Reasoning about Neural Network Activations: An Application in Spatial Animal Behaviour from Camera Trap Classifications*

Benjamin C. Evans[1*][0000−0001−5342−5895], Allan Tucker[1][0000−0001−5105−3506], Oliver R. Wearn[2][0000−0001−8258−3534], and Chris Carbone[2][0000−0002−9253−3765]

[1] Brunel University London, Uxbridge UB8 3PH, UK.
{Benjamin.Evans,Allan.Tucker}@brunel.ac.uk
[2] Institute of Zoology, Zoological Society of London, London, NW1 4RY, UK
Oliver.Wearn@gmail.com, Chris.Carbone@ioz.ac.uk

Abstract. Camera traps are a vital tool for ecologists to enable them to monitor wildlife over large areas in order to determine population changes, habitat, and behaviour. As a result, camera-trap datasets are rapidly growing in size. Recent advancements in Artificial Neural Networks (ANN) have emerged in image recognition and detection tasks which are now being applied to automate camera-trap labelling. An ANN designed for species detection will output a set of activations, representing the observation of a particular species (an individual class) at a particular location and time and are often used as a way to calculate population sizes in different regions. Here we go one step further and explore how we can combine ANNs with probabilistic graphical models to reason about animal behaviour using the ANN outputs over different geographical locations. By using the output activations from ANNs as data along with the trap's associated spatial coordinates, we build spatial Bayesian networks to explore species behaviours (how they move and distribute themselves) and interactions (how they distribute in relation to other species). This combination of probabilistic reasoning and deep learning offers many advantages for large camera trap projects as well as potential for other remote sensing datasets that require automated labelling.

Keywords: Animal Behavior · Convolutional Neural Networks · Bayesian Networks · Activation Based Reasoning.

1 Introduction

Artificial Neural Networks (ANNs), and in particular, Convolutional Neural Networks (CNNs) have superseded traditional statistical methods in a multitude of domains, ranging from speech recognition and synthesis [6], natural language processing [2] as well as image processing, detection and recognition tasks [1].

* Benjamin C. Evans work is funded by NERC (The Natural Environment Research Council)

Recently there has been growing interest by ecologists into the use of machine learning to assist with the growing task of labelling camera trap data [4]. A camera trap consists of an imaging device with an automatic trigger. These triggers are often in the form of a passive infrared sensor or timer which starts off the capture of a series of images, from here on in referred to as a sequence. Ecologists have increasingly used camera trap surveys to monitor and investigate species populations and animal behaviour without the need for physical capture of the animals [7]. With the decrease in the cost of camera technology and storage hardware, we've seen a rapid increase in size and extent of these datasets.

Traditionally an individual researcher, or their team, would process through each image taken during the survey, labelling each with the species seen and any behaviour that may be of interest to the study. As the datasets grew, researchers began to enlist citizen scientists to assist with the endeavour. This, however, still requires vast amounts of human time [4]. Which in turn has led ecologists to investigate alternative means in labelling, in particular, the use of machine learning to assist the endeavour in reducing human dependency and potentially speeding up the process of labelling. Initial work in species classification has utilised CNNs by feeding whole images into the model with individual neuron outputs representing each species occurring in a given image. This has shown promising results but has also identified certain issues with CNNs that are particularly apparent with camera trap data.

One major issue is that of overfitting. Given that each camera trap placed is usually left in the same location for the duration of a study, the background of all images from a single camera is the same. When there is a high density of a single species at one location, CNNs tend to focus on the background rather than the features of the animal within the image. To overcome this issue, Beery et al. have developed an object-detection model called MegaDetector which identifies 'animal', 'human' and 'vehicle' classes along with their predicted bounding box within the given image [1]. These detections can then be cropped in accordance to the predicted bounding box and passed to a species classifier for finer grain prediction, reducing the ability of the species classifier to overfit to the background of the images.

While there is still research to be done on fine-tuning the automation of the labelling process, there is a potential for reasoning about the labelling of images given the uncertainty of the classification to better understand the behaviour of different species. Thus, we propose a framework for reasoning about species behaviour, based on the activations of a CNN trained to identify species in different geographical locations.

Bayesian Networks (BNs) have been successfully used in many areas of research where data is characterised by uncertainty, including in ecology ([3], [11], [5]) . There has also been some work exploring the use of spatial Bayesian networks to analyse the movement of species [10]. In this paper, we investigate combining ANNs with BNs so that rather than reasoning about observations made by humans, we can automate the entire process of understanding ani-

mal behaviour starting from the image and ending with predictive models that explain complex spatial behaviours.

In the next section, we will firstly describe the BorneoCam dataset that we focus on for this study. Secondly, we shall present the deep-learning-based classifier for automating the labelling of images in the BorneoCam dataset. We then describe how we use these labels (in the form of ANN activations as data) to learn a Bayesian Network allowing us to reason and predict about the species sightings across a number of cameras spread over different geographical locations.

2 Methodology

2.1 Camera Trap Data

The BorneoCam dataset includes camera trap imagery from multiple surveys across northern Borneo which will be made publicly available at a later date. For the following experimentation, we utilise just one subset of the dataset relating to a single survey, identified as 'OG3' as shown in Figure 1. The region consists of 47 camera locations further split into North, West and East subregions each consisting of 15 to 16 cameras. The images in Figure 1 show a sample camera trap image along with a sample extraction/detection containing just the animal.

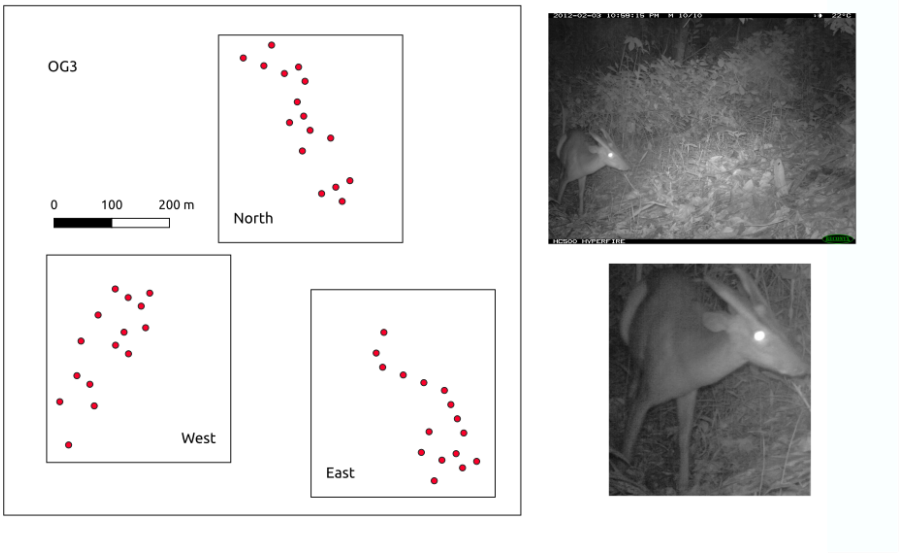

Fig. 1. OG3 Survey Site with Camera Locations from the BorneoCam Dataset, a sample camera trap image and detection extract of a Red muntjac.

2.2 Experiments

Firstly, the MegaDetector is used, an object detection model trained on a variety of datasets from around the world which identifies 'animal', 'human' and 'vehicle' classes along with their associated bounding box [1]. Through qualitative analysis, the results of the model seem reasonable enough to assume some degree of certainty that when an image has a species label and has one detection from the MegaDetector that the bounding box from the detection matches the region of image containing the species identified in the labelled set. With these one-detection, one-label images we are able to build a dataset to train a deep-learning-based classifier just on the region of an image that contains the animal limiting the risk of the classifier to overfit to the background of each location.

In order to demonstrate our approach, we train a CNN species classifier on the six species with the highest image frequencies from the dataset while merging 'Red muntjac' and 'Yellow muntjac' into a singular 'Muntjac' class. This is due to difficulties in classifying these species because of subtle differences in their visual characteristics. This presents us with 'Bearded Pig', 'Long-tailed porcupine', 'Southern pig-tailed macaque', 'Muntjac' and 'Spiny rat'.

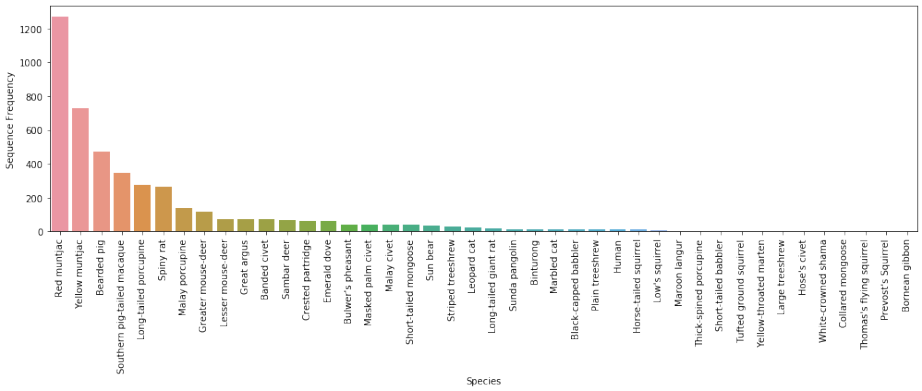

Fig. 2. Visualisation of sequence frequency by species in the OG3 region.

The survey in use from the BorneoCam dataset was set up so that each camera captures a sequence of ten images once the device has been triggered. Often it is found that the first image contains no or a small portion of the animal as the triggers viewing angle is wider than the camera's view. Thus, we take the second image of each sequence to obtain a higher likelihood of the animal in view. The resulting dataset is split based on camera location into 70% training, 5% validation and 25% testing subsets. Splitting based on camera location allows us to test that the model is able to generalise to images taken from previously unseen locations. Once the images are sorted into the relevant sets, the number

of images per species is normalised so that each class has a fair representation of samples.

A model is trained based on the ResNet50v2 architecture [12] with a classification head of one average pooling layer followed by a dropout layer of 0.2 and finally a dense layer using softmax activations. The model is trained utilising transfer learning from a model pretrained on the ImageNet dataset [8]. Categorical cross-entropy is used as the loss function and a learning rate of 1e-3 for five epochs is run with only the classification head trainable. Followed by five epochs with a learning rate of 1e-4 where layers above 150 are trainable. The ResNet50v2 architecture has been chosen based on prior experimentation as appears to be less prone to overfitting and provides a reasonable result in accuracy.

In order to build the BN from the activation data, the activations generated from the Resnet are pre-processed into intervals over the duration of the camera trap study. These intervals are derived by considering the highest activation of each predicted image as a sighting of the respective species. This is then recorded as a sighting at the interval the image capture falls within for the relevant camera location and species. The interval size can be determined based upon the idealised granularity ensuring that the interval is greater than the estimated time required for the species of interest to pass by two or more camera locations.

BNs are graphical models that encode the joint distribution of a dataset using a graphical structure to capture independent relations between variables and local conditional probability distributions. See Fig 3 for an example BN with five nodes where the probability distribution at each node is conditioned upon its parents.

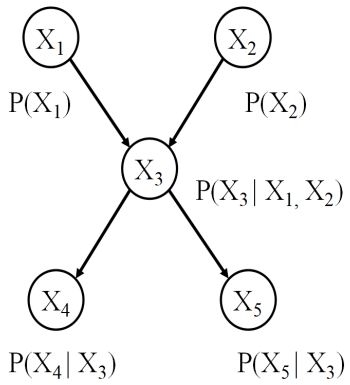

Fig. 3. A Bayesian Network with 5 nodes.

BNs can be inferred from data using score and search methods such as hill climbing with the log-likelihood metric, or constraint-based methods such as the

PC algorithm [9]. In this paper, these approaches were both explored to infer BN models from a subset of the processed activation data (70% as training data). Here, each BN node will represent one species at a particular spatial coordinate. The resulting structures were explored for spatial features and specific species interactions. The remaining 30% of activation data was used to test the BNs as predictive models. Prediction was then conducted on a location-by-location basis. It involved using inference where evidence was entered into the BN model based upon the activations of all surrounding locations and the BN model was used to predict the presence of species in each test location (in the form of a posterior probability distribution over all species).

Figure 4 presents an overview process diagram visualising the steps taken in the methodology.

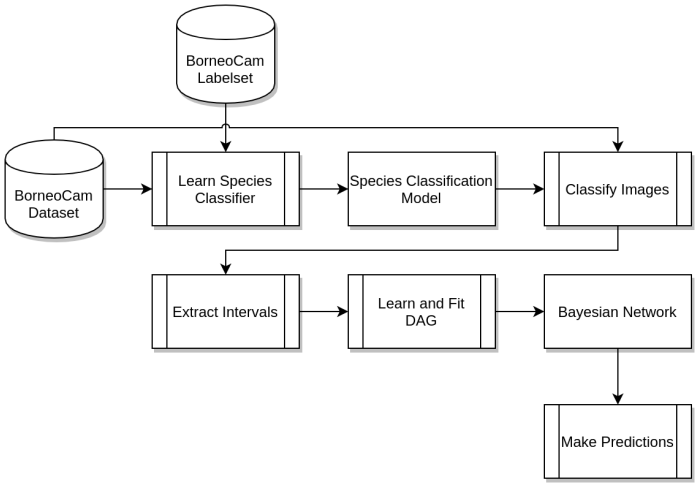

Fig. 4. Process diagram providing an overview of the methodology.

3 Results

First, we look at the ability of the CNN to automatically identify species from the BorneoCam data. Figure 5 shows the learning of the model with 5 epochs training the classification head and the last 5 epochs fine-tuning with layers 150+ trainable. Figure 6 shows the confusion matrices, visualising the difference between the labels and that of the predictions from the ResNet model. We can see that on the whole, the model performs well with relatively high numbers of correct classification (frequencies on the diagonal). The classes that are most confused are that of the long-tailed porcupine and the spiny rat. This could be

because both are relatively small, have a long tail and similar body shape. The long-tailed porcupine has a distinctive "brush" on the end of the tail, but this can be missing or difficult to see, so it's unlikely the ANN has learnt this feature.

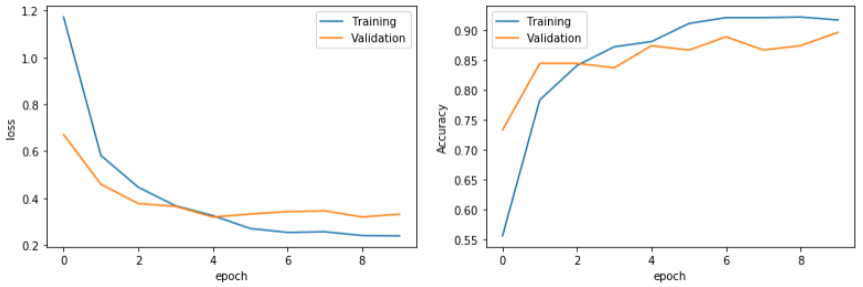

Fig. 5. Training/validation loss and accuracy by epoch.

Fig. 6. Confusion matrix of the Species Classifier.

Now we explore the use of the output activations as data that we want to reason about. To do this, interval pre-processing was used as explained in the methods section. Figure 7 shows the images taken within the OG3 region overtime giving us an overview of the total time period. The size of each circle represents the number of images taken for that location at a particular time. Notice that for a small number of cameras, there is potential malfunction, empty battery or full memory as the images stop recording after a fixed time. These cameras with missing data may need to be removed from further analysis, though if there are enough arcs learnt towards the nodes in the network where we are missing data it may be possible to predict the missing sightings.

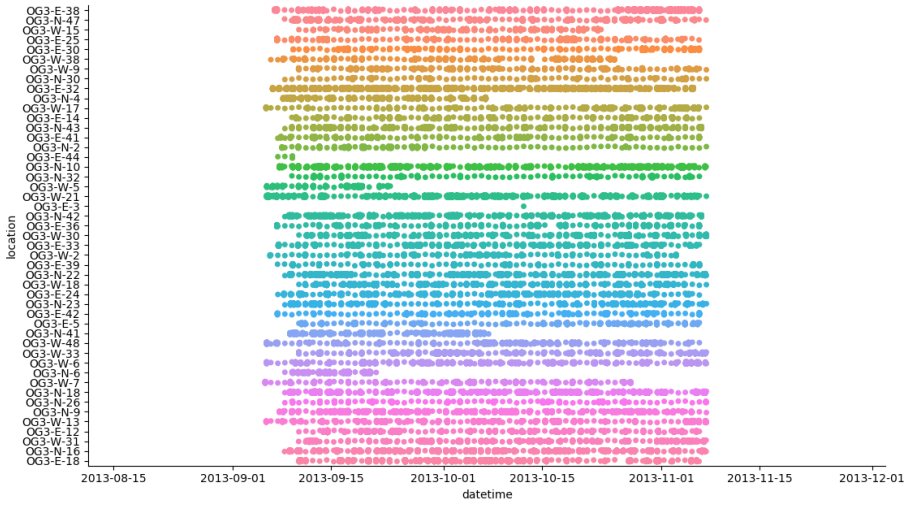

Fig. 7. Visualisation of images taken at each camera location over time. Colour relates to the individual cameras / rows.

Now we turn to the Bayesian network analysis. Table 1 shows the accuracy for each species and each location of a network learnt on the west subregion when tested on the test dataset. It is clear that some species/locations are more easily predicted than others which may infer a stronger relation, be it a higher number of arcs directed to the relevant nodes in the BN or that there is a strong pattern identified in the learning and fitting.

BNs are learnt for individual species and one network incorporating all species as seen in Figure 8. We can see that when incorporating additional species into the model that links are identified both between the same species at different locations but also between differing species. This in turn provides further information for making predictions and presents avenues of exploration into the underlying variables behind the inter-species behaviour patterns.

Species	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
Bearded Pig	0.934	0.989	0.204	0.942	0.989	0.965
Long Tailed Porcupine	0.906	0.986	0.15	0.917	0.986	0.95
Muntjac	0.821	0.965	0.262	0.835	0.965	0.895
Southern Pig Tailed Macaque	0.97	0.989	0.45	0.98	0.989	0.985
Spiny Rat	0.915	0.967	0.349	0.941	0.967	0.954

Table 1. OG3 West BN Prediction Statistics

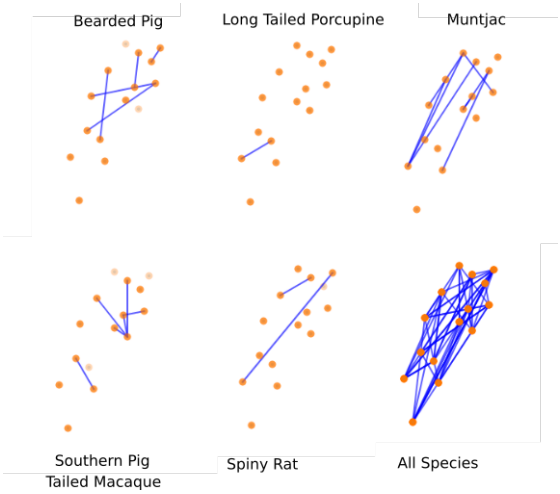

Fig. 8. BNs for each species (Bearded Pig, Long Tailed Porcupine, Muntjac, Southern Pig Tailed Macaque, Spiny Rat) and all species in OG3 West.

Table 2 and Figure 9 show the resulting BN structure over all regions for all species. It is clear that there are predictive relationships that span greater distances than the local ones discovered in Figure 8. These long distance relationships seem to involve all the species which may imply that there are regular activity patterns across the subregions where species are seen or not seen within the same day. This opens up further exploration into the factors at play, such as the time of day or weather conditions impacting behaviour.

Figure 10 shows a detail of the overall network structure that combines multiple species and locations whilst Table 2 summarises all inter-species links. Notice how some species interact far less than others (eg. spiny rat only seems to have some loose relationship to Muntjac, and no relationship to spiny rats in other locations), whilst other species interact a great deal with themselves across locations (eg. SPT Macaques indicating one or more of the species moving between locations) and with other species (SPT Macaques with Bearded Pigs indicating some avoidance or following behaviour).

We can also see in Figure 10 that the relationships being made can be reasoned about in terms of both species and location. In the zoomed in portion

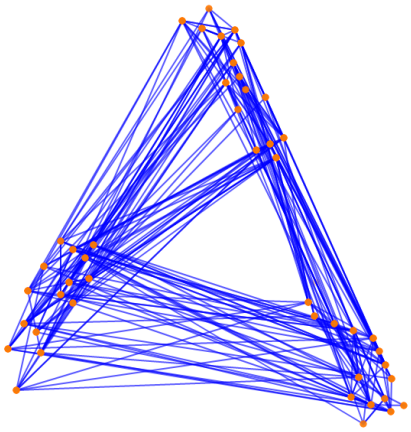

Fig. 9. The discovered BN across all 3 regions for all species.

of the network, the center-most Bearded Pig node has the highest frequency of relationships to nodes that are of the same species and the same (East) subregion, highlighting that the BN has learnt what appears to be logical relationships (likely to be the same animals as they move around a localised area).

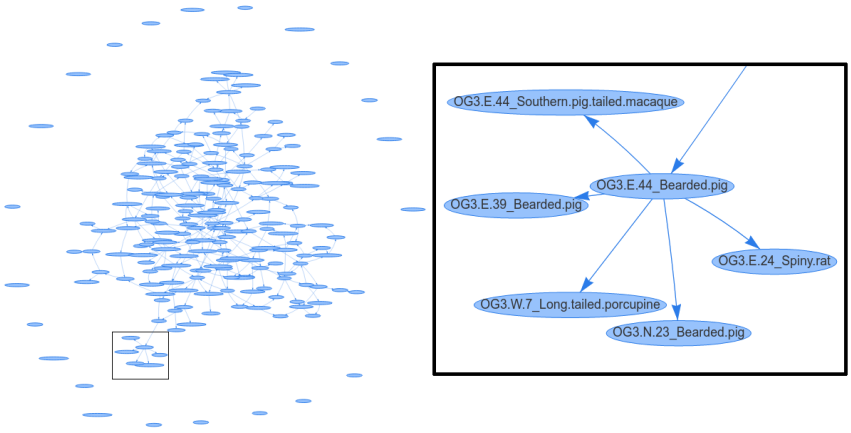

Fig. 10. Species Interaction Network (and detailed section as discussed).

	Bearded Pig	Long Tailed Porcupine	Muntjac	Southern Pig Tailed Macaque	Spiny Rat
Bearded Pig	14	10	9	5	10
Long Tailed Porcupine	4	4	16	10	7
Muntjac	11	15	18	11	8
Southern Pig Tailed Macaque	12	3	12	18	4
Spiny Rat	10	9	9	6	7

Table 2. Summary of BN arcs for OG3 with multiple species showing the total number of arcs between species with "from" on the Y axis and "to" on the X axis.

4 Conclusions

In this paper, we have explored a framework for reasoning about images that have been analysed by deep learners. We have applied a combination of deep learning (for image classification) and Bayesian networks (for spatial reasoning) to camera trap data, used by ecologists to better understand animal populations and behaviour. We have shown, using trap data from BorneoCam, that by treating deep learner label outputs (activations) as data and by combining them with time and spatial coordinates, we can build probabilistic models that can identify and predict specific species' behaviours and interactions. Whilst the deep learning classifier achieved accuracies of 0.92 for labelling species, the Bayesian network achieved accuracies of 0.97 for predicting whether a species would be present at a specific location given other information about nearby locations. This work is still in the early stages and there is a great deal of work that will be followed up including the incorporation of deep learning activations as measures of certainty in a species being present, the exploration of other datasets and the integration of geographical and climate features such as rivers and weather into the models.

References

1. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review. arXiv:1907.06772 [cs.CV] (2019), https://arxiv.org/abs/1907.06772
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North (2019). https://doi.org/10.18653/v1/n19-1423, http://dx.doi.org/10.18653/v1/n19-1423
3. Franco, C., Hepburn, L.A., Smith, D.J., Nimrod, S., Tucker, A.: A bayesian belief network to assess rate of changes in coral reef ecosystems. Environmental Modelling & Software **80**, 132 – 142 (2016). https://doi.org/https://doi.org/10.1016/j.envsoft.2016.02.029, http://www.sciencedirect.com/science/article/pii/S1364815216300494

4. Glover-Kapfer, P., Soto-Navarro, C.A., Wearn, O.R.: Camera-trapping version 3.0: current constraints and future priorities for development. Remote Sensing in Ecology and Conservation **5**(3), 209–223 (2019). https://doi.org/10.1002/rse2.106, https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.106
5. Maldonado, A., Uusitalo, L., Tucker, A., Blenckner, T., Aguilera, P., Salmerón, A.: Prediction of a complex system with few data: Evaluation of the effect of model structure and amount of data with dynamic bayesian network models. Environmental Modelling & Software **118**, 281 – 297 (2019). https://doi.org/https://doi.org/10.1016/j.envsoft.2019.04.011, http://www.sciencedirect.com/science/article/pii/S1364815218310338
6. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv:1609.03499 [cs.SD] (2016), https://arxiv.org/abs/1609.03499
7. Rowcliffe, J.M., Carbone, C.: Surveys using camera traps: are we looking to a brighter future? Animal Conservation **11**(3), 185–186 (2008). https://doi.org/10.1111/j.1469-1795.2008.00180.x, https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1795.2008.00180.x
8. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, prediction, and search (1993)
10. Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., Tucker, A.: Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. Ecological Informatics **30**, 142 – 158 (2015). https://doi.org/https://doi.org/10.1016/j.ecoinf.2015.10.003, http://www.sciencedirect.com/science/article/pii/S1574954115001648
11. Uusitalo, L.: Advantages and challenges of bayesian networks in environmental modelling. Ecological Modelling **203**(3), 312 – 318 (2007). https://doi.org/https://doi.org/10.1016/j.ecolmodel.2006.11.033, http://www.sciencedirect.com/science/article/pii/S0304380006006089
12. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017). https://doi.org/10.1109/cvpr.2017.634, http://dx.doi.org/10.1109/CVPR.2017.634