UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

AUTOMATIC IDENTIFICATION OF BAT SPECIES USING SUPERVISED LEARNING

Nuno Filipe Ladeira Costa Cláudio

Mestrado em Bioestatística

Dissertação orientada por:
Doutor Tiago André Marques
Doutora Maria Helena Mourino Silva Nunes

2017

Resumo

Recentemente, face à necessidade de encontrar alternativas aos combustíveis fósseis, assistiu-se a um aumento do investimento em energias renováveis, com destaque para a eólica. Contudo, a instalação de parques eólicos não está livre de impactos ambientais negativos, como a mortalidade direta de aves e quirópteros (i.e. morcegos). Os quirópteros desempenham um papel ecológico fundamental, sendo importantes polinizadores e dispersores de sementes de várias espécies vegetais de elevado valor ecológico e económico. Estes mamíferos têm igualmente um papel essencial no controlo de pragas, uma vez que se alimentam essencialmente de insetos (Basil, Vanitharani, & K, 2014; Jones, Jacobs, Kunz, Willig, & Racey, 2009). Dada a sua relevância ecológica, e por ocuparem uma ampla variedade de nichos ecológicos, os morcegos podem ser usados como bioindicadores (Jones et al., 2009). Por conseguinte, a sua conservação é da maior importância, pelo que urge avaliar a sua presença e distribuição. No entanto, estes estudos podem ser de difícil execução devido à pequena dimensão dos indivíduos, ao seu voo rápido e hábitos noturnos, e ainda, à existência de um elevado número de espécies, a maioria com elevadas semelhanças fisionómicas. Consequentemente, os programas de monitorização dependem muito da observação ou captura de indivíduos, sendo muitas vezes de difícil aplicação (Ochoa, O'Farrell, & Miller, 2000).

Os quirópteros dependem da ecolocalização para se orientar no espaço, nomeadamente na procura de alimento, emitindo continuamente vocalizações e interpretando os ecos (Schnitzler, Moss, & Denzinger, 2003). As vocalizações de ecolocalização podem ser usadas para efetuar a identificação de espécies, estratégia que, para além de ajudar a monitorar e avaliar o estado de conservação de uma população, está também em conformidade com as regulamentações ambientais (Ahlen & Baagøe, 1999). A identificação acústica assume-se, assim, como fundamental para avaliar o impacto dos parques eólicos nas populações de morcegos. A execução de trabalhos de monotorização em parques eólicos é crucial mas onerosa pois engloba grandes áreas, podendo tornar-se muito exigente em termos de força de trabalho e custos associados. Contudo, esta tarefa pode ser alivada através da utilização de detetores automáticos de ultrassons. Atualmente, estes aparelhos permitem o armazenamento de grandes quantidades de dados relativos à atividade dos morcegos (Jennings, Parsons, & Pocock, 2008). Como resultado, temos acesso a uma grande quantidade de informação que, no entanto, pode ser extremamente difícil para um (ser) humano interpretar e dar sentido.

A necessidade de agilizar o processo de identificação de espécies de morcegos levou-nos a desenvolver uma metodologia de identificação de espécies de morcegos da África do Sul, com recurso a metodologias de *machine learning*. A etapa seguinte passa, assim, por analisar os dados recolhidos e identificar as espécies presentes nas gravações. Para tal, é necessário introduzir métodos automáticos que permitam encontrar padrões nos dados, interpretá-los e tirar conclusões. Algoritmos de *machine learning* podem ser usados para executar esse tipo de análise. O *machine learning* é uma área que, *lato sensu*, cruza a programação computacional com a estatística e cujo objetivo é elaborar sistemas de

aprendizagem automática. Actualmente, o *machine learning* está em toda a parte e começa a ser um ponto central das nossas vidas (Domingos, 2015). De facto, permite-nos trabalhar problemas complexos de predição, como reconhecimento de voz, imagem, predição de séries temporais não-lineares e previsão em mercados financeiros, entre outros (Domingos, 2015).

Os algoritmos de *machine learning* são treinados através de um processo iterativo de *feeback* positivo e negativo. Em oposição à modelação estatística tradicional, o *machine learning* não efetua suposições sobre a distribuição subjacente dos dados, que é tratada como desconhecida (Breiman, 2001). Os métodos de *machine learning* dividem-se em duas categorias principais, aprendizagem supervisionada e não supervisionada. Na aprendizagem supervisionada, a variável dependente está associada a vetores de características que a descrevem. Os algoritmos de aprendizagem supervisionada "aprendem" as características da variável dependente num processo denominado treino. Um algoritmo de aprendizagem supervisionada tenta otimizar uma função (o modelo) para encontrar a combinação de características que ajudam a definir o valor da variável dependente. Os modelos resultantes poderão ser usados para prever novas observações. O tipo de predição feita varia de acordo com a natureza da variável dependente, a designar por y. Se y for uma variável categórica temos um problema de classificação, o qual implica identificar o grupo/classe a que pertence uma determinada observação. Caso y seja contínua, estamos perante um modelo de regressão que envolve estimar ou prever uma resposta. Por sua vez, na aprendizagem não supervisionada, a classe de cada observação é desconhecida e os algoritmos precisam de reconhecer padrões e encontrar grupos com características comuns. Estes métodos tendem, tipicamente, a ser utilizados numa abordagem mais exploratória.

A identificação acústica automática tem sido usada na identificação de uma ampla gama de espécies animais, por exemplo, aves (Peake & McGregor, 2001), mamíferos marinhos (Mellinger & Clark, 2000; Yack, Barlow, Rankin, & Gillespie, 2009) e insetos (Mankin, 2011). A identificação de espécies de morcegos a partir das vocalizações de ecolocalização tem sido feita através de diferentes abordagens, incluindo análise estatística multivariada (Vaughan, Jones, & Harris, 1997; Papadatou, Butlin, & Altringham, 2008), modelos de Markov (Skowronski & Harris, 2006), redes neurais (Parsons, Boonman, & Obrist, 2000), *support vector machine* (Redgwell, Szewczak, Jones, & Parsons, 2009) e *random forests* (Armitage & Ober, 2010a).

A Bioinsight, uma empresa de consultadoria ambiental, lançou o desafio de desenvolver um procedimento expedito de identificação de morcegos da África do Sul, no contexto dos parques eólicos. Neste trabalho, começámos por organizar uma base de dados de pulsos de ecolocalização dividida em conjunto de treino e de teste. O conjunto de treino foi utilizado para treinar modelos classificatórios, individuais para as espécies: *Chaerephon pumilus*, *Eptesicus hottentotus*, *Miniopterus natalensis*, *Neoromicia capensis*, *Sauromys petrophilus* e *Tadarida aegyptiaca*. Os modelos foram treinados com recurso a algoritmos de *machine learning*, nomeadamente, *random forest*, *support vector machine*, *eXtreme Gradient Boosting* e análise discriminante, com recurso a validação cruzada de modo a otimizar os parâmetros dos modelos. Seguidamente, foi feita a avaliação do poder preditivo dos modelos com recurso ao conjunto de teste e escolhidos os melhores modelos para cada espécie, tendo como base o poder preditivo, o equilibrio entre sensibilidade e especificidade, assim como a origem dos falsos positivos.

O objetivo deste trabalho consistiu em estabelecer uma sequência de modelos para identificar cada uma das espécies de interesse, com base nos resultados obtidos no conjunto de teste. A sequência determinada para a aplicação dos modelos estabelecida foi a seguinte:

1. *Mnat* RF
2. *Ehot* FDA
3. *Ncap* FDA
4. *Cpum* FDA
5. *Taeg* RF
6. *Spet* FDA

Esta sequência permitiu obter bons resultados no conjunto de teste, classificando corretamente as espécies em 95% das gravações. Foi igualmente elaborado um script em R que aplica os diversos modelos e fornece ao utilizador as espécies presentes em cada uma das gravações analisadas, gerando um *output* simples e informativo que pode ser editado pelo utilizador. Este sistema foi aplicado a um conjunto de 216 gravações proveniente de um projeto ativo da Bioinsight, tendo identificado corretamente 92% das gravações e indicando ser capaz de generalizar com sucesso para novos dados.

O trabalho aqui apresentado é, no nosso melhor conhecimento, o primeiro estudo que utiliza apredizagem supervisionada para identificar espécies de morcegos da África do Sul a partir de pulsos de ecolocalização, permitindo não só identificar espécies de interesse no contexto da monotorização de parques eólicos na África do Sul, mas fazê-lo de forma rápida e sistemática, em comparação com a identificação manual. Com recurso ao *script* elaborado é possível processar $\tilde{3}5$ 000 gravações em apenas 5 minutos. Num futuro próximo, este trabalho permitirá acelerar a identificação de espécies de morcegos da África do Sul e reduzir os custos associados, pois haverá menor necessidade de recorrer a especialistas externos para realizar identificação de gravações.

Os próximos passos deste trabalho devem focar-se em enriquecer a base de dados, nomeadamente as espécies menos representadas, como *Chaerephon pumilus*, *Sauromys petrophilus*, *Myotis tricolor* e *Neoromicia nanus*, *Miniopterus fraterculus* e *Taphozous mauritanus*. Em teoria, o melhoramento da base de dados permitirá que as predições se tornem mais precisas.

As limitações atuais de nossa abordagem incluem a propagação de erros uma vez que, se uma espécie é classificada incorretamente, não será possível reverter o erro porque os modelos são aplicados sequencialmente. Existe também incerteza sobre como se comportarão os modelos quando confrontados com espécies que não estão ainda presentes na base de dados e para as quais nenhum modelo foi treinado.

Palavras-chave: Morcegos; acústica passiva; *machine learning*; aprendizagem supervisionada; classicação automática.

Abstract

In recent years, given the need to find alternatives to fossil fuels, there has been an increase in the focus on renewable energies, especially wind power. However, the installation of wind farms is not free of negative environmental impacts, such as direct mortality of bats and birds. Bats are unique animals that play an important ecological role and their conservation is of the utmost importance. Acoustic identification is fundamental to assess the impact of wind farms on bat communities. The need to speed up the identification process of bat species led us to develop a methodology to identify species of bats from South Africa using machine learning methodologies. The Analook software was used to extract variables related to the frequency, slope, and duration of echolocation pulses from specialist-identified recordings. With this information a database was compiled and divided into training and test sets. The training set was used to train models to identify individual bat species, using an array of algorithms that included random forests, support vector machine, extreme gradient boosting and flexible discriminant analysis. The predictive power of the models was evaluated using the test set. In this work it was possible to obtain high identification accuracy rates for a set of species considered of greater interest in the context of impact studies in wind farms. In the near future, the methodology developed in this work will enhance the process of recording identification and reduce the associated costs. Moreover, gathering new records will improve the database, and allow for more precise predictions on the identification of bat species.

Keywords: bats; passive acoustics; machine learning; supervised learning; automatic classification.

Acknowledgements

Firstly, I would like to give a special thanks to Bioinsight for allowing me to develop this project.

I am grateful to my supervisors Tiago Marques and Helena Mouriño for their support.

A very special gratitude goes out to Sandra Rodrigues from Bioinsight who gave me valuable support and insight on bats.

Last but not the least, I would like to thank Ana for always being there for me.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area Under the Curve
Caret	Classification And REgression Training
CART	Classification and Regression Trees
CV	Cross-Validation
XGB	Extreme Gradient Boosting
FN	False Negative
FP	False Positive
FDA	Flexible Discriminant Analysis
FPR	False positives rate
FNR	False negatives rate
LDA	Linear Discriminant Analysis
ML	Machine Learning
QDA	Quadratic Discriminant Analysis
RF	Random Forests
SVM	Support Vector Machines
SMOTE	Synthetic Minority oversampling technique
TN	True Negatives
TP	True Positives
WAV	WAVEform audio format
XGBoost	eXtreme Gradient Boosting
ZC	Zero Crossing

Chapter 1

Introduction

1.1 Introduction to Bats

Bats are the second largest order of mammals and the only flying mammals (Tsang, Cirranello, Bates, & Simmons, 2016). These animals play key ecological and economic roles, as they are important pollinators and seed dispersers for several ecologically and economically important plants. Bats also play a major role in suppressing nocturnal insect populations, functioning as pest management agents (Jones et al., 2009).

In virtue of occupying a wide array of ecological niches, these mammals can be used as bioindicators, to monitor the ecosystem's health (Jones et al., 2009). Therefore, assessing bats' occurrence and distribution should be part of the effort to evaluate, conserve and monitor biodiversity (Lim & Engstrom, 2001). However, bats are difficult to study for being very small, fast flying and having nocturnal habits. Additionally, there are many species, the majority of which similar in appearance and with difficult to find roosts. Consequently, survey programs that rely on visual encounters or captures of individuals are difficult to apply to bats (Ochoa et al., 2000).

Monitoring bats is particularly important in the context of wind farms. Wind energy has become an increasingly important sector of the renewable energy industry (GAO, 2005). Wind power is widely known as a clean energy and considered to be an environmentally friendly source of energy with few negative impacts. Nevertheless, wind farms are responsible for bat and bird mortality resulting from collision and barotrauma for bats (Kunz et al., 2007; Baerwald, D'Amours, Klug, & Barclay, 2008; Drewitt & Langston, 2008; Cryan & Barclay, 2009). Aside from mortality, wind farms may also lead to habitat displacement (Arnett et al., 2007). Hence the importance of monitoring these areas to measure patterns of habitat used by local and migrating species, as well as the impact in the number of individuals using the area before and after the wind farm construction and operation.

Reliable bat identification can be achieved through handling, though contact and roost disturbance is a major contributor to population decline, being strictly regulated by the EuroBats agreement (EUROBATS, 1991). For this reason, non-intrusive means of detection and identification, which avoid handling, are preferable (Milne, 2002).

Microbats or insect-eating bats are known to continuously emit echolocation signals and analyze the returning echoes when orienting in space, searching for food, and/or approaching a target of interest (Schnitzler et al., 2003). Interestingly, echolocation calls can be used to identify bat species, this strategy in addition to helping to monitor and assess the conservation status of a population, is in compliance with environmental regulations (Ahlen & Baagøe, 1999). Megabats or the fruit-eating bats, are an exception since they find food by sight and smell and do not depend on echolocation (Springer, Teeling, Madsen, Stanhope, & de Jong, 2001).

1.2 Acoustic Identification

Reliable species identification is critical for survey and monitoring programs. For this propose acoustic surveys have indeed become an increasingly popular alternative to conventional bat survey methods (Ahlen & Baagøe, 1999). These have the advantage to be carried out in a wide range of habitats, allowing the detection of many species (Walters et al., 2012). Acoustic identification is thus a major help for bat conservation.

Identification of bat species through acoustics can be broadly divided into two categories: manual and automated. The manual approach identifies bats through ultrasonic calls relying on the visual matching of sonogram patterns. This method can also use observation of flight behavior, morphological characteristics and inspection of the habitat. Broadly, these procedures need a fair amount of time and multiple specialists making it also costly (Gaston & O'Neill, 2004; Jennings et al., 2008). Also, it can be quite subjective and a challenging task, as it fundamentally depends on the knowledge and experience of the investigator, as well as the availability of good reference calls (Gaston & O'Neill, 2004; Jennings et al., 2008). However, identification by experts allows to introduce a certain flexibility in identification, as well as the incorporation of features and patterns that may be difficult to quantify (Jennings et al., 2008).

By contrast, automated methods rely on the extraction of spectral and temporal parameters of the calls which can be used in statistical/mathematical analysis (Vaughan et al., 1997; Herr, Klomp, & Atkinson, 1997). Automation of species' identification offers advantages such as reproducibility, objective and quicker identification with potentially higher levels of accuracy (Gaston & O'Neill, 2004; Jennings et al., 2008).

Acoustic identification may become difficult since bats can vary the structure of their calls. Bat echolocation is relatively plastic as they can adjust their sonar signal according to the nature of the environment where they are foraging and their behavioral goals (Schnitzler et al., 2003; Siemers & Schnitzler, 2004). Moreover, bat echolocation may vary from individual to individual and species occupying similar foraging niches often produce similar calls (Noda, 1995).

1.3 Motivation and Objectives

Without a doubt, today we live in a truly globalized world. The economy is no exception to this increasingly evident trend, as goods, capital, services, technology and, information are traded at a worldwide scale. Therefore, to stay relevant and competitive, companies must seek new markets and opportunities. This is exactly what Bioinsight, an environmental consultancy company[1], is doing. Bioinsight launched the challenge to develop an expeditious procedure to identify bat species from South Africa, in the context of wind farms.

Despite producing clean energy, wind turbines are a deadly menace to bats (Amos, 2016). Surveying bats in these regions is a crucial, but onerous task, as it encompasses large areas and thus becoming too demanding in terms of workforce and associated costs. Efficient and reliable surveys can be

[1]Broadly, these companies ensure compliance with environmental regulations, so when projects with potential environmental impact are to be executed, they assess the impact on the fauna and flora.

accomplished by using automatic sound detectors, able to record large amounts of bat sounds. The goal of this work was to identify the species present in this type of recordings. We started by organizing a database of bat pulses from different species of interest and then trained classificatory models using state-of-the-art machine learning algorithms. The final objective was to establish a sequence to apply the best models, to identify each species, and write a script that can be applied by a user to generate an output with information regarding the species present in each recording.

1.4 Thesis Outline

The remaining chapters of this thesis are organized as follows:

- **Chapter 2** describes the construction of the call library prepared for this work, as well as the key variables associated with the bat pulses. In this chapter are also presented the bat species considered for this study.
- **Chapter 3** introduces the classificatory algorithms, and performance assessment metrics used in this work.
- **Chapter 4** explores the library of bat pulses prepared for this study, with details on how the different variables vary by species.
- **Chapter 5** details the process of data splitting for training and testing the model, and the rational behind model usage.
- **Chapter 6** presents and evaluates the models trained for the different species. Additionally, we also show how these models perform on the test set as well as on an external subset of recordings.
- **Chapter 7** discusses the results obtained on their strengths and flaws, in comparison with other works.
- **Chapter 8** sums up the main conclusions of this work.

Chapter 2

Bat Pulses' Database

The data used in this work comes from recordings made by field detectors, placed at different wind farms in South Africa. These devices are programmed to record when sound is detected in the frequency range between 12 and 192 kHz. Bats are known to emit calls ranging from 12 kHz to 160 kHz (Basil et al., 2014). Using bat detectors allows to monitor the activity of species that fly both above and below the canopy, provided their calls are loud enough (Fukui, Agetsuma, & Hill, 2004).

2.1 Audio Formats

The recordings were stored in the WAV format which were then converted to ZC using the Kaleidoscope software version 4.1.0a (Wildlife Acoustics). ZC express the transition of a signal waveform from positive to negative, i.e. number of times the sound wave crosses the zero-axis. The conversion to ZC is fast and simple as it does not rely on complex mathematical *formulae* to transform a digital signal into the frequency domain (Parsons et al., 2000).

Though this analysis is very efficient, it has as main disadvantage the fact that it only tracks the part of the signal with most energy associated with it, meaning that all remaining harmonic information in the signal is lost (Parsons et al., 2000; Armitage & Ober, 2010b). According to (Parsons et al., 2000) in some bat species, frequency may overlap between harmonics, meaning that the harmonic with most energy may change over the course of the call. This can lead to a misleading output signal due to the oscillation between harmonics. Another drawback of the utilization of ZC is its susceptibility to the presence of noise in the signal, making the interpretation of frequency data difficult (Parsons et al., 2000). On the plus side, ZC require less storage space.

2.2 Call Library

A key point before proceeding to create classification models is the selection of reference bat calls (Clement, Murray, Solick, & Gruver, 2014). The recordings used to construct the library had been previously identified by an expert, and are known to contain only one species in each audio file. Bat calls and their features were extracted with Analook version 4.2g (Titley Scientific). This software identifies calls present in the ZC files and measures different spectral features of the echolocation calls (see list in section 2.3).

Automation of call extraction and feature measurement allows to record precise and detailed data, avoiding researcher bias. Moreover, given the amount of data requiring processing it would be unpractical to perform such measurements by hand, as it is less precise and time-consuming process (Parsons & Jones, 2000; Scott, 2012).

Since pulse fragments or noise produced by non-bat sources may introduce undesired bias, thus these must be removed as much as possible using filters (Clement et al., 2014). Filter selection constitutes a key step of the acoustic identification process, as they can influence the content of the data set and the quality of the classification models trained (Clement et al., 2014). A good filter should prevent a classification model to be trained on noise, rather than true bat calls. Additionally, it is desirable that not too many true bat calls are excluded.

2.3 Variables Associated with the Pulses

The extracted parameters associated with the bat pulses are resumed next (see Figure 2.1 for a representation):

- Duration (**Dur**): The total time, in milliseconds, that a single call lasts.
- Time between calls (**Tbc**): The time, in milliseconds, between the start of one call and the start of the next call.
- Characteristic frequency (**Fc**): The frequency of the call at its lowest slope (kHz).
- Maximum frequency (**Fmax**): The highest frequency of the call (kHz).
- Minimum frequency (**Fmin**): The lowest frequency of the call (kHz).
- Mean frequency (**Fmean**).
- Frequency at the knee (**Fk**): (kHz).
- Time at the point of the characteristic frequency (**Tc**): in milliseconds.
- Time at the knee (**Tk**): in milliseconds.
- Slope at the start of the call (**s1**).
- Slope of the call at Fc (**Sc**).
- Duration of the portion of the call with the smallest slope (**Dc**): $Tc - Tk = Dc$.
- Call quality **Qual**: which is an averaged measure of the smoothness of the call.
- **Qk**, quality at the knee.
- Proportion of the maximum frequency with respect to the characteristic frequency (**Pmc**): $100 \times \frac{Fmax-Fc}{Fc}$.

2.4 Bat-Call Library Construction

The construction of the call library encompassed two steps of filtering to ensure call quality. The first, an Analook filter with the following specifications: Smoothness: 50; Body over: 1000 microseconds; Fmin: 10; Fmax: 180; Dur min: 2; Dur max: 50. The second step retained only the calls with $Qual \leqslant 0.30$. Qual is an averaged measure of the smoothness of the call, values lower than 0.30 are associated with good quality calls (Armitage & Ober, 2010a; Clement et al., 2014).

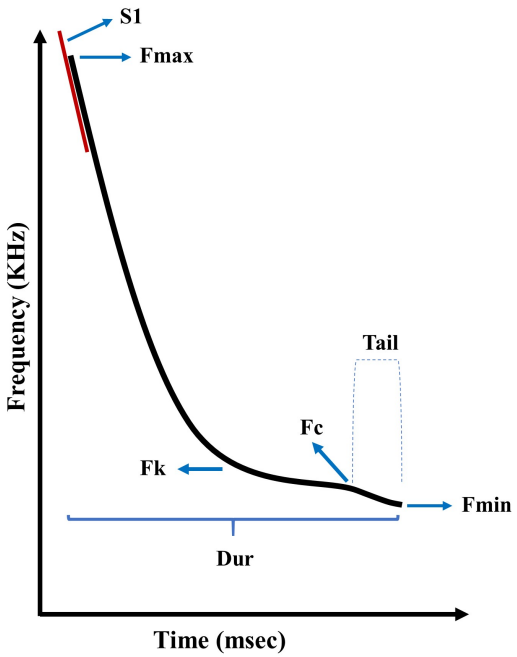

Figure 2.1: Schematic representation of a generic bat pulse.

2.5 Bat Species

The species present in the recordings used to construct the call library are resumed in Table 2.1.

Table 2.1: Bat species present in the reference library prepared for this study.

Abbreviation	Scientific name	Common name
Cpum	*Chaerephon pumilus*	Little Free-tailed Bat
Ehot	*Eptesicus hottentotus*	Long-tailed Greater Serotine Bat
Mfra	*Miniopterus fraterculus*	Lesser Long-fingered Bat
Mnat	*Miniopterus natalensis*	Schreiber's Long-fingered Bat
Mtri	*Myotis tricolor*	Temminck's Hairy Bat
Ncap	*Neoromicia capensis*	Cape Serotine Bat
Nnan	*Neoromicia nanus*	Banana Bat
Rcap	*Rhinolophus capensis*	Cape horse shoe bat
Rcli	*Rhinolophus clivosus*	Geoffroy's Horseshoe Bat
Rdar	*Rhinolophus darlingi*	Darling's Horseshoe Bat
Rsim	*Rhinolophus simulator*	Bushveld Horseshoe
Spet	*Sauromys petrophilus*	Flat-headed Free-tailed Bat
Taeg	*Tadarida aegyptiaca*	Egyptian Free-tailed Bat
Tmau	*Taphozous mauritianus*	Mauritian Tomb Bat

Chapter 3

Classification Models

3.1 Machine Learning: An Overview

Nowadays remote ultrasound detectors allow storage of large amounts of data on bat activity (Jennings et al., 2008). Additionally, automated recording allows multiple sites to be surveyed simultaneous. As a result we get access to a great deal of information, however it can be extremely hard for a human to interpret and make sense of all that data. Therefore, automatic methods are needed to uncover patterns in the data, interpret them and draw conclusions. Machine learning algorithms can be used to perform this type of analysis. Machine learning is a field that crosses computer and statistical research and whose goal is the design of automatic learning systems, nowadays it is around us and critical to our life (Domingos, 2015). In fact, it allows to work on complex prediction problems such as speech recognition, image recognition, nonlinear time series prediction, handwriting recognition and prediction on financial markets (Domingos, 2015).

Machine Learning algorithms are trained through an iterative process of providing positive and negative feedback on the results they give. Opposite to traditional statistical modeling, machine learning makes no assumptions about the underlying data distribution, which is treated as unknown (Breiman, 2001). Machine learning methods fall into two major categories, supervised and unsupervised.

In supervised learning, there is a dataset where the outcome variable is mapped to feature vectors that describe the label, i.e., there are instances where we know which group (of a number of potential groups) the observations belong to.

Supervised learning algorithms learn the classes' characteristics of the dependent variable by analyzing sample objects that have been previously annotated and classified, this step is known as training. A supervised learning algorithm attempts to optimize a function (the model) to find the combination of features that result in the target output. The resulting models will then be able predict the labels for new feature vectors. The type of prediction made varies according to the label type of data. So, given the target function, f, that maps each attribute set, $\mathbf{x}$, to one of the predefined class label, y, that is, $f : \mathbf{x} \rightarrow y$, then we have a classification problem. Classification entails identifying group membership. In case y is continuous, then it represents a regression model. Hence, the key characteristic that distinguishes classification models from regression models is the nature of the class label, y. Classification methods are mostly suited for predicting or describing data sets with binary or nominal categories (Tan, Steinbach, & Kumar, 2006).

Next, we introduce the notation for the attribute set and the class label. Let $\mathbf{X}$ be a matrix where each column represents one of the p independent variables, and each line represents an observation. Additionally, let $\mathbf{y}$ be a n-dimensional vector containing the value of each dependent observation. Let $\mathbf{X} = x_{ij}$, where x_{ij} represents the value of the i-th observation for the j-th variable, and the i-th observation is characterized by a p-dimensional vector, $\mathbf{x_i} = (x_{i1}, ..., x_{ip})$. In a classification problem,

the goal is to train a model with a given training set, $L = \{(x_1, y_1), ..., (x_{n_L}, y_{n_L})\}$, with n_L labeled observations which serve as a reference for the classification of new samples constituting the test set, $T = \{x_1, ..., x_{n_T}\}$, where n_T designates the number of samples of this set.

Contrasting with supervised learning, the main goal of unsupervised learning is to find similarities and relationships in the data, since there is not an explicit class label y. Unsupervised learning algorithms help in pattern recognition, and in the discovery of groups of similar examples within the data.

In this context, cluster analysis can be considered as unsupervised learning. In fact, cluster analysis is a technique to group the data into homogeneous classes (or clusters). It is based on information contained only in the data that describes the attribute set, $\mathbf{x}$, and their relationships. Clustering can, thus, be regarded as a form of classification in that it creates a labeling of objects with class (cluster) labels (Tan et al., 2006). An example in the bat sound context might include processing recordings in an attempt to find sounds from different species of bats, without knowing *a priori* exactly how many species are present. Lets suppose we identify 4 groups of sounds, these may represent 4 different bat species, or a variation of sounds from the same species. Therefore, without knowing the real labels it is impossible to be sure what the clusters really represent, consequently unsupervised methods are best suited for an exploratory approach of the data.

Automated acoustic identification has already been used on a wide range of animal species, from birds (Peake & McGregor, 2001), to marine mammals (Mellinger & Clark, 2000; Yack et al., 2009), and insects (Mankin, 2011). Classification of bat echolocation calls has been carried out using a variety of approaches, including multivariate statistical analysis (Vaughan et al., 1997; Papadatou et al., 2008), hidden Markov models (Skowronski & Harris, 2006), artificial neural networks (Parsons et al., 2000), support vector machines (Redgwell et al., 2009), and random forests (Armitage & Ober, 2010a).

3.2 Supervised Learning Methods Used in this Work

To classify bat sounds to species, we considered a set of supervised machine learning methods. In the next sections we will give a brief description of the most well-known methods. The algorithms are complex and diverse, it is therefore not feasible to cover them extensively, which is not the goal of the present work.

3.2.1 Discriminant Analysis

Linear Discriminant analysis (LDA) was established by Fisher, 1936, and Welch, 1939. The main goals of this method are to identify the variables that best differentiate two or more classes, and to construct a classification rule that allows predicting to which group a new sample belongs to. Fisher introduced the discriminant function as a linear combination of predictors such that the between-group variance was maximized relative to the within-group variance (Kuhn & Johnson, 2013). Welch took a Bayesian approach to minimize the total probability of misclassification, provided the predictors have a Normal distribution (Kuhn & Johnson, 2013). Fisher's approach makes no assumption regarding either the distribution of the explanatory random variables nor the covariance matrix. However, if the classes

are Gaussian with identical covariance, the LDA solution will be optimal for classification. If these two assumptions are not true, then such optimality is not guaranteed (Friedman, 2001).

To obtain an LDA solution is necessary that the covariance matrix is invertible, a unique solution exists only when this matrix is invertible. In practice this means that the data must contain more samples than predictors, and the predictors are independent (Friedman, 2001). LDA is a good method when applied to problems with linearly separable classes, but is severely impaired if the data do not present this characteristic (Kuhn & Johnson, 2013). Additionally, the normality assumption can be too restrictive, or even it does not apply, therefore it will be difficult to obtain an optimal solution.

This has led to a search for non-parametric solutions that overcome these restrictions. If we allow the covariance matrices to be different across the groups, it leads to quadratic discriminant analysis, where the primary repercussion is that the decision boundaries now become quadratically curvilinear in the predictor space (Friedman, 2001; Kuhn & Johnson, 2013). Quadratic Discriminant Analysis (QDA) holds the same assumptions as LDA except that the covariance matrix is not common to all classes (Friedman, 2001).

LDA and QDA minimize the total probability of misclassification if the data can truly be separated by hyperplanes or quadratic surfaces. Reality may be, however, that the data are best separated by structures somewhere between linear and quadratic class boundaries (Kuhn & Johnson, 2013). Therefore, penalization strategies can be applied to LDA models so there are no imposing restrictions on the data under study (Kuhn & Johnson, 2013). For instance an elastic-net strategy (Zou & Hastie, 2005) can be applied, implying the utilization of so-called L1 and L2 penalties, the first has the ability to eliminate predictors, while the second shrinks the coefficients of the discriminant functions towards zero. This conceptual framework is referred to as flexible discriminant analysis (FDA) (Friedman, 2001; Clemmensen, Hastie, Witten, & Ersbøll, 2011), and it is suited for problems such as signal and image classification, where a large number of highly correlated features exists (Friedman, Hastie, & Tibshirani, 2001).

3.2.2 Random Forests

The Random Forests (RF) method was created by Leo Breiman in 2001. RF represent an ensemble learning method (Breiman, 2001), which main idea is to combine a set of weak learners to create a strong learner that obtains better performance than any of its individual components. The building blocks of RF are classification trees, which are hierarchical structures consisting of nodes and directed edges (Tan et al., 2006). These trees are the result of a series of carefully crafted questions about the attributes of the test record. Each time we receive an answer, a follow-up question is asked until we reach a conclusion about the class label of the record (Tan et al., 2006). A decision tree has three types of nodes: root node, internal node, and leaf or terminal node (each assigned to a class label). Tree-based models have several advantages, in particular the creation of a set of conditions that are interpretable and easy to implement (Breiman, 2001). Also, by selecting many independent learners the variance of the overall ensemble is reduced relative to any individual learner in the ensemble (Breiman, 1996, 2001).

RF take on the concept of bagging, short for bootstrap aggregation, where each model in the ensemble is used to generate a prediction for a new sample (Breiman, 1996). Bagging is used to reduce the variance of an estimated prediction function (Friedman et al., 2001). In RF, each tree casts a vote for the classification of a new sample, so that the predictions of each tree are averaged to give the forest's prediction (Kuhn & Johnson, 2013). Another interesting feature of this algorithm is that it diminishes correlation between trees by introducing a random element into their construction, meaning that the predictors are randomly selected at each split (Breiman, 2001), this represents a tunning parameter of RF algorithm which is commonly referred to as m_{try} (Kuhn & Johnson, 2013).

Because of the logic in their construction, RF, can effectively handle many types of predictors (sparse, skewed, continuous, categorical, etc) without the need to pre-process them (Kuhn & Johnson, 2013). Furthermore, these models can effectively handle missing data and implicitly conduct feature selection and importantly, can deal with multicollinearity, characteristics that are desirable for many real-life modeling problems (Breiman, 2001; Kuhn & Johnson, 2013).

3.2.3 Support Vector Machines

Support Vector Machines (SVM) are a class of statistical models that can be used to solve both regression and classification problems. SVM were first developed by Vladimir Vapnik in the 1960s and in recent years have evolved and became one of the most flexible and effective machine learning tools available (Cortes & Vapnik, 1995). The basic idea behind SVM is to find a boundary, called a hyperplane, to partition data into groups of similar class values (Lantz, 2015). Sometimes, an infinity of hyperplanes may be found to accomplish this purpose, hence an alternate metric called the *margin* has been introduced (Cortes & Vapnik, 1995). In general terms, the *margin* stands for the distance between the classification boundary (i.e. a hyperplane that allows to separate the classes) and the closest training set points (Kuhn & Johnson, 2013). Here, we want to find the hyperplane that maximizes the margin allowing for the separation between classes, which is known as the maximum margin classifier (Kuhn & Johnson, 2013).

In its simplest approach, the linear SVM (Eq. 3.1), can be summarized as follows: having a training set D with n data points, where $\mathbf{x}_i$ represents the p-dimensional real vector corresponding to the predictors' values, and y_i the corresponding class label. For a two-class problem, y_i takes either values 1 or -1, for class 1 and class 2, respectively.

$$D = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \, \epsilon \, \mathbb{R}^p, \; y_i \, \epsilon \, \{-1, 1\}\} \, i = 1, 2, ..., n \tag{3.1}$$

$$\mathbf{w}.\mathbf{x} = 0 \tag{3.2}$$

Eq. 3.2 shows a generic representation of a hyperplane, in which $\mathbf{w}$ is a normal vector to the hyperplane. In case our data are linearly separable, two parallel hyperplanes that separate the two classes can be selected to find the largest distance between them, i.e. the margin. The margin is calculated as $\frac{2}{\|w\|}$, thus we want to find the parameters for the hyperplane that minimize the norm of w.

Solving the inequation of the hyperplane given the value of the classes will help us determine to which the data points belong to.

Interestingly, to calculate the margin only the data points closer to it are needed, therefore the maximum margin classifier is a function of only a subset of the training set points and these are referred to as the *support vectors*. This feature contrasts with discriminant analysis where the decision boundary is determined by the covariance of the class distributions and the positions of the class centroids (Friedman et al., 2001).

When there is no linear hyperplane that separates the classes from a given dataset, the data can be mapped to higher dimensions to find a hyperplane capable of splitting the dataset classes (Cortes & Vapnik, 1995). To learn nonlinear decision boundaries, specific functions denominated kernels are used, allowing the SVM model to produce extremely flexible decision boundaries. Some of the most used kernel functions are the Polynomial, Gaussian and Radial, but others exist. The kernel functions have parameters that control the complexity of the models and can be adjusted (Kuhn & Johnson, 2013).

3.2.4 Extreme Gradient Boosting

Developed by Tianqi Chen, eXtreme Gradient Boosting (XGBoost), is the most recent method used in this work (Chen & Guestrin, 2016). It can be used for supervised learning tasks such as regression and classification. This model is built on the principles of gradient boosting, a technique that produces predictions in the form of an ensemble of "weak" prediction models, typically tree-based models, in an iterative fashion. The term boosting designates an ensemble method in which new models are added to correct the errors made by previous existing models, additional models are added sequentially until no further improvements can be made. The XGBoost algorithm is focused on computational speed and model performance, by pushing the limit of computation resources, which is the reason why it become so largely used (Chen, 2015).

In XGBoost we want to determine the best tree parameters given the training data. Therefore, an objective function is defined and the propose is to measure the performance of the model, given a certain set of parameters denoted by Θ. The objective function $Obj(\Theta)$ (Eq. 3.3) consists of two parts, training loss, L, and regularization denoted by Ω.

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \tag{3.3}$$

The training loss measures the quality of the predictions of the model on the training data, while the regularization term controls the complexity of the model. The regularization term is necessary to ensure that the model is kept simple and predictive, since too much complexity may prevent good generalization.

As a boosting method, in XGBoost the learning is done additively, meaning that what has been learned until a certain point is fixed, and a new tree is added, at a time, to optimize the quality of the final model according to the objective that has been set, until no more improvement is possible. The complexity of the new tree is defined through a function that specifies its structure and other inner

characteristics. Additionally, at each step the optimization goal for the new tree is updated and the quality of the tree evaluated (Chen & Guestrin, 2016).

3.3 Performance Assessment

After a model has been trained and its parameters tunned, we need to evaluate its performance. Here we describe some metrics used in this work to evaluate model performance. The results of a classification model can have the following outcomes: True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the sum of all these terms equals N, the total number of samples. These are usually represented in the form of a confusion matrix (Table 3.1), from which several performance metrics can be derived.

Table 3.1: Confusion matrix.

		True values	
		Positive	Negative
Predicted values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- **Accuracy:** measures the proportion of correctly predicted values,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

- **Balanced Accuracy:** represents the number of correct classifications in each class, divided by the number of examples in each class, averaged over all classes. This measurement mitigates biases which could rise from unbalanced class sizes,

$$Balanced\ Accuracy = \frac{TP}{2(TP + FN)} + \frac{TN}{2(TN + FP)} \tag{3.5}$$

- **Precision:** measures the proportion of predicted positive cases that are properly predicted,

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

- **Sensitivity:** measures the proportion of positive cases that are correctly predicted,

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.7}$$

- **Specificity:** measures the proportion of false cases that are correctly predicted,

$$Specificity = \frac{TN}{TN + FP} \tag{3.8}$$

- **F-measure:** is the harmonic mean between precision and sensitivity.

$$F - Measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{3.9}$$

- **Cohen's Kappa:** is a metric that compares an Observed Accuracy with the Accuracy that would be observed by a random classification process (Cohen, 1960).

$$Kappa = 1 - \frac{1 - Acc_{random}}{1 - Acc_{observed}} \tag{3.10}$$

- **Receiver Operator Characteristic (ROC) Curve:** Sensitivity and specificity as well as other measures of classification performance from Table 3.1, depend on the single threshold used to classify a test result as positive. An ROC curve is a two-dimensional depiction of classifier performance, that evaluates the trade-off between specificity and sensitivity (Fawcett, 2006). The ROC curve is created by evaluating the class probabilities for the model across a continuum of thresholds. For each threshold, the sensitivity and the false-positive rate ($1 - specificity$) are plotted against each other. The closer this curve is to the upper left corner, the better the classifier's performance is (that is, maximizing the true positive rate while minimizing the false positive rate). The one-to-one line corresponds to a classifier that predicts classes at random. One advantage of using ROC curves to characterize models is that, since it is a function of sensitivity and specificity, the curve is insensitive to disparities in the class proportions (Provost, Fawcett, & Kohavi, 1998; Fawcett, 2006).

- **Area Under the ROC Curve (AUC):** the AUC reduces the performance information in the ROC to a single scalar value (Fawcett, 2006). The AUC varies between 0 and 1, where a 0.5 value represents an uninformative random predictor, and 1 a perfect classifier. This metric can be interpreted as the probability that a randomly selected positive sample will rank higher than a randomly selected negative sample (LeDell, Petersen, & van der Laan, 2015). AUC is a more discriminating performance measure than accuracy (Ling, Huang, & Zhang, 2003) as it is invariant to relative class distributions (Bradley, 1997). A rule of thumb to characterize the AUC is given by Hosmer Jr, Lemeshow, and Sturdivant, 2013:

  - $AUC = 0.5$: This suggests no discrimination.
  - $0.5 < AUC < 0.7$: Poor discrimination.
  - $0.7 \leqslant AUC < 0.8$: Acceptable discrimination.
  - $0.8 \leqslant AUC < 0.9$: Excellent discrimination.
  - $AUC \geqslant 0.9$: Outstanding discrimination.

These metrics are useful to evaluate model quality, however they do not give information on how a particular model will behave when challenged with data it has not "seen". Hence, to obtain robust models the following methods can be used:

- **Holdout:** To evaluate how well the model generalizes the new data, the original dataset is divided in two different subsets, using random subsampling without repetition, the training set and the test set. The training set is used to derive a model that is further used to predict the properties of the test set members, which were not used in the model development.

- **Cross-validation:** The fundamental goal of a classification model is to have a good predictive capacity, that is, a good generalization capacity for data that has not been used in the adjustment of the model. Cross-validation is a common method of validating a model (Arlot & Celisse, 2010). Cross-validated statistics can be used as criterion of robustness and predictive ability of the model. Additionally, it is used to optimize the parameters that impact the model in order to enable the algorithm to perform the best, according to the goals established.

  In this work K-fold cross validation was used. This is an iterative process by which the training set is partitioned into k subsets, one of these subsets is retained and the remainder used for training, and subsequently the withheld subset is used as an independent test set. The process is repeated, each time withholding a different subset as the test set, until all k subsets have been used. The k resampled estimates of performance are averaged across subsets and used as a criterion of robustness and predictive ability of the model.

  Cross validation allows to maximize the available training data and gives an almost unbiased estimate of the true error (Varma & Simon, 2006). The disadvantage of this method is that the training algorithm must be rerun from scratch k times making cross validation a computationally expensive process.

Chapter 4

Exploratory Data Analysis

4.1 Bat-Call Library Cleaning and Organization

Following extraction of all pulses from the recordings, used to construct the bat-call library, we ended up with 9130 pulses, from 503 recordings divided by 11 different species. After removing low quality pulses, Qual ≤ 0.30, the number of pulses decreases to 4079 pulses, from 388 recordings. In this step information is lost, nevertheless this step is essential to reduce bias resulting from noise and possible pulse fragments or other artifacts, making sure models will be trained on true bat pulses (Jennings et al., 2008).

Table 4.1: Summary of Fc by species, before cleaning.

	Fc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	10.75	35.24	26.76	25.24	26.06	27.87	139
Ehot	26.32	46.78	33.68	32.92	33.61	34.33	791
Mfra	65.04	66.12	65.58	65.04	65.58	66.12	4
Mnat	19.28	55.17	48.44	47.34	48.19	49.38	864
Mtri	32.79	48.19	37.61	36.36	37.38	37.91	33
Ncap	33.47	42.78	36.46	35.71	36.36	37.21	1213
Nnan	63.49	69.57	66.81	64.65	67.25	69.12	12
Rhin*	81.63	87.91	83.87	82.69	83.77	84.88	14
Spet	20.41	36.36	27.87	26.56	27.63	29.44	80
Taeg	18.18	39.22	21.77	20.46	21.56	22.60	883
Tmau	24.02	40.61	28.62	26.08	26.49	27.01	46

* All species from the *Rhinolophus* genus were grouped together, due to the few number of pulses of each individual species.

Fc is one of the most important variables for bat manual identification as species tend to have characteristic frequency ranges. Inspection of the density plots (Fig. 4.1) shows that pulses' Fc, in the different species, concentrates around certain values with low variance Fig. 4.1. This type of visualization is useful to detect abnormal values. Indeed, in *Cpum* and *Mnat* we detect some unusually low frequency clusters (Table 4.1, Fig. 4.1), which can turn out to be misleading when training the models. Hence, *Cpum* pulses below 20 kHz, and *Mnat* pulses below 42 kHz were discarded. Concerning *Spet*, a cluster of pulses with Fc bellow 25 kHz, which is unusual for this species, was detected (Table 4.1, Fig. 4.1) and thus removed from the library. Additionally, *Taeg* and *Ehot* display pulses with Fc greater than 30 and 36 kHz (Table 4.1, Fig. 4.1), respectively, that are not characteristic of these species, so these pulses were also removed from the library. The summary of Fc across species is displayed in Table 4.2. The summaries of the remaining variables are displayed in Appendix A.1.

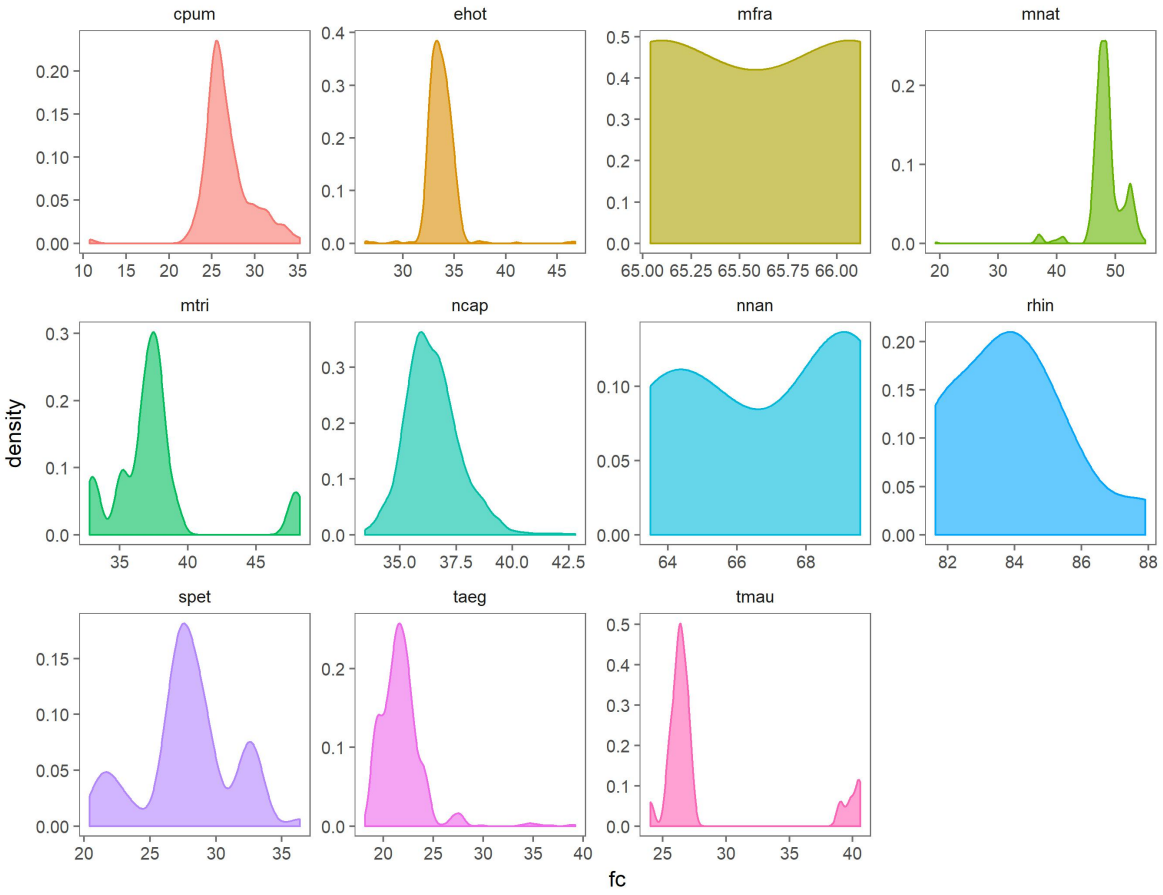

Figure 4.1: Density plots of the characteristic frequency (Fc) by species.

The resulting library is resumed in Table 4.3. A clear discrepancy is perceptible on the number of recordings and pulses per species. *Ncap*, *Taeg*, *Mnat* and *Ehot* are the most represented species in the database, and together account for over 90% of the total number of pulses in the library. These numbers are in striking contrast with those from the remaining species that together do not even add up to 10% of the calls. Due to the lack of pulses and/or recording, models for *Tmau* and *Mfra* were not trained. *Rhinolophus* spp were also excluded, since there are very few pulses from each individual species of these genus, and due to the different characteristics that these species present between them it is not advisable to group them.

Concerning *Rhinolophus spp*, these animals are difficult to detect since their calls can only be captured if in close distance to the detectors (Fukui et al., 2004). The *Rhinolophus* species present in our data display high frequency calls which are more rapidly absorbed by the air, hence the difficulty in detecting their sounds.

Table 4.2: Summary of Fc by species, after cleaning.

	Fc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	22.22	35.24	26.87	25.24	26.06	27.87	138
Ehot	26.32	35.87	33.59	32.92	33.61	34.33	781
Mfra	65.04	66.12	65.58	65.04	65.58	66.12	4
Mnat	45.20	55.17	48.79	47.34	48.19	49.38	836
Mtri	32.79	48.19	37.61	36.36	37.38	37.91	33
Ncap	33.47	42.78	36.46	35.71	36.36	37.21	1213
Nnan	63.49	69.57	66.81	64.65	67.25	69.12	12
Rhin*	81.63	87.91	83.87	82.69	83.77	84.88	14
Spet	25.81	36.36	29.01	27.30	28.27	29.96	67
Taeg	18.18	29.74	21.63	20.41	21.51	22.54	874
Tmau	24.02	40.61	28.62	26.08	26.49	27.01	46

* All species from the *Rhinolophus* genus were grouped together, due to the few number of pulses of each individual species.

Table 4.3: Number of pulses and recordings by species, after noise removal.

Species	**Number of files**	**Number of calls**	**% of calls**
Ncap	83	1213	30.19
Taeg	101	874	21.75
Mnat	94	836	20.81
Ehot	50	781	19.44
Cpum	22	138	3.43
Spet	6	67	1.67
Tmau	2	46	1.14
Mtri	9	33	0.82
Rhin	9	14	0.35
Nnan	4	12	0.30
Mfra	1	4	0.10

4.2 Explanatory Variables' Distribution by Species

Box plots are used to illustrate the distribution of each explanatory variable, which define the bat pulses, for the different species (Figures 4.2, 4.3, 4.4). Analysis of these plots gives us valuable clues to which variables allow the best discrimination between species. The optimal scenario of species identification would be to be able to draw vertical lines to achieve separation of the different species.

We observe that the predictors linked to pulse frequency seem to offer the best separation between species, specifically Fmin, Fmean, Fc and Fk (Figure 4.2). These variables are thus expected to be of importance for the models trained to identify bat pulses. Conversely, predictors Dur, Dc, Tbc, Tk, Tc, S1, Sc and Qk show a great deal of overlap across species, thus it is not expected that these will assume a key role to help distinguish between species (Figures 4.3 and 4.4).

Since there is no intuitive manner to distinguish the species with a few variables, resorting to machine learning algorithms might help us to identify patterns in the data allowing species identification.

4.3 Correlation Between Predictor Variables

Before proceeding to train models is it key to verify the existence of collinearity, the term that defines when two predictive variables have a substantial correlation with each other. If this happens between multiple predictors it is called multicollinearity. The existence of high correlation, between predictors, implies that redundant information will be passed to the models, adding nothing to their predictive capacity.

The correlation matrix between variables matrix was calculated and resumed with coloring according to its magnitude (Figure 4.5). High correlation between variables was observed, this finding was expected since, on the whole, predictors can be grouped into 3 categories: frequency, duration and slope. Hence, the variables are intrinsically related, i.e. **Fmax** and **Fmin**, or they are derived from others, as is the case of **Fmean** and **Pmc** (see Section 2.3).

Checking the correlation between the predictors, is important on the grounds that the algorithms have different tolerance for collinearity. While RF, XGBoost and FDA can accommodate high correlation between explanatory variables, QDA is highly susceptible to it.

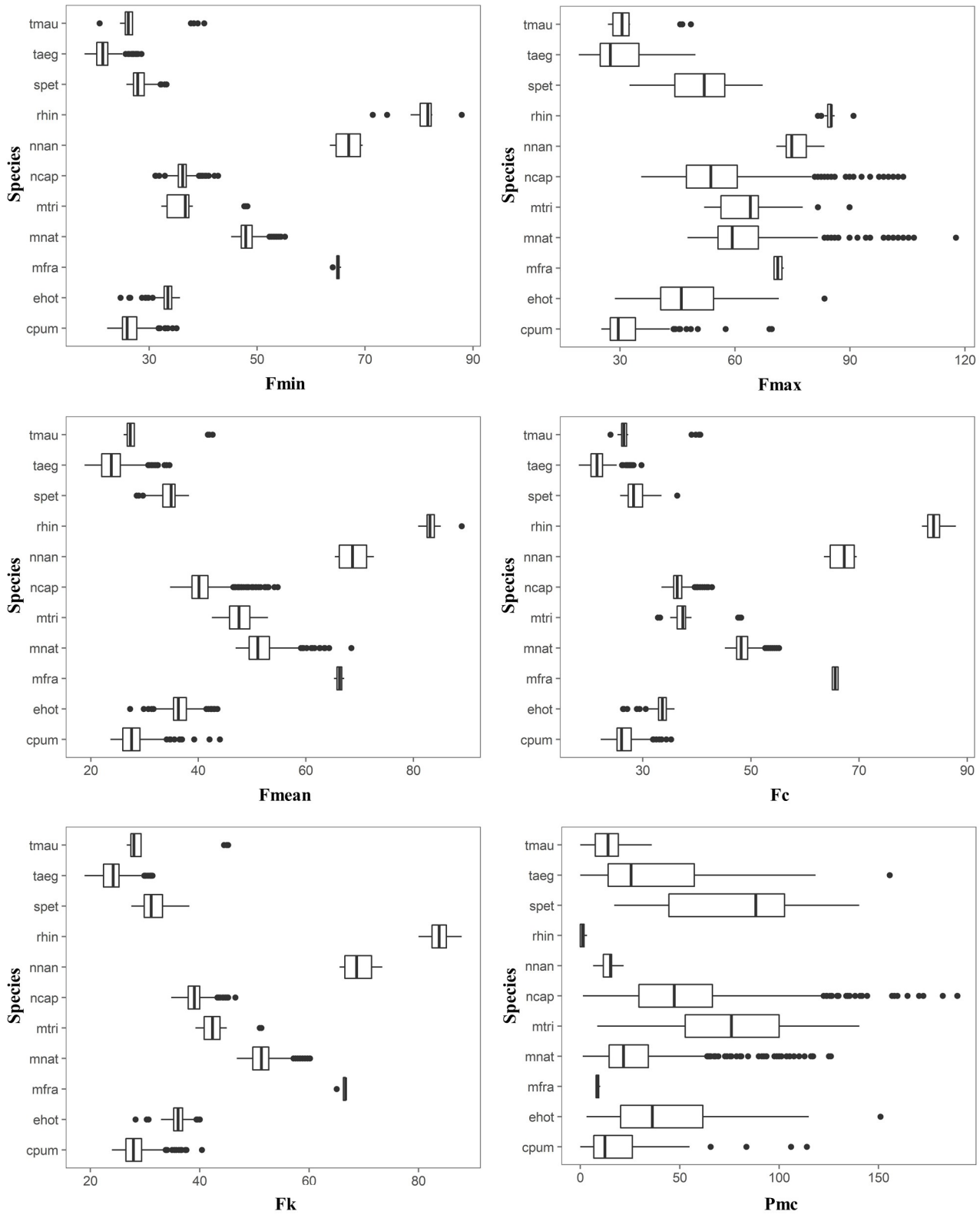

Figure 4.2: Box plots of the frequency variables associated with the bat pulses, by species.

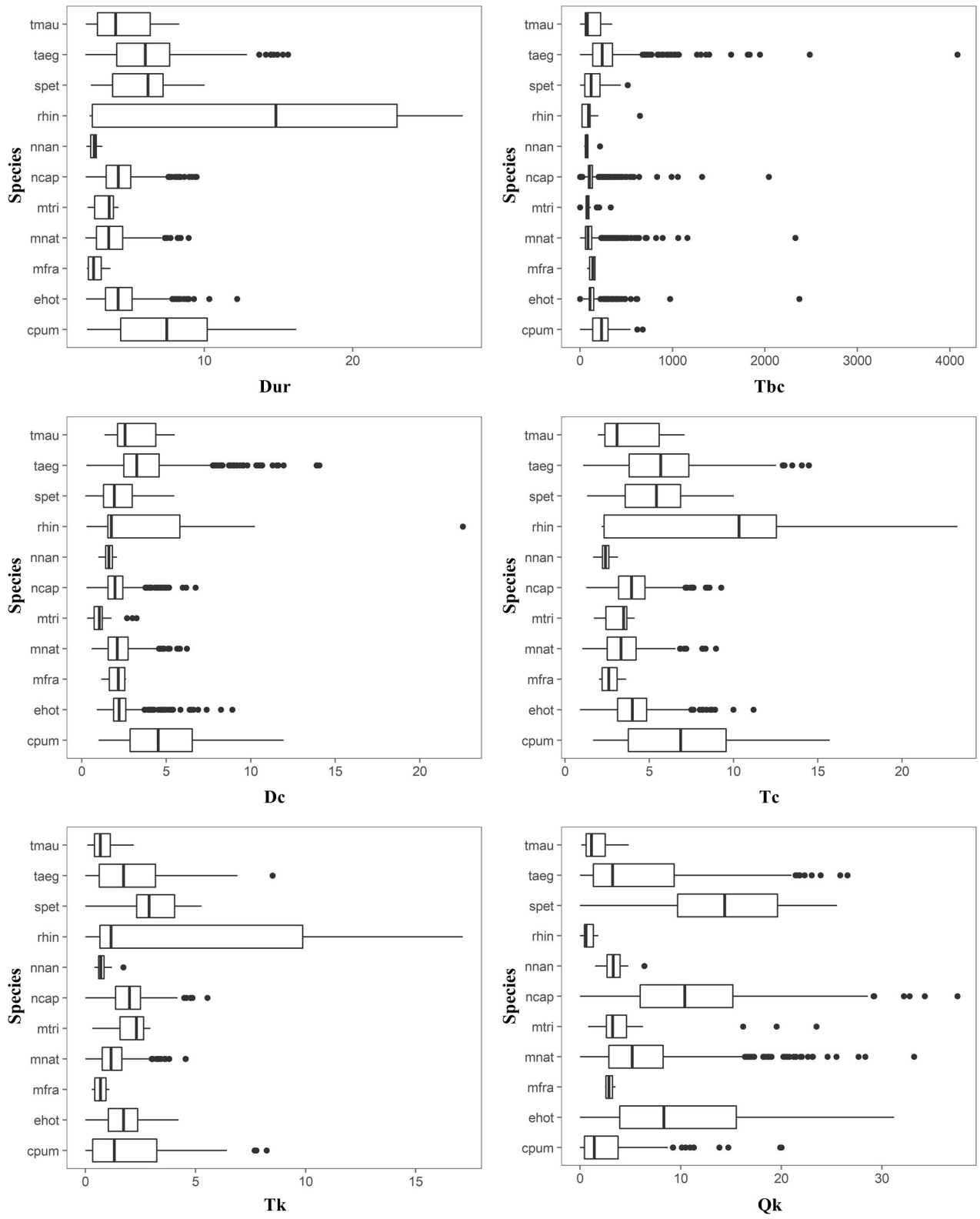

Figure 4.3: Box plots of the time variables associated with the bat pulses, by species.

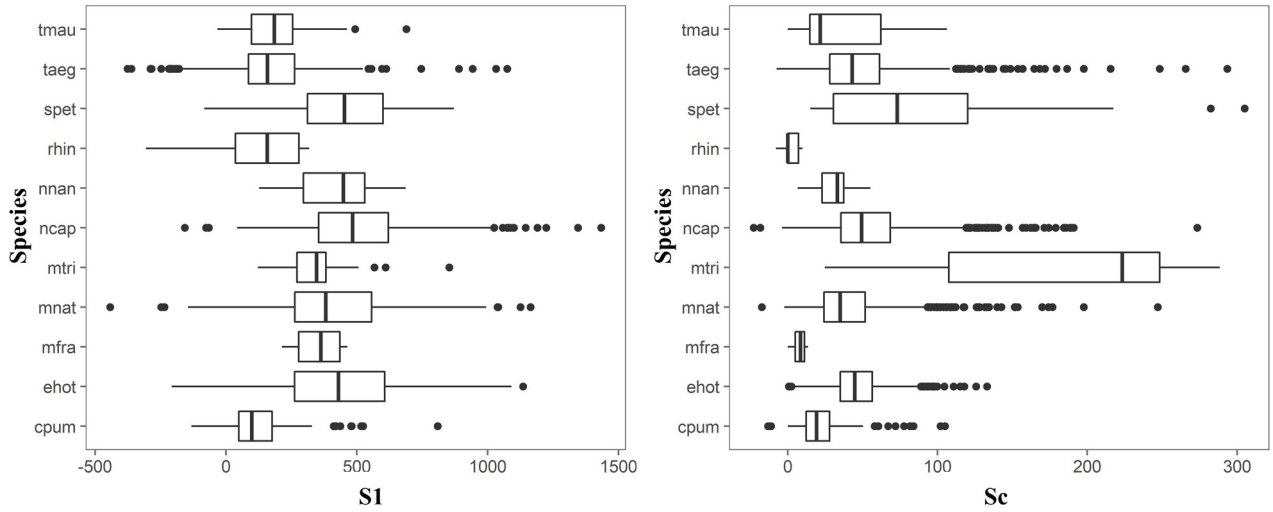

Figure 4.4: Box plots of the slope variables associated with the bat pulses, by species.

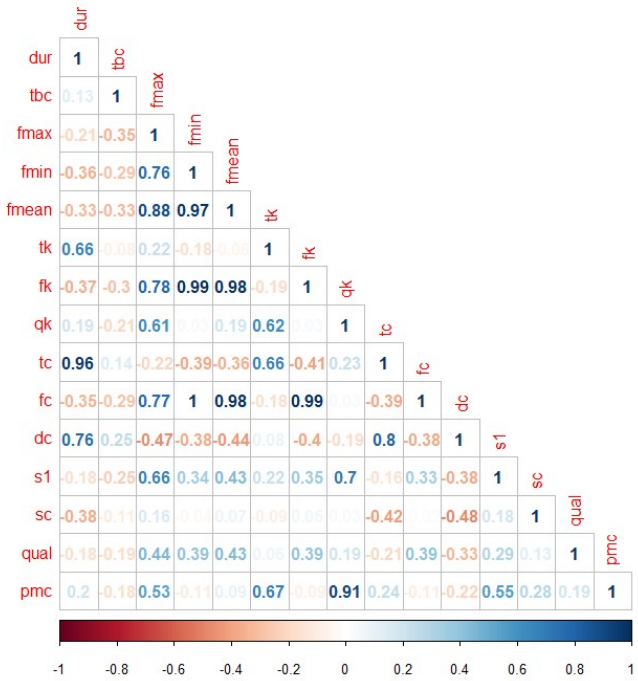

Figure 4.5: Correlation between the different variables associated with the bat pulses. Dark blue colors indicate strong positive correlations, dark red is used for strong negative correlations, and white implies no empirical relationship between the predictors

Chapter 5

Modeling Bat Species

5.1 Data Splitting

Data splitting, into a train and test set, is an important aspect of modeling. This allows to construct a classification model and evaluate how well the model extrapolates to new data (Kuhn & Johnson, 2013). Here, the data has been split at a 60:40 ratio, for train and test, respectively. Since our goal is to model discrete classes, i.e. bat species, the relative frequencies of each species can have a significant impact on the effectiveness of the model. Therefore, extra precaution must be taken since several bat species on the database have very low relative frequencies. This represents what is called an imbalance, hence the data is split using stratified random sampling to preserve the class distribution in the train and test sets (Kuhn & Johnson, 2013). Moreover, it was made sure that calls from the same audio file were not present in the same set, this allows to further test the capacity of the models to deal with new data, and prevent a possible overconfidence in the achieved performance.

5.2 Model Training

In this work a set of classificatory models were trained to identify bat species of interest for the motorizations carried out by Bioinsight in South African wind farms. To this end, different models were trained for each specie, using the following algorithms: Quadratic Discriminant Analysis (QDA), Flexible Discriminant Analysis (FDA), RF (Random Forests), eXtreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) with radial and polynomial kernels. The caret package (from Jed Wing et al., 2017), from the software R (R Core Team, 2013), was used to train the classification models, an grid search applied to tune up the parameters of the different models. The final parameters were chosen using resamples from a 10-fold cross-validation repeated 5 times[1], which was also used to assess the capacity of a model to classify previously unseen data and avoid overfitting[2].

5.3 Using the models to create a bat identification tool

As already stated in the objectives 1.3, the main goal of this work was to create a tool allowing fast identification of bats species of particular interest to Bioinsight's monitoring activity, in South Africa wind farms. It will consist of a sequence of binary classificatory models.

Firstly, the low quality or noise pulses are identified (see Table 5.1), then the sequence of classificatory models is applied to the pulses with $Qual \leqslant 0.30$, each attributing a label to the pulses

[1] In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged to produce a single estimation (Kuhn & Johnson, 2013).

[2] When a model corresponds too closely to the training data, and may therefore fail to fit additional data or predict future observations reliably.

which could be twofold: 1) the name of the species that the model was trained to identify, or "other" if the species' pulse did not, or could not, be identified as that species. The following model will only consider the pulses classified as "other", and so on. Lastly, when all models have been applied to the data, the script calculates the percentage of each species contained in every audio file as detailed in Figure 5.1. Subsequently, an output is generated in the CSV[3] format in which the user has the information for all recordings analyzed.

Table 5.1: Types of pulses defined for this study.

Classification	Characteristics
Blank or Noise	Dur = NA
Low quality or Noise	Qual > 0.30
Noise	Qual > 0.30 & Fmax < 15 kHz
Species' name or Other	Qual $\leqslant$ 0.30

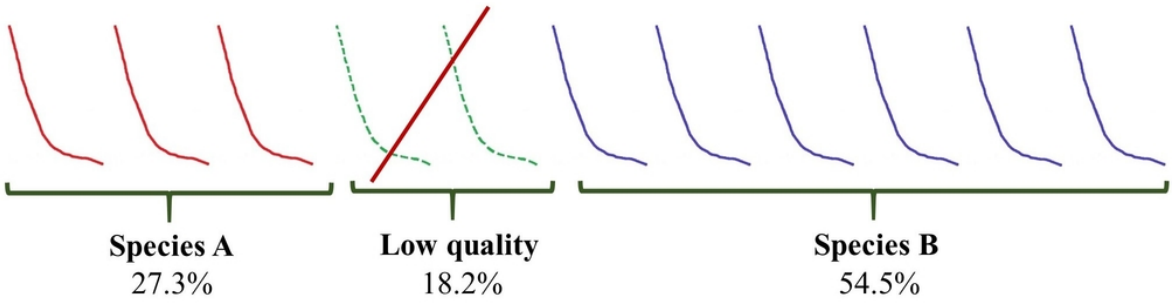

Figure 5.1: Calculating the percentage of each species in a recording. The pulses of low quality ($Qual > 0.30$) are exclude, and the relative frequency of each species calculated.

[3]Comma-separated values

Chapter 6

Results

In this section the models trained to identify the species present in the database will be discussed. We start by looking at 50 resamples of the cross-validation data from which the distribution of four different performance metrics - AUC, Cohen's Kappa, sensitivity and specificity - were compared between models, as these are good indicators of future predictive capacity of the models. Next, the best model to identify each species was chosen based on the performance in the test set, and finally a sequence was established to apply each model to be used as a tool to identify bat species.

6.1 *Tadarida aegyptiaca*

The distribution of four different performance metrics - AUC, Cohen's Kappa, sensitivity and specificity - were compared between models (Figure 6.1). AUC is a common way to compare models, generally it is a good indicator of the capacity of the make correct predictions (He & Ma, 2013; Kuhn & Johnson, 2013). The Cohen's Kappa is also helpful as it gives an idea of the concordance of the model prediction and the observed classes. Nevertheless, other metrics like sensitivity and specificity should be used to gather more information about the models giving valuable hints about future false negatives and false positives, respectively.

The values of the metrics displayed on Figure 6.1, for the different models trained, in the cross-validation data are close to 1 which may anticipate that the models have a good predictive capacity. Interestingly, models have few differences concerning the above mentioned metrics (Figure 6.1, see Table A.6 for more details), which does not allow to single out a specific model or models that will perform better.

Next, the models were tested on their capacity to identify *Taeg* pulses on the test set. The confusion matrix was calculated for each model and the results resumed in Table 6.1. Models show high balanced accuracy, as well as sensitivity and specificity. It is noteworthy that the balanced accuracy is higher than the no information rate, which is the proportion of the majority class. This shows that the results are better than if we had guessed all calls belonged to the majority class.

RF and XGBoost displayed the best trade-off between FP and FN, while discriminant analysis models (QDA and FDA) seem to be more prone to FP (Table 6.1). Further inspection of the predictions shows that the FP originate from *Cpum* calls (Table 6.2), indeed these species are both part of the family *Molossidae* and are phylogenetically close (Lamb et al., 2011).

RF and XGBoost models to identify *Taeg* pulses showed good results when applied to the test set, for future application RF will be preferred since it has less tunning parameters during the training stage.The caret package (from Jed Wing et al., 2017) has a function to calculate variable importance so that we can have information on which variables were the most informative in making distinctions

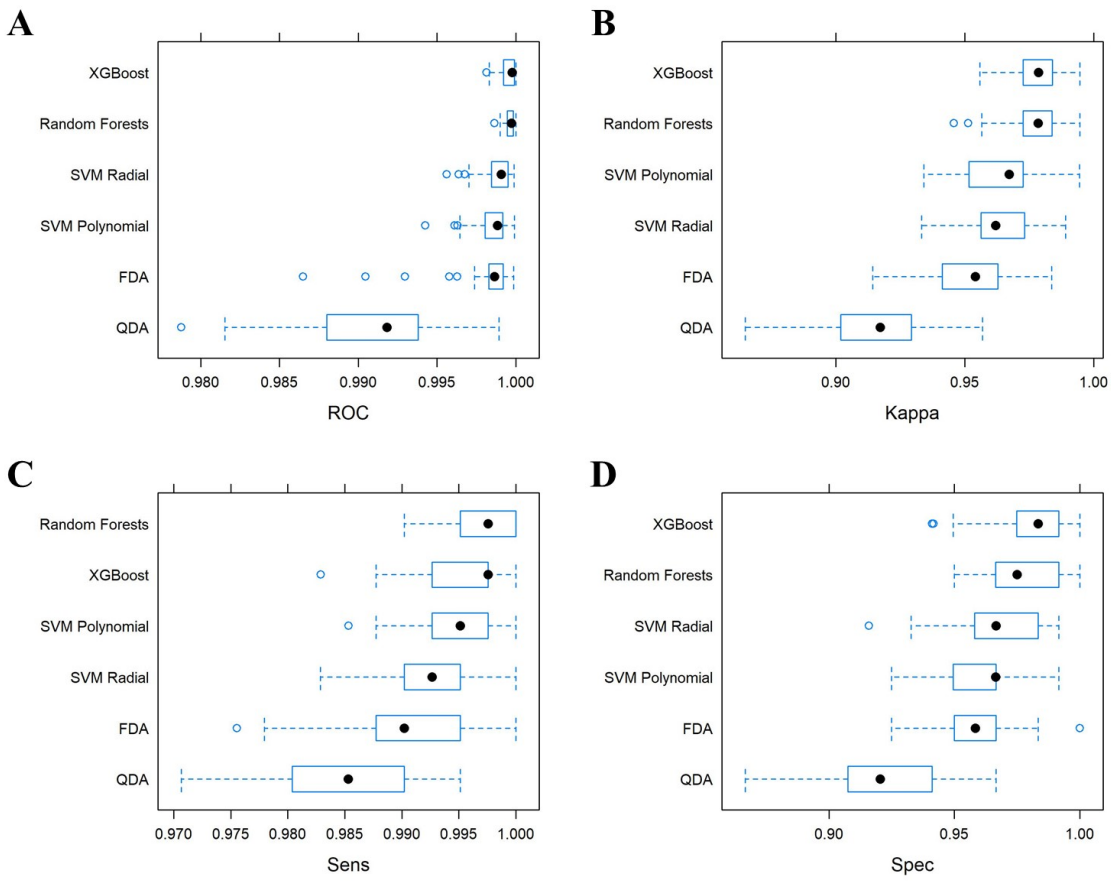

Figure 6.1: Performance estimation of the models trained to identify *Tadarida aegyptiaca* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

between classes [1]. In the RF model we have as most important predictors Fc, Fmin, Fk (Table 6.3), the same as XGBoost (Table 6.3), these models differ however in the importance of the remaining variables. As a curiosity, the FDA model only uses the variables Fc and Fmin (Table 6.3) to identify *Taeg* pulses, it is probably not enough information and hence the high number of false positives.

6.2 *Miniopterus natalensis*

Comparison of the cross-validation performance metrics, for the models trained to identify *Mnat* pulses, shows that, overall, the performance is high, with few differences between them (Figure 6.2), which tend to be not significant (Table A.7).

[1]Briefly, in Random Forests, variable importance reports the gain in prediction accuracy of one variable over the others. Similarly, in XGBoost variable importance reflects how using a certain feature to split the data labels improves accuracy relative to not using it. While for Flexible Discriminant Analysis, the importance of a variable depicts the absolute value of the coefficients of its variables (Kuhn, 2008).

Table 6.1: The confusion matrix and some performance metrics obtained for each model trained to identify *Tadarida aegyptiaca* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc.	Sens	Spec	Kappa	NIR
QDA	271	1073	29	5	0.98	0.98	0.97	0.92	0.80
FDA	274	1083	19	2	0.99	0.99	0.98	0.95	0.80
Random Forests	275	1092	10	1	0.99	1.00	0.99	0.97	0.80
XGBoost	274	1092	10	2	0.99	0.99	0.99	0.97	0.80
SVM Radial	269	1091	11	7	0.98	0.97	0.99	0.96	0.80
SVM Polynomial	273	1088	14	3	0.99	0.99	0.99	0.96	0.80

Table 6.2: Summary of the errors made by the models trained to identify *Tadarida aegyptiaca* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	-	0	0	0	29	0	0	0	0	0	5
FDA	-	0	0	0	19	0	0	0	0	0	2
Random Forests	-	0	0	0	10	0	0	0	0	0	1
XGBoost	-	0	0	0	10	0	0	0	0	0	1
SVM Radial	-	0	0	0	11	0	0	0	0	0	7
SVM Polynomial	-	0	0	0	7	0	0	0	0	0	4

Table 6.3: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Tadarida aegyptiaca* pulses.

RF		XGBoost		FDA	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
Fc	39.4	Fmin	25.4	Fc	100.0
Fmin	33.1	Fc	18.7	Fmin	22.3
Fk	27.6	Fk	18.2		
Sc	27.0	Fmean	17.3		
Fmean	22.9	Dc	3.6		
Fmax	15.7	Fmax	3.1		
Pmc	14.9	Sc	2.7		
Dc	14.4	Qk	2.7		
Tbc	12.9	Tbc	1.9		
Dur	12.0	Pmc	1.7		
Qk	11.3	S1	1.6		
Tk	10.6	Tk	1.5		
S1	10.5	Dur	1.3		
Tc	8.8	Tc	0.5		

The high performance of the models was confirmed when these were applied to the test set (Table 6.4). Aside from QDA, all models have a perfect prediction, making no mistakes classifying *Mnat* pulses present in the test set. Importantly, the balanced accuracy is higher than the no information rate, showing that the correct predictions are better than a random guess.

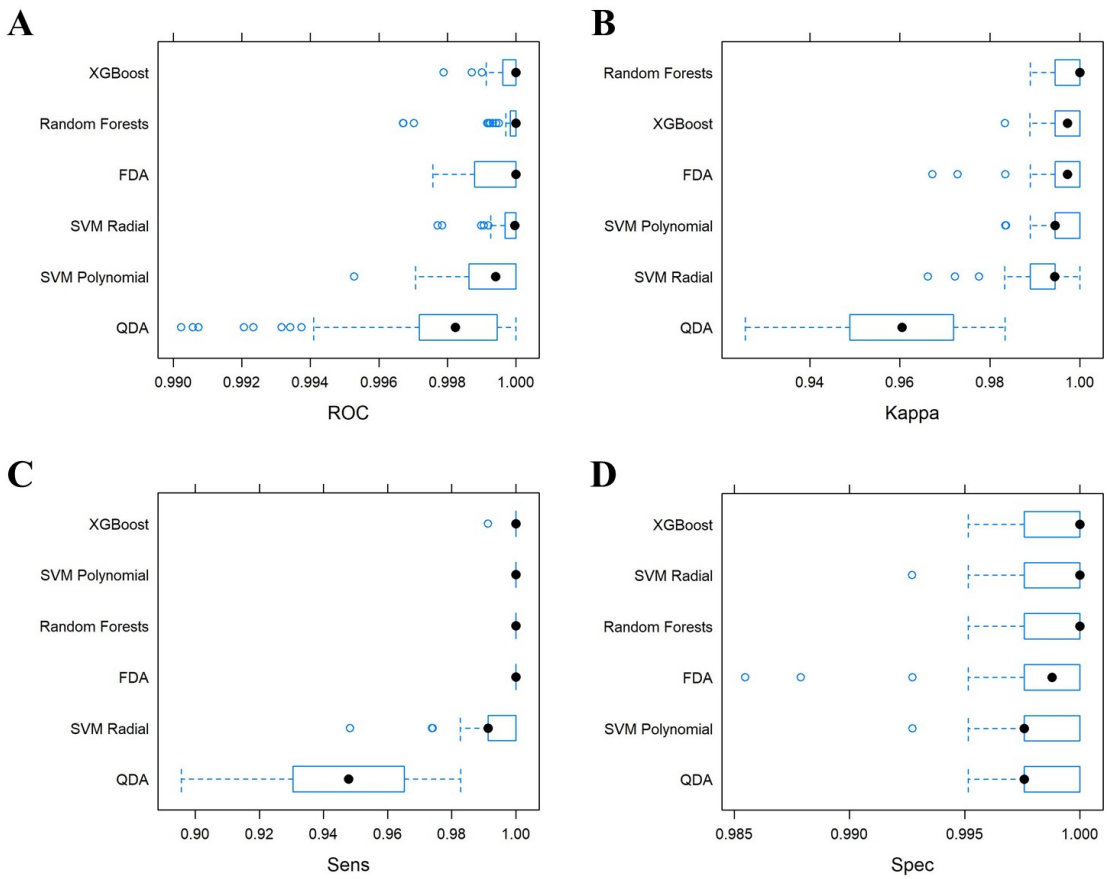

Figure 6.2: Performance estimation of the models trained to identify *Miniopterus natalensis* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

Since the AUC, sensitivity and specificity are quite similar between models, FDA, RF and XGBoost are preferred given that they are less computationally intensive than the SVM models. From these RF will be selected for future use since it has less tunning parameters relatively to XGBoost and, uses more information than FDA that only resources to Fmin and Fk to identify *Mnat* pulses (Table 6.5). If we inspect the most important variables in the RF and XGBoost, the variables associated with frequency are found (Table 6.5). Already in Section 4.2 from Chapter 4 we observed that for Fc, Fk and Fmin it is possibly to separate *Mnat* from the remaining species, except for *Mtri* that has an *outlier* that overlaps with *Mnat* (Figure 4.2).

6.3 *Neoromicia capensis*

Concerning the models to identify *Ncap* pulses, examination of the sampling distributions, from the cross-validation data, suggests high performance of the trained models (Figure 6.3), with few

Table 6.4: The confusion matrix and some performance metrics obtained for each model trained to identify *Miniopterus natalensis* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	248	1119	0	11	0.98	0.96	1.00	0.97	0.81
FDA	259	1119	0	0	1.00	1.00	1.00	1.00	0.81
Random Forests	259	1119	0	0	1.00	1.00	1.00	1.00	0.81
XGBoost	259	1119	0	0	1.00	1.00	1.00	1.00	0.81
SVM Radial	259	1119	0	0	1.00	1.00	1.00	1.00	0.81
SVM Polynomial	259	1119	0	0	1.00	1.00	1.00	1.00	0.81

Table 6.5: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Miniopterus natalensis* pulses.

RF		XGBoost		FDA	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
Fk	29.1	Fmean	35.6	Fmin	100.0
Fc	28.2	Fc	25.0	Fk	26.2
Fmin	27.9	Fk	10.8		
Fmean	18.8	Fmin	9.7		
Fmax	12.8	Fmax	6.2		
Pmc	11.6	S1	2.9		
Tc	11.2	Tk	2.6		
Sc	10.9	Qk	1.9		
Dur	10.1	Sc	1.6		
Tk	9.0	Pmc	1.2		
Qk	8.8	Dc	1.0		
Dc	8.7	Tbc	0.9		
Tbc	7.8	Tc	0.4		
S1	5.8	Dur	0.2		

differences between models (Table A.8). The QDA model contrasts by showing lower performance and higher variability.

When used to identify *Ncap* pulses in the test set, the models performed very well (Table 6.6) with the exception of QDA, which ends up confirming the results from the cross-validated data Figure 6.3. Once more is important to point out that the results obtained are better than a random guess, since the balanced accuracy is higher than the no information rate.

Further inspection of the predictions shows the FP in the XGBoost and RF models originate only from *Ehot* pulses, while in FDA they also derive from *Mtri*, the SVM Polynomial model adds *Taeg* to the list of FP Table 6.7. *Ncap*, *Ehot* and *Mtri* all belong to the *Vespertilionidae* and share some traits, thus these results are not surprising.

Considering the universe of FP, XGBoost and RF are less prone to confound other species with *Ncap*. However, the FDA trained model may allow for better detection of *Ncap* pulses, without risking too many FP. The performance will depend on the local abundance of *Mtri* bats in the scenarios where

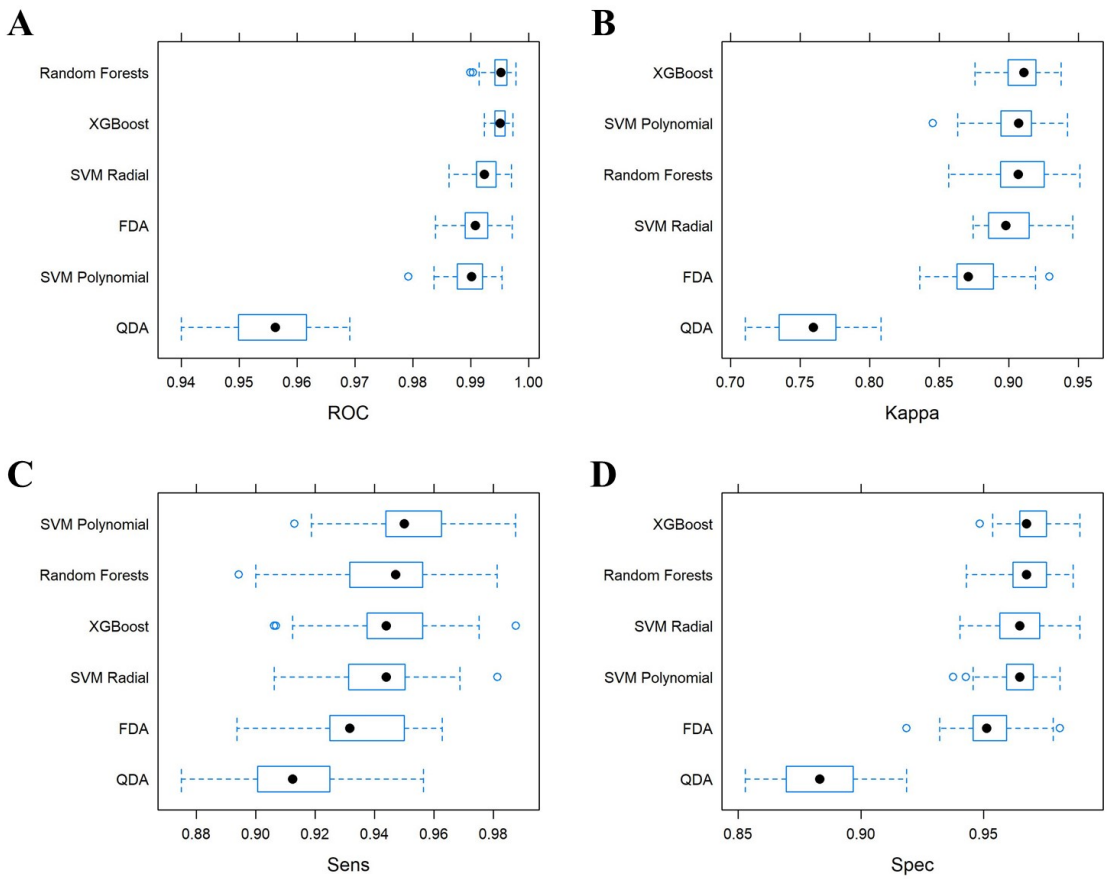

Figure 6.3: Performance estimation of the models trained to identify *Neoromicia capensis* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

the model will be used. The most important variables to identify *Ncap* pulses are similar in the RF, XGBoost and FDA models. These variables are mostly associated with frequency, specifically Fc and Fk are shared by the three models (Figure 6.8).

Table 6.6: The confusion matrix and some performance metrics obtained for each model trained to identify *Neoromicia capensis* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	390	735	232	21	0.85	0.95	0.76	0.62	0.70
FDA	403	923	44	8	0.97	0.98	0.95	0.91	0.70
Random Forests	394	926	41	17	0.96	0.96	0.96	0.90	0.70
XGBoost	388	921	46	23	0.95	0.94	0.95	0.88	0.70
SVM Radial	400	920	47	11	0.96	0.97	0.95	0.90	0.70
SVM Polynomial	405	922	45	6	0.97	0.98	0.95	0.91	0.70

Table 6.7: Summary of the errors made by the models trained to identify *Neoromicia nanus* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	0	0	-	229	0	0	3	0	0	0	21
FDA	0	0	-	40	0	0	4	0	0	0	8
Random Forests	0	0	-	41	0	0	0	0	0	0	17
XGBoost	0	0	-	46	0	0	0	0	0	0	23
SVM Radial	0	0	-	43	0	1	3	0	0	0	11
SVM Polynomial	2	0	-	40	0	0	3	0	0	0	6

Table 6.8: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Neoromicia capensis* pulses.

RF		XGBoost		FDA	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
Fc	61.0	Fmin	40.5	Fc	100.0
Fmin	48.1	Fc	25.1	Fk	45.5
Fk	43.5	Fk	18.0	Sc	13.8
Fmean	38.1	Sc	3.6	Qk	4.6
Sc	37.8	Qk	2.0		
Qk	22.1	Fmean	2.0		
Pmc	20.2	Tbc	1.6		
Tk	20.2	Dc	1.3		
Fmax	19.7	Tk	1.3		
Dc	19.2	Tc	1.1		
Tbc	19.0	Dur	1.0		
Tc	18.9	S1	1.0		
Dur	17.0	Pmc	0.9		
S1	14.7	Fmax	0.7		

6.4 *Eptesicus hottentotus*

Focusing now on the models trained to identify *Ehot* pulses, the sampling distributions of the performance metrics taken from the cross-validation suggest high performance for all models (Figure 6.4). Yet, no striking differences are observed on the CV statistics (Table A.9). Still, similarly to what has been observed on the QDA models, the performance is not as good, and high variability is shown (Figure 6.4).

The trained models were challenged on the test set Table 6.9. We are particularly interested in models that apart from displaying high (balanced) accuracy, also balance well the sensitivity and specificity. Indeed, all models, except QDA, satisfy this demand Table 6.9. Additionally, the balanced accuracy exceeds the no information ration, thus the prediction obtained are more than a random guess.

Analysis of the identification errors, in the test set, shows that FP come from *Ncap* and *Spet* in RF and XGBoost, while in SVM radial and FDA may also identify *Cpum* pulses as *Ehot*. Finally, SVM polynomial confounds *Taeg* pulses with *Ncap* (Table 6.10). RF model is more conservative compared

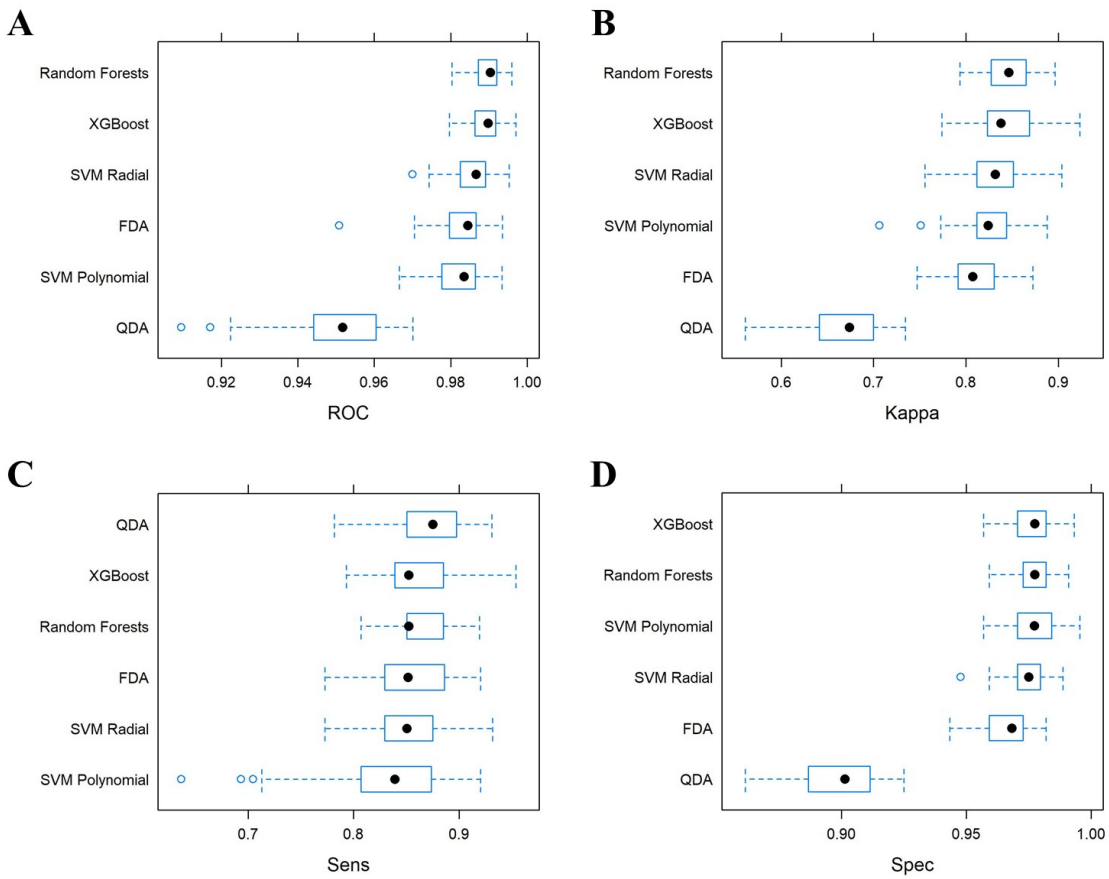

Figure 6.4: Performance estimation of the models trained to identify *Eptesicus hottentotus* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

to FDA, but the latter allows to identify a larger amount of *Ehot* pulses. Additionally, the FDA model displays a better balance between FP with FN. This may indicate that having more variables/information is not always beneficial, specifically if we want to differentiate *Ncap* from *Mnat* pulses. Furthermore, focusing on the errors table 6.7, in future uses of these models is necessary to gather information on the species expected to be present in the area. Mainly because The FDA and RF models seem to have different sensibilities to *Cpum* and *Spet* pulses. Interestingly, the variables that most contribute to the classification of the pulses are somehow similar between the RF and FDA models (Table 6.11).

6.5 *Chaerephon pumilus*

With *Cpum* we now enter a second phase, the analysis of models used to identify pulses from the rarer species. As previously mentioned *Cpum* pulses account for only 3% of the total bat pulses in our database (Table 4.3). Therefore, there is the risk that this information may be insufficient to train a model that will generalize properly for new data. In the cross-validation data, models trained to

Table 6.9: The confusion matrix and some performance metrics obtained for each model trained to identify *Eptesicus hottentotus* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	315	882	153	28	0.88	0.92	0.85	0.69	0.75
FDA	309	1015	20	34	0.94	0.90	0.98	0.89	0.75
Random Forests	298	1011	24	45	0.92	0.87	0.98	0.86	0.75
XGBoost	287	1011	24	56	0.91	0.84	0.98	0.84	0.75
SVM Radial	284	1024	11	59	0.91	0.83	0.99	0.86	0.75
SVM Polynomial	298	1017	18	45	0.93	0.87	0.98	0.87	0.75

Table 6.10: Summary of the errors made by the models trained to identify *Eptesicus hottentotus* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	0	0	150	-	0	3	0	0	0	0	28
FDA	0	0	7	-	1	12	0	0	0	0	34
Random Forests	0	0	17	-	0	7	0	0	0	0	45
XGBoost	0	0	17	-	0	7	0	0	0	0	56
SVM Radial	0	0	6	-	1	4	0	0	0	0	59
SVM Polynomial	2	0	8	-	1	6	0	1	0	0	45

Table 6.11: Variable importance for the Random Forest and FDA models, used to identify *Eptesicus hottentotus* pulses.

RF

Variable	**Importance**
Fc	55.8
Fmin	48.5
Fk	43.2
Fmean	35.4
Sc	31.2
Tk	23.6
Fmax	23.0
Tc	21.4
Qk	21.2
Dur	20.9
Pmc	19.3
Tbc	17.3
Dc	17.3
S1	15.8

FDA

Variable	**Importance**
Fc	100.0
Fmean	59.1
Sc	13.8
Fk	7.5

identify *Cpum* pulses display high values of AUC, except for SVM Radial Figure 6.5[A]. However, given the reduced number of *Cpum* pulses in the data, the focus must shift to models that better balance specificity and sensitivity. Models RF, XGBoost, QDA and FDA which present high AUC, have average values of sensitivity Figure 6.5[C], hence these models may be more prone to originate FN, the

same to say that they have less capacity to identify *Cpum* pulses. Conversely, the specificity is high, so few FP are expected Figure 6.5[D].

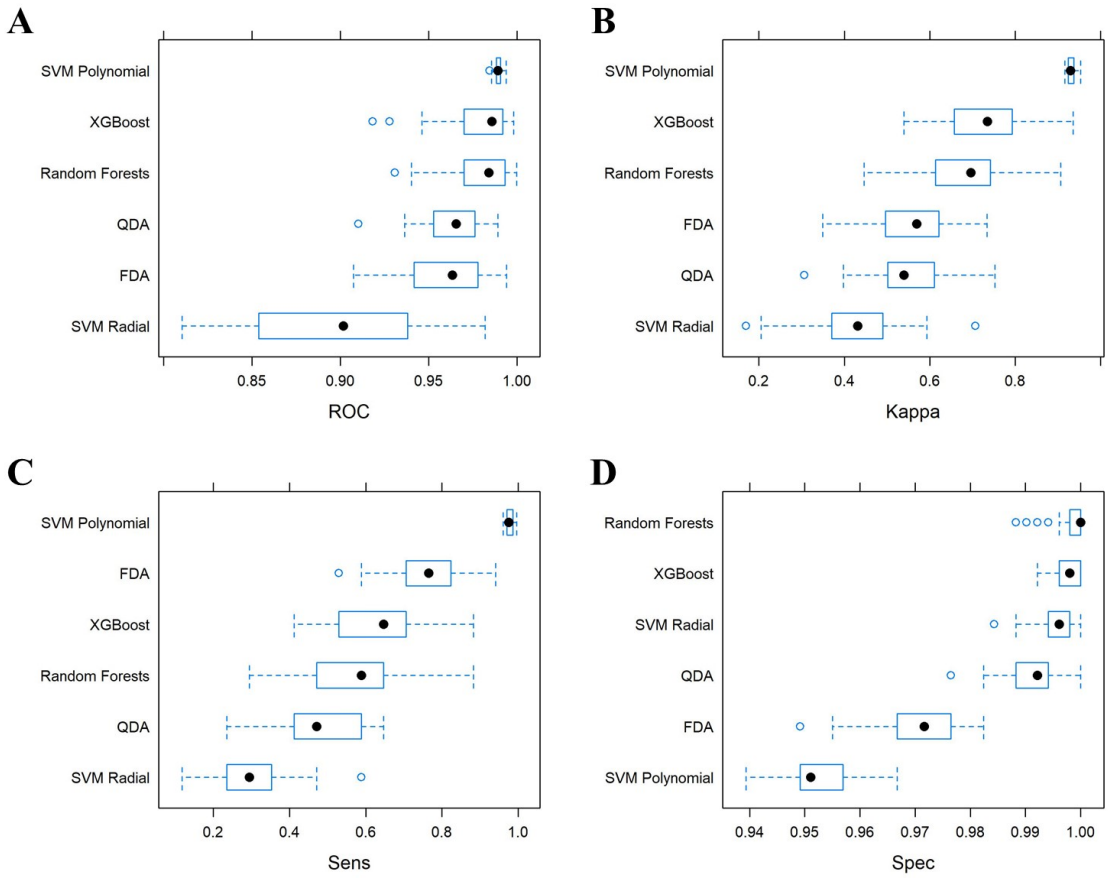

Figure 6.5: Performance estimation of the models trained to identify *Chaerephon pumilus* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

When the different models were used on the test set (Table 6.12), to identify *Cpum* pulses, some interesting results were observed. Firstly, the SVM Polynomial model did not generalize well for the new data, the 0.62 balanced accuracy is quite deceiving. The best accuracy was obtained with the FDA model, 0.90, followed by RF, 0.85 and XGBoost with 0.83. These models balance well the specificity and sensitivity, with FDA taking the lead, respectively, 1 and 0.81 (Table 6.12). These models need a careful analysis since the small number of *Cpum* pulses, particularly in the test set, may lead to considerable oscillations in the sensitivity. We need to be extra careful since the no information rate exceeds the accuracy.

Inspection of the errors reveals that the sole FP of the FDA model originates from *Rhin*. Generally, the models don't show many FP, their difficulty is really the identification of *Cpum* pulses. In order to increase the detectability of *Cpum* pulses, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was used to up-sample the *Cpum* pulses. A balanced train set was obtained synthesizing new pulses based on *Cpum*'s features on the original training

set alone, while the test set remains untouched. The models trained with up-sampling registered an increase in the sensitivity, compared with the model based on the original set, namely in the QDA, FDA and SVM polynomial (Table 6.12). Unfortunately, the increase in sensitivity comes at the expense of a high increase in FP. In the up-sampling models, the FP originate from *Taeg*, *Ehot* and *Spet* in the QDA model, and in FDA from *Taeg*, *Ncap*, *Spet* and *Rhin*, while in the SVM models they originate also from *Ncap* (Table 6.13). *Taeg*, *Cpum* and *Spet* are from the *Molossidae*, thus some of the FP observed are not totally unexpected.

The analysis of the variables' importance revealed a different pattern to what had been observed until now for the other species, the key variables were not only related to frequency but also to duration and slope (Table 6.14).

Table 6.12: The confusion matrix and some performance metrics obtained for each model trained to identify *Chaerephon pumilus* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	20	1320	5	23	0.78	0.57	1.00	0.67	0.96
FDA	43	1324	1	10	0.90	0.81	1.00	0.88	0.96
Random Forests	37	1325	0	16	0.85	0.70	1.00	0.82	0.96
XGBoost	35	1324	1	18	0.83	0.66	1.00	0.78	0.96
SVM Polynomial	13	1321	4	40	0.62	0.24	1.00	0.36	0.96
SVM Radial	20	1321	4	33	0.69	0.38	1.00	0.51	0.96
FDA UP	53	1288	37	0	0.99	1.00	0.97	0.73	0.96
QDA UP	52	1288	37	1	0.98	0.98	0.97	0.72	0.96
Random Forests UP	38	1323	2	15	0.86	0.72	1.00	0.81	0.96
SVM Radial UP	6	1321	4	47	0.56	0.11	1.00	0.18	0.96
SVM Polynomial UP	48	1291	34	5	0.94	0.91	0.97	0.70	0.96

Table 6.13: Summary of the errors made by the models trained to identify *Chaerephon pumilus* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	5	0	0	0	-	0	0	0	0	0	23
FDA	0	0	0	0	-	0	0	1	0	0	10
Random Forests	0	0	0	0	-	0	0	0	0	0	16
XGBoost	0	0	0	1	-	0	0	0	0	0	18
SVM Radial	0	0	0	1	-	3	0	0	0	0	33
SVM Polynomial	3	0	0	0	-	0	0	1	0	0	40
QDA UP	13	0	0	23	-	1	0	0	0	0	1
FDA UP	8	0	6	0	-	5	0	1	0	0	0
Random Forests UP	0	0	0	1	-	1	0	0	0	0	15
SVM Radial UP	0	0	0	4	-	3	0	0	0	0	47
SVM Polynomial UP	8	0	0	19	-	7	0	1	0	0	5

Table 6.14: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Chaerephon pumilus* pulses.

RF		XGBoost		FDA	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
Sc	52.8	Fc	19.9	Fc	100.0
Fc	49.6	Fmin	17.8	Fmin	66.6
Fmin	48.3	Dur	9.4	Fmax	66.6
Pmc	39.1	Sc	7.8	Dur	66.6
Qk	38.8	Fmean	6.5	Fmean	66.6
Dur	36.7	S1	5.9	Sc	42.6
Tbc	36.6	Fk	5.9	Fk	16.5
S1	35.7	Qk	5.7		
Tk	34.8	Fmax	4.4		
Fmean	33.1	Tbc	4.0		
Fk	32.4	Dc	3.6		
Tc	29.7	Pmc	3.4		
Fmax	28.1	Tk	3.3		
Dc	26.2	Tc	2.3		

6.6 *Sauromys petrophilus*, *Myotis tricolor* and *Neoromicia nanus*

Finally we are left with 3 species: *Spet*, *Mtri* and *Nnan*. As previously mentioned, these are very poorly represented in the dataset (Table 4.3). The outcomes, concerning pulses' identification, for the different species were quite distinct, while for *Mtri* and *Nnan* some models are able to perfectly classify all pulses, with *Spet* models the results turned out to be deceiving since the models fail to identify a large portion of the *Spet* pulses (Table 6.15). The FP originate almost exclusively from *Taeg* pulses, as has been mentioned previously these species are part of the *Molossidae* family, the pulses from these species are very similar (Lamb et al., 2011).

Similarly to what has been done for the identification of *Cpum*, SMOTE was used to try to boost the capacity of the models to identify *Spet* pulses. Only the FDA model was able to increase the number of correctly identified *Spet* pulses, however this comes at the expense of doubling the amount of FP (Table 6.15). The performance of the *Spet* is unsatisfactory due to the high number of FN, suggesting that more samples are needed to train the models if we want to be able to identify this specie in the future.

Concerning *Mtri* and *Nnan*, even tough some models were able to correctly predict all their pulses in the test set, one must be careful when using these models in the future since the models were trained with very few information. Proper cross-validation was used for the training, still these results may have been just fortuitous. Therefore, these models need to be reevaluated when more samples are available for *Mtri* and *Nnan*.

On the subject of the most important variables to identify the pulses of these species, we see that frequency-related variables are key to identify *Spet* and *Nnan* (Table 6.21 and 6.23), while for *Mtri* slope assumes the leading role (Table 6.22).

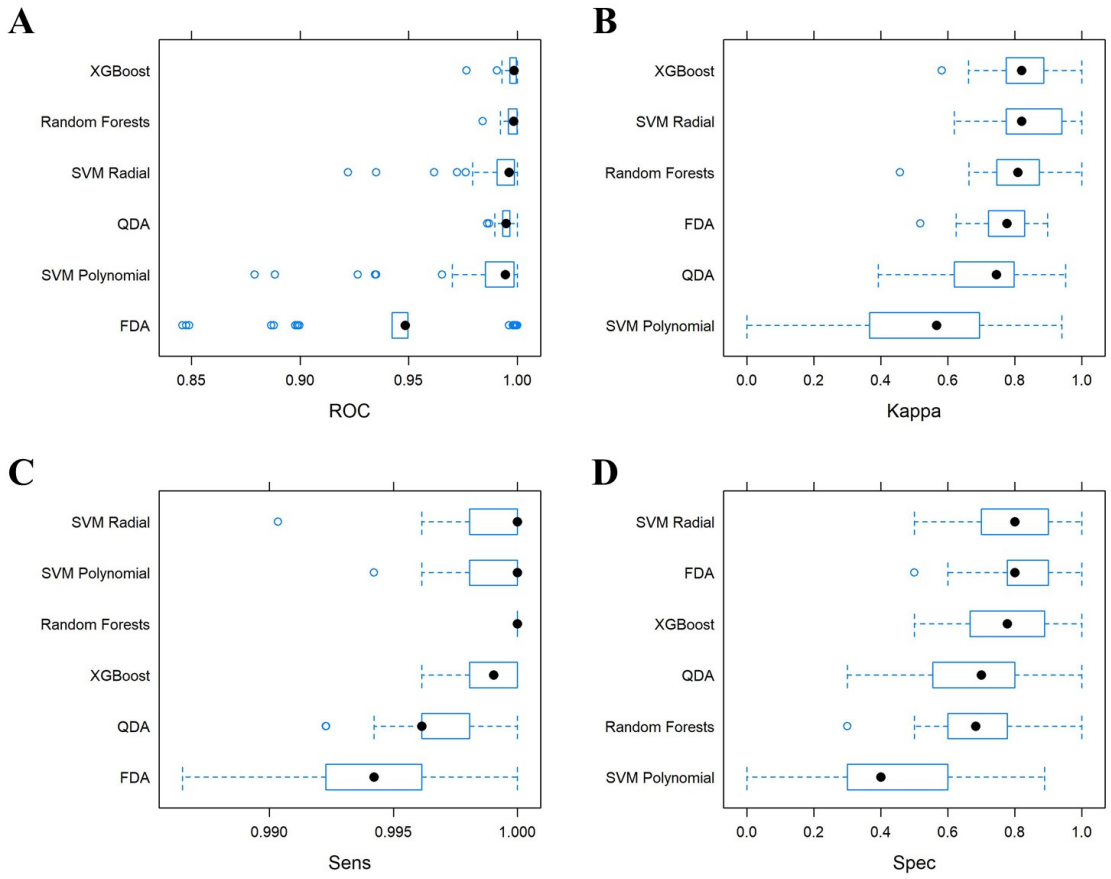

Figure 6.6: Performance estimation of the models trained to identify *Sauromys petrophilus* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

Table 6.15: The confusion matrix and some performance metrics obtained for each model trained to identify *Sauromys petrophilus* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	2	1357	3	16	0.55	0.11	1.00	0.17	0.99
FDA	6	1355	5	12	0.66	0.33	1.00	0.41	0.99
Random Forests	4	1359	1	14	0.61	0.22	1.00	0.34	0.99
XGBoost	4	1358	2	14	0.61	0.22	1.00	0.33	0.99
SVM Radial	6	1357	3	12	0.67	0.33	1.00	0.44	0.99
SVM Polynomial	4	1360	0	14	0.61	0.22	1.00	0.36	0.99
QDA UP	6	1358	2	12	0.67	0.33	1.00	0.46	0.99
FDA UP	12	1347	13	6	0.83	0.67	0.99	0.55	0.99
Random Forests UP	5	1359	1	13	0.64	0.28	1.00	0.41	0.99
SVM Polynomial UP	6	1355	5	12	0.66	0.33	1.00	0.41	0.99

Table 6.16: Summary of the errors made by the models trained to identify *Sauromys petrophilus* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	1	0	0	0	0	-	2	0	0	0	16
FDA	5	0	0	0	0	-	0	0	0	0	12
Random Forests	1	0	0	0	0	-	0	0	0	0	14
XGBoost	2	0	0	0	0	-	0	0	0	0	14
SVM Radial	3	0	0	0	0	-	0	0	0	0	12
SVM Polynomial	0	0	0	0	0	-	0	0	0	0	14
QDA UP	1	0	0	0	1	-	0	0	0	0	12
FDA UP	7	0	0	1	5	-	0	0	0	0	6
Random Forests UP	1	0	0	0	0	-	0	0	0	0	13
SVM Polynomial UP	5	0	0	0	0	-	0	0	0	0	12

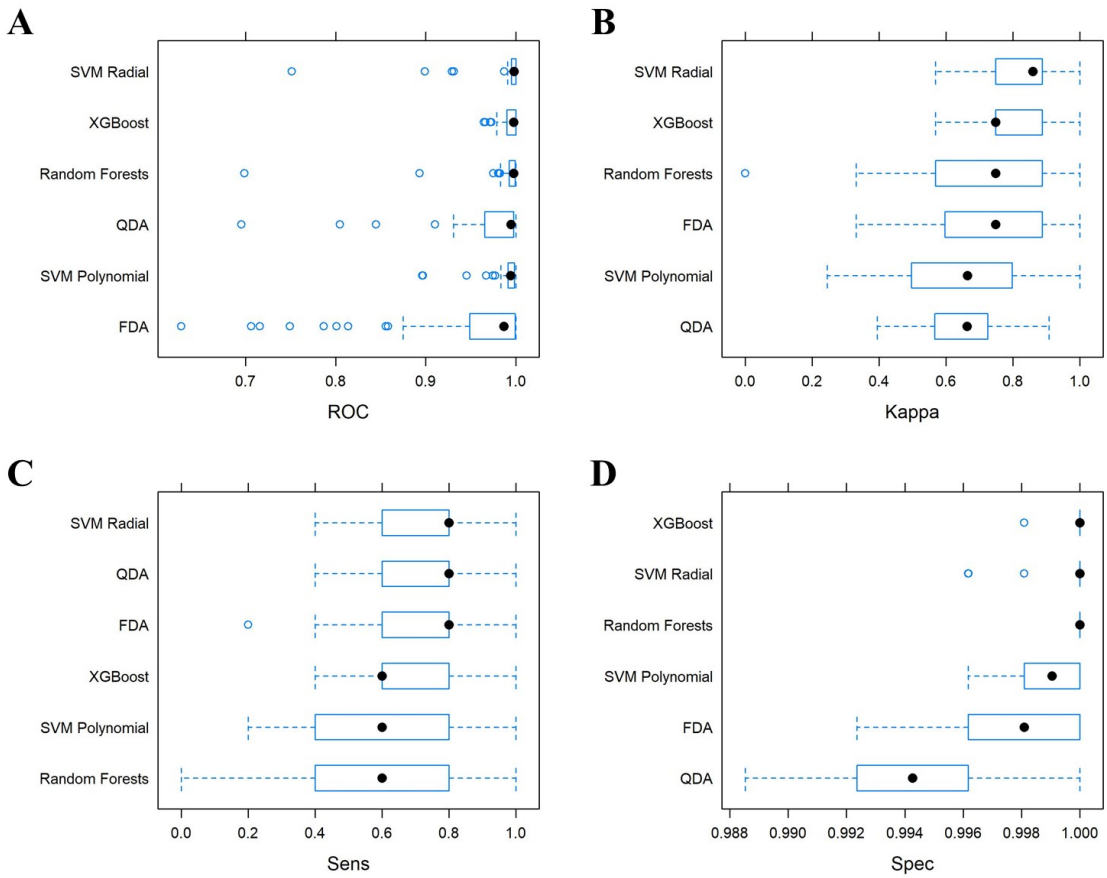

Figure 6.7: Performance estimation of the models trained to identify *Myotis tricolor* pulses. The performance measures used were: A) Area under the curve, B) Cohen's kappa, C) Sensitivity and D) Specificity.

Table 6.17: The confusion matrix and some performance metrics obtained for each model trained to identify *Myotis tricolor* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
QDA	5	1361	9	3	0.81	0.62	0.99	0.45	0.99
FDA	5	1367	3	3	0.81	0.62	1.00	0.62	0.99
Random Forests	8	1370	0	0	1.00	1.00	1.00	1.00	0.99
XGBoost	8	1370	0	0	1.00	1.00	1.00	1.00	0.99
SVM Radial	8	1367	3	0	1.00	1.00	1.00	0.84	0.99
SVM Polynomial	5	1369	1	3	0.81	0.62	1.00	0.71	0.99
QDA UP	8	1369	1	0	1.00	1.00	1.00	0.94	0.99
FDA UP	8	1337	33	0	0.99	1.00	0.98	0.32	0.99
Random Forests UP	8	1369	1	0	1.00	1.00	1.00	0.94	0.99
SVM Polynomial UP	8	1355	15	0	0.99	1.00	0.99	0.51	0.99

Table 6.18: Summary of the errors made by the models trained to identify *Myotis tricolor* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
QDA	1	0	8	0	0	0	-	0	0	0	3
FDA	2	0	1	0	0	0	-	0	0	0	3
Random Forests	0	0	0	0	0	0	-	0	0	0	0
XGBoost	0	0	0	0	0	0	-	0	0	0	0
SVM Radial	0	2	1	0	0	0	-	0	0	0	0
SVM Polynomial	0	0	1	0	0	0	-	0	0	0	3
QDA UP	0	0	1	0	0	0	-	0	0	0	0
FDA UP	1	15	17	0	0	0	-	0	0	0	0
Random Forests UP	0	0	1	0	0	0	-	0	0	0	0
SVM Polynomial UP	0	11	4	0	0	0	-	0	0	0	0

Table 6.19: The confusion matrix and some performance metrics obtained for each model trained to identify *Neoromicia nanus* pulses when applied on the test set. NIR, No information rate

Model	TP	TN	FP	FN	B. Acc	Sens	Spec	Kappa	NIR
FDA	5	1373	0	0	1.00	1.00	1.00	1.00	1.00
Random Forests	3	1373	0	2	0.80	0.60	1.00	0.75	1.00
SVM Radial	3	1373	0	2	0.80	0.60	1.00	0.75	1.00
SVM Polynomial	3	1373	0	2	0.80	0.60	1.00	0.75	1.00
FDA UP	5	1373	0	0	1.00	1.00	1.00	1.00	1.00
SVM Polynomial up	4	1373	0	1	0.90	0.80	1.00	0.89	1.00
Random Forests UP	3	1373	0	2	0.80	0.60	1.00	0.75	1.00

6.7 Models in Practice - Test Set

The main goal of this work was to identify bat species from their echolocation sounds. To achieve this we need to establish an order to apply the models we have just selected for each species. The

Table 6.20: Summary of the errors made by the models trained to identify *Neoromicia nanus* pulses, when applied to the test set.

Model	Taeg	Mnat	Ncap	Ehot	Cpum	Spet	Mtri	Rhin	Nnan	Other	FN
FDA	0	0	0	0	0	0	0	0	-	0	0
Random Forests	0	0	0	0	0	0	0	0	-	0	2
SVM Radial	0	0	0	0	0	0	0	0	-	0	2
SVM Polynomial	0	0	0	0	0	0	0	0	-	0	2
FDA UP	0	0	0	0	0	0	0	0	-	0	0
Random Forests UP	0	0	0	0	0	0	0	0	-	0	2
SVM Polynomial UP	0	0	0	0	0	0	0	0	-	0	1

Table 6.21: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Sauromys petrophilus* pulses.

RF

Variable	**Importance**
Fmin	30.8
Fc	26.6
Fk	24.4
Pmc	22.8
Fmean	22.3
Fmax	20.2
Sc	19.2
Qk	18.8
Dc	15.4
S1	15.2
Tk	12.7
Tbc	10.7
Tc	7.7
Dur	7.0

XGBoost

Variable	**Importance**
Fmean	24.0
Fmin	17.6
Fmax	13.3
Fc	12.2
Fk	11.8
Pmc	5.8
Sc	3.5
S1	2.8
Dc	2.6
Qk	2.1
Tbc	1.9
Dur	1.0
Tk	0.8
Tc	0.5

FDA

Variable	**Importance**
Fmean	100.0
Fmin	100.0
Pmc	47.6
Sc	6.1

rationale behind this strategy is to first apply the most accurate models, at the same time balancing the information about the origin of the FP to decrease misclassification. In this step, the test set is used as an experimental ground to calibrate the models' sequence.

The models trained to identify *Mnat* pulses presented an excellent performance in the test set, among these the RF-trained model seems to be the best option for future use. Therefore, Mnat-RF model will be the first in the sequence. RF model to identify *Taeg* pulses produce enthusiastic results with almost no FP and a good ability to identify *Taeg* pulses. Moreover, the FP derive only from *Cpum* (Table 6.2), thus a model to identify *Cpum* pulses must be used before proceeding to identify *Taeg*. Concentrating now on *Cpum* models, we have seen that a FDA-trained model offered the best performance on the test set (Table 6.12). This model can give rise to FP from *Rhinolophus* pulses, but this is not of great concern since species from this genus are quite rare, difficult to capture on record and surprisingly easy to identify without the need of classificatory models due to the distinctive shape of its echolocation pulses.

Table 6.22: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Myotis tricolor* pulses.

RF		XGBoost		FDA	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
Sc	31.0	Sc	41.1	Sc	100.0
Fmean	20.8	Fmean	12.1	Dc	29.4
Pmc	18.5	Fmin	7.2	Tc	21.3
Qk	16.8	Dc	6.6	Qk	9.1
Fk	16.7	Tc	6.0		
Fmax	15.5	Dur	5.7		
Dc	14.7	Tbc	4.2		
Fmin	13.5	Fk	3.9		
S1	12.5	Qk	3.4		
Tbc	12.3	Pmc	2.7		
Dur	12.1	Tk	2.7		
Fc	12.0	Fc	2.5		
Tc	11.9	Fmax	1.1		
Tk	11.8	S1	0.8		

Table 6.23: Variable importance for the Random Forest, XGBoost and FDA models, used to identify *Neoromicia nanus* pulses.

RF		FDA	
Variable	**Importance**	**Variable**	**Importance**
Fk	17.6	Fmin	100.0
Fmin	15.7	Fc	89.5
Fc	14.7		
Fmean	13.2		
Fmax	10.7		
Pmc	6.0		
Sc	5.0		
Tc	4.4		
Qk	4.1		
S1	4.0		
Tk	2.8		
Tbc	2.6		
Dc	2.5		
Dur	2.5		

Given the similarity between *Taeg*, *Cpum* and *Spet*, next we ponder when to use the *Spet* model. Indeed the models to detect *Spet* pulses, namely RF and FDA, generate false positives from *Taeg* pulses, therefore it is advisable that they are applied after *Taeg*. The model build using upsampling can be an alternative, there is however the inconvenient that it generates FP from *Taeg*, *Ehot* and *Cpum* pulses. Nevertheless, if the models to detect these species are applied before the occurrence of errors should

be reduced.

Finally we are left with *Ehot* and *Ncap*, which can be mistaken with each other. A FDA-trained model will be used to identify *Ncap*, since it has a better capacity according to what was obtained in the test set *Ncap* (Table 6.6). To identify *Ehot* pulses a FDA model will also be used, as it has the best balance between accuracy, FP and FN (Table 6.9). Still one needs to be careful concerning *Cpum* and *Spet* since FP can arise from pulses of these species (Table 6.10).

At the end the models' sequence is as follows:

1. *Mnat* RF
2. *Ehot* FDA
3. *Ncap* FDA
4. *Cpum* FDA
5. *Taeg* RF
6. *Spet* FDA

The results of the sequential application of the models resulted in the confusion matrix (Table 6.24) with an accuracy of 93.2%. In recordings this is translated to Table 6.25 with an accuracy of 95%.

Table 6.24: Confusion matrix obtained when the selected models were applied in the order defined in 6.7, to the test set.

		Predicted					
		Cpum	Ehot	Mnat	Ncap	Other	Taeg
Observed	Cpum	43	0	0	0	1	0
	Ehot	1	309	0	7	12	0
	Mnat	0	0	259	0	0	0
	Ncap	0	33	0	403	4	0
	Other	0	1	0	1	14	0
	Spet	0	0	0	0	5	1
	Taeg	9	0	0	0	0	275

6.8 Models in Practice - Real World Application

Finally, the models were tested in a "real" scenario. The sequence of application defined previously was used on a set of 216 recordings (Table 6.26) from a wind farm in the Wild Coast region of the Eastern Cape province, in South Africa. Firstly, potential noise and pulses that don't have enough quality were filtered out, since models were not trained with this kind of information it could give rise to unexpected effects/outcomes. Furthermore, this represents a good test on the rules that were established to remove noise.

Table 6.25: Confusion matrix, of the recordings, obtained when the selected models were applied in the order defined in 6.7, to the test set.

		Predicted						
		Cpum	Ehot	Mnat	Ncap	Other	Spet	Taeg
Observed	Cpum	7	0	0	0	1	0	0
	Ehot	0	18	0	1	0	1	0
	Mnat	0	0	37	0	0	0	0
	Ncap	0	1	0	31	0	0	0
	Other	0	1	0	1	6	0	0
	Spet	0	0	0	0	0	1	0
	Taeg	1	0	0	0	0	0	40

Table 6.26: Real data set. The "?" indicates that the bat specialist was not totally sure about the identity of the species.

Specie	Number of recordings
Noise	62
Taeg	55
Ncap	39
Ncap?	25
Mnat	14
Ncap or Mnat	8
Taeg?	5
Mnat?	2
Molossidae	2
Ncap and Mnat?	2
Mnat? or Ncap?	1
Molossidae or Mtri	1

The result of the use of the models to this set of recordings is resumed on (Table 6.27, and allowed to correctly identify 92% of the recordings. There were however some unexpected results, such as the *Ncap* model (FDA) wrongly classifying *Cpum* and *Mnat* pulses as *Ncap*. *Cpum* and *Mnat* confusion with *Ncap* also comes a bit unexpected, since the characteristics of these two species are quite distinct. Nevertheless, in the box plots (Figure 4.2) from Section 4.2 we can observe overlapping in quite a few variables(Figures 4.2, 4.3, 4.4), since the FDA model for *Ncap* relies on Fc, Fk, Sc and Qk.

Concerning the identification of *Taeg*, there were some FP resulting from *Ehot* pulses which were also not observed in the test with the test set. There were also a couple of recordings containing noise were identified pulses as *Taeg* and *Mnat*, reinforcing the need to remove noise before applying the models. These unexpected errors need to be taken into account in the future when working with large sets of recordings. This proceeding is capable of identifying around 35 000 recordings in about 5 minutes, and produces an output that can be edited and inspected by the user.

Table 6.27: Confusion matrix resulting from the identification of 216 recordings, from a wind farm in the Wild Coast region of the Eastern Cape province, in South Africa. The "?" indicates that the bat specialist was not totally sure about the identity of the species.

		Observed													
		Mnat	Mnat or Ncap	Mnat?	Mnat? and Ncap	Mnat? or Ncap?	Molossidae	Molossidae or Mtri	Ncap	Ncap and Mnat?	Ncap or Mnat	Ncap?	Noise	Taeg	Taeg?
Predicted	Cpum	0	0	0	0	0	0	0	0	0	0	1	0	0	2
	Cpum/Taeg	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Ehot	0	0	0	0	0	0	0	1	0	0	3	0	0	0
	Ehot/Cpum	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Ehot/Mnat	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	Ehot/Ncap	0	0	0	0	0	0	0	3	0	0	2	0	0	0
	Mnat	14	1	2	1	1	0	0	1	0	2	0	0	0	0
	Mnat/Ncap	0	0	0	0	0	0	0	6	1	5	0	0	0	0
	Mnat/Other	0	0	0	0	0	0	0	2	0	0	0	0	0	0
	Ncap	0	0	0	0	0	0	0	24	0	0	17	0	0	0
	Ncap/Ehot	0	0	0	0	0	0	0	0	0	0	2	0	0	0
	Noise	0	0	0	0	0	0	0	0	0	0	0	55	1	1
	Noise/Mnat	0	0	0	0	0	0	0	0	0	0	0	3	0	0
	Noise/Taeg	0	0	0	0	0	0	0	0	0	0	0	2	0	0
	Other	0	0	0	0	0	0	0	1	0	0	0	1	0	0
	Ehot/Spet	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	Mnat/Cpum	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Taeg	0	0	0	0	0	0	0	0	0	0	0	1	52	2
	Taeg/Spet	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Chapter 7

Discussion

Machine learning algorithms have shown successful identification of bat species from their echolocation calls. The methods used are quite diverse, such as artificial neural networks (Parsons & Jones, 2000; Preatoni et al., 2005; Armitage & Ober, 2010b; Redgwell et al., 2009), hidden Markov models (Skowronski & Harris, 2006), support vector machines (Armitage & Ober, 2010b; Redgwell et al., 2009), Random Forests (Armitage & Ober, 2010b; Zamora-Gutierrez et al., 2016) among others. These studies achieved high classification accuracies, even when the spectral and temporal characteristics of calls exhibit some degree of overlap. However, it is difficult to make a detailed side by side comparison of these studies, since different data and bat species are used. Still, in this chapter the most relevant aspects of the studies will be compared with the work here shown.

An important aspect of species identification with supervised machine learning, is to construct a proper representative library of bat pulses, covering all species of interest. Yet, it can prove to be a difficult task, especially in areas of high species richness. The database organized for this work contains only a small fraction of South Africa's bat species. Indeed, the main handicap of the work here presented was precisely the lack of bat pulses and/or recordings for some species. Fortunately, to assess the impact of wind farms on bats populations not all species need to be considered, as these are conditioned on whether the specie's habitat matches the area of the wind farm, its collision risk, menace status and on country or local regulations.

The composition of the database takes us to an important issue regarding supervised learning, the need for significant amounts of hand-labeled data to reliably train models. This should not constitute a major obstacle when working with bats for a long time, as large amounts of data can be gathered that with proper treatment these information can be assembled into a comprehensive database. Diversity is a key point here, since bats have an extreme flexibility in their echolocation pulses (Ahlen & Baagøe, 1999; Jones & Teeling, 2006; Fenton & Bell, 1981), thus reinforcing the need for specialist-labeled data to cover as much as possible the call variability of the species. Consequently, it is pivotal to enrich the database with calls, especially from the underrepresented species with moderate to high risk of collision, e.g. *Miniopterus fraterculus*, *Myotis tricolor*, *Neoromicia nanus* and *Taphozous mauritianus*. This will not solve all problems, since some species are difficult to identify, even for a well-trained specialist, but will be of great help to create more robust models. Though it is difficult to collect enough labeled recordings, for some species, strategies can be used to circumvent the problem. For example, to use models trained to identify these rare species, to detect new pulses that upon confirmation by a specialist may be used to retrain models. Given their scarcity, in a large project with a couple hundred thousand recordings, some positive samples may be found making the identification process more focused. Such approach is a long-term effort but that will pay out in the future.

Increasing the number of samples in training implies that more computational resources. Some models might take a long time to train, not only due to the number of samples, but also to find the right

tuning parameters. Still, models need only to be trained once, or eventually when enough new validated samples are included to the database. Furthermore, computers are getting more powerful by the day, and access to powerful cloud computing has been made easier and with increasingly affordable prices. Nonetheless, some measures can be adopted to help decrease computation requirements, namely to favor simpler models. As an example, in this work FDA models often revealed to be a good option to identify species, from the models here used FDA is one of the least computationally intensive methods. A simple model with enough quality data will often be a better option, since in the end the only important measure of a model is how well it predicts things, independently of the methods used. Besides, there is no need to hold on to a single model, as different models can be explored in varied contexts. The best example of this are the different studies using machine learning algorithms, to identify bat species, where good results have been obtained independent of the methods. To further limit computational demand, parallel computation can be used to make the most of the resources available. The *caret* package, used in this work to train the models, supports parallel computing. Additionally, and this is very important,it is necessary to gain as much information as possible about the area of study (Ahlen & Baagøe, 1999). No model can substitute this. In the present work, this was evident when cleaning the database from pulses with unexpected values, since having information on the pulses' characteristics help to filter out potential noise. Also, it is also key to have a good understanding of the species that inhabit, or pass through the areas of study, as this information helps to evaluate the results obtained by the models. However, not always is possible to gather detailed information about the terrain. Sometimes it can prove to be a difficult task, since regional species composition is dynamic and likely to change over time due to factors such as climate change (Lundy, Montgomery, & Russ, 2010).

There is no doubt bat calls are complex, since they not only have echolocation calls but also social calls (Jones & Siemers, 2011). Some bats display harmonics in their pulses, a feature that can sometimes be used to differentiate between species. The approach used in this work to identify bat species discards this information, because when the recordings are converted to zero-crossing only the most energetic pulse is retained (Parsons, 2001). This loss of information might result problematic if the number of harmonics is essential to distinguish or identify species. For instance, a species characterized by a high number of harmonics, without such information identification becomes more complex. Some studies such as Redgwell et al., 2009; Scott, 2012; Walters et al., 2012; Stathopoulos, Zamora-Gutierrez, Jones, and Girolami, 2014; Zamora-Gutierrez et al., 2016 have used methods that include the harmonics' information. In these, the spectral and temporal parameters are measured directly from the sonograms, thus allowing to extract more parameters, and increasing the amount of information. Commercial software is available to perform this, such as SonoBat used by Walters et al., 2012; Stathopoulos et al., 2014, which extracts up to 27 call parameters.

To have good models and prediction, non-pulses must be excluded, and also ensure that there is enough information to identify a certain species. Being able to exclude non-pulses is key either for model training and species' prediction. In his doctoral thesis Scott, 2012 developed an algorithm that locates echolocation calls in continuous recordings and subtracts the median noise signal to the complete extent of a recording. While the work of Stathopoulos et al., 2014 uses a different approach

to avoid noise, by capturing the bats and record them upon release. Bats are fragile and manipulation should be avoided, this method of recording and identification may be useful for specific proposes but not for large scale projects such as wind farm areas.

Here, like Armitage and Ober, 2010b, we considered as quality calls those where $Qual \leqslant 0.30$. As already mentioned, sonogram offer more information than the ZC format, thus being more advantageous when the signals are less intense. But does the inclusion of additional and more precise spectral parameters improve the results? Here one has to bear in mind the correlation between these new predictors, and variable selection will have to be performed (Zamora-Gutierrez et al., 2016). As more variables also implies more complex models, the complexity can be be controlled through regularization, as it constraints model fitting and reduces the effective degrees of freedom without reducing the actual number of parameters in the model. Regularization is a key aspect of flexible discriminant analysis (Clemmensen et al., 2011) which allowed to achieve good results in this work. Regularization is also important to deal with the high correlation observed between predictors in the data (Figure 4.5). Having highly correlated predictors means the same underlying information is being measured more than once. Effectively, FDA needs few variables to identify species, which may just be a direct consequence of the high correlation between variables. Decision-tree-based models are not susceptible to outliers and collinearity, hence the generally good results obtained with Random Forests.

On another note, here we could verify that regardless of the loss of harmonic information, the identifications carried out in this work were quite positive, so the loss of harmonic information and using the ZC format is not a major limitation. Interestingly, the different studies point to a particular set of key variables that match the ones we found here. Like in our study, Armitage and Ober, 2010b observed that frequency characteristics (Fmax, Fmin, Fk, Fc) were the most important in call identification, with Fc being the most differentiating (It occurs at the end of the flattest part of the echolocation call and shows very low intra-specific variation.). Other works such as Vaughan et al., 1997; Parsons, 2001; Russo and Jones, 2002; Redgwell et al., 2009 reached similar conclusions . More recently, Zamora-Gutierrez et al., 2016 also found Fc to have a key role in call identification. Thus, the use of ZC can actually lead to good results on bat species identification.

The methodology here used has the disadvantage of allowing error propagation, once a species is wrongly classified it will not be possible to revert such error as the models are applied sequentially. In the end, if a certain bat species is just too rare, or difficult to detect with field recorders, other methodologies must be used to determine its presence, e.g. such as bat capture.

Chapter 8

Conclusions

Economy expands to keep up with the growing needs of populations. Still, there is the need to comply with environmental regulations, even renewable energy infrastructures can have an impact on fauna and flora, as is the case of wind turbines. In this matter, environmental consultancy companies have a significant role. Among other tasks, these companies monitor species' abundance in the areas of the developments before, during and after their completion. In today's world companies and organizations need data and analytics to remain competitive. Coincidentally, more and more we are able to collect large amounts of information, although to use it to our advantage we must conduct a proper analysis. Here, our main focus is the identification of bat species from their echolocation calls, collected in large quantity by field recorders. Machine learning has greatly facilitated bat identification over the past years, since 1997 when for the first time bats were identified from their ultrasonic calls using a decision tree classification system (Herr et al., 1997).

Here we presented, to our best knowledge, the first study using supervised learning to identify South African bat species from their echolocation calls. The approach taken was very practical, resourcing to supervised learning algorithms already implemented in R, and data compiled by Bioinsight from projects in South Africa, we established a tool capable of identifying key bat species. These species include *Chaerephon pumilus*, *Eptesicus hottentotus*, *Miniopterus natalensis*, *Neoromicia capensis* and *Tadarida aegyptiaca*. For other species like *Sauromys petrophilus*, *Myotis tricolor* and *Neoromicia nanus*, we achieved promising identification models but that still need more information to turn them more robust and further testing.

Our identification tool, consisting of a system of models, achieved good results on the test set, correctly predicting the species in 95% of the recordings. The system was also able to correctly identify 92% in a set of recordings from a specific project, indicating it generalizes well for new data. Not only we were able to identify these species, but we did it in a faster and more systematic way, as compared with manual identification. Here a contribution is given to improve bat identification in South Africa, 35 000 records can be analyzed in just 5 minutes, and allowing to identify 6 or more bat species. The script creates an informative but simple output that can be modified by the user.

The present work will speed up the identification of bat species from South Africa by Bioinsight, and cut costs as there will be less need to send recordings to an external specialist to perform identifications (manual identification), these can be reserved just to special cases. The identification system developed is semi-automated, in which the user has still some degree of interaction. This is necessary to manage the database, retrain models if necessary, clean the data and control the quality of the predictions.

Here we can also conclude that by making use of well-known and R-implemented machine learning algorithms, it is possible to create a versatile tool to identify South African bats. Despite the loss of information resulting from the conversion of the records to zero-crossing, it can be said that the results here obtained were very positive, as this will allow saving a great deal of time. Additionally, this tool

can be further adapted to the needs of the user.

Current limitations of our approach include error propagation and uncertainty in respect to species that are not present in the database, and for which no models were trained.

Future work should focus on enriching the database, specially for the rarest species such as *Chaerephon pumilus*, *Sauromys petrophilus*, *Myotis tricolor* and *Neoromicia nanus*, *Miniopterus fraterculus* and *Taphozous mauritianus*. To this effect, models of rarer species can be used to search for pulses in the bulk data. This kind of work can be extended and adapted to other regions provided that a well-organized library of bat calls exists.

Appendix A

Tables Cross-Validation

A.1 Variable Summaries by Species

Table A.1: Summaries of the different variables, associated with the pulses, by species.

	Fmin						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	22.22	35.09	26.66	25.08	25.93	27.71	138
Ehot	24.69	35.71	33.44	32.79	33.47	34.19	781
Mfra	64.00	65.57	64.91	64.78	65.04	65.17	4
Mnat	45.20	55.17	48.54	47.06	47.90	49.08	836
Mtri	32.26	48.19	36.55	33.33	36.70	37.38	33
Ncap	31.13	42.78	36.24	35.40	36.20	36.87	1213
Nnan	63.49	69.57	66.76	64.65	66.97	69.12	12
Rhin	71.43	87.91	80.59	80.20	81.63	82.26	14
Spet	25.81	33.20	28.31	27.12	27.87	29.09	67
Taeg	18.06	28.57	21.48	20.30	21.39	22.35	874
Tmau	20.83	40.20	27.98	25.60	26.14	26.92	46

	Fmax						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	25.16	69.57	32.10	27.49	29.52	34.04	138
Ehot	28.67	83.33	47.96	40.61	45.98	54.42	781
Mfra	70.18	72.73	71.29	70.18	71.12	72.23	4
Mnat	47.62	117.65	61.86	55.56	59.26	66.12	836
Mtri	51.95	89.89	65.55	56.34	64.00	66.12	33
Ncap	35.56	103.90	54.94	47.34	53.69	60.61	1213
Nnan	70.80	83.33	76.11	73.39	74.77	78.62	12
Rhin	81.63	90.91	84.83	84.21	85.11	85.11	14
Spet	32.52	67.23	50.83	44.33	51.95	57.34	67
Taeg	19.28	49.69	29.62	24.84	27.49	34.93	874
Tmau	26.85	48.48	32.51	28.15	30.53	32.39	46

Table A.2: Summaries of the different variables, associated with the pulses, by species.

	Fc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	22.22	35.24	26.87	25.24	26.06	27.87	138
Ehot	26.32	35.87	33.59	32.92	33.61	34.33	781
Mfra	65.04	66.12	65.58	65.04	65.58	66.12	4
Mnat	45.20	55.17	48.79	47.34	48.19	49.38	836
Mtri	32.79	48.19	37.61	36.36	37.38	37.91	33
Ncap	33.47	42.78	36.46	35.71	36.36	37.21	1213
Nnan	63.49	69.57	66.81	64.65	67.25	69.12	12
Rhin	81.63	87.91	83.87	82.69	83.77	84.88	14
Spet	25.81	36.36	29.01	27.30	28.27	29.96	67
Taeg	18.18	29.74	21.63	20.41	21.51	22.54	874
Tmau	24.02	40.61	28.62	26.08	26.49	27.01	46

	Fk						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	23.95	40.40	28.70	26.51	27.87	29.38	138
Ehot	28.27	40.00	36.02	35.24	36.04	36.87	781
Mfra	65.04	66.67	66.26	66.26	66.67	66.67	4
Mnat	46.78	60.15	51.47	49.69	51.28	52.63	836
Mtri	39.22	51.28	42.71	40.82	42.33	43.72	33
Ncap	34.78	46.51	39.14	37.91	39.02	40.00	1213
Nnan	65.57	73.39	69.12	66.53	68.69	71.43	12
Rhin	80.00	87.91	83.81	82.47	83.77	85.11	14
Spet	27.49	38.10	31.60	29.85	31.13	33.20	67
Taeg	18.96	31.37	23.91	22.41	24.17	25.24	874
Tmau	26.67	45.20	30.90	27.42	27.97	29.28	46

	Pmc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	0.00	113.90	18.77	6.75	12.30	26.15	138
Ehot	3.10	151.00	42.85	20.30	36.20	61.60	781
Mfra	7.90	10.00	8.70	7.90	8.45	9.25	4
Mnat	1.20	126.00	26.81	14.50	21.70	34.20	836
Mtri	8.50	140.40	76.10	52.80	76.00	100.00	33
Ncap	1.30	189.70	50.65	29.50	47.20	66.40	1213
Nnan	6.40	21.80	13.98	11.57	15.15	15.75	12
Rhin	0.00	3.40	1.16	0.00	1.10	1.85	14
Spet	17.00	140.30	76.94	44.55	88.20	102.75	67
Taeg	0.00	155.60	36.68	14.00	25.50	57.30	874
Tmau	0.00	35.90	13.45	7.55	13.90	19.12	46

Table A.3: Summaries of the different variables, associated with the pulses, by species.

	Dur						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	2.11	16.20	7.62	4.37	7.47	10.21	138
Ehot	2.04	12.21	4.33	3.35	4.19	5.14	781
Mfra	2.09	3.68	2.71	2.19	2.54	3.06	4
Mnat	2.00	8.95	3.72	2.74	3.56	4.50	836
Mtri	2.14	4.23	3.34	2.62	3.59	3.88	33
Ncap	2.02	9.48	4.28	3.40	4.21	5.05	1213
Nnan	2.06	3.14	2.58	2.35	2.58	2.73	12
Rhin	2.27	27.42	13.82	2.46	14.82	22.99	14
Spet	2.36	10.02	5.70	3.83	6.20	7.23	67
Taeg	2.00	15.64	6.12	4.10	6.03	7.67	874
Tmau	2.02	8.31	4.56	2.81	4.03	6.37	46

	Tbc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	0.00	677.31	230.84	136.68	230.73	301.77	138
Ehot	0.00	2372.85	133.58	98.75	105.74	146.69	781
Mfra	75.91	163.21	127.96	102.50	136.35	161.81	4
Mnat	0.00	2329.94	118.16	58.87	85.41	126.16	836
Mtri	0.00	331.10	92.03	65.75	84.47	94.84	33
Ncap	0.00	2041.54	122.48	89.62	100.34	132.87	1213
Nnan	44.23	215.34	79.44	56.24	66.93	84.27	12
Rhin	15.93	646.21	115.44	21.02	90.67	108.99	14
Spet	0.00	515.40	145.98	51.82	118.21	214.33	67
Taeg	0.00	4083.35	276.36	136.93	238.73	350.17	874
Tmau	0.00	346.45	120.00	56.82	77.32	222.15	46

	Dc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	1.00	11.94	4.84	2.86	4.52	6.54	138
Ehot	0.89	8.91	2.34	1.88	2.21	2.61	781
Mfra	1.15	2.62	2.02	1.63	2.16	2.55	4
Mnat	0.59	6.22	2.22	1.56	2.10	2.74	836
Mtri	0.32	3.24	1.13	0.73	1.04	1.21	33
Ncap	0.28	6.73	2.06	1.55	1.96	2.43	1213
Nnan	0.98	2.08	1.60	1.42	1.60	1.81	12
Rhin	0.28	22.55	4.54	1.54	1.75	5.81	14
Spet	0.22	5.47	2.28	1.29	1.92	2.99	67
Taeg	0.27	14.06	3.75	2.49	3.25	4.58	874
Tmau	1.35	5.50	3.17	2.12	2.56	4.38	46

Table A.4: Summaries of the different variables, associated with the pulses, by species.

	Tc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	1.66	15.71	6.91	3.75	6.86	9.56	138
Ehot	0.89	11.18	4.06	3.12	3.99	4.84	781
Mfra	2.02	3.62	2.70	2.20	2.58	3.09	4
Mnat	1.02	8.95	3.46	2.50	3.31	4.21	836
Mtri	1.71	4.13	3.19	2.42	3.47	3.67	33
Ncap	1.26	9.26	4.00	3.18	3.94	4.74	1213
Nnan	1.66	3.14	2.39	2.21	2.40	2.60	12
Rhin	2.17	23.29	9.65	2.31	10.32	12.54	14
Spet	1.31	10.02	5.35	3.57	5.42	6.86	67
Taeg	1.07	14.47	5.77	3.80	5.67	7.35	874
Tmau	1.94	7.09	4.00	2.37	3.08	5.59	46

	Tk						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	0.00	8.23	2.07	0.33	1.31	3.24	138
Ehot	0.00	4.23	1.72	1.04	1.73	2.37	781
Mfra	0.29	1.08	0.68	0.42	0.68	0.93	4
Mnat	0.00	4.55	1.24	0.76	1.16	1.65	836
Mtri	0.31	2.95	2.06	1.57	2.31	2.64	33
Ncap	0.00	5.54	1.94	1.37	2.00	2.50	1213
Nnan	0.41	1.72	0.79	0.59	0.69	0.84	12
Rhin	0.00	17.14	5.11	0.65	1.16	9.87	14
Spet	0.00	5.28	3.07	2.33	2.89	4.05	67
Taeg	0.00	8.50	2.02	0.62	1.73	3.18	874
Tmau	0.09	2.19	0.82	0.41	0.67	1.13	46

	Qk						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	0.00	20.01	2.83	0.44	1.40	3.77	138
Ehot	0.00	31.20	10.10	3.94	8.31	15.53	781
Mfra	2.49	3.52	2.93	2.57	2.86	3.22	4
Mnat	0.00	33.18	6.31	2.86	5.17	8.25	836
Mtri	0.79	23.49	6.06	2.62	3.22	4.62	33
Ncap	0.00	37.50	11.07	5.97	10.39	15.18	1213
Nnan	1.48	6.38	3.42	2.68	3.28	3.97	12
Rhin	0.00	1.82	0.81	0.44	0.58	1.31	14
Spet	0.00	25.53	14.07	9.69	14.36	19.62	67
Taeg	0.00	26.57	5.60	1.32	3.22	9.34	874
Tmau	0.12	4.82	1.69	0.60	1.12	2.51	46

Table A.5: Summaries of the different variables, associated with the pulses, by species.

	S1						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	-132.48	808.65	126.11	49.18	98.28	175.69	138
Ehot	-207.50	1135.48	439.05	262.09	429.75	606.41	781
Mfra	213.53	464.20	350.42	277.39	361.98	435.00	4
Mnat	-442.55	1163.71	414.19	263.02	381.16	556.74	836
Mtri	122.06	852.97	384.57	270.88	345.70	381.47	33
Ncap	-157.37	1433.82	496.06	354.29	483.58	620.74	1213
Nnan	126.60	687.20	411.82	295.85	448.26	529.84	12
Rhin	-306.17	318.06	123.91	35.53	157.36	278.31	14
Spet	-82.73	871.33	449.04	311.62	452.67	599.75	67
Taeg	-376.35	1074.51	174.30	85.91	158.21	261.75	874
Tmau	-32.60	689.00	207.30	97.77	184.21	254.57	46

	Sc						
Species	**Min**	**Max**	**Avg**	**25%**	**50%**	**75%**	**N**
Cpum	-13.15	105.01	23.26	12.31	19.27	27.90	138
Ehot	0.70	133.10	47.10	35.13	44.67	56.51	781
Mfra	0.00	13.60	7.68	5.00	8.55	11.23	4
Mnat	-17.27	247.11	41.12	24.18	34.97	51.73	836
Mtri	24.82	288.42	193.32	107.53	223.30	248.38	33
Ncap	-22.66	273.40	55.85	35.59	49.37	68.44	1213
Nnan	6.63	55.18	30.22	22.92	33.13	37.44	12
Rhin	-7.84	9.90	1.89	0.00	0.00	7.15	14
Spet	15.10	305.07	84.63	30.48	73.14	120.20	67
Taeg	-7.36	293.39	48.77	28.10	42.97	61.32	874
Tmau	0.00	106.33	39.41	14.80	21.70	62.23	46

A.2 Model Comparison

The models trained to identify each species were compared, based on 50 resamples from the cross-validation data. Models were compared for performance measures such as the Area under the curve (AUC), Cohen's kappa, sensitivity and specificity. Since the cross-validation statistics were measured using identically resampled data sets, the statistical significance of the differences between models can be determined using a paired t-test (Kuhn & Johnson, 2013), where H0: difference = 0. Considering that multiple comparisons are being made, the p-value threshold for significance needs to be adjusted, for instance by using the Bonferroni correction (Bland & Altman, 1995), hence the confidence level α is divided by the number of tests made.

Table A.6: Comparison of the models trained to identify *Tadarida aegyptiaca*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	AUC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0069583	-0.0084648	-0.0076721	-0.0072478	-0.0083345
FDA	4.388e-13		-0.0015064	-0.0007138	-0.0002895	-0.0013762
RF	< 2.2e-16	0.0007485		0.0007926	0.0012170	0.0001302
SVM Rad	< 2.2e-16	0.2562017	1.407e-05		0.0004243	-0.0006624
SVM Poly	4.977e-15	1	3.587e-08	0.5129950		-0.0010868
XGBoost	< 2.2e-16	0.0014691	1	1.639e-05	4.569e-07	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0400114	-0.0647938	-0.0490173	-0.0499336	-0.0651738
FDA	6.897e-13		-0.0247825	-0.0090059	-0.0099222	-0.0251624
RF	< 2.2e-16	4.357e-11		0.0157765	0.0148603	-0.0003799
SVM Rad	2.019e-15	0.001508	4.934e-06		-0.0009163	-0.0161565
SVM Poly	< 2.2e-16	0.003988	1.566e-06	1		-0.0152402
XGBoost	< 2.2e-16	1.193e-14	1	4.799e-11	6.286e-08	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.005874	-0.012678	-0.008126	-0.010329	-0.011553
FDA	0.0003191		-0.006804	-0.002252	-0.004455	-0.005679
RF	1.625e-15	4.450e-08		0.004552	0.002349	0.001125
SVM Rad	1.360e-08	0.0848454	3.120e-07		-0.002203	-0.003427
SVM Poly	4.744e-12	0.0017625	0.0090420	0.0657783		-0.001224
XGBoost	8.796e-14	1.509e-06	1	3.680e-06	1	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0406359	-0.0558669	-0.0469846	-0.0411415	-0.0601891
FDA	1.157e-12		-0.0152311	-0.0063487	-0.0005056	-0.0195532
RF	< 2.2e-16	6.267e-05		0.0088824	0.0147255	-0.0043221
SVM Rad	6.658e-14	0.01244	0.15824		0.0058431	-0.0132045
SVM Poly	3.811e-14	1.00000	1.780e-05	0.54956		-0.0190476
XGBoost	< 2.2e-16	2.251e-15	1.00000	4.038e-08	2.859e-09	

Table A.7: Comparison of the models trained to identify *Miniopterus natalensis*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Radial	SVM Poly	XGBoost
QDA		-1.769e-03	-2.199e-03	-2.254e-03	-1.655e-03	-2.276e-03
FDA	0.001535		-4.297e-04	-4.855e-04	1.140e-04	-5.070e-04
Random Forests	2.200e-05	0.171638		-5.584e-05	5.436e-04	-7.735e-05
SVM Radial	1.022e-05	2.167e-05	1		5.995e-04	-2.151e-05
SVM Polynomial	0.002580	1	0.055663	0.004677		-6.210e-04
XGBoost	7.611e-06	4.100e-06	1	1	0.004393	

	Kappa					
Model	QDA	FDA	RF	SVM Radial	SVM Poly	XGBoost
QDA		-0.0346628	-0.0358617	-0.0306172	-0.0350897	-0.0351914
FDA	< 2.2e-16		-0.0011989	0.0040456	-0.0004270	-0.0005286
Random Forests	< 2.2e-16	1		0.0052445	0.0007719	0.0006703
SVM Radial	< 2.2e-16	0.0096036	0.0010165		-0.0044726	-0.0045742
SVM Polynomial	< 2.2e-16	1	1	0.0049700		-0.0001017
XGBoost	< 2.2e-16	1	1	0.0003701	1	

	Sensitivity					
Model	QDA	FDA	RF	SVM Radial	SVM Poly	XGBoost
QDA		-0.054610	-0.054610	-0.046474	-0.054610	-0.053571
FDA	< 2.2e-16		0.000000	0.008136	0.000000	0.001039
Random Forests	< 2.2e-16	NA		0.008136	0.000000	0.001039
SVM Radial	< 2.2e-16	2.147e-06	2.147e-06		-0.008136	-0.007097
SVM Polynomial	< 2.2e-16	NA	NA	2.147e-06		0.001039
XGBoost	< 2.2e-16	0.1532307	0.1532307	0.0001104	0.1532307	

	Specificity					
Model	QDA	FDA	RF	SVM Radial	SVM Poly	XGBoost
QDA		4.843e-04	-4.784e-05	-4.831e-05	2.907e-04	-4.843e-05
FDA	1		-5.321e-04	-5.326e-04	-1.936e-04	-5.327e-04
Random Forests	1	1		-4.702e-07	3.385e-04	-5.877e-07
SVM Radial	1	1	1		3.390e-04	-1.175e-07
SVM Polynomial	1	1	1	1		-3.391e-04
XGBoost	1	1	1	1	1	

Table A.8: Comparison of the models trained to identify *Neoromicia capensis*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-3.485e-02	-3.911e-02	-3.639e-02	-3.379e-02	-3.909e-02
FDA	< 2.2e-16		-4.252e-03	-1.536e-03	1.065e-03	-4.235e-03
RF	< 2.2e-16	1.405e-08		2.716e-03	5.317e-03	1.779e-05
SVM Rad	< 2.2e-16	0.0034911	4.185e-06		2.602e-03	-2.698e-03
SVM Poly	< 2.2e-16	1	2.630e-11	0.0008892		-5.300e-03
XGBoost	< 2.2e-16	< 2.2e-16	1	2.967e-10	4.114e-13	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.1187534	-0.1512897	-0.1431018	-0.1487991	-0.1521196
FDA	< 2.2e-16		-0.0325362	-0.0243484	-0.0300456	-0.0333661
RF	< 2.2e-16	6.030e-07		0.0081878	0.0024906	-0.0008299
SVM Rad	< 2.2e-16	5.206e-10	0.76598		-0.0056972	-0.0090177
SVM Poly	< 2.2e-16	2.084e-09	1.00000	1.00000		-0.0033205
XGBoost	< 2.2e-16	7.750e-15	1.00000	0.01093	1.00000	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0207057	-0.0310543	-0.0280598	-0.0371584	-0.0306747
FDA	6.595e-06		-0.0103486	-0.0073540	-0.0164526	-0.0099689
RF	3.034e-09	0.12417		0.0029946	-0.0061040	0.0003797
SVM Rad	1.027e-08	0.02266	1.00000		-0.0090986	-0.0026149
SVM Poly	2.305e-13	3.120e-05	1.00000	0.14529		0.0064837
XGBoost	1.622e-08	0.02424	1.00000	1.00000	0.96294	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0698270	-0.0857598	-0.0819510	-0.0812978	-0.0864637
FDA	< 2.2e-16		-0.0159328	-0.0121240	-0.0114708	-0.0166367
RF	< 2.2e-16	1.351e-07		0.0038088	0.0044620	-0.0007039
SVM Rad	< 2.2e-16	2.752e-12	1		0.0006532	-0.0045126
SVM Poly	< 2.2e-16	6.555e-06	0.167951	1		-0.0051659
XGBoost	< 2.2e-16	< 2.2e-16	1	0.001584	0.082806	

Table A.9: Comparison of the models trained to identify *Eptesicus hottentotus*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0334041	-0.0397311	-0.0362602	-0.0329686	-0.0394785
FDA	< 2.2e-16		-0.0063270	-0.0028561	0.0004354	-0.0060744
RF	< 2.2e-16	6.395e-05		0.0034709	0.0067625	0.0002526
SVM Rad	< 2.2e-16	0.003434	0.011638		0.0032915	-0.0032183
SVM Poly	< 2.2e-16	1	8.942e-08	0.204094		-0.0065098
XGBoost	< 2.2e-16	1.326e-11	1	6.939e-09	2.383e-05	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.142551	-0.178680	-0.163706	-0.156435	-0.174196
FDA	< 2.2e-16		-0.036129	-0.021155	-0.013884	-0.031645
RF	< 2.2e-16	5.009e-08		0.014975	0.022245	0.004484
SVM Rad	< 2.2e-16	3.356e-05	0.117851		0.007270	-0.010491
SVM Poly	< 2.2e-16	0.650811	0.003276	1		-0.017761
XGBoost	< 2.2e-16	3.248e-11	1	0.057524	0.164589	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.016437	0.009104	0.018239	0.040133	0.012492
FDA	0.585443		-0.007333	0.001803	0.023696	-0.003945
RF	1	1		0.009135	0.031029	0.003388
SVM Rad	0.265378	1	1		0.021894	-0.005747
SVM Poly	0.002841	0.360580	0.002172	0.494481		-0.027641
XGBoost	1	1	1	1	0.117746	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0667714	-0.0784341	-0.0752130	-0.0782077	-0.0776642
FDA	< 2.2e-16		-0.0116627	-0.0084417	-0.0114363	-0.0108928
RF	< 2.2e-16	2.798e-09		0.0032211	0.0002264	0.0007699
SVM Rad	< 2.2e-16	7.387e-12	0.61317		-0.0029946	-0.0024511
SVM Poly	< 2.2e-16	1.006e-06	1.00000	1.00000		0.0005435
XGBoost	< 2.2e-16	< 2.2e-16	1.00000	0.05117	1.00000	

Table A.10: Comparison of the models trained to identify *Chaerephon pumilus*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.0054346	-0.0153413	0.0665638	-0.0261305	-0.0146737
FDA	1		-0.0207759	0.0611293	-0.0315650	-0.0201082
RF	0.0027525	8.301e-05		0.0819051	-0.0107891	0.0006677
SVM Rad	6.162e-10	1.206e-08	5.045e-13		-0.0926943	-0.0812375
SVM Poly	2.661e-15	3.160e-11	0.0020329	1.138e-15		0.0114568
XGBoost	0.0001368	0.0009579	1	4.887e-12	0.0008126	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.01848	-0.14805	0.12851	-0.38597	-0.18208
FDA	1		-0.12958	0.14698	-0.36749	-0.16360
RF	2.766e-09	1.585e-08		0.27656	-0.23791	-0.03403
SVM Rad	1.251e-07	1.681e-09	3.310e-15		-0.51447	-0.31059
SVM Poly	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16		0.20388
XGBoost	5.027e-14	1.110e-11	1	< 2.2e-16	< 2.2e-16	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.27529	-0.08824	0.17529	-0.49048	-0.14353
FDA	2.424e-16		0.18706	0.45059	-0.21518	0.13176
RF	0.00332	1.902e-11		0.26353	-0.40224	-0.05529
SVM Rad	8.771e-10	< 2.2e-16	2.895e-12		-0.66577	-0.31882
SVM Poly	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16		0.34695
XGBoost	3.198e-08	1.149e-05	0.62514	< 2.2e-16	< 2.2e-16	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.0205479	-0.0067319	-0.0043836	0.0387867	-0.0058708
FDA	< 2.2e-16		-0.0272798	-0.0249315	0.0182387	-0.0264188
RF	1.067e-10	< 2.2e-16		0.0023483	0.0455186	0.0008611
SVM Rad	9.271e-05	< 2.2e-16	0.008239		0.0431703	-0.0014873
SVM Poly	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16		-0.0446575
XGBoost	1.694e-09	< 2.2e-16	1	0.457722	< 2.2e-16	

Table A.11: Comparison of the models trained to identify *Sauromys petrophilus*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.0503083	-0.0027198	0.0038998	0.0101192	-0.0028213
FDA	4.453e-10		-0.0530281	-0.0464085	-0.0401892	-0.0531296
RF	0.001178	1.245e-10		0.0066196	0.0128390	-0.0001014
SVM Rad	1	4.202e-10	0.062798		0.0062194	-0.0067210
SVM Poly	0.152561	8.927e-06	0.019780	1		-0.0129404
XGBoost	0.003115	3.949e-11	1	0.030053	0.012825	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.04433	-0.06920	-0.11809	0.20921	-0.10325
FDA	0.4118897		-0.02487	-0.07376	0.25354	-0.05892
RF	0.0863914	1		-0.04888	0.27841	-0.03405
SVM Rad	9.577e-06	3.462e-05	0.1243549		0.32730	0.01483
SVM Poly	6.788e-05	2.389e-07	7.323e-08	1.914e-10		-0.31246
XGBoost	0.0002317	0.0010560	0.6432186	1	1.179e-09	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		1.813e-03	-3.627e-03	-2.431e-03	-2.315e-03	-2.469e-03
FDA	0.0407290		-5.440e-03	-4.244e-03	-4.128e-03	-4.282e-03
RF	2.563e-16	9.203e-16		1.196e-03	1.311e-03	1.157e-03
SVM Rad	8.524e-06	2.351e-13	0.0003991		1.155e-04	-3.846e-05
SVM Poly	2.207e-06	2.201e-09	0.0001066	1		-1.540e-04
XGBoost	5.436e-08	2.652e-13	1.083e-06	1	1	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.11533	0.02067	-0.09111	0.27733	-0.06600
FDA	0.001136		0.13600	0.02422	0.39267	0.04933
RF	1	3.118e-06		-0.11178	0.25667	-0.08667
SVM Rad	0.056593	1	0.001034		0.36844	0.02511
SVM Poly	1.177e-07	4.270e-14	3.959e-07	4.656e-12		-0.34333
XGBoost	0.490685	0.021417	0.005078	1	3.352e-11	

Table A.12: Comparison of the models trained to identify *Myotis tricolor*, using the cross-validation data. The upper diagonal represent the difference of given performance measure between models, while the lower diagonal shows the p-value.

	ROC					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.0283505	-0.0159047	-0.0167712	-0.0174259	-0.0223703
FDA	0.775582		-0.0442552	-0.0451217	-0.0457764	-0.0507208
RF	1	0.046026		-0.0008665	-0.0015213	-0.0064656
SVM Rad	1	0.003710	1		-0.0006547	-0.0055991
SVM Poly	0.656442	0.014597	1	1		-0.0049443
XGBoost	0.110007	0.002249	1	1	1	

	Kappa					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-0.0994711	-0.1084522	-0.1667522	0.0154416	-0.1669743
FDA	0.0350947		-0.0089811	-0.0672811	0.1149128	-0.0675032
RF	0.0104231	1		-0.0583000	0.1238938	-0.0585221
SVM Rad	3.946e-06	0.0005726	1		0.1821938	-0.0002221
SVM Poly	1	0.0898133	0.0628445	7.756e-05		-0.1824159
XGBoost	1.396e-06	0.0043015	1	1	0.0001707	

	Sensitivity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		0.020	0.128	0.028	0.220	0.056
FDA	1		0.108	0.008	0.200	0.036
RF	0.0210640	0.2004550		-0.100	0.092	-0.072
SVM Rad	1	1	0.3745622		0.192	0.028
SVM Poly	8.566e-05	0.0010626	0.9874440	0.0008378		-0.164
XGBoost	1	0.7268418	1	1	0.0103034	

	Specificity					
Model	QDA	FDA	RF	SVM Rad	SVM Poly	XGBoost
QDA		-3.593e-03	-5.849e-03	-5.390e-03	-4.778e-03	-5.811e-03
FDA	1.411e-06		-2.255e-03	-1.797e-03	-1.185e-03	-2.217e-03
Random Forests	< 2.2e-16	1.706e-09		4.585e-04	1.070e-03	3.824e-05
SVM Radial	4.890e-16	1.843e-05	0.028747		6.119e-04	-4.203e-04
SVM Polynomial	2.249e-12	0.002266	6.470e-07	0.213809		-1.032e-03
XGBoost	< 2.2e-16	2.309e-09	1	0.095269	4.760e-06	

Bibliography

Ahlen, I. & Baagøe, H. J. (1999). Use of ultrasound detectors for bat studies in europe: experiences from field identification, surveys, and monitoring. *Acta Chiropterologica*, *1*(2), 137–150.

Amos, A. (2016). Bat killings by wind energy turbines continue. Retrieved from https://www.scientificamerican.com/article/bat-killings-by-wind-energy-turbines-continue/

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40–79.

Armitage, D. W. & Ober, H. K. (2010a). A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, *5*(6), 465–473.

Armitage, D. W. & Ober, H. K. (2010b). A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, *5*(6), 465–473.

Arnett, E. B., Inkley, D. B., Johnson, D. H., Larkin, R. P., Manes, S., andRuss Mason, A. M. M., ... Thresher, R. (2007). Impacts of wind energy facilities on wildlife and wildlife habitat. *Wildlife Society technical review*, *7*(2), 49.

Baerwald, E. F., D'Amours, G. H., Klug, B. J., & Barclay, R. M. R. (2008). Barotrauma is a significant cause of bat fatalities at wind turbines. *Current biology*, *18*(16), R695–R696.

Basil, G. S., Vanitharani, J., & K, J. (2014). An extensive review of methods of identification of bat species through acoustics. *International Journal of Computer Applications Technology and Research*, *3*(4), 186–192.

Bland, J. M. & Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *British Medical Journal*, *310*(6973), 170.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, T. (2015). Tianqi chen's answer to "what is the difference between the r gbm (gradient boosting machine) and xgboost (extreme gradient boosting)?" Retrieved from https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting

Chen, T. & Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

Clement, M. J., Murray, K. L., Solick, D. I., & Gruver, J. C. (2014, September 1). The effect of call libraries and acoustic filters on the identification of bat echolocation. *Ecology and Evolution*, *4*(17), 3482–3493.

Clemmensen, L., Hastie, T., Witten, D., & Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, *53*(4), 406–413.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cryan, P. M. & Barclay, R. M. R. (2009, December 15). Causes of bat fatalities at wind turbines: hypotheses and predictions. *Journal of Mammalogy*, *90*(6), 1330–1340.

Domingos, P. (2015). *The master algorithm*. Hachette Book Group USA.

Drewitt, A. L. & Langston, R. H. W. (2008). Collision effects of wind-power generators and other obstacles on birds. *Annals of the New York Academy of Sciences*, *1134*(1), 233–266.

Agreement on the Conservation of Populations of European Bats. (1991). Retrieved from http: //www.eurobats.org/official_documents/agreement_text

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fenton, M. B. & Bell, G. P. (1981). Recognition of species of insectivorous bats by their echolocation calls. *Journal of Mammalogy*, *62*(2), 233–243.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of human genetics*, *7*(2), 179–188.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (second edition). Springer series in statistics New York.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., . . . Hunt., T. (2017). *Caret: classification and regression training*. R package version 6.0-77.

Fukui, D., Agetsuma, N., & Hill, D. A. (2004). Acoustic identification of eight species of bat (mammalia: chiroptera) inhabiting forests of southern hokkaido, japan: potential for conservation monitoring. *Zoological Science*, *21*(9), 947–955.

GAO. (2005). *Impacts on wildlife impacts on wildlife and government responsibilities for regulating development and protecting wildlife*. United States Government Accountability Office.

Gaston, K. J. & O'Neill, M. A. (2004, April 29). Automated species identification: why not? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1444), 655–667.

He, H. & Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Herr, A., Klomp, N. I., & Atkinson, J. S. (1997). Identification of bat echolocation calls using a decision tree classification system. *Complexity International*, *4*, 1–9.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Jennings, N., Parsons, S., & Pocock, M. J. O. (2008). Human vs. machine: identification of bat species from their echolocation calls by humans and by artificial neural networks. *Canadian Journal of Zoology*, *86*(5), 371–377.

Jones, G., Jacobs, D., Kunz, T. H., Willig, M. R., & Racey, P. A. (2009). Carpe noctem: the importance of bats as bioindicators. *Endangered species research*, *8*(1-2), 93–115.

Jones, G. & Siemers, B. M. [Björn M.]. (2011). The communicative potential of bat echolocation pulses. *Journal of Comparative Physiology A*, *197*(5), 447–457.

Jones, G. & Teeling, E. C. (2006). The evolution of echolocation in bats. *Trends in Ecology & Evolution*, *21*(3), 149–156.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. Springer New York.

Kunz, T. H., Arnett, E. B., Erickson, W. P., Hoar, A. R., Johnson, G. D., Larkin, R. P., . . . Tuttle, M. D. (2007). Ecological impacts of wind energy development on bats: questions, research needs, and hypotheses. *Frontiers in Ecology and the Environment*, *5*(6), 315–324.

Lamb, J. M., Ralph, T. M. C., Naidoo, T., Taylor, P. J., Ratrimomanarivo, F., Stanley, W. T., & Goodman, S. M. (2011). Toward a molecular phylogeny for the molossidae (chiroptera) of the afro-malagasy region. *Acta Chiropterologica*, *13*(1), 1–16.

Lantz, B. (2015). *Machine learning with r*. Packt Publishing Ltd.

LeDell, E., Petersen, M., & van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic journal of statistics*, *9*(1), 1583.

Lim, B. K. & Engstrom, M. D. (2001). Bat community structure at iwokrama forest, guyana. *Journal of Tropical Ecology*, *17*(5), 647–665.

Ling, C. X., Huang, J., & Zhang, H. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai* (Vol. 3, pp. 519–524).

Lundy, M., Montgomery, I., & Russ, J. (2010). Climate change-linked range expansion of nathusius' pipistrelle bat, pipistrellus nathusii (keyserling & blasius, 1839). *Journal of Biogeography*, *37*(12), 2232–2242.

Mankin, R. W. (2011). Recent developments in the use of acoustic sensors and signal processing tools to target early infestations of red palm weevil in agricultural environments. *Florida Entomologist*, *94*(4), 761–765.

Mellinger, D. K. & Clark, C. W. (2000). Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, *107*(6), 3518–3529. eprint: http://dx.doi.org/10.1121/1.429434

Milne, D. J. (2002). *Key to the bat calls of the top end of the northern territory*. Parks and Wldlife Commission of the Northern Territory.

Noda, M. K. (1995). Flexible bat echolocation: the influence of individual, habitat and conspecifics on sonar signal design. *Behavioral ecology and sociobiology*, *36*(3), 207–219.

Ochoa, J., O'Farrell, M. J., & Miller, B. W. (2000). Contribution of acoustic methods to the study of insectivorous bat diversity in protected areas from northern venezuela. *Acta Chiropterologica*, *2*(2), 171–183.

Papadatou, E., Butlin, R. K., & Altringham, J. D. (2008). Identification of bat species in greece from their echolocation calls. *Acta Chiropterologica*, *10*(1), 127–143.

Parsons, S. (2001). Identification of new zealand bats (chalinolobus tuberculatus and mystacina tuberculata) in flight from analysis of echolocation calls by artificial neural networks. *Journal of Zoology*, *253*(4), 447–456.

Parsons, S., Boonman, A. M., & Obrist, M. K. (2000, November 1). Advantages and disadvantages of techniques for transforming and analyzing chiropteran echolocation calls. *Journal of Mammalogy*, *81*(4), 927–938.

Parsons, S. & Jones, G. (2000). Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology*, *203*(17), 2641–2656.

Peake, T. M. & McGregor, P. K. (2001). Corncrake crex crex census estimates: a conservation application of vocal individuality. *Animal Biodiversity and Conservation*, *24*(1), 81–90.

Preatoni, D. G., Nodari, M., Chirichella, R., Tosi, G., Wauters, L. A., & Martinoli, A. (2005). Identifying bats from time-expanded recordings of search calls: comparing classification methods. *Journal of Wildlife Management*, *69*(4), 1601–1614.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Icml* (Vol. 98, pp. 445–453).

R Core Team. (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Redgwell, R. D., Szewczak, J. M., Jones, G., & Parsons, S. (2009). Classification of echolocation calls from 14 species of bat by support vector machines and ensembles of neural networks. *Algorithms*, *2*(3), 907–924.

Russo, D. & Jones, G. (2002). Identification of twenty-two bat species (mammalia: chiroptera) from italy by analysis of time-expanded recordings of echolocation calls. *Journal of Zoology*, *258*(1), 91–103.

Schnitzler, H.-U., Moss, C. F., & Denzinger, A. (2003). From spatial orientation to food acquisition in echolocating bats. *Trends in Ecology & Evolution*, *18*(8), 386–394.

Scott, C. D. (2012). *Automated techniques for bat echolocation call analysis* (Doctoral dissertation, University of Leeds).

Siemers, B. M. [Bjorn M.] & Schnitzler, H.-U. (2004). Echolocation signal reflect niche differentiation in five sympatric congeneric bat species. *Nature*, *429*(6992), 657.

Skowronski, M. D. & Harris, J. G. (2006). Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. *The Journal of the Acoustical Society of America*, *119*(3), 1817–1833.

Springer, M. S., Teeling, E. C., Madsen, O., Stanhope, M. J., & de Jong, W. W. (2001). Integrated fossil and molecular data reconstruct bat echolocation. *Proceedings of the National Academy of Sciences*, *98*(11), 6241–6246.

Stathopoulos, V., Zamora-Gutierrez, V., Jones, K., & Girolami, M. (2014). Bat call identification with gaussian process multinomial probit regression and a dynamic time warping kernel. In *Artificial intelligence and statistics* (pp. 913–921).

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Pearson International Edition. Pearson Addison Wesley.

Tsang, S. M., Cirranello, A. L., Bates, P. J. J., & Simmons, N. B. (2016). The roles of taxonomy and systematics in bat conservation. In *Bats in the anthropocene: conservation of bats in a changing world* (pp. 503–538). Springer.

Varma, S. & Simon, R. (2006, February). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91.

Vaughan, N., Jones, G., & Harris, S. (1997). Identification of british bat species by multivariate analysis of echolocation call parameters. *Bioacoustics*, *7*(3), 189–207.

Walters, C. L., Freeman, R., Collen, A., Dietz, C., Fenton, M. B., Jones, G., ... Jones, K. E. (2012, October 1). A continental-scale tool for acoustic identification of european bats. *Journal of Applied Ecology*, *49*(5), 1064–1074.

Welch, B. L. (1939). Note on discriminant functions. *Biometrika*, *31*(1/2), 218–220.

Yack, T. M., Barlow, J., Rankin, S., & Gillespie, D. (2009). Integration of automated detection methods into noaa southwest fisheries science center (swfsc) acoustic marine mammal monitoring protocol. *The Journal of the Acoustical Society of America*, *125*(4), 2588–2588. eprint: http://dx.doi.org/10.1121/1.4783833

Zamora-Gutierrez, V., Lopez-Gonzalez, C., Gonzalez, M. C. M., Fenton, B., Jones, G., Kalko, E. K. V., ... Jones, K. E. (2016). Acoustic identification of mexican bats based on taxonomic and ecological constraints on call design. *Methods in Ecology and Evolution*, *7*(9), 1082–1091.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.