cows (Jensen, 2012; Saint-Dizier *et al*., 2015). The optimal logging interval for these behaviours had previously been determined to be ≤30s (Ledgerwood *et al*., 2010) provided that very short, possibly erroneous lying bouts were removed from the dataset. However, the majority of publications have used one-minute sampling intervals with performance equating to research that has used more frequent sampling intervals (Ito *et al*., 2009). Most research utilise *g*-force values generated by accelerometers with the addition of constants to facilitate data handling (Ledgerwood *et al*., 2010).

3.1.3 Classifying other biologically important behaviours using accelerometers

In addition to lying and standing, some authors have been able to successfully identify the feeding and ruminating behaviours of cows using head-mounted (Rayas-Amor *et al*., 2017) ear-mounted (Pereira *et al*., 2018) and neck-mounted accelerometers (Diosdado *et al*., 2015; Benaissa *et al*., 2019). Higher sampling frequencies are usually

required for feeding and ruminating behaviours. This is because of the frequent nature of these behaviours within a given time-interval, and the resolution required to effectively discriminate between behaviours that produce accelerometery signatures with variable frequencies and amplitudes. For example, Nielsen (2013) used a halter-mounted accelerometer set at a sampling frequency of 5 s to collect data on grazing and non-grazing events from cattle. A linear discriminant procedure was then used to discriminate grazing from non-grazing and a sensitivity and specificity of 84% and 80% was achieved respectively. Similarly, Watanabe *et al*. (2008) sampled accelerometer data at 1 s intervals and used a discriminant procedure to classify eating, ruminating and resting using features calculated from the raw accelerometery data (mean, variance and inverse coefficient of variation). The percentage of correct classifications was variable across axes and features used but the best performing combinations achieved >90% correct classification.

Using neck (González *et al*., 2015; Arcidiacono *et al*., 2017) and head-mounted accelerometers (Rayas-Amor *et al*., 2017), threshold classifiers have been developed using acceleration values. This strategy is relatively simple to undertake compared to more computationally intensive procedures that require extensive data pre-processing, algorithm development and refinement. Using ground-truth observations of biologically relevant behaviours, manually annotated data are analysed for descriptive statistics such as values of central tendency (mainly mean and median) and also for measures of variation such as the standard deviation. Thresholds can then be assigned to the acceleration data that represent the boundaries between each behaviour and an analysis undertaken on the discriminatory power of these thresholds on new data. These methodologies have been shown to discriminate well between behaviours such as feeding and standing (Arcidiacono *et al*., 2017) and also grazing and ruminating

(Rayas-Amor *et al*., 2017). These techniques share commonalities with that already established for discriminating between lying and standing behaviours using leg-mounted accelerometers (Ito *et al*., 2009; Ledgerwood *et al*., 2010). More is needed to test whether the classification success reported holds true for data collected over longer periods where pasture height and density and animal variation (e.g. posture and height) may affect success. This is essential to ensure that accurate estimations of feeding and ruminating times can be made with the possibility of using this information as reference for other important metrics of production and welfare such as energy expenditure (Green *et al*., 2009). Furthermore, since sampling interval varies in these studies (range = 4 Hz – 1 sample/30 s), a review of the literature would be valuable to measure the success of studies across the range of sampling intervals and techniques implemented.

3.1.4 Accelerometer sampling frequency

The sampling frequency of choice for acceleration data to a large extent depends on the behaviour(s) to be identified, the analysis to be undertaken post-hoc, the capacity of the instrument or the system to store data and perhaps, the effective battery life of the device. The benefit of maximising the sampling frequency is that the dataset can be subsampled to test algorithmic performance on a range of sampling intervals. Recently, sampling frequencies of up to 50 Hz (50 samples/s) have been reported (Diosdado *et al*., 2015) and studies often report sampling frequencies in the range of 1-20 Hz (Martiskainen *et al*., 2009; Dutta *et al*., 2015; González *et al*., 2015; Benaissa *et al*., 2017; Rahman *et al*., 2018).

The effect of sampling interval on classification performance is not clear because there is no standard sampling interval or analytical procedure for such data. The feature set used (Section 5.2), the algorithms or statistical procedures tested, and behaviours identified are some of the variables that differ greatly between studies.

Generally, however, the higher the sampling frequency, the better the classification performance for behaviours that occur often and for long periods of time (e.g. feeding). Furthermore, there is a risk that infrequent sampling intervals could lead to the loss of behavioural information and a misrepresentation of the frequency and duration of behaviours undertaken by the focal animal (González *et al*., 2015). Work is needed to collate the performance of various analytical strategies (e.g. features, algorithms and statistical analyses used) on the range of sampling intervals reported in the literature to determine the best procedures for classifying dairy cow behaviours using accelerometers. Furthermore, the performance of accelerometers in relation to position on the animal is required as this may influence the quality of the data signatures gathered and also the practicality of application at farm level.

3.2 GPS for location and behaviour classification

3.2.1 Background

GPS has been used extensively in animal research for many years (Turner *et al*., 2000; Agouridis *et al*., 2004; Woodroffe *et al*., 2016). The primary use of GPS has been to evaluate spatial location and the movement of animals in studies of species distribution and interaction (Turner *et al*., 2000; Barasona *et al*., 2014; Handcock *et al*., 2009). Some notable areas of interest include studying the effects of climate and disease on movement (Allred *et al*., 2013; Miguel *et al*., 2017). The use of animal-borne satellite receivers began to grow in the 1990s (Harris *et al*., 1990; Rodgers, 1996) evolving from their use as stand-alone sensors to the present day where GPS is often coupled with other sensors such as accelerometers, magnetometers and gyroscopes for multidimensional analysis of animal movement (Guo *et al*., 2009; Dutta *et al*., 2015).

3.2.2 GPS configuration for studies of animal behaviour

As with other sensor systems, GPS can be configured to collect time-stamped data at a predetermined rate. The data sample rate is usually determined by the objective of the study, which historically, was usually centred on knowing where the animal was, what it was interacting with and for how long. Coupled with a geographic information system, GPS data can be enhanced to provide further dimensionality to allow analysts further access into the behaviours of focal animals. Spatial location, speed of movement and altitude are examples of commonly collected variables that can be set to register at fixed intervals ranging from >1 Hz (Swain *et al.*, 2008) to once hourly (Perotto-Baldivieso *et al.*, 2012). GPS can also be configured to collect data at triggered intervals (Jurdak *et al.*, 2013), for example, when movement exceeds a certain predetermined threshold. This latter method can lead to significant savings in battery power and therefore a longer duration of observation. Variable fixes in GPS data can also occur when focal animals move in and out of satellite range and has been an important topic of discussion in the movement ecology literature as gappy data can present an additional challenge in animal behaviour inference (Jonsen *et al.*, 2007; Gurarie *et al.*, 2009). For terrestrial animals at least, the use of triggered (predetermined) sampling intervals in particular necessitates that GPS receivers are tested for their level of positional fix error (although this is also necessary with a regular sample rate). This is essential to maximise the chances of correctly classifying an instance of movement from an instance of non-movement from focal animals.

3.2.3 Errors in GPS fixes

While modern, domestic GPS receivers are able to provide accurate absolute positional fixes, several issues exist that can lead to inconsistencies and sometimes large variation in the absolute positional fix of each receiver at each *n*th sample when

measured against a surveyed mark. These issues have been described in detail by several authors in preparation for animal movement studies (DeCesare *et al*., 2005; Ganskopp and Johnson, 2007; Hurford, 2009) and must be accounted for if meaningful information is to be gathered.

Some of the main issues that can cause erroneous and highly variable positional fixes include satellite and receiver errors, atmospheric effects and multipath errors caused by large objects such as buildings. Screening methods have been used previously to remove erroneous GPS fixes using various strategies (Ganskopp and Johnson, 2007; Bjørneraas *et al*., 2010). Failing to account for these issues can lead to erroneous estimations of distances travelled by focal animals and the accurate identification of the patch in which they occupy at each timepoint (Hulbert and French, 2001). The data sampling rate can also affect these estimations and the selection of which will depend on the species under observation and the resolution required for monitoring. For example, it was found that GPS fixes taken every hour meant that the location of cattle within a 10,000 m^2 (1 ha) area could only be predicted with 30% accuracy. On the other hand, a sample rate of 10 s in a 100 m^2 patch led to an estimated location prediction error of 1% (Swain *et al*., 2008). If distance travelled is a study objective, then a more frequent sampling rate will also provide more accurate estimations of travel distance. This is because straight-line trajectories between longer samples will lead to greater uncertainty about animal activity (Pepin *et al*., 2004).

GPS receivers are usually supported with documentation on their expected levels of performance, the most notable performance measures in the context of animal movement studies being absolute horizontal and vertical accuracy, although others exist (e.g. hot, warm and cold start) (Low *et al*., 2015). GPS receivers used for research purposes should be tested for homogeneity. The following section discusses the tests

that were undertaken to ensure homogeneity between the study receivers used in Chapters 1-3.

3.2.4 Measuring uncertainty in GPS fixes

3.2.4.1 Background to error tests undertaken on study GPS receivers

A number of error metrics exist that can be used to determine the accuracy of a GPS receiver and the standards of the test procedures are detailed by the Institute of Navigation (ION, 1997). The six main error metrics are detailed in Table 1 (Section 3.2.4.3) and have been used previously to measure positional fix accuracy and receiver errors prior to animal research (Agouridis *et al.*, 2004). As the work undertaken in Chapters 1-3 used GPS receivers to model dairy cow behaviour, it was important to ensure that all receivers were tested as rigorously as possible to ensure homogeneity. Although it is possible to measure both the horizontal and vertical accuracies of GPS receivers, only horizontal accuracy was considered here. This is because it was hypothesised that the vertical accuracy of the receivers would be outside of that required for discriminating between the lying and standing behaviours of cows, and that grazing altitude would also be a factor to consider in this discrimination process. Nevertheless, there may be scope for exploring this variable in future.

The tests undertaken with the GlobalSat DG-100 GPS receivers in this work (GlobalSat® Technology Corp., Taiwan) were designed to test static accuracy in both open field (OF) and conditions that represented situations where signal obscurity due to common land features could be an issue. In this instance, the grazing platform at the research farm was largely unobstructed but hedgerows were present across the entire site. Therefore, a hedgerow (HR) static test was also undertaken. As well as providing information on the horizontal accuracy of each GPS receiver, the ultimate objective of the static tests was to determine if any GPS receivers were providing positional fixes

that were significantly different to the distribution of fixes provided by the whole sample. The null hypothesis was that all GPS receivers would provide a distribution of positional fixes that were not significantly different from one another. If any GPS receivers were found to violate this assumption, those receivers would be discarded from any animal experimentation.

3.2.4.2 GPS static accuracy tests

Thirty-six GPS receivers were available for static accuracy tests and data were collected between March 3 and March 20, 2014. Tests were undertaken to meet the requirements of the Institute of Navigation (ION, 1997) and the experimental methodology was adapted from Agouridis *et al*. (2004). OF tests were undertaken at Aberystwyth University dairy farm, Trawsgoed, Ceredigion in the centre of a 6-ha paddock that had no physical obstructions that could affect satellite signal such as farm buildings. The HR tests were undertaken along a 40 m portion of hedgerow running approximately east to west on the edge of the same paddock (Figure 3). The hedgerow (approximate dimensions: H = 2.2 m; W = 2 m) was predominantly hawthorn. For the OF tests, 35 wooden posts (length = 1,700 mm; width = 100 mm) were arranged in a 6 x 6 square grid pattern, 1 m apart (Figure 3: OF) and driven into the ground such that all were completely level with the post at the highest point of the arrangement. For the HR tests, each post was located 1 m apart along the hedgerow and levelled using the same method as that used for the OF arrangement (Figure 3: HR).

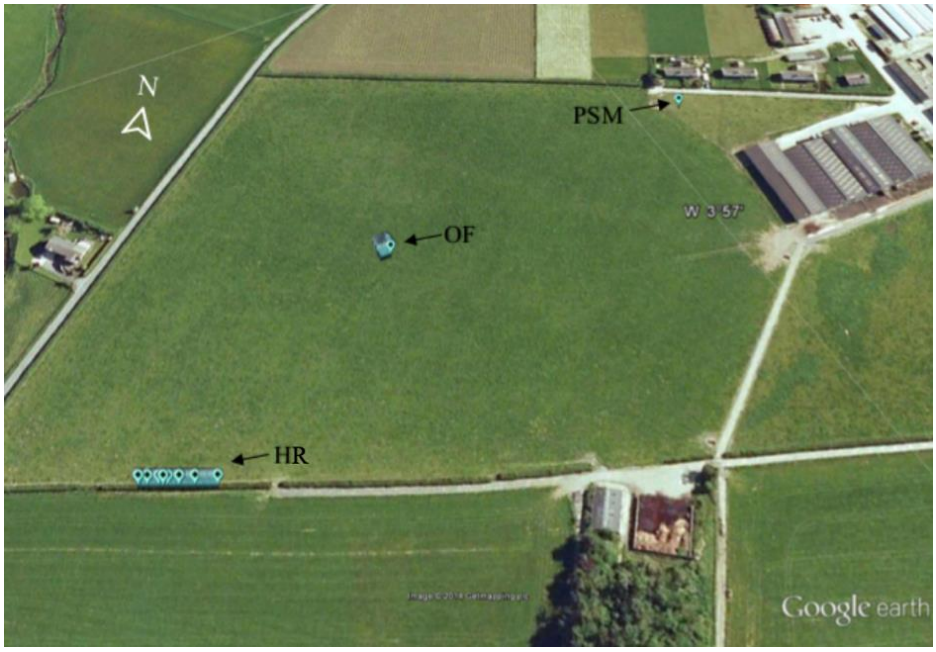


Figure 3. Image of 6-ha paddock where open field (OF) and hedgerow (HR) static GPS tests were undertaken. PSM = previously surveyed mark. Real-time kinematic-GPS base station was located at PSM and rover used to survey precise locations of all GPS-mounting posts (indicated by blue markers) in OF (36 posts) and HR (36 posts) tests.

A Trimble real-time kinematic GPS (RTK-GPS) system was used to accurately measure (reported accuracy ± 2cm) the location of each post in the study. The base station of the RTK-GPS system was located at a previously surveyed mark (Figure 3: PSM) and the rover was used to measure the location of each post at 1 s intervals for 1 min. The dataset was then averaged to gain the precise location of each post. This was done so that the location of the testing post could be subtracted from the coordinates provided by the experimental GPS receivers to calculate horizontal accuracy. This was done using equations 1 and 2:

$$\Delta long = {}^{\lambda}\text{GPS receiver} - {}^{\lambda}\text{post} \qquad (1)$$

$$\Delta lat = {}^{\varphi}\text{GPS receiver} - {}^{\varphi}\text{post} \qquad (2)$$

where $\Delta long$ is the difference in longitude between the surveyed post and experimental GPS receiver and Δlat is the difference in latitude between the surveyed post and experimental GPS receiver. In equations 1 and 2 λ is longitude and φ is latitude. The horizontal distance ΔH of each set of GPS coordinates to the reference point was then calculated using equation 3:

$$\Delta H = \sqrt{\Delta long^2 + \Delta lat^2} \qquad (3)$$

This information was used to:

(1) Calculate a range of error metrics that indicate the horizontal distance (m) from each GPS receiver that contain the specified percentage of GPS data points for each sampling interval (30 s, 10 s, and 5 s) for treatments OF and HR,

(2) Determine if statistical differences existed between the horizontal accuracies of GPS receivers within each tested sampling interval within treatment and remove receivers prior to animal work if necessary, and

(3) Determine if treatment and sampling interval had a significant impact on the GPS error distributions.

For both OF and HR experiments, the three sampling intervals tested were 30 s, 10 s and 5 s, undertaken in 3 separate trials lasting 26 hrs each. Tests were conducted over this length of time to minimise any biases in satellite location and movement that may occur if tests were conducted over shorter periods. Sampling intervals were chosen to reflect the potential resolution required for monitoring cattle behaviour at pasture (Section 3.2.3) and to test whether positional fixes differed significantly between each. For each test, GPS receivers were randomly selected and placed in waterproof bags before securing to the top of each wooden post using cable ties. Raw data were downloaded in spreadsheet format. Each row of data contained the date and timestamp, latitude, longitude, speed (mph) and altitude (m). Of the 26 hrs of data collected for each receiver, statistical analysis was undertaken on 24 hrs only. The first hour of data were discarded for each receiver to reduce the risk of erroneous GPS fixes due to cold start (Duncan *et al.*, 2009) and the final hour was discarded to allow for GPS removal.

3.2.4.3 Statistical analysis

Part 1 of the analysis used the formulae provided in Table 1 to compute the error metrics for each GPS receiver. These error metrics have been used as a standard for GPS error measurement in other studies (Agouridis *et al.*, 2004) but some can be expressed differently on occasion, for example R95 (Table 1) can sometimes be defined as the circular error probability at the 95% level (CEP-R95) (Meunier *et al.*, 2018). For part 2 of the analysis, the horizontal accuracies of the 36 GPS receivers were tested for normality within sampling interval for each treatment using the Shapiro-Wilk test. Depending on the result of this test, an appropriate one-way parametric or non-parametric test was chosen to check for differences between receivers. This was

followed by post-hoc testing if $P < 0.05$ and the process repeated for all sampling intervals and treatments. If any receivers were deemed to be erroneous in at least one test, they would be discarded from the sample before part 3 was undertaken. For part 3, the new dataset was retested for normality using the above procedure, and an appropriate one-way statistical test was chosen to check for significance between sampling intervals within treatment and also for differences in the mean error values between treatments. All statistical analyses were undertaken in R.

Table 1. Horizontal accuracy metrics used to measure GPS performance as defined by the Institute of Navigation (1997)

Measure	Formula	Probability	Definition
CEP (Circular error probability)	$0.62\sigma_y + 0.56\sigma_x$	50%	Radius of a circle centred at the true position containing 50% of the data points.
1 Sigma	$\sqrt{{\sigma_x}^2 + {\sigma_y}^2}$	68%	Square root of the average of the squared horizontal position errors.
R95	$R(0.62\sigma_y + 0.56\sigma_x)$	95%	Radius of a circle centred at the true position containing 95% of the data points. R = 2.08.
2 Sigma	$2\sqrt{{\sigma_x}^2 + {\sigma_y}^2}$	98%	Twice the value of 1 Sigma.

3.2.4.4 Results and discussion

The mean horizontal accuracies (m) for all 36 GPS receivers are shown in Table 2 for all sampling intervals within treatment. Some receivers recorded horizontal distances from the known location that were much greater than the distances recorded by the majority of receivers in each sampling interval (full set of horizontal accuracies available in Appendix 1). This is shown to an extent by high standard deviations in Table 2. Furthermore, Shapiro-Wilk tests for normality showed ($P < 0.001$) that receiver horizontal accuracies within sampling intervals were not normally distributed. A Kruskal-Wallis one-way ANOVA was undertaken ($P < 0.001$) followed by a pairwise Wilcoxon rank sum test that showed that a total of 11 receivers (1, 7, 9, 14, 15, 18, 25, 31, 33, 35, 36) produced error distributions that were significantly different ($P < 0.001$) from the others in at least one sampling interval. It was also found that these 11 receivers were consistently erroneous across treatments (OF and HR). This is an important observation as this meant that the initial sample of GPS receivers could no longer be considered replicates of each other and that removal of erroneous units was essential.

Table 2. Mean horizontal accuracies (m) of 36 GPS receivers across sampling intervals within treatment

Test	Horizontal error measures					
Open field	Min (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Max (100%)
30 s						
Mean	0.24	1.58	1.93	3.29	3.86	16.48
SE	0.02	0.15	0.18	0.30	0.35	5.63
SD	0.14	0.87	1.05	1.82	2.11	33.76
10 s						
Mean	0.13	1.86	2.43	3.87	4.85	34.05
SE	0.01	0.30	0.43	0.62	0.86	15.07
SD	0.08	1.79	2.59	3.73	5.19	90.39
5 s						
Mean	0.18	2.00	2.66	4.17	5.32	130.31
SE	0.02	0.52	0.84	1.09	1.67	110.05
SD	0.14	3.15	5.02	6.55	10.05	660.33
Hedgerow						
30 s						
Mean	0.10	1.81	2.29	3.76	4.59	19.95
SE	0.01	0.12	0.16	0.24	0.31	2.13
SD	0.06	0.70	0.94	1.47	1.88	12.76
10 s						
Mean	0.10	15.08	18.70	31.37	37.41	330.01
SE	0.01	13.30	16.47	27.65	32.94	306.69
SD	0.08	79.77	98.82	165.92	197.65	1840.17
5 s						
Mean	0.08	1.90	2.38	3.96	4.76	25.48
SE	0.01	0.09	0.12	0.19	0.23	5.08
SD	0.04	0.54	0.70	1.11	1.40	30.49

Subsequently, the data for each receiver deemed to be faulty were removed and the data from the remaining units (n = 25) were retested and normality was confirmed for all horizontal accuracy measures. The standardised datasets (without erroneous GPS receivers) for OF and HR can be seen in Appendix 2, but Table 3 shows the mean values for each sampling interval and treatment. Notably, the mean values for each sampling interval were more consistent, with a reduction in the variability of the data across both treatments. One-way ANOVA were undertaken to test for differences in the errors between sampling intervals within treatment, but none were found ($P > 0.05$). The mean (± SD) error at the R95 level for the OF test was 2.73 m (± 0.11 m) and in the HR test, this was 3.41 m (± 0.46 m). Despite absolute accuracy not having a direct impact on the methodologies employed in Chapters 1-3, this knowledge is useful as it highlights the capabilities and limitations of the receivers in use. It also meant that industry-standard values were available for comparing receivers and that if necessary, in future work, they could be used to identify the location of animals to within a given level of precision which could be important in studies of grazing ecology for example. Given that absolute positional fix accuracy can also change due to the movement of terrestrial animals, future studies should consider the effect of movement on GPS output and whether fine-scale behaviours are masked by any uncertainties in GPS measurements (Laube and Purves, 2011). This should be repeated using a range of sampling intervals to at least include those intervals tested here.

Table 3. Mean horizontal accuracies (m) of reduced set of 25 GPS receivers across sampling intervals within treatment

Test	Horizontal error measures					
Open field	Min (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Max (100%)
30 s						
Mean	0.24	1.31	1.60	2.73	3.21	7.23
SE	0.03	0.01	0.01	0.02	0.02	0.09
SD	0.15	0.04	0.05	0.08	0.10	0.47
10 s						
Mean	0.13	1.31	1.61	2.73	3.21	7.52
SE	0.02	0.01	0.01	0.02	0.03	0.10
SD	0.08	0.05	0.07	0.11	0.13	0.50
5 s						
Mean	0.19	1.32	1.62	2.74	3.23	9.74
SE	0.02	0.01	0.02	0.03	0.04	0.33
SD	0.12	0.07	0.09	0.15	0.18	1.66
Hedgerow						
30 s						
Mean	0.11	1.62	2.04	3.36	4.08	16.58
SE	0.01	0.05	0.07	0.10	0.13	1.78
SD	0.07	0.23	0.34	0.48	0.67	8.88
10 s						
Mean	0.09	1.58	1.96	3.28	3.92	14.70
SE	0.02	0.03	0.03	0.06	0.07	1.55
SD	0.08	0.13	0.17	0.28	0.35	7.76
5 s						
Mean	0.08	1.73	2.15	3.59	4.30	15.93
SE	0.01	0.06	0.08	0.12	0.16	2.14
SD	0.04	0.29	0.40	0.61	0.79	10.68

Finally, using the standardised group of receivers, one-way ANOVA were undertaken to test whether the observed pairwise differences between the data gathered in both OF and HR treatments were significantly different from each other. For this analysis, the horizontal accuracy values across sampling intervals were grouped given that sampling interval was found to have no impact on the error distribution of the

standardised dataset. Except at the minimum error level (P = 0.06), the observed differences in static accuracy between OF and HR receivers were highly significant (P < 0.001), likely because GPS signals were being reflected from the hedgerow leading to the phenomenon known as multipath error (de Weerd *et al*., 2015). At the R95 level, the HR data were on average 0.68 m further from the measured reference point compared to the OF data (Figure 4). The maximum values recorded for the HR receivers were also more variable than the maximum values recorded for the OF receivers (Table 3 and Figure 4).

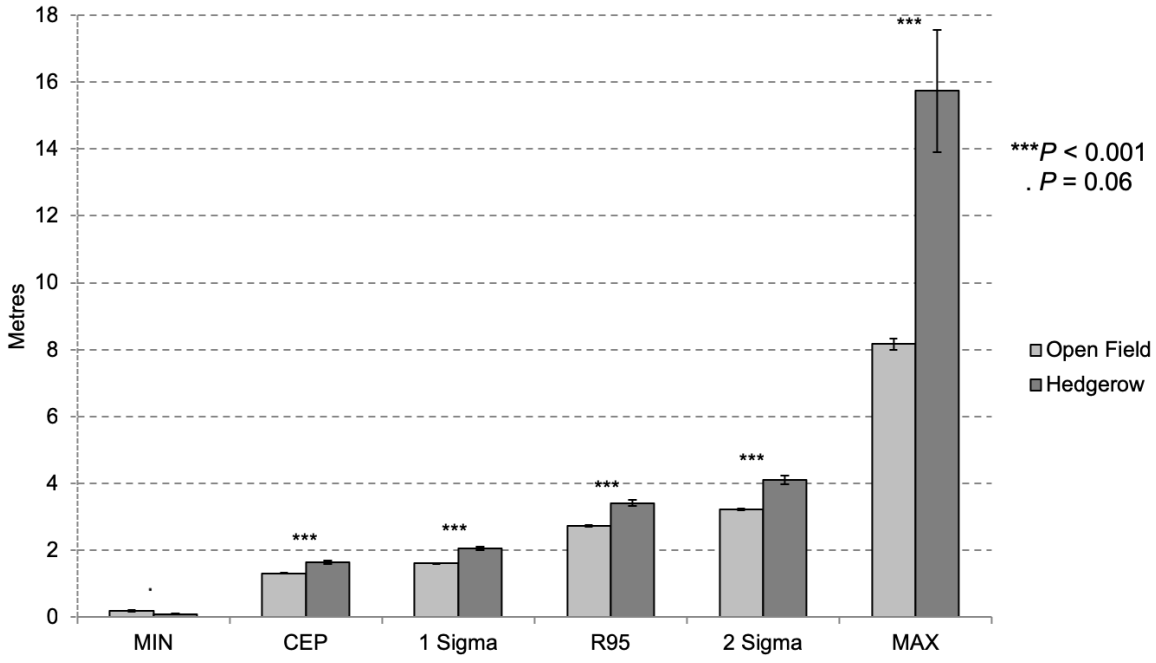


Figure 4. Pairwise comparisons of error measures of 25 standardised GPS receivers between open field and hedgerow treatments. Data are grand means across three sampling intervals (30 s, 10 s, 5 s). Bars are SE.

This meant that for any animal experimentation, only the sample of 25 standardised GPS receivers would be used. Consistency is important in the data gathered for modelling because any erroneous data appearing in training sets are likely to lead to unstable classification models (Lange *et al*., 2003). Given the significant impact of the hedgerow on the GPS error distributions, it was decided that all animals

used for subsequent data collection would be managed to avoid any objects that could affect positional fix accuracy. This is an obvious limitation to the methodology and account for such structures would be needed in future work. As such, all animal work described in Chapter 1 was undertaken in a 1-ha paddock located in the middle of the field used for the OF calibration experiment (Figure 3). More detail can be found in Chapter 1: *Grazing management and GPS*.

3.2.4.5 Summary of GPS static accuracy tests

The results indicate the importance of testing receiver static accuracy. This is evidently needed not only to standardise the GPS receivers used but also to test the impact of sampling interval and also physical structures on the data. It was concerning to find that some receivers produced significantly different error values to others. The documentation for the GlobalSat DG-100 GPS (https://www.gpscentral.ca/manuals/gs-dg100.pdf) states that the receiver has a horizontal 2DRMS accuracy of 10 m. This measure is equivalent to 2 Sigma (Agouridis *et al*., 2003) and it can be seen (Table 3) that the mean errors for the standardised sample of receivers was within this quoted distance (mean OF = 3.22 m; mean HR = 4.1 m). In addition, a dynamic test could have been undertaken to consider the impact of animal movement on error distribution and should perhaps be considered for future work (Schipperijn *et al*., 2014). Given that sampling interval was found not to affect receiver accuracy, it was decided that an interval of 5 s would be used for animal work as this would provide the greatest resolution for the battery power available. It was found that at this sampling interval (5 s); two AA batteries would provide enough power to gather data for 26 hrs.

To minimise the influence of noise or error in GPS samples, preprocessing methods can be implemented prior to further data manipulation. The application of moving-average windows or median filters that smooth data to remove outlying

instances are often used (Clapham *et al*., 2011; Studd *et al*., 2019). Fixed windows are passed over the data to generate a new dataset that represents the average (or median) of the original data from each fixed window. The size of the window is determined by the analyst and many window sizes can be tested. These methods have been shown to improve classification models and improve changepoint detection in both the PLF (Behmann *et al*., 2016; Lush *et al*., 2018) and the movement ecology literature (Meckley *et al*., 2014; Gurarie *et al*., 2016). Median filtering usually requires that thresholds are set where instances that are above (or below) these thresholds are replaced with the median of the window. Although not used in the chapters herein, data filtering techniques could be effective in removing erroneous GPS instances that could arise due to some of the challenges described in Section 3.2.

3.3 Other sensors and tools for behaviour classification

3.3.1 Muzzle pressure sensors

3.3.1.1 IGER recorder

While accelerometers and GPS occupy the majority of current PLF research in cattle behaviour classification (especially the former), other sensors have been in use for a number of years and are likely to feature more in future research. Pressure sensors are one such example. One of the first for use in ruminants (sheep) was developed by Penning (1983) that recorded jaw movements from the stretching and contraction of a noseband. They were able to classify grazing, ruminating and idling every minute but success on shorter bouts of behaviour was low. An updated version of this sensor (IGER recorder) was developed by Rutter *et al*. (1997) using a microcomputer recording system rather than a cassette recorder (Figure 5A). Post-processing of the data is then undertaken using bespoke software (Rutter, 2000). The overall level of agreement between manual observations and instances classified by the system was 91% for feeding, ruminating and 'other' behaviours.

3.3.1.2 Rumiwatch

More recently, a low-profile halter-based noseband pressure sensor; Rumiwatch (Figure 5B), has been developed for measuring the attributes associated with feeding in dairy cows (e.g. eating chews and prehension bites). Again, this technology uses bespoke software for post-processing. The most recent publication to examine the performance of this system (Rombach *et al*., 2018) found that it predicted the number of rumination boluses, rumination chews and total eating chews very well (prediction error < 0.10). However, the prediction errors for prehension bites and time spent in prehension and eating was higher (prediction error > 0.10). As previously discussed, (Section 2.3), one of the variables found to affect the performance of this system in this

most recent publication was the feed delivery method. Producing robust models and software that can deal with this type of variability will likely be a significant area of focus for researchers moving forward.

3.3.2 Acoustic sensors

Another well-established method for behaviour classification (particularly feeding) is acoustic monitoring. Ungar and Rutter (2006) found that when compared to the IGER recorder, a microphone placed on the head of cows was able to produce comparable classification results. Milone *et al*. (2009) used a wireless microphone attached to the heads of sheep (Figure 5C) and a hidden Markov model to automatically segment and classify chewing events produced from the consumption of two different forages. Overall, chewing events were correctly identified 82% of the time. More recently, Vanrell *et al*. (2018) used a head-mounted microphone and recording device (Figure 5D) and a two-stage segmentation and classification approach of foraging behaviours. The regularity of events was detected by the autocorrelation of the sound signature which was used to define the time boundaries of the events. Feeding attributes were then classified using features that are unique to each type of behaviour such as the number of pauses (higher during rumination compared to grazing). F-measures of 0.89 and 0.93 were achieved for rumination and grazing respectively.

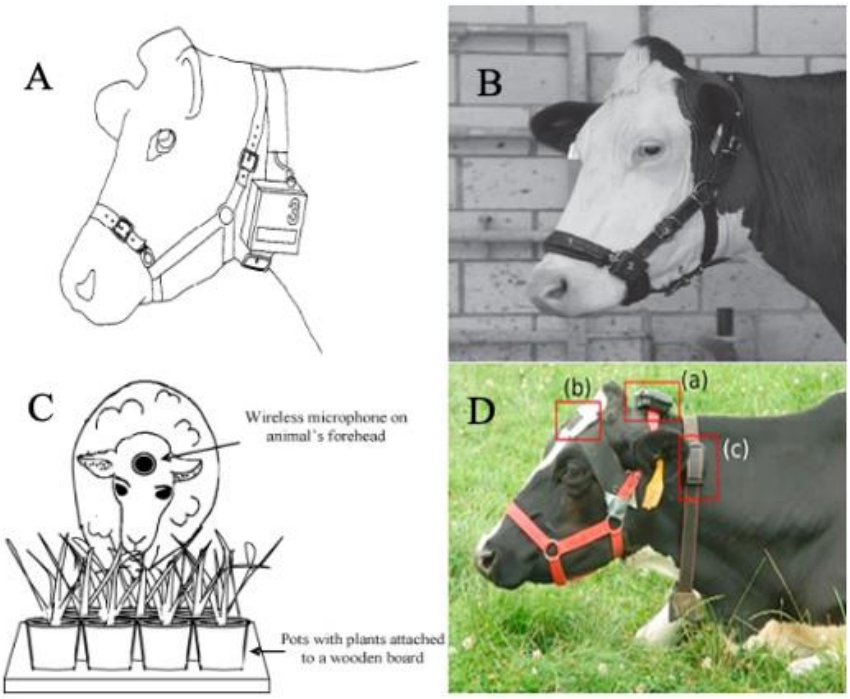


Figure 5. Examples of pressure and sound recording sensors used for ruminant behaviour classification. A = IGER recorder (pressure); B = Rumiwatch (pressure); C = head-mounted microphone; D = head-mounted microphone and recording device (b and a respectively). Images adapted from Rutter *et al*. (1997), Rombach *et al*. (2018), Milone *et al*. (2009) and Vanrell *et al*. (2018) respectively.

3.3.3 Local positioning sensor systems

A novel technique for measuring the location of cows in housed environments is gaining interest because of the insight that it can provide on the spatio-temporal behaviours of cows. Local positioning sensors (LPS) are in principle very similar to GPS in that they can provide information on the absolute position of cows, and given frequent sampling can be used to determine the movement of cows throughout a barn. The system comprises two main components; static receivers that are located within the barn and mobile sensors that can be fitted to cows that relay their positions at predetermined intervals to the static reveivers. The location of cows within the barn is calculated by triangulation with each static receiver, providing positioning in coordinate format (x, y). The location accuracy of the mobile sensor can be measured using methods similar to those reported in Table 1 and these analyses were undertaken by

Barker *et al*. (2018). They found that the location accuracy of the LPS system differed depending on whether cows were standing or lying, and this could not be improved upon using post-processing methods on the raw data. To mitigate some of the issues that could have arisen given inaccuracies in the measured location data, a moving average window was applied to the raw data to remove outliers.

These systems can be combined with other precision sensors to identify key behaviours throughout the housed environment. In conjunction with the LPS system, Barker *et al*. (2018) used a decision tree algorithm to classify the behaviour of housed dairy cows that were fitted with neck-mounted accelerometers. The LPS system was used to supplement the classification procedure to monitor the feeding behaviour of lame and non-lame cows. Diosdado *et al.* (2018) also used a similar system to monitor the space-use patterns of housed dairy cows. LPS is likely to be a very valuable tool to monitor cattle in housed environments as it can provide behavioural information that was previously inaccessible using other precision sensors such as GPS.

4. Animal movement modelling and path segmentation

4.1 An overview of methods in movement ecology

The development of GPS has permitted the collection of large amounts of data from a variety of species and as such, a variety of movement models have been applied and developed over recent years. At its simplest level, animal movement can be modelled using uncorrelated random walks with its fundamental application to the motion of pollen particles almost two centuries ago (Brown, 1828). In these models, the directionality of movement is completely independent of the previous direction moved and the location moved to depends only on the previous location (Codling *et al.*, 2008). Given the directional persistency that animals tend to demonstrate during movement, correlated random walks are often used to provide a more realistic representation of movement (Codling and Hill, 2005). Animals may also demonstrate a particular bias to certain targets during movement and these biases (either fixed throughout the whole movement path, or space and time dependent) can also be accounted for using biased random walks (Benhamou, 2006). Sampling interval will also affect the prediction of movement using random walk models and longer intervals will lead to the appearance of a more random pattern of behaviour (Codling and Hill, 2005). Step length (distance between successive relocations) is also an important feature of animal movement and generally, the step lengths of animals modelled using random walks have a finite variance although other methods (e.g. Lévy walks) have been proposed as models for animal movement where step lengths have an infinite variance (Viswanathan *et al.*, 2000). In these models, focal animals are assumed to switch between two or more walk patterns and can be useful when considering behaviour which may involve movement within a restricted area (Benhamou, 1992).

Composite correlated random walk is another random search model where area-restricted searches feature as one of the modelled behaviours (Benhamou, 2007). These two-behaviour models are typically used to account for 'intensive' and 'extensive' phases of behaviour, typically foraging events and the location of food patches (Dragon *et al*., 2012). Resources may be abundant but not spatially uniform and so it would be expected that animals would focus their attention to those patches of resources and thus exhibit what is known as an area-restricted search (Kareiva and Odell, 1986). These may be characterised by increased turning rates and slower speeds in GPS data (Fauchald and Tveraa, 2003). In reality, an animal may be responding to a number of environmental stimuli and thus, may demonstrate a number of different walks with different turning rates and speeds. To quantify differences in behaviours along a path, the first-passage time approach is an intuitive method which describes the time taken by animals to pass through a circle of a given radius (Johnson *et al*., 1992). An example of its application could include the time taken by a predator to locate its prey within a given area. First-passage time has been used extensively to model multi-species behaviours (Fauchald and Tveraa, 2003; Frair *et al*., 2005) and is being increasingly used for studying the movements of migratory and endangered species (Suryan *et al*., 2006) with further work examining its applicability to situations where the scale of area-restricted searches varies within the same movement path (Pinaud, 2008).

State-space models have also become a prominent methodology for predicting animal behaviour based on observations (e.g. location and speed) coupled with a process model which predicts the future state of the animal (e.g. behaviour) based on the observations (usually recorded by GPS or other sensors) (Jonsen *et al*., 2006; Breed *et al*., 2006). In mathematics, this assumption is known as the Markov condition and its principles are considered in Chapter 2 as a method for predicting cow behaviours. A

benefit of state-space models is that they combine three major phases of analysis, including error correction (e.g. correction of GPS movement paths based on assumptions of animal movement), calculation of movement metrics and pattern identification (statistical analysis). Furthermore, environmental information can be incorporated into these models leading to more realistic behavioural inferences (Jonsen *et al*., 2003). Indeed, understanding more about animal movement often requires path-level segmentation of movement trajectories to dissect paths into a number of discrete states that are assumed to represent different underlying behaviours (Edelhoff *et al*., 2016).

In many studies of animal movement, data segmentation is a necessary step for partitioning high frequency sensor data. Two main strategies exist in the PLF literature for the segmentation of sensor data, these are fixed-time segmentation (FTS) and variable-time segmentation (VS) with the former being used in the majority of PLF studies (Section 4.1.1). VS, on the other hand, despite having its roots in process control and fault detection has been very well developed over the last few decades for path segmentation of animal movement data where the behaviour of the focal species is not very well characterized largely due to observational difficulties. The selection of the appropriate segmentation strategy depends on a number of variables including the data collection methodology and the subsequent use of any derived movement model. The basic aim in each case is to characterize the behaviour in each segment from datasets that are ultimately samples of the underlying behavioural processes represented across time or both space and time (Calenge *et al*., 2009). From the perspective of animal movement ecology, path segmentation has been an important area of animal science for decades and the rapid increase in the technologies available for monitoring animal movement has led to the further development and use of sophisticated mathematical

solutions for path segmentation (Nathan *et al*., 2008). This has allowed scientists to learn more about the biological and environmental processes behind the behaviours that are driving the observed movement patterns (Killeen *et al*., 2014). To make behavioural inferences, path segments can be analysed for summary statistics that convey relevant information about individual movements and a large number of measures can be used to describe these segments (e.g. velocity, turning angle and step-length). It is important to note however that some segmentation methods do not provide segment summary statistics as they have been developed mainly for process control and do not require ecological context. It follows then that some methods require further post-processing to derive meaningful information from the resultant segments (Section 4.1.2.2.2 and Section 4.1.2.2.3). Regardless of the method employed for VS, three broad categories of analytical approaches exist, namely topology-based approaches (describe movement patterns quantitatively), timeseries (detect points in time where behaviour shift occurs) and state-space-based modelling approaches (identify hidden states), more of which are discussed in Section 4.1.2.2.

The following sections begin with a discussion on the use of FTS largely in the context of PLF data. VS is then discussed, providing background on the various techniques available and some examples of the use of VS both in the literature and on simulated data.

4.1.1 Fixed-time segmentation

4.1.1.1 Background

Many studies using ML techniques for automated livestock behaviour prediction have developed their classification models from sensor data partitioned using FTS (Martiskainen *et al*., 2009; Dutta *et al*., 2015; Rahman *et al*., 2018; Benaissa *et al*., 2019) including Chapter 1 herein (Williams *et al*., 2016). The classification models are usually derived from supervised datasets where the focal animal has been observed as ground-truth to the electronic data that were gathered. This data may have been collected from a single sensor or a combination of sensors placed on the animal. For modelling, FTS requires that data streams based on single behaviours (e.g. grazing) are taken and broken down into segments of a predetermined (fixed) size for the derivation of summary statistics or feature vectors; a set of explanatory variables extracted from each segment (Section 5.2). From the perspective of a movement ecologist interested in the underlying processes (hidden states) of a migratory species, there may be several considerations for choosing the appropriate segmentation strategy (Section 4.1.2) or indeed the general classification methodology, but for those interested in monitoring the behaviour of livestock (e.g. cattle) an FTS strategy is usually appropriate for ML. This is because the most important behaviours exhibited by cattle (from a farmer's perspective) are few and usually include feeding, ruminating, standing, lying, drinking and walking and possibly combinations of these. It is also well established that these behaviours correlate highly with animal performance and health (Liboreiro *et al*., 2015; Stangaferro *et al*., 2016; Kaufman *et al*., 2016) and that they tend to take place over long periods of time (minutes) (Thorup *et al*., 2016). Grooming, for example may not be as important and it may not be practically possible to identify from GPS or accelerometer data due to its intermittent nature and difficulty in obtaining enough

ground-truth observations as training data (González *et al.*, 2015; Ungar *et al.*, 2018). There may be other, species-specific behaviours that are useful to identify such as nursing (Yang *et al.*, 2018) and aggression (Lee *et al.*, 2016), which are important in pig-rearing systems and such behaviours may require the use of alternative sensors and video capture technologies. The relative ease of annotating sensor data gathered from domesticated animals means that comprehensive information can be collected on how the movement of the animal is reflected in each data signature. Conversely, an analyst interested in deciphering the behaviour of elusive animals can only speculate as to the behaviours that are taking place in each dataset. As such, several tools have been developed and refined to partition unsupervised datasets so that movement ecologists can gain a better insight into the behaviours of various species (Section 4.1.2).

4.1.1.2 Classifying data streams using fixed-time segmentation

When using a model developed using FTS, it must usually be deployed in a moving window of derivative size. Using useful and identifiable behaviours (e.g. grazing, lying, and standing) means that for the majority of the time, a moving window of fixed size is appropriate for cattle because the majority of behaviours are not sporadic in nature. Supervised classification with a robust and optimised classifier is therefore likely to yield good classification performance when deployed in real-time. However, a moving window will not always meet the exact behaviour transition points in a timeseries, and this can lead to misclassifications (Figure 6).

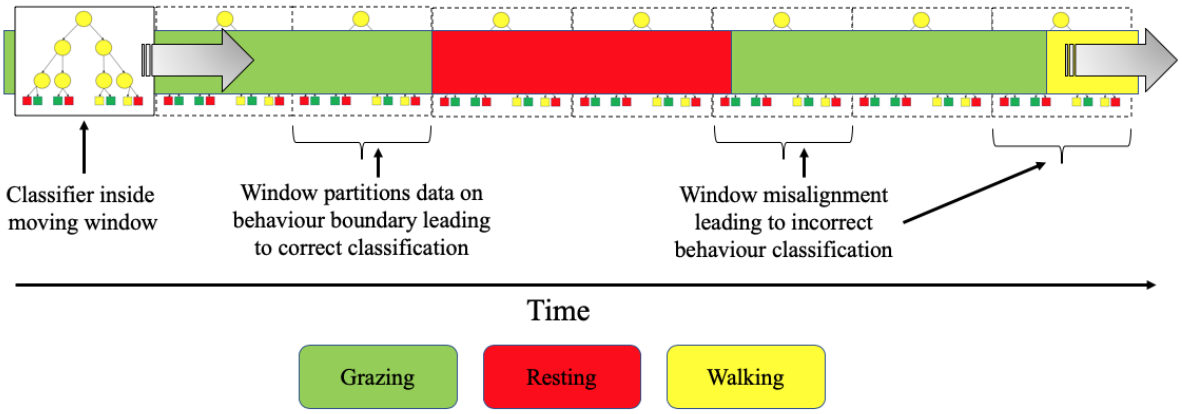


Figure 6. Fixed-time data segmentation and classification of cow behaviours. An analysis window (learned classification algorithm) moves along the timeseries, segmenting the dataset and extracting data features for behaviour classification. On some occasions, the analysis window will not segment at the exact behaviour transition boundaries and this will likely lead to behaviour misclassifications. Behaviours grazing, resting and walking indicate the ground-truth observations.

Deriving the optimal fixed window size is a function of the behaviours to be identified which in turn depends on the sensor used, the species of animal, its environment, amongst many other variables. Another important consideration is the temporal frequency of data capture. Frequent data capture (e.g. seconds) will allow for the definition of finer-scale behaviours but can be more prone to error due to spatial inaccuracies (GPS) and high sensitivity to movement (e.g. accelerometers) (de Weerd *et al*., 2015; Benaissa *et al*., 2019). However, lower sampling frequencies can reduce the variety of behaviours that can be identified (Benaissa *et al*., 2019). Sampling frequency is an important area of discussion in the movement ecology literature as in reality, animal movement (especially non-domestic species) often involves finite distances moved in finite times (Plank and Codling, 2009). In movement ecology, the discrete data points gatherd from continuous movements can be modelled but this discretisation can bias the movement path which can lead to issues such as over or underestimations in distances travelled (Nams, 2013). One of the key factors in

determining distances travelled is the tortuosity of the movement path which is usually species dependent and much work has been undertaken recently to try and make improved estimations of animal movement paths using various sampling frequencies and modelling strategies (Marcus Rowcliffe *et al*., 2012; Steenweg *et al*., 2018). For the study of livestock movement paths, these principles will be important to consider in future research, both where aggregation and FTS are required but also for modelling movement paths using techniques that are already well established in the movement ecology literature.

In PLF studies, the breadth of behaviours identified has increased over the last decade largely due to improved sensor capabilities and lower cost. Furthermore, analysts were mainly confined to sensors that only recorded spatial and temporal data at low frequencies mainly due to battery power and data capacity. As such, selecting an appropriate window size for FTS requires careful optimisation. However, analysts must recognise that performance can often be degraded when classifiers are deployed on continuous data streams because of window misalignment at transition points (and also due to segments not previously seen in the training set) (Bom *et al*., 2014). Unfortunately, this is often overlooked in livestock behaviour-classification studies. The majority of studies only undertake model training and testing (e.g. using cross-validation) on independent data segments and do not consider the classification of continuous data sequences (Rutten *et al*., 2013). Some are recognising this issue and have attempted to classify transition points automatically (Diosdado *et al*., 2015).

Chapter 2 discusses a method for correcting segments that are erroneously classified using FTS. This method was adapted from Feldman and Balch (2004) where the authors used the output of a kernel regression classifier as input to a hidden Markov model (HMM) to classify the movements of honeybees. The HMM was used to

statistically correct unlikely classifications from the classifier based on the known movements of honeybees. This led to an overall improvement in CA. Data from GPS receivers attached to cattle can be considered emissions of the true behaviours (hidden states) that are not visible (behaviour predictions are made from the data). Each hidden state therefore has a probability over all emissions. Probabilities can also be estimated for the transitions between states, for example, the probability of moving from a state of grazing to a walking state. HMMs use this information for estimating the likely state sequences. HMMs have been used extensively for speech recognition (Rabiner, 1989) and for behaviour prediction in animal research (Pedersen *et al*., 2011). In Chapter 2, an HMM was used to correct the state sequences produced from the model developed in Chapter 1 over several datasets. For an overview of HMMs and optimum state selection methods, see Pohle *et al*. (2017).

4.1.2 Variable segmentation

4.1.2.1 Background

Variable segmentation (VS) as a strategy for detecting significant changepoints in timeseries data has its origin in fault detection for industrial processes (Lai, 1995) but also has uses in monitoring medical conditions (Bosc *et al*., 2003), detecting climate change (Reeves *et al*., 2007) and in speech recognition (Chowdhury *et al*., 2012) amongst others. In some applications of movement ecology and for the majority of publications to date that have modelled livestock behaviour from discrete sensors, FTS is used to iteratively segment the timeseries using a predetermined segment size (Section 4.1.1). The objective of VS is to partition the timeseries based on significant changes in one or more properties of the dataset (Kawahara and Sugiyama, 2012) either retrospectively or in near real-time. For the majority of studies of animal movement ecology, retrospective methods (often known as ‘offline’ analyses; see Section 10.4 for

further discussion) are used on data recovered from animals to detect significant changepoints. Once the timeseries has been partitioned, expert opinion can be used to label the behaviours occurring between each partition point, but states can also be applied using statistical techniques such as clustering (Zhang *et al*., 2015). Given the background of VS as a technique for modelling multispecies behaviours from sensor data, Chapter 3 evaluates the use of VS for modelling dairy cow behaviour from a supervised dataset. Before discussing a small sample of VS methods, the next section provides some background on animal movement modelling and the range of methods discussed in the movement ecology literature.

4.1.2.2 Variable segmentation methods

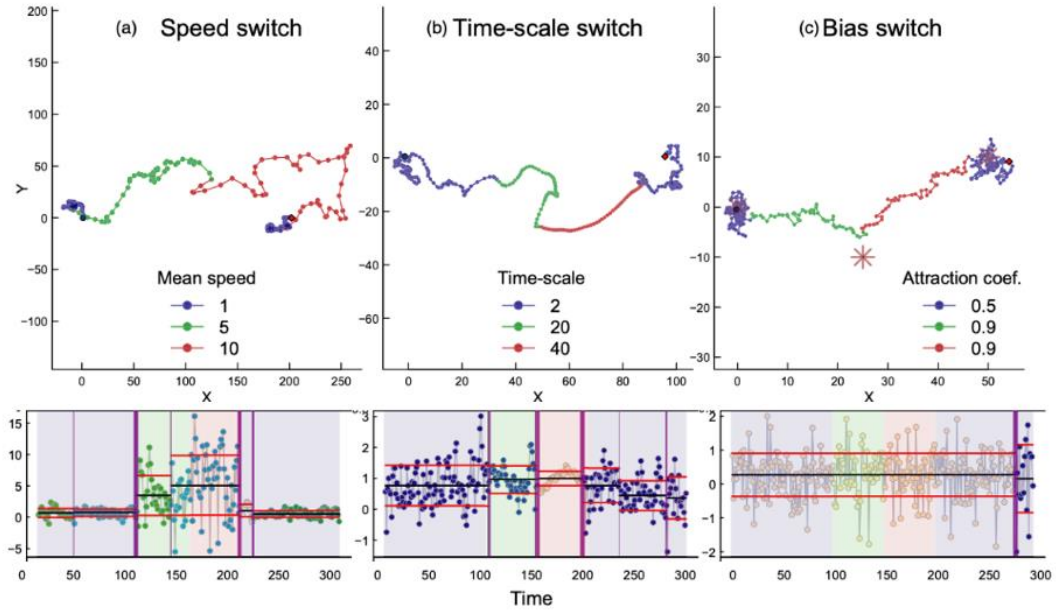
The field of movement ecology is well known for its development and use of changepoint detection methodologies for quantifying animal movement. These methods are fundamental in determining the behaviour of individuals and groups of animals and for studying the impact of ecosystem changes, infectious diseases and other perturbations on animal movement (Nathan *et al*., 2008). Fundamental to these analyses is the development of sensors that are able to effectively store and relay information about the focal animal at a timescale that best represents the overall movement of that animal. Movement paths can then be analysed for distinct elemental phases that can provide information on the internal states of animals as well as their motivations for movement. Methods that describe these movement patterns quantitatively are known as topology-based methods. These methods describe geometric properties of the data and group segments based on similarities in these properties. Typical methods that fall under this category of path analysis are thresholding or clustering methods (Van Moorter *et al*., 2010; Dzialak *et al*., 2015). Time-series analysis can be used to detect significant changes in a time-ordered set of observations and are widely used in ecology

and related disciplines (Gurarie *et al*., 2009; Madon *et al*., 2014). Some of these methods rely on user-modified parameters such as specifying the maximum number of changepoints in a sequence or the minimum length of the detected segments (Chapter 3). Some methods can also account for temporal correlation in sequential signal data (Gurarie *et al*., 2009). Finally, and as mentioned previously, state-space models provide an opportunity to identify hidden behavioural states based on the observed movement data (Chapter 2). In this method, a probabistic approach is used to predict the future state (behaviour) of the animal based on its current state which is derived from the emitted signals gathered from the on-animal sensor in many cases. Movement paths are then segmented based on state membership. Several VS methodologies have been developed and used to reveal insights into the spatial and temporal movement of animals but only a sample are discussed here, focussing on some of the most prominent segmentation methods in the movement ecology literature. Further VS methods are discussed in Section 10.4 and Chapter 2 describes the use of HMMs for modelling cow movement which share very similar properties with state-space models.

4.1.2.2.1 Behavioural changepoint analysis

The behavioural changepoint analysis (BCPA) is one example of a VS technique (Figure 7) that iteratively 'sweeps' an analysis window over a computed metric of the timeseries dataset (persistence velocity; V_p) derived from velocity, turning angle (angle between GPS fixes) and step length (distance between GPS fixes) (Gurarie *et al*., 2009). Changepoints are identified when a significant change is detected between the current analysis window and the previous using maximum likelihood methods. A change may be represented by a shift in a single or a combination of three properties of the data (mean, variance and autocorrelation) represented by seven different models. For example, a significant change in the mean of V_p may occur between two windows

but variance and autocorrelation stay the same. If no change is detected, then the null model is selected. BCPA has been used to identify the behaviours of Pacific black duck (*Anas superciliosa*) (McEvoy *et al*., 2015) and also to assess the interactions between greater sage-grouse (*Centrocercus urophasianus*) (Prochazka *et al*., 2017). The BCPA package is available in R (Gurarie, 2015). BCPA is widely used in the field of movement ecology and to the candidate's knowledge, it has not been used on livestock movement data to date.



Adapted from Gurarie *et al*. (2016)

Figure 7. Simulated tracks (a-c) each consisting of 300 equally spaced observations with changes in mean speed (a), tortuosity (b) and home-ranging behaviour (c). Bottom panels show the behavioural changepoint analysis (BCPA) applied to each track. Vertical bars (purple) indicate significant changepoints in the simulated tracks.

4.1.2.2.2 Changepoint model

The changepoint model (cpm) can be used to detect changes in the mean or variance of both parametric and non-parametric timeseries using retrospective changepoint detection (batch detection) or detection in almost real-time (sequential detection) (Ross, 2015). Examples of the single changepoint detection methods are shown in Figure 8. A number of test statistics are available for use in identifying significant changepoints depending on the distribution of the data (e.g. Student-*t* (Figure 8A) and Mann-Whitney (Figure 8B) tests). The cpm package (Ross, 2015), implemented in R has been used previously to segment accelerometer data from free-ranging crab plovers (*Dromas ardeola*) (Bom *et al*., 2014) and a golden eagle (*Aquila chrysaetos*) (Sur *et al*., 2017) prior to supervised ML of the variable segments. Bom *et al*. (2014) achieved significantly better classification performance for some behaviours

with VS compared to FTS, although for the majority of behaviours, the results were not significantly improved when VS was used.

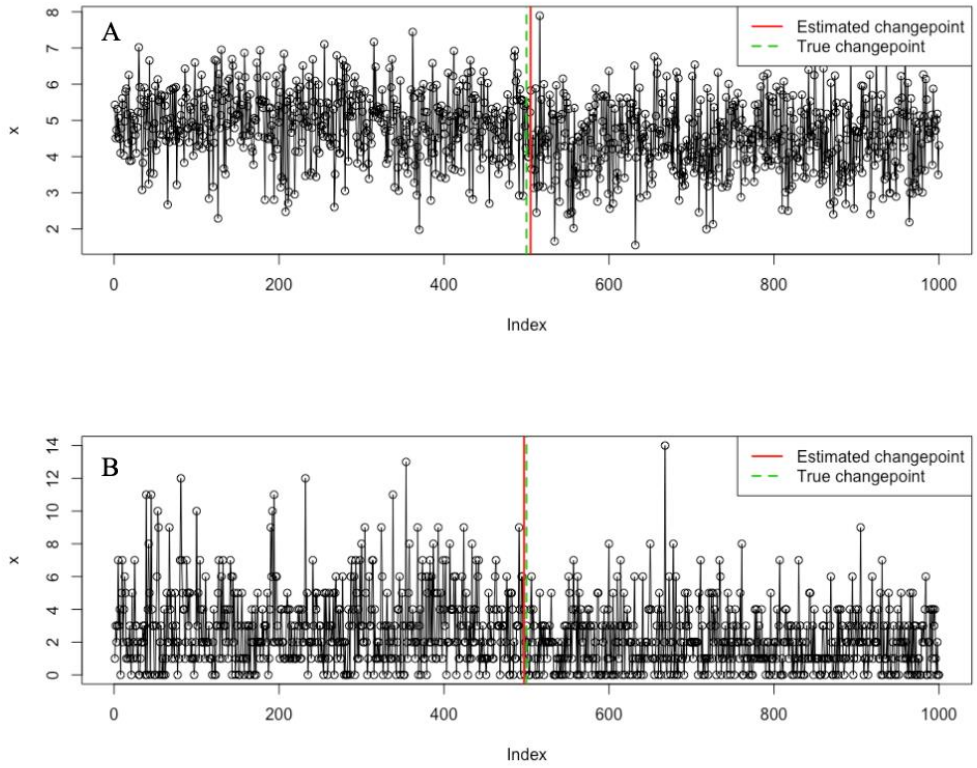


Figure 8. Simulated Gaussian (A) and binomial (B) timeseries of 1000 observations showing true and estimated changepoints as determined by parametric (Student-*t*) and non-parametric (Mann-Whitney) tests respectively. A single, true changepoint (change in mean) is located at observation 500 in each timeseries. Tests were undertaken using the changepoint model (cpm) package (Ross, 2015) in R.

4.1.2.2.3 Changepoint

A final example of VS that can be applied to animal movement data is implemented in the R package 'changepoint' (Killick and Eckley, 2014); a likelihood-based approach to changepoint detection. Changepoint provides a selection of algorithms for detecting multiple changes in the mean and/or variance of a timeseries as well as a selection of test statistics. Three multiple changepoint detection algorithms are available, namely, binary segmentation, segment neighborhood and pruned exact linear time (PELT). Binary segmentation is less computationally expensive than segment

neighborhood but a trade-off for higher speed is that binary segmentation is less exact in its estimations. PELT on the other hand is reported to share the exactness in its changepoint estimations as that of the segment neighborhood algorithm (Killick *et al*., 2012) but is computationally more efficient and has been the algorithm of choice for some (Madon and Hingrat, 2014; Ramasco *et al*., 2014; Shahriar *et al*., 2016).

Changepoint was used in Chapter 3, for VS of the transformed speed variable gathered from GPS receivers applied to dairy cows (Williams *et al*., 2019). The binary segmentation algorithm was used to search for changepoints in the variance of the timeseries primarily because in preliminary tests, this algorithm made fewer FP predictions than PELT and was computationally quicker than the segment neighborhood algorithm. Using the segmented data, feature-vectors were created for ML. Compared to FTS (Chapter 1; Williams *et al*., 2016), VS showed very promising results for classifying grazing, resting and walking which is in agreement with Bom *et al*. (2014).

4.2 Summary of segmentation methods

For PLF applications where high temporal resolutions are required for modelling infrequent, fine scale behaviours, VS techniques show great promise. More work is needed in PLF research to test this assumption using sensors such as accelerometers. Currently, the candidate is only aware of one other study that has used VS for identifying dairy cow behaviour (Shahriar *et al*., 2016) but there is potential for the use of such techniques in measuring variables associated with disease progression (e.g. temperature) and also for the prediction of calving (e.g. lying bout duration). For current research, most VS methods are based on offline analyses. Moving forward and with enhanced ability to relay sensor data in real-time, it may become possible to monitor animals in near real-time (online analysis) to detect behavioural changes

almost as they occur. Further unsupervised chagepoint detection methods are discussed in Section 10.4 but for a comprehensive review of many changepoint detection methods for timeseries data, please see Aminikhanghahi and Cook (2017).

5. Machine learning in dairy research

5.1 Background of machine learning

The availability of large and often complex datasets on farms is now providing the potential for the application of ML techniques in the dairy industry. The opportunity to exploit such data for information that may have been previously unattainable is very likely to lead to the development of new decision support tools at farm level (Kamilaris *et al.*, 2017). This data has largely stemmed from improvements in data capture technology. However, already established databases can be exploited by such techniques to reveal new insights, for example into disease phenotypes (Tsairidou *et al.*, 2018).

The dairy industry is not the only branch of livestock agriculture to benefit from these methods. Work has been undertaken to automatically identify the behaviours of sheep using accelerometers and ensemble algorithms (Mansbridge *et al.*, 2018), classify beef quality attributes using mass spectrometry data and support vector machine algorithms (Gredell *et al.*, 2019) and identify porcine disease using environmental sensor data and neural network algorithms (Cowton *et al.*, 2018). The power to exploit large datasets is continually improving and the emergence of big data technologies and powerful computing platforms is allowing for high throughput data analysis (Kamilaris *et al.*, 2017; Wolfert *et al.*, 2017).

Computerised management and decision support systems on dairy farms have been discussed for a number of years (Udomprasert *et al.*, 1990; Pietersma *et al.*, 1998). In a research context, ML for decision support has increased over the last 10 years for a

variety of dairy-related topics. A broad literature search using the keywords ‘machine learning’ and ‘dairy’ using the Web of Science (Thomson Reuters, 2019) database found 100 articles published since 1995. Thirty-nine articles were categorised into Agriculture and Dairy of which 29 (74%) were published since 2013. It is evident that the exploration of ML in the context of dairy management has become a very important field of research in the last 10 years and this trend is very likely to continue. Currently, the two most important research topics are automatic behaviour classification and the use of ML for predicting the optimal insemination time and likely outcomes in cattle.

5.2 Feature sets for machine learning

ML processes usually involve training a model on a feature set developed from a sampled raw dataset. Features are extracted to increase the dimensionality of the dataset and can take many forms (e.g. binary or numeric). The feature set is used as a collection of input parameters that an ML algorithm can learn from. In studies using accelerometers to classify the behaviours of focal animals, there are some well-established features that can be extracted that are common amongst research. For a given window of data (Section 4) and acceleration axis, the mean, skewness, kurtosis, standard deviation, maximum and minimum values are frequently computed for studies on livestock (Martiskainen *et al*., 2009; Dutta *et al*., 2015; Smith *et al*., 2016; Guo *et al*., 2018; Rahman *et al*., 2018) and other species (Yoda *et al*., 2001; Nathan *et al*., 2012; Bom *et al*., 2014; Williams *et al*., 2015; Sur *et al*., 2017). Other features that are also computed frequently include vectorial and overall dynamic body acceleration (VeDBA and ODBA respectively) (Diosdado *et al*., 2015; McClune *et al*., 2015; Benaissa *et al*., 2017; Lush *et al*., 2018). These are often used to isolate the movements of the animal and to calculate energy expenditure (Gleiss *et al*., 2011) and they are

particularly useful in distinguishing between behaviours with low and high movement (Diosdado *et al*., 2015).

To calculate ODBA and VeDBA the static component (ST) of acceleration and overall movement of the animal (dynamic body acceleration; DBA) are needed. To calculate ST, a running mean of fixed window size (e.g. 5 s) is used to smooth each acceleration axis (X, Y and Z). Next, DBA is calculated by subtracting ST from the raw acceleration data. From this, ODBA and VeDBA are calculated as:

$$ODBA = DBA_X + DBA_Y + DBA_Z \quad (4)$$

$$VeDBA = \sqrt{DBA_X^2 + DBA_Y^2 + DBA_Z^2} \quad (5)$$

Other features are sometimes reported for acceleration data and a compilation of research and their selected acceleration features can be seen in the publication by Pires *et al*. (2017). Once models are evaluated on the training set, they are then applied to a test set of previously unseen data. The test set may indicate the performance of the model in future classification or prediction tasks, but performance can be degraded if the test set contains instances that have not been previously seen in training (Valletta *et al*., 2017).

The rationale behind ML is that models can be improved over time as more data representative of the problem space is accrued. Models can then be measured on their performance using a variety of measures (Section 2.3.1). In part, the performance of any model may be influenced by the means that the raw data was sampled (e.g. automatic, manual), its overall quality (e.g. sampling frequency, precision of instrument) and the diversity of the extracted feature set (Valletta *et al*., 2017). When sensors such as GPS are used for raw data collection, feature diversity is usually greater, in part because feature-sets for acceleration data are well developed and tested

and also because GPS data signatures are fundamentally different to that of acceleration data. Examples of feature sets extracted from GPS data are documented by Godsk and Kjærgaard (2011) and Williams *et al*. (2016). The speed between consecutive GPS fixes for example can be exploited to extract some of the basic features analogous to that extracted from acceleration data such as mean, maximum and minimum speed. A further set of features unique to GPS can be extracted from positional coordinates such as the angularity of movement and the directional heading of the focal animal.

This is not to say that a diverse range of features cannot be extracted from acceleration data. Indeed, there is quite a variation in the number of features reported in some publications, and some have used 20 or more (Martiskainen *et al*., 2009; Resheff *et al*., 2014). The goal, however, should be to minimise the complexity of the resultant model to reduce the computational time required to process raw data in future. This can be achieved by reducing the size of the feature set so that the learning algorithm only considers features that carry the most important statistical properties for classification or prediction. This process is known as feature selection or dimensionality reduction and has been reviewed by Sorzano *et al*. (2014).

5.3 Supervised and unsupervised machine learning

ML tasks can be classified as either supervised or unsupervised. Supervised learning require that ML algorithms are provided with a set of training instances labelled with the corresponding outcome variable or class label. The algorithm then maps the dependence of each class label onto the feature set (explanatory variables) and models are learnt in such a way as to minimise the error over the entire dataset (Liakos *et al*., 2018). Trained classifiers are then deployed on a set of test instances to predict the corresponding class labels before evaluating the performance of the model.

In unsupervised learning tasks, the data provided to the ML algorithm is unlabelled and as such, the goal is to search for commonalities in the data and group each new instance based on whether a particular commonality is present or not (Valletta *et al*., 2017). Unsupervised learning can lead to the discovery of hidden patterns and there are two primary techniques, namely clustering and association; the latter used to search for rules that explain large portions of the dataset. One example is the use of association rule mining in linking cattle disease phenotypes to the symptoms for enhanced disease identification (Bhavsar and Arolkar, 2014). Clustering on the other hand has been used previously to identify distinct cattle behaviours (Fuzzy-C-means; a commonly used clustering algorithm in pattern recognition originally developed by Dunn (1973) and reviewed by Nayak *et al*. (2015)) on a feature set derived from neck-mounted accelerometers (Dutta *et al*., 2015). Using the distinct behaviour groupings from the unsupervised learning stage, the authors then proceeded to try and classify these groupings using a supervised ML technique, thus leveraging the unsupervised cluster analysis to reduce the number of redundant behaviour categories.

Kumar *et al*. (2018) used a deep convolutional neural network (a class of neural network algorithms) to extract and learn a set of discriminatory features from the muzzle images of cattle. This feature set was then used to classify individual cattle and the objective of such a system is to reduce the risk of cattle misidentification.

The experimental work herein used only supervised ML techniques. Despite, the limited use of unsupervised methods in the dairy sciences, particularly for cattle behaviour identification, some examples are emerging. Recent examples include the use of a clustering algorithm; *k*-means (Hartigan and Wang, 1979) to group segmented acceleration data (neck-mounted accelerometers) for the classification of heat events (Shahriar *et al*., 2016) and behaviour of housed dairy cows (Diosdado *et al*., 2015). The

next section will discuss a sample of ML algorithms, most of which are used in supervised classification tasks.

5.4 Machine learning algorithms

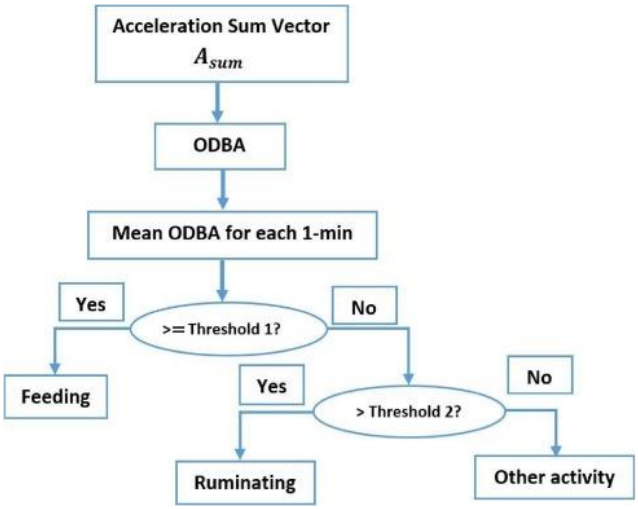
ML algorithms can be grouped by their learning style and two examples have already been discussed, namely supervised and unsupervised (Section 5.3). Other learning styles include semi-supervised learning where the input data contains a mixture of labelled and unlabelled data (Zhu and Goldberg, 2009) and reinforcement learning where an ML algorithm makes decisions based on feedback from the environment such as to maximise the reward in a particular situation (Sutton and Barto, 1998). Algorithms can also be grouped by their functional method and the discussion herein will relate to each algorithm in this manner. Data mining software packages such as WEKA (Section 5.5) primarily group algorithms according to their functional attributes but it is worth noting that some algorithms could in fact fall into more than one group depending on definition. The majority of the algorithms described in the following sections were used in Chapters 1 and 3 and despite small descriptions of each being present in those chapters, further mechanistic detail is provided here. What will not be discussed in the following sections are algorithms that are engineered for purpose by analysts. For example, simple decision tree algorithms can be created using user input. This can be done using biological intuition or by animal experimentation to find optimal thresholds in sensor data that may distinguish between the behaviours of interest. In essence, an algorithm for behavioural classification could be engineered as a next step following the work by Rayas-Amor *et al*. (2017) where grazing and ruminating were distinguished using specific acceleration thresholds. An example of this process was undertaken by Diosdado *et al*. (2015) who configured optimal thresholds for a decision tree classifier based on the optimal true-positive and false-positive values for specific behaviours across a range of acceleration values. Although

time-consuming, these methods are relatively simple and perhaps less prone to the complexities that can be introduced by some ML procedures.

Finally, it is worth noting that a diverse range of supervised ML algorithms exist, and a single section cannot provide an extensive and comprehensive insight into all. Where appropriate, algorithms are discussed in the context of their use in dairy science research.

5.4.1 Trees

Decision tree algorithms classify instances in a dataset by sorting them based on attribute (feature) values. A decision tree consists of nodes (root node and internal nodes), branches (connecting nodes) and leaves (leaf node). Nodes represent the attributes of the dataset and usually contain a test that compares a particular feature value with a constant, for example $Height \begin{matrix} \leq 1.8\text{m} \\ > 1.8\text{m} \end{matrix}$ where Height is the attribute representing a node and the two categories of choice represent the branches. Decision trees may contain multiple nodes that branch either to a subsequent node or to a leaf containing the class value. The classification of new instances involves movement down through the tree, beginning at the root node and selecting the appropriate pathway at each node until the instance reaches a leaf. The instance is then classified according to the label associated with that leaf (Figure 9).



Adapted from Benaissa *et al*. (2019)

Figure 9. Decision tree utilising only the feature overall dynamic body acceleration (ODBA) and two thresholds to distinguish between feeding, ruminating and other activities of cattle. ODBA = root node, ovals = internal nodes, behaviours (classes) = leaves. Internal nodes and leaves are connected by branches.

5.4.1.1 C4.5 (J48)

One of the most popular algorithms used to build decision trees is the C4.5 algorithm (Quinlan, 2014). J48 is the WEKA implementation of the C4.5 algorithm. When building a decision tree, the algorithm recursively partitions the data into subsets and measures the information gain resulting from choosing that particular attribute and thus measuring its importance in the dataset. The attribute with the highest information gain is used in each step before the partitioning process stops if resulting nodes contain single classes or if no further information gain is possible from the remaining attributes. Viazzi *et al*. (2013) used C4.5 to construct a decision tree classifier from video images of dairy cattle movement. Cows were categorised into three classes (not lame, lame and severely lame) and the decision tree had an overall CA of 76%. A similar decision tree was developed by Steensels *et al*. (2016) to classify dairy cattle as either sick or healthy using production parameters (e.g. milk yield and slope of lactation curve) and behaviour (e.g. activity and rumination). Their classifier achieved a CA of 78%.

5.4.1.2 Naïve Bayes tree

Naïve Bayes tree (NBTree) is a hybrid algorithm that combines decision trees and naïve Bayes classifiers (Section 5.4.3). The decision tree element of the algorithm segments the dataset and a naïve Bayes classifier is then deployed onto each segment (represented by leaves). The decision tree algorithm of NBTree is very similar to that of J48 (Kohavi, 1996).

5.4.1.3 Logistic model tree

Logistic model tree combines tree induction methods for classification with logistic regression. Logistic regression models are produced at every node in the tree using the LogitBoost algorithm (Friedman *et al*., 2000). Nodes are then split using the C4.5 algorithm. Cross-validation (Chapter 1: *Classifier Validation Strategy*) is used to

find the optimal number of LogitBoost iterations that do not overfit the training data before the tree is pruned (Landwehr *et al*., 2005).

5.4.2 Rules

Rule-based algorithms share a commonality with decision trees in that the resulting classifiers are logical in their interpretation. A set of rules can in fact be generated from a decision tree by following a particular path from the root node to a leaf in the tree. For example, taking the decision tree in Figure 9 (Benaissa *et al*., 2019) a set of rules can be generated to make the classifications:

IF mean ODBA ≤ Threshold 1 THEN Behaviour = Feeding; ELSE

IF mean ODBA < Threshold 2 THEN Behaviour = Ruminating; ELSE

Behaviour = Other activity

Rule-based algorithms however generate rules directly. As with nodes in a decision tree, rule sets may contain a series of tests (antecedents) based on different attributes of the dataset. These tests lead to a conclusion which is usually the class that is covered by any particular instance. Antecedents within rules can also be connected with AND (e.g. IF $a>3$ AND $b<4$ THEN x), and all tests within a particular rule must be met if the rule is to work for any particular instance. The aim is to construct the smallest set of rules that explain the assumptions behind the training data.

5.4.2.1 RIPPER (JRip)

One example of a rule-based algorithm implemented in WEKA is JRip which implements repeated incremental pruning to produce error reduction (RIPPER; Fürnkranz and Widmer, 1994); a method that produces rules quickly and efficiently. JRip learns propositional rules which it repeatedly grows and prunes. Rules are grown by adding conditions until the rule is perfect. Every possible value of each attribute is

assessed before selecting the condition with the greatest information gain. The conditions are pruned in the next phase according to a pruning metric. Finally, the ruleset is optimized where rules performing poorly on randomized data are deleted (Cohen, 1995). JRip was used in Chapter 1 (Williams *et al*., 2016) for modelling dairy cow behaviour. Nine rules were generated from the feature set (extracted from GPS data) and the rule set had an overall CA of 85% and F-measure of 76% for the behaviours grazing, resting and walking.

5.4.2.2 PART

PART (partial decision tree algorithm) uses the separate and conquer strategy of the JRip algorithm to generate a set of rules. Instances are then removed recursively from the training set that are covered by each rule, proceeding until no instances remain. Rules are generated using the decision tree approach of J48 (C4.5), using the leaf with the greatest amount of coverage as the new rule before the tree is discarded (Frank and Witten, 1998).

5.4.2.3 OneR

OneR (One rule) generates classification rules from a set of instances using a one-level decision tree, testing one particular attribute. Each value of every attribute generates a different set of rules. The algorithm works on the principle that it classes each branch with the class that occurs most frequently in the training data. The rule set for each attribute is then evaluated for its error rate by counting the number of instances that do not have the majority class (Holte, 1993).

Both PART and OneR were evaluated in Chapter 3 (Williams *et al*., 2019) as base learners for the classification of dairy cow behaviour from GPS data. However, to the candidate's knowledge, very few examples of the use of rule-based algorithms for classification tasks exist in the agricultural literature.

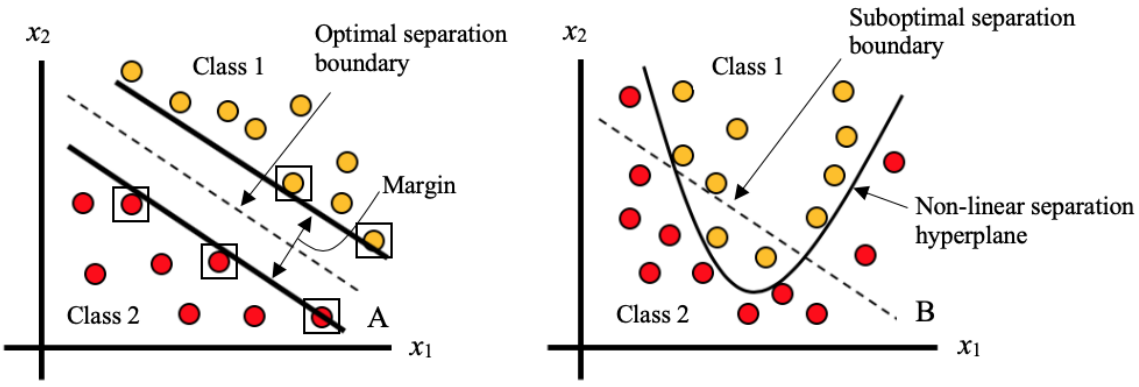
5.4.3 Bayes

Generally, Bayesian algorithms estimate the conditional probability distribution of attributes in a training set before assigning instances in the test set to the class with the highest posterior probability using Bayes' theorem. The naïve Bayes algorithm is the most commonly used algorithm that sits within this grouping and it assumes that all attributes are unrelated to each other and that no hidden attributes influence the prediction process (John and Langley, 1995). Another example is the Bayesian network algorithm (Bayes net). This algorithm classifies instances by computing the conditional probabilities of attributes (nodes) given values assigned to other nodes (Sebastiani *et al*., 2005). A Bayes net classifier can be described as a directed acyclic graph with nodes connected by arcs that represent the probabilistic dependencies between the nodes (Friedman *et al*., 1997).

Both naïve Bayes and Bayes net were used in a comprehensive evaluation alongside a number of algorithms in classifying the health outcomes of intensively grown calves after clinical diagnosis with bovine respiratory disease (Amrine *et al*., 2014). Over one-million animals were included in the dataset and over 100 variables were created for the feature set with multiple datasets created for evaluation. A discussion of the results relative to both classifiers is outside the scope of this section but those interested are referred to Amrine *et al.* (2014). Naïve Bayes was also used by Dutta *et al*. (2015) to classify dairy cow behaviour from neck-mounted accelerometer data, achieving very good results on average across five behaviour categories (CA = 90%; F-measure = 73%). Bayes net was evaluated for its use in predicting conception success in dairy cows with quite poor performance (AUC = 0.56) and was only marginally outperformed by a logistic regression model (AUC = 0.57) (Shahinfar *et al*., 2014).

5.4.4 Support vector machine

Support vector machine (SVM) can be applied to both classification and regression tasks (Vapnik, 1999). The aim is to partition the data into discrete classes (Figure 10A) using a hyperplane that maximises the margin between the data points and minimises the number of misclassifications on the training data (Vapnik, 1999). For non-linear tasks (Figure 10B), the input data can be transformed into a high dimensional feature space using a kernel function. This is used to solve inseparability problems associated with linear decision boundaries. A number of kernels can be used with SVM and examples include the polynomial kernel and radial basis function (RBF) kernel.



Adapted from Wang (2005)

Figure 10. Examples of linear (A) and non-linear support vector machine (SVM) (B) on binary classes. In A, the two classes are separated by a linear boundary (dashed) located midway between the two areas (black bars) that maximise the margin between the two classes. Support vectors (indicated by the instances surrounded by boxes) have a direct bearing on the location of the separation boundary. In B, a linear boundary (dashed) does not separate the classes without error. Two instances from Class 1 and three instances from Class 2 are misclassified. The classes are however separated with a quadratic curve. Generally, instead of a non-linear curve, an SVM will create a higher-dimensional feature space from the input data and attempt to solve the classification problem using a linear solution.

An SVM was used by Martiskainen *et al*. (2009) to classify a number of dairy cow behaviours which included standing, lying and feeding using data collected from neck-mounted accelerometers. Using a fixed window length of 10 s, 28 features were extracted from the acceleration data, including mean, standard deviation, skewness and kurtosis. The CA across all recorded behaviours using the SVM was >80%. Benaissa *et al*. (2019) evaluated the performance of an SVM in classifying feeding and ruminating behaviours from data gathered from a collar-mounted accelerometer and also from the halter-based RumiWatch system (Figure 5B). The data sampled at 10 Hz were partitioned into 1-min segments for the extraction of eight features. Compared to the RumiWatch system (CA = 91%), the SVM applied to the collar-mounted accelerometer output achieved an overall CA of 93%.

5.4.5 Artificial neural networks

Artificial neural networks (ANNs) are a group of algorithms that are loosely modelled on the human brain and are capable of ML as well as pattern recognition (Krogh, 2008). Architecturally an ANN is composed of three layers; an input and output layer of interconnecting nodes with one or more hidden layers (Figure 11). ANNs are used increasingly in the literature for complex tasks such as image and character recognition, and in the dairy sector in particular ANNs are being used for the automatic recognition of individual cows from images (Kumar *et al*., 2018; Zhao *et al*., 2019). Models are highly tuneable and can therefore become highly complex with many hidden layers and interconnecting pathways between hidden neurons.

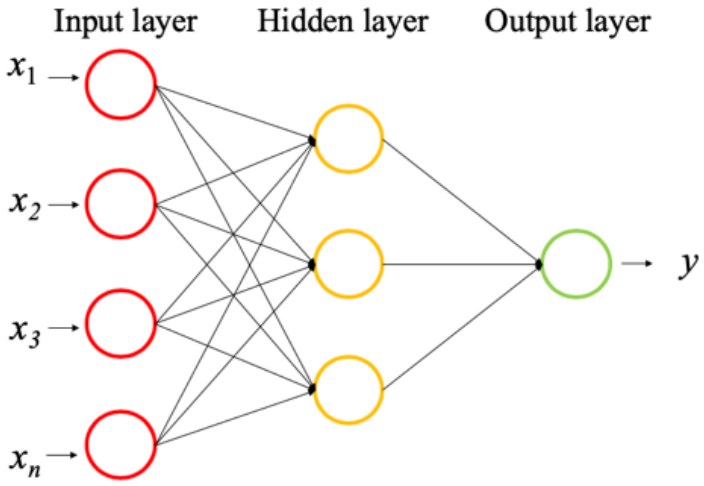


Figure 11. A feed-forward neural network architecture with an input layer, hidden layer and output layer. The hidden neurons process and capture non-linearity in the input variables passed to them from the input nodes. This information is then transferred to the output layer containing the modelled dependent variable(s). The hidden layer(s) of neurons in this architecture is required for dealing with non-linearly separable classes and thus is characteristic of the multilayer perceptron algorithm.

Although an extremely powerful ML process, complex ANNs can lead to data overfitting (Kim *et al*., 2005). One such ANN algorithm is the perceptron (Rosenblatt, 1958) but a number of algorithms are available. Multilayer perceptron (MP) (Figure 11 and Chapter 3) uses a self-correcting back propagation method for training which calculates the total error of the output nodes before adjusting the weights of the network to reduce the error in the output layer (Bishop, 1995). MP can be used for both linearly and non-linearly separable classes (differs to the simple perceptron which can only be used for linearly separable classes; Rosenblatt, 1958) using additional layers of neurons between the input and output layers (Basheer and Hajmeer, 2000). Despite taking the longest to train, MP performed very well at classifying dairy cow behaviour (Chapter 3) compared to other classifiers.

A further representation of ANNs are deep ANNs, often referred to as deep neural networks (DNNs). DNNs are very similar to ANNs and differ only in that they are composed of multiple hidden layers between the input and output layers of the network (Goodfellow *et al*., 2016). The convolutional neural network (CNN) is a DNN model that is being used more frequently in particular for image classification in the dairy sciences (e.g. Alvarez *et al*., 2018) owing to its ability to extract features from the dataset (image in this case) and map these back to the classes of interest (Hijazi *et al*., 2015). In contrast to ANNs, CNNs can learn complex problems quickly and if datasets are large enough, can result in improvements in CA (Schmidhuber, 2015). It is likely that CNNs will be used more frequently in future to assist with complex image processing tasks in agriculture and livestock science. An excellent review on the use of CNNs in agriculture is provided by Kamilaris and Prenafeta-Boldú (2018).

5.4.6 Ensembles

The use of ensembles of classifiers is becoming more popular in the agricultural sciences as in principle, the methodology can lead to improvements in prediction power over individual classifiers (Dietterich, 2000). In an ensemble, the individual decisions of all classifiers are combined in some way to classify new examples (Figure 12). Combining classifiers in this way is an attempt at reducing the probability of misclassifications that might occur from any one single model thus increasing the area of expertise in the system (Kuncheva and Whitaker, 2003). The more diverse (classifiers that make errors on different instances in the dataset) and accurate (classifier error rate is lower than if the classes were randomly assigned) the classifiers are in an ensemble, typically, the better the performance of the ensemble (Kuncheva and Whitaker, 2003). Several ensemble methods exist but the most commonly used are bagging (bootstrap aggregation), boosting and stacking (stacked generalisation).

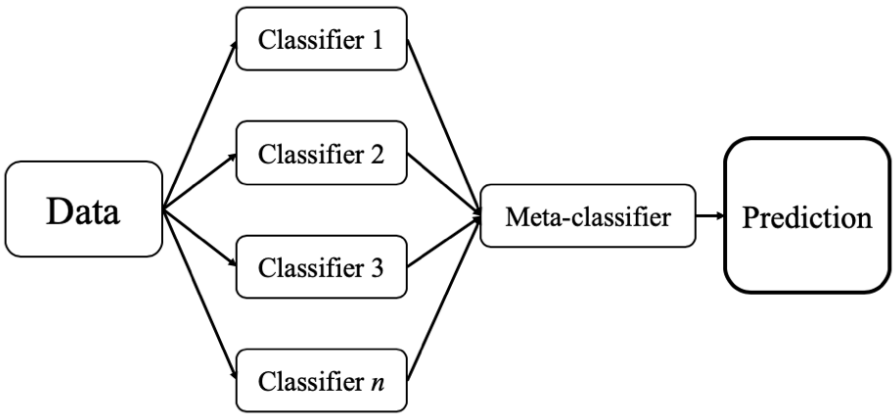


Figure 12. An ensemble classification framework where the collective decisions of individual classifiers learned from a training dataset are forwarded to a meta-level learner for modelling. The meta-classifier then makes the final class predictions.

5.4.6.1 Bagging

A bagging classifier fits base classifiers on randomly drawn subsets (instances) of the original data before aggregating their individual predictions to form a final prediction by either voting or averaging (Quinlan, 1996). Each base classifier is trained on a training set that is randomly drawn from the original training dataset before being replaced into the dataset (bootstrap replicates). The act of replacement means that some instances may be re-drawn and omitted from the training data. Each bootstrap replicate contains on average 63% of the original training set (Dietterich, 2000). Every run though the dataset results in a single classifier and the final classification of instances is derived by a majority vote (results are aggregated) of all of the predictions of the previously learned classifiers (Breiman, 1996). The predictability of bagging is improved by exploiting the instability of the learners used and so using stable learners such as naïve Bayes will not result in improvements in predictability over unstable learning methods such as tree learners (Breiman, 1996). Bagging with a tree learner was used by Dutta *et al*. (2015) and the classifier achieved an average CA of 96% in the classification of several dairy cow behaviours extracted from GPS and accelerometer data. In predicting insemination outcomes in Holstein cattle, bagging with tree learner was outperformed (AUC = 0.68) by a random forest classifier (AUC = 0.76) (Shahinfar *et al*., 2014). Random forest is in fact a type of bagging ensemble algorithm in itself (Breiman, 2001).

5.4.6.2 Boosting

Boosting is the act of iteratively learning weak classifiers that generally have performances that are only slightly better than random (Schapire, 2003). The addition of a weak learner to the ensemble leads to the re-weighting of instances in the dataset. More weight is added to instances that are misclassified so that new learners focus on

these erroneous instances (Schapire, 2003). This is repeated until the required number of iterations is complete or until the total weight of the misclassified instances exceeds more than 0.5 (Schapire, 1999). The final prediction is made by taking a vote based on the predictions of each classifier weighted for individual performance on each training set (Dietterich, 2000). Despite suiting a wide range of ML algorithms, boosting is particularly suited to decision trees. Few examples of the use of boosting algorithms exist in dairy research, but Dutta *et al*. (2015) used two variants (linear discriminant and tree learner) of the AdaBoost algorithm to classify cow behaviour that had marginally poorer performance compared to the other ensembles used in that study. They hypothesised that because boosting algorithms are prone to overfitting due to their training regimes (more emphasis on misclassified instances), they are susceptible to misclassifications on new instances.

5.4.6.3 Stacking

Stacking is the procedure of integrating a set of independently computed base classifiers into a higher-level meta-classifier with the aim of improving learning efficiency with collective decisions (Wolpert, 1992). The meta-classifier uses the predictions of the base classifiers as a set of attributes for predicting the class values (Van der Laan *et al*., 2007). Logistic regression and random forest algorithms often feature as meta-level learners (Healey *et al*., 2018; Arsov *et al*., 2019) and logistic regression in particular does not tend to overfit the data (Ting and Witten, 1999). Theoretically, the ensemble should perform at least as well as the best performing classifier nested within the group of base learners (Van der Laan *et al*., 2007).

Choosing the most suitable combination of base learners for an ensemble is a particular area of interest in the literature (Caruana *et al*., 2004; Tsoumakas *et al*., 2008). One of the most basic techniques is forward stepwise selection. Base learners are

sequentially added to an empty ensemble to maximise the performance of the ensemble based on an error metric such as the root mean squared error (RMSE). Other computational methods exist too and packages such as 'SuperLearner' (Polley *et al*., 2018) available for R can allow the user to fit multiple models at a time for performance assessment and fast optimisation of the ensemble using cross-validated risk (based on mean squared error). Options also exist for model customisation using various hyperparameters.

Given the recent widespread use of base classifiers in livestock behaviour classification, it was decided (Chapter 3) that a stacking ensemble should be tested to see whether improvements in performance could be achieved given the additional computational power required and also given the theory of their function. Other than Williams *et al*. (2019) (Chapter 3), the candidate is not aware of other literature that has assessed the performance of stacking ensembles in the dairy sciences, although Dutta *et al*. (2015) tested a number of other ensemble classifiers. Given the interest to date in the use of other ensemble methods, it is envisaged that stacking ensembles will feature in future behavioural classification and disease identification tasks. This will only hold as long as classifiers are computationally efficient and provide additional predictive power over standard base classifiers (Chapter 3).

5.5 Waikato Environment for Knowledge Analysis (WEKA) for machine learning

5.5.1 Background

The WEKA data mining workbench provides a range of ML tools and techniques across several graphical user interfaces (GUI). The majority of ML tasks can be undertaken using the Explorer interface. Feature sets can be visualised and analysed for summary statistics, and regression, classification and clustering tasks undertaken using a range of ML algorithms (e.g. Section 5.4). Furthermore, analysts can apply both supervised and unsupervised filters to datasets for selecting attributes, balancing classes and resampling datasets amongst many other features.

For classification tasks, multiple ML algorithms can be tested at once across a number of datasets using the Experimenter interface, and the results evaluated across several measures of performance, some of which were discussed in Section 2.3.1. The advanced option of the Experimenter interface allows users to configure its functionality for output such as per class statistics or test algorithms across multiple parameter values in succession. In both the Explorer and Experimenter, users can model data using either *k*-fold cross-validation or choose to split data into training and test sets (e.g. 66% training; 34% testing). In Chapters 1 and 3, cross-validation was used to learn the classifiers and the procedure is described in Chapter 1: *Classifier Validation Strategy*.

5.5.2 Classifier evaluation

The classification results from both the simple and advanced Experimenter interfaces can be statistically analysed using variations of the *t*-test for several evaluation metrics (e.g. comparing the CA of classifiers). However, the most appropriate test for comparing the results of multiple classifiers is a strongly discussed topic in the ML literature because of the risk of not meeting the assumptions of the *t*-

test and the risk of Type-1 errors in multiple comparisons (Demšar, 2006; Garcia and Herrera, 2008; García *et al*., 2010). As a result, some analysts turn to other statistical programs for analysing Experimenter results for tests that do not violate these assumptions. For these reasons, the performance of multiple classifiers in Chapters 1-3 was undertaken using statistical tests other than those provided by the Experimenter interface where necessary.

5.5.3 Waikato Environment for Knowledge Analysis (WEKA) for processing big data

Although not used for the work undertaken herein, two other notable functions of WEKA include the Knowledge Flow interface (KFI) and the Command Line interface (CLI). The KFI allows analysts to visualise the ML process by selecting a series of components from a tool bar and connecting them into a directed graph. It can also be used for multiple data processing streams that cannot be undertaken in the Explorer. The CLI allows users to execute all WEKA functions directly using commands, reducing the time taken for analyses. Furthermore, the size of the dataset that can be processed through the CLI (and KFI) is arbitrary whereas the Explorer is limited to the amount of computer memory available. This is particularly important when processing data that may fall within the domain of 'big data' (Section 5.1). Also, for ML tasks that demand the processing of very large datasets, wrappers for cluster-computing frameworks such as Apache Spark (a form of distributed computing) can be used. The package DistributedWekaSpark provides a platform for using Spark through WEKA's GUI which is particularly beneficial if the analyst is accustomed to using the WEKA software package. For a comprehensive discussion on the use of this integrated framework, see Koliopoulos *et al*. (2015). For a general discussion on the use of several platforms for big data processing, see Wu *et al*. (2014) and for a discussion on the use

of software for processing large, evolving data streams (Massive Online Analysis (MOA)) that can be implemented through WEKA, see Bifet *et al*. (2010).

6. Summary

6.1 Overview of PLF

The availability of large quantities of data from current conceptual sensor systems will undoubtedly impact the way in which livestock are managed in the future. At a relatively simple level, this is already occurring and much of it in the dairy industry. Currently, the majority of data driven tasks are occurring in areas that have had significant research attention over the last few decades and two of the major areas are fertility and mastitis management. These two areas also have a significant economic importance on farm. Increasingly however, more attention is being paid to lameness as more is understood about the complex and costly nature of lameness-causing diseases from the perspectives of production and welfare (Sadiq *et al*., 2017).

Other future data-driven aspects that can be expected to be common on dairy farms in the next 5-10 years will include sensor systems for individual feed, comfort and metabolic status management. This is because these variables can be measured using either existing hardware (addition of new algorithms) that farmers have already adopted or by retrofitting additional sensors to existing components. Based on the evidence, farmers that have already adopted certain technologies are likely to be more open to the adoption of additional technologies and so this leap will be relatively short for some. Predicting which sectors and systems make these adoptions however is difficult as farm economics, farmer age and the desire to automate are just some of the variables that will contribute to these decisions. What is clear is that investment in some sensor systems makes important economic sense to some producers and one of the best examples in the dairy sector is the use of animal-based sensors for automatic heat detection. Gauging the main drivers for adoption of PLF sensor systems in the UK dairy industry is needed, but some inference can probably be made from recent work

such as Lima *et al*. (2018) and Caja *et al*. (2016). Furthermore, the definition of, and what exactly constitutes PLF can sometimes be unclear in the literature, and this could be important when it comes to exploring some of the key drivers for adoption and studying farm resource use and efficiency in future. As found however, the adoption of any technology hinges on its ease of use and reliability.

6.2 Two main categories of PLF tools

PLF is a broad term, and as the literature suggests, covers a wide variety of applications. Wathes *et al*. (2008) define PLF as 'the management of livestock production using the principles and technology of process engineering' and continue by saying that 'it is the principal means by which 'smart' sensors will be used in livestock farming.' Smart sensors are of course integrated into a range of technologies. For example, AMS use sensors to locate teats prior to cluster attachment and to clean teats post-milking and are precise in that sense. There are other good reasons why AMS fall within the PLF paradigm. These systems often come integrated with sensors that measure milk variables or the weight of cows for ongoing precision management. Moving forward, it will be important to clearly define what is meant by PLF and the benefits of PLF will need to be communicated to farmers clearly for particular applications. It seems that some of the technologies that fall within the definition of PLF are first and foremost labour-saving technologies and this is how farmers probably perceive them. As documented by Steeneveld and Hogeveen (2015), farmers with an AMS made no conscious decisions to invest in the accompanying sensor system, which probably indicates that the main objective was to automate the milking process. PLF does encompass two main categories of sensor system however, and these can be categorised as on and off-animal sensor systems (Caja *et al*., 2016). The off-animal PLF technologies tend to be those that do something that could otherwise be done by

human means (e.g. AMS). On-animal sensors usually record something about the focal animal that probably could not have been recorded otherwise (e.g. rumination). However, it is easy to see how some of these sensor systems could fall under both categories. Table 4 gives examples of some of the sensor systems that sit within these two categories and are currently used in the dairy industry.

Table 4. Examples of off-animal and on-animal sensor systems currently in use in the dairy industry

Off-animal sensor system	On-animal sensor system
Automatic milking system	Oestrus sensor (pedometer / accelerometer)
Sorting gates	Rumination sensor (accelerometer / sound)
Gait analysis (video or image analysis)	Lameness detection (accelerometer)
Weighing scales (crush or race-fitted)	Temperature (bolus)
Milk electrical conductivity	Feeding sensor (noseband pressure sensor / accelerometer)
Feed dispensing system (forage and concentrate)	Calving sensor (accelerometer)
Forage pushing system	Rumen pH sensor (bolus)
Body condition score analyser	

6.3 Future PLF

GPS receivers fitted to pasture-based dairy cattle in future are in the majority of cases likely to be coupled with other discrete sensors that can add dimensionality to the data and decision processes. GPS in particular will likely be used in pasture-based systems given their limitations for use indoors although others have evaluated satellite-based devices that can to some extent be used indoors (Huhtala *et al*., 2007). As well as this, it is clear from the literature that other discrete sensors such as accelerometers are more sensitive to small perturbations in the position (e.g. head up vs. head down) of focal animals which can be highly valuable for determining the behaviour taking place. In addition to some of these other movement sensors that were not explored in Chapters

1-3, a new generation of biosensors are likely to feature in research activities and on farms in future (Neethirajan *et al.*, 2017). Biosensors will allow animal practitioners to measure physiological and immunological variables of livestock as well as measure how these parameters impact the behaviour of livestock. This will provide richer information, especially to veterinarians who often rely on subjective means of diagnostics and low-level herd data. But these sensors will provide benefits for farmers too, as long as sensors are integrated with management support and can provide straightforward answers to straightforward questions (Caja *et al.*, 2016).

The following chapters will discuss the use of GPS as a tool for monitoring the behaviour of dairy cows at pasture. Despite the literature having developed rapidly over the duration of this candidature, covering a breadth of sensor types for various applications, the principles and techniques employed in the following chapters are applicable to many of these developments.

6.4 Aims and objectives of this thesis

The aims and objectives of the following series of experiments were to model the behaviour of dairy cows at pasture to gain a better understanding of how some of the most notable behaviours are represented in high-frequency GPS data. Given the extensive use of GPS in the movement ecology literature it was necessary to test this technology with livestock. It was envisaged that novel insights would be gained as to the behavioural ecology of cattle and whether such behaviour models were likely to yield useful information for future long-term studies of welfare and performance. Given the establishment of such study principles (Godsk and Kjærgaard, 2011), and the importance of grazing livestock in the context of Welsh agriculture, this thesis was an attempt at providing a detailed methodology for the classification of pasture-based

cattle and to begin to learn more about specific attributes of cow behaviour that had not been previously published.

Herein, Chapter 1 discusses the development of a behavioural model of pasture-based dairy cows developed in an experimental grazing system designed to replicate a well-managed strip-grazing regime. The data used for modelling cow behaviours were gathered using high frequency GPS data and subsequently partitioned into fixed-time segments for classification by a series of ML algorithms. Chapter 2 then considers the application of an error-correcting technique for the predicted classes of the model developed in Chapter 1. This was undertaken using a hidden Markov model. Finally, Chapter 3 evaluates the use of a variable segmentation technique as an alternative to data segmentation at fixed intervals. In this chapter, ensemble classifiers were used as opposed to classifiers built using conventional ML algorithms. Ensembles were chosen due to limited exploration of these algorithms in the classification of dairy cow behaviour.

Chapter 1

The following chapter was accepted for publication in the Journal of Dairy Science and is therefore formatted to the requirements of the journal. The reference for the publication is:

Williams, M.L., Mac Parthaláin, N., Brewer, P., James, W.P.J. and Rose, M.T., 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *Journal of Dairy Science*, *99* (3), pp. 2063-2075.

Summary

Recording the minute-by-minute behaviour of dairy cows automatically will likely be a common feature of future dairy management. The benefits of automatic record keeping of cow behaviour are already being realised on some farms and has shown to be a crucial component of certain tasks such as fertility management. Sensors are being continually researched for their efficacy in predicting behaviours that are both economically important to farmers and that could be important for monitoring the welfare of dairy cows. For example, a discrete accelerometer sensor attached to cows can be used to automatically identify lying and standing behaviours. These records can then be used to monitor cow comfort on a continual basis. To try and contribute to this area of research in pasture-based systems, GPS data were gathered from a cohort of dairy cows at Trawsgoed dairy farm, Aberystwyth University for computational analysis. Machine learning is becoming an increasingly important branch of data analysis in many fields such as security for facial recognition tasks and in agriculture for automatic crop recognition. Machine learning techniques were applied to the GPS data for the automatic classification of the three most prevalent behaviours of cows; grazing resting and walking. The results showed that machine learning could be used

with good success for automatic behaviour identification from GPS data. Since this publication, others have shown how a combination of sensors applied to dairy cows can be used to gain more insight into daily behaviour. Far more is being learnt about the nature of cow behaviour using these techniques. It is very likely that many dairy farmers in future will be monitoring daily health and performance of cows using sensors and applications developed using similar methods.

Interpretive Summary

Williams

Identifying changes in the behavior of dairy cows indicative of disease is subjective and often difficult. Data from small, low-cost global positioning system (**GPS**) receivers mounted on cow-collars were used to predict the behaviors of pasture-based dairy cows. Here, we demonstrate the application of machine learning techniques and evaluation methods to rigorously test the performance of the predictive classification models derived from the raw GPS data. The most suitable model performed very well on independent test data (average classification accuracy = 0.85) for the behaviors grazing, resting and walking. This model will be used to study behavioral responses to disease in dairy cows to aid earlier disease identification.

A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques

M. L. Williams,[*] N. Mac Parthaláin,[†] P. Brewer,[‡] W. P. J. James,[*] M. T. Rose,[*1]

[*]Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom

[†]Department of Computer Science, Institute of Maths, Physics and Computer Science (IMPACS), Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DB, United Kingdom

[‡]Department of Geography and Earth Sciences, Aberystwyth University, Penglais Campus Ceredigion, SY23 3DB, United Kingdom

[1]Corresponding author: Dr Michael Rose. Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom. E-mail: mir@aber.ac.uk

ABSTRACT

A better understanding of the behavior of individual grazing dairy cattle will assist in improving productivity and welfare. Global positioning systems (GPS) applied to cows could provide a means of monitoring grazing herds while overcoming the substantial efforts required for manual observation. Any model of behavioral prediction using GPS needs to be accurate and robust by accounting for inter-cow variation as well as atmospheric effects. We evaluated the performance using a series of machine learning algorithms on GPS data collected from 40 pasture-based dairy cows over 4 mo. A feature extraction step was performed on the collected raw GPS data, which

resulted in 43 different attributes. The evaluated behaviors were grazing, resting and walking. Classifier learners were built using 10 times 10-fold cross validation and tested on an independent test set. Results were evaluated using a variety of statistical significance tests across all parameters. We found that final model selection depended upon level of performance and model complexity. The classifier learner deemed most suitable for this particular problem was JRip, a rule-based learner (classification accuracy = 0.85; false positive rate = 0.10; F-measure = 0.76; area under the receiver operating curve = 0.87). This model will be used in further studies to assess the behavior and welfare of pasture-based dairy cows.

Key words: GPS, grazing, behavior model, data mining

INTRODUCTION

It is well documented that the behavior of dairy cattle coincides with changes in health status. For example, Huzzey et al. (2007) used electronic feed bins to record feeding behavior in housed dairy cows. They found that feed intake and time spent feeding began to decrease 2 wk prior to clinical diagnosis of severe metritis. González et al. (2008) found that daily feeding time, number of visits to the feed bin, and feeding rate began to decrease as early as 30 d before lameness diagnosis in housed cows fed a silage ration. With larger herds and limited time, disease diagnosis becomes more difficult. Mobility scoring is one example of a subjective technique used by herdsmen to identify lameness and locomotor problems in dairy cows. Although cheap to disseminate, mobility scoring is time consuming and must be done regularly (Pluk et al., 2012; Van Nuffel et al., 2015). Another criticism of the technique is that it may fail to identify the early (sub-clinical) stages of lameness (Manske et al., 2002; Dyer et al., 2007). Reader et al. (2011) in their study reported that the milk yield of cows decreased by an average of 0.7 kg/day for approximately 7 wk before cows became visually lame. Furthermore, after recovery, the milk yield of lame cows remained lower for 4 wk. Identifying production disease as early as possible is therefore imperative to minimize welfare implications and production loss. Technology designed to identify lame cows using pressure plates to measure weight distribution for example has already been assessed (Bicalho et al., 2007). Although they concluded that more work was needed to refine the sensitivity of these devices, weight shifting by the cow may be visible by gait assessment and therefore these tools are likely to be effective in reducing the labour cost of mobility scoring. An approach to identify sub-clinical disease possibly by using behavioral changes before gait abnormalities are present may be more constructive.

Technology available for domestic and other commercial applications can be utilized to improve dairy cattle welfare and performance. For example, pedometers are effective in detecting increased activity during estrus, thus aiding dairy cattle fertility management (Roelofs et al., 2005). However, there is little scope for cattle behavioral classification using pedometer data. Accelerometers on the other hand have provided the ability to identify lying and standing behavior in grazed dairy cattle (Munksgaard et al., 2006; Nielsen et al., 2010). Nielsen (2013), using a 3-dimensional head-mounted activity logger to identify grazing behavior achieved a true positive classification rate of 84%. Dutta et al. (2015) achieved an average classification accuracy of 93% for grazing, searching, ruminating, resting and scratching when using a collar-mounted unit comprising of a global positioning system (**GPS**) receiver and 3-axis accelerometer. A behavioral recording system needs to be robust to the cow's environment and as accurate as possible. Pedometers attached to legs are open to damage and head-mounted accelerometers could be laborious to apply to cows. Furthermore, a system needs to be able to identify the main behaviors of the pasture-based dairy cow, preferably as a single, discrete and lightweight unit that is cheap to deploy in a commercial environment. As such, the scope for utilizing GPS collars alone to collect frequent temporal and positional data seems attractive.

The application of data mining and machine learning techniques in livestock behavioral studies to search for patterns in data that are unobservable by the human eye has been limited until recently. This has largely been because collecting a vast amount of behavioral data has been difficult. Such techniques have already been used to mine cattle disease databases to identify herd disease risk, for example (Ortiz-Pelaez and Pfeiffer, 2008), and to make breeding decisions based on the likelihood of conception from previous insemination data and disease history (Shahinfar et al., 2014). Applying

these techniques to behavioral data could present opportunities to learn more about the subtle changes that may occur over time during the onset of disease. The use of small, high data acquisition GPS receivers that are cheap and easy to apply to cattle may be a good incentive for pasture-based dairy practitioners provided the data are accurate and reliable. Furthermore, positional data could be extended to supply information on energy expenditure and pasture preference for grazing management as well as monitoring health status over time.

As far as we are aware, no published model of the main behaviors of pasture-based dairy cows from GPS data currently exists. The objective of this study was 2-fold. The first was to assess the ability of machine learning techniques in identifying the main behaviors grazing, resting and walking from GPS data. The second objective was to gain information about how these behaviors are represented within GPS data and to fully document this behavioral information as a novel model for the further study of welfare and production in the pasture-based dairy cow.

MATERIALS AND METHODS

Study Area

Data were collected at Aberystwyth University dairy farm, Trawsgoed, Ceredigion, United Kingdom, between March and August of 2014. The farm composed of approximately 200 Holstein and 150 Jersey-Friesian crossbred cows managed in a semi-intensive all-year-round calving system. Cows on the farm were rotated and strip-grazed on leafy swards of perennial ryegrass (*Lolium perenne*) between March and October on a grazing platform of approximately 100 ha with an altitude of between 70 and 250 m above sea level. Cows were allocated approximately 2,500 kg DM/ha of grass per area grazed leaving a grazing aftermath of approximately 1,500 kg of DM/ha. Cows in early lactation (0-120 DIM) were also buffer fed a TMR of grass [25 kg of

total fresh weight (**FW**)], maize silage (10 kg of FW), rolled wheat (3 kg of FW), barley straw (0.3 kg of FW) with the remainder comprising a rumen-protected fat supplement, dairy mineral blend and molasses during the evening. Early-lactation cows also received 3.60 kg of FW (± 0.81 kg) of concentrates per day. Cows were milked twice daily at 0500 and 1600 h.

Grazing Management and GPS

According to recent work (Kilgour, 2012) cows can exhibit up to 40 individual behaviors, though many of these are expressed in low abundance and for very short periods of time. Three main behaviors were identified and used in this work; grazing, resting and walking. Hancock (1954) reported that the main behaviors of pasture-based dairy cattle were grazing and resting. Due to the reported difficulty in distinguishing between lying and standing by others using GPS data (Homburger et al., 2014), we decided that collating these 2 behaviors in equal proportion and representing them as resting would lead to less complex decision rules during the machine learning phase. We also decided to include walking as this is also frequently reported in behavioral studies (Robert et al., 2009; Silper et al., 2015). Grazing was identified when the cow's head was lowered and tearing at the pasture whether walking or standing still. Browsing (walking with her head close to the pasture) was also included within grazing. Resting was identified when the cow was lying or when she stood still with her head raised. Walking was identified when the cow was walking or running with her head raised. A total of 40 early-lactation (50-120 DIM) Holstein cows were used. Average parity and milk yield (305 d) of cows were 2.8 ± 1.5 lactations (mean ± SD) and 7,414 ± 756 kg per cow respectively. Cows selected had normal gait (Whay et al., 2003) and showed no other obvious signs of ill health. Cows were otherwise randomly selected for behavioral observation over the period of study and observed between the

hours of 0800 and 1400 h. This time period was chosen simply because it allowed for the longest period of observation between milking.

Four cows were observed on each day of observation from a distance of approximately 40 m by 2 observers from within a vehicle. This allowed clear visibility of the behavior exhibited by the observed cows and minimized the impact that human presence may have had on their behavior. Observation days varied depending on observer availability and weather conditions. For example, cows were not observed on days where adverse weather conditions were forecast such as heavy rain. This meant that the number of days of observation and observation days themselves varied from week to week. On average, approximately 3 d were between sampling days. Both observers had previously spent 2 unrecorded observational days refining observational techniques to maximize agreement in behavioral and transitional definition. This was to ensure that high-quality time-stamped behaviors were recorded at and between each behavioral transition. Interobserver Kappa coefficients for grazing, resting and walking were 0.96, 0.99 and 1.00, respectively, for the unrecorded behavioral days. During each observation period, cows were equipped with low-cost GlobalSat DG-100 Data Logger GPS receivers (GlobalSat Technology Corp., Taiwan). Receivers were mounted on collars around the neck and cows were strip-grazed in a separate paddock at the same stocking density of the remainder of the herd (average stocking density approximately: 16 cows/ha). Choosing to graze the observation cows in a separate field to the rest of the herd was to ensure that cows could be fully observed at all times, reducing the risk of losing sight of the observation cows amongst others in the herd. The pasture of the observation paddock was also *Lollium perenne*. The sward was measured during the entire data collection period so that each strip provided 2,500 kg of DM/ha of grass. Cows were managed so that the grazing residue was approximately 1,500 kg of DM/ha.

This management regimen was the same for the entire farm and meant that pasture quality changed very little during the experimental period. The aim was to replicate a well-managed grazing system as closely as possible. This routine was enough for a total daily grazing and observational period of 6 h for each experimental period. Four new cows were introduced into a fresh strip of grass each day with the same pasture allowance. Each observer was allocated 2 cows for each period of observation. The GPS receivers were chosen randomly from a collection that was rigorously field tested under the specific guidelines of the Institute of Navigation (ION, 1997). From an initial 36 receivers, 11 were discarded due to non-normal positional fixes leaving 25 for the study. The GPS receivers were programmed to sample the positional coordinates every 5 s. From the calibration experiment (Section 3.2.4.2) this was considered the optimal logging interval for both data collection and power consumption. Using a Casio F-91W time clock (Casio Electronics Co. Ltd., Shibuya, Tokyo, Japan) synchronized to the GPS receivers, behaviors were manually observed and recorded at each transition, to the second they occurred for the whole 6-h observational period. Each 6-h period yielded 24 h of high-quality time-stamped cow activity from the 4 cows observed. An independent test data set was used on the final classifier models. This consisted of 14 h of data gathered from 4 previously unobserved cows managed in a separate, undulating paddock, with these cows mixing with the rest of the herd. Pasture was not managed as stringently in this paddock and, it contained a sward of varying density. This was a purposeful choice and was made to test the stability and robustness of the final models.

Data Sets

Data were downloaded from each GPS receiver as a comma separated variable (csv) file and stored in a spreadsheet program. The GPS data sets for each cow contained the record number of each logged position, time (s), latitude, longitude, speed (m/s) and altitude (m). The manually recorded field behaviors were matched alongside each row of GPS data. The compiled data set consisted of a total of 425 h (71 h of human observation) of behavioral and positional data. Behavioral sequences lasting at least 8 min were used for the analysis. This was done with the assumption that longer behavioral data sequences would provide a more reliable representation of the data sampled by the GPS units for each observed behavior. In total, 153 h of data was used for analysis: 28, 120, and 5 h for grazing, resting and walking, respectively.

Data Preprocessing and Preparation

Successive coordinates were exploited such that a variety of new features were extracted from the original data in 3 different phases. The approach detailed here is similar to that of earlier work (Godsk and Kjærgaard, 2011). However, the feature extraction and segmentation phases were modified. Briefly, the methodology employed is described below.

Phase 1

Raw, ground truth behavior-labeled and time-stamped GPS coordinates along with speed and altitude data were used to compute basic features for movement objects (**MO**). Contiguous data sequences for each behavioral class were selected from every data set represented by all 40 cows. An analysis of each consecutive GPS record allowed the calculation of the cow's distance traveled, speed, acceleration, absolute heading and bearing from one data entry to the next, every 5 s. Whether the cow was moving or not was determined using a simple adjustable Bayes induction filter. Three

input parameters were used for adjustment of the filter: minimum speed (minimum speed required for determining movement), history length (the number of previous GPS records taken into account) and heading threshold (the heading threshold required for determining movement). Minimum speed and heading threshold were determined for all 25 GPS receivers using data collected from a separate dynamic test under guidelines set out by the Institute of Navigation (ION, 1997). The average threshold values for all 25 GPS receivers were then calculated.

Phase 2

Data from phase 1 was grouped into segments of a predetermined size depending on the chosen segmentation strategy (i.e., the number of GPS instances to include when constructing each of the segments). Advanced features (**AF**) were then extracted for each segment based on the basic features from Phase 1. In total, 43 AF were extracted for phase 2 (Table 1). Of the behavioral classes under study, grazing and resting tended to be exhibited in the greatest proportion (Kilgour, 2012). Typically, dairy cattle graze in bouts of 7 min or less (Hejcmanová et al., 2009) with resting bouts lasting approximately 1 h. However, for this work, we explored 3 different time intervals for segmentation to investigate whether the segmentation strategy would have any effect on classifier performance: 160, 40 and 10 s. Thus, at a GPS sample rate of 5 s, each segmentation interval comprised 32, 8, and 2 MO, respectively. Data for each segmentation strategy were evaluated separately. Once the data had been segmented at the desired interval, each group of segmented data then formed a segment object (**SO**). The SO contained the data for all 43 AF. Table 2 shows the unbalanced datasets for the 3 respective segmentation strategies with the corresponding number of SO.

Table 1. Advanced features computed for each segment of GPS data to be used for behavioral classification

Advanced feature	Movement type (unit)
Accumulated time moving	Time (s)
Accumulated time nonmoving	Time (s)
Minimum speed	Speed (m/s)
Mean speed	Speed (m/s)
Maximum speed	Speed (m/s)
Accumulated distance moving	Distance (m)
Accumulated distance nonmoving	Distance (m)
Maximum distance moving	Distance (m)
Maximum distance nonmoving	Distance (m)
Movement percent left	Heading (%)
Movement percent right	Heading (%)
Movement percent forward	Heading (%)
Movement percent U-turn	Heading (%)
Percent nonmoving	Heading (%)
Rate of change moving and nonmoving	(no.)
Rate of change between any movement	(no.)
Rate of change left	(no.)
Rate of change right	(no.)
Rate of change forward	(no.)
Rate of change nonmoving	(no.)
Rate of change U-turn	(no.)
Minimum acceleration	Acceleration (m/s^2)
Maximum acceleration	Acceleration (m/s^2)
Mean acceleration positive	Acceleration (m/s^2)
Mean acceleration negative	Acceleration (m/s^2)
Accumulated acceleration positive	Acceleration (m/s^2)
Accumulated acceleration negative	Acceleration (m/s^2)
Changes between positive and negative acceleration	(no.)
Heading accumulated left	(no.)
Heading accumulated right	(no.)
Heading accumulated forward	(no.)
Heading accumulated nonmoving	(no.)
Heading accumulated U-turn	(no.)
Maximum heading change left	Heading (°)
Maximum heading change right	Heading (°)
Maximum heading change forward	Heading (°)
Maximum heading change nonmoving	Heading (°)
Maximum heading change U-turn	Heading (°)
Mean heading change per left	Heading (°)
Mean heading change per right	Heading (°)
Mean heading change per forward	Heading (°)
Mean heading change per nonmoving	Heading (°)
Mean heading change per U-turn	Heading (°)

We tested the data set where the decision classes (behaviors) were imbalanced and also balanced. The data instances were balanced by reducing the size of the behavioral category represented by the greatest number of SO (resting) to the same size as the category with the fewest SO. This is known as undersampling. We found that dairy cows exhibited very little walking behavior during the data collection phase and movement in itself was exhibited most of the time as grazing and cows spent the majority of their time grazing and resting. This could be due to the greater energy expenditure of walking thus this behavior may be limited especially in early-lactation cows (Dohme-Meier et al., 2014). Due to the great underrepresentation of walking in the data set, resting was reduced to the same level as grazing to preserve the integrity of the data set. Balancing the data instances was carried out using a subsampling method *SpreadSubsample* (Japkowicz and Stephen, 2002). This method produced a random undersample of the class with the greatest representation to a specified sample size. The size of the random undersampled class was predefined and was used to specify the number of SO for each individual dataset. Table 2 shows the class distribution for the data set balanced using random undersampling.

Table 2. Number of segment objects created for each behavior in unbalanced data sets and data sets balanced by random undersampling

Segmentation strategy[1] (movement object/segment)	Segment objects[2]					
	Unbalanced data sets			Balanced data sets		
	Grazing	Resting	Walking	Grazing	Resting	Walking
32	631	2,703	96	631	631	96
8	1,864	5,238	407	1,864	1,864	407
2	9,921	45,069	1,612	9,921	9,921	1,612

[1]Movement objects present per segment. For example, 32 movements equates to 160 s of GPS data (at a 5-s GPS sample rate) per segment object.
[2]Each segment object contained data for all advanced features for the expressed behavior.

Phase 3

The SO and their corresponding ground-truth behaviors were compiled and formatted such that they could be examined using the WEKA (Hall et al., 2009) data mining suite.

Classifier Learning

The popular data mining suite WEKA was used for the analysis of the data in this study. This suite of tools allows many different machine learning approaches to be used for comparative analysis and provides a variety of different metrics for assessing the performance of learners. The algorithms tested on the data used in this study were naïve Bayes (John and Langley, 1995), JRip (Cohen, 1995), J48 (Quinlan, 1993) and random forest (Breiman, 2001). Naïve Bayes is a simple technique for constructing classifiers, which are represented as vectors of feature values. The classifier considers each of the features to contribute independently (strong independence assumption) to the probability that an object belongs to a particular decision class. This is assumed regardless of any possible actual correlation(s) between features. The maximum

likelihood is used as a metric to decide which class a testing object belongs. J48 creates decision trees by choosing the most informative features and recursively partitioning the data into subtables based on their values. Each node in the tree represents a feature with branches from a node representing the alternative values this feature can take according to the current subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created, and this classification assigned. JRip learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, antecedents are added greedily until a termination condition is satisfied. Antecedents are then pruned in the next phase subject to a pruning metric. Once the ruleset is generated, a further optimisation is performed where rules are evaluated and deleted based on their performance on randomised data. Random forest is an ensemble learning method for classification. It works by constructing a collection ("forest") of (random) decision trees at training time and returning the class that is the mode of all of the classes of the individual trees. Random forest classifiers attempt to mitigate the tendency of decision trees to overfit the training data set. The reasoning behind these particular choices was to try to provide a realistic set of results and also to show the different characteristics of the learners themselves. Naïve Bayes is a pessimistic learner and therefore biased but stable (Bouckaert, 2008). J48 is an optimistic learner and it is therefore unbiased, suffers from high variance and is thus unstable (Bouckaert, 2008). JRip is a rule-based classifier and lies somewhere in between naïve Bayes and J48. Random forest attempts to reduce variance (and therefore total error).

Classifier Validation Strategy

Classifier performance was evaluated using stratified 10-fold cross-validation (**10-FCV**). In 10-FCV, the data set is randomly split into 10 folds or subsets of data where class representation is preserved in the same proportion (as far as possible) as the full data set. The first nine folds are then used to learn a classifier, whereas the tenth is used for validation. This process is then repeated each time using a single fold for validation with the remainder used for training. This continues until all 10 folds have been used for both training and validation resulting in an average classification accuracy (**CA**) and error rate. For the results generated here, 10-FCV was repeated for 10 different randomisations/runs (10 x 10-FCV) of the data. This helped to learn more robust classifiers and was done for all datasets in Table 2. Finally, the best performing classifier models were applied to a previously unseen independent test data set.

Selection of Advanced Features

In addition, the redundancy and relevance of the features extracted from the GPS data (AF) were investigated for the best performing classifiers to explore which features were most useful in classifying the 3 behaviors. Advanced features were evaluated using CFS (Hall, 1998) which selects subsets of features that are highly correlated with the class value and that have low correlation with each other. A "greedy hill-climbing" approach was used to perform the search through the space of AF. Classifier performance and stability was then evaluated using the best performing data set in Table 2 with 10-FCV.

Performance Evaluation

The overall CA was one of the metrics used to assess performance. This is the most commonly reported and intuitive metric for classifier performance. It serves as a general indicator of the efficiency of a model to correctly predict all of the behavioral classes:

$$\text{Classification accuracy} = \frac{TP + TN}{TP + FP + FN + TN},$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. However, as a stand-alone metric of performance of multi-class problems, CA does not compensate for success that is due to mere chance.

$$\text{False Positive Rate (}\mathbf{FPR}\text{)} = \frac{FP}{FP + TN},$$

which is the proportion of instances incorrectly labelled as positive instances.

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}},$$

where precision = TP / (TP + FP), and recall = TP / (TP + FN). The F-measure is defined as the harmonic mean of precision and recall.

Area under the receiver operating curve (**AUC**) was represented as

$$AUC = (TP + TN) / 2,$$

which is a measure of the discriminatory power of a classifier which measures the area under the receiver operating characteristic (**ROC**) curve (Fawcett, 2006). The ROC curve is constructed using the total true positives (y-axis) and the total true negatives (x-axis). A classifier generating more true positives and fewer false positives is preferable to the opposite. It should be noted that AUC is not without its problems

when used as a metric in and of itself for the classification problem (Hand, 2009) and this is the reason for including several other performance metrics above.

All statistical analysis was undertaken using GenStat Fourteenth Edition (VSN International, Hemel Hempstead, UK; footnotes below tables give information on the statistical tests used).

RESULTS

Classification accuracy was first compared with the unbalanced data sets (Table 2). Weighted average CA across all segmentation strategies was best for JRip and random forest (Table 3). Class imbalance can lead to the reinforcement of the majority class in the learned rules, thus leading to domination by a particular class (Kohavi, 1995). In this case, resting represented on average 76% of the SO in each dataset and had high CA (Table 3). Interestingly, walking, which represented the behavioral class with fewest instances (Table 2), had the highest comparable CA (Table 3); this was due to the easily distinguishable characteristics obtained for this class. Due to the relative underrepresentation of grazing in each data set and the apparent confusion between this class and resting, class-specific CA was generally poor for this behavior particularly for the 2 SO dataset. To address this, the data set was balanced by random undersampling (Table 2). The success of each classifier was determined based on relative performance for each segmentation strategy. The best classifiers were JRip and random forest when measured across all metrics regardless of the segmentation strategy employed (Table 4). The effect of balancing the class representation for the individual behaviors was investigated using only JRip and random forest because of their superior performance (Figure 1) when compared with naïve Bayes and J48. Additionally, performance for all metrics suffered as the data set was segmented into smaller sized SO and best performance was found when SO were composed of 32 MO (Table 4). This effect on performance is linked to the averaging effects of the inclusion of greater numbers of SO. Hence, larger segments produce more easily definable MO.

Table 3. Weighted classification accuracy of all classifier learners on unbalanced data for grazing, resting and walking for 32-, 8-, and 2-segment object strategies using 10-fold cross validation

Classifier	Segmentation strategy[1]	Classification accuracy			
		Grazing	Resting	Walking	Weighted average
Naïve Bayes	32	0.72	0.71	0.99	0.72
	8	0.71	0.64	0.99	0.68
	2	0.72	0.60	1.00	0.64
JRip	32	0.56	0.92	0.98	0.85
	8	0.48	0.89	0.98	0.79
	2	0.09	0.98	0.96	0.82
J48	32	0.56	0.90	0.98	0.84
	8	0.50	0.86	0.98	0.77
	2	0.16	0.95	0.95	0.81
Random forest	32	0.53	0.93	0.99	0.86
	8	0.51	0.88	0.99	0.79
	2	0.24	0.93	0.96	0.81

[1]Segmentation strategy was the number of movements [or global positioning system (GPS) instances] per data segment. For example, the 32-segment object segmentation strategy equates to 160 s of GPS data (at a 5-s GPS sample rate) per segment object.

Table 4. Weighted average classifier learner performance on data balanced by random undersampling under 32-, 8-, and 2 segment object strategy using 10-fold cross validation

	Naïve Bayes			JRip			J48			Random forest			
Metric[1]	32 SO[2]	8 SO	2 SO	32 SO	8 SO	2 SO	32 SO	8 SO	2 SO	32 SO	8 SO	2 SO	SEM
CA	0.73^{c}	0.71^{c}	0.69^{c}	0.81^{a}	0.76^{a}	0.73^{a}	0.78^{b}	0.73^{b}	0.72^{b}	0.81^{a}	0.76^{a}	0.72^{b}	0.002
FPR	0.23^{a}	0.22^{a}	0.23^{b}	0.17^{c}	0.20^{b}	0.23^{b}	0.19^{b}	0.22^{a}	0.24^{a}	0.16^{c}	0.20^{b}	0.24^{a}	0.002
F-Measure	0.72^{c}	0.70^{c}	0.69^{c}	0.81^{a}	0.76^{a}	0.73^{a}	0.78^{b}	0.73^{b}	0.72^{b}	0.81^{a}	0.76^{a}	0.72^{b}	0.002
AUC	0.82^{c}	0.82^{b}	0.79^{c}	0.85^{b}	0.83^{b}	0.80^{b}	0.81^{d}	0.76^{c}	0.78^{d}	0.90^{a}	0.86^{a}	0.81^{a}	0.002

[a-d]Means within a row with different superscripts are compared between classifiers for each corresponding segmentation strategy and differ ($P < 0.05$).
[1]Means were tested using ANOVA and Tukey's post-hoc test. CA = classification accuracy; FPR = false positive rate; AUC = area under the receiver operating curve.
[2]Segmentation strategy was the number of movements (or GPS instances) per data segment. For example, the 32-segment object segmentation strategy equates to 160 s of GPS data (at a 5-s GPS sample rate) per segment object.

Balancing the data sets by random undersampling led to an improvement in CA for grazing relative to the previous performance (JRip = +0.26, +0.28, and +0.68; random forest = +0.33, +0.28, and +0.52 for 32, 8, and 2 SO, respectively) but a reduction in performance for resting (JRip = -0.15, -0.18, and -0.33; random forest = -0.20, -0.20, and -0.28 for 32, 8 and 2 SO, respectively). The effect of class balancing the data on walking was negligible. It is clear that these two algorithms have a greater ability to classify behaviors where physical movement is a prominent feature of the data. Resting behavior was sometimes confused with grazing for all learners, probably because of the inherent error of the GPS device and therefore the perceived subsequent movement of cows between consecutive fixes (Lewis et al., 2007), when in fact they are resting.

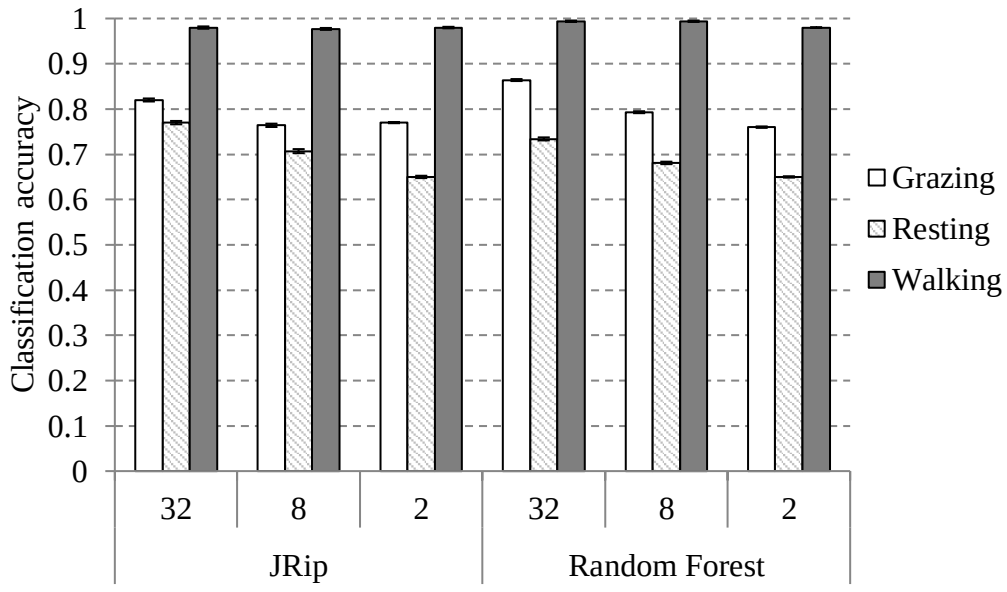


Figure 1. Average classification accuracy (±SE) of JRip and random forest classifier learners across all 3 behaviors according to segmentation strategy and data balanced by random undersampling.

Table 5. Rule set generated by the JRip classifier learner for classifying dairy cow behavior using advanced features created from GPS data

Rule	Behavior[1]
IF maximum speed ≥ 1.10 m/s	Walking (98.0/3.0)
IF mean acceleration positive ≥ 3.73E-6 m/s^2 AND maximum distance moving ≤ 2.17 m AND mean speed ≥ 0.10 m/s AND accumulated distance moving ≥ 11.92 m	Grazing (215.0/19.0)
IF maximum heading change nonmoving ≤ 258.69° AND minimum speed ≥ 0.04 m/s AND maximum distance moving ≤ 2.77 m AND mean heading change per right ≤ 144.61°	Grazing (121.0/12.0)
IF maximum heading change nonmoving ≤ 258.69° AND maximum distance moving ≤ 1.38 m AND minimum speed ≥ 0.04 m/s AND maximum distance moving ≤ 1.10 m AND mean heading change per right ≤ 232.63° AND rate of change between any movement ≤ 0.15	Grazing (35.0/3.0)
IF accumulated acceleration positive ≥ 4.59E-5 m/s^2 AND maximum distance moving ≤ 2.16 m AND mean heading change per right ≤ 140.80°	Grazing (104.0/26.0)
IF mean acceleration positive ≥ 3.73E-6 m/s^2 AND maximum distance moving ≤ 2.40 m AND mean acceleration negative ≥ 4.97E-6 m/s^2 AND maximum distance moving ≥ 1.24m	Grazing (54.0/13.0)
IF maximum heading change nonmoving ≤ 255.96° AND changes between positive and negative acceleration ≤ 5 AND maximum acceleration ≤ 1.12E-5 m/s^2	Grazing (39.0/8.0)
IF rate of change between any movement ≥ 0.11 AND maximum distance moving ≤ 1.60 m AND mean heading change per left ≥ 148.82°	Grazing (56.0/17.0)
Otherwise	Behavior = resting (636.0/106.0)

[1]Numbers in parentheses below behavioral classes indicate number of successful classifications/errors in the balanced training data.

Due to the superior performance of JRip and random forest when compared with other learners (Table 4), further validation and testing took place with these learners and the 32 SO data set only. During observation, cows generally exhibited each behavior for periods exceeding 3 min; therefore, the 32 SO strategy coincided well with average behavioral duration. An independent test set consisting 14 h of ground truth behavior-labeled and time-stamped GPS coordinates was used for final model testing that was segmented at 32 SO. The final model as derived from the validation phase for JRip is shown (Table 5) along with its respective success rates during modeling. It was interesting to note the induced rules and their antecedent values for each of the behaviors. Many of these were informative and because JRip generates rules that are transparent to human scrutiny, it offered an insight into how features of physiological movement of cows helped to characterise the behavior. For example, it could be seen that walking was so easy to classify that only a single antecedent was required to predict the behavior: maximum speed. This was both informative and realistic because if a healthy cow was walking in a given direction it would always exceed this threshold ($\geq$1.10 m/s). The rules also helped to reflect the complex nature of grazing behavior due to the rule clarity and number of actual rules. This complexity should not be interpreted as a poor reflection of the learner, but rather that grazing because of the physical nature, involves elements of both walking and resting since the cow may stop for a short period while it tears at pasture before moving on to the next area. Thus, many different antecedents were involved in building a model to represent it.

Table 6. Average performance of final models of JRip and random forest classifier learners on an independent, balanced 32-segment object test set

Metric[1]	JRip	Random forest	SEM	*P*-value
CA	0.85	0.83	0.008	NS[2]
FPR	0.10	0.17	0.010	NS
F-Measure	0.76	0.77	0.013	NS
AUC	0.87	0.92	0.009	NS

[1]Tested using 2 sample *t*-test. CA = classification accuracy; FPR = false positive rate; AUC = area under the receiver operating curve.
[2]NS = $P > 0.05$.

No significant differences were found in any of the performance measures (Table 6) for JRip and random forest on the independent test set. These results were encouraging because data for the test set was purposely retrieved from cattle grazing an undulating, densely stocked paddock where pasture conditions were not controlled as stringently as the area where the data were gathered for model building and validation. Classification accuracy of individual behaviors (Figure 2) was more balanced across the behaviors for JRip when compared with random forest. Walking behavior, in both examples, was classified with almost 100% accuracy.

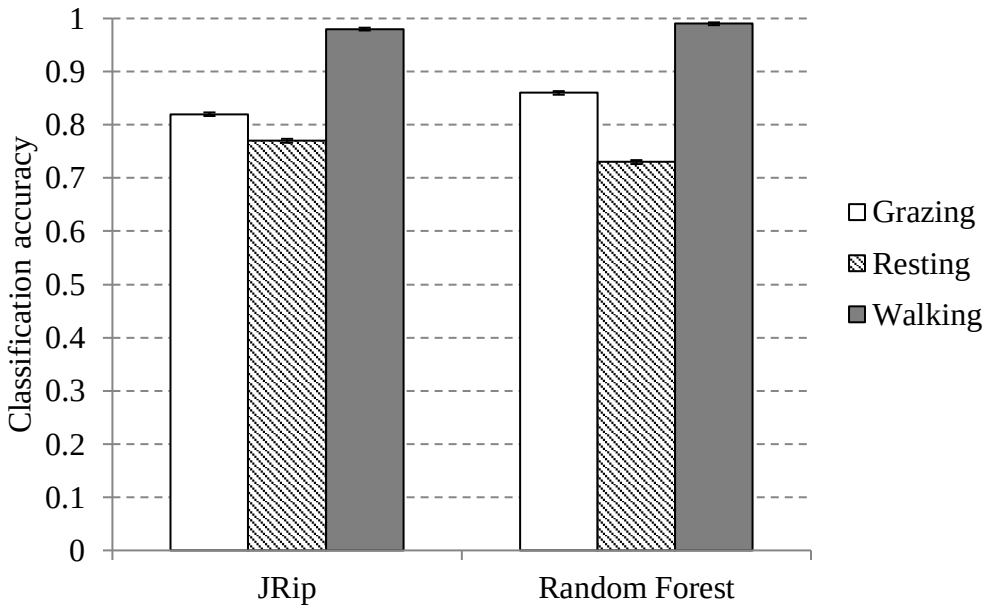


Figure 2. Average classification accuracy (±SE) of final models of JRip and random forest classifier learners across all 3 behaviors on an independent, balanced 32-segment object test set.

A feature selection and stability assessment was performed on the original 32 SO training data. Table 7 gives a comparison of the performance of JRip and random forest both with and without AF selection as tested by 10-FCV. No significant improvement or degradation was found in classifier performance when the AF with the highest contributions during modeling of the observed behaviors were used (Table 7). Figure 3 illustrates the average number of folds in which the most important AF appeared during 10-FCV (10 runs x 10 folds). Among the 13 selected AF (Figure 3), features 2, 5, 6, and 7 (i.e., maximum heading change nonmoving, minimum speed, mean speed and maximum speed) were selected in every fold. This indicated that it may not be necessary to consider extracting all of the AF listed in Table 1 for future work in this field. Although a marginal improvement was found in performance for random forest, this was nonsignificant. Furthermore, both classifiers had a small standard deviation and therefore good stability for all measures of performance. However, despite the comparable performance of JRip and random forest, the simplicity of the rule set generated by JRip (Table 5) is advantageous because it provides a model which is transparent and humanly interpretable through the use of intuitive if-then rules.

Table 7. Weighted average performance and stability (±SD) of JRip and random forest classifier learners before and after selection of advanced features on original 32-segment object data set balanced by random undersampling using 10-fold cross validation

	JRip				Random forest			
Metric[1]	All features	Selected features	SEM	*P*-value	All features	Selected features	SEM	*P*-value
CA	0.81 (3.66)	0.81 (3.74)	0.337	NS[2]	0.81 (3.37)	0.82 (3.06)	0.306	NS
FPR	0.17 (0.03)	0.16 (0.03)	0.003	NS	0.16 (0.03)	0.15 (0.03)	0.003	NS
F-Measure	0.81 (0.04)	0.81 (0.04)	0.004	NS	0.81 (0.03)	0.82 (0.03)	0.003	NS
AUC	0.85 (0.03)	0.85 (0.03)	0.003	NS	0.90 (0.03)	0.91 (0.03)	0.003	NS

[1]Tested using 2-sample *t*-test. CA = classification accuracy; FPR = false positive rate; AUC = area under the receiver operating curve.
[2]NS = $P > 0.05$.

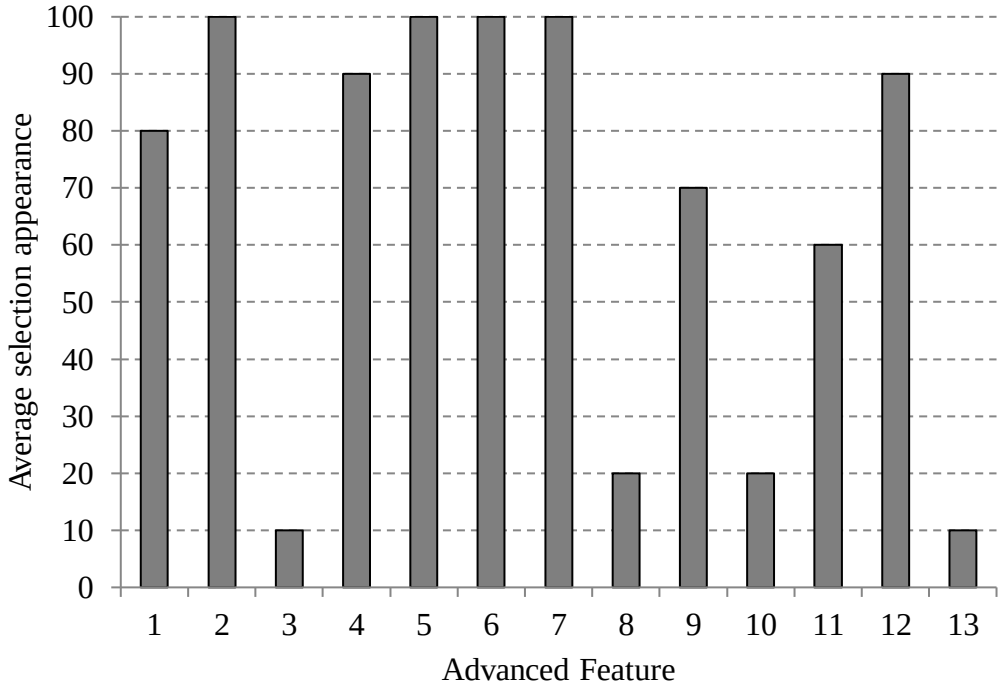


Figure 3. Average selection appearances of the most valuable advanced features. The advanced features from 1-13 correspond to (1) movement percent forward, (2) maximum heading change nonmoving, (3) mean heading change percent left, (4) mean heading change percent right, (5) minimum speed, (6) mean speed, (7) maximum speed, (8) mean acceleration positive, (9) mean acceleration negative, (10) accumulated acceleration positive, (11) accumulated acceleration negative, (12) accumulated distance moving, (13) maximum distance moving.

DISCUSSION

The objective of this work was to build a robust model of the most notable behaviors of the pasture-based dairy cow: grazing, resting, and walking and to fully document how this was derived. We have achieved greater performance than other comparable studies using GPS only to classify the most commonly represented behavioral activities (Homburger et al., 2014). Homburger et al. (2014) used a logging interval of 20 s to record positional fixes achieving an overall CA of 77%. Longer fix intervals can create uncertainty about the location and the subsequent intra-sample behavior exhibited by cattle (Swain et al., 2008). This can lead to inaccuracies in distance traveled and speed between fixes (Pépin et al., 2004). Speed and distance metrics featured heavily in the final model generated using the JRip learner for grazing and walking behaviors in this work and indeed this was reflected in the learned rules.

Cattle could of course be grazing or walking in several different directions in the period between samples, limiting the use of the extracted features to distinguish behaviors. The final models in this study were based on 40 cows and 153 h of time-stamped cow behavior gathered over a period of 4 mo. To our knowledge, this is the most extensive data set used for the preparation of such models. We believe this was sufficient to account for conditions out of our experimental control such as atmospheric effects and dilution of GPS precision (ION, 1997). Furthermore, the models achieved high performance when tested on an independent test set where cows were grazed outside experimentally controlled conditions. Unaccounted for in the training data were multipath effects, errors in GPS output created by topographical features and buildings, for example (Cai et al., 2014). However, the independent test set contained instances where cows were standing next to trees, hedgerows, and buildings, suggesting that overall performance in the test set may have been higher had cows been restricted to

areas where GPS signal-to-noise ratio was lower. We deem the behaviors under study not only to be the most easily definable when using temporal movement metrics produced from GPS receivers, but perhaps also the most important for monitoring cow health in pasture-based cows. Lying times have been shown to increase in both severely lame and clinically lame cows (Sepúlveda-Varas et al., 2014) and standing time has been shown to be higher in cows with ketosis (Itle et al., 2015). It seems that behavior is also affected by age, parity and disease. For example, Charlton et al. (2015) found that cows with hock and knee injuries spent less time lying compared to cows without injuries and Steensels et al. (2012) found that lying times increased significantly with age in housed cows. To our knowledge, little work exists in the long-term monitoring of behavioral changes associated with disease in pasture-based dairy cows and is therefore the long-term objective of the current work. Information has already been gathered on the effect of disease on the feeding behavior of housed cows (Huzzey et al., 2007; González et al., 2008) and therefore more is needed for long-term assessment of behavior in cows at pasture. Inter-cow comparisons are likely to be of lower value than long-term intra-cow behavior for disease diagnostics (González et al., 2008; Reader et al., 2011).

As well as the most notable behaviors, other authors (Dutta et al., 2015) have had success in classifying other, less frequently exhibited behaviors, for example, rubbing, scratching and licking. These behaviors are often combined into a single class due to their relatively poor representation in general bovine behavior. Recent success has been achieved in classifying data from other cow-mounted sensors. For example, Nielsen (2013) successfully classified cattle grazing (sensitivity = 83.63%; specificity = 90.20%; precision = 85.75%) when data from a 3-dimensional head-mounted activity sensor were coupled with data from a 3-dimensional leg-mounted accelerometer.

Robert et al. (2009) classified lying and standing behaviors of 15 calves to a very high accuracy (CA = 99 and 98%, respectively) using accelerometers however, less success was achieved in classifying walking (CA = 68%). We achieved high CA with JRip and random forest on the independent test set (JRip = 0.85; random forest = 0.83). Both JRip and random forest also achieved a false discovery rate (0.10 versus 0.17, respectively) comparable with other published work (Dutta et al., 2015). In comparison with such work, Pluk et al. (2012) and Viazzi et al. (2014) describe methods of automatically identifying changes in the posture of dairy cows using video recordings to automatically detect lameness with some success. Whereas this type of work sets a strong benchmark for making gait classifications more reliable, it is less time consuming and reduces the subjectivity associated with manual gait assessment and does little to detect the early onset of lameness. Cows are often described as stoic prey animals and may not display obvious gait abnormalities during the early onset of a lameness problem for example. This may account for the little change that is seen between the lower end of mobility scoring systems and it may only be when the pain has become unbearable that cows eventually show signs of an altered gait. As shown by Reader et al. (2011), decreases in milk yield were evident 7 wk before the manifestation of the later diagnosed lameness.

Due to its simplicity, CA is the most commonly used evaluation metric for model performance. However, it does not compensate for results that are due to chance (Ben-David, 2008). The strength of a model can also be defined by the area under the receiver operating characteristic (ROC) curve, known as the AUC. Graphically, in the ROC space the diagonal line y = x represents a classification strategy that randomly guesses a class (Fawcett, 2006). Larger AUC values are preferred and represent a ROC

curve that is positioned toward the top left corner of the ROC space. JRip and random forest performed very well on the independent test set (0.87 versus 0.92, respectively).

Despite the similar performance of JRip and random forest, the simplicity of JRip and comparable execution time makes it the favoured choice of the 2 for this particular work. JRip uses a global optimization phase on an initial set of rules built from scratch, and consequently, the resulting model has reduced complexity. Previous studies (Swain et al., 2008; Forin-Wiart et al., 2015) found that increasing the GPS fix rate provided a more accurate account of the movement of animals. Fix rate did not have a significant effect on absolute accuracy during static accuracy tests (Section 3.2.4.2); therefore, the only limitation to data collection was battery power. We found that the optimal fix rate for maximum data acquisition and power consumption for the GlobalSat DG-100 GPS receiver was 5 s (Section 3.2.4.2). This allowed for approximately 17,000 fixes to be recorded in 24 h on each receiver equating to approximately 28% of the capacity of the receiver. We have already measured the stationary accuracy of these receivers in conditions where for example the receiver may come into close proximity with hedgerows. For the application of our model in the further study of dairy cow behavior in uncontrolled conditions, these factors will be taken into account.

Practically, the simplicity of placing each receiver onto cows was as simple as placing any other management tool on cow collars for example a pedometer for measuring activity in reproductive management. With minimal positional adjustment required, the use of small GPS receivers may be useful in a commercial environment. We believe that the versatility and the high level of performance attained in the present work constitute a strong case for the use of small, discrete, low cost (approximately $60 US) GPS receivers mounted on neck collars. Using a variety of features to describe the

behavior of cattle from GPS data, we have produced a humanly interpretable model of the main behaviors that can be used for further study. This model extends on the success of others in identifying cattle behavior (de Weerd et al., 2015; Dutta et al., 2015) and uses movement features that can be further exploited in understanding the behavior of cows at pasture from GPS only. Nevertheless, power consumption is a weakness of current GPS receivers and more work on this aspect is required in their development for practical application. As well as behavioral identification, the flexibility of GPS would also allow herdsmen to retrieve positional fixes as well as data on distances travelled by cows and their energy expenditure. This could assist in making more informed decisions on future nutritional management as well as the possibility of monitoring health status over time.

CONCLUSIONS

The results of this work illustrate the application of data mining techniques to features extracted from temporal GPS data. Furthermore, it highlights the wealth of data that can be gathered from small, low-cost GPS receivers alone. The success rate of our final model using JRip highlights its ability to classify the main behaviors exhibited by pasture-based dairy cows in a way that is transparent to human scrutiny. The next step will be to identify GPS data sequences representative of behavioral transitions. This model can then be used to track temporal changes in the behavior of pasture-based dairy cows afflicted by disease to work towards providing a more objective means of early disease identification.

ACKNOWLEDGEMENTS

The authors thank the farm staff at Aberystwyth University Research Farm (Aberystwyth, Ceredigion, UK) for their cooperation. The sponsorship of M. Williams by Coleg Cymraeg Cenedlaethol (Carmarthen, Wales, UK) is gratefully acknowledged.

REFERENCES

Ben-David, A. 2008. About the relationship between ROC curves and Cohen's kappa. Eng. Appl. Artif. Intell. 21: 874-882. http://dx.doi.org/10.1016/j.engappai.2007.09.009

Bicalho, R. C., S. H. Cheong, G. Cramer, and C. L. Guard. 2007. Association between a visual and an automated locomotion score in lactating Holstein cows. J. Dairy Sci. 90: 3294-3300. http://dx.doi.org/10.3168/jds.2007-0076

Bouckaert, R. R. 2008. Practical bias variance decomposition. *AI 2008: Adv. in Artif. Intell.* Springer Berlin Heidelberg, 247-257.

Breiman, L. 2001. Random Forests. Machine Learning. 45: 5-32.

Cai, C., Y. Gao, L. Pan, and W. Dai. 2014. An analysis on combined GPS/COMPASS data quality and its effect on single point positioning accuracy under different observing conditions. Adv. Space Res. 54: 818-829. http://dx.doi.org/10.1016/j.asr.2013.02.019

Charlton, G. L., V. Bouffard, J. Gibbons, E. Vasseur, D. B. Haley, D. Pellerin, and A. M. de Passillé. 2015. Can automated measures of lying time help assess lameness and leg lesions on tie-stall dairy farms? Appl. Anim. Behav. Sci. http://dx.doi.org/10.1016/j.applanim.2015.02.011

Cohen, W. W. 1995. Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123.

de Weerd, N., F. van Langevelde, H. van Oeveren, B. A. Nolet, A. Kölzsch, H. H. Prins, and W. F. de Boer. 2015. Deriving animal behavior from high-frequency GPS: tracking cows in open and forested habitat. PloS One. doi: 10.1371/journal.pone.0129030.

Dohme-Meier, F., L. D. Kaufmann, S. Görs, P. Junghans, C. C. Metges, H. H. van Dorland, and A. Münger. 2014. Comparison of energy expenditure, eating pattern and physical activity of grazing and zero-grazing dairy cows at different time points during lactation. Livestock Science. 162: 86-96. http://dx.doi.org/10.1016/j.livsci.2014.01.006

Dutta, R., D. Smith, R. Rawnsley, G. Bishop-Hurley, J. Hills, G. Timms, and D. Henry. 2015. Dynamic cattle behavioral classification using supervised ensemble classifiers. Comput. Elect. Ag. 111: 18-28. http://dx.doi.org/10.1016/j.compag.2014.12.002

Dyer, R. M., N. K. Neerchal, U. Tasch, Y. Wu, P. Dyer, and P. G. Rajkondawar. 2007. Objective determination of claw pain and its relationship to limb locomotion score in dairy cattle. J. Dairy Sci. 90: 4592-4602. http://dx.doi.org/10.3168/jds.2007-0006

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27: 861-874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.

Forin-Wiart, M. A., P. Hubert, P. Sirguey, and M. L. Poulle. 2015. Performance and Accuracy of Lightweight and Low-Cost GPS Data Loggers According to Antenna Positions, Fix Intervals, Habitats and Animal Movements. PloS One. 10, e0129271. doi: 10.1371/journal.pone.0129271.

Godsk, T., and M. B Kjærgaard, M. B. 2011. High classification rates for continuous cow activity recognition using low-cost GPS positioning sensors and standard machine learning techniques. Adv. in Data Mining. Applications and Theoretical Aspects: 174-188. Springer Berlin Heidelberg.

González, L. A., B. J. Tolkamp, M. P. Coffey, A. Ferret, and I. Kyriazakis. 2008. Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. J. Dairy Sci. 91: 1017-1028. http://dx.doi.org/10.3168/jds.2007-0530.

Hall, M. A. 1998. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update; SIGKDD explorations, 11: 10-18. http://dx.doi.org/10.1145/1656274.1656278.

Hancock, J. 1954. Studies of grazing behavior in relation to grassland management I. Variations in grazing habits of dairy cattle. The Journal of Agricultural Science. 44: 420-433. http://dx.doi.org/10.1017/S0021859600045287

Hand, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 77.1: 103-123.

Hejcmanová, P., M. Stejskalová, V. Pavlů, and M. Hejcman. 2009. Behavioral patterns of heifers under intensive and extensive continuous grazing on species-rich pasture in the Czech Republic. Appl. Anim. Behav. Sci. 117: 137-143. http://dx.doi.org/10.1016/j.applanim.2009.01.003.

Homburger, H., M. K. Schneider, S. Hilfiker, and A. Lüscher. 2014. Inferring Behavioral States of Grazing Livestock from High-Frequency Position Data Alone. PloS One 9: e114522. doi:10.1371/journal.pone.0114522.

Huzzey, J. M., D. M. Veira, D. M. Weary, and M. A. G. Von Keyserlingk. 2007. Prepartum behavior and dry matter intake identify dairy cows at risk for metritis. J. Dairy Sci. 90: 3220-3233. http://dx.doi.org/10.3168/jds.2006-807.

ION. 1997. ION STD 101: Recommended Test Procedures for GPS Receivers. Alexandria, Va. Institute of Navigation Standards.

Itle, A. J., J. M. Huzzey, D. M. Weary, and M. A. G. von Keyserlingk, M. A. G. 2015. Clinical ketosis and standing behavior in transition cows. J. Dairy. Sci. 98: 128-134. http://dx.doi.org/10.3168/jds.2014-7932

Japkowicz, N., and S. Stephen. 2002. The class imbalance problem: A systematic study. Intell. Data Anal. 6: 429-449.

John, G. H., P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345.

Kilgour, R. J. 2012. In pursuit of "normal": A review of the behavior of cattle at pasture. Appl. Anim. Behav. Sci. 138: 1-11. http://dx.doi.org/10.1016/j.applanim.2011.12.002

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI), Montreal, Quebec, Canada August 20–25 Morgan Kaufmann Publishers, pp. 1137–1145.

Lewis, J. S., J. L. Rachlow, E. O. Garton, and L. A. Vierling. 2007. Effects of habitat on GPS collar performance: using data screening to reduce location error. J. Appl. Ecol. 44: 663-671. http://dx.doi.org/10.1111/j.1365-2664.2007.01286.x.

Manske, T., J. Hultgren, and C. Bergsten. 2002. Prevalence and interrelationships of hoof lesions and lameness in Swedish dairy cows. Prev. Vet. Med. 54: 247-263. http://dx.doi.org/10.1016/S0167-5877(02)00018-1

Munksgaard, L., C. Reenan, and R. Boyce. 2006. Automatic monitoring of lying, standing and walking behavior in dairy cattle. J. Dairy. Sci. Suppl. 89: 304.

Nielsen, L. R., A. R. Pedersen, M. S. Herskin, and L. Munksgaard. 2010. Quantifying walking and standing behavior of dairy cows using a moving average based on output from an accelerometer. Appl. Anim. Behav. Sci. 127: 12-19. http://dx.doi.org/10.1016/j.applanim.2010.08.004

Nielsen, P. P. 2013. Automatic registration of grazing behavior in dairy cows using 3D activity loggers. Appl. Anim. Behav. 148: 179-184. http://dx.doi.org/10.1016/j.applanim.2013.09.001.

Ortiz-Pelaez, Á., and D. U. Pfeiffer. 2008. Use of data mining techniques to investigate disease risk classification as a proxy for compromised biosecurity of cattle herds in Wales. BMC Vet. Res. 4: 24. http://dx.doi:10.1186/1746-6148-4-24

Pépin, D., C. Adrados, C. Mann, and G. Janeau. 2004. Assessing real daily distance travelled by ungulates using differential GPS locations. J. Mammal 85: 774-780.

Pluk, A., C. Bahr, A. Poursaberi, W. Maertens, A. Van Nuffel, and D. Berckmans. 2012. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. J. Dairy. Sci. 95: 1738-1748. http://dx.doi.org/10.3168/jds.2011-4547

Quinlan, R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Reader, J. D., M. J. Green, J. Kaler, S. A. Mason, and L. E. Green. 2011. Effect of mobility score on milk yield and activity in dairy cattle. J. Dairy Sci. 94: 5045-5052. http://dx.doi.org/10.3168/jds.2011-4415

Robert, B., B. J. Whit, D. G. Renter, and R. L. Larson. 2009. Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle. Comput. Elect. Ag. 67: 80-84. http://dx.doi.org/10.1016/j.compag.2009.03.002.

Roelofs, J. B., F. J. van Eerdenburg, N. M. Soede, and B. Kemp. 2005. Pedometer readings for estrous detection and as predictor for time of ovulation in dairy cattle. Theriogenology 64:1690-1703. http://dx.doi.org/10.1016/j.theriogenology.2005.04.004

Sepulveda-Varas, P., D. M. Weary, and M. A. G. von Keyserlingk. 2014. Lying behavior and postpartum health status in grazing dairy cows. J. Dairy. Sci, 97: 6334-6343. http://dx.doi.org/10.3168/jds.2014-8357

Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. Weigel, 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. J. Dairy. Sci. 97: 731-742. http://dx.doi.org/10.3168/jds.2013-6693

Silper, B. F., I. Robles, A. M. L. Madureira, T. A. Burnett, M. M. Reis, A. M. de Passillé, and R. L. A. Cerri. 2015. Automated and visual measurements of estrous behavior and their sources of variation in Holstein heifers. I: Walking activity and behavior frequency. Theriogenology. 84: 312-320. http://dx.doi.org/10.1016/j.theriogenology.2014.12.029

Steensels, M., C. Bahr, D. Berckmans, I. Halachmi, A. Antler, and E. Maltz. 2012. Lying patterns of high producing healthy dairy cows after calving in commercial herds as affected by age, environmental conditions and production. Appl. Anim. Behav. Sci. 136: 88-95.

Swain, D. L., T. Wark, and G. J. Bishop-Hurley. 2008. Using high fix rate GPS data to determine the relationships between fix rate, prediction errors and patch selection. Ecol. Modell. 212: 273-279. http://dx.doi.org/10.1016/j.ecolmodel.2007.10.027.

Van Nuffel, A., W. Saeys, B. Sonck, J. Vangeyte, K. C. Mertens, B. De Ketelaere, and S. Van Weyenberg. 2015. Variables of gait inconsistency outperform basic gait Variables in detecting mildly lame cows. Livestock Science. 177: 125-131. http://dx.doi.org/10.1016/j.livsci.2015.04.008

Viazzi, S., C. Bahr, T. Van Hertem, A. Schlageter-Tello, C. E. B. Romanini, I. Halachmi, and D. Berckmans. 2014. Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. Comput. Elect. Ag. 100: 139-147. http://dx.doi.org/10.3168/jds.2012-5806

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2003. Assessment of the welfare of dairy cattle using animal-based measurements: direct observations and investigation of farm records. Vet. Rec. 153: 197-202. http://dx.doi.org/10.1136/vr.153.7.197.

Chapter 2

The following chapter was accepted for publication in the journal Computers and Electronics in Agriculture and is therefore formatted to the requirements of the journal. The reference for the publication is:

Williams, M.L., James, W.P. and Rose, M.T., 2017. Fixed-time data segmentation and behavior classification of pasture-based cattle: Enhancing performance using a hidden Markov model. *Computers and Electronics in Agriculture*, *142*, pp. 585-596.

Summary

The success of models such as that developed in Chapter 1 used in real-time to classify behaviour events as they occur depends on many factors. One of these is how well the model copes with behaviour irregularity. In an ideal scenario, if cows undertake a long period of grazing, followed by a long period of resting, then models can usually identify these behaviours with high accuracy. Indeed, when data are collected to develop these models, the data are often free of irregular behaviour patterns such as very small bouts of grazing and resting (e.g. seconds to minutes) as these do not represent the majority of behaviours. Furthermore, constraints with the technologies and methodologies used often mean that it can be very difficult to model behaviour with such precision which can lead to some errors in the real-time behaviour recognition process. The model developed in Chapter 1 is also open to such anomalies. Therefore, the aim of this chapter was to try and account for some of the errors that can occur when classification models are used to recognize continuous streams of data gathered in real-time. We used a hidden Markov model for this purpose, connected to the classification model developed in Chapter 1. Hidden Markov models use probabilities to decide whether an agent has moved from one behaviour state to another

using estimated probability distributions. These include the probability of a cow transitioning from, grazing to resting for example, and also the probability that a cow is really grazing when sensor data are suggesting that she is grazing. This chapter demonstrates that this technique can correct classification errors and provide a better estimate of cow behaviour. This is a continuing area of research in cow behaviour classification.

Fixed-time data segmentation and behavior classification of pasture-based cattle: Enhancing performance using a hidden Markov model

M. L. Williams,[*] W. P. James, M. T. Rose

Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom

[*]Corresponding author: Manod Lewis Williams. Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom. E-mail: maw90@aber.ac.uk

Highlights

- A hidden Markov model was developed to improve cattle behavior classification.
- The hidden Markov model significantly reduced segment classification errors.
- Classifier and hidden Markov model together can improve cattle behavior prediction.

ABSTRACT

It is often difficult to monitor dairy cow behavior where grazing contributes a significant proportion of dairy cow diets and where cow contact is reduced. We previously developed a behavioral model of the pasture-based dairy cow that requires incoming, transformed GPS data collected from cattle to be partitioned into segments of a fixed length prior to behavioral classification into grazing, resting or walking. However, fixed-time segmentation presents a problem during behavior classification because segment boundaries may not be located precisely at the point of behavioral transition, leading to classification errors. The objective of this work was to try to overcome this problem by statistically correcting the behavioral predictions. This was

achieved using a hidden Markov model trained using 90 h of supervised data gathered from a previously studied cohort of dairy cattle. The statistical probabilities of the behaviors predicted by the classifier being the true (hidden) behaviors exhibited by cows and also the probability of transition between behaviors was used to statistically modify the predicted output sequences from the classifier. Using 51 h of behavior-labelled validation data we report a significant mean improvement in the classification of grazing, resting and walking behaviors of Holstein dairy cattle (overall classification accuracy = 0.85 (CI = 0.83 – 0.87) vs. 0.94 (CI = 0.92 – 0.95)) for the classifier alone and after the application of the hidden Markov model to the predicted behaviors respectively. To further test our combined models, buffer fed, healthy, early lactation (mean ± SD; 43 ± 20.9 DIM) primiparous (n = 12) and multiparous (n = 12) pasture-based Holstein dairy cattle were fitted with a GlobalSat® DG-100 GPS and monitored every other day for 10 days for the proportion of time spent grazing, resting and walking. Over the 10-day observation period, the predicted mean daily duration of grazing, resting and walking for primiparous cows was 344.86 min (CI = 319.04 – 370.68), 752.99 min (CI = 725.25 – 780.74) and 42.15 min (CI = 31.35 – 52.95) respectively. Multiparous cows were predicted to spend on average 392.33 min (CI = 366.51 – 418.16) grazing, 714.19 min (CI = 686.45 – 741.94) resting and 33.48 min (CI = 22.68 – 44.28) walking. These results corroborate other studies that have measured the activity of pasture based-dairy cows and provide confidence in the predictive ability of the combined models.

Keywords: Dairy cattle; Hidden Markov model; Automated measures; Classification; Transition detection

1. INTRODUCTION

It is now well established that automatic technologies can provide valuable information about dairy cow diet selection (Gregorini et al., 2015) and health status (Charlton et al., 2016). Monitoring pasture-based dairy cattle however has historically been very difficult but advances in the capability of data gathering tools such as data loggers and global positioning systems (GPS) now allow for the simultaneous collection of many variables sampled over high temporal resolutions (Bailey et al., 2015). As such, there has recently been a move towards utilizing such high-frequency data for behavioral modelling and inference (González et al., 2015; Williams et al., 2016). On small spatial scales, supervised methods can be used where direct observation of the study population is undertaken to gather behavioral information which can be subsequently used to recognize behaviors in spatio-temporal data (Dutta et al., 2015). We previously developed a rule-based behavioral model of pasture-based cattle (Williams et al., 2016) using a supervised behavioral classification approach with data gathered from GPS receivers set to record spatial and temporal information at a high sample rate (5 s). The model consisted of nine rules, allowing for the classification of the three main behaviors exhibited by pasture-based dairy cattle; grazing, resting and walking.

Fixed-time segmentation of the temporal GPS data was undertaken for training and testing the classifier similarly to Bom et al. (2014). The size of the data segments was determined according to a trade-off between the average duration of the three behaviors at any point in time and the performance of the classifier on data segments of varying size during model optimization. However, fixed-time segmentation can be problematic. Biological variability and environmental factors will dictate the temporal duration of specific behaviors. Therefore, when models are deployed for behavior

prediction, segments are unlikely to be partitioned exactly at the point of behavioral transition. In other words, data segments could contain data represented by two separate behaviors. Thus, for transition detection and classifier error correction we turn to hidden Markov models (HMMs). Traditionally used in speech recognition tasks (Rabiner, 1989) HMMs are now extensively used in animal research for behavioral inference particularly with elusive species which may be difficult to observe (Pedersen et al., 2011; McKellar et al., 2014). In these examples, HMMs are useful in providing information about different modes of animal movement and its consequences at the individual and population level. They have also been demonstrated to be effective in the behavioral inference of agriculturally important species such as cattle (Guo et al., 2009) and sheep (Milone et al., 2009). An HMM requires a set of discrete unobservable states that follow a Markovian property where future states depend only on the current (Rabiner, 1989). These states are interpretable through a series of observations that are emitted through for example data collected at a high frequency such as that gathered from data loggers. Emissions such as step length (distance between successive positional fixes) and turning angle are then used to infer logical behavioral states. Guo et al. (2009) as an example inferred foraging behavior from high angular speed and low directional speed from GPS data.

In this work however, rather than use an HMM as a behavioral inference tool (Guo et al., 2009; Dean et al., 2012) using raw GPS movement metrics, we instead applied an HMM for use in behavior transition detection by connecting it to our classifier. By doing so, behavior emissions produced by the classifier could be statistically modified by the HMM. For example, if a data segment quantitatively indicates that the exhibited behavior is grazing but this segment occurs in the middle of

a sequence of resting segments the HMM will recognize this as a misclassification due to its statistical improbability and correct it.

This paper is organized into two parts. First, we report the performance of our classification and HMM label-correcting method in identifying the behaviors of pasture-based cattle using supervised test datasets. Secondly, we show the results of the deployed models in a simple experiment undertaken to predict the activity budgets of primiparous and multiparous dairy cows, paying particular attention to the results of other published work to contextualize our predictions.

2. MATERIALS AND METHODS

2.1. Behavioral classification

The first step in the behavior recognition process is to produce a series of predicted behavior labels (emissions) to feed into an HMM. The behavior classification step can be broken down into three distinct phases; movement analysis, segmentation & feature extraction and behavior prediction.

1. Movement analysis:

The movement analysis computes the basic information required for feature extraction such as the distance travelled by cows (m), speed (m/s), acceleration (m/s^2) and turning angle (degrees) between contiguous coordinates and speed data. In this work, the GPS sample rate was set to 5 s because this was found to provide the best trade-off between battery power and movement resolution.

2. Segmentation & feature extraction:

Once the movement analysis is complete the entire dataset is broken down into segments of a predetermined size. Here, the size of each segment was set so that each contained 32 movements or contiguously gathered data instances (160 s of data). For example, a GPS receiver set to gather positional coordinates at a 5 s sample rate would gather approximately 17, 280 positional fixes in 24hrs yielding 540 segments. We opted for a 160 s segmentation strategy because it yielded the best classification performance in our previous work (Williams et al., 2016). Once segmented a number of features can be extracted from the data that are derived from the information gathered in the movement analysis. The feature extraction phase gathers more information from the data to help define the decision classes which in this instance are grazing (where the cow is standing and actively ingesting plant material or browsing), resting (where the

cow is either standing stationary or lying and not exhibiting the additional features of grazing) and walking (where the cow is purposefully moving usually with a high degree of directional persistence). For classification, a total of 13 features were extracted from each segment in this work, examples of which include minimum, mean and maximum speed (m/s) and the rate of directional changes per segment. The full list of features extracted from each segment can be found in Williams et al. (2016); Fig. 3.

3. Behavior prediction:

Finally, using all of the extracted features, each data segment is classified into grazing, resting or walking using a rule-based model of cow behavior. The output of the classification process is in fact the emission sequence that subsequently feeds into the HMM which is described in the next section.

2.2. Hidden Markov model

An HMM can be used to infer the underlying hidden state or the true behavior that cannot be seen being expressed by the focus cow and was chosen for its potential to correct classification errors which may arise due to fixed-time data segmentation. Here, the hidden states were grazing (G_S), resting (R_S) and walking (W_S). Three components are required to train an HMM:

1. State transition probabilities:

These are the probability values for the transitions between the hidden states where γij denotes the probability of the cow switching from state i (at any time t) to state j (at time $t + 1$).

2. Emission probabilities:

These are the probability values of the predicted emissions grazing (G_E), resting (R_E) and walking (W_E) given the hidden state. For this work, the emissions are the predicted labels given by the classifier which are a product of the features of the GPS data such as distance travelled (m) and turning angle (degrees). The classifier uses many and often complex combinations of such variables. For example, one sub-component of the full model that can be used to explain a three-minute grazing segment is:

IF segment mean positive acceleration ≥ 3.73E−6 m/s^2 AND maximum distance moved between GPS fix within segment ≤ 2.17 m AND segment mean speed ≥ 0.10 m/s AND segment accumulated distance moving ≥ 11.92m then Behavior = grazing

Therefore, the probability values of the emissions were initialized into the HMM using the true positive performance values of the classifier. For example, grazing was misclassified as resting at a rate of 0.18 by the classifier but was never misclassified as

walking. Therefore, the probability of the hidden state G_S given the emitted label G_E was 0.82. This method meant that reverting to the original, raw movement metrics for behavioral inference was not required as the emissions are in fact a product of a combination of movement metrics.

3. Initial state probabilities:

Which are the initial behavioral probabilities where π_0 (start probability vector) is the probability of starting in state G_S, R_S and W_S.

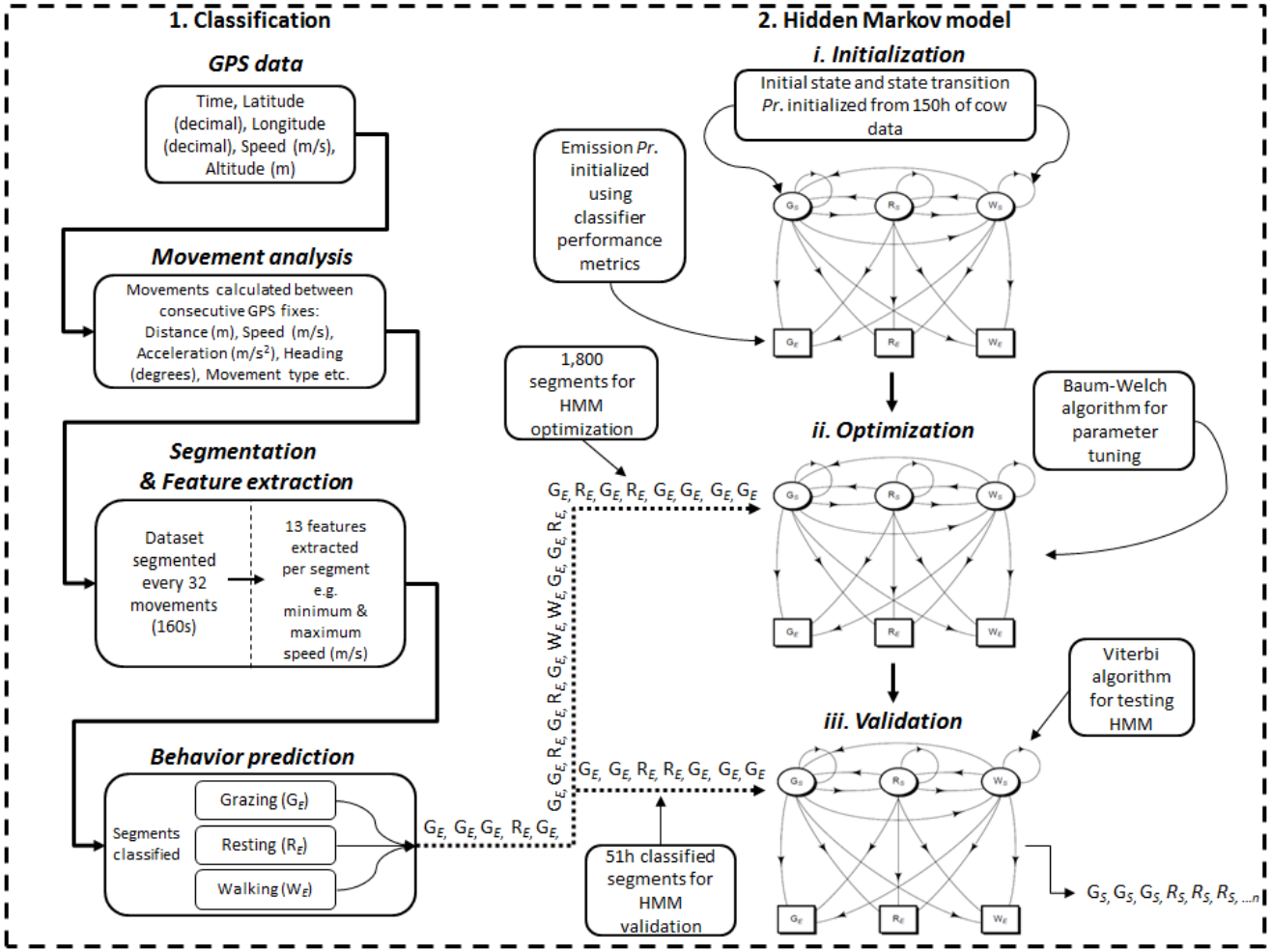


Fig. 1a. Segment classification (1) and hidden Markov model (2) initialization (2.i), optimization (2.ii) and validation (2.iii). G_S, R_S, W_S = Grazing state, Resting state, Walking state. G_E, R_E, W_E = Grazing emission, Resting emission, Walking emission.

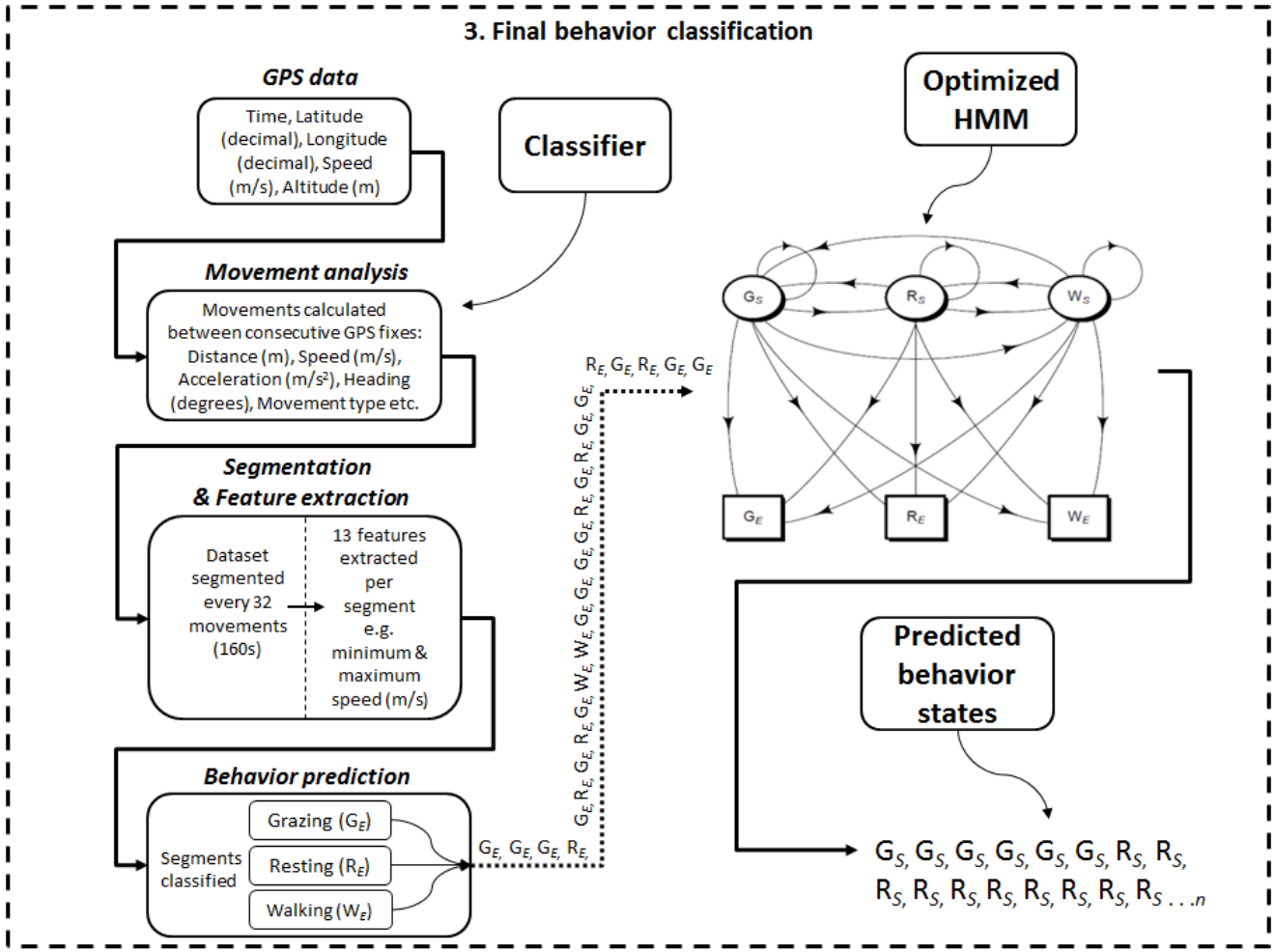


Fig. 1b. Full cattle behavior prediction workflow showing segment classifier and optimized hidden Markov model. G_S, R_S, W_S = Grazing state, Resting state, Walking state. G_E, R_E, W_E = Grazing emission, Resting emission, Walking emission.

2.2.1. Initialization

Initial state and state transition probabilities for the HMM were calculated from 150 h of behavior labelled (supervised) data sampled every 5 s that was used to create the classifier model (Williams et al., 2016). The forty Holstein cows were managed according to a typical dairy grazing regime where cows were given a daily allocation of approximately 2,500 kg DM/ha of pasture comprising of mainly perennial ryegrass (*Lolium perenne*) which was grazed to a residual of 1,500 kg DM/ha. One of the challenges in building a single HMM for use as a statistical label-correcting tool is the potential for over or under compensating for variability in the behavior of cattle. For example, an HMM could be used to re-classify a segment label t to match the next segment $t + 1$ if the statistical probability of t is very low. Such errors, where behavioral states may be intermittent could occur with data gathered from younger cattle where behaviors are perhaps more unpredictable (Kutzer et al., 2015) or if cattle are affected by disease (González et al., 2008), exhibiting reproductive behavior (Dolecheck et al., 2015) or interacting with herd mates (Chebel et al., 2016). This will be considered for future work, however, we considered creating individual HMM's impractical for monitoring large herds of cattle and opted for a single HMM here. Therefore, using data from as many cows as possible to initialize the parameters of the HMM was to reduce the likelihood of over fitting the HMM and to increase the chance of creating a more general model.

2.2.2. Optimization

The transition, emission and initial probability estimates for the HMM were optimized using the Baum-Welch algorithm (Baum et al., 1970) of the HMM package (Himmelmann, 2010) in R (version 3.1.0; R Core Team, 2014). This function requires an initial HMM and a sequence of emissions. For the optimization phase we used all of

the available data from cows tracked from previous work in 2014 and hence there was no specific reason for the selected datasets for optimization other than data exhaustion. Sequences of predicted emissions (for example; $G_E \rightarrow G_E \rightarrow R_E \rightarrow R_E \rightarrow W_E \rightarrow n$) were generated by running our classifier over 20 individual datasets each totaling 270 min of GPS data gathered from 20 randomly selected healthy, Holstein cows that had not been previously used for the parameter initialization phase. This resulted in a total of 1,800 emitted segments (behaviors) passed to the Baum-Welch algorithm in sequences of 90 segments as output from the prediction model. The Baum-Welch algorithm iteratively adjusts the original HMM model parameters to maximize the probability of obtaining the emissions fed to the algorithm (Rabiner, 1989). After each run of the Baum-Welch algorithm on each sequence of segments the transition and emission probability estimates of the HMM were re-estimated until no further improvement in the model parameters were reached. This was repeated until all sequences had been processed. We then used the Viterbi algorithm from the same package to test the HMM on sequences of emissions from data with known behavioral states. Using the optimized probability estimates, the Viterbi algorithm computes the hidden state sequence that best fits the sequence of emissions (Forney, 1973), here provided by the classifier.

2.2.3. Validation

To test our HMM we used a total of 51 h of behavior-labelled GPS data previously gathered from 9 randomly selected healthy Holstein dairy cows that were between parities 1 and 8 at Aberystwyth University dairy farm, Trawsgoed, Ceredigion, between March and August 2014. The data used here were not previously used to initialize or optimize the parameters of the HMM and this was the total amount of validation data available. Each cow had been directly observed by two trained observers situated approximately 40 m from the observational paddock for the

behaviors grazing, resting and walking (Kappa = 0.96, 0.99 and 1.00 respectively) according to the methodology of Williams et al. (2016). Data from each cow were downloaded as comma separated variable (csv) files and saved in a spreadsheet program. Each dataset contained the record number of each logged position, time (s), latitude, longitude, speed (m/s), and altitude (m). Individual datasets from each cow were arranged so that a single continuous dataset of 32,783 observations taken at 5 second intervals was compiled for segmentation, feature extraction and classification. Each cow contributed 5 hours of behavior-labelled data equating to 3,600 ± 14 (mean ± SD) observations and therefore 112 ± 0.81 (mean ± SD) three-minute segments for classification. Datasets were passed to the behavioral classifier for labelling before being transferred for statistical correction by the HMM. The entire workflow is summarized in Figs. 1a and 1b.

The prediction performance of both methods was evaluated using overall classification accuracy ((TP + TN) / (TP + TN + FP + FN) where TP = true positives, TN = true negatives, FP = false positives, FN = false negatives); sensitivity (TP / (TP + FN)); specificity (TN / (FP + TN)); F1 (2TP / (2TP + FP + FN)); precision (TP / (TP + FP)); and balanced accuracy (TP / (TP + FN)) + (TN / (TN + FP)) / 2. Balanced accuracy was included as it is intuitively simple where predictive performance is measured independently for each class and aggregated. Furthermore, where class distribution is unbalanced considering only overall classification accuracy can be misleading (Brodersen et al., 2010). To compare the sensitivity, specificity and classification accuracies of both classification methods the resulting classification matrices from the validation phase (containing the classifier predictions versus the actual behaviors) were combined to form a 2 x 2 confusion matrix for statistical comparison. The resulting 2 x 2 matrix contains 1; the number of instances that were

correctly classified by the classifier and also by the combined classifier and HMM (bottom right cell of the matrix) 2; the number of instances that were correctly classified by the classifier but the classifier and HMM failed to correctly identify (bottom left cell of the matrix) 3; the number of instances that the classifier alone failed to identify correctly but the classifier and HMM correctly identified (top right cell of the matrix) and 4; the number of instances that the classifier failed to correctly identify and that the classifier and HMM also failed to identify (top left cell of the matrix). For this comparison a McNemar's test was undertaken (McNemar, 1947; Kim and Lee, 2017).

2.3. Cattle activity analysis

Finally, we used our classification and HMM to predict the duration of grazing, resting and walking exhibited by two groups of cows in a commercial dairy herd. We chose to compare the activities of primiparous (n = 12) and multiparous cows (n = 12). Previous authors (Sepúlveda-Varas et al., 2014; Westin et al., 2016) reported differences in the behavioral frequencies of primiparous and multiparous cows and it was therefore decided that such a comparison could yield useful information on the predictive ability of our models. GPS data were collected from 24 healthy Holstein dairy cows in a longitudinal study during August 2016 at the same farm. The sample size was dictated by the number of GPS units available each day. At the start of the experiment the mean (± SD) number of days since calving was 43 (± 22.67) and 43 (± 19.12) for primiparous and multiparous cows respectively. Data collection took place over a period of 20 days. Study cows were fitted around the neck with a randomly selected GPS receiver from a sample of 24 units that had been rigorously field tested under the specific guidelines of the Institute of Navigation (ION, 1997). Collars were placed onto cows every other day at approximately 06:30 h. GPS collars were left in

place to collect data for a period of 24hrs, logging positional coordinates at a 5 s sample rate. Immediately after placing the collars on the cows, the study cows returned to the main group of 120 cows where the entire group was buffer fed a total mixed ration of grass silage (25 kg of total fresh weight (FW)), maize silage (10 kg FW), rolled wheat (3 kg FW), barley straw (0.3 kg FW) with the remainder comprising a rumen-protected fat supplement, dairy mineral blend and molasses prior to returning to the designated grazing paddock. Buffer feeding was being used by the farm in order to compensate for the shortfall in grass availability. All cows were returned to the designated grazing paddock by 08:00 h and were removed from the paddock 8 h later at 16:00 h for the afternoon milking. On arrival back at the farm in the afternoon, cows were offered the same buffer silage ration before and after milking. Cows were returned to the same grazing paddock by 18:00 h where they spent the following 11hrs. GPS collars were removed from the study cows for processing during the subsequent morning milking period and replaced 24hrs later. Each 24hr data collection period yielded 19hrs of GPS data from each grazing paddock for behavioral analysis from each study cow. All cows were allocated a new grazing paddock at the start of each day which was part of the routine herd management. Cows were set stocked in each grazing paddock which provided approximately 2,000 kg DM/ha of *Lolium perenne*. Pasture quality was not managed stringently during the experimental period and so varied throughout each paddock. Given the nature of the data retrieval process, 20 consecutive experimental days yielded 10 days of data from ten different grazing paddocks for predictive analysis. In developing a similar predictive model Guo et al. (2009) collected four consecutive days of cow data whereas Alsaaod et al. (2012) recommended the collection of approximately 14 days of consecutive data for classifying lame cattle.

Here, we deemed that 10 days of data would provide a realistic representation of the behavior of cattle given the grazing rotation at the farm.

2.4. Data handling

Data from each cow for the 10 days of data collection were downloaded as csv files before applying our behavioral prediction methodology to each dataset from each cow (Figs. 1a and 1b). Time spent (minutes) in each behavioral category (grazing, resting and walking) was computed for each cow on each day. In total, there were 16 missing cow days of data (6% of total) in the full dataset where GPS units had failed either due to impact or water penetration. To deal with missing instances, we separately regressed the time spent in each behavioral category on all experimental days (10d) for each cow with missing values. Missing values were then replaced by values calculated for the missing day from the coefficients of each regression model. A similar method was undertaken by Thorup et al. (2016) to deal with missing data when estimating the feed intake of cattle.

2.5. Statistical analysis

All analyses were performed using R (R Core Team, 2014). Descriptive statistics were first computed for the raw behavioral predictions of the daily time spent in each behavioral category for both primiparous and multiparous cows combined for the duration of the experiment. Hourly time budgets were also calculated as means of the proportion of time spent in each behavioral category for both groups over the 10-day period. Time budgets were created from available GPS records and did not include any data where missing values had been computed previously. The activities of the two groups of cows over the course of the 10 days were then analyzed. The incidence of the variable walking was highly positively skewed and so data were transformed using the natural log + 1 to meet the assumptions of normality prior to modelling for this

behavior. Data for both grazing and resting were found to meet the assumptions of normality. We used the lme4 package in R for linear mixed effects analysis (Bates et al., 2007). In an initial screening of the effects under study, we constructed three linear mixed effects models for parity (1), parity + day (2) and parity + day + parity x day (3) using cow as the random effect in each model and tested these on the daily duration of grazing, resting and walking. The full model with interaction was found to fit the data best ($P < 0.001$) for grazing and resting whereas no improvement in fit was seen over the main effects model (parity + day) for the behavior walking ($P = 0.63$). Results are presented as least squares means with standard errors for each behavior.

3. RESULTS

3.1. Estimated model parameters

The final state transition and emission probability matrices as well as the initial probability values for the trained HMM are shown in Table 1. As an example of the transition probabilities, it can be seen that the probability of a cow remaining in a state of grazing (G_S) at time $t + 1$ is 0.953. If a cow is in a state of resting (R_S) the probability of her being in a state of grazing (G_S) at $t + 1$ is 0.03. In the classifier building phase there was difficulty in distinguishing between lying and standing (where a cow is stood still) behaviors, which is why a single category 'resting' was created. Because 98% of the original training data for the classifier model contained instances where the cow was physically lying this explains why the probability of a transition between resting and walking ($R_S \rightarrow W_S$) and $W_S \rightarrow R_S$ is 0. In other words, the probability that a cow will rise and immediately start walking or transfer to a lying position immediately from walking is very low and in the HMM parameter optimization phase this became zero. For the emission probability values, the probability of a cow being in a real state of grazing (G_S) when the classifier predicted (emitted) grazing (G_E) was 0.799 whereas if walking (W_E) was emitted the probability that the cow was in fact in R_S was 0.004. Objects can appear to be moving even when GPS receivers are stationary which can be explained by receiver accuracy and multipath effects (Ryan et al., 2004; Ganskopp and Johnson, 2007). The initial probability values represented the probability of finding a cow in each behavioral state at the start of a dataset. Cows were most likely (0.80) to be in an initial state of G_S.

Table 1. Complete hidden Markov model for cattle behavioral analysis.

State transition probability	G_{S+1}[b]	R_{S+1}	W_{S+1}
G_S[a]	0.9529	0.0151	0.0321
R_S	0.0301	0.9699	0.0000
W_S	0.0249	0.0000	0.9751
Emission probability	G_E[c]	R_E	W_E
G_S	0.7989	0.2011	0.0000
R_S	0.2010	0.7950	0.0040
W_S	0.0000	0.0197	0.9803
Initial probability			
π[d]	0.8000	0.1500	0.0500

[a] G_S = the cow is in a true state of grazing (R_S = resting state; W_S = walking state) at time t.
[b] G_{S+1} = the cow will be in a state of grazing in the next segment at time $t + 1$.
[c] G_E = the predicted emission is grazing (R_E = resting emission; W_E = walking emission).
[d] π = initial probability of behavioral states.

3.2. Prediction performance

Table 2 shows the resulting confusion matrices generated from the validation phase showing the performance of the classifier and the combined classifier and HMM. It can be seen that the number of correctly identified segments increased with the addition of the HMM for grazing (+ 46 correctly identified segments) and resting (+ 47). No improvement in prediction performance was seen for walking behavior however (71 segments correctly identified by both techniques). This may have been due to the high classification accuracy of the prediction model for this behavior and that cows were highly likely to remain in a state of walking in the next segment (*Pr*: 0.98; Table 1) which is then undisputed as a prediction by the HMM.

Table 2. Confusion matrices generated from the validation dataset showing the classification performance of the classifier and the combined classifier and hidden Markov model versus the true behavioral segments in the dataset.

	Prediction method Classifier			
True behavior	Grazing	Resting	Walking	∑
Grazing	288	88	3	379
Resting	62	510	1	573
Walking	0	1	71	72
∑	350	599	75	1024
	Classifier + hidden Markov model			
Grazing	334	42	3	379
Resting	15	557	1	573
Walking	1	0	71	72
∑	350	599	75	1024

Table 3 shows the performance of the classifier and associated 95% CI's and also performance after the addition of the HMM to the predicted emissions on the compiled validation dataset for the behaviors grazing, resting and walking. There was an improvement in average performance across all of the measured parameters (sensitivity, specificity, precision, F1, balanced accuracy) for grazing and resting behaviors with the use of the HMM. The results for walking corroborate that seen in Table 2 in that no improvement in these parameters were seen for this behavior using the HMM. To test whether the overall improvement seen by using the HMM was statistically meaningful a 2 x 2 confusion matrix was created from the data provided in Table 2. Table 4 shows the resulting 2 x 2 matrix with the number of correct and incorrect classifications by the classifier and combined classifier and HMM. Overall, the McNemar's test revealed that the HMM significantly improved the identification of the behavioral labels of the segmented validation dataset compared to the classifier alone ($\chi^2 = 88.1$; $P < 0.001$).

Table 3. Performance (and 95% CI) of the classifier on the validation dataset versus the combined performance of the classifier and hidden Markov model.

Prediction method	Behavior	Sensitivity	Specificity	Precision	F1	Balanced accuracy	Overall accuracy
Classifier	Grazing	0.76 (0.73 – 0.79)	0.90 (0.88 – 0.92)	0.82 (0.80 – 0.85)	0.79 (0.76 – 0.82)	0.83 (0.81 – 0.85)	NA
	Resting	0.89 (0.87 – 0.91)	0.80 (0.77 – 0.83)	0.85 (0.83 – 0.88)	0.87 (0.85 – 0.89)	0.85 (0.83 – 0.87)	NA
	Walking	0.99 (0.98 – 0.99)	1.00 (0.99 – 1.00)	0.95 (0.93 – 0.96)	0.97 (0.95 – 0.98)	0.99 (0.98 – 1.00)	NA
Average		0.88 (0.86 – 0.90)	0.90 (0.88 – 0.92)	0.87 (0.85 – 0.89)	0.88 (0.86 – 0.90)	0.89 (0.87 – 0.91)	0.85 (0.83 – 0.87)
Classifier + hidden Markov model	Grazing	0.88 (0.86 – 0.90)	0.98 (0.97 – 0.98)	0.95 (0.94 – 0.97)	0.92 (0.90 – 0.93)	0.93 (0.91 – 0.95)	NA
	Resting	0.97 (0.96 – 0.98)	0.91 (0.89 – 0.92)	0.93 (0.91 – 0.95)	0.95 (0.94 – 0.96)	0.94 (0.92 – 0.96)	NA
	Walking	0.99 (0.98 – 0.99)	1.00 (0.99 – 1.00)	0.95 (0.93 – 0.96)	0.97 (0.95 – 0.98)	0.99 (0.98 – 1.00)	NA
Average		0.93 (0.91 – 0.95)	0.96 (0.95 – 0.97)	0.94 (0.92 – 0.96)	0.95 (0.94 – 0.96)	0.95 (0.94 – 0.96)	0.94 (0.92 – 0.95)

Sensitivity = true positive rate; Specificity = true negative rate; Precision = positive predictive value; F1 = harmonic mean of precision and sensitivity; Balanced accuracy = sensitivity + specificity / 2; Overall accuracy = fraction of correctly classified instances.

Table 4. Confusion matrix showing the observed number of correct and incorrect outcomes for the classifier and classifier and hidden Markov model combined.

Allocation	Classifier + hidden Markov model			
	Incorrect	Correct	∑	McNemar's test χ^2 statistic and *P*-value
Classifier				
Incorrect	61	93	154	
Correct	1	869	870	
∑	62	962	1024	$\chi^2 = 88.1$; $P<0.001$

To visually demonstrate the effect of the classifier and HMM in predicting the behaviors of cattle, we randomly selected two 5 h sequences of data from the compiled dataset represented by two separate cows and plotted their behaviors. Fig. 2 shows two example path plots produced by two cows that exhibited all three behaviors under study. It can be seen that grazing intensity was highest in the northern regions of these plots where the fresh strip of pasture was located. Evidently, both cows made a single trip to the water source which was located in the southerly region before returning to graze and finally rest. Each data point within each path plot corresponds to a single observation sampled every 5 s (32 observations = 1 segment). Plots (a) show the efficacy of the classifier alone in predicting the behaviors. Plots (b) and (c) respectively illustrate the predicted behaviors after the application of the HMM to the classified segments and the visually verified behaviors. Series 1(a) shows that 79% of the segments were correctly identified by the classifier (73 out of 93 segments). Application of the HMM to this dataset resulted in all segments being correctly classified (1b vs. 1c). The second series in Fig. 2 shows the path plot of the second cow. In this example the classifier alone (2a) correctly identified 83% of the segments (77 out of 93 segments) while the HMM improved the classified labels to 97% (total = 90 out of 93 segments correct). The majority of misclassifications for the classifier during

model construction occurred between the behaviors grazing and resting. This can be seen in both series 1 and 2 (plot a) where some grazing instances were labelled as resting. This usually occurs where the motion of grazing takes place very slowly. These segments are usually found in amongst grazing segments and the HMM corrects these segments due to their low statistical probability of occurrence. However, for series 2, the HMM misidentified three grazing segments as resting.

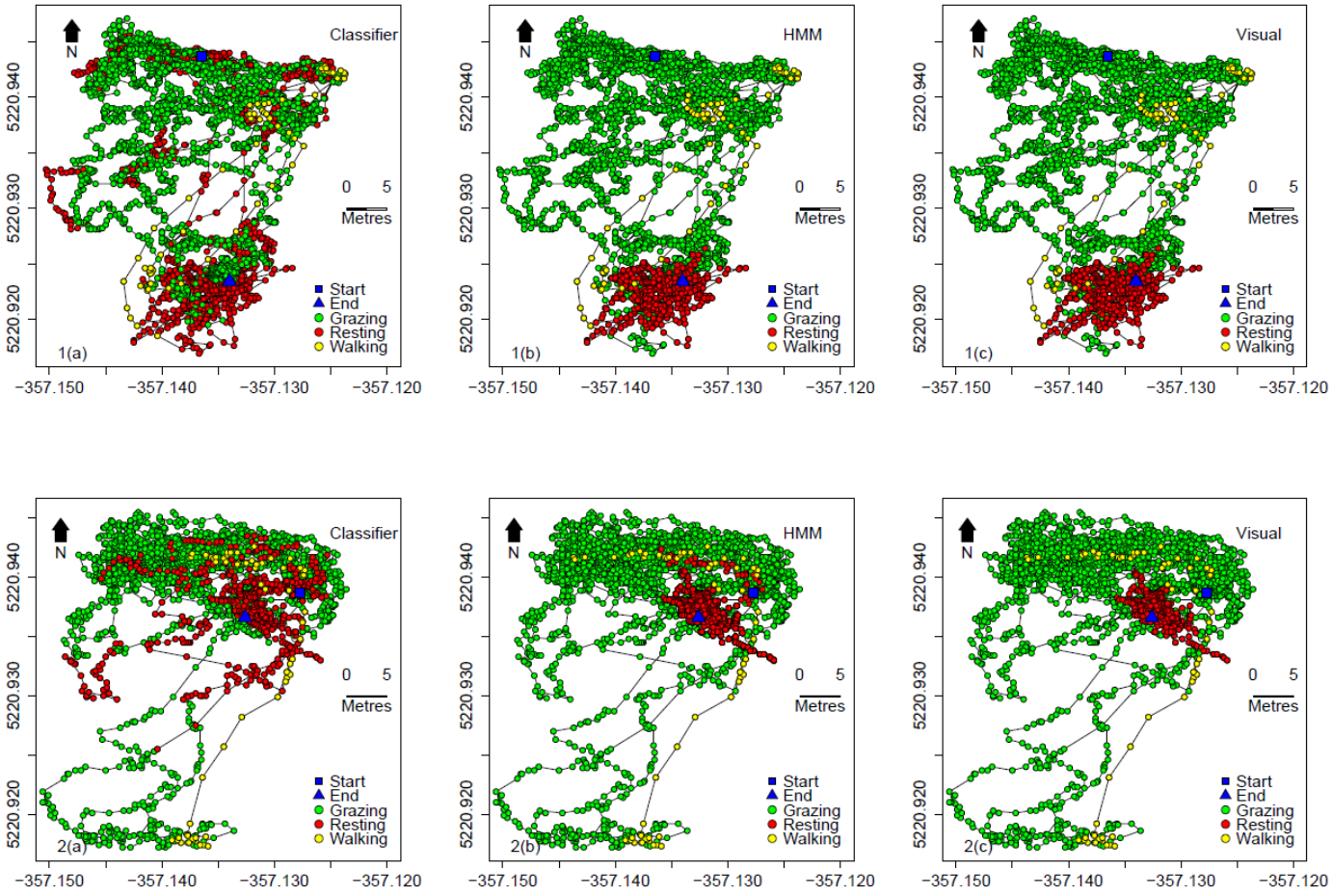


Fig. 2. Two five-hour path plots of the behaviors of two cows as predicted by the classifier model (a), after smoothing the predictions with the hidden Markov model (b) and the visually verified behaviors (c). Each data point corresponds to a GPS coordinate sampled at 5 s intervals.

3.3. Group activity analysis

For the 10-day experimental period, all cows (n = 24; 240 observations) were predicted to have spent a median duration of 357 min/d grazing (25^{th} – 75^{th} percentile = 294 – 448 min/d), 738 min/d resting (25^{th} – 75^{th} percentile = 656 – 822 min) and 24 min/d walking (25^{th} – 75^{th} percentile = 9 – 48 min). Fig. 3 shows the variation found in all three behavioral categories across the experimental period for all cows. Fig. 4 shows the predicted hourly allocation of behavior averaged over the whole 10-day experimental period for both study groups. Between the hours of 08:00 h and 16:00h (am period) and 18:00 h and 05:00 h (pm period) it was predicted that primiparous and multiparous cows respectively allocated on average 37% and 40% of their time to grazing during the am period and 22% and 25% during the pm period. It was predicted that primiparous and multiparous cows proportionally allocated 58% and 54% of their time to resting during the am period and 74% and 71% during the pm period. Walking was represented in the smallest proportion with primiparous and multiparous cows predicted to have allocated 6% of their time to this behavior during the am period and 4% during the pm period.

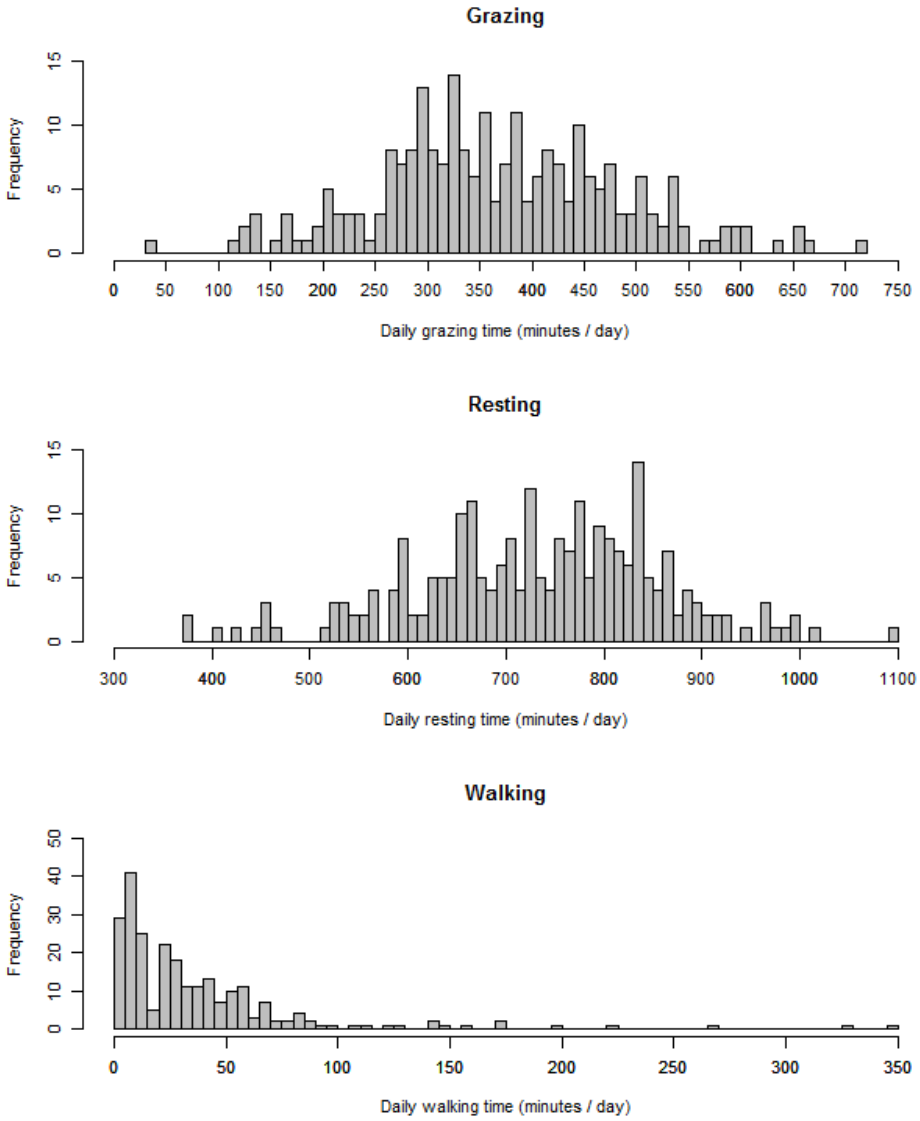


Fig. 3. Frequency distribution of the daily duration (min/d) of grazing, resting and walking of 24 cows as predicted from 10 days of experimental data (n = 240 observations).

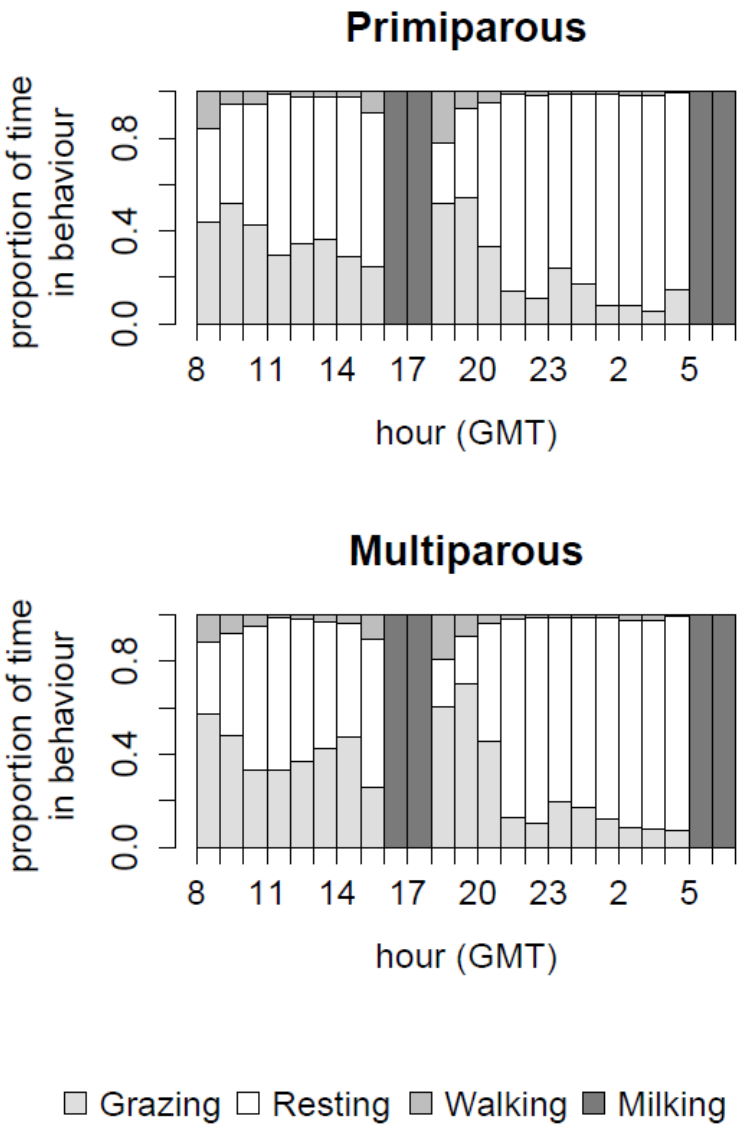


Fig. 4. Proportion of time spent grazing, resting and walking by multiparous and primiparous cows at pasture as predicted from 10 days of experimental data. Means for each cow were taken over the 10 days and are expressed as a grand mean for both groups.

With reference to the predicted behaviors, the mixed effects analysis showed that multiparous cows grazed on average 48 min/d longer ($P = 0.01$) over the course of the 10-day observation period and rested for approximately 39 min/d less ($P = 0.05$, Table 5). There was no significant difference in the daily duration of walking between the two groups (Table 5). There was a significant effect of day on the duration of grazing and resting compared to the first day of observation (Table 5) but these differences could not be reconciled under the experimental conditions. There was no significant effect of day however on the duration of walking over the 10 days compared to the first day of observation (Table 5). There was a significant interaction between parity and day with primiparous cows showing significant day on day variation in the daily duration of grazing (SD = 86.83 min/d; $P < 0.001$) and resting (SD = 85.36min/d; $P < 0.05$). No such variation was found for the multiparous group for grazing (SD = 38.51 min/d; $P > 0.05$) or resting (SD = 43.07 min/d; $P > 0.05$).

Table 5. Results of linear mixed-effects models on the daily duration of grazing, resting and walking (min/d) by primiparous and multiparous cows over the 10-day experimental period.

		Grazing				Resting				Walking[b]			
Variable	Level	Coefficient	SE	95% CI	*P* –value	Coefficient	SE	95% CI	*P* –value	Coefficient	SE	95% CI	*P* –value
Parity	Primiparous	Reference[a]	–	–	–	Reference	–	–	–	Reference	–	–	–
	Multiparous	47.48	12.57	22.84 – 72.12	0.01	-38.8	13.50	-65.26 – -12.34	0.05	-0.43	0.19	-0.80 – -0.06	0.12
Day	1	Reference	–	–	–	Reference	–	–	–	Reference	–	–	–
	2	-63.38	19.60	-101.80 – -24.96	0.31	101.92	22.29	58.23 – 145.61	0.03	-0.69	0.23	-1.14 – -0.24	0.28
	3	64.75	19.60	26.33 – 103.17	0.28	-51.00	22.29	-94.69 – -7.31	0.80	0.08	0.23	-0.37 – 0.53	1.00
	4	10.29	19.60	-28.13 – 48.71	1.00	29.46	22.29	-14.23 – 73.15	0.99	-0.78	0.23	-1.23 – -0.33	0.14
	5	22.75	19.60	-15.67 – 61.17	0.98	-21.21	22.29	-64.90 – 22.48	0.99	0.14	0.23	-0.31 – 0.59	1.00
	6	109.58	19.60	71.16 – 148.00	0.002	-77.50	22.29	-121.19 – -33.81	0.23	-0.30	0.23	-0.75 – 0.15	0.99
	7	104.13	19.60	65.71 – 142.55	0.003	-68.21	22.29	-111.90 – -24.52	0.41	-0.61	0.23	-1.06 – -0.16	0.48
	8	-3.29	19.60	-41.71 – 35.13	1.00	27.71	22.29	-15.98 – 71.40	0.99	-0.29	0.23	-0.74 – 0.16	0.99
	9	73.25	19.60	34.83 – 111.67	0.14	-66.33	22.29	-110.02 – -22.64	0.45	0.11	0.23	-0.34 – 0.56	1.00
	10	107.88	19.60	69.46 – 146.30	0.002	-71.00	22.29	-114.69 – -27.31	0.35	-0.75	0.23	-1.20 – -0.30	0.18
Intercept		365.42	26.53	313.42 – 417.42	<0.001	736.00	30.18	676.85 – 795.15	<0.001	3.17	0.26	2.66 – 3.68	<0.001

[a] Reference level; [b] Log transformed values.

Between parity groups (Fig. 5), a highly significant difference was found in the duration of grazing ($P < 0.001$) on day 2 with resting tending to differ significantly between the two groups on this day ($P = 0.09$). Again, this effect could not be reconciled under the experimental conditions. On average, it was predicted that primiparous cows spent 345 min/d, 753 min/d and 42 min/d grazing, resting and walking respectively. Multiparous cows on the other hand were predicted to have spent on average 392 min/d, 714 min/d and 34 min/day performing the same behaviors respectively (Table 6).

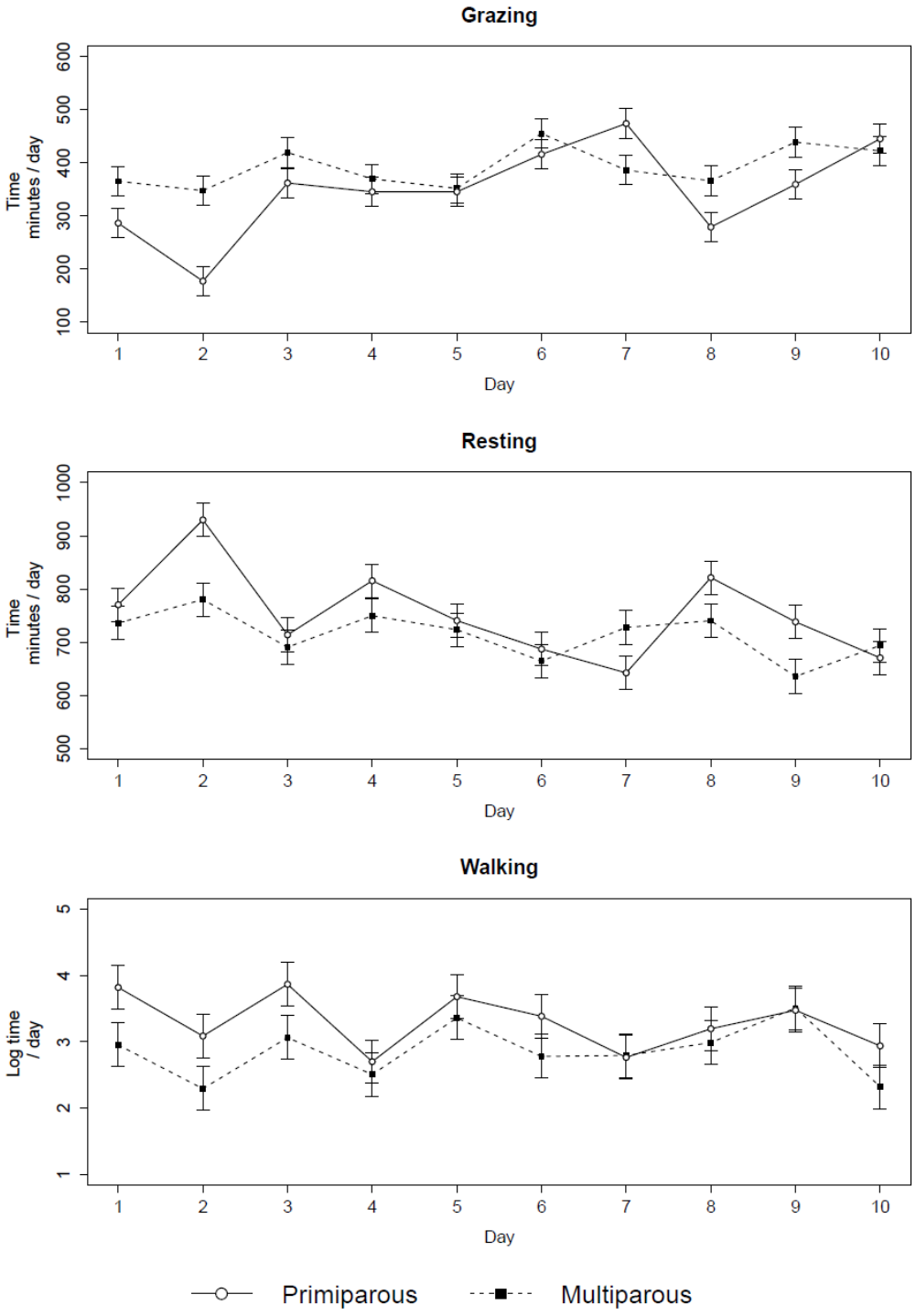


Fig. 5. Least squares means of daily duration (min/d) of grazing, resting and walking activities of primiparous and multiparous cows over the 10-day experimental period.

Table 6. Least squares means (and 95% CI) of time spent (min/d) in each behavior by primiparous and multiparous cows over the 10–day experimental period.

Variable	Primiparous	Multiparous	*P* – value
Grazing (min/d)[a]	344.86	392.33	0.01
	(319.04 – 370.68)	(366.51 – 418.16)	
Resting (min/d)[a]	752.99	714.19	0.05
	(725.25 – 780.74)	(686.45 – 741.94)	
Walking (min/d)[b]	42.15	33.48	0.12
	(31.35 – 52.95)	(22.68 – 44.28)	

[a] Values are from mixed effects models accounting for the effect of day, the interaction between parity and day and the random effect of cow.
[b] Values are from mixed effects models accounting for the effect of day and parity and the random effect of cow. Values are back transformed where the natural log + 1 transformation was used.

4. DISCUSSION

4.1. Behavior recognition

The application of an HMM to the predicted output of our behavioral model of pasture-based cattle allowed us to significantly improve our experimental predictions of cattle activity. Initializing the HMM using the state and emission probabilities gathered from the performance of the original behavioral prediction model meant that reverting to using untransformed GPS movement metrics such as velocity or turning angles to train the HMM was not necessary. This approach to the behavioral prediction problem meant that the sequences of emissions as predicted by the classifier alone could be statistically corrected using the HMM. The limitation of classifying segmented data alone is that each segment is treated independently of the next and does not consider the sequence of behavior-labelled segments as a whole. The HMM in this instance was useful in improving the outcome of the classification task by contextualizing the behavioral sequences, taking a probabilistic approach to the likelihood of occurrence of each classified label. This led to a mean improvement in the overall classification accuracy of 9%. This level of performance is comparable to others that have had success in classifying the behaviors of cattle. For example, using data gathered from a combined accelerometer and magnetometer and an ensemble classifier Dutta et al. (2015) achieved an average classification accuracy of 96% (range 92% – 98%) for the behaviors grazing, searching, ruminating, resting and scratching. The improvement in classification in this work was only recognized for the behaviors grazing and resting where classification errors were more likely to occur. However, the test data revealed that the HMM was not completely effective in statistically modifying the predicted grazing and resting labels of the classifier. On occasion, emitted sequences may be intermittent, and the exact point of transition may be detected earlier or later by the

HMM when the statistical likelihood of the transition is met. This is problematic for the methodology and refining the HMM to be robust to such anomalies is challenging. One solution to this would be to train the HMM on a greater number of behavioral sequences from a greater number of cattle than that used here in order to better generalize the HMM and to incorporate the biological variability of individuals. Foraging behavior of cattle for example has been shown to vary greatly in other work that has explored the use of an HMM to infer the behavior of cattle (Guo et al., 2009). A second method could be to tailor individual HMM to groups of cattle, for example, in order to account for the variation in behavior between age or health status (González et al., 2008; Kutzer et al., 2015).

The behavior walking was distinct in that it could be identified by high directional persistency and velocity by the classification model and the final HMM matrix reflected this. For example, the likelihood that a cow would remain in a state of walking in the next segment was very likely (0.98). Thus, the final classification of walking segments using the HMM was well recognized. Imperfections for this behavior may have been due to erroneous GPS fixes where cows appeared to move very quickly over a particular area. Coupled with the high probability of an observed walking segment being the true state (0.98) this meant that the HMM made adjustments to the behaviors flanking each side of the predicted walking segment rather than adjusting the walking segment itself. Further refinement of the HMM would help in reducing this problem. The high behavioral state probabilities of the HMM could however be seen as a reflection of the way in which contemporary dairy cattle behave. Using automated techniques to identify activities immediately limits the user in the number of behaviors that can be identified. However, it is well documented, that given the energy demand of dairy cattle for functional maintenance and milk production that the main behaviors

exhibited can be partitioned into the three under study with grazing and resting occupying the majority of the time budget (Kilgour, 2012).

4.2. Time budgets

To contextualize and test our models, we created a simple longitudinal experiment to study the activities of two sets of cattle grouped by parity. Our predictions corroborate other work in that the time allocated to grazing, resting and walking each day varies greatly between cows and herds (Gomez and Cook, 2010; Westin et al., 2016; Solano et al., 2016). Here we reported a median daily grazing time over all cows of 357 min/d (mean = 369 min/d). Arachchige et al. (2013) reported a mean grazing time of 312 min/d when cows were supplemented with silage at pasture. Williams et al. (2006) reported a grazing time of between 384 and 426 min/d when cows were supplemented with cereal grain and pasture hay. On the other hand, cows not supplemented at pasture grazed for a mean daily duration of 438 min/d in the study by Soca et al. (2014) and for 527 min/d in the study of Dohme–Meier et al. (2014). Supplementation likely had an impact on the amount of time cows spent grazing when at pasture in the current study. This was reflected in the time spent resting where cows were predicted to have spent a median duration of 738 min/d at rest (mean = 734 min/d). On average, the greatest proportion of cows housed in tie-stalls in the study by Charlton et al. (2016) rested for 780 min/d whereas Solano et al. (2016) found that on average, cows spent 649 min/d resting. Cows housed on farms with automatic milking systems rested for a median daily duration of 684 min/d (25^{th} – 75^{th} percentile = 582 – 774 min/d) in the study by Westin et al. (2016). Both lying and feeding behavior have been used to indicate cow comfort and changes in the health and welfare status of cows (O'Driscoll et al., 2008; González et al., 2008; Charlton et al., 2016) but relatively little work has been focused explicitly on measuring the resting times of cattle at pasture.

However, Hernandez-Mendo et al. (2007) found that cows on pasture spent less time resting compared to cows housed indoors (654 min/d vs. 732 min/d). At pasture, the time required for searching and consuming forage will depend on its accessibility and quality (Gregorini, 2012) which will ultimately govern the time that cattle are able to allocate to other behaviors such as resting. However, the activity results of this study are comparable with others in the literature.

An analysis of the time spent in each behavioral category per hour revealed the distinct diurnal grazing pattern that is well documented for cattle (Gregorini, 2012; Sheahan et al., 2013). Both primiparous and multiparous groups spent the majority of their time grazing around sunrise and sunset with major bouts of grazing occurring after both milking periods. A small peak of grazing occurred at midnight with the majority of time thereafter allocated to resting prior to the morning milking period. We found that primiparous cows spent more time resting compared to multiparous cows. This differs to the study of Sepúlveda-Varas et al. (2014) who found that primiparous cows for the first 15 days after calving spent less time lying per day compared to multiparous cows. However, primiparous cows gradually increased their lying times thereafter, until there was no difference in lying duration between both groups. This has also been seen in other studies examining lying duration (Nielsen et al., 2000; Vasseur et al., 2013). Proportionally therefore, multiparous cows grazed for longer overall in the present study but did not significantly change the amount of time allocated to grazing and resting over the 10-day observation period. The daily behavior of primiparous cows however varied significantly between days. Although there was no direct control over pasture quality and field size during this study, we hypothesize that the daily variation shown in grazing and resting behaviors of the primiparous group could be explained by lack of grazing experience. Speculatively, older cows may have been dedicating more

time towards grazing the best areas of each paddock while the younger cows simply went to rest.

5. CONCLUSIONS

The use of an HMM to probabilistically recognize behavioral change points in the output of our behavioral classifier was successful in improving the accuracy of the predicted behaviors. This improvement was only realized however for the behaviors grazing and resting but highlighted the benefit and computational efficiency of using an HMM to contextualize sequences of behavior labels. By comparing our results with the wider literature, we were able to identify that our predictions were in close association with the most recent studies that measured the duration of time spent by healthy cows in each of the studied behaviors. Furthermore, we found that our predictions coincided with the documented diurnal behavior patterns of cattle. On occasion however, GPS multipath effects of grazing near woodland meant that cattle would appear to move very quickly which the classifier identified as walking (high speed and high directional persistency). As a consequence, the HMM would fail to recognize this as a misclassification because the likelihood of a cow exhibiting walking behavior following grazing is very high. Eliminating this type of error however would require the refinement of the classifier itself to deal with such multipath effects which would ultimately lead to an improved sequence of emissions being fed to the HMM. This is important moving forward because gaining an accurate estimation of the exhibited proportion of each behavior is vital in the context of precision livestock husbandry.

This method provides an efficient solution to fixed-time data segmentation and classification, eliminating the need for separate change point detection algorithms for segmentation. The results also provide confidence in the combined use of a classifier and HMM and that data gathered from GPS receivers alone can be used effectively to track bovine spatio-temporal behavior in both precision livestock husbandry and behavioral research contexts.

ACKNOWLEDGEMENTS

The authors would like to thank the farm staff at Aberystwyth University Research Farm for their cooperation and Emma Davies for assisting with cow observation. The sponsorship of M. Williams by Coleg Cymraeg Cenedlaethol is gratefully acknowledged.

REFERENCES

Alsaaod, M., C. Römer, J. Kleinmanns, K. Hendriksen, S. Rose-Meierhöfer, L. Plümer, and W. Büscher. 2012. Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. Appl. Anim. Behav. Sci. 142: 134-141. https://doi.org/10.1016/j.applanim.2012.10.001.

Arachchige, A. D., A. D. Fisher, M. J. Auldist, W. J. Wales, and E. C. Jongman. 2013. Effects of different systems of feeding supplements on time budgets of cows grazing restricted pasture allowances. Appl. Anim. Behav. Sci. 148: 13-20. http://dx.doi.org/10.1016/j.applanim.2013.08.002.

Bailey, D. W., M. B. Stephenson, and M. Pittarello. 2015. Effect of terrain heterogeneity on feeding site selection and livestock movement patterns. Anim. Prod. Sci. 55: 298-308. http://dx.doi.org/10.1071/AN14462.

Bates, D., D. Sarkar, M. D. Bates, and L. Matrix. 2007. The lme4 package. R package version 2: 74.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41: 164-171.

Bom, R. A., W. Bouten, T. Piersma, K. Oosterbeek, and J. A. van Gils. 2014. Optimizing acceleration-based ethograms: the use of variable-time versus fixed-time segmentation. Mov. Ecol. 2: 1. http://dx.doi.org/10.1186/2051-3933-2-6.

Brodersen, K. H., C. S. Ong, K. E. Stephan, and J. M. Buhmann. 2010. The balanced accuracy and its posterior distribution. Proc IAPR Int Conf Pattern Recogn. IEEE. 10.1109/ICPR.2010.764.

Charlton, G. L., V. Bouffard, J. Gibbons, E. Vasseur, D. B. Haley, D. Pellerin, J. Rushen, and A. M. de Passillé. 2016. Can automated measures of lying time help assess lameness and leg lesions on tie-stall dairy farms? Appl. Anim. Behav. Sci. 175: 14-22. http://dx.doi.org/10.1016/j.applanim.2015.02.011.

Chebel, R. C., P. R. Silva, M. I. Endres, M. A. Ballou, and K. L. Luchterhand. 2016. Social stressors and their effects on immunity and health of periparturient dairy cows. J. Dairy Sci. 99: 3217-3228. http://dx.doi.org/10.3168/jds.2015-10369.

Dean, B., R. Freeman, H. Kirk, K. Leonard, R. A. Phillips, C. M. Perrins, and T. Guilford. 2012. Behavioral mapping of a pelagic seabird: combining multiple sensors and a hidden Markov model reveals the distribution of at-sea behavior. J. R. Soc. Interface rsif20120570. http://dx.doi.org/10.1098/rsif.2012.0570.

Dohme-Meier, F., L. D. Kaufmann, S. Görs, P. Junghans, C. C. Metges, H. A. van Dorland, R. M. Bruckmaier, and A. Münger. 2014. Comparison of energy expenditure, eating pattern and physical activity of grazing and zero-grazing dairy cows at different time points during lactation. Livest. Sci. 162: 86-96. http://dx.doi.org/10.1016/j.livsci.2014.01.006.

Dolecheck, K. A., W. J. Silvia, G. Heersche, Y. M. Chang, D. L. Ray, A. E. Stone, B. A. Wadsworth, and J. M. Bewley. 2015. Behavioral and physiological changes around estrus events identified using multiple automated monitoring technologies. J. Dairy Sci. 98: 8723-8731. http://dx.doi.org/10.3168/jds.2015-9645.

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

Forney, G. D. 1973. The viterbi algorithm. Proc. IEEE. 61: 268-278. http://dx.doi.org/10.1109/PROC.1973.9030.

Ganskopp, D. C., and D. D. Johnson. 2007. GPS error in studies addressing animal movements and activities. Rangeland Ecol. Manage. 60: 350-358. http://dx.doi.org/10.2111/1551-5028(2007)60[350:GEISAA]2.0.CO;2.

Gomez A., and N. B. Cook. 2010. Time budgets of lactating dairy cattle in commercial freestall herds. J. Dairy Sci. 93: 5772-5781. http://dx.doi.org/10.3168/jds.2010-3436.

González, L. A., B. J. Tolkamp, M. P. Coffey, A. Ferret, and I. Kyriazakis. 2008. Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. J. Dairy Sci. 91: 1017-1028. http://dx.doi.org/10.3168/jds.2007-0530.

González, L. A., G. J. Bishop-Hurley, R. N. Handcock, and C. Crossman. 2015. Behavioral classification of data from collars containing motion sensors in grazing cattle. Comput. Electron. Agric. 110: 91-102. http://dx.doi.org/10.1016/j.compag.2014.10.018.

Gregorini, P. 2012. Diurnal grazing pattern: its physiological basis and strategic management. Anim. Prod. Sci. 52: 416-430. http://dx.doi.org/10.1071/AN11250.

Gregorini, P., J. J. Villalba, F. D. Provenza, P. C. Beukes, and J. M. Forbes. 2015. Modelling preference and diet selection patterns by grazing ruminants: a development in a mechanistic model of a grazing dairy cow, MINDY. Anim. Prod. Sci. 55: 360-375. http://dx.doi.org/10.1071/AN14472.

Guo, Y., G. Poulton, P. Corke, G. J. Bishop-Hurley, T. Wark, and D. L. Swain. 2009. Using accelerometer, high sample rate GPS and magnetometer data to develop a cattle movement and behavior model. Ecol Modell. 220: 2068-75. http://dx.doi.org/10.1016/j.ecolmodel.2009.04.047.

Hernandez-Mendo, O., M. A. Von Keyserlingk, D. M. Veira, and D. M. Weary. 2007. Effects of pasture on lameness in dairy cows. J. Dairy Sci. 90: 1209-1214. http://dx.doi.org/10.3168/jds.S0022-0302(07)71608-9.

Himmelmann, L. 2010. HMM—Hidden Markov models. URL http://cran. r-project. org/web/packages/HMM/index. html

ION. ION STD 101: 1997 Recommended Test Procedures for GPS Receivers. Alexandria, Va. Institute of Navigation Standards.

Kilgour, R. J. 2012. In pursuit of "normal": A review of the behavior of cattle at pasture. Appl. Anim. Behav. Sci. 138: 1-1. http://dx.doi.org/10.1016/j.applanim.2011.12.002.

Kim, S. and W. Lee. 2017. Does McNemar's test compare the sensitivities and specificities of two diagnostic tests? Stat Methods Med Res. 1: 142-154. https://doi.org/10.1177/0962280214541852.

Kutzer, T., M. Steilen, L. Gygax, and B. Wechsler. 2015. Habituation of dairy heifers to milking routine - Effects on human avoidance distance, behavior and cardiac activity during milking. J. Dairy Sci. 98: 5241-5251. http://dx.doi.org/10.3168/jds.2014-8773.

McKellar, A. E., R. Langrock, J.R. Walters, and D.C. Kesler. 2014. Using mixed hidden Markov models to examine behavioral states in a cooperatively breeding bird. Behav. Ecol. 26: 148-157. https://doi.org/10.1093/beheco/aru171.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 2: 153-157.

Milone, D. H., H. L. Rufiner, J. R. Galli, E. A. Laca, and C. A. Cangiano. 2009. Computational method for segmentation and classification of ingestive sounds in sheep. Comput. Electron. Agric. 65: 228-237.

Nielsen, B. L., R. F. Veerkamp, and A. B. Lawrence. 2000. Effects of genotype, feed type and lactational stage on the time budget of dairy cows. Acta Agric. Scand. A Anim. Sci. 50: 272-278. http://dx.doi.org/10.1080/090647000750069467.

O'Driscoll, K., L. Boyle, and A. Hanlon. 2008. A brief note on the validation of a system for recording lying behavior in dairy cows. Appl. Anim. Behav. Sci. 111: 195-200. http://dx.doi.org/10.1016/j.smallrumres.2015.04.002.

Pedersen, M. W., T.A. Patterson, U.H. Thygesen, and H. Madsen. 2011. Estimating animal behavior and residency from movement data. Oikos, 120: 1281-1290.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. 77: 257-286. http://dx.doi.org/10.1109/5.18626.

R Core Team. 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ryan, P. G., S. L. Petersen, G. Peters, and D. Grémillet. 2004. GPS tracking a marine predator: the effects of precision, resolution and sampling rate on foraging tracks of African Penguins. Mar. Biol. 145: 215-223. http://dx.doi.org/10.1007/s00227-004-1328-4.

Sepúlveda-Varas, P., D. M. Weary, and M. A. Von Keyserlingk. 2014. Lying behavior and postpartum health status in grazing dairy cows. J. Dairy Sci. 97: 6334-6343. http://dx.doi.org/10.3168/jds.2014-8357.

Sheahan, A. J., R. C. Boston, and J. R. Roche. 2013. Diurnal patterns of grazing behavior and humoral factors in supplemented dairy cows. J. Dairy Sci. 96: 3201-3210. http://dx.doi.org/10.3168/jds.2012-6201.

Soca, P., H. González, H. Manterola, M. Bruni, D. Mattiauda, P. Chilibroste, and P. Gregorini. 2014. Effect of restricting time at pasture and concentrate supplementation on herbage intake, grazing behavior and performance of lactating dairy cows. Livest. Sci. 170: 35-42. http://dx.doi.org/10.1016/j.livsci.2014.07.011.

Solano, L., H. W. Barkema, E. A. Pajor, S. Mason, S. J. LeBlanc, C. G. Nash, D. B. Haley, D. Pellerin, J. Rushen, A. M. de Passillé, and E. Vasseur. 2016. Associations between lying behavior and lameness in Canadian Holstein-Friesian cows housed in freestall barns. J. Dairy Sci. 99: 2086-2101. http://dx.doi.org/10.3168/jds.2015-10336.

Thorup, V. M., B. L. Nielsen, P. E. Robert, S. Giger-Reverdin, J. Konka, C. Michie, and N. C. Friggens. 2016. Lameness affects cow feeding but not rumination behavior as characterized from sensor data. Front. Vet. Sci.3. http://dx.doi.org/10.3389/fvets.2016.00037.

Vasseur, E., J. Rushen, D. B. Haley, and A. M. de Passillé. 2013. Sampling cows to assess lying time for on-farm animal welfare assessment. J. Dairy Sci. 95: 4968-4977. http://dx.doi.org/10.3168/jds.2011-5176.

Westin, R., A. Vaughan, A. M. de Passillé, T. J. DeVries, E. A. Pajor, D. Pellerin, J. M. Siegford, E. Vasseur, and J. Rushen. 2016. Lying times of lactating cows on dairy farms with automatic milking systems and the relation to lameness, leg lesions, and body condition score. J. Dairy Sci. 99: 551-561. http://dx.doi.org/10.3168/jds.2015-9737.

Williams, Y. J., W. J. Wales, P. T. Doyle, A. R. Egan, and C. R. Stockdale. 2006. Effects of grain or hay supplementation on the chewing behavior and stability of rumen fermentation of dairy cows grazing perennial ryegrass-based pasture in spring. Anim. Prod. Sci. 45: 1519-1528. http://dx.doi.org/10.1071/EA04084.

Williams, M. L., N. Mac Parthaláin, P. Brewer, W. P. James, and M. T. Rose. 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. J. Dairy Sci. 99: 2063-2075. http://dx.doi.org/10.3168/jds.2015-10254.

Chapter 3

The following chapter was accepted for publication in the journal Biosystems Engineering and is therefore formatted to the requirements of the journal. The reference for the publication is:

Williams, M.L., James, W.P. and Rose, M.T., 2019. Variable segmentation and ensemble classifiers for predicting dairy cow behaviour. *Biosystems Engineering*, *178*, pp. 156-167.

Summary

An alternative technique to partitioning data with fixed window sizes for machine learning is to extract information from behaviours for the whole duration that they occur (variable segments). With fixed window sizes of 10 s duration for example (Chapter 2), a two-minute sequence of grazing behaviour would be broken down into 12 segments. With variable segmentation, the two-minute sequence is treated as one continuous segment for feature extraction. One benefit of this method is that models are based on sequences that are more representative of cow behaviour. A second benefit is that real-time sequences of data can be partitioned using separate changepoint algorithms. The behaviour classifier then takes these variable sequences and classifies them. The success of the technique depends on collecting as many sized segments as possible so that the classifier will generalise well on new data. The success of classification also depends heavily on the ability of the changepoint algorithm to identify the points at which the behaviour changes. To validate variable segmentation, a number of standard machine learning algorithms were evaluated in this chapter as well as an ensemble algorithm. Ensemble algorithms are gaining a lot of attention in other fields such as medical diagnosis. The reported benefits of ensembles are that they can result in improved classification performance over standard algorithms (base learners).

This is because they make collective decisions based on the inputs of often several other classifiers built from standard machine learning algorithms. Given their popularity in other fields, the aim was to see whether an ensemble classifier could provide improved classification performance over classifiers built from standard algorithms on data partitioned using variable segmentation.

Variable segmentation and ensemble classifiers for predicting dairy cow behaviour

M. L. Williams,[*] W. P. James, M. T. Rose

Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom

[*]Corresponding author: Manod Lewis Williams. Institute of Biological, Environmental and Rural Science, Aberystwyth University, Penglais Campus, Ceredigion, SY23 3DA, United Kingdom. E-mail: maw90@aber.ac.uk

Highlights

- Variable segmentation of GPS data gathered from dairy cattle was undertaken
- >90% of behavioural changepoints were correctly identified
- The best base learner achieved 96% classification accuracy on the derived segments
- Classification accuracy was not improved when Stacking algorithm was used
- Variable segmentation is a promising strategy for cattle behaviour identification

ABSTRACT

Automatically classifying cattle behaviour using high frequency data usually involves segmentation of data with fixed window sizes for feature extraction. Machine learning algorithms can then be used for supervised modelling of the most biologically important behaviours using these segments. In this work, variable segmentation was applied to GPS data gathered from 30 dairy cows at pasture. Using these segments, the performance of 13 machine learning algorithms (base learners) implemented in WEKA were compared using default parameters in classifying grazing, resting and walking.

Two Stacking ensembles (WEKA implementation of Super Learner) were then derived. The first ensemble contained the best performing base learners. The second was an optimised version derived using a manual ensemble selection method. Both versions of the ensemble were evaluated on an independent test set derived from 10 cows. Overall, the variable segmentation strategy identified 90.2% of changepoints. On the training set, all base learners achieved classification accuracies and F-measures ≥0.90. Optimising the Stacking ensemble led to no further improvement in F-measure (full ensemble = 0.93; optimised ensemble = 0.92) on the test set. The ensembles performed well but base learners utilising boosting algorithms (e.g. simple logistic; logistic model trees) performed as well as the more computationally expensive ensembles. Variable segmentation and ensemble classifiers are promising strategies for classifying the behaviour of dairy cows. However, more work is needed to fully explore and evaluate the potential of ensembles because some base learners may perform equally if not better in some contexts.

Keywords: Variable segmentation; Changepoint; Ensemble classifier; Dairy cattle behaviour classification; Machine learning; Stacking algorithm

Nomenclature: *BN*, Bayes net algorithm; *BinSeg*, Binary segmentation method for data segmentation; *CA*, Classification accuracy; *cpt.mean*, Function to identify changes in the mean of data; *cpt.meanvar*, Function to identify changes in the mean and variance of data; *cpt.var*, Function to identify changes in the variance of data; *LR*, Logistic regression algorithm; *ML*, Machine learning; *MP*, Multilayer perceptron algorithm; *NB*, Naïve Bayes algorithm; *SL*, Simple logistic algorithm; *SMO*, Sequential minimal optimisation; *SVM*, Support vector machine algorithm.

1. INTRODUCTION

The utility of precision technologies in supporting the management of livestock is becoming increasingly evident and is providing the opportunity for practitioners to gain a deep insight into livestock behaviour (Rutten, Velthuis, Steeneveld, & Hogeveen, 2013). Information gathered on a minute-by-minute basis can be used as an additional tool for monitoring performance (Borchers et al., 2016), welfare (Meen et al., 2015), and animal: environment interactions (Fournel, Rousseau, Laberge, 2017) and may provide opportunities for more objective decision support on farms (Neethirajan, Tuteja, Huang, & Kelton, 2017).

Many behavioural recognition studies that gather data using precision technologies such as GPS and accelerometers attempt to classify biologically important behaviours using supervised machine learning (ML) techniques (Diosdado et al., 2015), where the dataset is labelled by manual observation of the behaviours that took place. The next challenge is to partition the dataset prior to the ML phase into appropriately sized segments. Segment size is often determined by some combination of criteria such as the length of time that the target behaviours tend to take place for, and the sampling interval of the device used. The majority of behavioural classification studies segment data at fixed intervals (Martiskainen et al., 2009; Williams, Mac Parthaláin, Brewer, James, & Rose, 2016) followed by a feature extraction step which yields summary statistics that can be used for class discrimination in the modelling phase. A limitation of segmenting temporal data at fixed intervals is the likelihood that some segments will contain data representing more than a single behaviour resulting in misclassifications (Bom, Bouten, Piersma, Oosterbeek, & van Gils, 2014; Williams, James, & Rose, 2017). Consequently, an alternative segmentation methodology has been used in some studies where the temporal data sequences are partitioned at statistically significant

changepoints that are indicative of changes in behaviour (Bom et al., 2014). These methodologies are often used in unsupervised classification schemes in circumstances where there is no prior knowledge of the behaviour that took place in each segment (Valletta, Torney, Kings, Thornton, & Madden, 2017). Expert opinion is then used to carefully describe the likely behaviour that is taking place between each partition point.

To our knowledge, only one other study has used variable segmentation as a strategy to partition data gathered from dairy cattle where it was used to detect heat events (Shahriar et al., 2016). The first objective of our study was to see whether a variable-time segmentation strategy could be used to effectively partition supervised GPS data gathered from dairy cattle. The segmentation strategy used herein was that developed by Killick and Eckley (2014) and is implemented in R (R Core Team, 2014) as the package 'changepoint.' This package provides several options for segmentation, but we opted for the strategy which detects changes in the variance of the time series which has also been successfully implemented in studies of ocean wave heights (Killick, Eckley, Ewans, & Jonathan, 2010) and a species of migratory bird (Madon & Hingrat, 2014).

The second objective of this work was to explore the performance of several ML algorithms implemented in the WEKA data mining suite (Witten, Frank, Hall, & Pal, 2016) on the segmented data for modelling the behaviour of dairy cows. As well as using standard base learners we also wanted to see whether an ensemble classifier could provide better classification performance. The ensemble algorithm used in this work was Stacking; an algorithm which uses a second level meta-learner trained on the output predictions of a set of base learners. Ensemble algorithms have gained popularity in agricultural applications (Escalante, Rodriguez, Cordero, Kristensen, & Cornou, 2013; González-Recio, Rosa, & Gianola, 2014) and have been shown to be

very powerful tools in other subject areas for their superior performance in many cases (Petersen et al., 2015).

2. MATERIALS AND METHODS

2.1. Datasets

GPS data were collected from 40 Holstein dairy cows in 2014 that were managed according to a typical UK dairy grazing regime. Cows were given a daily allocation of approximately 2,500 kg [DM] ha^{-1} of pasture comprising mainly of perennial ryegrass (*Lolium perenne*) which was grazed to a residual of 1,500 kg [DM] ha^{-1} (Williams et al., 2016). Timestamped GPS data was sampled every 5 s because this sample rate was previously found in calibration experiments to provide the best resolution for behaviour recording and power consumption. The GPS data included latitude, longitude and speed. The recorded behaviours were grazing, resting and walking which represent the majority of the time budget of dairy cows (Kilgour, 2012). Grazing was identified when the cow was either walking or standing still with the head lowered and biting the pasture and also if the cow was walking with her head close to the pasture in search of grass. Resting was identified when the cow was lying or stood still with her head raised. Previously Williams et al. (2016) found that separately identifying lying and standing using GPS alone is very difficult and it was therefore decided to combine these behaviours and identify 'resting' as a cow either lying or standing. Walking was identified when the cow was walking or running with her head raised. Approximately six hours of behaviour-labelled data recorded by a single observer was available for each cow for use in this study.

2.2. Variable segmentation

For the purpose of segmentation in this work, the R package 'changepoint' (Killick & Eckley, 2014) which has previously been successful in segmenting data gathered from a migratory bird (Madon & Hingrat, 2014) and marine species (Patel et al., 2015) was chosen. The changepoint package provides users with three main parameterised options for segmentation. Users can choose to search for changes in the mean (cpt.mean()), variance (cpt.var()) or both (cpt.meanvar()) in a continuous timeseries. Changes in the variance of our data were selected for identification. This was simply because upon visual inspection of the majority of sequences, the magnitude of the change in variance between behaviours was usually greater than the changes in the mean. For the changepoint search method the binary segmentation strategy (BinSeg) was implemented. This method iteratively splits the data where changes are identified and searches each new segment for further changes until no further changes are found. In order to implement the cpt.var() function, the data must meet the assumptions of constant mean and normality.

2.2.1. Data transformation

The data was searched for changepoints in the variance of speed (m h^{-1}) gathered from the GPS receivers. The overwhelming majority of this data was heavily tailed, so the Box-Cox transformation was applied to each sequence of speed data before any changepoint estimation took place (Box & Cox, 1964). Data was then visually inspected using histograms and confirmed for normality using the Kolmogorov-Smirnov test (Smirnov, 1939). In order to meet the assumption of constant mean for the application of the cpt.var() function, data were further transformed using first-differencing and again checked for normality using the aforementioned procedures.

2.2.2. Segment building

A training set of segments was built using data from 30 of the cows in the full sample. The changepoint function cpt.var() was operated over each speed sequence from each cow and the GPS data between each estimated changepoint boundary used for feature extraction. From these segments a total of 43 features were extracted, such as turning angle and acceleration, that could be used for the ML phase. For a full account of the features used see Williams et al. (2016). In total, 100 segments each were created for the behaviours grazing and resting and 47 segments were compiled for the behaviour walking. This was labelled as the training set. An independent test set for subsequent model evaluation was compiled using the remaining 10 cows in the full sample. The test set contained 42 segments each for grazing and resting and 12 segments for the behaviour walking. The number of segments created for each category represented the availability of data and that proportionally, cows spent less time exhibiting walking compared to grazing and resting.

2.3. Machine learning

For this phase, the Experimenter GUI of the software package WEKA version 3.9 (Witten et al., 2016) was used. A number of algorithms are available for ML, but the selection of candidates can often be difficult and depend on the data to be modelled, the error risk and computational efficiency (Vilalta & Drissi, 2002). A set of base learners and an ensemble learner were used to test whether using an ensemble added additional prediction power over the best performing individual base learners. Each phase is described below.

2.3.1. Base learner selection

To provide a diverse representation of base learners, the performance of learners selected from four main domains was tested (Bayes, Functions, Rules, Trees) in the WEKA library using *k*-fold cross validation with $k = 10$ on the training set. The best performing from each domain were then used as input into an ensemble. Those selected formed our library of learners.

2.3.2. Ensemble classifier

The Stacking algorithm (sometimes referred to as a 'Super Learner') uses a library of learners which are again trained (and tested) initially using 10-fold cross validation to produce a set of predicted values on the training set. A meta learner (within the Stacking algorithm) is then trained on the outputs of the base learners (Van der Laan, Polley, & Hubbard, 2007). A logistic regression as the meta learner was used for the ensembles as this has been used previously (Whalen & Pandey, 2013) and it has been shown to help avoid overfitting (Ting & Witten, 1999).

2.3.3. Ensemble selection

A manual ensemble selection method was used. The best individual base learners from section 2.3.1 were first evaluated for their performance as individual predictors in empty ensembles. Performance was measured as the cross-validated risk using the estimated root mean squared error (RMSE). The performance among each individual base learner was then statistically evaluated (section 2.5) and only the best performing were added to the ensemble. While recognising that this method disregards the value of each individual base learner to the ensemble and that other methods are available (e.g. Caruana, Niculescu-Mizil, Crew, & Ksikes, 2004), it has been found that this manual selection method with an ensemble comprised of few base learners (e.g. <14) performs as well as other methods (Whalen & Pandey, 2013) that add predictors according to their performance contribution to the ensemble (Caruana et al., 2004).

2.3.4. Ensemble performance

The performance of (1) the optimised ensemble with selected base learners (section 2.3.3) was evaluated against the performance of (2) the ensemble with all of the best base learners from section 2.3.1 using a variety of metrics (section 2.5) including training time.

2.4. Machine learning algorithms

The ML algorithms used here were not an exhaustive list but those selected were from five main domains (Bayes, Functions, Rules, Trees, Meta) within the WEKA software in order to provide a diverse range of algorithms and also because some have performed well previously (Díez-Pastor, Rodríguez, García-Osorio, & Kuncheva, 2015; Williams et al., 2016). Algorithms were selected from each domain and kept within these domains for statistical comparison for the sake of simplicity, despite there being differences in algorithm mechanics within each domain. Algorithms are grouped in the

WEKA library largely according to their mathematical attributes. Due to the large number of hyperparameter tuning options available for most of the algorithms used, it was decided to test each one using default parameters. Below is a brief summary of each algorithm tested.

Bayes – two algorithms were chosen from this folder; naïve Bayes (NB; John & Langley, 1995) and Bayes net (BN; Friedman, Geiger, & Goldszmit, 1997).

Functions – this folder contains several algorithms including functions for linear regression modelling, logistic regression, support vector machines and neural networks. Selected were; logistic regression (LR; Le Cessie & Van Houwelingen, 1992), simple logistic (SL; Friedman, Hastie, & Tibshirani, 2000), support vector machine (SVM; Vapnik, 1999) SMO (sequential minimal optimisation) and the neural network, multilayer perceptron (MP; Bishop, 1995).

Rules – the algorithms selected were JRip (Cohen, 1995), PART (Frank & Witten, 1998) and OneR (Holte, 1993).

Trees – three decision tree algorithms were selected from this category, naïve Bayes tree (NBTree; Kohavi, 1996), J48 (Quinlan, 1993) and an algorithm for building logistic model trees (LMT; Friedman et al., 2000; Landwehr, Hall, & Frank, 2005).

Meta – the ensemble algorithm chosen was Stacking, often referred to in the literature as 'Super Learner' (Petersen et al., 2015; Van der Laan et al., 2007). Stacking uses a set of base learners to firstly classify instances from the training set which are subsequently channelled into a meta level training set to produce a meta classifier (Dzeroski & Zenko, 2004; Wolpert, 1992).

2.5. Classifier performance

The overall performance of classifiers was evaluated using the classification accuracy (CA):

$$\text{CA} = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. However, class imbalance can lead to over optimistic performance (Stąpor, 2018). Therefore, as well as CA we evaluated each learner using sensitivity (TP / (TP + FN)), specificity (TN / (FP + TN)), precision (TP / (TP + FP)), and F-measure (2TP / (2TP + FP + FN)) which is the harmonic mean of precision and sensitivity.

For the statistical comparison of base learners within each WEKA category (section 2.3.1) we used the F-measure. This metric has previously been shown to be highly correlated with other metrics of performance such as CA (Ferri, Hernández-Orallo, & Modroiu, 2009) and is useful in cases where datasets are imbalanced (Díez-Pastor et al., 2015). For WEKA algorithm groups with more than two learners, the Friedman non-parametric two-way analysis of variance (Friedman, 1937) with Iman-Davenport extension (Iman & Davenport, 1980) was used to test the null hypothesis that all base learners within each group performed the same. If α was ≤0.05 we used the Nemenyi *post-hoc* test (Nemenyi, 1962) which is the non-parametric equivalent of the Tukey test for ANOVA for pairwise comparisons (Demšar, 2006). Groups with only two base learners were tested using a *t*-test for repeated cross validation (Stąpor, 2018).

Next, in order to find the optimal base learner combination for the ensemble, we used the RMSE performance of the best base learners was used from above for

statistical comparison (section 2.3.3). The RMSE was derived from the individual inclusion of each base learner in empty ensembles on the training set. Statistical comparison was done, again using the extended Friedman test and Nemenyi *post-hoc* test if α was ≤0.05. For the RMSE statistics we also provide the bootstrapped 95% CI (Carpenter & Bithell, 2000). We chose bootstrapped CI's because of the non-Gaussian distribution of the RMSE data. These were calculated in R using the software package 'boot' (Canty & Ripley, 2012). For non-statistical comparison, we also provide the training time for each learner in empty ensembles.

Finally, the best performing base learners were put forward for the optimised ensemble learner for a performance comparison against the full (non-optimised) ensemble (section 2.3.4). Both ensembles were evaluated on their F-measures, using a *t*-test for repeated cross validation.

3. RESULTS

3.1. Data transformation

Figure 1A shows the data transformation procedure for an example supervised dataset from a single cow with two behavioural changepoints, the first at instance 1,393 and the second at instance 1,847. Panels a-c show the raw speed data, a histogram of raw speed with distributional curve and an autocorrelation function (acf) plot of the speed data respectively. Some important requirements of the changepoint detection algorithm are that the data are normally distributed and have a constant mean. The Box-Cox transformation was used to achieve data normality for each dataset before first-differencing to remove data mean fluctuations (Fig. 1A; panel d). Transformation and first-differencing meant that the data met the assumption of normality as closely as possible (Fig. 1A; panel e). A potential issue for the changepoint algorithm is the detection of false positive changepoints. Data gathered at a high sample rate is susceptible to high positive autocorrelation (Perotto-Baldivieso et al., 2012) and the procedure of first-differencing was also an attempt at removing as much of this autocorrelation as possible. Figure 1A panel f shows an example of a transformed dataset after removing the autocorrelation showing majority removal of this phenomenon.

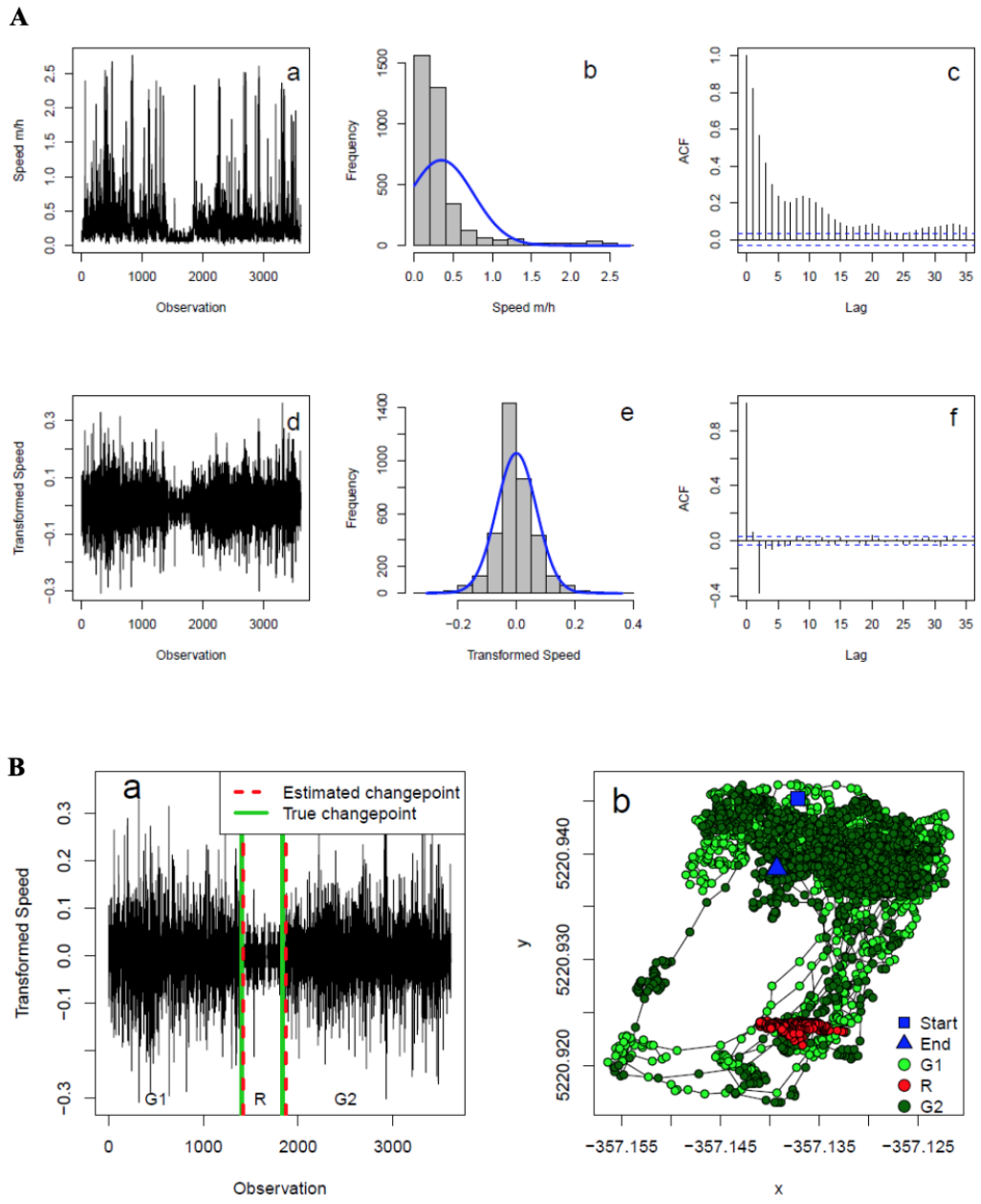


Fig. 1 - A. GPS data transformation procedure. Panels a - c show a sample of raw speed data collected by GPS from a dairy cow at pasture, a histogram of speed distribution and an autocorrelation plot of speed respectively. Panels d - f show the speed data after Box-Cox transformation and first-differencing, a histogram of the transformed speed data and resulting autocorrelation plot respectively. **B**. Example of variable segmentation (a) and behaviour prediction (b). Panel a shows Box-Cox transformed and first-differenced speed data from a dairy cow and also the estimated and true changepoints defining the first grazing bout (G1), resting bout (R) and a second grazing bout (G2). Panel b shows the spatial data of the same cow plotted to define the behaviour bouts estimated by the changepoint method in panel a.

3.2. Performance of the changepoint model framework

Based on the supervised data collected for the 30 cows, on average (±SD), the changepoint algorithm identified changepoints correctly to within 4.45min (±5.23) of the true changepoint taking place with 90.2% of changepoints being identified. Figure 1B, panel a gives an example of the true and estimated changepoints in a dataset of 3,600 observations (5 h) after applying the changepoint algorithm. The resultant pathplot is shown in Fig. 1B panel b showing the spatial data and the two distinct grazing bouts (G1, G2) surrounding the single resting bout (R). Despite the overall average delay in segmentation at the exact transition point across all datasets, an observation of the data showed that in the overwhelming majority of cases, changes in the variation of speed as logged by the GPS receivers occurred more slowly in relation to the time point where the cow actually changed behaviour. Examples of true and estimated changepoints are shown in Fig. 2. Panels a, c, d and e show behavioural transitions from grazing to resting states while panels b and f show examples of cows moving from resting to grazing. In most cases, the changepoint algorithm was identifying points of significant change in variance when it empirically occurred (as it should) and not necessarily at the timestamp where the behaviour actually changed. Some segments during compilation were therefore contaminated with false positive behaviours equating to on average (±SD) 68.6 (±69.8) GPS instances. This can be seen as a fundamental weakness of the GPS receiver itself in adjusting to changes in movement rather than a weakness of the changepoint algorithm in identifying changepoints. The average number of instances per segment for each behaviour exhibited in the supervised training set was compiled and duration calculated for the 30 cows. The mean time (±SD) cows spent per segment exhibiting grazing resting and

walking was 15.58min (±14.30), 60.1min (±37.46) and 0.38min (±0.24) respectively, highlighting the variation exhibited by different cows in these behaviours.

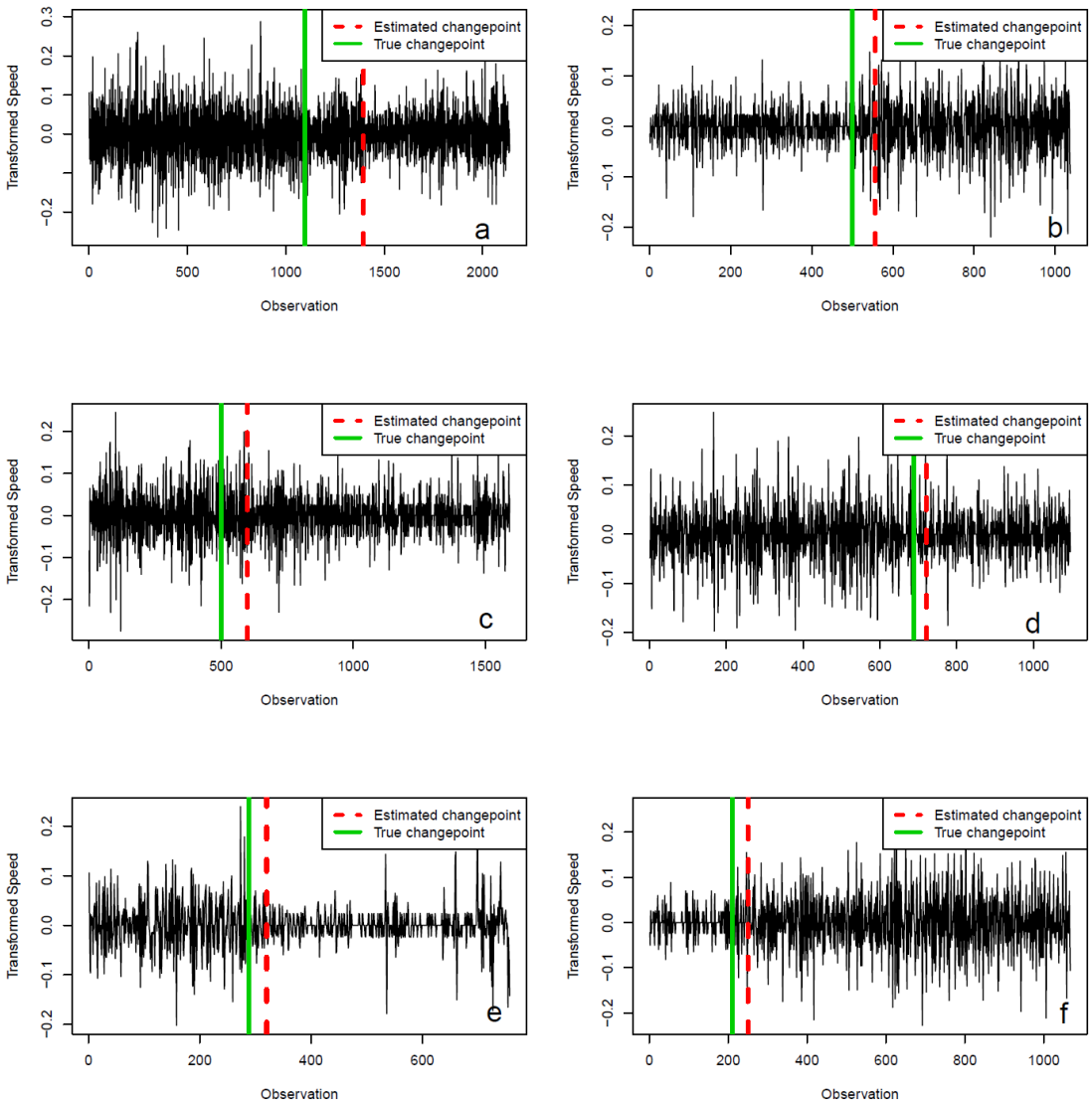


Fig. 2 - True and estimated behavioural changepoints of dairy cows as predicted by the changepoint method. Panels a, c, d and e show cows moving from grazing to resting states. Panels b and f show cows moving from resting to grazing states. The Y axis is the speed variable collected by GPS receivers after Box-Cox transformation. The X axis represents the timestamped observation number sampled every 5 s.

3.3. Performance and selection of base learners

Overall, each base learner performed very well on the training set across all evaluation metrics with all achieving >0.90 CA. Table 1 shows the performance of each base learner within each category as defined in the WEKA library. Even the less complex algorithm OneR which selects a single rule with the smallest total error from a list generated for each feature in the dataset achieved very good performance (CA = 0.90; sensitivity = 0.88; specificity = 0.92; precision = 0.92; F-measure = 0.90). However, as with the LR classifier and the SVM with RBF Kernel the sensitivity for this learner was lower. Despite this, the OneR learner was trained in a fraction of the time (training time = 0.01min) compared to both MP and NBTree which both took >19min to train. The more complex classifiers seemed to perform the best overall however which included the neural network algorithm MP. All classifiers returned very high specificity (≥0.92) and precision (≥0.92) metrics. The differences between the F-measures of each classifier and the results were statistically tested for significance as shown in Table 1. In the Functions category, both SL and MP outperformed LR and were thus selected for further analysis as components in the ensemble. BN outperformed the NB classifier in the Bayes category. For the decision tree classifiers only LMT was selected for further analysis in the full ensemble. In the rules category, PART was selected only, despite JRip significantly outperforming OneR on the training set. In an attempt to ensure algorithm diversity, JRip was not included. Of the two SVM tested, the SVM with PolyKernel function performed better than the equivalent classifier with RBFKernel function and was thus selected for further analysis.

Table 1 - Classification performance of base learners within each WEKA domain on the training set tested using 10-fold cross-validation. Tests were performed on a Windows 7, Intel Core i5-3470 CPU machine with 3.2 GHz.

WEKA library domain	Base learner	Classification accuracy	Sensitivity	Specificity	Precision	F-measure	Training time (min)
Functions	Simple logistic	0.97	0.96	0.97	0.97	0.96[a]	0.86
	Logistic	0.92	0.89	0.94	0.94	0.91[b]	0.16
	Multilayer perceptron	0.96	0.95	0.97	0.97	0.96[a]	19.14
Bayes	Naïve Bayes	0.93	0.91	0.94	0.94	0.92[b]	0.01
	Bayes net	0.95	0.94	0.96	0.96	0.95[a]	0.02
Decision trees	NBTree	0.95	0.93	0.96	0.96	0.94[b]	19.31
	J48	0.94	0.92	0.95	0.95	0.93[b]	0.02
	LMT	0.96	0.95	0.97	0.97	0.96[a]	2.35
Rules	JRip	0.93	0.91	0.94	0.94	0.92[b]	0.05
	PART	0.94	0.93	0.95	0.95	0.94[a]	0.05
	OneR	0.90	0.88	0.92	0.92	0.90[c]	0.01
Support vector machines	SVM + PolyKernel	0.94	0.93	0.95	0.95	0.94[a]	0.12
	SVM + RBFKernel	0.91	0.89	0.93	0.93	0.91[b]	0.24

[a-c]F-measures for classifiers within the column and within each WEKA library domain with different superscripts are significantly different ($P < 0.05$).

3.4. Ensembles

3.4.1. Ensemble optimisation

The RMSE performance and bootstrapped 95% CI of base learners evaluated in empty Stacking ensembles on the training set is shown in Table 2. Overall, LMT, MP and SL (mean RMSE = 0.16) significantly outperformed ($P < 0.05$) BN, PART and the SVM with PolyKernel function (mean RMSE = 0.21) and were thus chosen to form the optimised ensemble. It was interesting to note that the three best performing algorithms were indeed those that either implemented the LogitBoost algorithm (SL and LMT) or backpropagation method of classification (MP). A closer analysis of the time taken to train each learner showed that the empty ensemble learner with the MP algorithm took 63 min to train compared to a total training time for every other learner of 11.20 min (Table 2). This represents a substantial computational cost in training the ensemble with nested MP algorithm. Comparatively, the empty ensemble with LMT algorithm took 7.53 min and the empty ensemble with SL took 3.06 min for training.

Table 2 - Root mean squared error performance (±95% bootstrapped CI) of selected base learners in empty ensemble classifiers on the training set evaluated using 10-fold cross-validation. Tests were performed on a Windows 7, Intel Core i5-3470 CPU machine with 3.2 GHz.

Base learner	Training time (min)	RMSE (±95% CI)
Bayes net	0.09	0.20 (±0.01)[a]
LMT	7.53	0.15 (±0.02)[b]
Multilayer perceptron	63	0.17 (±0.02)[b]
PART	0.16	0.21 (±0.02)[a]
Simple logistic	3.06	0.15 (±0.01)[b]
SVM + PolyKernel	0.36	0.22 (±0.01)[a]

[a-b]Different superscripts indicate significant differences in the RMSE performances of base learners ($P < 0.05$).

3.4.2. Ensemble performance on training and independent test sets

The performance per class of the full and optimised Stacking ensemble classifiers on the training and test set is shown in Table 3. Overall, the CA of both learners on the training set was very high with both achieving 0.96 on average across all classes. Similarly, performance was excellent for sensitivity, specificity and precision and there was no significant difference in the F-measure of both classifiers. However, as would be expected, the training time of the optimised classifier was significantly shorter (-6 min; $P < 0.001$) compared to the classifier with the full complement of base learners. Therefore, the only apparent benefit of optimisation was the marginal saving in training time of the optimised ensemble.

Performance on the test set was marginally poorer compared to the training set for both ensemble classifiers with both achieving an overall CA of 0.93. The major per class loss in CA on the test set was for the behaviour walking. One distinct

disadvantage of the test set was the relatively small number of instances, especially for this particular class (n = 12 instances). The full and optimised versions of the classifier suffered an overall average 3% loss in CA and very similarly in the F-measure on the test set compared to the training set. On the test set, there was no significant difference between the F-measure performances of both ensembles ($P > 0.05$). There was however a marginal difference in testing time between the ensembles with the optimised ensemble being marginally faster ($P < 0.001$).

The computational time required to teach the ensemble classifiers on the training set averaged 220 min (Table 3). This raises the question of whether this additional time can be justified by the performance of the ensembles in this work. Table 1 also shows the computational time for training each of the base learners on the training set. It can be seen that in comparison with the base learners, training the ensembles comes at greater computational cost with very little performance improvement over the best base learners selected (Table 1; Table 3). In fact, the F-measure achieved on the training set with the SL classifier (0.96) was the same as that achieved on the training set by both ensembles. SL also had a marginally higher CA (0.97). For a fraction of the computational power, SL, LMT and MP performed at least as well as both ensembles on this particular training set.

Table 3 - Individual class performance of full and optimised ensemble classifiers on training (10-fold cross-validation) and test sets. Tests were performed on a Windows 7, Intel Core i5-3470 CPU machine with 3.2 GHz.

Training set	Classification accuracy	Sensitivity	Specificity	Precision	F-measure	Training time (min)
Full ensemble (6 base learners)						
Grazing	0.94	0.94	0.94	0.94	0.94	74
Resting	0.94	0.91	0.96	0.96	0.93	74
Walking	0.99	0.99	1.00	1.00	0.99	75
Average	0.96	0.95	0.97	0.97	0.95[*]	$\sum$ 223[†]
Optimised ensemble (3 base learners)						
Grazing	0.95	0.96	0.94	0.94	0.95	74
Resting	0.95	0.92	0.97	0.97	0.94	71
Walking	0.99	0.99	1.00	1.00	0.99	72
Average	0.96	0.96	0.97	0.97	0.96[*]	$\sum$ 217[†]
Test set						Testing time (s)
Full ensemble (6 base learners)						
Grazing	0.94	0.95	0.93	0.93	0.94	0.06
Resting	0.94	0.93	0.94	0.94	0.93	0.04
Walking	0.92	0.83	1	1	0.91	0.04
Average	0.93	0.90	0.96	0.96	0.93[**]	$\sum$ 0.14[‡]
Optimised ensemble (3 base learners)						
Grazing	0.94	0.93	0.94	0.94	0.93	0.04
Resting	0.92	0.93	0.90	0.90	0.92	0.03
Walking	0.92	0.83	1	1	0.91	0.02
Average	0.93	0.90	0.95	0.95	0.92[**]	$\sum$ 0.09[‡]

[*]No significant difference ($P > 0.05$) between F-measure of full and optimised ensemble on training set.
[**] No significant difference ($P > 0.05$) between F-measure of full and optimised ensemble on test set.
[†]Training time differs significantly ($P < 0.001$) between full and optimised ensemble on training set.
[‡]Testing time differs significantly ($P < 0.001$) between full and optimised ensemble on test set.

4. DISCUSSION

4.1. Variable segmentation

The first aim of this study was to test whether a variable segmentation technique could be used to partition GPS data collected from dairy cows into recognisable segments for ML. Similar segmentation approaches have been used previously with good success for the segmentation of data gathered by accelerometers from crab plovers (*Dromas ardeola*) (Bom et al., 2014) and golden eagles (*Aquila chrysaetos*) (Sur et al., 2017). Such variable segmentation techniques fundamentally differ to the commonly used fixed-time segmentation method. The latter partitions data at fixed intervals which can lead to behavioural overlap and misclassification (Bom et al., 2014) but does not require changepoint inference. The benefit of the former, applied in its usual context is that it can be applied without supervision of the dataset (in fact it is largely used on elusive species). The sensitivity of such methods to data fluctuations can also lead to a broader range of behaviours being identified (although this could also be detrimental to the task). Thus, it has been proposed that variable segmentation techniques are probably best suited to species where supervision of the training set is not possible and where behaviours are likely to persist for shorter periods, and change rapidly (Sur et al., 2017).

To our knowledge, variable segmentation of data gathered from cattle has only been explored in one other study which was to classify heat events in dairy cows (Shahriar et al., 2016). From the results herein, combined with ground-truth data, variable segmentation could potentially lead to improvements over methods utilising fixed segments for cattle behaviour recognition (Williams et al., 2016). This is indeed because dairy cow behavioural bouts of grazing and resting generally take place for long periods of time (section 3.2). Provided that sensors are sufficiently able to detect

changes represented by these behaviours, from our experience, change detection algorithms such as the one used in this work are very useful. The sensitivity of the changepoint algorithm can also be tuned to better fit individual datasets and more work is required in exploring these parameters for further evaluation. The success of the algorithm will also depend on the study species (e.g. cattle or sheep), the sampling interval of the device used to capture the data as well as the location of the sensor on the animal. For example, better results may be achieved with accelerometers where users need not account for satellite positional fix error or multipath effects as is usually the case with GPS (Ganskopp & Johnson, 2007). Furthermore, the changepoint algorithm was somewhat let down in this work by the fact that there was a delay in the empirical transition point in many cases, leading to contaminated segments. However, the algorithm was still successful in identifying >90% of changepoints.

Overall, devices such as accelerometers are likely to provide access to a wider range of behaviours, going beyond that which is often limited to GPS such as grazing, resting and walking. For example, rumination is a very important metric for consideration for cow health and production, but GPS is not sufficiently sensitive enough to capture this information and further dimensionality is needed (Rayas-Amor et al., 2017). Variable segmentation shows very good potential in identifying transition points in data gathered from livestock using precision technologies that can provide greater resolution. Supervised applications of variable segmentation methods provide a good opportunity to properly evaluate these algorithms and to optimise the algorithm used to search for transition points. The binary segmentation (BinSeg) algorithm was found to be very successful in its current application but other researchers are encouraged to evaluate the full range of options in future work.

4.2. Base learner evaluation

The WEKA ML suite provides a range of algorithms for training and testing. The second aim of this work was to evaluate a sample of these for model learning on variable data segments with the ultimate goal of using the best performing classifiers as components in a Stacking ensemble. Overall it can be stated unequivocally that all base learners performed very well in cross-validation on the training set across all metrics analysed with the SL classifier performing the best (CA = 0.97; sensitivity = 0.96; specificity = 0.97; precision = 0.97; F-measure = 0.96). Training time for SL was marginally longer however compared to the majority of the other algorithms tested (training time = 0.86 min). SL fits logistic regression models using the LogitBoost algorithm and supports automatic attribute selection (Landwehr et al., 2005) making it computationally efficient, usually producing compact classifiers and has been found to perform very well previously (Alickovic & Subasi, 2014). Similarly, the LMT algorithm performed very well on the training set (CA = 0.96; sensitivity = 0.95; specificity = 0.97; precision = 0.97; F-measure = 0.96) but taking longer to train (2.35 min).

The performance of all learners was equal to or marginally better compared to similar studies (Dutta et al., 2015; Williams et al., 2016) that have classified dairy cow behaviour. One important benefit which became apparent with variable segments was that the algorithms used were able to form simple decision rules based on fewer attributes compared to segments of fixed size (Williams et al., 2016). For example, OneR was able to classify instances with an overall CA of 0.90 on the training set using the single attribute of the accumulated number of non-moving instances (instances where the animal was not empirically moving). This attribute amongst others helped simplify the classification task compared to fixed segments where more complex

models were developed (Williams et al., 2016). The more complex algorithms, although performing well, were computationally expensive and it is questionable whether the additional performance of such algorithms, (e.g. MP) outweighs the computational cost in training.

Manual ensemble selection was used to select the most appropriate base learners for optimising the ensemble. While other methods were available, this method has been used previously (Whalen & Pandey, 2013) and found to have been appropriate for ensembles with few base learners (e.g. <14). However, it is recommended that other methods are also tested, such as iteratively adding base learners that maximise the performance of the ensemble thus accounting for the value of each base learner to the full ensemble. Whichever method is used, in order for the ensemble to perform better than any of its individual members, algorithm diversity is an important consideration and care must be given to avoid overfitting which can be problematic with some methods (Whalen & Pandey, 2013).

4.3. Ensemble evaluation

Compared to single classification algorithms, ensemble algorithms are being used more frequently in agricultural contexts (Chaudhary, Kolhe, & Kamal, 2016; Dutta et al., 2015; Hill, Connolly, Reutemann, & Fletcher, 2014) because of their efficacy. The final aim of this work was to evaluate the performance of selected base learners in an ensemble classifier. The ensemble chosen was Stacking. An optimised (best 3 base learners) and full ensemble (6 best base learners) were compared for performance. Despite better performance with some of the base learners in empty Stacking ensembles, no statistically significant improvement was found between the F-measures of the full and optimised versions of the ensemble in neither training nor test set. However, there was a difference in the time required to teach the models on the

training set (full ensemble = 223 min vs. optimised ensemble = 217 min) and in execution time on the test set (full ensemble = 0.14 s vs. optimised ensemble = 0.09 s). This is a very marginal gain in computational efficiency, but the performance of neither classifier fully justifies its use when compared to the performances of the single base learners that could be trained in a fraction of the time while achieving comparable performance.

The loss in performance found for both classifiers in the test set was largely due to the loss in sensitivity for the class walking, probably because of the small sample size and similar instances not appearing in the training set. When selecting an ML algorithm, evidently the context of the work is important and the means by which the data are collected and segmented may well impact performance. Thus, we would recommend that algorithms such as SL and LMT are used for further analysis on similar datasets for behavioural classification with potential for use as single classifiers or integrated in an ensemble. Based on their individual performances here these algorithms are more efficient than the ensembles and base learner combinations tested. SL and LMT use a boosting algorithm and despite similar performance to MP on the training set this contributes toward better computational efficiency in training. We would not recommend the use of MP on similar data gathered by GPS simply based on the training time, however, it must be noted that we did not attempt to tune any hyperparameters in this work and so more experimentation would be required before firm conclusions could be made as to the potential performances of all algorithms tested. The options for hyperparameter tuning are almost non-exhaustive and outside the scope of this work. Methods for automatic hyperparameter selection do exist however and could be utilised in future (Thornton, Hutter, Hoos, & Leyton-Brown, 2013). It is well known though that if class decision boundaries are complex then the

convergence of the backpropagation method of MP is slow (Chaudhuri & Bhattacharya, 2000). Others have also experienced excellent performance with MP as well as lengthy training times (Nookala, Orsu, Pottumuthu, & Mudunuri, 2013).

More work is needed to fully explore the capabilities of ensemble classifiers for similar applications in livestock behaviour prediction. For example, accelerometers provide an opportunity to gather data at higher sampling rates, thus potentially allowing for subtler and often inconspicuous behaviours to be identified. Examples of which include transitional behaviours (e.g. moving from lying to standing) that usually occur very quickly but can provide important cues that variable segmentation techniques could firstly identify. Diosdado et al. (2015) had some success in classifying overall behavioural transitional events with dairy cows but distinguishing between lying to standing events and standing to lying events was more difficult. Future work should aim to elucidate the power of ensemble algorithms in this regard as well as for identifying other behaviours that may be useful for monitoring health and performance such as overall feed intake and bite rate (González, Tolkamp, Coffey, Ferret, & Kyriazakis, 2008). With an increase in the amount of data available for analysis in the livestock sector (Wolfert, Ge, Verdouw, & Bogaardt, 2017) training time, model complexity and performance are examples of variables that need careful consideration in the selection of the most appropriate algorithm for the task at hand.

5. CONCLUSIONS

Variable segmentation techniques present an efficient framework for partitioning data for classifying dairy cattle behaviour using ML algorithms. This technique may be more efficient and potentially lead to enhanced behaviour recognition than data segmented at fixed intervals. This is likely to be especially true if used on data captured at frequent intervals and with sensors that are sufficiently able to detect changes quickly (e.g. seconds). This latter point is important because as we found in this work there was often a delay in the empirical transition point, for example when a cow moved from a state of grazing to a state of resting. This resulted in changes being detected later than the true changepoint; a problem that could be mitigated if accelerometers were used in conjunction with the GPS receivers. The sensor used is also an important consideration. As discussed, variable changepoint detection algorithms tend to be used most where it is not possible to view the species directly for supervision of the training data. Such applications are often aimed at elucidating behaviours that may not require the resolution needed for livestock monitoring. Examples include deciphering the dispersal patterns of terrestrial (Gurarie et al., 2016) and marine species (Patel et al., 2015) where data is likely to be infrequently sampled. In these examples GPS is usually sufficient. In modern livestock husbandry where long-term behaviour monitoring is important for disease recognition (González et al., 2008), efficient and accurate tools are required for detecting subtle changes in behaviour and robust models are needed for classifying those behaviours. For this purpose, it was established that the base learners explored here performed very well. Ensemble algorithms also show great promise in precision livestock husbandry and more work is needed in exploring the different options available for farm-level tasks

that may benefit from the added power of ensemble learning that is already being discovered in other areas.

ACKNOWLEDGEMENTS

The authors would like to thank the farm staff at Aberystwyth University Research Farm for their cooperation. The sponsorship of M. L. Williams by Coleg Cymraeg Cenedlaethol is gratefully acknowledged. Declarations of interest: none.

REFERENCES

Alickovic, E., & Subasi, A. (2014). Effect of Multiscale PCA De-noising in ECG Beat Classification for Diagnosis of Cardiovascular Diseases. *Circuits, Systems, and Signal Processing*. https://doi.org/10.1007/s00034-014-9864-8

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. *Clarendon Press*. https://doi.org/10.2307/2965437

Bom, R. A., Bouten, W., Piersma, T., Oosterbeek, K., & van Gils, J. A. (2014). Optimizing acceleration-based ethograms: The use of variable-time versus fixed-time segmentation. *Movement Ecology*. https://doi.org/10.1186/2051-3933-2-6

Borchers, M. R., Chang, Y. M., Proudfoot, K. L., Wadsworth, B. A., Stone, A. E., & Bewley, J. M. (2017). Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2016-11526

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*. https://doi.org/10.2307/2287791

Canty, A., & Ripley, B. (2012). boot: Bootstrap R (S-Plus) functions. *R package version*, *1*(7).

Carpenter, J., & Bithell, J. (2000). Bootstrp intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Twenty-first international conference on Machine learning - ICML '04*. https://doi.org/10.1145/1015330.1015432

Chaudhary, A., Kolhe, S., & Kamal, R. (2016). A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2016.03.026

Chaudhuri, B. B., & Bhattacharya, U. (2000). Efficient training and improved performance of multilayer perceptron in pattern classification. *Neurocomputing*. https://doi.org/10.1016/S0925-2312(00)00305-2

Cohen, W. W. (1995). Fast effective rule induction. *Proceedings of the Twelfth International Conference on Machine Learning*. https://doi.org/10.1.1.50.8204

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*. https://doi.org/10.1016/j.jecp.2010.03.005

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C., & Kuncheva, L. I. (2015). Random Balance: Ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2015.04.022

Diosdado, J. A. V., Barker, Z. E., Hodges, H. R., Amory, J. R., Croft, D. P., Bell, N. J., & Codling, E. A. (2015). Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. *Animal Biotelemetry*. https://doi.org/10.1186/s40317-015-0045-8

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2014.12.002

Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning*. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

Escalante, H. J., Rodriguez, S. V., Cordero, J., Kristensen, A. R., & Cornou, C. (2013). Sow-activity classification from acceleration patterns: A machine learning approach. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2013.01.003

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2008.08.010

Fournel, S., Rousseau, A. N., & Laberge, B. (2017). Rethinking environment control strategy of confined animal housing systems through precision livestock farming. *Biosystems Engineering*. https://doi.org/10.1016/j.biosystemseng.2016.12.005

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. *Proceedings of the Fifteenth International Conference on Machine Learning*. https://doi.org/1-55860-556-8

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1937.10503522

Friedman, N., Geiger, D., & Goldszmit, M. (1997). Bayesian Network Classifiers. *Machine Learning*. https://doi.org/10.1023/a:1007465528199

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*. https://doi.org/10.1214/aos/1016218223

Ganskopp, D. C., & Johnson, D. D. (2007). GPS error in studies addressing animal movements and activities. *Rangeland Ecology and Management*. https://doi.org/10.2111/1551-5028(2007)60[350:GEISAA]2.0.CO;2

González-Recio, O., Rosa, G. J. M., & Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*. https://doi.org/10.1016/j.livsci.2014.05.036

González, L. A., Tolkamp, B. J., Coffey, M. P., Ferret, A., & Kyriazakis, I. (2008). Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2007-0530

Gurarie, E., Bracis, C., Delgado, M., Meckley, T. D., Kojola, I., & Wagner, C. M. (2016). What is the animal doing? Tools for exploring behavioural structure in animal movements. *Journal of Animal Ecology*. https://doi.org/10.1111/1365-2656.12379

Hill, M. G., Connolly, P. G., Reutemann, P., & Fletcher, D. (2014). The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2014.08.011

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*. https://doi.org/10.1023/A:1022631118932

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the fbietkan statistic. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610928008827904

John, G. H., & Langley, P. (1995). Estimating Continuous Distribution in Bayesian Classifiers. In *UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*.

Kilgour, R. J. (2012). In pursuit of "normal": A review of the behaviour of cattle at pasture. *Applied Animal Behaviour Science*. https://doi.org/10.1016/j.applanim.2011.12.002

Killick, R., & Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of statistical software*, *58*(3), 1-19.

Killick, R., Eckley, I. A., Ewans, K., & Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*. https://doi.org/10.1016/j.oceaneng.2010.04.009

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. https://doi.org/citeulike-article-id:3157868

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*. https://doi.org/10.1007/s10994-005-0466-3

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191-201. https://doi.org/10.2307/2347628

Madon, B., & Hingrat, Y. (2014). Deciphering behavioral changes in animal movement with a "multiple change point algorithm- classification tree" framework. *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2014.00030

Martiskainen, P., Järvinen, M., Skön, J. P., Tiirikainen, J., Kolehmainen, M., & Mononen, J. (2009). Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science*. https://doi.org/10.1016/j.applanim.2009.03.005

Meen, G. H., Schellekens, M. A., Slegers, M. H. M., Leenders, N. L. G., van Erp-van der Kooij, E., & Noldus, L. P. J. J. (2015). Sound analysis in dairy cattle vocalisation as a potential welfare monitor. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2015.08.028

Neethirajan, S., Tuteja, S. K., Huang, S. T., & Kelton, D. (2017). Recent advancement in biosensors technology for animal and livestock health management. *Biosensors and Bioelectronics*. https://doi.org/10.1016/j.bios.2017.07.015

Nemenyi, P. (1962). Distribution-free multiple comparisons. In *Biometrics* (Vol. 18, No. 2, p. 263). 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210: INTERNATIONAL BIOMETRIC SOC.

Nookala, M., Orsu, N., Pottumuthu, B. K., & Mudunuri, S. B. (2013). Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification. *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*. https://doi.org/0120884070, 9780120884070

Patel, S. H., Morreale, S. J., Panagopoulou, A., Bailey, H., Robinson, N. J., Paladino, F. V., & Spotila, J. R. (2015). Changepoint analysis: A new approach for revealing animal movements and behaviors from satellite telemetry data. *Ecosphere*. https://doi.org/10.1890/ES15-00358.1

Perotto-Baldivieso, H. L., Cooper, S. M., Cibils, A. F., Figueroa-Pagán, M., Udaeta, K., & Black-Rubio, C. M. (2012). Detecting autocorrelation problems from GPS collar data in livestock studies. *Applied Animal Behaviour Science*. https://doi.org/10.1016/j.applanim.2011.11.009

Petersen, M. L., LeDell, E., Schwab, J., Sarovar, V., Gross, R., Reynolds, N., & Bangsberg, D. R. (2015). Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. *Journal of Acquired Immune Deficiency Syndromes*. https://doi.org/10.1097/QAI.0000000000000548

Quinlan, J. (1993). *C4. 5: programs for machine learning*. *Machine Learning*. https://doi.org/10.1016/S0019-9958(62)90649-6

R Core Team. (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing: Vienna, Austria*. https://doi.org/10.1017/CBO9781107415324.004

Rayas-Amor, A. A., Morales-Almaráz, E., Licona-Velázquez, G., Vieyra-Alberto, R., García-Martínez, A., Martínez-García, C. G., & Miranda-de la Lama, G. C. (2017). Triaxial accelerometers for recording grazing and ruminating time in dairy cows: An alternative to visual observations. *Journal of Veterinary Behavior*, *20*, 102-108. https://doi.org/10.1016/j.jveb.2017. 04.003

Rutten, C. J., Velthuis, A. G. J., Steeneveld, W., & Hogeveen, H. (2013). Invited review: Sensors to support health management on dairy farms. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2012-6107

Shahriar, M. S., Smith, D., Rahman, A., Freeman, M., Hills, J., Rawnsley, R., & Bishop-Hurley, G. (2016). Detecting heat events in dairy cows using accelerometers and unsupervised learning. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2016.08.009

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, *2*(2), 3-14.

Stąpor, K. (2018). Evaluating and comparing classifiers: Review, some recommendations and limitations. In *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-59162-9_2

Sur, M., Suffredini, T., Wessells, S. M., Bloom, P. H., Lanzone, M., Blackshire, S., & Katzner, T. (2017). Improved supervised classification of accelerometry data to distinguish behaviors of soaring birds. *PLoS One*. https://doi.org/10.1371/journal.pone.0174785

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*. https://doi.org/10.1145/2487575.2487629

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.594

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour*. https://doi.org/10.1016/j.anbehav.2016.12.005

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.2202/1544-6115.1309

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*. https://doi.org/10.1109/72.788640

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*. https://doi.org/10.1023/A:1019956318069

Whalen, S., & Pandey, G. (2013). A comparative analysis of ensemble classifiers: Case studies in genomics. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2013.21

Williams, M. L., Mac Parthaláin, N., Brewer, P., James, W. P. J., & Rose, M. T. (2016). A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2015-10254

Williams, M. L., James, W. P., & Rose, M. T. (2017). Fixed-time data segmentation and behavior classification of pasture-based cattle: Enhancing performance using a hidden Markov model. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2017.11.010

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. *Data Mining: Practical Machine Learning Tools and Techniques*.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*. https://doi.org/10.1016/j.agsy.2017.01.023

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*. https://doi.org/10.1016/S0893-6080(05)80023-1

Discussion

8. Review of Chapter 1: Williams, M.L., Mac Parthaláin, N., Brewer, P., James, W.P.J. and Rose, M.T., 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. Journal of Dairy Science, 99(3), pp.2063-2075.

8.1 Overview

This was a contribution to the PLF literature by using computational techniques that are well recognised in other disciplines such as mathematics and computer science although the field of dairy genetics has been making use of these techniques for a number of years (Weigel *et al*., 2017). A comprehensive dataset was gathered from cows grazed in an experimental paddock, providing a total of 153 h of data for analysis. Several classifiers were developed and evaluated for their performance in correctly recognising the behaviours of dairy cows in pasture-based systems. The best classifiers were those created using a type of ensemble algorithm (random forest) (Breiman, 2001) and a rule-based classifier developed using the JRip (RIPPER) algorithm (Fürnkranz and Widmer, 1994). The difference in performance was marginal between these two classifiers, but JRip provided intuitive rules that could be easily interpreted and offered an insight into the complex way in which the classified behaviours manifest themselves in the GPS data. Despite the high classification performance of random forest and its popularity for use in the ML literature, random forest often produces highly complex and large decision rules. This often makes it difficult to interpret and so gaining biological intuition from the output of this algorithm can be very difficult. This was one of the main reasons why JRip was selected here given that the rules produced were very easy to interpret and could be scrutinised visually. In PLF applications it may be that this type of output will not be required by farmers for management decisions. However, in future, it may be useful to experiment with various algorithms to see whether additional information can be gathered such as that produced here with JRip.

The rule-set developed by the JRip algorithm demonstrated the variety of different ways in which cows may graze (as determined by GPS), with seven rules explaining the occurrence of grazing behaviour. These rules can intuitively provide analysts with a better understanding of the movement ecology of grazing cows. In the experimental paddock at least, these rules may represent the seven most common ways in which cows were found to be grazing and could be useful in future studies for understanding more about grazing ecology. For example, it may be that grazing behaviour can actually be defined by a finite set of rules and this could be very useful in modelling the impact of grazing on various landscapes and systems. It would be very interesting to undertake further work in this area to see whether differences occur between the movements of different animals grouped by age or breed perhaps. Care is needed in the interpretation of the rule-set developed here as JRip undertakes a pruning step which means that the rules presented here were the most optimal for representing the various behaviours. For the purpose of classification, the final model showed very good performance. Furthermore, despite an original feature set of 43 features, a feature analysis showed that there were 13 features that appeared to be the most useful to the JRip algorithm. Feature selection is very useful so that classifiers can be learned without the need for added computational complexity.

An important consideration for the future will be to ask how a variety of managerial scenarios affect classification performance. Can models developed in controlled conditions extend to correctly classify behaviour (especially grazing) when cows are managed in suboptimal conditions? This will be true no matter what sensor or modelling technique is used.

8.2 A finite behaviour set

It was evident that only a limited number of behaviours could be modelled and predicted with acceptable performance using GPS data. One of the most notable disappointments in the preliminary modelling stages was the realisation that no clear distinction could be made between the behaviours lying and standing. These of course have biological value as it is well known that patterns in these behaviours are indicative of certain issues such as lameness (Walker *et al*., 2008). Indeed, one of the initial objectives of this candidature was to develop a behaviour model for a long-term experiment to study patterns in the behaviours of dairy cows that could be indicative of disease onset. This is not to say that this would not be possible with the final model developed in Chapter 1, but it certainly reduces the opportunity to evaluate a breadth of important behaviours. The number of correct and misclassified instances in the development stages for the behaviours grazing, lying, standing and walking can be seen later in this section (unbalanced classification performance shown in Table 5).

It was clear that additional sensors would be required to discriminate between instances of lying and standing despite having extracted a number of different features from the dataset. A GPS collar with integrated accelerometer for example could be very useful in distinguishing between lying and standing behaviours. Indeed, in future, it may be that GPS will be used simply for gaining a positional fix for the animal with accelerometers being used for behaviour classification. Nevertheless, it was decided to merge the instances of lying and standing into a single category (resting) and despite not being the optimal outcome, resting was identifiable in the data. A search of the literature showed that others have also faced this problem and have combined behaviours as a consequence (Guo *et al*., 2009; Owen-Smith *et al*., 2012; de Weerd *et al*., 2015). To the candidate's knowledge, only Godsk and Kjærgaard (2011) have

successfully discriminated between lying and standing behaviours of cows with acceptable performance using only GPS. They achieved CAs of 77% and 76% for lying and standing respectively.

The merging of behaviours is not uncommon in other studies that have used different sensors for behaviour classification. The decision to merge certain behaviours essentially depends on three factors, (1) the ability of an algorithm to discriminate between the behaviours (2) the biological importance of the behaviours to the study objective or to the industry and (3) the representation of those behaviours in the training set. González *et al*. (2015) developed a decision tree classifier from data collected from GPS and accelerometers fitted to steers. The identified behaviours were ruminating, foraging, travelling and resting, with head shaking, scratching and grooming grouped as 'other active behaviours' because of their low occurrence in the training data. They also noted that this category had a lower sensitivity compared to the other evaluated categories. Following on from the rationale of Dutta *et al*. (2015), an unsupervised cluster analysis could also be undertaken to search for structure in the feature-space for aiding decisions on the number of classes that can be identified in the supervised stage. This technique will be considered as this work progresses into the future.

8.3 Behaviour representation

8.3.1 Grazing and lying dominate

Gaining an even distribution of training data across all behaviour categories was challenging in the present work. Even the behaviour standing was underrepresented compared to grazing and lying; the two most abundant behaviours in the dataset. Very rarely did cows stand for periods exceeding 1 min and it was very uncommon to find cows standing for at least 3 min which was the threshold required for feature extraction.

Walking was also very sparsely represented. Indeed, it became clear based on observations that the cohort of dairy cows under observation did not often walk purposefully, at least to the extent that it was described in the methodology (Chapter 1: *Grazing Management and GPS*). Instead, cows often grazed, or browsed their way to another location. In the experimental paddock, cows would often walk to the water trough, but unfortunately, this represented the main instance of walking that could be recorded. Having said that, given the persistence velocity and directionality represented by a bout of walking, this behaviour was actually very easy to discriminate by the ML algorithms tested. Lying and standing on the other hand were characterised by similar features such as acute turning angles and small movements between fixes. Future modelling should try and ensure that rare behaviours are well represented in datasets providing that these behaviours are of biological value. For behaviours like standing, this could require that observers search only for this behaviour in order to balance behaviour classes. Furthermore, observing animals in a standard-sized paddock may be useful for gaining greater instances of walking, although catching sight of focal cows would be more difficult within a larger cohort of cattle.

8.3.2 Data balancing techniques

Balancing datasets so that the most represented behaviours were equal helped improve the discriminatory power and therefore performance of the algorithms tested, and this was true across all segment sizes. To balance the majority classes, random under-sampling was used (Batista *et al*., 2004). Under-sampling had very little, if any effect on the efficacy of the majority of classifiers to discriminate walking from the other behaviours because of its strong feature characteristics (high speed and directionality). For grazing and resting, balancing meant that the performance of classifiers was more even. For classes that have high representation in a dataset, some

algorithms can favour their presence and sometimes this leads to high classification success of the majority class and poorer performance for the others (Longadge, and Dongre, 2013). Having said that, the naïve Bayes classifier was resilient to this phenomenon (Chapter 1: Table 3) and this has been documented previously for naïve Bayes classifiers (Provost, 2000). As well as under-sampling, other techniques can be used which include cost-sensitive classification techniques and over-sampling. Cost-sensitive classification is used frequently in medical diagnostics (Yang *et al*., 2009; Hsu *et al*., 2015). For example, when building classifiers for diagnosing a type of cancer, the training dataset will most often include more instances of benign tumours compared to malignant. Thus, using cost-insensitive classification on this training set could lead to the development of classifiers that miss TP instances of malignant tumours which is far more costly than incorrectly classifying a benign tumour as malignant. Cost-sensitive classification accounts for the importance of each class despite its representation by adding weight to each class (Ling and Sheng, 2011).

Although cost-sensitive classification was not tested in Chapter 1, an over-sampling technique was initially tested. Over-sampling can be performed in WEKA using SMOTE (synthetic minority over-sampling technique) which creates synthetic samples of the selected class using a *k*-nearest neighbour technique (Chawla *et al*., 2002). The original, compiled dataset for the classes grazing, standing, lying and walking was highly imbalanced with 631, 196, 2,507 and 96 segment instances respectively. Such imbalances can lead to poor classification performance because algorithms will favour the majority class (Chawla *et al*., 2003). Table 5 shows the results of a random forest classifier on this unbalanced dataset. Overall, the classifier had a strong bias towards the class lying, resulting in a misleadingly high CA (Section 2.3.1.2).

Table 5. Classification performance of random forest classifier on full, unbalanced dataset. Results show high bias towards the class lying. Note a high overall classification accuracy largely due to the high number of correctly classified lying instances

Random forest	Grazing	Standing	Lying	Walking	TP	FP	Precision	F1	Overall CA
Grazing	355	1	275	0	0.56	0.05	0.68	0.61	0.83
Standing	29	26	141	0	0.13	0.001	0.89	0.23	
Lying	134	2	2,371	0	0.94	0.45	0.85	0.89	
Walking	0	0	0	96	1.00	0	1.00	1.00	

Data balancing is essential in these circumstances. Here, balancing could first be undertaken by reducing the size of the lying class to equal that of the grazing class (631 segments each). Re-modelling using random forest produces the results in Table 6. Despite poorer but misleading CA (does not fully represent the per class results), the performance of the resulting model improves the classification success for grazing greatly, and marginally for the class standing.

Table 6. Classification performance of random forest classifier on dataset balanced for the classes grazing and lying (631 instances each) using random under-sampling

Random forest	Grazing	Standing	Lying	Walking	TP	FP	Precision	F1	Overall CA
Grazing	567	2	62	0	0.89	0.23	0.72	0.80	0.76
Standing	79	32	85	0	0.16	0.006	0.80	0.27	
Lying	137	6	487	1	0.77	0.15	0.76	0.77	
Walking	0	0	0	96	1.00	0.001	0.99	0.99	

Over-sampling can be used to raise the number of standing instances only as it seems that walking is discriminated well by random forest. Raising the percentage of standing instances by 200% using 5 nearest neighbours provides a more balanced

dataset with 631, 588, 631 and 96 instances for grazing, standing, lying and walking respectively. Table 7 shows the results of a random forest classifier on the new dataset balanced using SMOTE. Per class performance increases greatly for the class standing but reduces marginally for grazing and lying. Overall, CA improves, but only slightly.

Despite great improvement in performance for identifying standing instances, care needs to be taken when class instances are synthetically increased in number because new data are generated from existing instances which means that it is unlikely that the technique will introduce much variance into the dataset (Chawla *et al*., 2002). In the present example, if the original dataset balanced by under-sampling includes only 196 instances for grazing, lying and standing (under-sampling to equal that of standing) it can be seen in Table 8 that performance is actually reduced and the classification performance for standing and lying is almost identical.

Table 7. Classification performance of random forest classifier on dataset balanced for the classes grazing, standing and lying using synthetic minority over-sampling technique

Random forest	Grazing	Standing	Lying	Walking	TP	FP	Precision	F1	Overall CA
Grazing	526	51	54	0	0.83	0.13	0.75	0.70	0.78
Standing	59	467	62	0	0.79	0.10	0.77	0.78	
Lying	113	88	429	1	0.68	0.08	0.78	0.73	
Walking	0	0	0	96	1.00	0.001	0.99	0.99	

Some authors have used SMOTE to increase minority class instances (Homburger *et al*., 2014; Krug *et al*., 2015) when working with livestock data. It seems that an effective method of using SMOTE as recommended by Chawla *et al*. (2002) is to combine under-sampling and over-sampling in the same dataset, which reduces the initial bias towards the majority class to favour the minority class. On the example

dataset here, under-sampling the majority classes by 50% and over-sampling the minority class by 50% does improve the prediction accuracy of both lying and standing, and, on average their TP rate improves by 20% whilst grazing TP rate improves by 14%. Overall, the CA improves to 80% with more even performance across classes (confusion matrix not shown). Whilst this could serve as an effective technique to demonstrate potential classifier capability, it should perhaps be used as an exercise to supplement additional data collection for minority class instances so that a rigorous assessment can be made of classifiers on real-world data.

Table 8. Classification performance of random forest classifier on dataset balanced by random under-sampling of grazing and lying classes to match original number of standing instances (*n* = 196)

Random forest	Grazing	Standing	Lying	Walking	TP	FP	Precision	F1	Overall CA
Grazing	146	38	12	0	0.74	0.18	0.62	0.67	0.65
Standing	52	103	41	0	0.52	0.18	0.52	0.52	
Lying	37	54	104	1	0.53	0.10	0.66	0.58	
Walking	0	0	0	96	1.00	0.002	0.99	0.99	

8.4 Limitations of GPS receivers

8.4.1 GPS hardware

The GPS receivers themselves were a source of variation which became evident in the preliminary trials (Section 3.2.4) and this phenomenon must be considered for future work with satellite-based receivers to maximise the chances of consistent data. No further stationary tests were undertaken during the remainder of the candidature to ensure ongoing consistency so no comment can be made as to whether the experimental sample of receivers (n = 25) remained homogenous for the duration of the work which took place between 2014-2016. With regard to the integrity of the hardware, the receivers used were not specifically designed for use with livestock and so were open to physical damage (physical damage was very rare) and water ingress during the work which sometimes (very rarely) led to data loss. During the stationary tests, each receiver was wrapped in a polythene bag before each launch to prevent water ingress. To ensure consistency, the receivers were also wrapped in polythene bags before attaching to cow collars which led to reduced data loss.

One major limitation of GPS receivers is power consumption which increases in-line with the sampling interval. High sampling frequency is usually desired in animal movement studies because it maximises the information gathered about the study species. Furthermore, to decrease the size and weight of receivers, traditional receivers often require small batteries which leads to a trade-off between the number of locations that can be generated and battery life. In marine species, significant savings can be made in battery power as positional fixes are taken only when the animal surfaces (Bestley *et al*., 2016). A similar means of saving battery power can be undertaken with terrestrial animals where GPS can be programmed to log only when the animal is moving. Longer sampling intervals (minutes rather than seconds) can also be used with

terrestrial animals and gaps in GPS fixes filled using inertial movement sensors such as an accelerometer and an electronic compass to make estimates on where the animal was and what it was doing (Tomkiewicz *et al*., 2010). It has been found that vegetation density and time to satellite fix also increase battery drain (Fischer *et al*., 2018). Fundamentally, with re-capture loggers such as those used here, the receiver has to be removed from the animal for data processing and so the capacity of the onboard memory card is also an issue with high-frequency data collection. Bio-telemetry devices that transmit data to a central processing station can be more cost-effective in terms of power consumption, particularly if accelerometers are used for behaviour classification and that the classification algorithm is computationally simple (Kwong *et al*., 2012; Abbasi *et al*., 2014).

In this work, a 5 s sampling interval was chosen as a result of the static receiver tests (no significant difference in horizontal error between tested intervals) and also because it provided the best resolution for the battery power available. This meant however, that batteries would only provide 24 hrs of power, meaning that daily battery changes were required in subsequent cow studies (Chapter 2). On board GPS in purpose-built tracking collars for animal studies (e.g. Lotek LiteTrack series; Biotrack ltd., Wareham, UK) are usually used for low-level behaviour inference or solely for positional fixes and an integrated accelerometer is usually used for behaviour classification if required. This means that batteries last weeks, to months at a time, relative to the frequency of positional fixes. It seems likely that this will be a common hardware choice in future especially if rechargeable solutions are provided that improve unit longevity.

8.4.2 Sources of GPS error

As well as multipath error (signals reflect off structures) and atmospheric effects (delay in signal as it passes through atmosphere) (Kos *et al*., 2010; Li *et al*., 2015), satellite constellation is another source of error that cannot be directly controlled for and requires post-processing of the GPS data (Polojärvi *et al*., 2011). For the GPS receivers used in the current work, the accompanying software allowed for differential correction of the accuracy of the positional coordinates by enabling the EGNOS (European geostationary navigation overlay service) function. Such errors are important to consider so that accurate estimations can be made on the movement of individual animals if the objectives include monitoring individual health. While it was realised in the present work that even small structures could lead to multipath errors (even with EGNOS enabled) and that this was accounted for in the experimental design, future studies should assess the extent of the errors caused by physical barriers. The impact of tree cover, farm buildings and other structures should be fully assessed so that horizontal error can be accounted for appropriately in the post-processing phase (Ganskopp and Johnson, 2007). Figure 13 gives an example of how GPS error distributions can be plotted to visualize the standard horizontal error metrics. Shown in Figure 13 is GPS receiver #32 (located at red cross) scheduled to sample at 30 s intervals. Its estimated CEP and R95 are also shown which were 1.31 m and 2.74 m respectively. Error-drift can be seen in the top right-hand area of the plot. This was present in 32 of the 36 receivers tested.

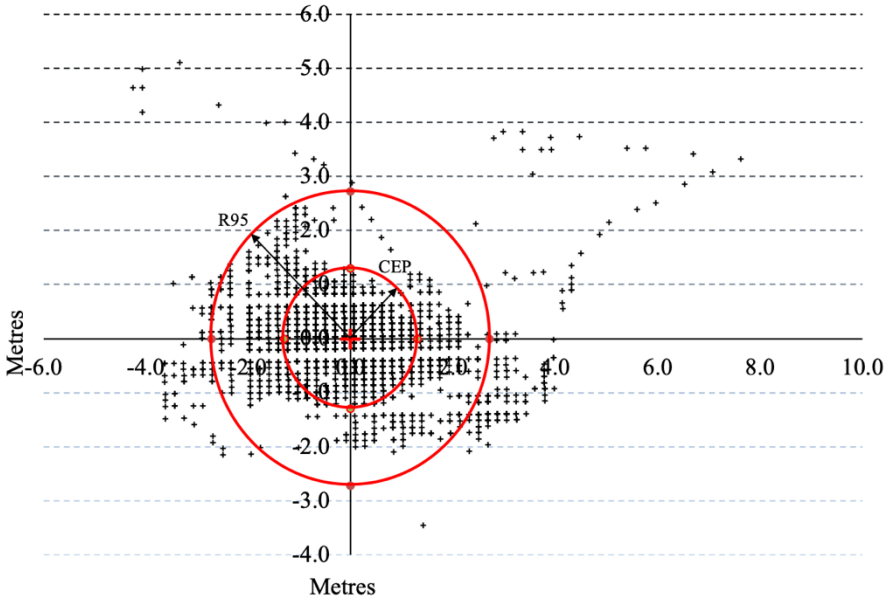


Figure 13. Horizontal error distribution of GPS receiver #32 logging at 30 s intervals showing CEP (1.31 m) and R95 (2.72 m). Plot shows 2,881 coordinates with some coordinate drift evident in the right-hand corner of the plot. This phenomenon was present in 32 of the 36 GPS receivers in the original sample. Red cross at point 0,0 indicates location of GPS receiver.

8.4.3 Implications of sensor variation to other studies

The candidate is not aware of other work that has undertaken preliminary trials of other sensors such as accelerometers to ensure homogeneity between units. Speculatively, it may be that more variability is introduced by the animals themselves compared to that introduced by precision engineered sensors such as accelerometers. With that and while in pursuit of models that can generalise well between animals, the candidate is only aware of one study (Diosdado *et al.*, 2015) that has assessed the degree of variability in accelerometer signatures between animals for the same behaviour. They found that variability (measured as sensitivity) between individual cows for feeding behaviour was low (78.49-100%) but sensitivity for lying behaviour showed greater variation (21.82-100%). They also found that misclassifications occurred for the same behaviour within cows on successive days and they hypothesised that this may have been down to small variations in the location of the sensor on the

neck of the animal. However, it seems that there are no studies that have assessed performance for the same animal in different environments. In supervised classification exercises, the aim is to gather as much training data as possible so that a pooled sample is accrued for ML. However, in most circumstances this information is gathered at once, perhaps over a number of days (Nielsen, 2013) or weeks (Robert *et al*., 2009) where little variation is represented in physical (e.g. feed height at feed fence or pasture) or biological factors (e.g. height of the animal) and so more is needed to assess the degree of variability represented by these variables.

8.5 Dealing with biological and environmental variability in future studies

It is well established that a range of sensors are capable of capturing meaningful information about cattle and that this data can be modelled to make accurate predictions of behaviour states. Moving forward, there seems to be two clear objectives.

The first objective should be to ascertain whether daily variation in the environment of cows has an impact on the predictability of established models. This is particularly important for attributes of feeding behaviour which could be affected by variables such as pasture quality, pasture density, silage height at the feed fence and the components of a total mixed ration. Some researchers modelling the precise attributes of feeding behaviour are attempting to account for some of these variables using acoustic monitoring (Milone *et al*., 2009; Galli *et al*., 2011). It remains to be seen whether accelerometers can be used to distinguish between different attributes of feeding (e.g. tearing pasture, swallowing and regurgitating), and it may be that other sensors such as acoustics are required for this. However, combining sensor systems is likely to provide greater resolution for the behaviour classification process. As well as using accelerometers for feeding and resting behaviours, additional information can be inferred from GPS or LPS location data across a pasture or within a barn environment.

For example, location can be important for monitoring grazing areas that are at a higher risk of parasite burden (Falzon *et al*., 2013) or used for monitoring displacement activity in barns which could lead to more information on disease progression (Gygax *et al*., 2007). Perhaps in future, farmers will have to decide on what level of surveillance they need for their particular farming system. For example, it may be enough to know that a cow was feeding between 9-10am (e.g. classification by accelerometers) on one farm but another farmer may require a higher level of behaviour discrimination and may want to know how much feed was actually consumed between 9-10am (e.g. classification by noseband pressure sensor).

Once models are considered robust to these daily variations, the second objective should be to monitor the long-term impact of other variables such as weather, topography, aspect and stocking density (particularly important in pasture-based systems) on the behaviour of cows. If health monitoring is one of the objectives of such surveillance systems, then it will be important to account for the variability introduced by such factors to reduce the risk of FP alarms to a particular issue. For example, if early lameness detection is incorporated into the surveillance system, then the algorithm should account for variables that may reduce or increase the amount of behaviour undertaken beyond that expected within the normal range of individual cows. Already established products should ideally be accounting for such variation.

9. Review of Chapter 2: Williams, M.L., James, W.P. and Rose, M.T., 2017. Fixed-time data segmentation and behavior classification of pasture-based cattle: Enhancing performance using a hidden Markov model. *Computers and Electronics in Agriculture, 142*, pp.585-596.

9.1 Overview

Chapter 2 described an attempt at improving on the performance of the model developed in Chapter 1. When models such as that described in Chapter 1 are developed, behaviours are usually modelled independently by extracting features from groups of single behaviours. As such, transitions between behaviours are not often considered. Indeed, the candidate is only aware of a small number of publications (Robert *et al*., 2009; González *et al*., 2015; Diosdado *et al*., 2015) that have directly attempted to model the transitionary behaviours of cattle. Of these, only Diosdado *et al*. (2015) successfully classified non-specific transitions (combination of standing up and lying down transitions). However, discriminating between standing up and lying down was more difficult. The other publications noted the difficulty of classifying transitions given the very rapid nature of a transitionary event.

Fixed-time modelling of behavior can therefore lead to reductions in overall CA when deployed on continuous streams of data and future attempts to model cow behavior should certainly consider transitionary events. In the current study, rather than use an HMM to model the behaviour of cows outright which has been done previously (Guo *et al*., 2009), it was decided to develop an HMM for error correction and this significantly improved the overall classification performance. The assumption behind the Markov process is that any given future state depends only on the current state and not on any previous states. Transitions between the selected finite states are then governed by a set of transition probabilities. Many classification approaches (e.g. ML) assume independence between individual observations but in sequential GPS or accelerometer data there is a natural dependence between behaviour observations

(Leos-Barajas *et al*., 2017). The intervals between data points vary widely between studies and in reality, the sampling interval will affect the methods used and the inferences that can be made from movement data. Ideally, the temporal scale of observations should be selected before sensor deployment based on the behavioural characteristics of the focal animal. Often, the studied behaviours outlast the sample rate of the sensor and therefore, studies where serial dependance is assumed in the data has often been dealt with using HMMs (Ward *et al*., 2006; Mannini *et al*., 2011). HMMs provide a means of accounting for strong autocorrelation in sensor data rather than neglecting this feature. If sampling rate varies or samples are taken at random times then HMMs are not suitable as the Markov process breaks down (Patterson *et al*., 2017). With this in mind, it was envisaged that an HMM would be able to deal with transitionary events in a more effective manner. The classified segments were used effectively as the movement path which was then assumed to follow the Markov assumption. This was a more intuitive means of dealing with transitionary events given the independence assumed between segments by the ML classification process (Chapter 1).

Prior to using this methodology, an attempt was made at modelling transitionary events using ML techniques. In agreement with the previous publications, these events were very rapid, and the majority occurred faster than the GPS sampling interval (< 5 s).

9.2 Modelling transitionary events

9.2.1 Transition segments

An attempt was made at modelling transitions by sampling a window equivalent in size to that determined as the optimum in Chapter 1 (32 movements) and centralizing the point of transition within the window (16 movements either side of the transition point) (Figure 14).

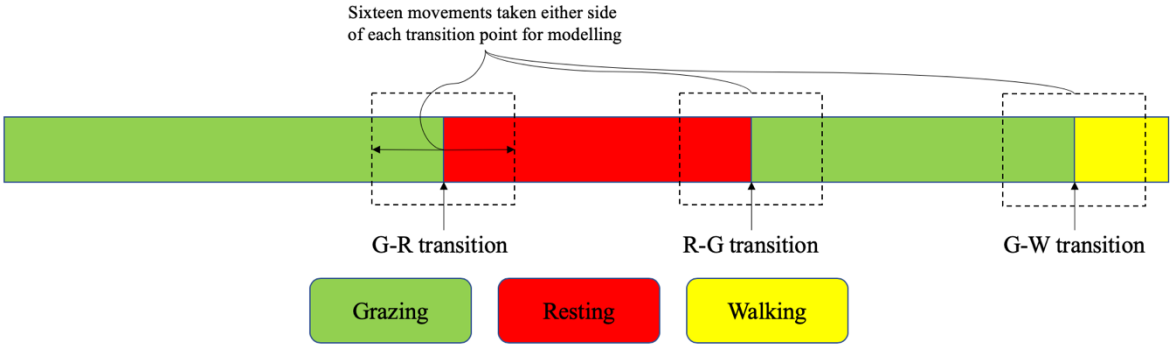


Figure 14. Method for collecting transition events. Event windows were 32 movements in length (160 s) with the point of transition at the center of each window. Window length was chosen to coincide with optimal segmentation strategy as determined in Chapter 1.

The rationale, based on observations, was that cows were likely preparing for the transition sometime in advance and that after the transition, cows would still be exhibiting some movement. The most prevalent transitions were those between grazing and resting (G-R) and resting and grazing (R-G) and so, segments were compiled to represent only these two transitionary groups. In total, 76 G-R segments and 84 R-G segments were compiled and tested with several algorithms for classification performance. Table 9 shows the confusion matrices and performance of JRip and random forest classifiers using 10-fold cross-validation on the unbalanced transition dataset. It can be seen that neither model discriminated between transitions very well. Furthermore, balancing the number of instances in each class (n = 76 instances per class) resulted in almost a 10% reduction in CA for JRip and only a 6% improvement in CA for random forest (confusion matrices and results not shown).

Table 9. Confusion matrices and performance of JRip and random forest classifiers on unbalanced transition dataset. G-R indicate grazing to resting transitions. R-G indicate resting to grazing transitions

JRip	G-R	R-G	TP	FP	Precision	F1	Overall CA
G-R	25	51	0.32	0.31	0.49	0.39	0.52
R-G	26	58	0.69	0.67	0.53	0.60	
Random forest							
G-R	32	44	0.42	0.52	0.45	0.43	0.48
R-G	38	46	0.54	0.57	0.51	0.52	

Given the poor performance, it was decided to merge the transitions into a single category (transition) to see whether classifiers could discriminate between this category and the various behaviour classes. The number of transition segments totalled

152 instances. Grazing and resting were represented equally (n = 631 instances per class) and walking was represented by 96 instances. Results of the reconfigured dataset are shown in Table 10. Far better performance was realised having merged the transitionary segments, which to some extent was not expected given that half of each transition segment is represented by either one of the two behaviours modelled (grazing and resting). The results in Tables 9 and 10, before and after transition merger corroborate that found by Diosdado *et al*. (2015) in that non-specific transition events can be detected with very acceptable performance.

Table 10. Confusion matrices and performance of JRip and random forest classifiers on dataset balanced for grazing and resting and including the combined transitionary behaviours (Transition)

JRip	Grazing	Resting	Walking	Transition	TP	FP	Precision	F1	Overall CA
Grazing	499	124	0	8	0.79	0.17	0.76	0.77	
Resting	149	475	3	4	0.75	0.15	0.77	0.76	0.79
Walking	0	2	94	0	0.97	0.002	0.96	0.97	
Transition	4	10	0	138	0.90	0.009	0.79	0.79	
Random forest									
Grazing	558	70	0	3	0.88	0.18	0.77	0.82	
Resting	154	474	2	1	0.75	0.09	0.85	0.75	0.83
Walking	1	0	95	0	0.99	0.001	0.97	0.98	
Transition	9	10	0	133	0.87	0.11	0.84	0.83	

Despite good transition classification performance of the combined transition set, the fundamental issue with trying to recognise data segments as already established is that unless the moving window aligns perfectly with the point of transition, then classification performance if likely to suffer (Section 4.1.1.2 and Figure 6). The analyst is subsequently faced with the same challenge as previously. No attempt was made therefore at deploying this model on sequences of data given the issue with transition boundaries. However, it was concluded that discrimination of non-specific transitions was indeed possible, and in fact, on a per class basis, success was higher compared to classes grazing and resting with the JRip classifier.

One solution to this issue would be to develop a classifier that can undertake an initial 'sweep' of the timeseries data to define the transitionary events first, before classifying behaviour events within the partitioned data. An example of which was undertaken by Diosdado *et al*. (2015) using thresholds for determining transitions in the Y-axis of acceleration data. Their methodology achieved very good determination of non-specific transitionary events (sensitivity = 95.45%; precision = 87.50%). In essence, the initial identification of transition events using this type of method is akin to using VS methods in a timeseries and is one of the main reasons why VS was attempted in Chapter 3.

9.2.2 HMM for transition inference

HMMs have been used for a number of years to model the movement of a variety of animals (Franke *et al*., 2006; Langrock *et al*., 2012; Pohle *et al*., 2017) including cows (Guo *et al*., 2009) and even for recognizing bovine call types (Jahns, 2008). HMMs are typically used outright to infer the path of behavioural states using a set of probabilities, these being the initial state probabilities, the probabilities of transitions between states, and the probability of observation given the hidden state. Guo *et al*. (2009) modelled the stay regions of cattle (areas between travel sites) using HMMs. They used a hierarchical combination of HMMs, producing time-dependent models for each individual cow. Data were collected from GPS and accelerometers and the authors related the directional and angular speeds of cows to three categories of behaviour; foraging, bedding and relocating. The benefit of modelling each individual cow was that they could make tailored predictions of the behaviour of each animal as they found that some cows exhibited large variation in their behavioural attributes. Including time of day also accounted for the fact that behaviour probability distributions can differ depending on time period.

In principle, developing an HMM for detecting transitions in classified sequences (Chapter 2) is the same as for outright behaviour modelling. When initializing the HMM described in Chapter 2, both initial state and transition probabilities were calculated from 150 h of data gathered during the observational study undertaken as part of Chapter 1. The emission probabilities (probability of observation) however were inferred from the classified instances of the algorithm developed in Chapter 1 in conjunction with the known behaviour states. Finally, when run on sequences of predicted behaviours, it was possible for the HMM to detect whether the classified instances were representative of the likely behaviour sequence

that cows had actually undertaken. Any classifications deemed erroneous by the HMM were then corrected.

The HMM in Chapter 2 was developed using the R package 'HMM' (Himmelmann, 2010) which, in hindsight is slightly more limited in its functionality compared to other packages developed for this purpose such as depMixS4. This package provides capabilities for selecting the optimal number of hidden states (not necessarily required for the present study) and also further functions for optimizing model parameters and for model selection (Visser and Speekenbrink, 2010). Several other packages exist that also provide similar functionalities including HiddenMarkov (Harte, 2015) and msm (Jackson, 2011). A more recent package which provides a workflow process that is specifically aimed at the analysis of ecological and animal movement data is moveHMM (Michelot *et al*., 2016). A range of software exists for fitting HMMs, and for future work, some of these alternative packages will be explored.

A limitation of the technique in this work was that, as noted by Guo *et al*. (2009), the behaviour of cows can vary greatly, and the HMM could make corrections to classified instances that were in fact true representations of the behaviour state. This could be problematic for example, if the HMM was modelled on higher parity cows but applied to younger cows that exhibit more erratic behaviour characterized by short behaviour bouts. Furthermore, there were occasions where erroneous GPS instances would give the impression that cattle had moved very quickly and the HMM would identify this as a walking instance. Data smoothing techniques may be able to help in this respect as part of a pre-processing phase to reduce the number of instances that are highly unlikely to be true representations of cow behaviour. More also needs to be done to test and optimize different HMM configurations to suit different groups of animals

and possibly optimize for the time of day. This could be especially challenging if the aim of the analysis is to track changes in behaviour which may be indicative of disease progression.

10. Review of Chapter 3: Williams, M.L., James, W.P. and Rose, M.T., 2019. Variable segmentation and ensemble classifiers for predicting dairy cow behaviour. *Biosystems Engineering*, *178*, pp.156-167.

10.1 Overview

Given the few examples of VS techniques applied to cattle sensor data, it was necessary to test whether this technique could provide improvements in classification performance over FTS. Although no statistical comparison was made between the classification performances of algorithms derived from both segmentation techniques, the performance of classifiers was very strong under VS and probably better than those derived from FTS. This certainly seems like a promising strategy for future use, but VS must be sufficiently sensitive enough to detect subtle changes in data. This of course depends on the sensitivity required, but there were instances where the Changepoint method did not detect behaviour transitions. As reported, the method correctly identified approximately 90% of transitions and one would speculate that this could be improved upon if different sensors were used to capture the movement data. Given the various algorithms and tuning parameters available in the Changepoint package, more needs to be done to fully explore its potential in detecting transitions in data collected from cows. Furthermore, many other packages exist which could also be evaluated on performance.

The evaluated base learners performed very well, particularly the neural network and SL classifiers. No significant performance improvement was realised by using the Stacking ensemble. This does question whether these are necessary in this context, especially given the additional computational time required for training, although this is strongly dependent on the base learner configuration. However, once models are developed, testing time is usually less of an issue. Stacked generalization (Stacking) was used as the ensemble here, but others exist (Section 5.4.6) and should be

evaluated in future with VS. Furthermore, future studies should look to evaluate any biological insights or important information that can be gained from such ML algorithms on sensor data as was undertaken in Chapter 1. Such insights are lacking in the study of livestock movement ecology. Gaining a better understanding could enhance the recognition of non-normal behaviours more objectively and also lead to learning novel information about the interactions of livestock with the environment and with conspecifics. In Chapter 3, the output model of each ML algorithm was not evaluated for this purpose. Often, ML models are complex and only interpretable by their performance metrics on test data. More work is needed to recognize the functionality of models from a pattern-recognition and data-mining perspective that may be useful to movement ecologists. However, from a practical standpoint, most of the time, it is only the power of the model to recognise the behaviour of interest that is important.

In this work, the best algorithms performed as well as the ensembles explored by Dutta *et al.* (2015) and it may be that with improvements in transition recognition and enhanced algorithm hyperparameter tuning that even better performance can be achieved.

10.2 Variation in GPS data and implications for changepoint detection

Despite a good level of transition detection using Changepoint, there were often instances where the variance of the GPS data did not consistently change at the point of transition. This will eventually lead to some inaccuracies in the duration of behaviour expression. Mitigating this issue with GPS could be difficult given the sources of variability that are difficult to control for (Section 8.4.2). There are options in the Changepoint package to test the data for changes in the mean and also for changes in both the mean and the variance. We explicitly chose to detect changes in the variance

of speed because in preliminary tests, the magnitude of variance change was greater across behaviours than that of the mean. Also, variance change-detection was generally more effective, if more computationally intensive given that data transformations are required.

There were also instances in the data where, despite a visually observed transition having taken place, no change was detected by the changepoint algorithm (Figure 15). This could have occurred because of the way in which particular cows graze. If cows move very slowly while grazing, then there may have been very little difference between their moving speed and the speed of movement when they were stationary (resting). This could be influenced by pasture height or density for example leading to little requirement for cows to move forward while grazing (Gibb *et al.*, 1997). GPS signal quality could also have led to such anomalies driven by some of the factors already discussed. Another reason could be that the GPS receiver itself may have been damaged. GPS receivers were not re-tested for horizontal accuracy as this was not deemed necessary at the time (2014-2016). With this, it seems that GPS does introduce a significant amount of variation that is very difficult to control for and as such, data from other sensors should be tested in the same context for efficacy of transition detection.

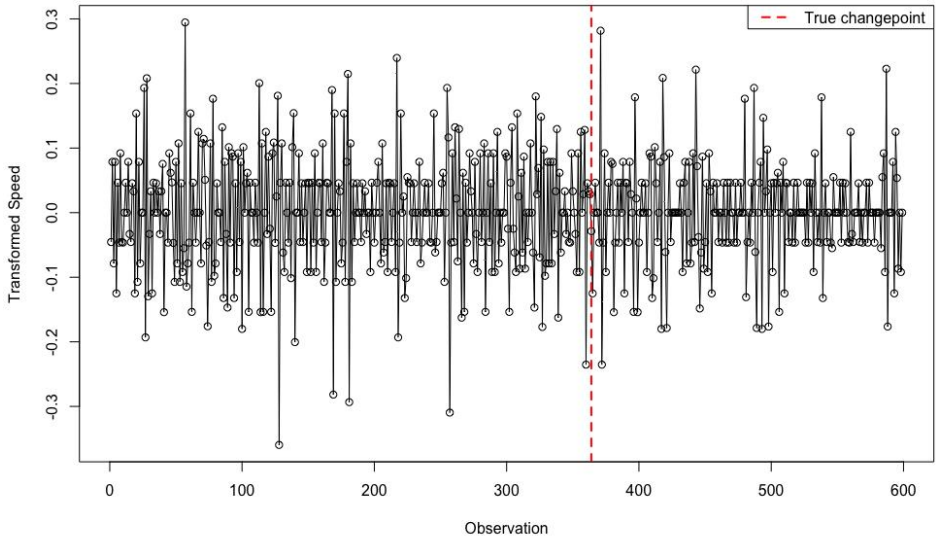


Figure 15. Example timeseries of Box-Cox transformed and first-differenced speed data containing 600 observations, each sampled at 5 s intervals. No changepoint was detected in this timeseries using the variance changepoint detection algorithm cpt.var. The transition was present at observation 364 where the cow moved from grazing to resting (vertical dashed line).

Further investigation could be undertaken in future to optimize the change-detection procedure from the perspective of the function itself. There are options to select different algorithms for change-detection, and others have had good success using the PELT algorithm for multiple changepoints (Madon & Hingrat, 2014). Users can also select the optimal penalty threshold to prevent overfitting and state the assumed distribution of the data. Clearly, there are many parameters that can be tested, and this exercise was outside the scope of this work. Few papers are available on the use of this package on other species and certainly, Chapter 3, as far as the candidate is aware is one of only two examples that have discussed its use for detecting changes in data collected from cattle (Shahriar *et al*., 2016).

10.3 Application of the changepoint method (Killick and Eckley, 2014) to accelerometer data

As an example, as to the efficacy of the changepoint method used in Chapter 3 on accelerometer data gathered from housed dairy cows, data were taken from the supporting material provided by Diosdado *et al.* (2015). The BinSeg algorithm of the changepoint function was used to search for changes in the mean of the data sampled at 1-min intervals. The data provided were the mean VeDBA (Section 5.2) and the mean static component of acceleration in the Y-axis (SCAY). In total there are 34 behaviour transitions in this dataset. Because the authors omitted drinking (amongst other behaviours) from their classification scheme, these behaviours were not considered for this exercise and so only transitions between feeding, lying and standing were included. The total number of transitions considered therefore were 31. The Changepoint function (cpt.mean) was configured so that Q = 40 (maximum number of changepoints to search for) and pen.value = 0.01 (theoretical type I error). For more information on the various algorithm configurations, see Killick and Eckley (2014). Figure 16 shows the results of the algorithm run on both datasets.

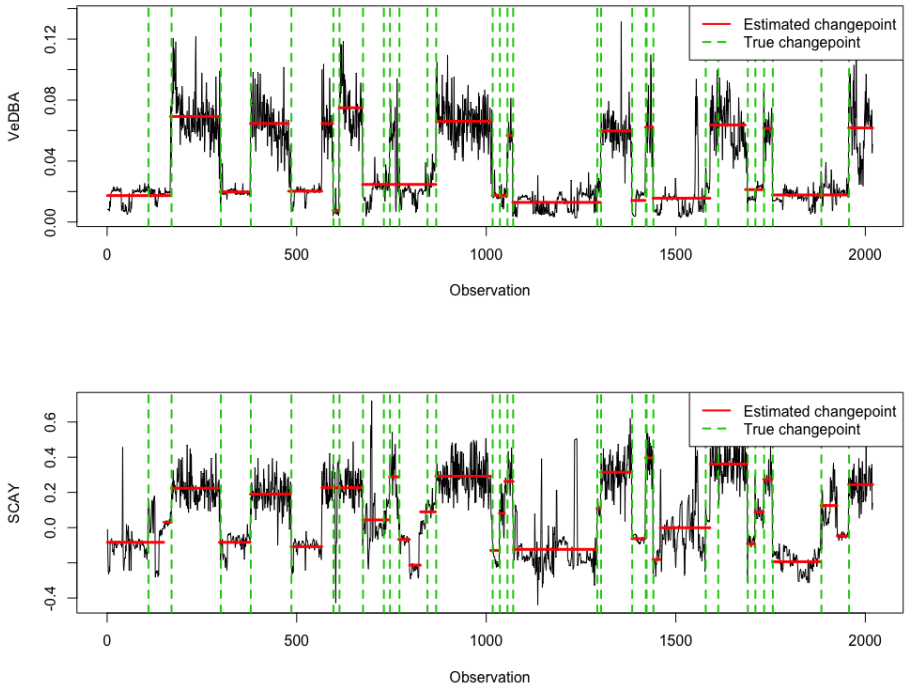


Figure 16. Changepoint detection algorithm used to search for changes in the mean vectorial dynamic body acceleration (VeDBA) and static component of acceleration in the Y-axis (SCAY) of data captured from neck-mounted accelerometers attached to the necks of housed dairy cows. Horizontal red lines indicate underlying fitted means (represent estimated behaviour boundaries). Vertical green dashes indicate boundaries of true transitions between behaviours. Dataset included 2,019 observations taken at 1-min intervals. Data taken from Diosdado *et al*. (2015).

The results show that the changepoint algorithm was effective in detecting behaviour transitions in the acceleration data. Overall, it seems that the algorithm was more sensitive to changes in the SCAY of the data and that this feature may be a more effective representation for detecting sensitive changes in the behaviour of dairy cows. The results also suggest that the algorithm underestimated the number of changepoints in the VeDBA data and overestimated on occasion in the SCAY data. Further supplementary data taken at different sampling intervals are provided with the publication (Diosdado *et al*., 2015), and it may be worth exploring the success of such changepoint algorithms on this data as well as larger datasets.

10.4 Other unsupervised changepoint detection methods

Changepoint detection algorithms can be categorized into two main types namely, 'online' and 'offline'. Offline algorithms retrospectively analyse data for significant changepoints and the majority of the techniques discussed earlier in this thesis fall under this category of algorithms although some can be adapted for online changepoint detection e.g. cpm (Section 4.1.2.2.2). Online algorithms on the other hand attempt to detect significant changepoints in almost real-time, processing each data point as it arrives. However, no changepoint algorithm can work in absolute real-time as algorithms require new data to be processed before a changepoint can be detected.

The choice of whether to process in real-time or retrospectively will be context-dependent and determined by factors such as the level of importance placed on any changes that may occur in the measured process (e.g. fault detection or falls in elderly individuals). Examples where both online and offline detection techniques discussed in the literature beyond movement ecology include methods used for human activity analysis (Brahim-Belhouari and Bermak, 2004), speech recognition (Rybach *et al.*, 2009) and medical condition monitoring (Bosc *et al.*, 2003). Here, likelihood and probabilistic-based methods are discussed which have potential for use in both PLF and movement ecology studies.

10.4.1 Likelihood ratio segmentation methods

Likelihood ratio methods (as used in Chapter 3, Section 2.2; Killick and Eckley, 2014) are based on comparing probability densities between two consequent intervals (Page, 1954; MacEachern *et al.*, 2007). A cumulative sum control chart (CUSUM) is another example of a likelihood ratio method. CUSUM can be used for detecting small shifts and variations in continuous data streams that deviate from the process target using a subsequence of retrospective data (Amiri *et al.*, 2012). If the variation in the

recorded measure is within the expected boundary then the process is considered to be operating on target (Amiri *et al.*, 2012). Conversely, if a sample deviates beyond the allowable boundary then the process is considered to be operating away from the target value. CUSUM control charts were used by Pastell *et al.* (2008) to detect changes in the weight distribution of dairy cows for the detection of lameness in a milking robot. This is an example of how a statistical process method can be used to monitor individual animals over time using an off-animal sensor that is not supported by a complex algorithm for classification. The benefit of these systems is that they evaluate long-term deviations in the measured output and so do not assume that all individuals fit a particular model of behavior e.g. a general ML classification algorithm for grazing. In this particular work, load cells were used to measure individual limb weight and CUSUM charts were developed to track weight distribution (mean ± SD) over time. For this application, approximately 15 days of historic data were needed for the system to make informed judgements as to the normal leg-load distribution and variation for individual cows. A similar process was applied to accelerometer data gathered from pregnant sows to detect farrowing (Pastell *et al.*, 2016). Here, trend and seasonal components were extracted from accelerometer data (neck-mounted accelerometers) using a dynamic linear modelling approach before applying a CUSUM chart to detect activity increases. The authors reported an average positive detection of farrowing activity (mean ± SD) 13 ± 4.8 h before farrowing with a sensitivity and specificity of 96.7% and 100% respectively.

10.4.2 Probabalistic segmentation methods

Two notable probabilistic changepoint detection methods exist, namely Bayesian and Gaussian methods. The majority of Bayesian changepoint detection methods had previously focused on offine, retrospective timeseries analysis (Smith,

1975; Stephens, 1994) but online methods have since been developed (Adams and Mackay, 2007). Online Bayesian techniques aim to generate an accuracte distribution of the forthcoming and unseen data in the timeseries using only past observed data. The posterior distribution is estimated over the timeseries data elapsed since the last detected changepoint. This elapsed time is defined as the 'run length' and increases by 1 if no changepoint is detected and returns to 0 when the state changes. The run length distribution is based on Bayes' theorem and takes a set of observations within the run as well as prior, likelihood and recursive components. The probability that the current datum belongs to the current state is represented by the likelihood term. After calculating the run length distribution and updating the corresponding statistics, a changepoint is detected if the run length has the highest probability in the distribution (Aminikhanghahi and Cook 2017).

A Gaussian process is another example of a probabilistic timeseries analysis method and can be defined as a collection of random variables, any number of which having a joint Gaussian distribution. In the case of Gaussian processes, timeseries data are defined as a noisy version of Gaussian distribution function values (observations are corrupted versions of the function values). The distribution function is specified by mean zero and a covariance function. A Gaussian process changepoint algorithm estimates the predictive distribution at a given point in the timeseries using previous timeseries values (Chandola and Vatsavai, 2010). Probability values are then computed for incoming data using the reference distribution and a threshold is used to determine when new observations do not follow the predicted distribution (Chandola and Vatsavai, 2011).

Several other changepoint detection methods exist but only a small number have been discussed in this thesis. Readers are referred to Aminikhanghahi and Cook (2017) for a comprehensive overview of methods.

10.5 Ensemble algorithm selection methods

The method undertaken to select the best base learners for the Stacking ensemble was a manual method described by Whalen and Pandey (2013). The weakness of this approach is that despite selecting the best available base learners on test data, the method does not consider the impact of each of these classifiers on the performance of the ensemble. A better approach could be to iteratively add the best classifiers to the ensemble and select those that improve the performance of the ensemble. This alternative method could become very difficult to perform manually especially if a number of candidate base learners are considered. Others have undertaken automatic base learner selection based on optimizing a certain evaluation function (Fan *et al*., 2002; Martınez-Munoz and Suárez, 2004). These techniques will be considered in future for ensemble optimization using packages such as that developed by Polley *et al*. (2018). In fact, a number of different ensemble selection methods exist and Tsoumakas *et al*. (2008) have reviewed some of these methods, namely, search-based methods, clustering-based methods and ranking-based methods.

Initially selecting a subset of the best base learners was thought necessary in the current work in order to reduce the pool of candidates for the ensemble. Indeed, this is a common method amongst publications (Whalen and Pandey, 2013). In hindsight, it may have been the case that some of those that performed the weakest may have provided additional diversity to the ensemble and may have improved the classification performance of the final ensemble. Diversity is of course an important characteristic of base learners and Whalen and Pandey (2013) demonstrate a technique of empirically

measuring diversity using Yule's *Q*-statistic (Yule, 1900). This statistic can be used to generate a contingency table with the data based on the classified labels of a pair of classifiers. With this, both classifiers can be assessed for similarity in their classified instances and the information used as a tool to support the level of diversity within the ensemble.

11. Conclusion

11.1 Future research

Progress is being made quickly in PLF research. There is no doubt that most future livestock production will, at least in developed regions, make use of precision-based means of measuring something about a particular component of the system. PLF methods are already being used in many intensively managed systems. Given that livestock production systems are at the forefront of some of the biggest challenges facing humanity such as global climate change (Rojas-Downing *et al*., 2017), PLF methods and technologies could help producers manage their systems more efficiently but also demonstrate the accountability of livestock industries to these wider challenges (Tullo *et al*., 2018). Having examined the literature, it is evident that the next decade will bring many advances in the ability to quickly recognize constraints on livestock production systems and solutions to quickly manage them. Advances in sensor hardware, data processing and pattern recognition will probably mean that data-driven management decisions will be possible at the level of the individual animal which could be beneficial for a range of welfare and production issues.

With regard to the dairy industry, broad areas where future research should be focused include (1) understanding the motivation of UK farmers to adopt (or not) PLF technologies. Limited research is available in this area, particularly from the perspective of UK farmers. Research should also look to (2) measure the impact that PLF interventions have had on farms both from an empirical and qualitative opinion-based perspective. This will be important to assess the impact of PLF intervention. Herd size (3) should also be assessed for relationships between performance and welfare as the evidence for these relationships is either not fully conclusive in existing research or is currently unavailable from a UK perspective. It would also be very useful

to see whether PLF adoption is more likely in larger herds. Finally (4), more research is needed to assess the behaviour of cows using various sensor systems and how management regimes, disease and other factors such as separation manifest in the data.

11.2 Combining sensors for livestock behaviour and movement

In this thesis, GPS receivers were used for data collection throughout and were applied to pasture-based cattle. Monitoring animals with sensors that can combine position and behaviour prediction could be very valuable in the context of Welsh livestock production systems and indeed any system where grazed grass is an important component of the diet. GPS could be important for future livestock monitoring to gain an understanding of livestock behaviour trends and also to better understand the movement of livestock. It is clear that more needs to be done to realise the benefits of GPS in this repect. Chapter 1 showed that a simple rule-based model could provide an insight into the movement associated with grazing. One of the weaknesses of GPS is the error associated with the estimates of location and movement due to issues such as satellite position relative to the receiver, and multipath errors due to buildings and other large objects. This could lead to erroneous movement models including those that represent behaviours such as grazing. Other sensors that have and will probably continue to improve estimates of behaviour in combination with GPS include accelerometers and magnetometers. Magnetometers were not discussed in this thesis previously, but these sensors can be built to fit livestock and indeed many species in discrete on-animal housing. Magnetometers are used to measure posture from the static data component with respect to the Earth's magnetic field and also dynamic movement. They have been used previously in conjunction with GPS and accelerometers for the classification of dairy cow behaviour at pasture (Dutta *et al*., 2015) and are becoming more popular in movement ecology (Chakravarty *et al*., 2019). Magnetometers are

reportedly able to record some movements that accelerometers can miss such as low acceleration behaviours (Williams *et al*., 2015). In Williams *et al*. (2017), the authors found some large differences between accelerometers and magnetometers in their capabilities to recognise certain behaviours. In future, it will be worth considering how these two sensor types can be leveraged for their complimentarity for recognising the various behaviours of livestock including the most fundamental such as grazing, lying and standing.

For housed production systems, LPS is likely to be an important tool for future use in monitoring behavioural interactions and localisation. Given that GPS is not fit for purpose in housed systems, LPS will probably be important from a research perspective in aiding the diagnosis of disease and for supporting other interesting work on herd behavior such as hierarchy and the interplay between social interactions and production disorders. In combination with on-animal sensors, both GPS and LPS will be able to provide a level of dimensionality not seen previously in dairy production systems and allow analysts the opportunity to learn far more about the dynamics of the production environment. There is also potential for LPS in grazed systems, allowing triangulation of animal location and reducing the need for a power-intensive system such as GPS and the uncertainties that come with it. Given these opportunities, there is great potential for future research in both extensive and intensive dairy production systems that were previously unavailable.

11.3 Final recommendations

Data-driven dairying is not new; in fact, performance recording has always been a key factor in improving system performance. The major difference now is that the data available are vast and dealing with this information can, at times seem a difficult task. With greater engineering capabilities, data can be processed quickly, but a key requirement will be that this information is useful to the individuals who work with livestock on a daily basis. Indeed, this project has highlighted the complexities involved with processing a relatively small quantity of data in order to make effective use of it in behaviour recognition. There are also several highlighted areas that will require further work in future so that cow-mounted sensor data can be fully understood and exploited. These challenges include:

(1) Finalizing a selection of behaviours that are biologically useful and realistic to identify and also measure automatically. These behaviours may vary depending on application, system type and farm objectives.

(2) Producing models of behaviour that are deployable and capable of classifying behaviours in continuous time (outside of the training and testing environment). These models need not necessarily follow the same format as those proposed in this thesis (developed via ML). Methods developed in the field of movement ecology are likely to be less computationally intensive (e.g. behavioural changepoint detection and localization). In conjunction with supervised behaviour classification methods, these techniques could provide an efficient means of behaviour recognition or to decipher patterns in data indicative of perturbations in normal behaviour.

(3) Measuring the impact of biological and environmental variability on the predictability of derived models.

(4) Supporting sensor systems with useful advice and information on the trajectory that farm managers should take given a particular event.

It is clear that the wealth of data available will allow researchers and farmers to gain a deep insight into the mechanisms that drive performance in production systems in future. This will be necessary to better inform farmers of any changes that require a managerial response and inform quality and welfare assurance. Sensor systems will also be important for allowing consumers to better understand livestock production, helping to provide confidence in an age where the story behind the food item is just as important as the quality of the product on the plate. Sensor systems will also allow for the precise monitoring of inputs and outputs which includes environmentally degrading by-products. This will help the livestock industry meet some of its global challenges. The applications of the techniques discussed in this thesis and in the wider literature also go beyond those of livestock production systems and in fact, many of the techniques are derived from the study of animal movement ecology and behaviour. The methods and works discussed in this thesis are also applicable to a range of other species. This includes zoo or park-kept animals (Whitham and Miller, 2016) that may require a level of observation necessary for precisely monitoring behaviour and behaviour change, or for optimizing conditions for those animals (Wijers *et al*., 2018).

Specifically, the experimental chapters in this thesis have contributed to better understanding cattle movement by combining methods used in animal movement ecology with those already being applied to livestock production research. The context of the work is unique to pasture-based livestock systems which are an important component of the UK agricultural economy. Simultaneously knowing the location and behaviour of livestock in extensive systems could help lead to more sustainable agricultural systems by making better and more efficient use of resources. Achieving

trust in such systems will require extensive work in improving the longevity of sensors, battery power and also demonstrating to farmers that the technology works. The information gained in this thesis could help direct further research in cattle grazing ecology. Specifically, it is hoped that it will make livestock scientists interested in PLF aware of the challenges that have and are being faced in the field of animal movement ecology and help bridge the gap between the two disciplines.

12. References

Abbasi, A.Z., Islam, N. and Shaikh, Z.A., 2014. A review of wireless sensors and networks' applications in agriculture. *Computer Standards & Interfaces, 36* (2), pp. 263-270.

Abeni, F., Petrera, F. and Galli, A., 2019. A Survey of Italian Dairy Farmers' Propensity for Precision Livestock Farming Tools. *Animals, 9* (5), pp. 202.

Adams, R.P. and MacKay, D.J., 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Agouridis, C.T., Stombaugh, T.S., Workman, S.R., Koostra, B.K. and Edwards, D.R., 2003. Examination of GPS collar capabilities and limitations for tracking animal movement in grazed watershed studies. In *2003 ASAE Annual Meeting*, pp. 1. American Society of Agricultural and Biological Engineers.

Agouridis, C.T., Stombaugh, T.S., Workman, S.R., Koostra, B.K., Edwards, D.R. and Vanzant, E.S., 2004. Suitability of a GPS collar for grazing studies. *Transactions of the ASAE, 47* (4), pp. 1321.

AHDB Dairy, 2019[a]. *Producer numbers*. [Online]. Available at: https://dairy.ahdb.org.uk/resources-library/market-information/farming-data/producer-numbers/#.XPGViVNKhUM. [Accessed 12/5/2019].

AHDB Dairy, 2019[b]. *Average herd size*. [Online]. Available at: https://dairy.ahdb.org.uk/resources-library/market-information/farming-data/average-herd-size/#.XNgEvxNKhUM. [Accessed 12/5/2019].

AHDB Dairy, 2019[c]. *Average milk yield*. [Online]. Available at: https://dairy.ahdb.org.uk/resources-library/market-information/farming-data/average-milk-yield/#.XNgGEBNKhUM. [Accessed 12/5/2019].

AHDB Dairy, 2015. *Dairy statistics. An insider's guide 2015*. [Online]. Available at: https://dairy.ahdb.org.uk/non_umbraco/download.aspx?media=21932. [Accessed 12/5/2019].

AHDB Dairy, 2013. *Summary Report. The structure of the GB dairy farming industry – what drives change?* [Online]. Available at: https://dairy.ahdb.org.uk/non_umbraco/download.aspx?media=14129. [Accessed 12/5/2019].

Allred, B.W., Fuhlendorf, S.D., Hovick, T.J., Dwayne Elmore, R., Engle, D.M. and Joern, A., 2013. Conservation implications of native and introduced ungulates in a changing climate. *Global Change Biology, 19* (6), pp. 1875-1883.

Alvarez, J.R., Arroqui, M., Mangudo, P., Toloza, J., Jatip, D., Rodríguez, J.M., Teyseyre, A., Sanz, C., Zunino, A., Machado, C. and Mateos, C., 2018. Body condition estimation on cows from depth images using Convolutional Neural Networks. *Computers and Electronics in Agriculture, 155*, pp. 12-22.

Aminikhanghahi, S. and Cook, D.J., 2017. A survey of methods for time series change point detection. *Knowledge and Information Systems, 51* (2), pp. 339-367.

Amiri, A. and Allahyari, S., 2012. Change point estimation methods for control chart postsignal diagnostics: a literature review. *Quality and Reliability Engineering International, 28* (7), pp. 673-685.

Amrine, D.E., White, B.J. and Larson, R.L., 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. *Computers and Electronics in Agriculture, 105*, pp. 9-19.

Archer, S.C., Mc Coy, F., Wapenaar, W. and Green, M.J., 2013. Association of season and herd size with somatic cell count for cows in Irish, English, and Welsh dairy herds. *The Veterinary Journal, 196* (3), pp. 515-521.

Arcidiacono, C., Porto, S.M.C., Mancino, M. and Cascone, G., 2017. Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioural activities from accelerometer data. *Computers and Electronics in Agriculture, 134*, pp. 124-134.

Arnott, G., Ferris, C.P. and O'Connell, N.E., 2017. Welfare of dairy cows in continuously housed and pasture-based production systems. *Animal, 11* (2), pp. 261-273.

Arsov, N., Pavlovski, M. and Kocarev, L., 2019. Stacking and stability. *arXiv preprint arXiv:1901.09134*.

Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science, 83*, pp. 1064-1069.

Augustin, M.A., Udabage, P., Juliano, P. and Clarke, P.T., 2013. Towards a more sustainable dairy industry: Integration across the farm–factory interface and the dairy factory of the future. *International Dairy Journal, 31* (1), pp. 2-11.

Ball, G., Aljabar, P., Arichi, T., Tusor, N., Cox, D., Merchant, N., Nongena, P., Hajnal, J.V., Edwards, A.D. and Counsell, S.J., 2016. Machine-learning to characterise neonatal functional connectivity in the preterm brain. *Neuroimage, 124*, pp. 267-275.

Barasona, J.A., Latham, M.C., Acevedo, P., Armenteros, J.A., Latham, A.D.M., Gortazar, C., Carro, F., Soriguer, R.C. and Vicente, J., 2014. Spatiotemporal interactions between wild boar and cattle: implications for cross-species disease transmission. *Veterinary Research, 45* (1), pp. 122.

Barkema, H.W., Von Keyserlingk, M.A.G., Kastelic, J.P., Lam, T.J.G.M., Luby, C., Roy, J.P., LeBlanc, S.J., Keefe, G.P. and Kelton, D.F., 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. *Journal of Dairy Science, 98* (11), pp. 7426-7445.

Barker, Z.E., Diosdado, J.V., Codling, E.A., Bell, N.J., Hodges, H.R., Croft, D.P. and Amory, J.R., 2018. Use of novel sensors combining local positioning and acceleration to measure feeding behavior differences associated with lameness in dairy cattle. *Journal of Dairy Science, 101* (7), pp. 6310-6321.

Basheer, I.A. and Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43* (1), pp. 3-31.

Batista, G.E., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6* (1), pp. 20-29.

Behmann, J., Hendriksen, K., Müller, U., Büscher, W. and Plümer, L., 2016. Support Vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors. *Geoinformatica, 20* (4), pp. 693-714.

Benaissa, S., Tuyttens, F.A., Plets, D., De Pessemier, T., Trogh, J., Tanghe, E., Martens, L., Vandaele, L., Van Nuffel, A., Joseph, W. and Sonck, B., 2019. On the use of on-cow accelerometers for the classification of behaviours in dairy barns. *Research in Veterinary Science, 125*, pp. 425-433.

Benaissa, S., Tuyttens, F.A., Plets, D., Cattrysse, H., Martens, L., Vandaele, L., Joseph, W. and Sonck, B., 2019. Classification of ingestive-related cow behaviours using RumiWatch halter and neck-mounted accelerometers. *Applied Animal Behaviour Science, 211*, pp. 9-16.

Benhamou, S., 1992. Efficiency of area-concentrated searching behaviour in a continuous patchy environment. *Journal of Theoretical Biology, 159* (1), pp. 67-81.

Benhamou, S., 2006. Detecting an orientation component in animal paths when the preferred direction is individual-dependent. *Ecology, 87* (2), pp. 518-528.

Benhamou, S., 2007. How many animals really do the Lévy walk? *Ecology*, *88* (8), pp. 1962-1969.

Bestley, S., Jonsen, I., Harcourt, R.G., Hindell, M.A. and Gales, N.J., 2016. Putting the behavior into animal movement modeling: Improved activity budgets from use of ancillary tag information. *Ecology and Evolution*, *6* (22), pp. 8243-8255.

Bhavsar, A.R. and Arolkar, H.A., 2014, March. Multidimensional association rule-based data mining technique for cattle health monitoring using wireless sensor network. In *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 810-814. IEEE.

Bicalho, R.C., Cheong, S.H., Cramer, G. and Guard, C.L., 2007. Association between a visual and an automated locomotion score in lactating Holstein cows. *Journal of Dairy Science*, *90* (7), pp. 3294-3300.

Bifet, A., Holmes, G., Kirkby, R. and Pfahringer, B., 2010. Moa: Massive online analysis. *Journal of Machine Learning Research*, *11*, pp. 1601-1604.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Oxford university press.

Bjørneraas, K., Van Moorter, B., Rolandsen, C.M. and Herfindal, I., 2010. Screening global positioning system location data for errors using animal movement characteristics. *The Journal of Wildlife Management*, *74* (6), pp. 1361-1366.

Bom, R.A., Bouten, W., Piersma, T., Oosterbeek, K. and van Gils, J.A., 2014. Optimizing acceleration-based ethograms: the use of variable-time versus fixed-time segmentation. *Movement Ecology*, *2* (1), pp. 6.

Borchers, M.R., Chang, Y.M., Tsai, I.C., Wadsworth, B.A. and Bewley, J.M., 2016. A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. *Journal of Dairy Science*, *99* (9), pp. 7458-7466.

Borchers, M.R. and Bewley, J.M., 2015. An assessment of producer precision dairy farming technology use, prepurchase considerations, and usefulness. *Journal of Dairy Science*, *98* (6), pp. 4198-4205.

Bosc, M., Heitz, F., Armspach, J.P., Namer, I., Gounot, D. and Rumbach, L., 2003. Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage*, *20* (2), pp. 643-656.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30* (7), pp. 1145-1159.

Brahim-Belhouari, S. and Bermak, A., 2004. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, *47* (4), pp. 705-712.

Breed, G.A., Bowen, W.D., McMillan, J.I. and Leonard, M.L., 2006. Sexual segregation of seasonal foraging habitats in a non-migratory marine mammal. *Proceedings of the Royal Society B: Biological Sciences*, *273* (1599), pp. 2319-2326.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, *24* (2), pp. 123-140.

Breiman, L., 2001. Random forests. *Machine Learning*, *45* (1), pp. 5-32.

Britt, J.H., Cushman, R.A., Dechow, C.D., Dobson, H., Humblot, P., Hutjens, M.F., Jones, G.A., Ruegg, P.S., Sheldon, I.M. and Stevenson, J.S., 2018. Invited review: Learning from the future - A vision for dairy farms and cows in 2067. *Journal of Dairy Science*, *101* (5), pp. 3722-3741.

Brown, R., 1828. XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, *4* (21), pp. 161-173.

Caja, G., Castro-Costa, A. and Knight, C.H., 2016. Engineering to support wellbeing of dairy animals. *Journal of Dairy Research*, *83* (2), pp. 136-147.

Calamari, L.U.I.G.I., Soriani, N., Panella, G., Petrera, F., Minuti, A. and Trevisi, E.R.M.I.N.I.O., 2014. Rumination time around calving: An early signal to detect cows at greater risk of disease. *Journal of Dairy Science*, *97* (6), pp. 3635-3647.

Calenge, C., Dray, S. and Royer-Carenzi, M., 2009. The concept of animals' trajectories from a data analysis perspective. *Ecological Informatics*, *4* (1), pp. 34-41.

Caraviello, D.Z., Weigel, K.A., Craven, M., Gianola, D., Cook, N.B., Nordlund, K.V., Fricke, P.M. and Wiltbank, M.C., 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. *Journal of Dairy Science*, *89* (12), pp. 4703-4722.

Carpentier, L., Berckmans, D., Youssef, A., Berckmans, D., van Waterschoot, T., Johnston, D., Ferguson, N., Earley, B., Fontana, I., Tullo, E. and Guarino, M., 2018. Automatic cough detection for bovine respiratory disease in a calf house. *Biosystems Engineering*, *173*, pp. 45-56.

Caruana, R., Niculescu-Mizil, A., Crew, G. and Ksikes, A., 2004. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine Learning*, pp. 18. ACM.

Chambers, A.R.M., Hodgson, J. and Milne, J.A., 1981. The development and use of equipment for the automatic recording of ingestive behaviour in sheep and cattle. *Grass and Forage Science*, *36* (2), pp. 97-105.

Chandola, V. and Vatsavai, R.R., 2010, September. Scalable Time Series Change Detection for Biomass Monitoring Using Gaussian Process. In *CIDU*, pp. 69-82.

Chandola, V. and Vatsavai, R.R., 2011. A Gaussian process based online change detection algorithm for monitoring periodic time series. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 95-106. Society for Industrial and Applied Mathematics.

Chapinal, N., Liang, Y., Weary, D.M., Wang, Y. and Von Keyserlingk, M.A.G., 2014. Risk factors for lameness and hock injuries in Holstein herds in China. *Journal of Dairy Science*, *97* (7), pp. 4309-4316.

Charlton, G.L., Bouffard, V., Gibbons, J., Vasseur, E., Haley, D.B., Pellerin, D., Rushen, J. and de Passillé, A.M., 2016. Can automated measures of lying time help assess lameness and leg lesions on tie-stall dairy farms? *Applied Animal Behaviour Science*, *175*, pp. 14-22.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, pp. 321-357.

Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K.W., 2003. SMOTEBoost: Improving prediction of the minority class in boosting. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 107-119. Springer, Berlin, Heidelberg.

Chowdhury, M.F.R., Selouani, S.A. and O'Shaughnessy, D., 2012. Bayesian on-line spectral change point detection: a soft computing approach for on-line ASR. *International Journal of Speech Technology*, *15* (1), pp. 5-23.

Clapham, W.M., Fedders, J.M., Beeman, K. and Neel, J.P., 2011. Acoustic monitoring system to quantify ingestive behavior of free-grazing cattle. *Computers and Electronics in Agriculture*, *76* (1), pp. 96-104.

Clements, R., 2015. Tackling lameness in sheep and evaluating progress. *Veterinary Record*, *176* (20), pp. 519-520.

Codling, E.A. and Hill, N.A., 2005. Sampling rate effects on measurements of correlated and biased random walks. *Journal of Theoretical Biology*, *233* (4), pp. 573-588.

Codling, E.A., Plank, M.J. and Benhamou, S., 2008. Random walk models in biology. *Journal of the Royal Society Interface*, *5* (25), pp. 813-834.

Cohen, J., 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, W.W., 1995. Fast effective rule induction. In *Machine Learning Proceedings 1995*, pp. 115-123. Morgan Kaufmann.

Cole, J.B. and VanRaden, P.M., 2018. Symposium review: Possibilities in an age of genomics: The future of selection indices. *Journal of Dairy Science*, *101* (4), pp. 3686-3701.

Corner-Thomas, R.A., Kenyon, P.R., Morris, S.T., Ridler, A.L., Hickson, R.E., Greer, A.W., Logan, C.M. and Blair, H.T., 2015. Influence of demographic factors on the use of farm management tools by New Zealand farmers. *New Zealand Journal of Agricultural Research*, *58* (4), pp. 412-422.

Cowton, J., Kyriazakis, I., Plötz, T. and Bacardit, J., 2018. A combined deep learning gru-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors. *Sensors*, *18* (8), pp. 2521.

DeCesare, N.J., Squires, J.R. and Kolbe, J.A., 2005. Effect of forest canopy on GPS-based movement data. *Wildlife Society Bulletin*, *33* (3), pp. 935-941.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, pp. 1-30.

de Weerd, N., van Langevelde, F., van Oeveren, H., Nolet, B.A., Kölzsch, A., Prins, H.H. and de Boer, W.F., 2015. Deriving animal behaviour from high-frequency GPS: tracking cows in open and forested habitat. *PloS One*, *10* (6), p.e0129030.

Dewhurst, R.J., Shingfield, K.J., Lee, M.R. and Scollan, N.D., 2006. Increasing the concentrations of beneficial polyunsaturated fatty acids in milk produced by dairy cows in high-forage systems. *Animal Feed Science and Technology*, *131* (3-4), pp. 168-206.

Dietterich, T. G., 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1-15. Springer, Berlin, Heidelberg.

Diosdado, J.A.V., Barker, Z.E., Hodges, H.R., Amory, J.R., Croft, D.P., Bell, N.J. and Codling, E.A., 2015. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. *Animal Biotelemetry*, *3* (1), pp.15.

Diosdado, J.A.V., Barker, Z.E., Hodges, H.R., Amory, J.R., Croft, D.P., Bell, N.J. and Codling, E.A., 2018. Space-use patterns highlight behavioural differences linked to lameness, parity, and days in milk in barn-housed dairy cows. *PloS One*, *13* (12), p.e0208424.

Doyle, L.P., Gordon, A.W., Abernethy, D.A. and Stevens, K., 2014. Bovine tuberculosis in Northern Ireland: risk factors associated with time from post-outbreak test to subsequent herd breakdown. *Preventive Veterinary Medicine*, *116* (1-2), pp. 47-55.

Dragon, A.C., Bar-Hen, A., Monestiez, P. and Guinet, C., 2012. Comparative analysis of methods for inferring successful foraging areas from Argos and GPS tracking data. *Marine Ecology Progress Series*, *452*, pp. 253-267.

Drake, M.A., 2007. Invited review: Sensory analysis of dairy foods. *Journal of Dairy Science*, *90* (11), pp. 4925-4937.

Duncan, M.J., Badland, H.M. and Mummery, W.K., 2009. Applying GPS to enhance understanding of transport-related physical activity. *Journal of Science and Medicine in Sport*, *12* (5), pp. 549-556.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3 (3), pp. 32-57.

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G. and Henry, D., 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, *111*, pp. 18-28.

Dzialak, M.R., Olson, C.V., Webb, S.L., Harju, S.M. and Winstead, J.B., 2015. Incorporating within-and between-patch resource selection in identification of critical habitat for brood-rearing greater sage-grouse. *Ecological Processes*, *4* (1), pp. 5.

Edelhoff, H., Signer, J. and Balkenhol, N., 2016. Path segmentation for beginners: an overview of current methods for detecting changes in animal movement patterns. *Movement Ecology*, *4* (1), pp.21.

Falzon, G., Schneider, D., Trotter, M. and Lamb, D.W., 2013. A relationship between faecal egg counts and the distance travelled by sheep. *Small Ruminant Research*, *111* (1-3), pp.171-174.

Fan, W., Chu, F., Wang, H. and Yu, P.S., 2002. Pruning and dynamic scheduling of cost-sensitive ensembles. *American Association for Artificial Intelligence*, pp. 146-151.

Fauchald, P. and Tveraa, T., 2003. Using first-passage time in the analysis of area-restricted search and habitat selection. *Ecology*, *84* (2), pp.282-288.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, *27* (8), pp. 861-874.

Feldman, A. and Balch, T., 2004. Representing honey bee behavior for recognition using human trainable models. *Adaptive Behavior*, *12* (3-4), pp. 241-250.

Ferguson, J.D. and Skidmore, A., 2013. Reproductive performance in a select sample of dairy herds. *Journal of Dairy Science, 96* (2), pp. 1269-1289.

Firk, R., Stamer, E., Junge, W. and Krieter, J., 2002. Automation of oestrus detection in dairy cows: a review. *Livestock Production Science, 75* (3), pp. 219-232.

Fischer, M., Parkins, K., Maizels, K., Sutherland, D.R., Allan, B.M., Coulson, G. and Di Stefano, J., 2018. Biotelemetry marches on: A cost-effective GPS device for monitoring terrestrial wildlife. *PloS One, 13* (7), p.e0199617.

Frair, J.L., Merrill, E.H., Visscher, D.R., Fortin, D., Beyer, H.L. and Morales, J.M., 2005. Scales of movement by elk (*Cervus elaphus*) in response to heterogeneity in forage resources and predation risk. *Landscape Ecology, 20* (3), pp. 273-287.

Frank, E. and Witten, I.H., 1998. Generating accurate rule sets without global optimization. Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Franke, A., Caelli, T., Kuzyk, G. and Hudson, R.J., 2006. Prediction of wolf (*Canis lupus*) kill-sites using hidden Markov models. *Ecological Modelling, 197* (1-2), pp. 237-246.

Friedman, N., Geiger, D. and Goldszmidt, M., 1997. Bayesian network classifiers. *Machine Learning, 29* (2-3), pp. 131-163.

Friedman, J., Hastie, T. and Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics, 28* (2), pp. 337-407.

Fuller, F., Huang, J., Ma, H. and Rozelle, S., 2006. Got milk? The rapid rise of China's dairy sector and its future prospects. *Food Policy, 31* (3), pp. 201-215.

Fürnkranz, J. and Widmer, G., 1994. Incremental reduced error pruning. In *Machine Learning Proceedings,* pp. 70-77. Morgan Kaufmann.

Galli, J.R., Cangiano, C.A., Milone, D.H. and Laca, E.A., 2011. Acoustic monitoring of short-term ingestive behavior and intake in grazing sheep. *Livestock Science*, *140* (1-3), pp. 32-41.

Ganskopp, D.C. and Johnson, D.D., 2007. GPS error in studies addressing animal movements and activities. *Rangeland Ecology & Management*, *60* (4), pp. 350-358.

Garcia, S. and Herrera, F., 2008. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, *9*, pp. 2677-2694.

García, S., Fernández, A., Luengo, J. and Herrera, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180* (10), pp. 2044-2064.

Gargiulo, J.I., Eastwood, C.R., Garcia, S.C. and Lyons, N.A., 2018. Dairy farmers with larger herd sizes adopt more precision dairy technologies. *Journal of Dairy Science*, *101* (6), pp. 5466-5473.

Gibb, M.J., Huckle, C.A., Nuthall, R. and Rook, A.J., 1997. Effect of sward surface height on intake and grazing behaviour by lactating Holstein Friesian cows. *Grass and Forage Science*, *52* (3), pp. 309-321.

Gleiss, A.C., Wilson, R.P. and Shepard, E.L., 2011. Making overall dynamic body acceleration work: on the theory of acceleration as a proxy for energy expenditure. *Methods in Ecology and Evolution*, *2* (1), pp. 23-33.

Godsk, T. and Kjærgaard, M.B., 2011. High classification rates for continuous cow activity recognition using low-cost GPS positioning sensors and standard machine learning techniques. In *Industrial Conference on Data Mining*, pp. 174-188. Springer, Berlin, Heidelberg.

González, L.A., Bishop-Hurley, G.J., Handcock, R.N. and Crossman, C., 2015. Behavioral classification of data from collars containing motion sensors in grazing cattle. *Computers and Electronics in Agriculture*, *110*, pp. 91-102.

González, L.A., Tolkamp, B.J., Coffey, M.P., Ferret, A. and Kyriazakis, I., 2008. Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. *Journal of Dairy Science*, *91* (3), pp. 1017-1028.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT press.

Gredell, D.A., Schroeder, A.R., Belk, K.E., Broeckling, C.D., Heuberger, A.L., Kim, S.Y., King, D.A., Shackelford, S.D., Sharp, J.L., Wheeler, T.L. and Woerner, D.R., 2019. Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data. *Scientific Reports*, *9* (1), pp. 5721.

Green, J.A., Halsey, L.G., Wilson, R.P. and Frappell, P.B., 2009. Estimating energy expenditure of animals using the accelerometry technique: activity, inactivity and comparison with the heart-rate technique. *Journal of Experimental Biology*, *212* (4), pp. 471-482.

Guo, Y., Poulton, G., Corke, P., Bishop-Hurley, G.J., Wark, T. and Swain, D.L., 2009. Using accelerometer, high sample rate GPS and magnetometer data to develop a cattle movement and behaviour model. *Ecological Modelling*, *220* (17), pp. 2068-2075.

Guo, L., Welch, M., Dobos, R., Kwan, P. and Wang, W., 2018. Comparison of grazing behaviour of sheep on pasture with different sward surface heights using an inertial measurement unit sensor. *Computers and Electronics in Agriculture*, *150*, pp. 394-401.

Gurarie, E., Andrews, R.D. and Laidre, K.L., 2009. A novel method for identifying behavioural changes in animal movement data. *Ecology Letters*, *12* (5), pp. 395-408.

Gurarie, E., Bracis, C., Delgado, M., Meckley, T.D., Kojola, I. and Wagner, C.M., 2016. What is the animal doing? Tools for exploring behavioural structure in animal movements. *Journal of Animal Ecology*, *85* (1), pp. 69-84.

Gurarie, E., 2015. bcpa: behavioral change point analysis of animal movement. R package version 1.1.

Gygax, L., Neisen, G. and Bollhalder, H., 2007. Accuracy and validation of a radar-based automatic local position measurement system for tracking dairy cows in free-stall barns. *Computers and Electronics in Agriculture*, *56* (1), pp. 23-33.

Hand, D.J. and Till, R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45* (2), pp. 171-186.

Handcock, R., Swain, D., Bishop-Hurley, G., Patison, K., Wark, T., Valencia, P., Corke, P. and O'Neill, C., 2009. Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collars and satellite remote sensing. *Sensors*, *9* (5), pp. 3586-3603.

Harris, R.B., Fancy, S.G., Douglas, D.C., Garner, G.W., Amstrup, S.C., McCabe, T.R. and Pank, L.F., 1990. *Tracking wildlife by satellite: current systems and performance* (No. 30). US Department of the Interior, Fish and Wildlife Serivce.

Harte, D., 2015. HiddenMarkov: Hidden Markov Models. *Statistics Research Associates, Wellington. R package version*, pp. 1-8.

Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS136. A K-means algorithm. *Journal of the Royal Statistical Society Series C* (*Applied Statistics)*, *28* (1), pp. 100-108.

Healey, S.P., Cohen, W.B., Yang, Z., Brewer, C.K., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Hughes, M.J., Kennedy, R.E. and Loveland, T.R., 2018. Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment*, *204*, pp. 717-728.

Higaki, S., Miura, R., Suda, T., Andersson, L.M., Okada, H., Zhang, Y., Itoh, T., Miwakeichi, F. and Yoshioka, K., 2019. Estrous detection by continuous measurements of vaginal temperature and conductivity with supervised machine learning in cattle. *Theriogenology*, *123*, pp. 90-99.

Hijazi, S., Kumar, R. and Rowen, C., 2015. Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*.

Himmelmann, L., 2010. HMM: Hidden Markov Models. *R package version*, *1*.

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11* (1), pp. 63-90.

Homburger, H., Schneider, M.K., Hilfiker, S. and Lüscher, A., 2014. Inferring behavioral states of grazing livestock from high-frequency position data alone. *PLoS One*, *9* (12), p.e114522.

Hostiou, N., Fagon, J., Chauvat, S., Turlot, A., Kling-Eveillard, F., Boivin, X. and Allain, C., 2016. Impact of precision livestock farming on work and human-animal interactions on dairy farms. A review. *Biotechnologie, Agronomie, Société et Environnement/Biotechnology, Agronomy, Society and Environment*, *21* (4), pp. 268-275.

Hsu, J.L., Hung, P.C., Lin, H.Y. and Hsieh, C.H., 2015. Applying under-sampling techniques and cost-sensitive learning methods on risk assessment of breast cancer. *Journal of Medical Systems, 39* (4), pp. 40.

Huhtala, A., Suhonen, K., Mäkelä, P., Hakojärvi, M. and Ahokas, J., 2007. Evaluation of instrumentation for cow positioning and tracking indoors. *Biosystems Engineering, 96* (3), pp. 399-405.

Hulbert, I.A. and French, J., 2001. The accuracy of GPS for wildlife telemetry and habitat mapping. *Journal of Applied Ecology*, *38* (4), pp. 869-878.

Hurford, A., 2009. GPS measurement error gives rise to spurious 180 turning angles and strong directional biases in animal movement data. *PLoS One, 4* (5), p.e5632.

Huzzey, J.M., Veira, D.M., Weary, D.M. and Von Keyserlingk, M.A.G., 2007. Prepartum behavior and dry matter intake identify dairy cows at risk for metritis. *Journal of Dairy Science, 90* (7), pp. 3220-3233.

Ingvartsen, K.L., Dewhurst, R.J. and Friggens, N.C., 2003. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livestock Production Science*, *83* (2-3), pp. 277-308.

ION. 1997. ION STD 101: Recommended Test Procedures for GPS Receivers. Alexandria, Va. Institute of Navigation Standards.

Ito, K., Weary, D.M. and Von Keyserlingk, M.A.G., 2009. Lying behavior: Assessing within-and between-herd variation in free-stall-housed dairy cows. *Journal of Dairy Science*, *92* (9), pp. 4412-4420.

Ito, K., Von Keyserlingk, M.A.G., LeBlanc, S.J. and Weary, D.M., 2010. Lying behavior as an indicator of lameness in dairy cows. *Journal of Dairy Science, 93* (8), pp. 3553-3560.

Jackson, C.H., 2011. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software, 38* (8), pp. 1-29.

Jahns, G., 2008. Call recognition to identify cow conditions - A call-recogniser translating calls to text. *Computers and Electronics in Agriculture, 62* (1), pp. 54-58.

Jensen, M.B., 2012. Behaviour around the time of calving in dairy cows. *Applied Animal Behaviour Science, 139* (3-4), pp. 195-202.

John, G.H. and Langley, P., 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pp. 338-345. Morgan Kaufmann Publishers Inc.

Johnson, A.R., Wiens, J.A., Milne, B.T. and Crist, T.O., 1992. Animal movements and population dynamics in heterogeneous landscapes. *Landscape Ecology, 7* (1), pp. 63-75.

Jonsen, I.D., Myers, R.A. and Flemming, J.M., 2003. Meta-analysis of animal movement using state-space models. *Ecology, 84* (11), pp. 3055-3063.

Jonsen, I.D., Myers, R.A. and James, M.C., 2006. Robust hierarchical state–space models reveal diel variation in travel rates of migrating leatherback turtles. *Journal of Animal Ecology, 75* (5), pp. 1046-1057.

Jonsen, I.D., Myers, R.A. and James, M.C., 2007. Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model. *Marine Ecology Progress Series, 337*, pp. 255-264.

Jurdak, R., Sommer, P., Kusy, B., Kottege, N., Crossman, C., Mckeown, A. and Westcott, D., 2013. Camazotz: multimodal activity-based GPS sampling. In *Proceedings of the 12th international conference on Information processing in sensor networks*, pp. 67-78. ACM.

Kamilaris, A., Kartakoullis, A. and Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture, 143*, pp. 23-37.

Kamilaris, A. and Prenafeta-Boldú, F.X., 2018. A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, *156* (3), pp. 312-322.

Kareiva, P. and Odell, G., 1987. Swarms of predators exhibit “preytaxis" if individual predators use area-restricted search. *The American Naturalist*, *130* (2), pp. 233-270.

Kashiha, M., Bahr, C., Ott, S., Moons, C.P., Niewold, T.A., Ödberg, F.O. and Berckmans, D., 2014. Automatic weight estimation of individual pigs using image analysis. *Computers and Electronics in Agriculture, 107*, pp. 38-44.

Kaufman, E.I., LeBlanc, S.J., McBride, B.W., Duffield, T.F. and DeVries, T.J., 2016. Association of lying behavior and subclinical ketosis in transition dairy cows. *Journal of Dairy Science*, *99* (9), pp. 7473-7480.

Kawahara, Y. and Sugiyama, M., 2012. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5* (2), pp. 114-127.

Khalaf, M., Hussain, A.J., Keight, R., Al-Jumeily, D., Fergus, P., Keenan, R. and Tso, P., 2017. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. *Neurocomputing, 228,* pp. 154-164.

Killeen, J., Thurfjell, H., Ciuti, S., Paton, D., Musiani, M. and Boyce, M.S., 2014. Habitat selection during ungulate dispersal and exploratory movement at broad and fine scale with implications for conservation management. *Movement Ecology, 2* (1), pp. 15.

Killick, R., Fearnhead, P. and Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107* (500), pp. 1590-1598.

Killick, R. and Eckley, I., 2014. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, *58* (3), pp. 1-19.

Kim, H.T., Choi, H.L., Lee, D.W. and Yoon, Y.C., 2005. Recognition of individual Holstein cattle by imaging body patterns. *Asian-Australasian Journal of Animal Sciences*, *18* (8), pp. 1194-1198.

King, M.T.M., Dancy, K.M., LeBlanc, S.J., Pajor, E.A. and DeVries, T.J., 2017. Deviations in behavior and productivity data before diagnosis of health disorders in cows milked with an automated system. *Journal of Dairy Science*, *100* (10), pp. 8358-8371.

Klein, H. G., Isselbacher, K. J., Braunwald, E., Wilson, J. D., Martin, J. B., Fauci, A. S. and Kasper, D. L., 1994. Blood groups and blood transfusion. *Harrison's principles of internal medicine. 13th ed. New York: McGraw-Hill*, 1788-93.

Koeck, A., Loker, S., Miglior, F., Kelton, D.F., Jamrozik, J. and Schenkel, F.S., 2014. Genetic relationships of clinical mastitis, cystic ovaries, and lameness with milk yield and somatic cell score in first-lactation Canadian Holsteins. *Journal of Dairy Science*, *97* (9), pp. 5806-5813.

Kohavi, R., 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, *96*, pp. 202-207.

Koliopoulos, A.K., Yiapanis, P., Tekiner, F., Nenadic, G. and Keane, J., 2015. A parallel distributed weka framework for big data mining using spark. In *IEEE International Congress on Big Data*, pp. 9-16. IEEE.

Kos, T., Markezic, I. and Pokrajcic, J., 2010. Effects of multipath reception on GPS positioning performance. In *Proceedings ELMAR-2010*, pp. 399-402. IEEE.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, pp. 8-17.

Krogh, A., 2008. What are artificial neural networks? *Nature Biotechnology, 26* (2), pp. 195.

Krug, C., Haskell, M.J., Nunes, T. and Stilwell, G., 2015. Creating a model to detect dairy cattle farms with poor welfare using a national database. *Preventive Veterinary Medicine, 122* (3), pp. 280-286.

Kumar, S., Pandey, A., Satwik, K.S.R., Kumar, S., Singh, S.K., Singh, A.K. and Mohan, A., 2018. Deep learning framework for recognition of cattle using muzzle point image pattern. *Measurement, 116*, pp. 1-17.

Kuncheva, L.I. and Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51* (2), pp. 181-207.

Kwong, K.H., Wu, T.T., Goh, H.G., Sasloglou, K., Stephen, B., Glover, I., Shen, C., Du, W., Michie, C. and Andonovic, I., 2012. Practical considerations for wireless sensor networks in cattle monitoring applications. *Computers and Electronics in Agriculture, 81*, pp. 33-44.

Lai, T.L., 1995. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Methodological), 57* (4), pp. 613-644.

Landwehr, N., Hall, M. and Frank, E., 2005. Logistic model trees. *Machine Learning, 59* (1-2), pp. 161-205.

Lange, T., Braun, M.L., Roth, V. and Buhmann, J.M., 2003. Stability-based model selection. In *Advances in Neural Information Processing Systems*, pp. 633-642.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D. and Morales, J.M., 2012. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, *93* (11), pp. 2336-2342.

Laube, P. and Purves, R.S., 2011. How fast is a cow? Cross-scale analysis of movement data. *Transactions in GIS*, *15* (3), pp. 401-418.

Leach, K.A., Whay, H.R., Maggs, C.M., Barker, Z.E., Paul, E.S., Bell, A.K. and Main, D.C.J., 2010. Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. *Research in Veterinary Science*, *89* (2), pp. 311-317.

Lee, J., Jin, L., Park, D. and Chung, Y., 2016. Automatic recognition of aggressive behavior in pigs using a kinect depth sensor. *Sensors*, *16* (5), pp. 631.

Ledgerwood, D.N., Winckler, C. and Tucker, C.B., 2010. Evaluation of data loggers, sampling intervals, and editing techniques for measuring the lying behavior of dairy cattle. *Journal of Dairy Science*, *93* (11), pp. 5129-5139.

Leiber, F., Holinger, M., Zehner, N., Dorn, K., Probst, J.K. and Neff, A.S., 2016. Intake estimation in dairy cows fed roughage-based diets: An approach based on chewing behaviour measurements. *Applied Animal Behaviour Science*, *185*, pp. 9-14.

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T.A., Watanabe, Y.Y., Murgatroyd, M. and Papastamatiou, Y.P., 2017. Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, *8* (2), pp. 161-173.

Liakos, K., Busato, P., Moshou, D., Pearson, S. and Bochtis, D., 2018. Machine learning in agriculture: A review. *Sensors*, *18* (8), pp. 2674.

Liboreiro, D.N., Machado, K.S., Silva, P.R., Maturana, M.M., Nishimura, T.K., Brandão, A.P., Endres, M.I. and Chebel, R.C., 2015. Characterization of peripartum rumination and activity of cows diagnosed with metabolic and uterine diseases. *Journal of Dairy Science, 98* (10), pp. 6812-6827.

Li, X., Zhang, X., Ren, X., Fritsche, M., Wickert, J. and Schuh, H., 2015. Precise positioning with current multi-constellation global navigation satellite systems: GPS, GLONASS, Galileo and BeiDou. *Scientific Reports, 5*, pp. 8328.

Lima, E., Hopkins, T., Gurney, E., Shortall, O., Lovatt, F., Davies, P., Williamson, G. and Kaler, J., 2018. Drivers for precision livestock technology adoption: A study of factors associated with adoption of electronic identification technology by commercial sheep farmers in England and Wales. *PloS One, 13* (1), p.e0190489.

Ling, C.X. and Sheng, V.S., 2011. Cost-Sensitive Learning and the Class Imbalance Problem. *Encyclopedia of Machine Learning: Springer*, pp. 231-235.

Longadge, R. and Dongre, S., 2013. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.

Lowe, S.A. and ÓLaighin, G., 2014. Monitoring human health behaviour in one's living environment: a technological review. *Medical Engineering & Physics, 36* (2), pp. 147-168.

Low, Z.N. and Law, C.L., 2015. Improving Time to First Fix for GPS Receivers. *School of Electrical & Electronic Engineering Nanyang Technological University*.

Lush, L., Wilson, R.P., Holton, M.D., Hopkins, P., Marsden, K.A., Chadwick, D.R. and King, A.J., 2018. Classification of sheep urination events using accelerometers to aid improved measurements of livestock contributions to nitrous oxide emissions. *Computers and Electronics in Agriculture, 150*, pp. 170-177.

MacEachern, S.N., Rao, Y. and Wu, C., 2007. A robust-likelihood cumulative sum chart. *Journal of the American Statistical Association*, *102* (480), pp. 1440-1447.

Madon, B. and Hingrat, Y., 2014. Deciphering behavioral changes in animal movement with a 'multiple change point algorithm-classification tree' framework. *Frontiers in Ecology and Evolution*, *2*, pp. 30.

Mannini, A. and Sabatini, A.M., 2011. Accelerometry-based classification of human activities using Markov modeling. *Computational Intelligence and Neuroscience*, *2011*, pp. 4.

Mansbridge, N., Mitsch, J., Bollard, N., Ellis, K., Miguel-Pacheco, G., Dottorini, T. and Kaler, J., 2018. Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep. *Sensors*, *18* (10), pp. 3532.

Marchesini, G., Mottaran, D., Contiero, B., Schiavon, E., Segato, S., Garbin, E., Tenti, S. and Andrighetto, I., 2018. Use of rumination and activity data as health status and performance indicators in beef cattle during the early fattening period. *The Veterinary Journal*, *231*, pp. 41-47.

Marcus Rowcliffe, J., Carbone, C., Kays, R., Kranstauber, B. and Jansen, P.A., 2012. Bias in estimating animal travel distance: the effect of sampling frequency. *Methods in Ecology and Evolution*, *3*(4), pp. 653-662.

Martınez-Munoz, G. and Suárez, A., 2004. Aggregation ordering in bagging. In *Proc. of the IASTED International Conference on Artificial Intelligence and Applications*, pp. 258-263. Citeseer.

Martiskainen, P., Järvinen, M., Skön, J.P., Tiirikainen, J., Kolehmainen, M. and Mononen, J., 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science*, *119* (1-2), pp. 32-38.

Maselyne, J., Saeys, W. and Van Nuffel, A., 2015. Quantifying animal feeding behaviour with a focus on pigs. *Physiology & Behavior*, *138*, pp. 37-51.

Mathie, M.J., Coster, A.C., Lovell, N.H. and Celler, B.G., 2004. Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. *Physiological Measurement*, *25* (2), pp. R1.

McClune, D.W., Marks, N.J., Delahay, R.J., Montgomery, W.I. and Scantlebury, D.M., 2015. Behaviour-time budget and functional habitat use of a free-ranging European badger (*Meles meles*). *Animal Biotelemetry*, *3* (1), pp. 7.

McEvoy, J.F., Roshier, D.A., Ribot, R.F. and Bennett, A.T., 2015. Proximate cues to phases of movement in a highly dispersive waterfowl, *Anas superciliosa*. *Movement Ecology*, *3* (1), pp. 21.

Meckley, T.D., Holbrook, C.M., Wagner, C.M. and Binder, T.R., 2014. An approach for filtering hyperbolically positioned underwater acoustic telemetry data with position precision estimates. *Animal Biotelemetry*, *2* (1), pp. 7.

Meen, G.H., Schellekens, M.A., Slegers, M.H.M., Leenders, N.L.G., Van Erp-Van der Kooij, E. and Noldus, L.P., 2015. Sound analysis in dairy cattle vocalisation as a potential welfare monitor. *Computers and Electronics in Agriculture*, *118*, pp. 111-115.

Menzies, D., Patison, K.P., Corbet, N.J. and Swain, D.L., 2018. Using Walk-over-Weighing technology for parturition date determination in beef cattle. *Animal Production Science*, *58* (9), pp. 1743-1750.

Mert, A., Kılıç, N., Bilgili, E. and Akan, A., 2015. Breast cancer detection with reduced feature set. *Computational and Mathematical Methods in Medicine*.

Meunier, B., Pradel, P., Sloth, K.H., Cirié, C., Delval, E., Mialon, M.M. and Veissier, I., 2018. Image analysis to refine measurements of dairy cow behaviour from a real-time location system. *Biosystems Engineering*, *173*, pp. 32-44.

Michelot, T., Langrock, R. and Patterson, T.A., 2016. moveHMM: an R package for the statistical modelling of animal movement data using hidden Markov models. *Methods in Ecology and Evolution, 7* (11), pp. 1308-1315.

Miguel, E., Grosbois, V., Fritz, H., Caron, A., de Garine-Wichatitsky, M., Nicod, F., Loveridge, A.J., Stapelkamp, B., Macdonald, D.W. and Valeix, M., 2017. Drivers of foot-and-mouth disease in cattle at wild/domestic interface: Insights from farmers, buffalo and lions. *Diversity and Distributions, 23* (9), pp. 1018-1030.

Milone, D.H., Rufiner, H.L., Galli, J.R., Laca, E.A. and Cangiano, C.A., 2009. Computational method for segmentation and classification of ingestive sounds in sheep. *Computers and Electronics in Agriculture, 65* (2), pp. 228-237.

Morgan-Davies, C., Lambe, N., Wishart, H., Waterhouse, T., Kenyon, F., McBean, D. and McCracken, D., 2018. Impacts of using a precision livestock system targeted approach in mountain sheep flocks. *Livestock Science, 208,* pp. 67-76.

Mottram, T., 2016. Animal board invited review: precision livestock farming for dairy cows with a focus on oestrus detection. *Animal, 10* (10), pp. 1575-1584.

Nakas, C.T., Alonzo, T.A. and Yiannoutsos, C.T., 2010. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Statistics in Medicine, 29* (28), pp. 2946-2955.

Nams, V.O., 2013. Sampling animal movement paths causes turn autocorrelation. *Acta Biotheoretica, 61* (2), pp. 269-284.

Nathan, R., Getz, W.M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D. and Smouse, P.E., 2008. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences, 105* (49), pp. 19052-19059.

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M. and Getz, W.M., 2012. Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *Journal of Experimental Biology, 215* (6), pp. 986-996.

Nayak, J., Naik, B. and Behera, H.S., 2015. Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In *Computational Intelligence in Data Mining, 2*, pp. 133-149. Springer, New Delhi.

Neethirajan, S., 2017. Recent advances in wearable sensors for animal health management. *Sensing and Bio-Sensing Research, 12*, pp. 15-29.

Nielsen, P.P., 2013. Automatic registration of grazing behaviour in dairy cows using 3D activity loggers. *Applied Animal Behaviour Science, 148* (3-4), pp. 179-184.

Ortiz-Pelaez, Á. and Pfeiffer, D.U., 2008. Use of data mining techniques to investigate disease risk classification as a proxy for compromised biosecurity of cattle herds in Wales. *BMC Veterinary Research, 4* (1), pp. 24.

Owen-Smith, N., Goodall, V. and Fatti, P., 2012. Applying mixture models to derive activity states of large herbivores from movement rates obtained using GPS telemetry. *Wildlife Research, 39* (5), pp. 452-462.

Page, E.S., 1954. Continuous inspection schemes. *Biometrika, 41* (1/2), pp. 100-115.

Pastell, M. and Madsen, H., 2008. Application of CUSUM charts to detect lameness in a milking robot. *Expert Systems with Applications, 35* (4), pp. 2032-2040.

Pastell, M., Hietaoja, J., Yun, J., Tiusanen, J. and Valros, A., 2016. Predicting farrowing of sows housed in crates and pens using accelerometers and CUSUM charts. *Computers and Electronics in Agriculture, 127*, pp. 197-203.

Patterson, T.A., Thomas, L., Wilcox, C., Ovaskainen, O. and Matthiopoulos, J., 2008. State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23* (2), pp. 87-94.

Patterson, T.A., Parton, A., Langrock, R., Blackwell, P.G., Thomas, L. and King, R., 2017. Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. *AStA Advances in Statistical Analysis*, *101* (4), pp. 399-438.

Pedersen, M.W., Patterson, T.A., Thygesen, U.H. and Madsen, H., 2011. Estimating animal behavior and residency from movement data. *Oikos*, *120* (9), pp. 1281-1290.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S. and Turini, F., 2018. Open the black box data-driven explanation of black box decision systems. *arXiv preprint arXiv:1806.09936*.

Penning, P.D., 1983. A technique to record automatically some aspects of grazing and ruminating behaviour in sheep. *Grass and Forage Science*, *38* (2), pp. 89-96.

Pépin, D., Adrados, C., Mann, C. and Janeau, G., 2004. Assessing real daily distance traveled by ungulates using differential GPS locations. *Journal of Mammalogy*, *85* (4), pp. 774-780.

Pereira, G.M., Heins, B.J. and Endres, M.I., 2018. Validation of an ear-tag accelerometer sensor to determine rumination, eating, and activity behaviors of grazing dairy cattle. *Journal of Dairy Science*, *101* (3), pp. 2492-2495.

Perotto-Baldivieso, H.L., Cooper, S.M., Cibils, A.F., Figueroa-Pagán, M., Udaeta, K. and Black-Rubio, C.M., 2012. Detecting autocorrelation problems from GPS collar data in livestock studies. *Applied Animal Behaviour Science*, *136* (2-4), pp. 117-125.

Pietersma, D., Lacroix, R. and Wade, K.M., 1998. A framework for the development of computerized management and control systems for use in dairy farming. *Journal of Dairy Science*, *81* (11), pp. 2962-2972.

Pinaud, D., 2008. Quantifying search effort of moving animals at several spatial scales using first-passage time analysis: effect of the structure of environment and tracking systems. *Journal of Applied Ecology*, *45* (1), pp. 91-99.

Pires, I.M., Garcia, N.M., Pombo, N., Flórez-Revuelta, F. and Spinsante, S., 2017. Pattern recognition techniques for the identification of Activities of Daily Living using mobile device accelerometer. *arXiv preprint arXiv:1711.00096*.

Plank, M.J. and Codling, E.A., 2009. Sampling rate and misidentification of Lévy and non-Lévy movement paths. *Ecology*, *90* (12), pp. 3546-3553.

Pohle, J., Langrock, R., van Beest, F.M. and Schmidt, N.M., 2017. Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22* (3), pp. 270-293.

Polley, E., LeDell, E., Kennedy, C., Lendle, S. and van der Laan, M., 2018. Package 'SuperLearner'.

Polojärvi, K., Colpaert, A., Matengu, K. and Kumpula, J., 2011. GPS collars in studies of cattle movement: cases of northeast Namibia and north Finland. In *Engineering Earth*, pp. 173-187. Springer, Dordrecht.

Powell, J.M., MacLeod, M., Vellinga, T.V., Opio, C., Falcucci, A., Tempio, G., Steinfeld, H. and Gerber, P., 2013. Feed–milk–manure nitrogen relationships in global dairy production systems. *Livestock Science*, *152* (2-3), pp. 261-272.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Prochazka, B.G., Coates, P.S., Ricca, M.A., Casazza, M.L., Gustafson, K.B. and Hull, J.M., 2017. Encounters with pinyon-juniper influence riskier movements in Greater Sage-Grouse across the Great Basin. *Rangeland Ecology & Management*, *70* (1), pp. 39-49.

Provost, F., 2000. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, *68*, pp. 1-3. AAAI Press.

Provost, F. and Domingos, P., 2000. Well-trained PETs: Improving probability estimation trees. *Raport instytutowy IS-00-04, Stern School of Business, New York University*.

Qin, Z., Zhang, C., Wang, T. and Zhang, S., 2010. Cost sensitive classification in data mining. In *International Conference on Advanced Data Mining and Applications*, pp. 1-11. Springer, Berlin, Heidelberg.

Quinlan, J.R., 1996. Bagging, boosting, and C4. 5. In *AAAI/IAAI, 1*, pp. 725-730.

Quinlan, J.R., 2014. *C4. 5: programs for machine learning*. Elsevier.

R Core Team. 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77* (2), pp. 257-286.

Rahman, A., Smith, D.V., Little, B., Ingham, A.B., Greenwood, P.L. and Bishop-Hurley, G.J., 2018. Cattle behaviour classification from collar, halter, and ear tag sensors. *Information Processing in Agriculture*, *5* (1), pp. 124-133.

Ramasco, V., Biuw, M. and Nilssen, K.T., 2014. Improving time budget estimates through the behavioural interpretation of dive bouts in harbour seals. *Animal Behaviour*, *94*, pp. 117-134.

Rayas-Amor, A.A., Morales-Almaráz, E., Licona-Velázquez, G., Vieyra-Alberto, R., García-Martínez, A., Martínez-García, C.G., Cruz-Monterrosa, R.G. and Miranda-de la Lama, G.C., 2017. Triaxial accelerometers for recording grazing and ruminating time in dairy cows: An alternative to visual observations. *Journal of Veterinary Behavior*, *20*, pp. 102-108.

Reeves, J., Chen, J., Wang, X.L., Lund, R. and Lu, Q.Q., 2007. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, *46* (6), pp. 900-915.

Resheff, Y.S., Rotics, S., Harel, R., Spiegel, O. and Nathan, R., 2014. AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements. *Movement Ecology*, *2* (1), pp. 27.

Robert, B., White, B.J., Renter, D.G. and Larson, R.L., 2009. Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle. *Computers and Electronics in Agriculture*, *67* (1-2), pp. 80-84.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C. and Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12* (1), pp. 77.

Rodgers, A.R., 1996. A GPS-based telemetry system. *Wildlife Society Bulletin*, *24*, pp. 559-566.

Rodríguez Alvarez, J., Arroqui, M., Mangudo, P., Toloza, J., Jatip, D., Rodriguez, J.M., Teyseyre, A., Sanz, C., Zunino, A., Machado, C. and Mateos, C., 2019. Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. *Agronomy*, *9* (2), pp. 90.

Rodriguez-Jimenez, S., Haerr, K.J., Trevisi, E., Loor, J.J., Cardoso, F.C. and Osorio, J.S., 2018. Prepartal standing behavior as a parameter for early detection of postpartal subclinical ketosis associated with inflammation and liver function biomarkers in peripartal dairy cows. *Journal of Dairy Science, 101* (9), pp. 8224-8235.

Rojas-Downing, M.M., Nejadhashemi, A.P., Harrigan, T. and Woznicki, S.A., 2017. Climate change and livestock: Impacts, adaptation, and mitigation. *Climate Risk Management, 16*, pp. 145-163.

Rombach, M., Münger, A., Niederhauser, J., Südekum, K.H. and Schori, F., 2018. Evaluation and validation of an automatic jaw movement recorder (RumiWatch) for ingestive and rumination behaviors of dairy cows during grazing and supplementation. *Journal of Dairy Science, 101* (3), pp. 2463-2475.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65* (6), pp. 386.

Ross, G.J., 2015. Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software, 66* (3), pp. 1-20.

Russell, R.A. and Bewley, J.M., 2013. Characterization of Kentucky dairy producer decision-making behavior. *Journal of Dairy Science, 96* (7), pp. 4751-4758.

Rutten, C.J., Velthuis, A.G.J., Steeneveld, W. and Hogeveen, H., 2013. Invited review: Sensors to support health management on dairy farms. *Journal of Dairy Science, 96* (4), pp. 1928-1952.

Rutten, C.J., Kamphuis, C., Hogeveen, H., Huijps, K., Nielen, M. and Steeneveld, W., 2017. Sensor data on cow activity, rumination, and ear temperature improve prediction of the start of calving in dairy cows. *Computers and Electronics in Agriculture, 132*, pp. 108-118.

Rutter, S.M., Champion, R.A. and Penning, P.D., 1997. An automatic system to record foraging behaviour in free-ranging ruminants. *Applied Animal Behaviour Science*, *54* (2-3), pp. 185-195.

Rutter, S.M., 2000. Graze: a program to analyze recordings of the jaw movements of ruminants. *Behavior Research Methods, Instruments, & Computers*, *32* (1), pp. 86-92.

Rybach, D., Gollan, C., Schluter, R. and Ney, H., 2009. Audio segmentation for speech recognition using segment features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4197-4200. IEEE.

Sadiq, M.B., Ramanoon, S.Z., Shaik Mossadeq, W.M., Mansor, R. and Syed-Hussain, S.S., 2017. Association between lameness and indicators of dairy cow welfare based on locomotion scoring, body and hock condition, leg hygiene and lying behavior. *Animals*, *7* (11), pp. 79.

Saint-Dizier, M. and Chastant-Maillard, S., 2012. Towards an automated detection of oestrus in dairy cattle. *Reproduction in Domestic Animals*, *47* (6), pp. 1056-1061.

Saint-Dizier, M. and Chastant-Maillard, S., 2015. Methods and on-farm devices to predict calving time in cattle. *The Veterinary Journal*, *205* (3), pp. 349-356.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G., 1986. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*.

Sakamoto, K.Q., Sato, K., Ishizuka, M., Watanuki, Y., Takahashi, A., Daunt, F. and Wanless, S., 2009. Can ethograms be automatically generated using body acceleration data from free-ranging birds? *PloS One*, *4* (4), pp. e5379.

Schapire, R.E., 1999. A brief introduction to boosting. In *Ijcai*, *99*, pp. 1401-1406.

Schapire, R.E., 2003. The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification*, pp. 149-171. Springer, New York, NY.

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C.D. and Troelsen, J., 2014. Dynamic accuracy of GPS receivers for use in health research: a novel method to assess GPS accuracy in real-world settings. *Frontiers in Public Health*, *2*, pp. 21.

Schirmann, K., Weary, D.M., Heuwieser, W., Chapinal, N., Cerri, R.L.A. and Von Keyserlingk, M.A.G., 2016. Rumination and feeding behaviors differ between healthy and sick dairy cows during the transition period. *Journal of Dairy Science*, *99* (12), pp. 9917-9924.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural Networks*, *61*, pp. 85-117.

Scholten, M.T., De Boer, I.J.M., Gremmen, B. and Lokhorst, C., 2013. Livestock farming with care: towards sustainable production of animal-source food. *NJAS – Wageningen Journal of Life Sciences*, *66*, pp. 3-5.

Sebastiani, P., Abad, M. and Ramoni, M.F., 2005. Bayesian networks for genomic analysis. *Genomic Signal Processing and Statistics*, *2*, pp. 281-320.

Sepúlveda-Varas, P., Proudfoot, K.L., Weary, D.M. and von Keyserlingk, M.A., 2016. Changes in behaviour of dairy cows with clinical mastitis. *Applied Animal Behaviour Science*, *175*, pp. 8-13.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P. and Weigel, K., 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *Journal of Dairy Science*, *97* (2), pp. 731-742.

Shahriar, M.S., Smith, D., Rahman, A., Freeman, M., Hills, J., Rawnsley, R., Henry, D. and Bishop-Hurley, G., 2016. Detecting heat events in dairy cows using accelerometers and unsupervised learning. *Computers and Electronics in Agriculture, 128*, pp. 20-26.

Sharma, S. and Kaushik, A., 2019. Views of Irish Farmers on Smart Farming Technologies: An Observational Study. *AgriEngineering*, *1* (2), pp. 164-187.

Smith, A.F.M., 1975. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika, 62* (2), pp. 407-416.

Smith, D., Rahman, A., Bishop-Hurley, G.J., Hills, J., Shahriar, S., Henry, D. and Rawnsley, R., 2016. Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Computers and Electronics in Agriculture*, *131*, pp. 40-50.

Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pp. 1015-1021. Springer, Berlin, Heidelberg.

Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45* (4), pp. 427-437.

Soriani, N., Panella, G. and Calamari, L.U.I.G.I., 2013. Rumination time during the summer season and its relationships with metabolic conditions and milk production. *Journal of Dairy Science*, *96* (8), pp. 5082-5094.

Sorzano, C.O.S., Vargas, J. and Montano, A.P., 2014. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

Spoliansky, R., Edan, Y., Parmet, Y. and Halachmi, I., 2016. Development of automatic body condition scoring using a low-cost 3-dimensional Kinect camera. *Journal of Dairy Science, 99* (9), pp. 7714-7725.

Stangaferro, M.L., Wijma, R., Caixeta, L.S., Al-Abri, M.A. and Giordano, J.O., 2016. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part I. Metabolic and digestive disorders. *Journal of Dairy Science*, *99* (9), pp. 7395-7410.

Stąpor, K., 2017, May. Evaluating and comparing classifiers: review, some recommendations and limitations. In *International Conference on Computer Recognition Systems*, pp. 12-21. Springer, Cham.

Steeneveld, W. and Hogeveen, H., 2015. Characterization of Dutch dairy farms using sensor systems for cow management. *Journal of Dairy Science, 98* (1), pp. 709-717.

Steeneveld, W., Vernooij, J.C.M. and Hogeveen, H., 2015[a]. Effect of sensor systems for cow management on milk production, somatic cell count, and reproduction. *Journal of Dairy Science*, *98* (6), pp. 3896-3905.

Steeneveld, W., Hogeveen, H. and Lansink, A.O., 2015[b]. Economic consequences of investing in sensor systems on dairy farms. *Computers and Electronics in Agriculture, 119*, pp. 33-39.

Steensels, M., Antler, A., Bahr, C., Berckmans, D., Maltz, E. and Halachmi, I., 2016. A decision-tree model to detect post-calving diseases based on rumination, activity, milk yield, BW and voluntary visits to the milking robot. *Animal, 10* (9), pp. 1493-1500.

Steensels, M., Maltz, E., Bahr, C., Berckmans, D., Antler, A. and Halachmi, I., 2017. Towards practical application of sensors for monitoring animal health; design and validation of a model to detect ketosis. *Journal of Dairy Research, 84* (2), pp. 139-145.

Steenweg, R., Hebblewhite, M., Whittington, J., Lukacs, P. and McKelvey, K., 2018. Sampling scales define occupancy and underlying occupancy–abundance relationships in animals. *Ecology*, *99* (1), pp. 172-183.

Stephens, D.A., 1994. Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43* (1), pp. 159-178.

Stewart, M., Wilson, M.T., Schaefer, A.L., Huddart, F. and Sutherland, M.A., 2017. The use of infrared thermography and accelerometers for remote monitoring of dairy cow health and welfare. *Journal of Dairy Science*, *100* (5), pp. 3893-3901.

Studd, E.K., Boudreau, M., Majchrzak, Y., Menzies, A., Peers, M., Seguin, J., Lavergne, S., Boonstra, R., Murray, D., Boutin, S. and Humphries, M., 2019. Use of acceleration and acoustics to classify behavior, generate time budgets, and evaluate responses to moonlight in free-ranging snowshoe hares. *Frontiers in Ecology and Evolution*, *7*, pp. 154.

Sur, M., Suffredini, T., Wessells, S.M., Bloom, P.H., Lanzone, M., Blackshire, S., Sridhar, S. and Katzner, T., 2017. Improved supervised classification of accelerometry data to distinguish behaviors of soaring birds. *PloS One*, *12* (4), p.e0174785.

Suryan, R.M., Sato, F., Balogh, G.R., Hyrenbach, K.D., Sievert, P.R. and Ozaki, K., 2006. Foraging destinations and marine habitat use of short-tailed albatrosses: a multi-scale approach using first-passage time analysis. *Deep Sea Research Part II: Topical Studies in Oceanography*, *53* (3-4), pp. 370-386.

Sutton, R.S. and Barto, A.G., 1998. Introduction to reinforcement learning. Vol. 135.

Swain, D.L., Wark, T. and Bishop-Hurley, G.J., 2008. Using high fix rate GPS data to determine the relationships between fix rate, prediction errors and patch selection. *Ecological Modelling*, *212* (3-4), pp. 273-279.

Tharwat, A., 2018. Classification assessment methods. *Applied Computing and Informatics*.

Thompson, R., Matheson, S.M., Plötz, T., Edwards, S.A. and Kyriazakis, I., 2016. Porcine lie detectors: Automatic quantification of posture state and transitions in sows using inertial sensors. *Computers and Electronics in Agriculture*, *127*, pp. 521-530.

Thomson Reuters, 2019. Web of Science. [Online] Available at: http://www.webofknowledge.com/ [Accessed 2019].

Thorup, V.M., Munksgaard, L., Robert, P.E., Erhard, H.W., Thomsen, P.T. and Friggens, N.C., 2015. Lameness detection via leg-mounted accelerometers on dairy cows on four commercial farms. *Animal*, *9* (10), pp. 1704-1712.

Thorup, V.M., Nielsen, B.L., Robert, P.E., Giger-Reverdin, S., Konka, J., Michie, C. and Friggens, N.C., 2016. Lameness affects cow feeding but not rumination behavior as characterized from sensor data. *Frontiers in Veterinary Science*, *3*, pp. 37.

Ting, K.M. and Witten, I.H., 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, *10*, pp. 271-289.

Tomkiewicz, S.M., Fuller, M.R., Kie, J.G. and Bates, K.K., 2010. Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365* (1550), pp. 2163-2176.

Tsairidou, S., Allen, A., Banos, G., Coffey, M., Anacleto, O., Byrne, A.W., Skuce, R.A., Glass, E.J., Woolliams, J.A. and Doeschl-Wilson, A.B., 2018. Can We Breed Cattle for Lower Bovine TB Infectivity? *Frontiers in Veterinary Science*, *5*, pp. 310.

Tsoumakas, G., Partalas, I. and Vlahavas, I., 2008. A taxonomy and short review of ensemble selection. In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications*, pp. 1-6.

Tullo, E., Finzi, A. and Guarino, M., 2018. Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy. *Science of the Total Environment*, *650*, pp. 2751-2760.

Turner, L.W., Udal, M.C., Larson, B.T. and Shearer, S.A., 2000. Monitoring cattle behavior and pasture use with GPS and GIS. *Canadian Journal of Animal Science*, *80* (3), pp. 405-413.

Udomprasert, P. and Williamson, N.B., 1990. The DairyCHAMP program: a computerised recording system for dairy herds. *The Veterinary Record*, *127* (10), pp. 256-262.

Ungar, E.D. and Rutter, S.M., 2006. Classifying cattle jaw movements: comparing IGER behaviour recorder and acoustic techniques. *Applied Animal Behaviour Science*, *98* (1-2), pp. 11-27.

Ungar, E.D., Nevo, Y., Baram, H. and Arieli, A., 2018. Evaluation of the IceTag leg sensor and its derivative models to predict behaviour, using beef cattle on rangeland. *Journal of Neuroscience Methods*, *300*, pp. 127-137.

Valletta, J.J., Torney, C., Kings, M., Thornton, A. and Madden, J., 2017. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, *124*, pp. 203-220.

Van De Gucht, T., Saeys, W., Van Nuffel, A., Pluym, L., Piccart, K., Lauwers, L., Vangeyte, J. and Van Weyenberg, S., 2017. Farmers' preferences for automatic lameness-detection systems in dairy cattle. *Journal of Dairy Science*, *100* (7), pp. 5746-5757.

Van De Gucht, T., Saeys, W., Van Meensel, J., Van Nuffel, A., Vangeyte, J. and Lauwers, L., 2018. Farm-specific economic value of automatic lameness detection systems in dairy cattle: From concepts to operational simulations. *Journal of Dairy Science*, *101* (1), pp. 637-648.

Van der Laan, M.J., Polley, E.C. and Hubbard, A.E., 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6* (1).

Van Diggelen, F., 1998. Innovation: GPS accuracy-lies, damn lies, and statistics. *GPS WORLD*, *9*, pp. 41-45.

Van Erp-Van der Kooij, E., Van de Brug, M. and Roelofs, J. B., 2016. Validation of Nedap Smarttag Leg and Neck to assess behavioural activity level in dairy cattle. In *Precision Dairy Farming Conference*, pp. 321-326.

Van Hertem, T., Rooijakkers, L., Berckmans, D., Fernández, A.P., Norton, T. and Vranken, E., 2017. Appropriate data visualisation is key to Precision Livestock Farming acceptance. *Computers and Electronics in Agriculture*, *138*, pp. 1-10.

Van Moorter, B., Visscher, D.R., Jerde, C.L., Frair, J.L. and Merrill, E.H., 2010. Identifying movement states from location data using cluster analysis. *The Journal of Wildlife Management*, *74* (3), pp. 588-594.

Van Nuffel, A., Zwertvaegher, I., Van Weyenberg, S., Pastell, M., Thorup, V., Bahr, C., Sonck, B. and Saeys, W., 2015. Lameness detection in dairy cows: Part 2. Use of sensors to automatically register changes in locomotion or behavior. *Animals*, *5* (3), pp. 861-885.

Vanrell, S.R., Chelotti, J.O., Galli, J.R., Utsumi, S.A., Giovanini, L.L., Rufiner, H.L. and Milone, D.H., 2018. A regularity-based algorithm for identifying grazing and rumination bouts from acoustic signals in grazing cattle. *Computers and Electronics in Agriculture*, *151*, pp. 392-402.

Vapnik, V.N., 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10* (5), pp. 988-999.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C.E.B., Pluk, A., Halachmi, I., Lokhorst, C. and Berckmans, D., 2013. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *Journal of Dairy Science*, *96* (1), pp. 257-266.

Visser, I. and Speekenbrink, M., 2010. depmixS4: an R package for hidden Markov models. *Journal of Statistical Software*, *36* (7), pp. 1-21.

Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Havlin, S., Da Luz, M.G.E., Raposo, E.P. and Stanley, H.E., 2000. Lévy flights in random searches. *Physica A: Statistical Mechanics and its Applications*, *282* (1-2), pp. 1-12.

Voss, B., Laue, H.J., Hoedemaker, M. and Wiedemann, S., 2016. Field-trial evaluation of an automatic temperature measurement device placed in the reticulo-rumen of pre-weaned male calves. *Livestock Science*, *189*, pp. 78-81.

Walker, S.L., Smith, R.F., Routly, J.E., Jones, D.N., Morris, M.J. and Dobson, H., 2008. Lameness, activity time-budgets, and estrus expression in dairy cattle. *Journal of Dairy Science*, *91* (12), pp. 4552-4559.

Walton, E., Casey, C., Mitsch, J., Vázquez-Diosdado, J.A., Yan, J., Dottorini, T., Ellis, K.A., Winterlich, A. and Kaler, J., 2018. Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. *Royal Society Open Science*, *5* (2), pp. 171442.

Wang, L. ed., 2005. *Support vector machines: theory and applications*, 177. Springer Science & Business Media.

Ward, J.A., Lukowicz, P., Troster, G. and Starner, T.E., 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28* (10), pp. 1553-1567.

Watanabe, S., Izawa, M., Kato, A., Ropert-Coudert, Y. and Naito, Y., 2005. A new technique for monitoring the detailed behaviour of terrestrial animals: a case study with the domestic cat. *Applied Animal Behaviour Science*, *94* (1-2), pp. 117-131.

Watanabe, N., Sakanoue, S., Kawamura, K. and Kozakai, T., 2008. Development of an automatic classification system for eating, ruminating and resting behavior of cattle using an accelerometer. *Grassland Science*, *54* (4), pp. 231-237.

Wathes, C.M., Kristensen, H.H., Aerts, J.M. and Berckmans, D., 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? *Computers and Electronics in Agriculture*, *64* (1), pp. 2-10.

Weigel, K.A., VanRaden, P.M., Norman, H.D. and Grosu, H., 2017. A 100-Year Review: methods and impact of genetic selection in dairy cattle—from daughter–dam comparisons to deep learning algorithms. *Journal of Dairy Science*, *100* (12), pp. 10234-10250.

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS One*, *12* (4), p.e0174944.

Werkheiser, I., 2018. Precision livestock farming and farmers' duties to livestock. *Journal of Agricultural and Environmental Ethics*, *31* (2), pp. 181-195.

Westin, R., Vaughan, A., de Passillé, A.M., Devries, T.J., Pajor, E.A., Pellerin, D., Siegford, J.M., Vasseur, E. and Rushen, J., 2016. Lying times of lactating cows on dairy farms with automatic milking systems and the relation to lameness, leg lesions, and body condition score. *Journal of Dairy Science*, *99* (1), pp. 551-561.

Whalen, S. and Pandey, G., 2013. A comparative analysis of ensemble classifiers: case studies in genomics. In *2013 IEEE 13th International Conference on Data Mining*, pp. 807-816. IEEE.

Whitham, J.C. and Miller, L.J., 2016. Using technology to monitor and improve zoo animal welfare. *Animal Welfare*, *25* (4), pp. 395-409.

Wijers, M., Trethowan, P., Markham, A., Du Preez, B., Chamaillé-Jammes, S., Loveridge, A. and Macdonald, D., 2018. Listening to lions: animal-borne acoustic sensors improve bio-logger calibration and behaviour classification performance. *Frontiers in Ecology and Evolution*, *6*, pp. 171.

Williams, H.J., Shepard, E.L.C., Duriez, O. and Lambertucci, S.A., 2015. Can accelerometry be used to distinguish between flight types in soaring birds? *Animal Biotelemetry*, *3* (1), pp. 45.

Williams, H.J., Holton, M.D., Shepard, E.L., Largey, N., Norman, B., Ryan, P.G., Duriez, O., Scantlebury, M., Quintana, F., Magowan, E.A. and Marks, N.J., 2017. Identification of animal movement patterns using tri-axial magnetometry. *Movement Ecology*, *5* (1), pp. 6.

Williams, M.L., Mac Parthaláin, N., Brewer, P., James, W.P.J. and Rose, M.T., 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *Journal of Dairy Science*, *99* (3), pp. 2063-2075.

Williams, M.L., James, W.P. and Rose, M.T., 2019. Variable segmentation and ensemble classifiers for predicting dairy cow behaviour. *Biosystems Engineering*, *178*, pp. 156-167.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Woodroffe, R., Donnelly, C.A., Ham, C., Jackson, S.Y., Moyes, K., Chapman, K., Stratton, N.G. and Cartwright, S.J., 2016. Badgers prefer cattle pasture but avoid cattle: implications for bovine tuberculosis control. *Ecology Letters*, *19* (10), pp. 1201-1208.

Wolf, R., Barkema, H.W., De Buck, J., Slomp, M., Flaig, J., Haupstein, D., Pickel, C. and Orsel, K., 2014. High herd-level prevalence of Mycobacterium avium subspecies paratuberculosis in Western Canadian dairy farms, based on environmental sampling. *Journal of Dairy Science*, *97* (10), pp. 6250-6259.

Wolfert, S., Ge, L., Verdouw, C. and Bogaardt, M.J., 2017. Big data in smart farming–a review. *Agricultural Systems*, *153*, pp. 69-80.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5* (2), pp. 241-259.

Wu, X., Zhu, X., Wu, G. Q. and Ding, W., 2014. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26* (1), pp. 97-107.

Yang, F., Wang, H.Z., Mi, H. and Cai, W.W., 2009. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, *10* (1), pp. S22.

Yang, A., Huang, H., Zhu, X., Yang, X., Chen, P., Li, S. and Xue, Y., 2018. Automatic recognition of sow nursing behaviour using deep learning-based segmentation and spatial and temporal features. *Biosystems Engineering*, *175*, pp. 133-145.

Yoda, K., Naito, Y., Sato, K., Takahashi, A., Nishikawa, J., Ropert-Coudert, Y., Kurita, M. and Le Maho, Y., 2001. A new technique for monitoring the behaviour of free-ranging Adelie penguins. *Journal of Experimental Biology*, *204* (4), pp. 685-690.

Youden, W.J., 1950. Index for rating diagnostic tests. *Cancer*, *3* (1), pp. 32-35.

Yule, G.U., 1900. VII. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *194*, (252-261), pp. 257-319.

Zehner, N., Niederhauser, J.J., Schick, M. and Umstatter, C., 2019. Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. *Computers and Electronics in Agriculture*, *161*, pp. 62-71.

Zhang, J., O'Reilly, K.M., Perry, G.L., Taylor, G.A. and Dennis, T.E., 2015. Extending the functionality of behavioural change-point analysis with k-means clustering: a case study with the little penguin (*eudyptula minor*). *PloS One*, *10* (4), p.e0122811.

Zhao, K., Jin, X., Ji, J., Wang, J., Ma, H. and Zhu, X., 2019. Individual identification of Holstein dairy cows based on detecting and matching feature points in body images. *Biosystems Engineering*, *181*, pp. 128-139.

Zhu, X. and Goldberg, A.B., 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3* (1), pp. 1-130.

Zuidhof, M.J., Fedorak, M.V., Ouellette, C.A. and Wenger, I.I., 2017. Precision feeding: Innovative management of broiler breeder feed intake and flock uniformity. *Poultry Science*, *96* (7), pp. 2254-2263.

Appendix 1

Open field treatment

Table 11. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 30 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.20	4.20	5.18	8.74	10.37	76.47
2	0.07	1.33	1.65	2.76	3.30	6.59
3	0.15	1.32	1.62	2.74	3.24	7.07
4	0.32	1.31	1.61	2.73	3.21	7.59
5	0.05	1.33	1.62	2.78	3.24	6.43
6	0.22	1.29	1.56	2.67	3.12	6.60
7	0.23	1.42	1.75	2.96	3.50	9.20
8	0.30	1.28	1.56	2.67	3.11	7.26
9	0.09	5.27	6.33	10.96	12.67	187.77
10	0.11	1.40	1.71	2.91	3.43	7.86
11	0.36	1.31	1.61	2.73	3.22	6.92
12	0.21	1.29	1.57	2.69	3.14	7.05
13	0.07	1.32	1.62	2.74	3.24	7.39
14	0.48	1.30	1.60	2.71	3.21	7.17
15	0.15	3.58	4.30	7.44	8.59	81.74
16	0.25	1.34	1.64	2.79	3.27	7.47
17	0.34	1.24	1.52	2.59	3.03	6.54
18	0.39	1.28	1.56	2.67	3.13	7.22
19	0.61	1.29	1.58	2.69	3.16	6.90
20	0.09	1.34	1.64	2.78	3.27	7.63
21	0.26	1.39	1.71	2.88	3.41	6.85
22	0.30	1.29	1.57	2.68	3.15	7.08
23	0.14	1.30	1.60	2.70	3.21	7.18
24	0.42	1.31	1.59	2.72	3.18	7.44
25	0.28	1.41	1.73	2.94	3.45	7.63
26	0.57	1.40	1.70	2.91	3.39	7.27
27	0.22	1.31	1.60	2.74	3.20	7.54
28	0.33	1.33	1.62	2.77	3.25	7.65
29	0.08	1.26	1.53	2.63	3.06	7.88
30	0.35	1.26	1.56	2.63	3.11	7.48
31	0.38	1.59	1.96	3.32	3.92	11.59
32	0.14	1.30	1.59	2.71	3.18	8.36
33	0.09	1.28	1.55	2.67	3.09	7.24
34	0.09	1.26	1.53	2.62	3.06	6.77
35	0.18	1.48	1.83	3.07	3.66	9.27
36	0.14	1.31	1.60	2.73	3.21	7.02
Mean	0.24	1.58	1.93	3.29	3.86	16.48
SE	0.02	0.15	0.18	0.30	0.35	5.63
SD	0.14	0.87	1.05	1.82	2.11	33.76

Table 12. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 10 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.04	2.02	2.43	4.21	4.86	28.66
2	0.08	1.30	1.59	2.70	3.18	7.03
3	0.08	1.30	1.60	2.71	3.21	7.46
4	0.13	1.38	1.70	2.87	3.41	7.86
5	0.32	1.33	1.62	2.76	3.25	6.79
6	0.04	1.21	1.48	2.52	2.95	6.92
7	0.12	1.35	1.66	2.80	3.32	7.60
8	0.22	1.34	1.64	2.79	3.28	8.71
9	0.09	9.68	12.08	20.14	24.15	438.20
10	0.12	1.26	1.54	2.62	3.09	7.74
11	0.07	1.30	1.60	2.71	3.20	7.34
12	0.25	1.30	1.60	2.71	3.21	7.08
13	0.25	1.31	1.61	2.73	3.21	7.34
14	0.19	1.23	1.52	2.57	3.03	7.41
15	0.08	1.15	1.41	2.38	2.82	6.38
16	0.06	1.33	1.62	2.77	3.25	8.24
17	0.17	1.35	1.65	2.81	3.31	7.06
18	0.10	1.49	1.82	3.10	3.64	8.63
19	0.13	1.30	1.59	2.70	3.17	7.80
20	0.09	1.34	1.64	2.79	3.28	7.67
21	0.23	1.35	1.66	2.82	3.32	7.21
22	0.15	1.24	1.52	2.58	3.04	7.65
23	0.05	1.26	1.55	2.62	3.10	7.54
24	0.03	1.30	1.58	2.71	3.17	7.62
25	0.32	1.33	1.62	2.77	3.24	7.95
26	0.20	1.34	1.63	2.79	3.26	7.66
27	0.12	1.36	1.66	2.82	3.32	7.57
28	0.07	1.36	1.67	2.84	3.34	7.93
29	0.04	1.44	1.78	3.00	3.57	8.48
30	0.13	1.22	1.50	2.53	2.99	6.62
31	0.05	6.54	9.32	13.60	18.64	317.70
32	0.12	1.29	1.58	2.68	3.16	7.35
33	0.20	2.53	3.13	5.27	6.25	31.42
34	0.09	1.26	1.54	2.62	3.08	7.31
35	0.06	6.13	10.95	12.75	21.89	176.56
36	0.12	0.70	1.25	1.46	2.50	7.24
Mean	0.13	1.86	2.43	3.87	4.85	34.05
SE	0.01	0.30	0.43	0.62	0.86	15.07
SD	0.08	1.79	2.59	3.73	5.19	90.39

Table 13. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 5 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.04	1.83	2.28	3.80	4.55	41.74
2	0.41	1.18	1.44	2.45	2.88	7.69
3	0.19	1.32	1.63	2.75	3.26	10.09
4	0.16	1.40	1.74	2.91	3.47	10.92
5	0.10	1.28	1.56	2.65	3.12	7.88
6	0.47	1.21	1.48	2.51	2.95	8.17
7	0.27	1.36	1.69	2.84	3.38	12.70
8	0.15	1.31	1.60	2.72	3.20	9.79
9	0.09	3.91	5.00	8.12	10.01	181.56
10	0.11	1.33	1.61	2.76	3.23	8.67
11	0.25	1.28	1.58	2.67	3.17	8.68
12	0.34	1.36	1.66	2.83	3.33	10.68
13	0.07	1.25	1.53	2.61	3.06	8.50
14	0.03	20.12	31.71	41.86	63.42	3976.69
15	0.06	2.92	3.69	6.07	7.38	138.57
16	0.15	1.39	1.71	2.89	3.42	10.71
17	0.23	1.26	1.54	2.63	3.08	9.64
18	0.10	1.48	1.80	3.09	3.60	13.11
19	0.21	1.33	1.62	2.76	3.24	10.76
20	0.10	1.39	1.71	2.88	3.41	10.66
21	0.08	1.33	1.63	2.77	3.27	8.48
22	0.08	1.33	1.64	2.77	3.29	10.73
23	0.14	1.29	1.59	2.69	3.19	8.68
24	0.19	1.27	1.56	2.64	3.11	8.26
25	0.12	1.33	1.62	2.77	3.24	8.35
26	0.02	1.47	1.80	3.05	3.61	10.82
27	0.42	1.24	1.52	2.59	3.04	8.63
28	0.16	1.41	1.72	2.93	3.45	10.90
29	0.04	1.41	1.74	2.94	3.47	15.19
30	0.34	1.23	1.51	2.57	3.01	7.39
31	0.17	1.80	2.24	3.75	4.47	31.20
32	0.25	1.40	1.72	2.90	3.44	11.03
33	0.03	1.82	2.21	3.78	4.42	20.48
34	0.10	1.27	1.55	2.64	3.11	10.61
35	0.06	1.23	1.51	2.57	3.03	12.88
36	0.68	1.36	1.68	2.83	3.35	10.36
Mean	0.18	2.00	2.66	4.17	5.32	130.31
SE	0.02	0.52	0.84	1.09	1.67	110.05
SD	0.14	3.15	5.02	6.55	10.05	660.33

Hedgerow treatment

Table 14. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 30 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.12	5.15	6.68	10.71	13.36	61.26
2	0.06	1.49	1.83	3.09	3.66	11.56
3	0.10	1.65	2.09	3.43	4.18	13.44
4	0.03	1.64	2.05	3.42	4.10	13.40
5	0.12	2.10	2.83	4.37	5.67	34.10
6	0.04	1.72	2.20	3.57	4.40	24.39
7	0.09	1.63	2.07	3.40	4.14	21.11
8	0.08	2.22	2.91	4.61	5.81	31.92
9	0.09	3.46	4.50	7.20	9.00	53.99
10	0.10	1.58	1.95	3.29	3.89	13.38
11	0.03	1.51	1.87	3.14	3.74	19.26
12	0.11	1.35	1.70	2.82	3.41	15.13
13	0.24	1.99	2.53	4.15	5.06	39.79
14	0.09	1.61	2.02	3.35	4.04	28.43
15	0.03	1.72	2.21	3.59	4.41	24.19
16	0.10	1.74	2.22	3.61	4.44	27.38
17	0.05	1.49	1.89	3.09	3.78	7.64
18	0.18	1.38	1.70	2.88	3.40	6.60
19	0.06	1.41	1.77	2.93	3.54	8.47
20	0.10	1.65	2.06	3.42	4.12	20.07
21	0.07	1.50	1.89	3.12	3.77	9.47
22	0.07	1.49	1.86	3.10	3.72	10.51
23	0.19	1.75	2.21	3.64	4.42	21.11
24	0.06	1.59	2.00	3.32	4.01	10.02
25	0.02	1.43	1.81	2.98	3.62	10.28
26	0.16	1.38	1.69	2.87	3.38	7.24
27	0.07	1.33	1.66	2.78	3.32	9.27
28	0.31	1.47	1.82	3.06	3.64	10.60
29	0.20	1.86	2.40	3.87	4.80	17.90
30	0.07	1.41	1.74	2.93	3.48	12.59
31	0.09	2.33	3.08	4.84	6.16	34.36
32	0.04	1.68	2.11	3.49	4.21	17.38
33	0.08	2.48	3.21	5.17	6.41	26.72
34	0.17	1.37	1.71	2.86	3.41	8.41
35	0.10	1.95	2.50	4.05	4.99	22.77
36	0.02	1.48	1.86	3.08	3.73	14.01
Mean	0.10	1.81	2.29	3.76	4.59	19.95
SE	0.01	0.12	0.16	0.24	0.31	2.13
SD	0.06	0.70	0.94	1.47	1.88	12.76

Table 15. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 10 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.12	4.02	5.26	8.36	10.52	96.34
2	0.06	1.58	1.96	3.29	3.91	18.87
3	0.13	1.54	1.94	3.21	3.89	12.89
4	0.03	1.64	2.06	3.42	4.12	12.86
5	0.12	1.50	1.89	3.12	3.77	24.67
6	0.04	1.57	1.95	3.26	3.90	17.13
7	0.09	1.80	2.29	3.74	4.58	21.48
8	0.08	1.75	2.18	3.64	4.36	8.37
9	0.09	4.48	5.72	9.32	11.44	162.08
10	0.04	1.66	2.04	3.45	4.07	14.78
11	0.03	1.48	1.81	3.09	3.62	18.71
12	0.11	1.37	1.73	2.85	3.45	14.14
13	0.03	1.94	2.45	4.03	4.90	46.11
14	0.09	1.69	2.08	3.51	4.16	15.27
15	0.03	1.29	1.60	2.69	3.21	8.56
16	0.10	1.58	1.94	3.28	3.88	11.46
17	0.05	1.71	2.14	3.55	4.27	16.27
18	0.09	1.37	1.72	2.84	3.44	8.09
19	0.06	1.68	2.08	3.50	4.16	10.68
20	0.10	1.61	1.96	3.34	3.93	10.80
21	0.04	1.66	2.05	3.45	4.10	8.34
22	0.07	1.56	1.97	3.25	3.95	14.99
23	0.17	1.60	1.98	3.33	3.97	9.80
24	0.06	1.58	1.98	3.28	3.97	9.13
25	0.30	1.49	1.83	3.09	3.65	7.65
26	0.02	1.39	1.73	2.89	3.46	8.14
27	0.06	1.43	1.77	2.97	3.55	9.12
28	0.27	1.40	1.73	2.91	3.45	11.48
29	0.11	1.63	2.01	3.39	4.02	20.12
30	0.07	1.57	1.91	3.28	3.83	10.57
31	0.12	480.40	595.15	999.22	1190.30	11062.80
32	0.04	1.75	2.22	3.64	4.45	19.33
33	0.08	3.56	4.40	7.40	8.79	90.99
34	0.17	1.36	1.68	2.82	3.36	8.19
35	0.10	1.89	2.34	3.92	4.68	25.10
36	0.36	1.44	1.81	2.99	3.62	15.13
Mean	0.10	15.08	18.70	31.37	37.41	330.01
SE	0.01	13.30	16.47	27.65	32.94	306.69
SD	0.08	79.77	98.82	165.92	197.65	1840.17

Table 16. Mean horizontal accuracies of each of the 36 GPS receivers sampled at 5 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
1	0.12	3.55	4.58	7.38	9.17	103.49
2	0.06	1.83	2.28	3.82	4.56	10.07
3	0.07	1.52	1.89	3.17	3.78	11.68
4	0.03	1.57	1.93	3.26	3.86	8.47
5	0.12	1.94	2.43	4.03	4.86	35.09
6	0.04	1.40	1.73	2.91	3.47	17.39
7	0.09	1.54	1.90	3.21	3.80	16.67
8	0.08	1.90	2.38	3.95	4.75	10.94
9	0.09	2.87	3.71	5.96	7.41	34.65
10	0.13	1.54	1.91	3.20	3.82	14.38
11	0.03	1.34	1.63	2.80	3.25	9.65
12	0.11	1.48	1.87	3.08	3.75	16.01
13	0.15	2.32	2.97	4.82	5.94	55.11
14	0.09	2.18	2.74	4.53	5.47	36.25
15	0.03	1.48	1.80	3.08	3.60	12.26
16	0.10	1.70	2.09	3.54	4.18	17.40
17	0.05	1.71	2.12	3.56	4.23	8.82
18	0.09	2.62	3.19	5.46	6.37	144.19
19	0.06	2.01	2.47	4.19	4.94	10.50
20	0.10	1.91	2.39	3.97	4.78	13.67
21	0.04	1.78	2.20	3.69	4.40	9.46
22	0.11	1.72	2.13	3.59	4.26	12.42
23	0.17	1.58	1.92	3.29	3.85	11.24
24	0.13	1.92	2.31	3.99	4.62	11.62
25	0.02	1.58	1.95	3.30	3.89	10.98
26	0.02	1.31	1.63	2.73	3.27	7.41
27	0.06	1.71	2.07	3.56	4.13	10.21
28	0.07	1.40	1.73	2.91	3.47	13.57
29	0.11	2.54	3.33	5.29	6.66	27.66
30	0.07	1.75	2.17	3.64	4.33	17.13
31	0.09	3.25	3.99	6.75	7.98	106.22
32	0.04	1.86	2.40	3.88	4.81	29.33
33	0.08	2.54	3.25	5.29	6.49	16.95
34	0.17	1.45	1.80	3.01	3.61	9.01
35	0.10	2.29	2.95	4.77	5.90	19.07
36	0.10	1.46	1.83	3.03	3.65	18.29
Mean	0.08	1.90	2.38	3.96	4.76	25.48
SE	0.01	0.09	0.12	0.19	0.23	5.08
SD	0.04	0.54	0.70	1.11	1.40	30.49

Appendix 2

Open field treatment

Table 17. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 30 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.07	1.33	1.65	2.76	3.30	6.59
3	0.15	1.32	1.62	2.74	3.24	7.07
4	0.32	1.31	1.61	2.73	3.21	7.59
5	0.05	1.33	1.62	2.78	3.24	6.43
6	0.22	1.29	1.56	2.67	3.12	6.60
8	0.30	1.28	1.56	2.67	3.11	7.26
10	0.11	1.40	1.71	2.91	3.43	7.86
11	0.36	1.31	1.61	2.73	3.22	6.92
12	0.21	1.29	1.57	2.69	3.14	7.05
13	0.07	1.32	1.62	2.74	3.24	7.39
16	0.25	1.34	1.64	2.79	3.27	7.47
17	0.34	1.24	1.52	2.59	3.03	6.54
19	0.61	1.29	1.58	2.69	3.16	6.90
20	0.09	1.34	1.64	2.78	3.27	7.63
21	0.26	1.39	1.71	2.88	3.41	6.85
22	0.30	1.29	1.57	2.68	3.15	7.08
23	0.14	1.30	1.60	2.70	3.21	7.18
24	0.42	1.31	1.59	2.72	3.18	7.44
26	0.57	1.40	1.70	2.91	3.39	7.27
27	0.22	1.31	1.60	2.74	3.20	7.54
28	0.33	1.33	1.62	2.77	3.25	7.65
29	0.08	1.26	1.53	2.63	3.06	7.88
30	0.35	1.26	1.56	2.63	3.11	7.48
32	0.14	1.30	1.59	2.71	3.18	8.36
34	0.09	1.26	1.53	2.62	3.06	6.77
Mean	0.24	1.31	1.60	2.73	3.21	7.23
SE	0.03	0.01	0.01	0.02	0.02	0.09
SD	0.15	0.04	0.05	0.08	0.10	0.47

Table 18. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 10 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.08	1.30	1.59	2.70	3.18	7.03
3	0.08	1.30	1.60	2.71	3.21	7.46
4	0.13	1.38	1.70	2.87	3.41	7.86
5	0.32	1.33	1.62	2.76	3.25	6.79
6	0.04	1.21	1.48	2.52	2.95	6.92
8	0.22	1.34	1.64	2.79	3.28	8.71
10	0.12	1.26	1.54	2.62	3.09	7.74
11	0.07	1.30	1.60	2.71	3.20	7.34
12	0.25	1.30	1.60	2.71	3.21	7.08
13	0.25	1.31	1.61	2.73	3.21	7.34
16	0.06	1.33	1.62	2.77	3.25	8.24
17	0.17	1.35	1.65	2.81	3.31	7.06
19	0.13	1.30	1.59	2.70	3.17	7.80
20	0.09	1.34	1.64	2.79	3.28	7.67
21	0.23	1.35	1.66	2.82	3.32	7.21
22	0.15	1.24	1.52	2.58	3.04	7.65
23	0.05	1.26	1.55	2.62	3.10	7.54
24	0.03	1.30	1.58	2.71	3.17	7.62
26	0.20	1.34	1.63	2.79	3.26	7.66
27	0.12	1.36	1.66	2.82	3.32	7.57
28	0.07	1.36	1.67	2.84	3.34	7.93
29	0.04	1.44	1.78	3.00	3.57	8.48
30	0.13	1.22	1.50	2.53	2.99	6.62
32	0.12	1.29	1.58	2.68	3.16	7.35
34	0.09	1.26	1.54	2.62	3.08	7.31
Mean	0.13	1.31	1.61	2.73	3.21	7.52
SE	0.02	0.01	0.01	0.02	0.03	0.10
SD	0.08	0.05	0.07	0.11	0.13	0.50

Table 19. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 5 s intervals in open field treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.41	1.18	1.44	2.45	2.88	7.69
3	0.19	1.32	1.63	2.75	3.26	10.09
4	0.16	1.40	1.74	2.91	3.47	10.92
5	0.10	1.28	1.56	2.65	3.12	7.88
6	0.47	1.21	1.48	2.51	2.95	8.17
8	0.15	1.31	1.60	2.72	3.20	9.79
10	0.11	1.33	1.61	2.76	3.23	8.67
11	0.25	1.28	1.58	2.67	3.17	8.68
12	0.34	1.36	1.66	2.83	3.33	10.68
13	0.07	1.25	1.53	2.61	3.06	8.50
16	0.15	1.39	1.71	2.89	3.42	10.71
17	0.23	1.26	1.54	2.63	3.08	9.64
19	0.21	1.33	1.62	2.76	3.24	10.76
20	0.10	1.39	1.71	2.88	3.41	10.66
21	0.08	1.33	1.63	2.77	3.27	8.48
22	0.08	1.33	1.64	2.77	3.29	10.73
23	0.14	1.29	1.59	2.69	3.19	8.68
24	0.19	1.27	1.56	2.64	3.11	8.26
26	0.02	1.47	1.80	3.05	3.61	10.82
27	0.42	1.24	1.52	2.59	3.04	8.63
28	0.16	1.41	1.72	2.93	3.45	10.90
29	0.04	1.41	1.74	2.94	3.47	15.19
30	0.34	1.23	1.51	2.57	3.01	7.39
32	0.25	1.40	1.72	2.90	3.44	11.03
34	0.10	1.27	1.55	2.64	3.11	10.61
Mean	0.19	1.32	1.62	2.74	3.23	9.74
SE	0.02	0.01	0.02	0.03	0.04	0.33
SD	0.12	0.07	0.09	0.15	0.18	1.66

Hedgerow treatment

Table 20. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 30 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.06	1.49	1.83	3.09	3.66	11.56
3	0.10	1.65	2.09	3.43	4.18	13.44
4	0.03	1.64	2.05	3.42	4.10	13.40
5	0.12	2.10	2.83	4.37	5.67	34.10
6	0.04	1.72	2.20	3.57	4.40	24.39
8	0.08	2.22	2.91	4.61	5.81	31.92
10	0.10	1.58	1.95	3.29	3.89	13.38
11	0.03	1.51	1.87	3.14	3.74	19.26
12	0.11	1.35	1.70	2.82	3.41	15.13
13	0.24	1.99	2.53	4.15	5.06	39.79
16	0.10	1.74	2.22	3.61	4.44	27.38
17	0.05	1.49	1.89	3.09	3.78	7.64
19	0.06	1.41	1.77	2.93	3.54	8.47
20	0.10	1.65	2.06	3.42	4.12	20.07
21	0.07	1.50	1.89	3.12	3.77	9.47
22	0.07	1.49	1.86	3.10	3.72	10.51
23	0.19	1.75	2.21	3.64	4.42	21.11
24	0.06	1.59	2.00	3.32	4.01	10.02
26	0.16	1.38	1.69	2.87	3.38	7.24
27	0.07	1.33	1.66	2.78	3.32	9.27
28	0.31	1.47	1.82	3.06	3.64	10.60
29	0.20	1.86	2.40	3.87	4.80	17.90
30	0.07	1.41	1.74	2.93	3.48	12.59
32	0.04	1.68	2.11	3.49	4.21	17.38
34	0.17	1.37	1.71	2.86	3.41	8.41
Mean	0.11	1.62	2.04	3.36	4.08	16.58
SE	0.01	0.05	0.07	0.10	0.13	1.78
SD	0.07	0.23	0.34	0.48	0.67	8.88

Table 21. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 10 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.06	1.58	1.96	3.29	3.91	18.87
3	0.13	1.54	1.94	3.21	3.89	12.89
4	0.03	1.64	2.06	3.42	4.12	12.86
5	0.12	1.50	1.89	3.12	3.77	24.67
6	0.04	1.57	1.95	3.26	3.90	17.13
8	0.08	1.75	2.18	3.64	4.36	8.37
10	0.04	1.66	2.04	3.45	4.07	14.78
11	0.03	1.48	1.81	3.09	3.62	18.71
12	0.11	1.37	1.73	2.85	3.45	14.14
13	0.03	1.94	2.45	4.03	4.90	46.11
16	0.10	1.58	1.94	3.28	3.88	11.46
17	0.05	1.71	2.14	3.55	4.27	16.27
19	0.06	1.68	2.08	3.50	4.16	10.68
20	0.10	1.61	1.96	3.34	3.93	10.80
21	0.04	1.66	2.05	3.45	4.10	8.34
22	0.07	1.56	1.97	3.25	3.95	14.99
23	0.17	1.60	1.98	3.33	3.97	9.80
24	0.06	1.58	1.98	3.28	3.97	9.13
26	0.02	1.39	1.73	2.89	3.46	8.14
27	0.06	1.43	1.77	2.97	3.55	9.12
28	0.27	1.40	1.73	2.91	3.45	11.48
29	0.11	1.63	2.01	3.39	4.02	20.12
30	0.07	1.57	1.91	3.28	3.83	10.57
32	0.04	1.75	2.22	3.64	4.45	19.33
34	0.17	1.36	1.68	2.82	3.36	8.19
36	0.36	1.44	1.81	2.99	3.62	15.13
Mean	0.09	1.58	1.96	3.28	3.92	14.70
SE	0.02	0.03	0.03	0.06	0.07	1.55
SD	0.08	0.13	0.17	0.28	0.35	7.76

Table 22. Mean horizontal accuracies of each of the 25 standardised GPS receivers sampled at 5 s intervals in hedgerow treatment

GPS receiver	Minimum (0%)	CEP (50%)	1 Sigma (68%)	R95 (95%)	2 Sigma (98%)	Maximum (100%)
2	0.06	1.83	2.28	3.82	4.56	10.07
3	0.07	1.52	1.89	3.17	3.78	11.68
4	0.03	1.57	1.93	3.26	3.86	8.47
5	0.12	1.94	2.43	4.03	4.86	35.09
6	0.04	1.40	1.73	2.91	3.47	17.39
8	0.08	1.90	2.38	3.95	4.75	10.94
10	0.13	1.54	1.91	3.20	3.82	14.38
11	0.03	1.34	1.63	2.80	3.25	9.65
12	0.11	1.48	1.87	3.08	3.75	16.01
13	0.15	2.32	2.97	4.82	5.94	55.11
16	0.10	1.70	2.09	3.54	4.18	17.40
17	0.05	1.71	2.12	3.56	4.23	8.82
19	0.06	2.01	2.47	4.19	4.94	10.50
20	0.10	1.91	2.39	3.97	4.78	13.67
21	0.04	1.78	2.20	3.69	4.40	9.46
22	0.11	1.72	2.13	3.59	4.26	12.42
23	0.17	1.58	1.92	3.29	3.85	11.24
24	0.13	1.92	2.31	3.99	4.62	11.62
26	0.02	1.31	1.63	2.73	3.27	7.41
27	0.06	1.71	2.07	3.56	4.13	10.21
28	0.07	1.40	1.73	2.91	3.47	13.57
29	0.11	2.54	3.33	5.29	6.66	27.66
30	0.07	1.75	2.17	3.64	4.33	17.13
32	0.04	1.86	2.40	3.88	4.81	29.33
34	0.17	1.45	1.80	3.01	3.61	9.01
Mean	0.08	1.73	2.15	3.59	4.30	15.93
SE	0.01	0.06	0.08	0.12	0.16	2.14
SD	0.04	0.29	0.40	0.61	0.79	10.68