# chapter 2 - RF model

Dajun Wang

11/9/2019

## 1 Introduction

## 2 Material and methods

### 2.1 Random forest model construction

Random Forests (RF) is a relatively novel and powerful machine learning algorithm that has been reported to work well with complex ecological data that cannot be easily fitted with traditional methods such as generalized linear models (**???**). Through the use of tri-axial accelerometery measurements (i.e., acceleration values), RF models can also be used to predict unobservable behaviours in wild, free-rangubg animals based on information collected from examining captive animals. To do this, the RF model must first be trained to recognize and predict labelled behaviours in a 'training' dataset with decision trees constructed with suitable predicor variables (Table 1) used for classifying and predicting behaviours (Table 2).

For the classification and labelling of known behaviours in accelerometery data, I collared 11 healthy adult dogs of three different dog breeds: German shepherds (n = 9, six males and three females, age range: 1–8 years, mean age: 5 years old, Rottweiler (n = 1, male, age: 8 years old) and Golden Retriever (n = 1, male, age: 1 year old) with an accelerometer-equipped wildlife tracking collar (Type 1C-heavy, E-obs GmbH; Grünwald, Germany). The wildlife tracking collar used in this chapter is also used for collaring and tracking free-roaming dogs in

the subsequent chapters. All eleven dogs were well-trained individuals that could perform the selected repertoire of movement behaviours (Table 2) solely with verbal instructions from their trainer while being off-leash.

Table 2: Description of the behavioural tasks performed by kennel club-trained domestic dogs.

| Behaviour | Description |
|---|---|
| Stand | Individual under study is standing on all four limbs with no visible locomotion. Resting posture of the dog's head will vary between individuals. |
| Lie | Individuals under study is lying prone on the ground with all limps extended out. For some individuals, the dog may be lying fully on its side. In all recordings, there was no visible locomotion. |
| Sit | Individual under study is sitting on its haunches with no visible locomotion. Similar to Stand, head posture may vary per individual but body posture remains relatively similar. |
| Eat | Individual under studying is eating or drinking from a bowl whilst standing. No visible locomotion but the posture of the head is angled approximately 45° downwards. |

| Behaviour | Description |
|---|---|
| Walk | Individual under study is walking freely without leash at the same speed of the handler. |
| Forage | Similar to Walk but the position of the dog's head is angled downwards as it sniffs for hidden treats on the ground. |
| Run | Individual under study is sprinting after a tossed ball. Task is considered completed once the tossed ball has been received by the dog. |

For the purpose of collecting data to train the RF models (i.e., training dataset), the wildlife tracking collar was programmed to sample continuously at 10 hz, and the collar was mounted such that the x-, y-, and z- axes of the tri-axial acclerometer were parallel to the median (surge), the dorsal (sway), and the dorsal (heave) planes of the animal, respectively. The design and alignment of the collar provided the tri-axial acceleromenter the capacity to measure the surge (back and forth movement on the x-axis), sway (left and right motion on the y-axis) and heave (up and down on the z-axis) motion of the animal. Data was collected in October 2016 and May 2019 in a grassy field to a) simulate the environment where the intended free-roaming dogs for study are typically found, and b) provide sufficient ground for the dogs to move continuously. The terrain selected for this component was relatively flat to avoid compromising the gravitional forces acting on the heave axis.

All behavioural tasks (Table 2) were performed continuously for at least 2 min in a stipulated sequence, except for Forage, Eat and Run as these tasks are either too energetically taxing or behaviourally inconsistent. Each dog was tasked to perform the entire repertoire of behaviours at least twice (with consent from their trainer), and approximately 300 minutes of

video-recorded behavioural observations were obtained. The collected accelerometery data was viewed with an acceleration viewer (http://www.movebank.org, version 33) and the time-stamps of the collected accelerometer measurements were synchronized and labelled manually to the time-stamps of the video recordings of each performed behavioural task. This complete labelled dataset was subsequently used in the preparation and construction of the predictive RF model.

In preparation of the training dataset, the labelled acceleration measurements were binned into two-second windows (or 'bursts') to accomodate at least two full strides for motion-based dog behaviours (i.e., walking, foraging) without influence from un-intentional behavioural transitions. Subsequently, series of summary statistics (mean, min, max, kurtosis, skewness, range, standard deviation; Table 1) for all three axes were applied onto each labelled burst of accelerometer measurements to characterize the accelerometery measurements of the different behavioural tasks and describe the predictor variables used in the RF model.

Table 1: Predictor variables and their statistic labels used for predicting dog behavioural tasks in the random forest models.

| Statistics label | Predictor variables | Description |
|---|---|---|
| Mean | mean.x mean.y mean.z | The calculated mean of the acceleration measurements within each burst for axes x, y and z. |
| Min | min.x min.y min.z | The smallest acceleration measurement within each burst for axes x, y and z. |
| Max | max.x max.y max.z | The largest acceleration measurement within each burst for axes x, y and z. |

| Statistics label | Predictor variables | Description |
|---|---|---|
| Kurtosis | kurt.x kurt.y kurt.z | The relative flatness of the acceleration measurements within each burst for axes x, y and z. |
| Skew | skew.x skew.y skew.z | The relative skewness of the acceleration measurements within each burst for axes x, y and z. |
| Range | range.x range.y range.z | The calculated difference between the largest and smallest acceleration measurements of axes x, y, and z. |
| Standard deviation | sd.x sd.y sd.z | The calculated standard deviations of the acceleration measurements of axes x, y, and z. |

The `randomForest` package (Liaw and Wiener 2002) in `R` was then used to prepare and construct the RF model used for predicting dog behaviours. The labelled dataset was subsampled at a 80% proportion as a training dataset to prepare the RF model, while the remaining 20% was used as a testing and validation dataset. The default number of trees ($n$ = 500) were used in the RF model construction as the Out-of-Bag (OOB) error estimates, a method of measuring prediction error rates, was found to be consistently stable (Figure 1). To refine and improve the RF model's predictive capacity, I cross validated the RF models by comparing the prediction accuracies of RF models constructed ($n$ = 21; Figure 3) with a range of mTry values mirroring the number of described predictor variables (Table 1).

Following which, the best predictive RF model was constructed and the classification error rate and mean prediction accuracy of each behavioural task in the model were tallied and tabled (Figure 3).

## 2.2 Accelerometer threshold accelerometer-informed GPS sampling

The wildlife tracking collar is equipped with the capacity to sample GPS relocations dynamically based on the accelerometer measurements collected in real-time. With the use of a selected accelerometer threshold (ACT), reseachers are able to selectively intensify the sampling of GPS re-locations only when the animal is moving (hence higher variance in accelerometer readings). This feature not only reduces the energy and memory consumption during the research period, it also allows the researcher to identify movement paths or patterns that are ecologically interesting (e.g., predation or harrassement events).

For this function to work realistically, I used the above constructed RF model to identify the variances of the predicted movement behaviours and selected a variance value (Table 4) for the axis that best represents movement or motion in a dog (i.e., Z axis; the acceleration in the heave axis typified by the rolling gait in dogs). Following which, I ground truth the selected variance by walking four dogs (Mongrel breeds, 4 females, age range 2-7 years old) with the wildlife tracking collar equipped. All dogs were walked continuously for approximately ten minutes, and with the studied dog resting (with the collar on) only in the beginning and end of each walk. The dogs were rested in a non-motion movement behaviour (e.g., lying, sitting or standing) to simulate the resting behaviour of free-roaming dogs in the wild. The GPS and ACC dataset were then retrieved from the collar, and the occurence of quick- and long-burst gps relocations were plotted sequentially against the variance of the accelerometer measurements collected from all three axes (Figure 5).

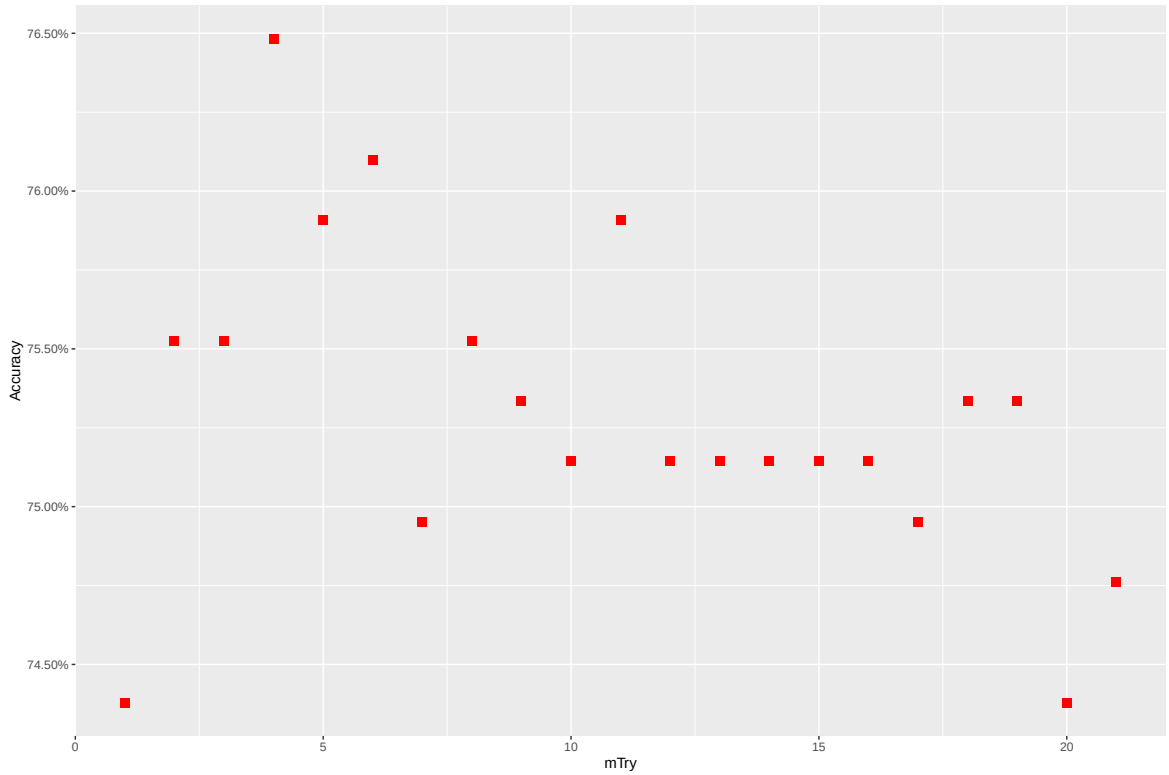



Figure 3: The prediction accuracy of the RF model.

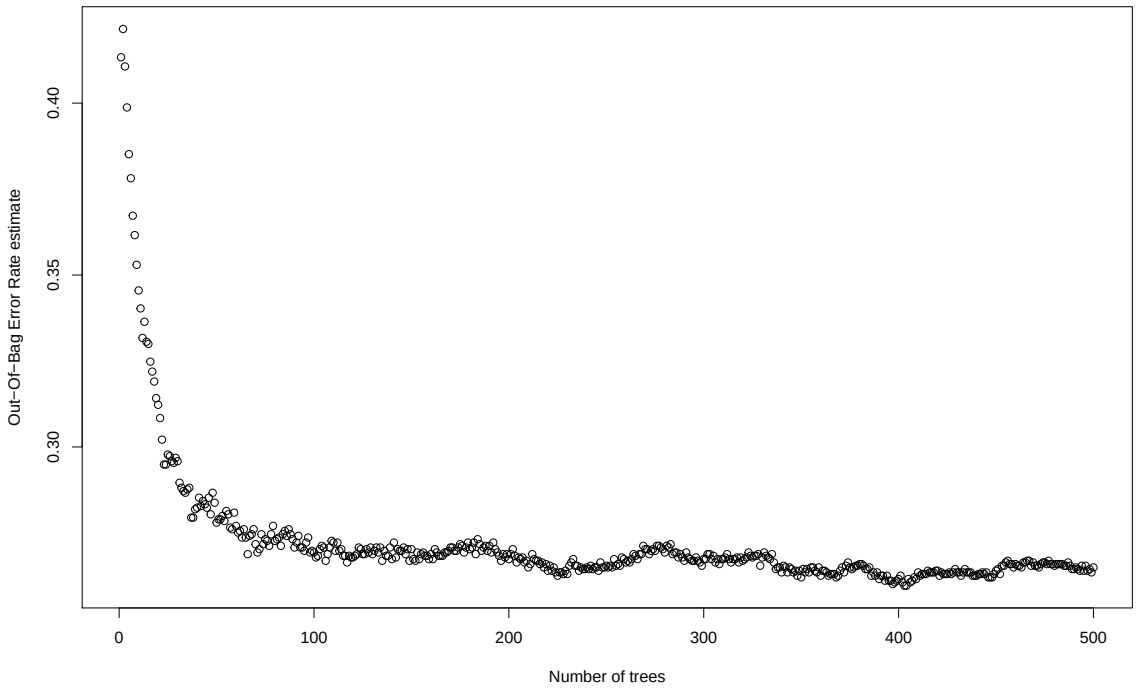
Figure 4: The importance of each predictor variable in the RF model.

Table 3: The prediction accuracy of the RF model.

```
##
## ================================================
##         eat forage lie run sit stand walk accuracy
## ------------------------------------------------
## eat     34     6    13   0   7     2    4    51.50%
## forage   1    66     6   5   3     1   44    52.40%
## lie      4    10   302   1  67    25   25    69.60%
## run      0     2     1 353   0     0   52    86.50%
## sit      3    11    87   0 231    14   10    64.90%
## stand    2     4    23   2  20   113   17    62.40%
```

```
## walk    0    9     4  54   7    3   424   84.60%
## ------------------------------------------------
```

Figure 1: The Out-Of-Bag error estimates for the predictive RF model plotted gainst the number of trees ($n$) used.




both models showed relatively consistent prediction accuracy rates with 500 trees.

```
##
## ================================================
##        eat forage lie run sit stand walk accuracy
## ------------------------------------------------
## eat     4    2     8   0   1    0    2    23.50%
## forage  0    15    2   4   1    1    9    46.90%
## lie     1    1    79   1  16    7    4    72.50%
## run     0    0     0  89   0    0    14   86.40%
```

```
## sit     0    0    10   2  68    3    7    75.60%
## stand   0    0     4   0   6   28    8    60.90%
## walk    0    5     1   5   2    1  112    88.90%
## ----------------------------------------------
```

As with most RF models, overfitting does occur when the RF model is tasked to predict on the subsampled testing dataset. The most stark differences lie in the prediction of the 'Eat' behaviour where the RF model's prediction accuracy fell from 51.5% to 23.5%.

The ROC curve is a curve of probability, and a AUC near to 1 (higher AUC) is a good measure of separability. The RF model appears to be relatively sensitive to predicting the dogs' movement behaviours, with some distinction defined between motion and non-motion behaviours. For example, the 'Run' has the highest AUC followed by 'Walk' whereas non-movement based behaviours such as 'Sit'.

The model has some difficulty telling apart

the mean prediction accuracy of the raw model was 75.53%. looking at the prediction accuracy of the individual behaviours, the RA model predicted poorly for the behaviours 'eat' and 'forage' and middingly for 'stand'. 'Eat' was most often mis-classifed All other behaviours were predicted relatively well.

Table 4: The mean variance of acceleromenter measurements for predicted movement behaviours in dogs.

```
##
## ============================================================
##   behaviour      mean.X           mean.Y           mean.Z
## ------------------------------------------------------------
## 1     2     4382.1600553725  5236.04653271207 3097.25599648845
## 2     4     5966.96777758207 3639.99184859353 7489.65092405363
## 3     5     10536.6964273547 15618.6782670397 7963.08467735117
```

```
## 4      1      26128.2932034003 23022.5935070752 22260.3063053903
## 5      6      56863.5952330838 34096.0773644017 42351.5690677523
## 6      3      335325.958796157 152629.014611539 326456.066690359
## ------------------------------------------------------------
```

Figure 5: The sampling of GPS re-locations based with an accelerometer threshold of 10,000 in the Z axis. The red, blue and green lines represents the X, Y and Z axis.

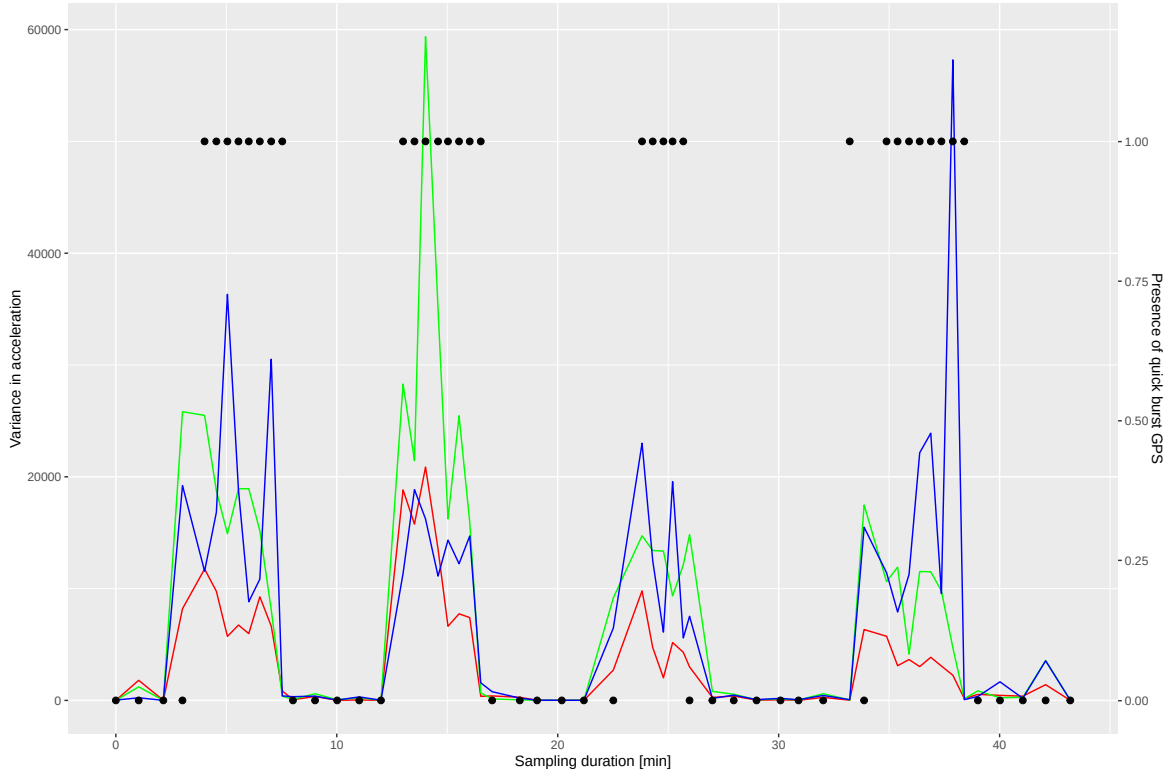



Figure 5: The sampling of GPS re-locations based with an accelerometer threshold of 10,000 in the Z axis. The red, blue and green lines represents the X, Y and Z axis.

# 3 Results

# 4 Discussion

# 5 Conclusion

## References

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 6.