Available online at www.sciencedirect.com

ScienceDirect

Computers and Electronics in Agriculture 56 (2007) 46–59

Computers and electronics in agriculture

www.elsevier.com/locate/compag

Robust classification of animal tracking data

Mac Schwager[a,*], Dean M. Anderson[b], Zack Butler[c], Daniela Rus[a]

[a] Distributed Robotics Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] USDA-ARS, Jornada Experimental Range, Las Cruces, NM 88003, United States
[c] Computer Science Department, Rochester Institute of Technology, Rochester, NY 14623, United States

Received 2 June 2006; received in revised form 3 November 2006; accepted 5 January 2007

Abstract

This paper describes an application of the K-means classification algorithm to categorize animal tracking data into various classes of behavior. It was found that, even without explicit consideration of biological factors, the clustering algorithm repeatably resolved tracking data from cows into two groups corresponding to active and inactive periods. Furthermore, it is shown that this classification is robust to a large range of data sampling intervals. An adaptive data sampling algorithm is suggested for improving the efficiency of both energy and memory usage in animal tracking equipment.

Keywords: Cluster analysis; GPS; Animal tracking; Adaptive sampling; Sensor networks

1. Introduction

The use of mobile sensors promises a data-rich future for the animal behavioral sciences. However, with large data sets also comes the burden of data analysis and interpretation. In the future, the human analyst will rely on computational techniques to meaningfully interpret large data sets and discover meaningful statistical trends. In this work we propose one simple computational tool to help animal researchers interpret large data sets.

Specifically, we introduce the use of the K-means classification algorithm (Arabie et al., 1996) to classify animal tracking data without *a priori* information into two categories. These categories are found to correspond to periods of animal activity and inactivity. We use position and head angle data from several cows to demonstrate how this algorithm could be employed in a behavioral study involving free-ranging animals. Furthermore, we investigate a particularly useful property of this classification algorithm—its robustness with respect to data sample rates. This robustness suggests that the classifier can be used to drive an adaptive sample rate data collection system, which will have the potential to save energy without sacrificing the information content of the collected data. Robustness of the algorithm to different sample rates is also demonstrated using cow tracking and head angle data.

We have not attempted to carry out a conclusive study of free-ranging cow activity and inactivity in this work, but, rather, suggest a method in general terms and use cow data to demonstrate how the algorithm can be useful in a typical animal behavioral study. It should also be noted that the applicability of the method is not tied to the specific example or equipment described herein. The method can equally be employed to classify tracking data from any species, including

* Corresponding author. Tel.: +1 617 253 6532; fax: +1 617 253 6849.
E-mail addresses: schwager@mit.edu (M. Schwager), deanders@nmsu.edu (D.M. Anderson), zjb@cs.rit.edu (Z. Butler), rus@csail.mit.edu (D. Rus).

0168-1699/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2007.01.002

humans, regardless of the tracking instrument used. Indeed, our broader research agenda is to develop computational approaches for studying groups of various kinds of dynamical agents with social interactions. We expect that such dynamical groups can be studied and modelled using physical data, and ultimately controlled.

1.1. Relationship to current state of knowledge

Several recent animal behavior studies have made use of automated data collection systems to allow for computationally intensive post-processing analysis (Butler et al., 2006; Bishop-Hurley et al., 2005; Anderson, 2001). An interdisciplinary group at Princeton University has developed a network for monitoring the movements and social groupings of Zebras in Kenya (Juang et al., 2002). Researchers at the University of California, San Diego have used pattern recognition algorithms to identify and classify worm behavior (Geng et al., 2003). Statistical clustering techniques have also been adapted to classify dolphin whistle calls (McCowan, 1995), and to identify animal aggregations (Strauss, 2001).

More directly related to our work, some recent studies have been conducted to infer activity states of free-ranging animals using global positioning system (GPS) tracking data. Ganskopp (2001) found that the Euclidean distance between successive GPS fixes was not sufficient to infer cattle activity states using a regression analysis. Schlecht et al. (2004) compared classification results from human observers with those found from discriminant analysis of GPS data and found that the two were in agreement for 71% of the data. This work required an initial calibration step using a portion of the collected data to manually select discriminants that were then used to classify the remainder of the data. Finally, Ungar et al. (2005) used GPS position and acceleration data from cattle to infer states of activity. They found that discriminant analysis classification agreed with human observation for at least 74% of the data while regression tree classification agreed with human observation for at least 84% of the data. The K-means algorithm discussed in this paper differs from the above statistical methods in that it performs classification in a highly autonomous fashion. In particular it requires only minimal data preparation and does not require an off-line calibration step, nor does it require the human analyst to pre-select suitable discriminants on which to base a regression model. In the machine learning community, K-means is considered a simple example of an unsupervised learning algorithm, whereas the above methods would be considered supervised algorithms (Duda et al., 2001).

This research points toward a potential application of the K-means algorithm to help ease power and memory requirements for data collection equipment. Current research has identified the difficulty of balancing data resolution with technical limitations, particularly on-animal power and memory requirements (Anderson, 2006). In order to mitigate power and memory constraints, most commercial on-animal GPS devices are restricted to a maximum position recording rate of once every 5 min (this trend is changing, however (Clark et al., 2006)). This rather infrequent data acquisition rate may cause significant amounts of information about landscape utilization to be missed, especially during foraging (Schlecht et al., 2004) and possibly even during non-foraging travel (Estevez and Christian, 2005). This is a problem of critical importance since most current animal behavior rangeland research is focused on understanding animal distribution in order to accurately characterize landscape utilization and ultimately manage utilization (Anderson, 2006; Bailey, 2005; DelCurto et al., 2005; Bailey, 2004; Bailey et al., 2001; Hulbert and French, 2001; Turner et al., 2000; Coppolillo, 2000; Bailey et al., 1996; Coughenour, 1991; Pinchak et al., 1991; Roath and Krueger, 1982).

Thus, choosing an appropriate data sampling rate appears to be a principle factor in understanding free-ranging animal behavior. This highlights the advantage of creating a data collection system that has an adaptive sample rate. With such a system, data may be collected infrequently during periods for which high sample rates do not help to determine land utilization, while data can be sampled more frequently during periods that are more critical to understanding land utilization. The results of this paper suggest that the K-means algorithm would be well suited to trigger a change in data sample rates in such a data collection system because its classification qualities are insensitive to sample rates.

2. Methods

Our technical approach was as follows: (1) data from several free-ranging cows were collected; (2) data were then analyzed using the K-means algorithm to classify animal behavior into active and inactive states; (3) biological factors associated with activity and inactivity were compared with the autonomous classification to determine if the categories corresponded with actual periods of activity and inactivity; (4) the effects of decreasing sample rates on the resulting sampled path was investigated for active and inactive states; (5) the effects of decreasing sample rates on the K-means classification algorithm was investigated for active and inactive states.

2.1. Tracking data

2.1.1. Data collection location

Free-ranging cow data were collected in a 466 ha (1103 ac) area (Paddock 10B), on the U.S. Department of Agriculture, Agricultural Research Service's (USDA-ARS), Jornada Experimental Range (JER). This site has an undulating topography of predominantly sandy soil with a mean elevation of 1260 m (4134 ft) above sea level(106°43.263″W, 32°34.297″N) located approximately 37 km (23 mi) north of Las Cruces, NM. The climate of the region is typical of arid rangeland having a long-term mean precipitation approaching 230 mm (9.1 in.) with 52% occurring between July and September. Mean maximum ambient air temperatures vary from a high of 36 °C (96.8 °F) in June to below 13 °C (55.4 °F) in January.

The major grass species found in Paddock 10B are *Bouteloua eriopoda* (Torr.) Torr., *Sporobolus flexuosus* (Thurb.) Ryudb. and *Aristida purpurea* Steud. Shrubs include *Prosopis juliflora var glandulosa* (Torr.) Cock, *Yucca elata* Engelm. and *Gutierrezia sarothrae* (Pursh) Britt. + Rusky. The few low-lying areas with heavier soils are dominated by *Scleropogon brevifolious* Phil, and *Sporobolus airoides* (Torr.) Torr.

2.1.2. The animals

Three free-ranging mature beef cattle of Hereford and Hereford × Brangus genetics, labeled Cow 1, Cow 2, and Cow 3, were monitored during two periods in 2004. Data were collected in 2004 beginning April 26 at 17:12 h through April 28 at 08:55 h (Trial 1), and again from May 17 at 12:50 h through May 19 at 07:44 h (Trial 2). During these intervals, uncorrected global positioning system (GPS) data were recorded with Directional Virtual Fence (DVF^{TM}) devices capable of providing sensory cues to alter the animal's direction of movement on the landscape (Anderson, 2006; Anderson et al., 2004; Anderson and Hale, 2001; Anderson, 2001). However, during these two trials, the animals received no cues from the DVF^{TM} devices.

The data included date, time, Universal Transverse Mercator (UTM) location together with the horizontal and vertical angle of the animal's head. Head angles were measured in degrees from a reference position corresponding to the head being level with the animal's backbone while looking straight ahead. A downward tilted head angle was denoted as negative and an upward tilted head angle positive. Similarly, a left tilting head angle was denoted as negative while a right tilting head angle was positive. Head angles can be indicative of certain kinds of behavior. For instance, while foraging, the cow's head is likely to be angled downward toward the ground and while resting it is likely to be looking straight ahead or elsewhere (Anderson, 2006). An electronic magnetometer within the DVF^{TM} device provided these head angle measurements. During Trial 2, the instrument worn by Cow 1 broke leaving a total of only five complete data sets from Trial 1 and Trial 2 upon which these analyses were made.

For each trial, data consisted of approximately 3000 entries for each cow collected at intervals alternating between 43 and 53 s, with some additional sampling irregularities. Data points that located the animal outside of the paddock fences were removed manually as obvious outliers, however, no other data processing was carried out beyond what is stated explicitly below.

2.2. The K-means classifier

The K-means algorithm was used to classify each animal's data into two categories without using *a priori* information. The resulting categories were examined and were determined to correspond to clearly defined periods of activity and inactivity. The algorithm was applied with the intention of testing the simplest possible approach, thus special considerations were not made to filter out inherent GPS measurement noise.

Specifically, we used the K-means classification algorithm to group the data into two categories. The classification was performed over three dimensions of the collected data: speed, horizontal head angle, and vertical head angle. The speed was calculated from a numerical differentiation according to the formula:

$$s_k = \frac{\|x_{k+1} - x_k\|}{t_{k+1} - t_k} \tag{1}$$

where s is the speed in m/s, x the GPS position vector, t the time stamp from the on board clock, k indicates the data index number and $\|\cdot\|$ denotes the Euclidean distance. The horizontal and vertical head angles, θ_{h} and θ_{v}, respectively, were collected directly from the DVF$^{\mathrm{TM}}$. We note that time of day was not considered in this analysis, nor was temporal auto-correlation of the data, that is, all data points for a single cow are simply put into a set for the calculations irrespective of their temporal separation from one another. Our intention was to determine what clusters could be found in the data using the most naive possible approach.

Before being used for classification, the data were normalized by subtracting the mean from each dimension, s, θ_{h}, and θ_{v}, and dividing each dimension by its standard deviation, In particular:

$$z_k = \frac{y_k - \mu}{\sqrt{(1/k)\sum_k (y_k - \mu)^2}}, \tag{2}$$

where $\mu = (1/k)\sum_k y_k$ is the mean, y represents the dimension, and z represents the dimension after normalization. This standard procedure was used to remove effects from constant offsets and scaling. Each data entry was arranged as a vector $v_i = [s_i\ \theta_{\mathrm{h}_i}\ \theta_{\mathrm{v}_i}]$, and the resulting normalized data vectors were used for classification.

Algorithm 1: K-means classification

```
Require: Initial mean positions, $\mu_j, j = 1, \ldots, K$
Require: A norm, $\|\cdot\|$, by which to measure the distance between any two data vectors
Require: A desired final tolerance, $e$, for the means
repeat
  Group each data vector with its closest mean (the one for which $\left\| v_i - \mu_j \right\|$ is smallest)
  Compute the centroid of each group according to $(1/n_j)\sum_{N_j} v_i$, where $N_j$ is the set of indices and $n_j$ is the size of the group
  Take these new centroids as the new mean positions $\mu_j$
until $\left\| \mu_j^{\mathrm{new}} - \mu_j^{\mathrm{old}} \right\| < e$ for all $j = 1, \ldots, K$
```

The K-means algorithm was carried out using the Euclidean norm with each dimension equally weighted. The algorithm is listed here as Algorithm 1. This algorithm is guaranteed to stop for arbitrarily small e values. Unfortunately, the final mean positions and resulting groupings are known to be sensitive to the initial mean positions chosen by the user. A more elaborate discussion of K-means and other classification methods can be found in Aldenderfer and Blashfield (1984), Arabie et al. (1996), and the original presentation of the algorithm is given in McQueen (1967). The algorithm was carried out as described above independently on the data for each individual cow.

3. Results

K-means repeatably produced two clusters from the data sets, clearly delineated into a category with low speed and high head angle, which we call inactive, and one with high speed and low head angle, which we call active. The classifier identifies the horizontal head angle (right versus left) as being inconsequential to the classes. Classification results for the three cows during Trial 1 are shown in Figs. 1 and 2, and results for the two cows in Trial 2 are shown in Figs. 3 and 4. Note that the classification is cohesive among periods of time. That is, adjacent data points were very likely to be in the same category despite the inherent random GPS position errors in the data. Thus the algorithm appears to be sufficiently robust with respect to random position errors and is not hindered by the fact that temporal relationships between successive data points were not used in the input to the classifier. Also, note that the periods of activity and inactivity are similar for all animals within the same trial. The classification was carried out independently for each animal, thus the classifier has no knowledge of the context of the experiment nor of the location of one cow relative to another. Therefore, the correlations that exist among animals serve a means of validating our results. This interpretation suggests that the classifier has identified biologically relevant behavioral categories, since activity levels within a group of gregarious animals mutually influence one another (Smith, 1998; Immelmann and Beer, 1989).

To further investigate whether or not the active and inactive states identified by the classifier were biologically relevant, we compared statistics reflecting aggregation behavior within the two categories. In particular, the mean distance from each cow to the herd centroid and its standard deviation show that the animals were significantly and reliably closer together during periods of common inactivity than during activity. These data are shown in Fig. 5. The clear difference in the placement and activity of the herd during these periods gives evidence that the classifier has

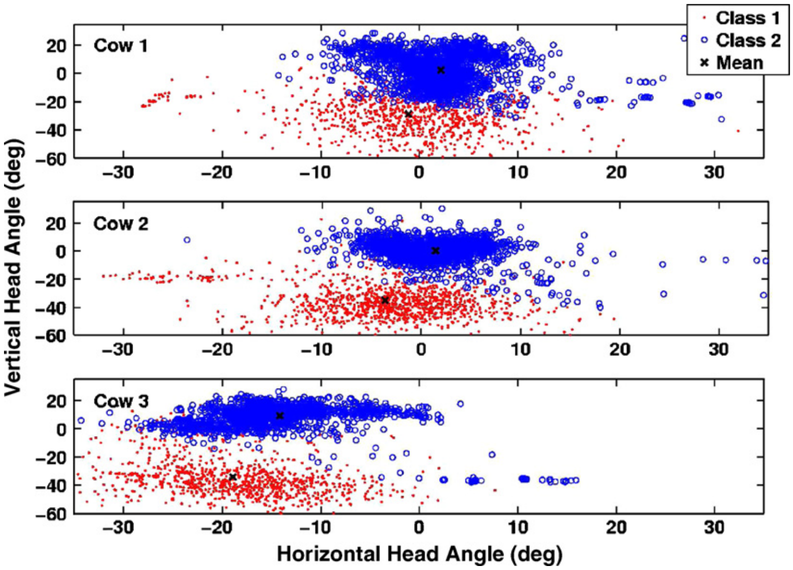

Fig. 1. K-means classification results are shown for the head angles of the three cows over the course of Trial 1. The class shown by solid dots is interpreted as the active category, while the class shown by open circles is interpreted as the inactive category. Class means are shown by a solid "×". The active category has a lower vertical head angle than the inactive category on average, suggesting foraging behavior. The data for Cow 3 indicate that the collar worn by the animal was approximately 15° off center from the animal's spine.

identified relevant biological states. Again, we stress that the classifier worked independently upon data from each cow.

3.1. The problem of validation

There are often subtle problems associated with using automated statistical processing techniques. Janik (1999) showed that humans presented with the task of classifying dolphin whistles produced far more uniform classifications

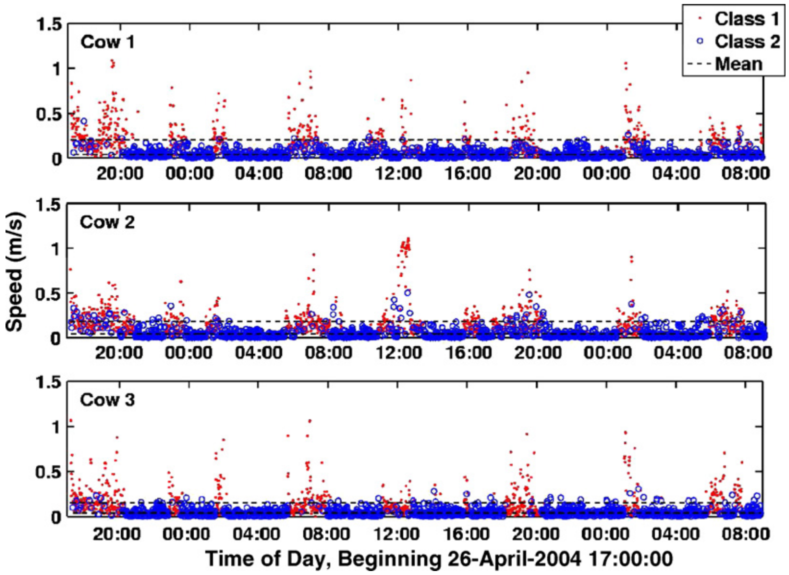

Fig. 2. K-means classification results are shown for the speed of three cows during Trial 1. The class shown by solid dots is interpreted as the active category, having a higher mean speed, while the class shown by open circles is interpreted as the inactive category. Class means are shown by a dotted line.

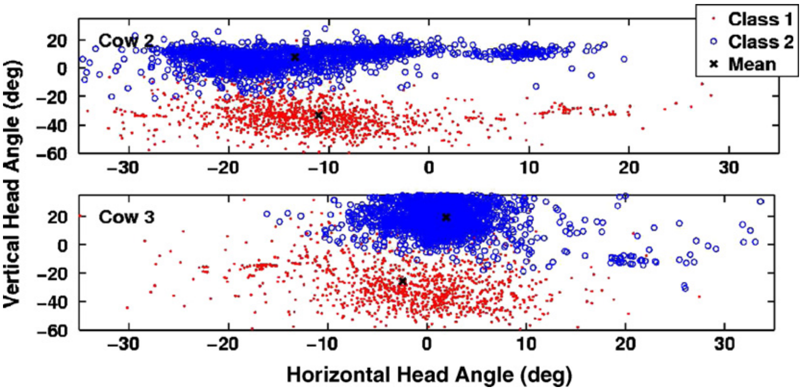

Fig. 3. K-means classification results are shown for the head angles of the two cows over the course of Trial 2. Data for Cow 1 is unavailable as the device failed during the trial. The class shown by solid dots is interpreted as the active category, while the class shown by open circles is interpreted as the inactive category. Class means are shown by a solid "×". The active category has a lower vertical head angle than the inactive category on average, suggesting foraging behavior. The data for Cow 2 indicate that the collar worn by the animal was approximately 12° off center from the animal's spine.

than clustering algorithms presented with the same data. Caution must be exercised in the application of such algorithms to make sure that results are not sensitive to arbitrary or insignificant factors. On the other hand, some findings suggest that discrepancies between human and statistical classification methods are more likely due to human error (Rutter et al., 1997). Indeed, the "true" activity state of an animal is not a precisely defined concept.

In this work we have deliberately avoided attempts to define a "true" activity state by not comparing the classification results with results from human observers. Instead, we appeal to the clear differentiation of similar categories among different animals and different time periods produced by the algorithm. When we use the term "active class" we specifically mean that class defined by the unsupervised classifier, which has a relatively high mean rate of travel. Thus we are not seeking in this work to confirm the classification results with human observation, but to demonstrate the regularity of the K-means classifier as an autonomous agent. Indeed, our proposed application of the algorithm as a trigger for adaptive data sampling is not concerned with maintaining fidelity to human classification, but rather, with determining when it is safe to use a low sample rate verses a high sample rate-likely a task for which a human observer would not be well suited.

3.2. Changing sample intervals

Next, the effect of changing sampling intervals during inactive and active periods was investigated. We developed an intuitive measure by which to quantify the loss of information when sampling the same cow path at different sample

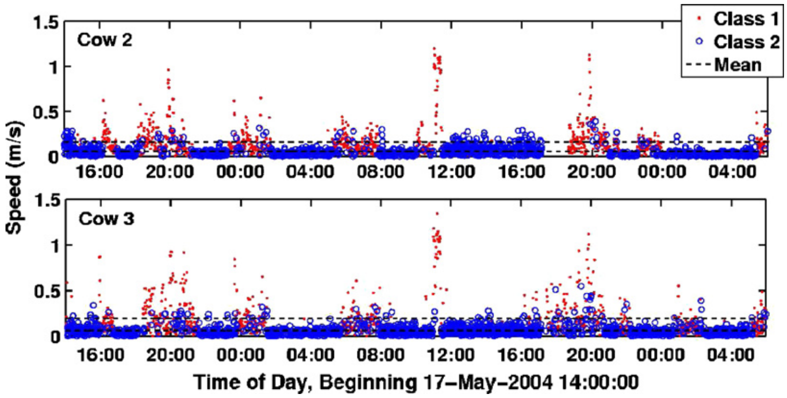

Fig. 4. K-means classification results are shown for the speed of two cows during Trial 2. The class shown by solid dots is interpreted as the active category, having a higher mean speed, while the class shown by open circles is interpreted as the inactive category. Class means are shown by a dotted line.

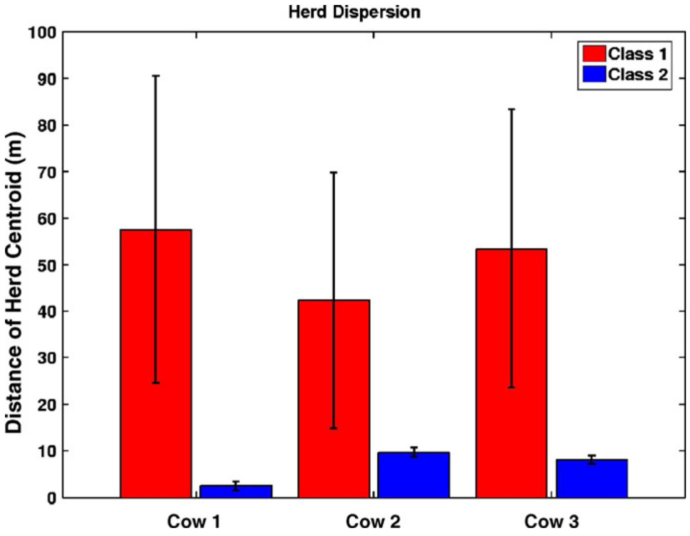

Fig. 5. The mean distance and standard deviation of each cow from the herd centroid is shown for Trial 1 during the two categories identified by the K-means classifier. Class 1 is interpreted as active, while Class 2 is interpreted as inactive. The significant difference between the two categories suggests that the activity states identified by the classifier are biologically relevant.

rates. Namely, for a given path X^{s} with high sample frequency, we regularly subsample it (i.e. taking every *n*th data point, for increasing values of *n*) to produce a low frequency data set X^{l}. Then, for each data point, X_k^{s}, we measure how accurately it would be approximated by X^{l}. This is done by linearly interpolating X^{l} at the time of X_k^{s} to get an estimated point X_k^{l} and measuring the Euclidean distance $\left\|X_k^{\mathrm{s}} - X_k^{\mathrm{l}}\right\|$. These distances are then averaged over the length of X^{s} to determine the overall loss of path information in X^{l}. This analysis is shown in schematic form in Fig. 6.

Note that this measure is in no way related to the literal distance covered by a path. It is, however, closely related to the tortuosity of a path. A more tortuous path will suffer a greater loss of information when it is subsampled. Thus

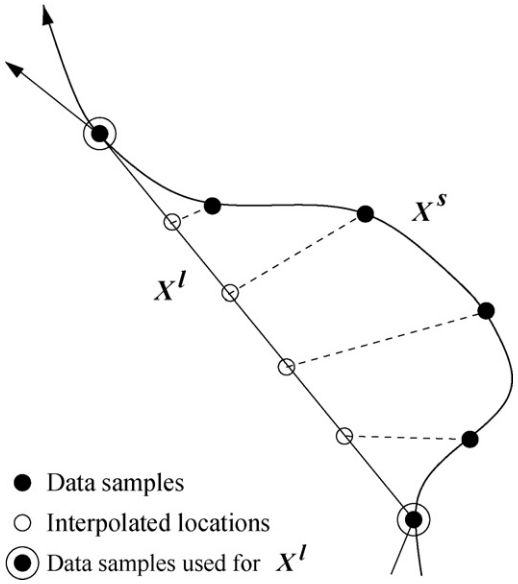

Fig. 6. A schematic depicting the procedure used for calculating path error for paths with data points removed. The solid dots indicate positions sampled, the open circles indicate interpolated positions for which no sample was taken, and the dotted lines show the distances between a recorded data point on the true path and the point obtained from interpolation on the subsampled path. The path error is the mean of all such lengths over a total path.

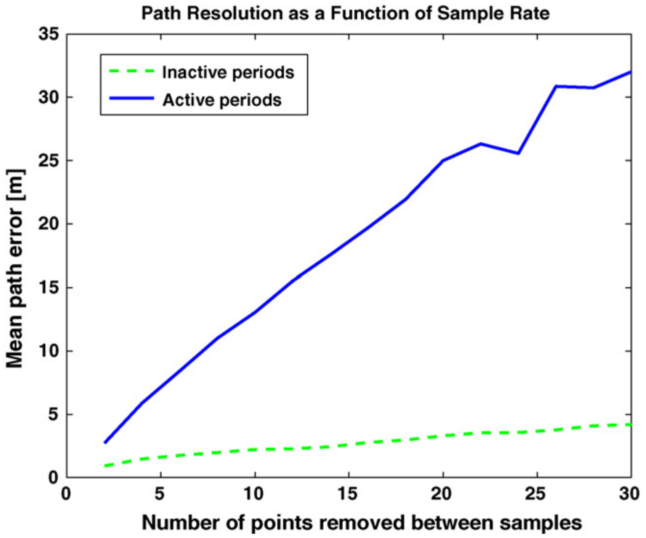

Fig. 7. The path error is shown as a function of sampling frequency. The error metric used is described in Fig. 6. The path error for inactive and active periods are given by the dotted and solid lines, respectively. The plot suggests that during inactive periods, a relatively low sample rate can be used without introducing path error, while for active periods, a relatively high sample rate is required to accurately resolve an animal's path.

the measure described above will have a steep increase as a function of sample rate for a tortuous path. Conversely, a perfectly straight path will suffer no loss of information from subsampling as long as we keep at least two data points from the path. Thus the measure described above will have a flat slope as a function of sample rate for a perfectly straight path.

For the data collected in this study, the actual sample interval alternates between 43 and 53 s per data entry. This was used as the baseline path X^s for comparison, with increasing fractions of the data points removed from the baseline data set to simulate the effect of increasingly longer sample times. For each segment of the subsampled path X^1 the distance error was assigned to one of two tallies, depending on the classification of the two nearest points in X^s. That is, if both subsampled points were classified as active, the distance error along the segment was added to the total error for active points, and likewise for inactive points. The distance measures were then averaged over all cows in Trial 1 to produce Fig. 7, which shows the path error for increasing sample intervals for both an inactive (dashed line) and active (solid line) path. The figure demonstrates that for an inactive path little information is lost by using a comparatively long sample interval, while for an active path the information about the path depends strongly on the sample rate. This can be seen more directly in Figs. 8 and 9, which show the appearance of representative sections of inactive and active paths with increasing sample intervals. Clearly the inactive path changes minimally, in fact it appears almost as a stationary dot, while the change to the active path is more severe. The measure of path information loss, calculated as described above, is displayed on each plot. The figures suggest that during active periods, samples should be taken frequently, perhaps on the order of once per minute or more, to sufficiently capture the animal's activity. During inactive times, sample periods can be significantly longer without loss of resolution. A wealth of field research supports this finding (Ganskopp, 2001; Schlecht et al., 2004; Estevez and Christian, 2005).

The effect of increasing sample rates on the classification results was also investigated. The original input to K-means was subsampled similarly to the path error study by retaining every *n*th data point, for increasing values of *n*. Note that the data are not exactly the same, as the speed value used for classification is not inherent in the tracking data, but rather calculated based on the difference in position between consecutive readings, as in (1). Thus, in the subsampled data, the speed is recalculated based on the data points in the smaller set. We then examined the changes in the clusters produced as the data set was increasingly more coarsely sampled. Fig. 10 shows the Euclidean distance between the final mean values of the clusters and the means produced by the full data set as the sample interval was increased. This is significant because the location of the final mean values define the characteristics of the two classes identified by the classifier. The figure shows that errors in the mean positions are minor for sample intervals up to 20 times as long as the original sample interval (about 16 min between each sample). Figs. 11 and 12 show directly the classification results of various time intervals. It is evident that despite the increasingly sparse data, the final categories found by

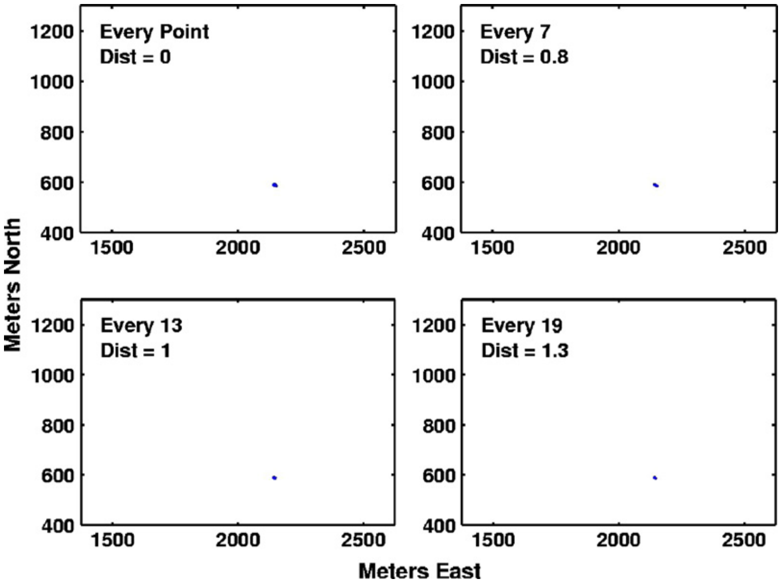

Fig. 8. An inactive path segment sampled at different frequencies. Data are from Cow 1, Trial 1, April 27, 2004, 02:12–04:52. The cow moved so little during inactivity that the path appears as a dot. It is clear that the path does not change appreciably with increasing sample intervals.

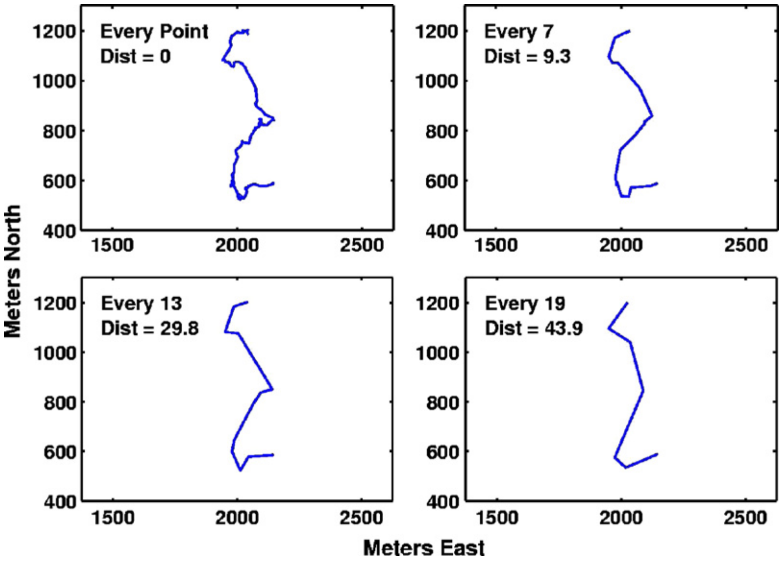

Fig. 9. An active path segment sampled at different frequencies. Data are from Cow 1, Trial 1, April 27, 2004, 05:24–07:35. It is clear that increasing sample intervals causes significant path information to be lost.

the classifier are similar. Additionally, the reliability of the classification degrades gradually as sampling intervals are increased beyond 20 times the original interval.

4. Discussion

4.1. Biological significance

The most important biological reason to understand how animals use landscapes will be to prevent over and under utilization, and to understand how animal utilization of rangeland can provide proactive information for implementing management practices (Heitschmidt and Taylor, 1991). Proper utilization remains the second biggest challenge in

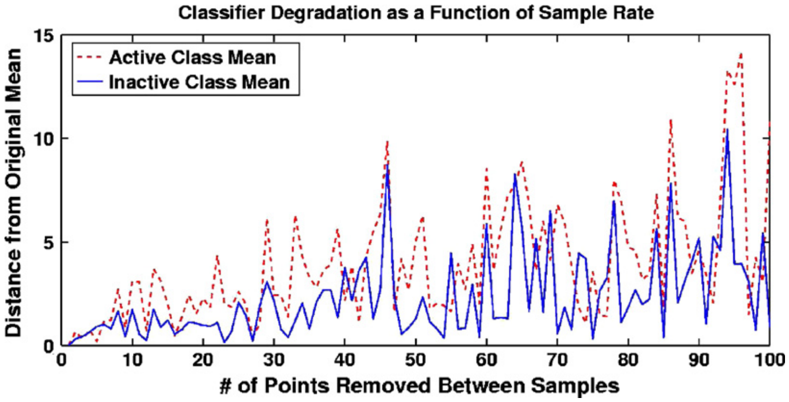

Fig. 10. The change in class means are shown as sampling interval is increased. Original mean refers to the class means computed from all data points; distance is a dimensionless quantity in the normalized space of the classification. The change in means for the active and inactive classes are given by the dotted and solid line, respectively. Notice that the class means change insignificantly for data sets with only 5% of the original data (20 points removed between samples).

the management of free-ranging herbivores (Holechek et al., 1998). By taking data at the optimal sampling rate for free-ranging herbivory based on different activity states, we will have at our disposal the information necessary to make informed proactive management decisions. This will especially be true when virtual fencing (Anderson, 2006) becomes commercially available, making real-time management possible.

The analysis developed in this paper can classify location data into two coarse categories, active and inactive. When these data are combined with other biotic as well as abiotic information about the landscape, we would expect finer grained categorizations of behavior to be possible with this algorithm. The classification approach used here was intentionally made without using any *a priori* biological knowledge. The fact that biologically relevant categories, i.e. periods of activity and inactivity, were produced is encouraging. Furthermore, the duration and frequency of the active category appears to agree with periods of activity previously reported for cattle (Wagnon, 1963; Squires, 1981; Arnold and Dudzinski, 1978).

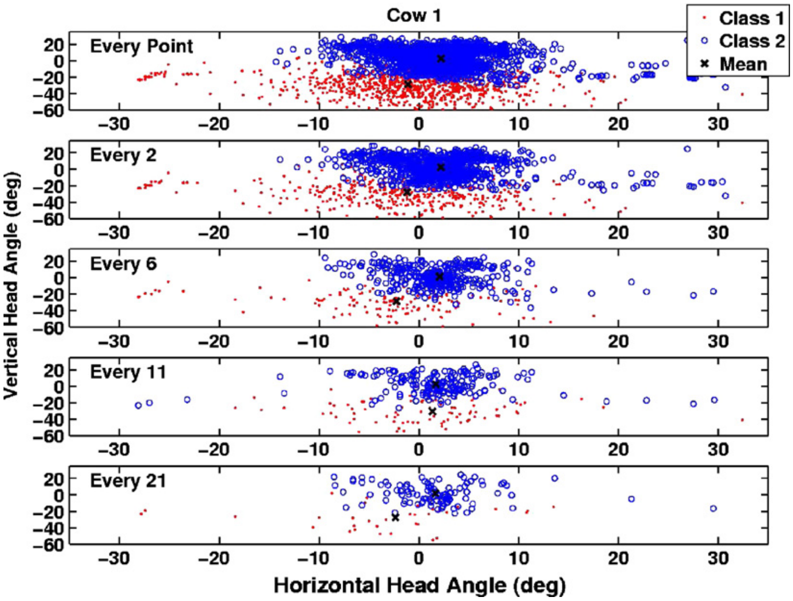

Fig. 11. Classifier results are shown for the head angles of Cow 1 during Trial 1 with increasing sample intervals. The class shown by solid dots is interpreted as the active category, while that shown by open circles is interpreted as the inactive category. Class means are shown by a solid “×”. Notice that the classes and mean locations change little despite a significant decrease in the amount of data used for classification.

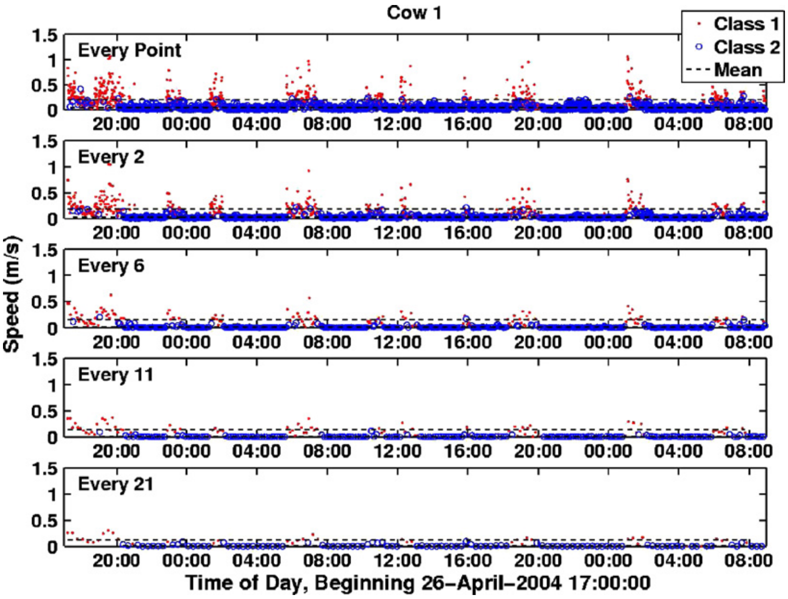

Fig. 12. Classifier results are shown for the speeds of Cow 1 during Trial 1 with increasing sample intervals. The class shown by solid dots is interpreted as the active category, while the class shown by open circles is interpreted as the inactive category. Class means are shown by a dotted line. Notice that the classes and mean locations change little despite a significant decrease in the amount of data used for classification.

In addition, similar analysis for the data obtained during a trial for which the animals received cues from the DVF^{TM} device Anderson (2006) show similar categories of behavior (see Fig. 13). This indicates that the presence of cuing does not significantly alter the results of classification. We present these as preliminary results for data classification during cuing, since more thorough investigation is required. However, these results suggest that a classification algorithm could be used with a virtual fencing system in the future.

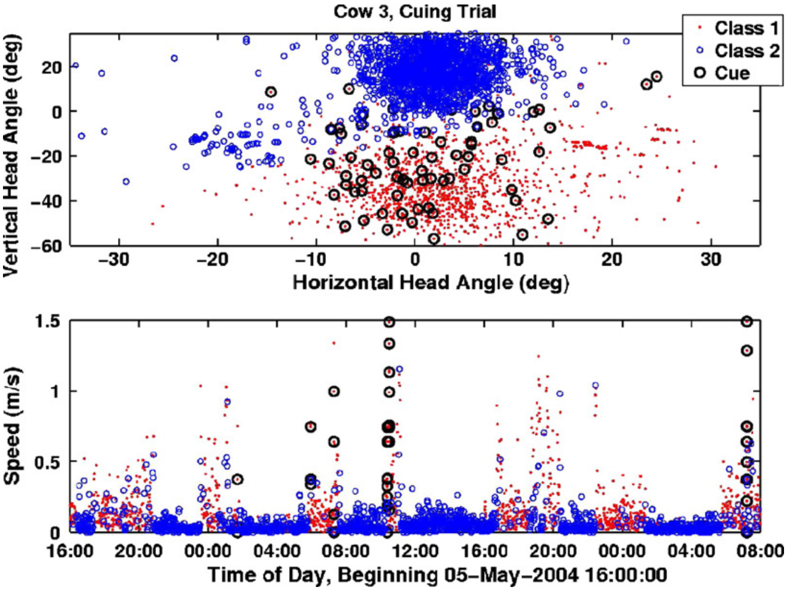

Fig. 13. Classification for a data set in which Cow 3 experienced Directional Virtual Fencing (DVF^{TM}) cuing is shown. The class depicted with solid dots is interpreted as the active category, while the class depicted with open circles is interpreted as the inactive category. Data points for which cuing was administered by sound, electric stimulation, or both are indicated by bold open circles. The results suggest cuing does not significantly affect the clear definition of an active and inactive category. Also notice that cuing occurs almost exclusively during periods of activity, as would be expected. The few points in the inactive category for which cuing was administered were categorized incorrectly.

We expect that incorporating additional biological knowledge into the algorithm will enable the computation of finer-grained classification and more complete real-time understanding of the animals' behavior. For example, it is clear that biotic and abiotic landscape parameters (density of forage, topography, wind speed and direction, etc.) affect cattle (Hafez and Bouissou, 1975; Squires, 1981). By adding these dimensions to our data set, we may be able to evaluate how activity varies in a correlated way with landscape parameters. Besides individual behavior, we are also interested in the relative positions of the animals and the movement of the herd. Including this information in a classification algorithm could lead to intraspecial, as well as interspecial, animal behavior models to characterize how animal behavior changes as a function of neighbor proximity and herd dispersion or aggregation.

4.2. Other data collection devices

Previous GPS-based animal behavioral research projects have used both modified and off-the-shelf GPS units. For instance, Schlecht et al. (2004) used customized backpacks with modified GPS units which recorded animal positions every 10 s, while Ungar et al. (2005) used commercially available Lotek$^{\mathrm{TM}}$2000 and 2200LR collars that logged data at 20 and 5 min intervals, respectively, combined with on-collar accelerometer measurements to aid in classification. Our data were collected from DVF$^{\mathrm{TM}}$ devices (Anderson, 2006), which are not yet commercially available. This breadth of devices and sample intervals used for collecting data raises the question of whether or not the devices themselves may influence the results of studies.

Our analysis of path lengths above attempts to addresses this question. It is always advantageous to collect data as frequently as possible. Statistically, more frequent data sampling can only improve the fidelity of an analysis. Thus data collection rates ought to be limited only by device energy and memory restrictions ultimately imposed by cost constraints. It would seem that logging data at 10 s intervals is sufficient to capture the most erratic of cow paths with great detail (Schlecht et al., 2004), however energy and memory would be wasted on periods during which the animal is inactive. Conversely, logging data every 20 min (Ungar et al., 2005) will lose much fine-grained information about a cow's path, but will give devices a long lifetime in the field. The researcher ought to collect data at a resolution appropriate for the objectives of the study. In any case, the K-means algorithm has been shown above to be insensitive to sample rates, thus it can be used safely with any reasonable data collection device to classify activity levels.

4.3. Adaptive sampling

The robustness of the K-means algorithm to sample rates ranging from once per minute (or less) to once every 16 min suggests that the K-means algorithm can be used to drive a variable sample interval data collection device—one which records samples at a high frequency during periods of activity and records at a lower frequency during periods of inactivity. The results show that during inactive periods there is little benefit to a high sample rate, while during active periods a high sample rate is required to resolve important properties of an animal's path.

From an engineering point of view, these results suggest guidelines for animal device design as well as experiment design. Both memory and power could be used more efficiently if the data collecting device were able to change its sample rate given the activity state of the animal. Fortunately, the K-means classifier has an on-line variant, which can be used to classify data as it is obtained. The identified category of the incoming data can then be used to dictate the future sample rate at which data will be recorded. Such an adaptive sampling algorithm could provide significant gains in over-all device lifetime.

5. Conclusion

The K-means classification algorithm was used to cluster GPS tracking data obtained from mature cows into two categories representing activity and inactivity. We found that the categories computed by the algorithm were consistent among animals within a trial for two different trials. These results do not address questions about cattle behavior at large, but, rather, demonstrate that the K-means algorithm can be useful in determining animal activity states in an autonomous way. This research also suggests that a sufficiently high data sampling rate is important for characterizing land utilization by cattle during periods of activity, while sample rate is not important during periods of inactivity. Lastly we have demonstrated that K-means classification is robust with respect to sample intervals of various length. These results suggest that an on-line K-means algorithm can be used to identify between periods of activity and inactivity and subsequently adjust the data sampling rate accordingly. Such an adaptive sampling scheme will provide improved resolution while conserving memory and energy in future data collection devices.

Acknowledgments

This work was supported in part by NSF grant IIS-0513628 and the SWARMS MURI grant. The authors also thank the staff of the USDA-ARS Jornada Experimental Range, especially Mr. Roy Libeau, biological animal sciences technician Jornada Experimental Range and Ms. Barbara Nolen, GIS administrator Jornada Experimental Range for their dedicated commitment to the DVF^{TM} research program that provided these data.

References

Aldenderfer, M.S., Blashfield, R.K., 1984. Cluster Analysis. Sage Publications, Newbury Park, CA.

Anderson, D.M., 2001. Virtual fencing—a prescription range animal management tool for the 21st century. In: Sibbald, A., Gordon, I. (Eds.), Proceedings of the Conference Tracking Animals with GPS. Macaulay Land Use Research Institute, Aberdeen, Scotland, pp. 85–94.

Anderson, D.M., 2006. Virtual fencing—a concept into reality. In: Spatial Grazing Behaviour Workshop Proceedings, Rockhampton, Qld., CSIRO.

Anderson, D.M., Hale, C.S., 2001. Inventors; The United States of America as represented by the Secretary of Agriculture. Animal control system using global positioning and instrumental animal conditioning. U.S. Patent 6,232,880.

Anderson, D.M., Nolen, B., Fredrickson, E., Havstad, K., Hale, C.S., 2004. Representing spatially explicit directional virtual fencing DVF^{TM} data. In: The 24th Annual ESRI International User Conference Proceedings.

Arabie, P., Hubert, L.J., Soete, G.D., 1996. Clustering and Classification. World Scientific, Singapore.

Arnold, G.W., Dudzinski, M.L., 1978. Ethology of Free-ranging Domestic Animals. Elsevier Scientific, New York.

Bailey, D.W., 2004. Management strategies for optimal grazing distribution and use of arid rangelands. J. Anim. Sci. 82, E147–E153.

Bailey, D.W., 2005. Identification and creation of optimum habitat conditions for livestock. Range. Ecol. Manage. 58, 109–118.

Bailey, D.W., Gross, J.E., Laca, E.A., Rittenhouse, L.R., Coughenour, M.B., Swift, D.M., Sims, P.L., 1996. Mechanisms that result in large herbivore grazing distribution patterns. J. Range Manage. 49, 386–400.

Bailey, D.W., Kress, D.D., Anderson, D.C., Boss, D.L., Miller, E.T., 2001. Relationship between terrain use and performance of beef cows grazing foothill rangeland. J. Anim. Sci. 79, 1883–1891.

Bishop-Hurley, G.J., Swain, D.L., Anderson, D.M., Corke, P., Sikka, P., Crossman, C., 2005. Understanding interactions between autonomous animal control and temperament when cattle are subjected to virtual fencing applications. In: Horizons in Livestock Sciences Redesigning Animal Agriculture. Gold Coast, Queensland, p. 23.

Butler, Z., Corke, P., Peterson, R., Rus, D., 2006. From robots to animals: virtual fences for controlling cows. Int. J. Robot. Res. 25 (5/6).

Clark, P.E., Johnson, D.E., Kniep, M.A., Jermann, P., Huttash, B., Wood, A., Johnson, M., McGillivan, C., Titus, K., 2006. An advanced, low-cost, GPS-based animal tracking system. Range. Ecol. Manage. 59, 334–340.

Coppolillo, P.B., 2000. The landscape ecology of pastoral herding: spatial analysis of land use and livestock production in east Africa. Hum. Ecol. 28, 527–560.

Coughenour, M.B., 1991. Spatial components of plant-herbivore interactions in pastoral, ranching, and native ungulate ecosystems. J. Range Manage. 44, 530–542.

DelCurto, T., Porath, M., Parsons, C.T., Morrison, J.A., 2005. Management strategies for sustainable beef cattle grazing on forested rangelands in the pacific northwest. Range. Ecol. Manage. 58, 119–127.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley and Sons, New York.

Estevez, I., Christian, M.C., 2005. Analysis of the movement and use of space of animals in confinement: the effect of sampling effort. Appl. Anim. Behav. Sci. 97, 221–240.

Ganskopp, D., 2001. Manipulating cattle distribution with salt and water in large arid-land pastures: a GPS/GIS assessment. Appl. Anim. Behav. Sci. 73, 251–262.

Geng, W., Cosman, P., Huang, C., Schafer, W., 2003. Automated worm tracking and classification. In: Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, pp. 2063–2068.

Hafez, E.S.E., Bouissou, M.F., 1975. The behaviour of cattle. In: Hafez, E.S.E. (Ed.), The Behaviour of Domestic Animals. Williams and Wilkins, Baltimore.

Heitschmidt, R.K., Taylor Jr., C.A., 1991. Livestock production. In: Heitschmidt, R.K., Stuth, J.W. (Eds.), Grazing Management an Ecological Perspective. Timber Press, Portland, Oregon, pp. 161–177.

Holechek, J.L., Pieper, R.D., Herbel, C.H., 1998. Range Management Principles and Practices, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Hulbert, I.A.R., French, J., 2001. The accuracy of GPS for wildlife telemetry and habitat mapping. J. Appl. Ecol. 38, 869–878.

Immelmann, K., Beer, C., 1989. A Dictionary of Ethology. Harvard University Press, Cambridge, MA.

Janik, V.M., 1999. Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods. Anim. Behav. 57, 133–143.

Juang, P., Oki, H., Wang, Y., Martonosi, M., Peh, L.-S., Rubenstein, D., 2002. Energy efficient computing for wildlife tracking: design and early experiences with zebranet. In: Proceedings of the Conference on Architectural Support for Programming Languages and Operating Systems, San Jose, CA.

McCowan, B., 1995. A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (*Delphinidae*, *Tursiops truncatus*). Ethology 100, 177–193.

McQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 281–297.

Pinchak, W.E., Smith, M.A., Hart, R.H., Waggoner, J.W., 1991. Beef cattle distribution patterns on foothill range. J. Range Manage. 44, 267–275.

Roath, L.R., Krueger, W.C., 1982. Cattle grazing and behavior on a forested range. J. Range Manage. 35, 332–338.

Rutter, S.M., Champion, R.A., Penning, P.D., 1997. An automatic system to record foraging behaviour in free-ranging ruminants. Appl. Anim. Behav. Sci. 54, 185–195.

Schlecht, E., Hulsebusch, C., Mahler, F., Becker, K., 2004. The use of differentially corrected global positioning system to monitor activities of cattle at pasture. Appl. Anim. Behav. Sci. 85, 185–202.

Smith, B., 1998. Moving ’em: A Guide to Low Stress Animal Handling. The Gaziers Hui, Kamela, HI.

Squires, V., 1981. Livestock Management in the Arid Zone. Inkata Press, Melbourne.

Strauss, R.E., 2001. Cluster analysis and the identification of aggregations. Anim. Behav. 61, 481–488.

Turner, L.W., Udal, M.C., Larson, B.T., Shearer, S.A., 2000. Monitoring cattle behavior and pasture use with GPS and GIS. Can. J. Anim. Sci. 80, 405–413.

Ungar, E.D., Henkin, Z., Gutman, M., Dolev, A., Genizi, A., Ganskopp, D., 2005. Inference of animal activity from GPS collar data on free-ranging cattle. Range. Ecol. Manage. 58, 256–266.

Wagnon, K.A., 1963. Behavior of beef cows on a California range. Tech. Rep. 799, California Agricultural Experiment Station Bulletin.