ASSESSING AND CORRECTING FOR THE EFFECTS OF SPECIES MISCLASSIFICATION DURING PASSIVE ACOUSTIC SURVEYS OF CETACEANS

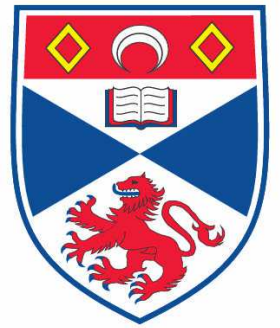
Marjolaine Caillat

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2013

Assessing and correcting for the effects of species misclassification during passive acoustic surveys of cetaceans

Marjolaine Caillat



This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy

School of Biology and School of Mathematics and Statistics,

at the University of St Andrews

May 2013

1. Candidate's declarations:

I, Marjolaine Caillat hereby certify that this thesis, which is approximately 59 000. words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2008 and as a candidate for the degree of PhD in the School of Biology in October 2009 in; the higher study for which this is a record was carried out in the University of St Andrews between 2008 and 2013.

I, Marjolaine Caillat, received assistance in the writing of this thesis in respect of grammar, spelling and syntax, which was provided by Angelica Studeny, Rene Swift and Janine Illian.

Date 14/05/2013 signature of candidate …….

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 14/05/2013 signature of supervisor …….

3. Permission for electronic publication: (*to be signed by both candidate and supervisor*)

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Access to printed copy and electronic publication of thesis through the University of St Andrews.

Date 14/05/2013 signature of candidate …… signature of supervisor …….

Abstract

In conservation ecology, abundance estimates are an important factor from which management decisions are based. Methods to estimate abundance of cetaceans from visual detections are largely developed, whereas parallel methods based on passive acoustic detections are still in their infancy. To estimate the abundance of cetacean species using acoustic detection data, it is first necessary to correctly identify the species that are detected. The current automatic PAMGUARD Whistle Classifier used to automatically identify whistle detection of cetacean species is modified with the objective to facilitate the use of these detections to estimate cetacean abundance. Given the variability of cetacean sounds within and between species, developing an automated species classifier with a 100% correct classification probability for any species is unfeasible. However, through the examples of two case studies it is shown that large and high quality datasets with which to develop these automatic classifiers increase the probability of creating reliable classifiers with low and precise misclassification probability.

Given that misclassification is unavoidable, it is necessary to consider the effect of misclassified detections on the number of observed acoustic calls detected and thus on abundance estimates, and to develop robust methods to cope with these misclassifications. Through both heuristic and Bayesian approaches it is demonstrated that if misclassification probabilities are known or estimated precisely, it is possible to estimate the true number of detected calls accurately and precisely. However, misclassification and uncertainty increase the variance of the estimates. If the true numbers of detections from different species are similar, then a small amount of misclassification between species and a small amount of uncertainty in the probabilities of misclassification does not have a detrimental effect on the overall variance and bias of the estimate. However, if there is a difference in the encounter rate between species calls associated with a large amount of uncertainty in the probabilities of misclassification, then the variance of the estimates becomes larger and the bias increases; this in return increases the variance and the bias of the final abundance estimate. This study despite not bringing perfect results highlights for the first time the importance of dealing with the problem of species misclassification for cetacean if acoustic detections are to be used to estimate abundance of cetaceans.

Acknowledgments

The first person I would like to thank is my main supervisor Dr Douglas Gillespie. Our adventure started by chance 8 years ago…when he accepted to take one student with no knowledge about acoustic to do some classification work on harbour porpoises. Since then he has always been there for me; offering me jobs, encouraging me to take my time each time I had babies, proposing this great PhD project to me. During this PhD, he was always available almost in the hour I needed him, for simple questions as well as for moments of brainstorming, in which I learned a lot. For all that a simple THANK YOU is only the visible part of the iceberg representing my gratitude. Then I want to thank my second supervisor Dr Len Thomas, for his help throughout this PhD. He taught me to be more rigorous and to pay attention to every detail. A thank you to Professor John Harwood who made it possible to start this great adventure.

Amongst others, I wish to thank in particular Professor Peter Jupp and Dr Ruth King for their technical help and for Janine Illian for correcting some of my chapters. I wish to thank the NERC for funding this PhD and the precious extra time I needed.

An unquantifiable thanks to all my friends who supported me during this time, they are numerous but some deserve to be mentioned especially: René for his cookies, teas, rhubarb pies and other little attentions each time I was working hard and for his help, always associated with a positive comments when correcting my English. Thanks to Debbie, Sophie, Glenna and Ester for their help with the technical parts of the Bayesian statistics. Violaine, Marion for your motivation and encouragements. Susan and Catarina for the huge support with the boys.

And finally the people without whom I am not sure I would have been able to finish this PhD. Angelika, for her endless moral support and rigorous English corrections. Roberta and her magical needles, which, during these 4 years, helped me to keep my energy levels high, my stress levels low and who always managed to say the right word to help me to feel stronger and more confident. To both of you thanks a million.

Merci à ma maman, mon papa, mon grand-père, pour leur soutien inconditionnel malgré les plus de 1000km qui nous séparent. Merci à eux et à Edith et Alain pour être venu vous occuper des enfants quand on en avait le plus besoin et merci pour toutes les ondes positives et énergisantes envoyées.

And finally the last, but most important, thanks to Guillaume, Elliot and Matisse. In different ways, all of you helped me to cope, to put things in perspective, to make me laugh and forget when it was difficult, to make me feel loved with all your cuddles when I was feeling rubbish and stupid. And thanks to YOU, Guillaume, who supported me on a daily basis. In the last 6 months you helped so much at home, without complaining once, in order to allow me to work evenings and week-ends, you looked after our children, house, garden, and me brilliantly. “Thanks to the three of you” is such a small word, it means much more to me.

Finally, I would like to dedicate this thesis to my Bonne-Maman, who left our world when I started my master project in 2005 and from which this adventure started. I am not sure I would have been able to come to St Andrews where everything started without her. Pendant ces 25 années à tes côtés, tu ne m'as jamais jugé et tu m'as toujours supporté, fait confiance quel que soit le chemin que je prenais. Tu rêvais de me voir Docteur, ce livre, ce travail représente un pas de plus vers cet objectif, même si ce n'est pas tout à fait le même genre de docteur.

A ma Bonne Maman …

Table of content

Abstract i

Acknowledgments ii

Chapter 1: General Introduction 1
1.1. Background 1
1.2. Abundance estimation a tool for management strategy 3
1.2.1. Generalities 3
1.2.2. Abundance estimation of cetaceans 5
1.3. Thesis outline 13
1.4. Overview of Bayesian theory 17
1.4.1. Introduction 17
1.4.2. Bayes' theorem 17
1.4.3. Elements of Bayesian analysis 18
1.4.4. Mixing 21
1.4.5. Burn-in and convergence 21
1.4.6. Parameter inferences 22

Part I. Classification 23

Chapter 2: Measuring the variability of an automatic whistle classifier 24
2.1. Introduction 24
2.1.1. Misclassification 24
2.1.2. Uncertainty in the estimates of classification probability 25
2.1.3. PAMGUARD Whistle Classifier 26
2.2. Methods 29
2.2.1. Data 29
2.2.2. PAMGUARD Whistle Classifier modifications 29
2.2.3. Models 30
2.3. Results 33
2.3.1. Data description 33
2.3.2. Model selection 34
2.3.3. Comparison of the variance with the version of the PWC described in Gillespie et al. (2013) 35
2.4. Discussion 36

Chapter 3: Classification of data from a reliable training dataset 38
3.1. Introduction 38
3.2. Methods 40
3.2.1. Creation of the classifiers 40
3.2.2. Classification of unidentified data 44
3.3. Results 47
3.3.1. Training dataset 47
3.3.2. Selection of the optimal fragment and section length 47
3.3.3. Training of the classifiers 49
3.4. Discussion 54

Chapter 4: Classification of data from a less reliable training dataset 57
4.1. Introduction 57
4.2. Datasets 59
4.2.1. Visual survey 59
4.2.2. Acoustic survey 60
4.3. Methods 63
4.3.1. Creation of the training datasets 63
4.3.2. Creation of the classifiers 65
4.3.3. Classification of new data 66
4.4. Results 68

4.4.1. Visual detection selection....68
4.4.2. Acoustic Data....68
4.4.3. Link between Acoustic and Visual observations....71
4.4.4. Parameter optimisation....72
4.4.5. Classifier Training....73
4.4.6. Classification of new data with the classifiers....77
4.5. Discussion....86
4.5.1. Parameters influencing the performance of a classifier....86
4.5.2. Consequences of a lack of training data....86
4.5.3. Consequences of a lack of the accuracy of species identification....87
4.5.4. Advantages of acoustic detections over visual detections....88
Chapter 5: Classification: General Discussion....91
5.1. Introduction....91
5.2. Parameters influencing on the quality of the classifier....91
5.2.1. Size of the training dataset....92
5.2.1. Reliability of the visual observation....92
5.2.2. Characteristics of the classification groups....93
5.2.3. False positive detections....93
5.3. Defining the robustness of a classifier....94
5.4. Specificity of the PAMGUARD Whistle Classifier....94
5.5. Recommendations of creating a good whistle classifier....95
Part II. Misclassification....97
Chapter 6: A heuristic method to estimate the number of acoustic detections in the presence of species misclassification....98
6.1. Introduction....98
6.2. The classification process....100
6.3. Methods....101
6.3.1. Models tested....102
6.3.2. Analytical approach....103
6.3.3. Data simulation....104
6.4. Results....105
6.4.1. No uncertainty in the confusion matrix....105
6.4.2. Uncertainty in the confusion matrix....106
6.5. Discussion....108
Chapter 7: A Bayesian method to estimate the number of acoustic detections in the presence of species misclassification....112
7.1. Introduction....112
7.2. Methods....113
7.2.1. Data....113
7.2.2. Overview....114
7.2.3. Likelihood functions....115
7.2.4. Prior distributions....116
7.2.5. Investigated scenarios....120
7.2.6. Posterior inference....121
7.2.7. Statistical versus biological significance....123
7.3. Results....123
7.4. Discussion....134
7.4.1. Consequences of unequal number of detections between species....134
7.4.2. Results of the prior sensitivity analysis....135
7.4.3. Impact of classification scenarios....135
7.4.4. Rare species....136
7.4.5. Criticism of the model and conclusion....136

Chapter 8: Methods to estimate the number of acoustic detections in the presence of species misclassification applied to real data 138
8.1. Introduction 138
8.2. Methods 139
8.2.1. Heuristic methods 139
8.2.2. Bayesian methods 140
8.2.3. Description of the results 147
8.3. Results 148
8.3.1. Dataset 1: French training dataset classified with the *3Sp* Spanish classifier 148
8.3.2. Dataset 2: Training data of 5Sp classifier 150
8.3.3. Dataset 3: Data recorded from the DO1 EAR deployment in the Moray Firth S.A.C 151
8.3.4. Dataset 4: Data recorded from EARs (E17,A20,A21) deployed in the MORL-BOWL wind farm sites 153
8.4. Discussion 155
Chapter 9: Dealing with species misclassification: General discussion 158
9.1. Equal versus unequal detections between species 158
9.2. Prior sensitivity 159
9.3. Misclassification probabilities 159
9.4. Rare species 160
9.5. Limitations of the methods 161
9.6. Abundance estimation using misclassified observed detections 162
9.7. Conclusion 164
References 158
Appendix A. Appendix for chapter 3 ..... A-1
Appendix B. Appendix for chapter 4 ..... B-14
Appendix C. Appendix for chapter 6. ..... C-23
C.1 Analytic estimate of the bias and variance of the true number of detected calls when there is no uncertainty in the values of the confusion matrix. C-23
C.2: Tables C-25
Appendix D. Appendix for chapter 7 ..... D-29

List of figures

Figure 1-1: Trace plots representing a good (a) and a slow (b) mixing of the MCMC chains. These plots were extracted from (King et al., 2010, p131). 21

Figure 2-1: Schematic diagram of the PAMGUARD Whistle classifier training process during the B bootstraps. 27

Figure 2-2: Last run of the PAMGUARD Whistle classifier training process. 28

Figure 2-3: Training process of the modified PAMGUARD Whistle classifier. Note the testing dataset has been divided in two so it is possible to measure a between and within variance. 30

Figure 2-4: Variances of the classification probabilities (V_{ij}) for a given classification probabilities (p_{ij}) and a training sampling size (S). S is the proportion of the sections used to train the classifier: half of the sections used to train the classifier (black open circles), a quarter of the sections (red triangle) and an eighth of the sections (blue cross). Symbolised with a black cross are the variances as function of probabilities obtained from a Dirichlet distribution directly. 34

Figure 2-5 : Observed data (open symbols) versus predicted (lines) and extrapolation (bold black triangles) with full dataset. Each colour represents a sampling size as described in previous figure. 35

Figure 3-1 Example screen grab showing whistle contours extracted from recordings of bottlenose dolphins using the PAMGUARD Whistle and Moan detector module. Frequency (kHz) is on the y-axis and time (10 seconds) is on the x-axis). The different colours show the contours identified by the WMD (clicks are also visible above 6 kHz). (SMRU ltd et al., 2011) 43

Figure 3-2: Quality coefficient *Q* of the *2Sp* classifier for varying fragment lengths (averaged over section lengths between 10 and 60 fragments) and varying section lengths (averaged over fragment lengths between 5 and 39 bins). 48

Figure 3-3: Quality coefficient *Q* of the *5Sp* classifier for varying fragment lengths (averaged over section lengths between 10 and 60 fragments) and varying section lengths (averaged over fragment lengths between 5 and 39 bins). 48

Figure 3-4: Screen capture from PWC of a "rubbing" false detection. Frequency is on the y-axis (0 to 24 kHz) and time (5.58seconds) is on the x-axis. The different colours show the contours generated by the PAMGAURD whistle detector. (SMRU ltd et al., 2011) 51

Figure 4-1: Total Contour Length per minutes for A) the French dataset and B) the Spanish data set. Figures on the right are zoomed to 40 s with the vertical line being placed at 4 s of contour length per minute. 62

Figure 4-2: Schematic diagram of the data selection and decision process. (i) Selection of visual data with a high confidence of species identification. (ii) Detection and selection of whistles contour and discard of the false positive contours. (iii) Creation of the training dataset by assigning sightings to recordings. (iv) Training of the classifiers with the datasets. (v) Testing of the classifiers on identified data.(vi) use of the classifier to identify new data. 67

Figure 4-3: Distribution of the false positive detections (FD) and whistle (W) contour lengths for each category of $L_{m.}$ The * indicates if the mean difference between the false detection contours and whistle contours was significant with a probability (p)<0.05. 70

Figure 4-4: Quality coefficient *Q* for varying fragment lengths (averaged over section lengths between 10 and 30 fragments); and varying section lengths (averaged over fragment lengths between 5 and 15 bins) used to classify five groups of species (bottlenose dolphins, common dolphins, common/striped dolphins, pilot whales and striped dolphins) from both the French and Spanish datasets. 73

Figure 6-1: Expected true number of detections for each species, from simulation without uncertainty within the confusion matrix: for equal data scenarios Sc1a to Sc1e (A) and for unequal data scenarios Sc2.a to Sc2.e (B). Solid bars show the standard deviation and the dotted line the true number of detections..... 106

Figure 6-2: CV of the expected true number of detections for unequal data for each scenario (Sc2b to Sc2e), with different misclassification probabilities and no uncertainty in the confusion matrix. The y axis is on the log10 scale..... 107

Figure 6-3: Mean of the CV of the expected true number of detections across the ten scenarios Sc1a to Sc1e (A) and Sc2a to Sc2e (B) for each species and each level of uncertainty of the confusion matrix values, no uncertainty, low uncertainty and high uncertainty. The y axis is on the log10 scale. 107

Figure 7-1: Trace plots showing MCMC sample values for parameters *v* (y-axis) vs. sample iteration (x-axis, after thinning), obtained with classification scenario Sc1.b and prior *V1*. 124

Figure 7-2: Auto-correlation plots of the posterior distributions of parameters v_j obtained from the model with classification scenario Sc1.b and prior *V1*. Note that samples were thinned, so that a lag of 1 corresponds to 4 MCMC sample iterations..... 124

Figure 7-3: Trace plots for each chain obtaine in the analysis of model with scenario Sc.e and prior V2 for each species. Within-chain mixing is good but the chains arere not converging towards the same values. 125

Figure 7-4: Relative bias (beanplots) and mean relative bias (bold lines) for each species for models where *p* is assumed known (Models A) or is estimated (Models B), and for models with equal and unequal data. The thin lines crossing the entire plot (close to zero) are the mean across the four species. For Models A and equal data each beanplot is computed from 10 values (the relative bias for 5 scenarios times 2 priors on v), for unequal data each beanplot is computed from 15 values (the relative bias for 5 scenarios times 2 priors on v). For models B each bean plot is computed from 24 and 36 values from equal and unequal data, respectively (relative bias for 4 scenarios times 3 priors on p times 2 or 3 prios on v, respectively). 127

Figure 7-5: Beanplots of the relative bias of the estimates as function of the priors on the parameters *v* (V1 to V3) and species. The bold lines are the mean relative bias for each beanplot whereas the dotted line is the mean across all beanplots.128

Figure 7-6: Beanplots of the relative bias for each species as function of the classification scenarios used in the Models B with unequal data. The bold lines are the mean of the relative bias for classification scenarios and the dotted lines are the average relative bias across the four classification scenarios. Each beanplot is computed from 9 values (three priors on *p* $\times$ three priors on *v*)..... 130

Figure 7-7: Beanplots of the relative bias distributions for Models B without classification scenarios Sc2.e with unequal data as a function of the *v* and *p* priors for each species. Each beanplot is made from 4 values (4 classification scenarios), the bold line being the mean relative bias of the bean plot. The dotted lines are the mean across all the prior combinations. 131

Figure 7-8: Bean plots of mean CV for all models as a function of species identities. The Y axis is displayed on the log scale. 134

Figure D-1 : Beanplots of the estimates relative bias as a function of the priors on *v* and *p* in the models for each species. The bold lines are the mean relative bias for each beanplot whereas the dotted lines are the mean of the relative bias across all models for one species. D-33

List of Tables

Table 1-1: Whistle frequency ranges for the species used in this thesis, with the location of recordings, the frequency limit of the recording system and the references. 16

Table 2-1: Number of sections S_j for each species used to train the classifier. The number of sections is dependent on the proportion of the data used to train the classifier. The first classifier used half of all the sections, the second a quarter and the third an eighth, whereas for the prediction, 100% of the sections are used to train the classifier. 33

Table 2-2: Δ AIC, AIC and residual sum of squares for the three models 35

Table 2-3 Estimated standard deviation by the least squares model 3 if 100% of the data were used to train the classifier. Values in brackets show the measured standard deviation by the PWC of Gillespie et al., (2013) when 2/3 of the data are used to train the classifier. BND=Bottlenose dolphin COD=common dolphin, RSD =Risso's dolphin, WBD=white beaked dolphin and WSD= white sided dolphin 36

Table 3-1: Main stages to create a whistle classifier and to apply it on new data using the PWC. 40

Table 3-2: Training dataset and the general location and sources which collected them. 42

Table 3-3: Groups of species classified for both classifiers. *2Sp* classifier discriminated Bottlenose dolphins from all other species pooled, whereas *5Sp* classifier discriminated between all five species. 42

Table 3-4: Details of EAR deployments from (SMRU ltd et al., 2011) 45

Table 3-5: Number of whistle contours extracted for each species in the training data set. 47

Table 3-6: Confusion matrix of the 2Sp classifier. The classification probabilities are the probabilities observed when 80% of the training data are used to train the classifier. The standard deviation (in %, within the brackets) is an estimation if 100% of the data were used to train the classifier. BND=bottlenose dolphins, Other=all other species. *p* being the p-value of a t.test with the alternative hypothesis being the true difference in mean is not smaller than by chance. 49

Table 3-7: Confusion matrix for the 5Sp classifier. The classification probabilities were the probabilities observed when 80% of the training data were used to train the classifier. The standard deviation (in % within bracket) was an estimation if 100% of the data were used to train the classifier. BND=bottlenose dolphins, COD=common dolphins, RSD=Rissos' dolphins, WBD=white beaked dolphins, WSD= white side dolphins. *p* being the p-value of a t.test with the alternative hypothesis being the true difference in mean is not smaller than by chance. 50

Table 3-8: Comparison of the EARs recording classification by the *2Sp* and *5Sp* classifier. Only EARs deployments with dolphins encounters have been processed with the 5Sp classifier. 53

Table 4-1: Numbers of visual detections with a high or assumed high confidence level on the French and Spanish vessels. ... 68

Table 4-2: Summary of the numbers (*n*) of recordings in total, with all the acoustic detections and when the false positive detections (FD) have been removed. Also summary of the total number of acoustic detections contours as well as the number of acoustic contours used for the rest of the analysis when the false positive detections were removed. 69

Table 4-3: Proportion (%) of detection contour lengths within the false detections (FD) and whistles (W) minutes below the contour length for each dataset. 71

Table 4-4: Numbers of whistles contours used in the whistle classifier for each species and datasets. 72

Table 4-5: *2Sp* French classifier with the classification probabilities when the classifier was trained with 80% of the French dataset. Standard deviations (%) are within the brackets. Species codes are the same as in table 1-4, with CSD = COD +STD + C&S pooled. P = p-value of a one-tailed t-test to test, the null hypothesis that the results were obtained purely by chance, $H0 = 50\%$ 74

Table 4-6: *4Sp* French classifier confusion matrix: Classification probabilities of the classifiers trained with 80% proportion of the French dataset. Standard deviations are within the brackets. p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H0 = 25\%$..... 74

Table 4-7: *3Sp* Spanish classifier confusion matrix with the classification probabilities when the classifier was trained with 70% of the Spanish dataset. Standard deviations (%) are within the brackets. Species codes are the same as in table 1-4, with CSD = COD +STD + C&S pooled. P = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H0 = 33\%$ 75

Table 4-8: *5Sp* Spanish confusion matrix with the classifiers trained with 70%n of the Spanish dataset. Standard deviations are within the brackets. P = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H0 = 20\%$. 76

Table 4-9: North Atlantic classifier confusion matrix with the classifiers trained with 80% of the dataset. Standard deviations are within the brackets. p = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H0 = 33\%$. 76

Table 4-10: Classification result of the French dataset classified with the Spanish and North Atlantic classifiers. 78

Table 4-11: Encounters classification of the French acoustic dataset classified with the Spanish classifier. 79

Table 4-12: Encounters classification of the French acoustic dataset classified with the North Atlantic classifier (*N.Atlantic*).. 79

Table 4-13: Encounter identification of the French acoustic detection, not associated with a visual detection, using the French and North Atlantic classifiers. COD=Common dolphin, C&S=Common and Striped dolphin, FPW=Pilot whale, STD=Striped dolphin, unidentified=when a section contain the same maximum classification probabilities between several species.81

Table 4-14: Spanish data classified with the *3Sp*, *2Sp* French classifiers and by the *North Atlantic* classifiers. CD=Common dolphin, CS=Common and Striped dolphin, PW=Pilot whale, SD=Striped dolphin. 82

Table 4-15: Encounters of the Spanish data classified with the *3Sp*, *2Sp* French classifiers. 83

Table 4-16: Encounters of the Spanish data classified with the *North Atlantic* classifier 84

Table 4-17: Classification result of the Spanish acoustic dataset classified using the Spanish and North Atlantic classifiers. Number of sections classified with the corresponding classification probability (%).CD=Common dolphin, CS=Common and Striped dolphin, PW=Pilot whale, SD=Striped dolphin. The number in bracket is the number of sections classified similarly by the Spanish and the North Atlantic classifiers. 85

Table 4-18: Summary of the numbers of encounter classified per species for the training dataset and the number of encounters classified by the classifier but for which the species identification was not known. 89

Table 6-1: The five different confusion matrixes (a - e) used during the simulation studies. Confusion matrix a is the identity matrix (no misclassification), b and c both have a high correct classification probabilities, but differ in that the misclassification probabilities of b are equal between species, whereas they are different in c. Confusion matrices d and e both have low rates of correct classification and again differ in that misclassification is equal between species in d, but varies in e. 102

Table 6-2 Summary of the scenarios tested in the simulation study; similar misclassification probabilities means that elements of the confusion matrix outside the diagonal are the same between species (scenarios x.b and scenarios x.d), whereas for different misclassifications rates, they are different between species (scenarios x.c and scenarios x.e). 103

Table 6-3: Examples of Dirichlet α parameters used for species A for each scenario. For the remaining species parameters α were the same but in different order to match the confusion matrices. 105

Table 7-1: Summary of the notation used in previous chapter. 114

Table 7-2: Prior parameters of the negative binomial prior distribution $V1, V2$ and $V3$..... 117

Table 7-3: Summary of all the investigated Bayesian models. Sc1.x and Sc2.x correspond to the scenarios of misclassification (Scx.a ,Scx.b, Scx.c, Scx.d, Scx,e) described in chapter 6. The prior parameters were described in the section 7.2.4. MH (Metropolis Hastings) and GS (Gibbs sampler) are the MCMC alogithms used in the models..... 120

Table 7-4 : Mean and standard deviation (in brackets) of the relative bias across all Models A and all Models B when the same number of detections for each species was simulated (Sc1,equal data) and when different number of detections between species were simulated (Sc2, unequal data). 126

Table 7-5: Mean relative bias (%) and their standard deviation in brackets across the priors on *v* for each species. The different colours represent the different level of misclassification: No misclassification (white), low misclassification (light grey) and high misclassification (dark grey)..... 129

Table 7-6: Mean CV for models with scenarios Sc1.a to Sc1e with equal data priors V1 and V2 and Models A and Models B. 132

Table 7-7: Mean CV for unequal data for the four species for the different classification scenarios (Sc2.b to Sc2.c), priors on parameters *v* (V1 to V3) and no priors (Models A P0) or priors on parameters *p* (Models B P1 to P3). 133

Table 8-1: Summary of the methods used to estimate the true number of sections. For each method the type of confusion matrix (*C*) used in the models is described: PAMG. mean is the mean confusion matrix given by the PWC at the end of a classification process whereas PAMG. samples is the confusion matrices of each bootstrap of the classification process, Dirichlet dist. is the confusion matrices generated from a Dirichlet distribution. Initial values indicates whether the initial values are needed (Y) or not (N) for the method. prior on parameters *v* and *p* describe the parameters needed for the prior distributions on *v* and *p* in the Bayesian models. 138

Table 8-2: Abundances estimation from the CODA visual survey (CODA, 2009) for each species (BND (bottlenose dolphin), COD (common dolphin), C&S (common and stripped dolphin), STD (stripped dolphin), FPW (long and short finned pilot whale), and each classification group. The encounter proportion for each classification group is the abundance for that classification group as a proportion of the total abundance of the 5 species..... 142

Table 8-3: Number of contours classified for each species and each classification group. The contour proprotion is the proportion of contours for a classification group relatively to the total number of contours of the 5 species. 142

Table 8-4: Mean and variance parameters (with associated CV in parentheses) of the prior distributions on parameters *v* for each species or classification group (CSD is common and striped dolphins). The number of observed sections *n* and the parameters α of the Dirichlet distribution for the prior distribution on the *p* parameters are also summarised. 143

Table 8-5: Number of contours *n* classified for each classification group (bottlenose dolphin (BND), common dolphin (COD), Risso's dolphin (RSD), white beaked dolphin (WBD), white side dolphin (WSD). The contour rate is the proportion of contours for a classification group relatively to the total number of contours of the 5 species. 144

Table 8-6: Mean and variance parameters (with associated CV in parentheses) of the prior distributions for each classification group (classification abbreviation similar to previous table) . Number of observed sections and the parameters α of the Dirichlet distribution for the prior distribution on the *p* parameters are also summarised. 144

Table 8-7: Number of observed sections detected by the DO1 deployment in the S.A.C, as well as the mean and variance parameters (with associated CV in parentheses) of the prior distributions on parameters *v* for each classification group (classification abbreviation similar to previous table). 145

Table 8-8: Number of observed sections detected by the EARs deployed in the MORL_BOWL sites, as well as the mean and variance parameters (with associated CV in parentheses) of the prior distributions on the parameters *v* for each classification group (classification abbreviation similar to previous table). 146

Table 8-9: *3Sp* Spanish confusion matrix, with the classification probabilities and their standard deviation (in brackets), discriminating bottlenose dolphins (BND), common and striped dolphins (CSD) and long and short finned pilot whales (FPW.) 147

Table 8-10: Confusion matrix, with the classification probabilities and their standard deviation (in brackets), of the *5Sp* classifier discriminating bottlenose dolphin (BND), common dolphin (COD), Risso's dolphin (RSD), white beaked dolphin (WBD) and white sided dolphin (WSD). 147

Table 8-11: Mean, CV and 95% credible interval (CI) of the estimated true number of sections for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors on parameters *v* estimated from the visual survey (p. f.surv) or from the proportion of whistle contours per species (p. f. cont) with variance parameters such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species. 149

Table 8-12: Mean, CV and 95% credible interval (CI) of the estimated true number of sections for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species 151

Table 8-13: Mean, CV and 95% credible interval (CI) of the estimated true number of sections detected by the DO1 deployment, for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species 152

Table 8-14: Mean, CV and 95% credible interval (CI)l of the estimated true number of sections detected by the EARs deployed in the MORL_BOWL sites, for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species 154

Table A-1 Classification result of the EAR data classified with the *2Sp* classifier: Encounters time: time of the first section of the encounter. n= total numbers of sections within each encounters of bottlenose dolphins (nBND) and other dolphins (nOTHER). p is the average probability of a section to be classified as bottlenose dolphins (pBND) or as other dolphins (pOTHER). Classified as: final classification of the encounter after observation by the manual observed. When all the contours within an encounter are false detections then the encounters was classified as a false detection (FD) encounters. A-1

Table B-1: Classification result of the French data classified with the 5Sp and 3Sp Spanish classifiers: BND=bottlenose dolphins, COD=common dolphins, C&S=common/striped dolphins, FPW=pilot whales, STD=Striped dolphins, n=number of sections per encounter, p= classification probabilities per classification group. Class as= classification result by the 5Sp classifier in comparison to the classification result by the 3Sp classifier B-14

Table B-2: Classification results of the Spanish data classified with the 4Sp and 2Sp French classifiers with n, nCOD, nC&S, nFPW, nSTD, nCSD being respectively the total number of sections per encounters for all species and the number of section for common dolphins, common/stripped dolphins, pilot whales, striped dolphins and commons and striped together. pCOD, pC&D, pFPW, pSTD, pCSD being the classification probabilities per classification group and Class as is the classification result per encounter. B-16

Table B-3: Classification of the French encounters, not associated with visual detections, with the 2Sp French classifier and the North Atlantic classifier. n=number of sections per encounter in total (n) and per classification groups (nCSD,nFPW etc..). p=classification probability per classification groups (pCSD,pFPW...).VisualDet=statute of the visual team during the encounters: On effort=visual team was on effort but they did not detect the animals, Off effort=the visual team was Off effort, sonar or electric = description of the sound generating false detections, species name at time=when a species has been observed by the visual team close to the encounter time. B-19

Table B-4: Classification of the Spanish encounters (not associated with visual detections) with the 3Sp Spanish classifier and the North Atlantic classifier. n=number of sections per encounter in total (n) and per classification groups (nBND,nCSD etc..). p=classification probability per classification groups (pBND,pC&S…). VisualDet=statute of the visual team during the encounters: On effort=visual team was on effort but they did not detect the animals, Off effort=the visual team was Off effort, sonar or electric = description of the sound generating false detections, species name at (time)=when a species has been observed by the visual team close to the encounter time. B-20

Table C-1: Analytically derived mean of the expected true number of calls, $\mathbf{E}[\mathbf{v}]$, and coefficient of variation (CV, expressed as a percentage). C-25

Table C-2: Simulation result, without uncertainty in the confusion matrix, of the mean of the estimates of the true number of calls $\mathbf{E}[\mathbf{v}]$, and coefficient of variation (CV, expressed as a percentage). C-26

Table C-3: Simulation result, with a low level of uncertainty in the confusion matrix, of the mean of the estimates of the true number of calls $\mathbf{E}[\mathbf{v}]$, and coefficient of variation (CV, expressed as a percentage). C-27

Table C-4: Simulation result, with a high level of uncertainty in the confusion matrix, of the means of the estimates of the true number of calls $\mathbf{E}[\mathbf{v}]$, and coefficient of variation (CV, expressed as a percentage). C-28

Table D-1: Values of the parameters α for the Dirichlet prior distributions, for each species, each scenario and each set of priors for p. The parameters α were selected such that: the CV of the correct classification probabilities (diagonal element) was equal to 1%, 40% and 77% and the means of the prior distribution were equal to the classification probabilities of the scenarios for P1 and P2 whereas for prior P3, the mean distribution was equal to 0.25 for each species. D-29

Table D-2: Convergence test results for each model A and species. Y indicates that the chains for the corresponding species converged. The 0 value in Sc2.d and Sc2.e with prior V3 indicates that the posterior distribution for species D had stopped converging and the mean of this posterior distribution was 0. Grey cells indicate models that were found to be sensitive to the initial values of the Markov chains. D-31

Table D-3: Summary of the convergence test results for the Posterior distribution of the parameters *v* and *p* for all models B. Y indicates that the chains for the corresponding species converged whereas N indicates they did not converged.. Sc=Scenario for the different confusion matrices (Scx a to Scx e) and for the equal(Sc1.) and unequal dataset Sc2. The grey cells indicate models sensitive to the initial values of the Markov chains. D-32

Chapter 1: General Introduction

1.1. Background

The viability of many populations in all taxonomic groups is threatened by anthropogenic disturbances, such as habitat loss and degradations, harvesting (for hunting or gathering for food, medicine, fuel and material), diseases, accidental mortalities due to interaction with human activities, pollution and/or climate change (Schipper et al., 2008; Stuart et al., 2004). To protect them, environmental managers and policy makers have the responsibility to seek advice and gather information from scientists to create policies and to organise management actions which will hopefully help the preservation and conservation of these natural ecosystems. Ecosystems are complex, non-linear and influenced by stochasticity, it is thus difficult for scientists who try to understand them to predict their natural dynamism accurately. Anthropogenic disturbances and current management strategies add other levels of complexity; and given this complexity the outcome of scientific analysis and advice contains numerous sources of uncertainty that environmental managers and policy makers need to consider when they make decisions. By identifying the origin of these uncertainties, characterising them, quantifying them and finally understanding their impact on particular management actions scientists will help decision-makers to make cost-effective decisions to minimise potential risks to the environment. Four sources of uncertainty are commonly recognised (Akçakaya et al., 2000; Boyd et al., 2010; Harwood and Stokes, 2003):

Natural uncertainty: This uncertainty is a consequence of the natural demographic and environmental stochasticity (Akçakaya et al., 2000; Harwood and Stokes, 2003).

Measurement error: This uncertainty is a consequence of inaccuracy and imprecision during data collection or in the estimation of the parameter of interest. Most of the time only a sample of the observations of interest is collected. The choice of the sampling strategy or the method of statistical inference used to estimate the parameter of interest from the observations generates this uncertainty.

Model error: Models are regularly used to describe complex natural processes, to better understand their mechanism and/or to predict how this system will change in the future. Given the complexity of natural processes, models can only be an approximation of reality and thus they provide an incomplete representation of the reality. Model errors come from the differences between the model and the reality. (Harwood and Stokes, 2003).

Implementation errors: These errors are a consequence of errors in the management strategy, for example these could arise from delays in the establishment of protected areas, inadequate protection within them, imperfect policy implementation and/or unpredicted changes which generate a failure to reach the management objectives (Ellison, 1996).

Given these uncertainties, risk assessment frameworks have been defined to help in the decision process. A risk is the probability that a hazardous outcome will happen and risk assessment is the quantification of this probability (Rowe, 1977). If there was no uncertainty then a scientist would be certain about the outcome and there would be no risk. The role of conservation scientists is to use robust methods to measure the probability of an outcome that will characterise and incorporate all these uncertainties. The role of environmental managers and policy makers in view of the uncertain outcome is to decide if the risk is acceptable or not, and if it is not, to propose new management strategies which will minimise the risk and optimise the balance between social, economic and ecological objectives.

Complex mathematical models, often called "operating" models, have been developed for this purpose (Harwood and Stokes, 2003). These models are a combination of three types of models: a *process model*, describing the underlying biological process with factors influencing this process, an *observation model*, illustrating the data collection and analysis, and finally a *management model*, simulating the effect of management decisions on the biological model (Harwood and Stokes, 2003). These models are used to test the performance of different management options.

Either a frequentist or a Bayesian statistical framework can be used to develop such models. However the interpretation of the result will be different depending on the statistical approach used. To illustrate these differences, consider a model M describing a system of interest. Conventional frequentist statisticians will establish whether the null hypothesis (H_0 : data x come from the model M) is rejected or failed to be rejected at α-level of significance. This framework does not give information about the actual probability of obtaining the model given the data $(M|x)$ (Ellison, 1996). A Bayesian framework, based on the Bayes' theorem (Bayes and Price, 1763)(Eq1-1) estimates this probability:

$$P(M|x) \propto P(x|M).P(M) \qquad (1\text{-}1)$$

Where $P(x|M)$ is the likelihood function describing the data, $P(M|x)$ is the posterior distribution representing the probability of obtaining the model given the data and current information about the model M ($P(M)$).

The Bayesian framework provides decision-makers a probability of the outcome. In this particular example, the outcome is the probability that the model describes the system given the data and prior knowledge of the model *M*. It is then the responsibility of the decision makers to interpret this outcome (probability) within a risk assessment framework to determine if the risk of damage or disturbance is too high, and that potential irreversible damage will occur to the system. It is the presence of the prior distribution that makes Bayesian inference a good tool to be used during the risk assessment procedure. The choice of the prior distribution variance for a given parameter generates its level of uncertainty. Running a sensitivity analysis comparing the outcome of similar models with different prior variances will help in evaluating the consequences of parameter uncertainty, and to select the model generating the lowest acceptable risk. If models are sensitive to the prior, then every effort should be made to collect more information to reduce the variance of the priors (Harwood and Stokes, 2003).

1.2. Abundance estimation a tool for management strategy

1.2.1. Generalities

Article 1.a of the European Habitats Directive defines conservation as "a series of measures required to maintain or restore the natural habitats and the populations of species of wild fauna and flora at a favourable status". The notion of "a favourable status" for a species is defined in Article 1.i and refers to the idea of "maintaining a population or species on a long-term basis as a viable component of its natural habitats", neither reducing nor likely to reduce their habitat range and ensuring there is and will continue "to be a sufficiently large habitat to maintain its population on a long-term basis" (European Union, 1992). Risk assessment methods such as Population Viability Analysis (Gilpin and Soulé, 1986) used to predict the probability of extinction within a particular interval of time, and Management Strategy Evaluation (Punt, 1992) used in fisheries management to evaluate the expected performance of harvest strategies, are frequently used in conservation biology to achieve the Habitats Directive objectives.

Maintaining, and/or restoring a population, predicting a probability of extinction or measuring the impact of harvesting strategies, all require knowledge of the size of the current

population. Conservation strategies consist of measuring trends in population size, and considering if this size is large enough to insure the existence of the population in the long term.

Ideally to measure the exact population size of a species it would be necessary to detect all the animals within the population. In practice due to, for example, the behaviour, the habitat or the distribution range of species, it is rarely possible to do so. Thus, abundance must be estimated from a sample of the population. Once the number of animals has been counted and identified, it is then necessary to extrapolate these counts to estimate the abundance of the species. Given that in the majority of the counts not all animals can be detected, an intuitive estimator of abundance, $\widehat{N}$, assuming the entire habitat range of the species is surveyed, is given by:

$$\widehat{N} = \frac{n}{\hat{P}}$$

where n is the number of animals detected and $\hat{P}$ represents the estimated probability of detecting an animal (Buckland et al., 2001). Depending on the approaches used to detect and count individuals, $\hat{P}$ can be estimated by different methods (Borchers et al., 2004; Buckland et al., 2001). One common method used to estimate $\hat{P}$ is the distance sampling theory described in detail by (Buckland et al., 2001, 2004). This theory is based on the principle that the probability of detecting an animal decreases with the distance between the animal and the observer. This method consists of surveying randomly placed transects (line transect sampling) or randomly placed points (point transect sampling) (Borchers et al., 2004; Buckland et al., 2001) and counting the number of animals detected along them, and measuring the distance between the animal and the line or point. The basic formula to estimate abundance becomes:

$$\widehat{N} = \frac{nA}{aP_a} \qquad (1\text{-}2)$$

with n being the number of detected animals, a is the surveyed area (area of all the transects), A is the total area of interest and finally P_a is the mean probability of detecting an animal. In this formula only P_a is unknown. In the simplest model, the only factor influencing P_a is assumed to be distance from transect or sample point, i.e. $g(x)$ being a function linking the probability of detecting an animal to its distance from the line or point. This basic theory is based on four key assumptions (Buckland et al. 2001):

1. all animals directly on the transect line or at the sample point are certain to be detected (g(0)=1;
2. animals do not move before detection in reaction to the observer or observation platform;
3. the distance of a detected animal from the transect or sample point is measured accurately;
4. detections are independent events.

1.2.2. Abundance estimation of cetaceans

Depending on the species for which this abundance estimation method is being used, some or all these assumptions can be violated. With cetacean species, the four key assumptions of the distance sampling theory are violated.

Marine mammals and particularly cetaceans spend all their time in the water and most of the time underwater (Boyd et al., 2010). They can be visually detected when they come to the surface to breath. However, during their underwater time some species are extremely vocal (Richardson et al., 1995). Odontocete species produce vocalisations generally grouped into three categories: whistles (frequency modulated sounds which vary with time), clicks (very short broad band sounds), and long pulsed sounds also referred to as burst pulse calls (Richardson et al 1995). Depending on species these vocalisations can be detected up to few tens of kilometres. Baleen whale species produce sounds (moans, calls) detectable up to several hundreds of kilometres (Sirovic et al., 2007).

Violation of assumption 1

Cetaceans that are on the transect line or at the sampling point may be missed because of availability bias or perception bias:

- Availability bias happens when the animal is not detectable. With visual detections, the situation happens when cetaceans are under the water. For acoustic detections, availability bias occurs if the species does not vocalise or chooses not to vocalise, for example sperm whales and beaked whales vocalise essentially during their dive (Barlow and Taylor, 2005; Johnson et al., 2006), for humpback whale, males vocalise mainly during breeding season whereas female vocalise very rarely (Vu et al., 2012).
- Perception bias occurs when the animal is detectable but missed by the observer. With visual detection this situation happens when the animals are at the surface but are not

detected by observers. With acoustic detections this type of bias occurs when animals are vocalising but are not detected (acoustic missed detection). Acoustically, a missed detection can arise if the vocalisation is not loud enough to be detected, or if the vocalisation is very directional and the animal is not pointing the hydrophone (Zimmer et al., 2008).

Violation of assumption 2

This assumption is regularly violated as many cetacean species have been observed avoiding the survey platform (Au and Perryman, 1982; Barlow, 1988) or are attracted to it (Buckland and Turnock, 1992). These behaviours have consequences for both visual and acoustic detections.

Violation of assumption 3

The accuracy of the distance measurement is dependent on the reliability of the method used to estimate the distance. It will nearly always be an estimate as it is difficult at sea to have an exact measurement (Gillespie et al., 2010; Leaper et al., 2010).

Violation of assumption 4

This assumption is violated in situation where, for example, species live in groups. Thus, if one animal is detected then the probability of detecting other individuals within the group may rise after the first detection because it is difficult for the observer not to look harder in the area of the first detection.

1.2.2.a Abundance estimation from visual detections

Nevertheless, distance sampling theory is one of the most common methods used to estimate abundance of cetaceans (Boyd et al., 2010). This is possible because a lot of work has been carried out to make distance sampling methods robust to these violations when being used with visual detections.

Line transect sampling has been combined with capture-recapture theory to estimate g(0) despite the perception bias of the observers (Borchers et al., 2004; Borchers and Samara, 2007; Buckland et al., 2004; Skaug and Schweder, 1999). In this approach, different observers survey the same area from two independent platforms. Each observer records their detections, and detections from all observers are then compared. Detections that have been made by both observers are recorded as duplicates, and correspond to recaptures in capture-recapture theory (e.g Canadas et al., 2005; Hiby and Hammond, 1989; Hiby, 1999).

Other methods have been developed to deal with the problem of availability bias for visual detections. For cetaceans this availability bias is often dependent on the surfacing behaviour of the species. If the animals become available for an instant only, then it is necessary to account for this in the abundance estimation formula by adding a component modelling the probability of being available while within a detectable range (Buckland et al., 2004; Skaug and Schweder, 1999). If the animal is available for detection for some time and its availability changes when it is within detectable range (for example a sperm whale can stay at the surface for up to 10 minutes before diving for periods of 50 minutes or more) then its availability is classified as 'intermittent' (Buckland et al., 2004). In this situation the component of the abundance estimation function modelling availability should model the process of becoming available and the duration of availability. Borchers and Samara (2007) developed a line transect sampling method using a hidden Markov model to deal with intermittent availability. Their method modelled the probability of detecting an animal available at time *t*, as a function of its probability of being available at time *t-1*.

Buckland and Turnock (1992) developed a survey approach to accommodate violations of the second and third assumptions. Using their approach, two independent platforms, the tracker and primary platforms, survey different areas ahead of the vessel to account for responsive movements by the animals to the approaching survey platform. The tracker platform uses high power binoculars (Big Eyes) to survey an area well ahead of the vessel with the objective of detecting animals before they respond to the boat (Buckland and Turnock, 1992; Hedley et al., 1999). This estimation method deals with the responsive and/or random animal movement and reduces the dependence between detection which can rise from un-modelled variables such as animal surfacing behaviour (Hedley, 2000).

The violation of the fourth assumption is not important in practice as robust methods have been developed to deal with it (Buckland et al., 2010).

1.2.2.b Abundance estimation from passive acoustic detections

1.2.2.b.i Visual versus Acoustic detections

For some species, detecting cetaceans by the sounds they produce is often a more efficient method than detecting them visually, and practically it offers many advantages over visual methods. Acoustic detections are independent of daylight and they are less dependent on environmental conditions (visual detections are dependent on distance of visibility, sun glare, sea state) (Palka, 1996). Another advantage is that the acoustic detection process can be fully

automated (Baumgartner et al., 2008; Gillespie and Chappell, 2002; Mellinger and Clark, 1997; Mellinger et al., 2011) while automated detection and species recognition form visual recordings is in its infancy. However, in practice it is only in the last decade or two, with the improvement of underwater recording systems and computer technology that the interest in using acoustics to detect cetaceans has rapidly grown. Passive acoustic monitoring (PAM), the recording and analysis of sounds emitted by species, is more widely used than active acoustics to detect cetaceans. Passive acoustic methods may use stationary hydrophones (autonomous or cabled) (Mellinger and Clark, 1997; Sousa-Lima et al., 2013), which can record what is happening in a specific area over a longer period of time and at a relatively low cost, or towed hydrophones, which allow a wider spatial coverage and can be used in association with visual observations.

1.2.2.b.ii Abundance estimation from fixed hydrophones: cue counting methods

Using acoustics to detect cetaceans is a relatively recent innovation and consequently estimating abundance from acoustic detections is in its infancy (Marques et al., 2013). Currently most of the methods used to estimate abundance of cetaceans from acoustic detections are based on distance sampling theory used for visual detections. This method needs to be modified before it can be properly used with the acoustic detections. Indeed the basic formula 1-2 in distance sampling is based on the number of animals *n* visually detected. With acoustic detections, one animal can produce numerous vocalisations in a short period of time. Animal abundance can be estimated using *cues***,** where cues are defined as instantaneous availability events (Buckland et al., 2004). Acoustic detections, particularly vocalisations from cetaceans, can thus easily be defined as cues when estimating animal abundance because they are not produced continuously. The description of a cue can be species dependent; for example a blue whale (*Balaenoptera musculus*) call is considered a cue (Moore et al., 1998), whereas for humpback whales (*Megaptera novaeangliae)* a song unit is considered a cue (Swartz et al., 2003). For echolocating species, one click of a beaked whale is considered as a cue by Marques et al. (2009) whereas Kyhn et al. (2012) used a click train from a harbour porpoise as a cue. For whistling species a cue could be considered to be one whistle (Ansmann et al., 2007). However with cetaceans a cue does not have to be necessary the vocalisation produced by the animals, Moretti et al. (2010) used the acoustic component at the beginning of a Beaked whale dive as a cue whereas for example Hiby

(1985) used the surface behaviour of great whales to estimate their abundances. From these cues, Hiby (1985) developed a cue counting theory to improve the detection of whales during line transect surveys. This theory is derived from distance sampling theory and is often referred to as the cue-counting distance sampling method (Buckland et al., 2001; Marques et al., 2011). If only cues are used in the abundance estimation formula then it is the abundance of cues, that will be estimated and not the abundance of the population. To overcome this issue several approaches have been proposed for estimating abundance from acoustic detections / acoustic cues. These approaches fall into two broad categories; firstly, those dealing with acoustic cues from stationary hydrophones, and secondly those dealing with detections from towed hydrophones.

Marques et al (2011) proposed a method based on cue counting theory to estimate the density of right whales (*Eubalaena japonica)* in the Bering Sea detected by stationary hydrophones:

$$\widehat{D} = \frac{n_u\left(1 - \hat{f}_p\right)}{a_c \hat{P} T \hat{r}} \qquad (1\text{-}3)$$

where n_u was the number of detected right whales calls in *T* hours within the covered area a_c, $\hat{r}$ represented the call rate per individual, $\hat{P}$ the detection probability within a_c and $\hat{f}_p$ corresponded to the estimated proportion of false positive detections. In this formula $n_u\left(1 - \hat{f}_p\right)$ corresponded to the true number of calls detected and $\frac{n_u\left(1-\hat{f}_p\right)}{T\hat{r}}$ measured the number of individual *n* of the abundance estimation equation 1-2 with visual detections.

The density estimation $\widehat{D}$ (and consequently the abundance estimation) was dependent on the estimation of three parameters ($\hat{P}$, $\hat{r}$ and $\hat{f}_p$) which required independent analysis to be obtained.

To estimate the detection function $\hat{P}$ it is necessary to estimate the distance of the vocalising animal to the hydrophones. In this paper they used a predictive acoustic propagation model from a single hydrophone to estimate the distance. Other authors have used a variety of physical or mathematical models to estimate the distance of the vocalising animal from a single hydrophone (McDonald and Fox, 1999) or array of hydrophones (Harris, 2012; Thode et al., 2012); these include hyperbolic techniques, waveguide models (Wiggins et al., 2004) and multipath propagation models (Tiemann et al., 2004).

Marques et al., (2013), in a review of passive acoustic density estimation methods, recommended that cue rates should be estimated in the survey area, while the survey was

being conducted, and over a large and random sample of animals. Indeed acoustic cue rates have been shown to vary as a function of time of day, (Boisseau et al., 2008; Gordon et al., 2000; Matthews et al., 2001), group size (Ansmann et al., 2007), and behaviour. These sources of variation make cue rate a difficult parameter to estimate accurately.

In equation 1-3, $\hat{f}_p$ is a parameter estimating the probability of false positive detections. Detections are classified as false positives when they have been identified by the detector as vocalisations made by the species of interest, but in reality these sounds were not. A false detection is generally generated by the presence of a sound with characteristics similar to the sound of interest such that the detector cannot differentiate them**.** These sounds could be either other biological sounds made by another species or associated with the environment or it could be anthropogenic sounds such as boat noise, electrical noise, sonars, or echo sounders. If false positive detections are not identified and removed, the number of vocalisations from the species of interest will be over-estimated.

1.2.2.b.iii Abundance estimation from towed hydrophones

Abundance of cetaceans estimated from towed hydrophones has been estimated principally for sperm whales (Barlow and Taylor, 2005; Borchers et al., 2007; Lewis et al., 2007) and porpoise species (Gerrodette et al., 2011; Gillespie et al., 2005). For both species the method used was the same as used for visual line transect theory and visual detections. These species have some of the most distinctive vocalisations of all the cetacean species making them easy to detect automatically with low false positive detection rates. Similarly their vocalisation rates are very predictable.

Sperm whales produce clicks almost continuously during their dive with a constant inter-click interval. This regularity makes it is easy for a manual operator to identify individuals in the same way as a visual operator does with a surfacing animal. This regularity also allows measurements of the bearing (angle between the hydrophones and the vocalising animal) to be estimated by measuring the time delay between detections at a pair of hydrophones. The intersection point of consecutive bearings is then used to estimate the distance between the animal and the track line (Leaper et al., 2000). However this distance could be a source of bias when used in distance sampling theory. Indeed, ideally the distance needed to have a robust abundance estimate is the horizontal distance projected to the surface and not the perpendicular distance to the transect lines. To obtain horizontal distance the depth of the

animal needs to be known. In the current literature, most of the abundance estimation studies (Leaper et al., 2000; Lewis et al., 2007) using towed hydrophones ignored this factor whereas some authors demonstrate that in their study the average distance between the animal and the hydrophones is such that the difference in slant and horizontal distance is small enough to be ignored (Barlow and Taylor, 2005).

Porpoises do not produce regular clicks but do produce frequent sequences of clicks (click trains) (Linnenschmidt et al., 2013). It is generally easy for an operator looking at a display of bearing versus time to identify click trains and thus individuals. However, there are difficulties in estimating the number of individuals accurately, particularly when there is a group of several animals. In this case the detection unit can be a group and a new parameter needs to be added to Eq. 1-3 specifying the average group size. To the best of my knowledge no abundance estimate of species other than sperm whales or porpoises have been made using data from towed hydrophones only. A current study on minke whale is ongoing (Norris et al., 2010).

1.2.2.b.iv Classifiers

As well as modifying visual distance sampling theory, estimating abundance from acoustic detections also requires improvements in the methods used to detect and identify (classify) sounds. Visual detection and identification is dependent on environmental conditions, species behaviour and observer competence. Although, reliable and consistent automatic detection systems have been developed for marine mammal vocalisations (Baumgartner et al., 2008; Gillespie and Chappell, 2002; Mellinger and Clark, 1997; Mellinger et al., 2011), that are largely unaffected by most environmental conditions, these detectors lack the ability to immediately identify the vocalising species. While some species produce easily identifiable vocalisations, e.g. sperm whale, harbour porpoise, humpback whale, blue whale, the majority do not, and produce vocalisations that are difficult to differentiate (Oswald et al., 2003; Rendell et al., 1999).

In the early days of passive acoustic detection, species were identified by listening and observing the spectrogram of their recorded sounds (Clark et al., 1996; Thomas et al., 1986). This process is time consuming and only possible if the observer is very familiar with the entire vocal repertoire of each species, or if the species has very specific vocal characteristics. With the improvement of passive acoustic monitoring systems, it is now common to record

terabytes of acoustic data after only a few weeks of recording. With such large data volumes, manual identification and classification of individual cues is not practical or feasible. Over the last two decades classifiers (Gillespie and Caillat, 2008; Gillespie et al., 2013; Nanayakkara et al., 2007; Oswald et al., 2007; Roch et al., 2007) have been developed to automatically identify species from their vocal characteristics. One advantage of these automatic classifiers is their ability to classify gigabytes of data in few hours. On the other hand, they may not be as accurate as a human operator. The accuracy of identification is a species specific problem, and some sounds are more difficult to identify than others, for example, whistles from pelagic delphenid species are more difficult to identify to species than blue whale calls.

Among the current classifiers developed it is possible to identify three common stages for the creation of these classifiers. For each stage different methods specific to the classifiers can be used:

1. *Feature or variable extraction*: For each species to be identified / discriminated, some physical characteristics are extracted from vocalisations recorded concurrently as an observer was visually identifying the species. For click vocalisations these parameters can be peak frequency (maximum frequency), click length, frequency bandwidth (Gillespie and Caillat, 2008; Soldevilla et al., 2008). For whistle vocalisations, peak frequency, number of inflexion points within the whistle, start and end frequency are commonly used (Oswald et al., 2003; Rendell et al., 1999). Parameter extraction can be done manually (Oswald et al., 2003) or automatically (Gillespie et al., 2013)
2. *Statistical selection of the most appropriate classification algorithm*: Once these variables are extracted a statistical method is used to find the best algorithm which will identify each species. Some of the most used methods are linear discriminate function analysis (Gillespie et al., 2013; Oswald et al., 2003), neural network process (Mellinger, 2008; Potter and Mellinger, 1993; Thode et al., 2012) and tree classification (Gillespie and Caillat, 2008; Oswald et al., 2003).
3. *Efficiency testing*: Finally, this algorithm is tested with data where the species has previously been reliably identified, to measure and report the efficiency of the classifier. When only two species are classified, this efficiency can be represented by a curve called Receiver Operating Characteristic (ROC) (Fawcett, 2006), representing the false negative versus the false positive rates. When more than two species are classified, the accuracy of the classification can be illustrated by the correct

classification probability of each sound only (Gillespie and Caillat, 2008; Roch et al., 2007; Soldevilla et al., 2008) or it can be expressed in a matrix called *confusion matrix*. This confusion matrix has the useful advantage of expressing both correct classification probabilities and misclassification probabilities. In a square confusion matrix of dimension $m \times m$, m representing the number of classification groups, here the number of species to discriminate, and each element of the matrix p_{ij} is the probability of classifying species j (column) as species i (rows). In particular, the entries for $i=j$ represent the probabilities of correctly classifying a species (success) and the off-diagonals ($i \neq j$) are probabilities of incorrectly classifying species j as species i (failures or misclassification). A small p_{ij}, $\forall$ i≠j, means a low misclassification probability of species j as species i while a large $p_{ij,}$ $\forall$ i≠j, means a high misclassification probability. On the other hand, a small p_{ij}, $\forall$ $i=j$, means a low correct classification probability of species j and vice versa for a high p_{ij}, $\forall$ $i=j$, Hence, the confusion matrix is given as

$$C = \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mj} & \cdots & p_{mm} \end{pmatrix}$$

where $\sum_j p_{ij} = 1 \forall 1 \leq j \leq m$. The confusion matrix quantifying the misclassification between species is a precious tool to be able to measure the false positive detection probabilities for each species.

Once created a classifier is used to associate to new acoustic detections a species of the classification group of the classifier.

1.3. Thesis outline

The aim of this PhD is to modify current methods for classifying whistle vocalisations and to develop new methods for estimating the correct number of whistle vocalisations detected by a hydrophone, with an objective of using these detections to estimate animal abundance.

Several studies have previously estimated abundance of cetacean species from clicks or calls (e.g. Gerrodette et al., 2011; Gillespie et al., 2005; Marques et al., 2011, 2009; McDonald and Fox, 1999), but to the best my knowledge this has never been done using whistles.

Most of the odontocete cetaceans produce whistles. Whistles are a frequency modulated signal emitted mainly for communication. These sounds are highly variable within (Rendell et al., 1999) and between species (Rendell et al., 1999; Steiner, 1981). So the first challenge in using whistles to estimate abundance is to identify the species producing the detected whistles. In the first part of this PhD (chapters 2 to 5) whistle classifiers are developed. The objective of these chapters was not to develop yet another whistle classifier technique, but to identify those parameters influencing the quality of the classifier, and to establish a method to quantify the uncertainty of the classification probabilities due to measurement error. In chapter 2, the current PAMGUARD Whistle Classifier (http://www.pamguard.org/) developed by Gillespie et al. (2013) is modified to develop a new method to quantify the uncertainty of the classification probabilities. In chapters 3 and 4 this modified classifier is applied to data to identify which features of an acoustic dataset are important to obtain a reliable classifier. The datasets in chapters 2 and 3 were compiled from data recorded from towed hydrophones towed by several small survey platforms operating around the coast of Scotland with the specific objective of developing a classifier to identify the presence of bottlenose dolphin (*Tursiops truncatus)* (a protected species in European waters) in some potential wind farm sites. The dataset for chapter 4 was collected from towed hydrophones during a large scale survey organised to assess the impact of bycatch on some cetacean species with a view of providing recommendations on safe bycatch limits for the common dolphin (*Delphinus delphis*) (CODA, 2009). Chapter 5 is a general discussion around the previous three chapters.

From the literature a summary of the vocalisation frequency range of all the species classified within these three chapters are presented in Table 1-1. Papers of data collected from wild animals and in the North Atlantic and preferably close to the British isles were preferred when possible. When no reference was found with these criteria then references from data collected in other oceans are used. All the referred studies used different type of hydrophones with the maximum frequency detected specified in the table. This table highlights the large overlap of the whistle frequency ranges between species.

The nature and the variety of whistles, means that a perfect classifier will never exist, and that a confusion matrix of a classifier will always have misclassification probabilities greater than zero (non-diagonal elements $\neq 0$). The consequence of such classification probabilities is that the observed number of detections for a species *i* after classification is the sum of detections

correctly identified plus detections of other species misidentified as species *i.* So the observed number of detections of species *i* is a biased estimate of the actual number of detections of that species and should then not be used directly to estimate abundance. So the second part of this PhD (chapters 6 and 8) proposes three methods evaluated on simulated data (chapters 6 and 7) and then applied to real data (chapter 8) to estimate the true number of detections for each species from the observed detections after classification. The principal objective of these chapters was to investigate the impact of different misclassification probabilities and varying amounts of uncertainty within the confusion matrix, on the reliability and precision of estimated true number of detections. In chapter 6 analytical and heuristic methods are used to conduct this investigation whereas a Bayesian framework is used in chapter 7. Finally chapter 8 applies these methods to some of the data used in chapters 3 and 4.

Chapter 9 summarised the results of the three chapters in a general discussion about the impact of misclassification and concludes this thesis.

Table 1-1: Whistle frequency ranges for the species used in this thesis, with the location of recordings, the frequency limit of the recording system and the references.

	Whistles frequency range (kHz)	**Recorder frequency limit (kHz)**	**Location**	**References**
Bottlenose dolphin *Tursiop truncatus*	7.3-16.2	20	North Atlantic	(Steiner, 1981)
Common dolphin *Delphinus delphis*	3.56-23.51	48	British Isles	(Ansmann et al., 2007)
Striped dolphin *Stenella coeruleoalba*	8.1-14.8	22	Tropical East Pacific	(Oswald et al., 2003)
Short finned pilot whale *Globicephala macrorhynchus*	6.32-8.69	15	Caribbean	(Rendell et al., 1999)
Long finned pilot whale *Globicephala melas*	4.15-8.86 2.821-4.72	15 20	Mediterranean Atlantic	(Rendell et al., 1999) (Steiner, 1981)
White beaked dolphin *Lagenorhynchus albirostris*	3-35	44	Iceland	(Rasmussen and Miller, 2002)
White sided dolphin *Lagenorhynchus acutus*	8.21-12.14	20	Atlantic	(Steiner, 1981)
Risso's dolphin *Grampus griseus*	6.63-13.41	15	Azores	(Rendell et al., 1999)

1.4. Overview of Bayesian theory

1.4.1. Introduction

The objective of the second part of this thesis is to estimate the true number of whistle detections for several species from the number of observed whistle detections, an unknown number of which are misclassified. For a fixed survey period, the true number of acoustic detections is mainly dependent on three parameters, the number of individuals producing sounds, the call rate of the species and the detection probability of the whistle detector. Some prior knowledge about these different parameters is sometimes available from previous surveys or analysis. Although this prior knowledge can occasionally be very accurate, it is most of the time very vague. As explained above in section 1.1, Bayesian methods provide a well-adapted framework to analyse the impact of uncertainty of model parameters on the precision of the outcome variables. This section provides a detailed description of the principles of Bayesian theory.

1.4.2. Bayes' theorem

The Bayesian approach was first introduced at the end of the eighteen century by mathematicians, such as Bernoulli, Bayes and Laplace (Fienberg, 1992).

Bayesian statistics make inference about a parameter θ conditioned on the observed data X and on some knowledge about θ which is assumed to be gained prior to the observation of the data (Gelman et al., 2004). The data are seen as fixed and the inference on θ is based on the posterior distribution, $\pi(\theta|X)$, which is the conditional probability of θ given X. This posterior distribution comes from the application of the Bayes' Theorem (Bayes and Price, 1763).

$$\pi(\theta|X) = \frac{f(X|\theta)\rho(\theta)}{f(X)}, \qquad (1\text{-}4)$$

where $f(X|\theta)$ is the likelihood (as it is used in classical frequentist statistics), $\rho(\theta)$ represents the prior distribution and $f(X)$ is the function of the data, independent of θ. Because the data are considered as fixed, $\rho(X)$ can be considered as constant and Eq 1-4 can be formulated as (Gelman et al., 2004)

$$\pi(\theta|X) \propto f(X|\theta)\rho(\theta)$$

1.4.3. Elements of Bayesian analysis

1.4.3.a Prior distribution

The prior distribution, $\rho(\theta)$, represents the initial knowledge that we have about the parameter of interest, before the data are observed. In the absence of prior knowledge, an uninformative or "vague" prior is used. Uninformative priors are selected such that they have a suitable large variance (Gelman et al., 2004). In this case, the inference on the parameter θ depends mainly on the data. However, when information about θ that has been gained independently of the data (for example, from experts' opinion and/or previous studies) is available, the prior distribution can be chosen such that it is 'informative'_ in other words, a suitable prior probability distribution is selected that expresses the available information as accurately as possible. The data, via the likelihood function, will help refine the prior distribution to obtain the posterior distribution. If the data are sufficiently informative, the actual choice of the prior should have little influence on the posterior distribution that is derived in the end.

When the posterior distribution is of the same family of probability distribution as the prior, then the prior is called a *conjugate* prior for the likelihood. The Dirichlet distribution is an example of a conjugate prior for the multinomial likelihood (Gelman et al., 2004). Conjugate priors are a useful tool in Bayesian analysis as they facilitate the use of a Gibbs sampler (see section 1.4.3.c.ii).

It is possible to conduct a prior sensitivity analysis to assess the sensitivity of the outcome of the Bayesian analysis, e.g. the mean of the posterior distribution(s) or some other summary statistic, with respect to the choice of the prior distribution. A simple prior sensitivity analysis consists of varying the parameters of the prior distribution, for example by systematically increasing or decreasing its variance. Subsequently, the differences observed to the posterior distribution of the parameter of interest are introduced by these different variances (King et al., 2010). Prior sensitivity is not regarded as a problem in itself, but it may indicate problems such as parameter redundancy (over-parameterisation of the model, so it is not possible to estimate all the parameters in the model) or overly restrictive prior assumptions (King et al., 2010).

1.4.3.b Posterior distribution

The posterior distribution incorporates all the information about the parameter of interest. When the model contains more than 2 or 3 parameters the posterior distribution becomes very complex (King et al., 2010). As a consequence, the information regarding one single parameter is obtained from the marginal posterior distribution of this individual parameter rather than the joint distribution. The marginal posterior is derived by integrating over the rest of the parameters ("integrating out"). For example, if $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_n\}$, the posterior marginal distribution of θ_1 is given by (Gelman et al., 2004):

$$\pi(\theta_1|X) = \int \pi(\theta|X) d\theta_2 \dots d\theta_n.$$

This integration is often complex and difficult if not impossible to derive explicitly. The introduction of the Markov chain Monte Carlo (MCMC) integration methods (Smith and Gelfand, 1992) made it possible to obtain an estimate of this marginal posterior distribution without too much difficulty.

1.4.3.c Bayesian computation: Markov Chain Monte Carlo

MCMC methods are a combination of Markov chain theory (Gilks et al., 1995) and Monte Carlo integration (Morgan, 1984). They are based on the idea of constructing a sequence of values (a Markov chain) whose distribution converges towards the posterior distribution, if the chain is run for long enough and if the conditions of aperiodicity and irreducibility are met (King et al., 2010). The characteristic of the Markov chain is that the distribution of a given value, θ^t, depends only on the previous value, θ^{t-1}. Thus, if there is a sequence, θ^t, with $t = 1,2,3 \dots$, starting at θ^0 then, for each *t*, $\theta^t \sim T_t(\theta^t|\theta^{t-1})$, with T_t being a transition distribution that depends on the iteration *t*. A key element is to define an appropriate transition distribution such that the Markov chain converges to a unique stationary distribution, namely the posterior distribution of the parameter θ (Gelman et al., 2004).

Once it has converged to the stationary distribution, the sequence of values can be used to obtain empirical (Monte Carlo) estimates of the posterior distribution of θ (King et al., 2010).

In this thesis, two types of MCMC algorithms are used to sample from the posterior distribution: the Metropolis-Hasting algorithm (Hastings, 1970; Metropolis et al., 1953) and the Gibbs Sampler (Geman and Geman, 1984).

1.4.3.c.i The Metropolis-Hasting (MH) algorithm

The MH algorithm involves three main steps:

1. selection of initial parameter values θ^t with *t*=0. This could be done either from a starting distribution $p_0(\theta)$ or from a set of starting values dispersed around a crude approximate of the estimate (Gelman et al., 2004);
2. generation, at iteration *t*, of a candidate value θ^* via a specified proposal density distribution $q(\theta^* | \theta^t)$;
3. determination of whether or not the new candidates values are accepted as a t^{th} +1 element of the chain, through the use of an acceptance function $\alpha(\theta^t, \theta^*)$:

$$\alpha(\theta^t, \theta^*) = min(1, \frac{\pi(\theta^*|x)q(\theta^t|\theta^*)}{\pi(\theta^t|x)q(\theta^*|\theta^t)})$$

Then either the candidate value $\theta^*$is accepted with a probability $\alpha(\theta^t, \theta^*)$ and set $\theta^{t+1} = \theta^*$, or it is rejected and $\theta^{t+1} = \theta^t$.

Block updates

With the MH algorithm it is possible to update either one parameter at a time using the single update Metropolis-Hasting algorithm or to do a multi-parameter update called a block parameter update. This last method is often used when there is high correlation between some parameters which can generate slow converging (King et al., 2010), although it can be difficult to specify a suitable multi-dimensional proposal distribution. Due to the nature of the Bayesian models developed in this thesis, some parameters are highly correlated and those parameters are updated in a block.

1.4.3.c.i The Gibbs Sampler

The Gibbs sampler algorithm is a particular case of a MH algorithm where the acceptance probability is always 1. The proposal distribution for a given parameter is the conditional posterior distribution of that parameter (King et al., 2010). Gibbs samplers are easily implemented when conjugate priors are adopted in the model, as the posterior conditional distributions with such priors are of standard form.

For a vector of parameter $\boldsymbol{\theta^t} = (\theta_1, \dots, \theta_d)$ at a state *t* of the Markov chain iteration, each θ_d^t in turn is sampled from the conditional distributions as follows (Gelman et al., 2004).

$$\theta_d^{t+1} \sim \pi(\theta_d | \theta_{-d}^t, x),$$

Where θ_{-d}^t represents all the components of θ, except θ_d, at their current values *t*.

1.4.4. Mixing

The proposal distribution is one of the factors that determine the mixing speed of a chain. If the candidate parameter value drawn for the proposal distribution is too far from the current values (large step) the acceptance rate of candidate value will be low, resulting in a chain that frequently fails to move and thus poor mixing, it will thus take longer to reach the stationary distribution. On the other hand, if the step between the current draw and the candidate is too small, the acceptance rate is going to be high but it will take a long time to move over the parameter space and so for the chain to reach the stationary distribution (King et al., 2010).

Observation of time-series trace plots representing the parameter values for each iteration are a good indicator of the mixing speed (Figure 1-1). A "grassy" plot is sign of good mixing plot (Figure 1-1.a) whereas a plot where a "plateau" can be observed (Figure 1-1.b) is a sign of slow mixing.

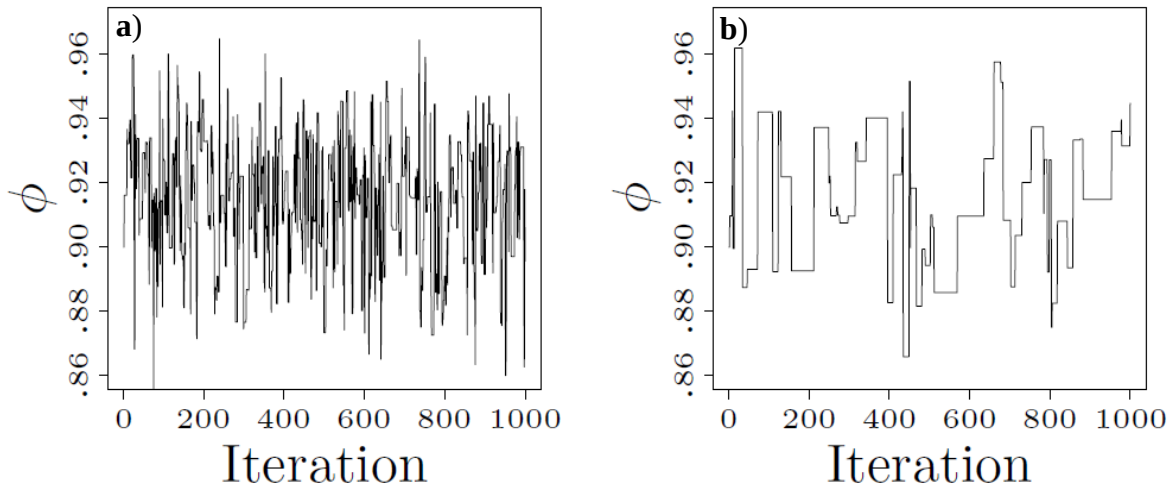


Figure 1-1: Trace plots representing a good (a) and a slow (b) mixing of the MCMC chains. These plots were extracted from (King et al., 2010, p131).

1.4.5. Burn-in and convergence

To be sure that the sample used to obtain inference for the parameter of interest rises from the posterior distribution, the chain needs to have reached convergence to the stationary distribution. In practice this means that observations from the start of the chain are discarded, to use only observations once the chain as converged (King et al., 2010). This initial part of the chain discarded is called burn-in period.

Numerous diagnostic methods have been developed to test if convergence is reached (Cowles and Carlin, 1996). The Brooks-Gelman-Rubin (BGR) (Brooks and Gelman, 1998), convergence diagnostic is one of the most popular and will be used in this thesis. Their diagnostic is based on performing an analysis of variance between different chains starting from different over-dispersed starting points. They looked at the ratio of the within-chain over the between-chain variance and defined a reduction factor $\hat{R}_c$. If this factor is close to 1 it can be said that the chain has converged (Brooks and Gelman, 1998). However a reduction factor greater than 1.2, means the chains have failed to converge (Gelman et al., 2004). This diagnostic test only gives an indication if the chain has converged toward a common distribution; they do not indicate if they have converged toward the correct stationary distribution (and indeed no test can do this).

1.4.6. Parameter inferences

Once the chains have converged then it is possible to obtain empirical estimate of any posterior summaries of interest. When data are generated from simulation point summary statistic such as the mean, median, mode can be used, to measure the error between the posterior point estimate and the expected parameter true value. But a point estimate by itself is not very meaningful: information on the uncertainty of the point estimate is very important. In this thesis the coefficient of variation (CV) is the statistic used to measure the uncertainty of the point estimate. Coefficient of variation measures the standard deviation of the posterior samples relatively to the mean of the samples.

Part I. Classification

Chapter 2: Measuring the variability of an automatic whistle classifier

2.1. Introduction

Whistles produced by odontocete cetaceans are highly variable between and within species. Comparative analyses of whistle characteristics have shown that factors such as taxonomy, morphology and natural selection pressure can explain some of the variation between species (Rendell et al., 1999; Steiner, 1981), whereas variation within species is correlated with population structure, environmental heterogeneity and/or behaviour (Rendell et al., 1999). The variability in whistles can be useful to identify odontocete species as some whistle features are characteristic to each species.

To identify species by their whistles, whistle classifiers have been developed (chapter 1 2.2.b.iv, *Classifier* p11). These classifiers are created using data for which species' identities are known (training data) and the performance of the classifier is presented by a $m \times m$ confusion matrix with each element p_{ij} giving the probability of classifying species j as species i (see chapter 1 p13). Once created, these classifiers and their trained species categories are subsequently used to identify whistles in new acoustic data.

2.1.1. Misclassification

For species living in the same type of environment and/or being closely related to each other, the confusion matrix is expected to have misclassification probabilities higher than 0 because the similarity in vocalizations between species makes it difficult to tell them correctly apart (Steiner, 1981). In reality none of the whistle classifiers developed to date (Datta and Sturtivant, 2002; Gillespie et al., 2013; Oswald et al., 2007) are able to identify any odontocete species perfectly. A consequence of misclassification is that the observed number of detections as identified by the classifier for each species, $\boldsymbol{n} = (n_1, \dots, n_j, \dots, n_m)$, contains (after classification) correctly identified detections as well as misidentified detections. Chapters 6 and 7 demonstrate that it is possible to estimate the true number of detections of each species from the observations, if it is assumed that there is no uncertainty on the classification probabilities $p_{ij.}$ However if the classification probabilities have uncertainty attached to them, chapter 7 of this thesis demonstrate that estimating the true number of detections for each species becomes much more challenging and estimates can be very imprecise, even if the variance of the classification probabilities is small.

The most common method to estimate abundance from acoustic data is based on cue counting theory according to which the number of counted cues has to be multiplied by some factors such as cue probabilities, to estimate the number of individuals (chapter 1: 2.2.b.ii, p*8*). Clearly, if the number of counted cues is biased then the final abundance estimate will be biased as well. Furthermore, if the number of counted cues contains uncertainty then the final abundance estimate will also contain this uncertainty. It is important for management decisions to know the precision of the abundance estimates, so understanding and measuring the uncertainty of the observed number of cues (here detected whistles), and consequently uncertainty of the true number of cues, is essential. The uncertainty of the number of observed cues comes in part from uncertainty in the classification probabilities of the classifiers.

2.1.2. Uncertainty in the estimates of classification probability

Given the method used to develop a classifier (chapter1: 2.2.b.iv p11), the classification probabilities of the confusion matrix are only estimates of the true classification probabilities. Indeed conceptually, the classification probabilities p_{ij} are estimated from two sampling processes both of which generate uncertainty in the estimation of $\hat{p}_{ij}$:

Uncertainty from the training process: The vocalisations used as training data to create the classifier are a sample of the entire set of vocalisations that could be used to train the classifier – i.e., the vocalisations across all the populations for which the classifier can be used to produce an acoustic abundance estimate. Consequently, there is uncertainty as to the performance of the classification algorithm that arises from this sampling process.

Uncertainty from the testing process: An additional source of uncertainty arises when attempting to measure the performance of the classifier, regardless of how it was trained. To exactly evaluate performance, the classifier would have to be tested on the entire set of possible vocalisations. However this is clearly not possible in practice, and a small set of testing data is used, which can be regarded as a sample from the entire set. Hence additional uncertainty about the classifier performance arises from this sampling process.

2.1.3. PAMGUARD Whistle Classifier

The PAMGAURD whistle classifier (PWC) developed by Gillespie et al. (2013) is the only whistle classifier to my knowledge that measures the uncertainty in the classification probabilities. Their classification process is organised in six main stages (*i* to *vi*) and tries to classify groups of whistle contours organised in section rather than individual whistle contour. A whistle contour being a representation in time and frequency of the whistle detected (Fig3.1). The details of the process is described in Gillespie et al., (2013), Figure 2-1 and the following lines give only a summary of the main stages of this process. (*i*) For each species, detected whistle contours are divided into small units (called fragments). For each of this fragment 3 parameters are extracted: the mean frequency; the slope of the frequency change over time and the curvature of the fragment. (*ii*) For each species a separate random start is taken within the fragments; 2/3 of the fragments read consecutively from that point are used to train the classifier whereas the remainder is used for testing. (*iii*) Within the training and testing dataset, fragments are grouped into consecutive sections, containing a number of fragments ordered by date and time. While the distribution of the three primary parameters extracted for each fragment overlaps largely between species, they also have a markedly different shapes (Gillespie et al 2013). Therefor by accumulating these fragments in section it is possible to build a distribution of those primary parameters from which a secondary set of parameters, being the mean, the standard deviation and the skew, of each distribution of the primary parameters is calculated, giving a final of 9 parameters extracted for each section. (*iv*) A Linear Discriminate function Analysis (LDA) using those 9 parameters is applied to the training data (made of sections from each species); the output of this method is a linear combination of the section's parameters. (*v*) Based on this linear combination, for each section in the test data, a relative probability is assigned to each classification group (each species of the training data) of the classifier such that the sum of the probabilities across the classification group is one. The classification of the section corresponds to the classification group with the higher probability. (*vi*) The outcome of this classification is compared with the test data representing the truth and as a result of this comparison a confusion matrix (*C*) is derived.

Gillespie et al. (2013) repeats stages (*ii)* to (*vi) B* times. For each repetition a new random selection of training and testing dataset is generated. After the *B* bootstraps are done, the final LDA algorithm is calculated using the entire training dataset (Figure 2-2). It is thus not possible to derive the confusion matrix from this last run. This last run is done to create a classifier algorithm with the maximum data possible.

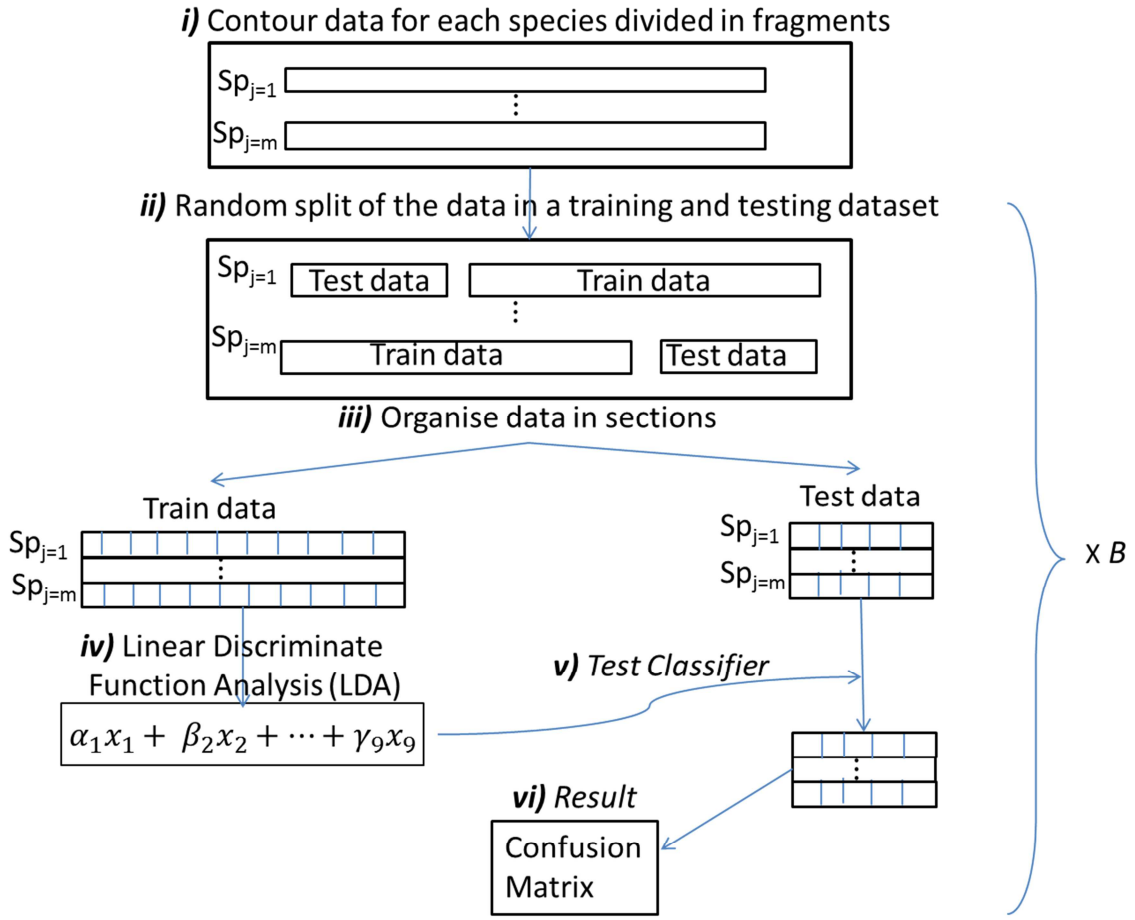


Figure 2-1: Schematic diagram of the PAMGUARD Whistle classifier training process during the B bootstraps.

The final confusion matrices shown in Gillespie et al., (2013) are each an average over the *B* confusion matrices created and the variability is estimated by measuring the standard deviation over the *B* bootstraps of the $\hat{p}_{ij}$. This estimate of the variability contains several sources of uncertainty in one measurement: uncertainty from the training process, uncertainty from the testing process and uncertainty from the bootstrap method used, which is close to a moving-block bootstrap method. Ideally when developing a classifier one should try to minimize uncertainty. To do so, the first stage is to identify and quantify as many sources of uncertainty as possible.

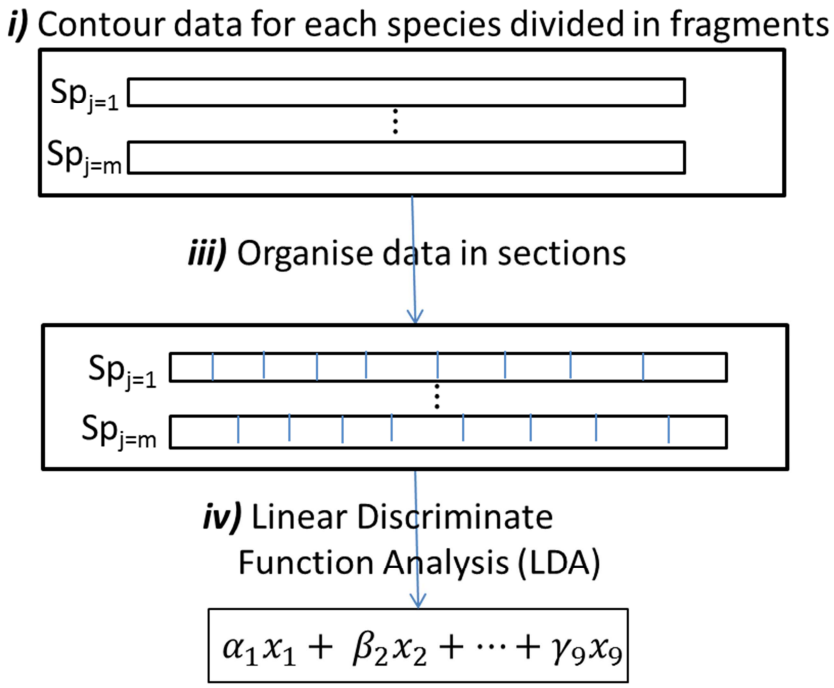


Figure 2-2: Last run of the PAMGUARD Whistle classifier training process.

As explained in the description of the PWC, to create a classifier one part of the training data is used to train the classification algorithm and the second part is used to test this algorithm. Given that the quantity of uncertainty (from the training and testing process) is linked to the sample size of the data, a trade-off between the proportion of the training data used to train and test the algorithm needs to be found. A large proportion of the data used to train the classifier will decrease the training uncertainty but increase the testing uncertainty and vice versa when a small proportion of the data is used to train the classifier. The optimum classifier should be obtained when all the training data are only used to train the classifier algorithm (it is what is done with the PWC during the last run of the PWC process). In this configuration the testing process uncertainty is removed and the training uncertainty is minimized.

In this chapter, the uncertainty of the training process is described using Nonlinear Least Square models which are used to predict the final uncertainty when all the data have been used for training. To do so a modification of the PWC is proposed that enables us to measure independently the two sources of uncertainty (testing and training) generated by the sampling process.

2.2. Methods

2.2.1. Data

Whistle contours used in this chapter were extracted by the PAMGUARD Whistle and Moans detector (Gillespie et al. 2013) from data of the MORL_BOWL project presented in detail in chapter 3. The MORL_BOWL dataset consisted of whistle detections from 5 species (Bottlenose dolphin, *Tursiops truncatus*, Common dolphins, *Delphinus delphis*, Risso's dolphin, *Grampus griseus*, White beaked dolphins, *Lagenorhynchus albirostris*, and Stripped dolphins, *Stenella coeruleoalba*) recorded along the Scottish coasts. Each acoustic recording was associated with a visual detection confirming the species identification.

2.2.2. PAMGUARD Whistle Classifier modifications

The PWC was modified such that it was possible to measure the training variability independently of the testing variability. In the PWC described by Gillespie et al., (2013), the data are divided in one training and one testing dataset. In this improved method, the PWC was modified to divide the data in one training and two test datasets (Figure 2-3). Despite this difference, the classification process was exactly the same, divided into 6 main stages as outlined above.

With a=two test datasets per bootstrap replicate it was possible to measure the variance between each bootstrap replicate (between variance) and the variance within each bootstrap replicate (within variance). The between component of variance should capture the training uncertainty generated by the different training data used at each bootstrap replicate whereas the within component of variance should capture the testing uncertainty generated by the different test data used in each bootstrap replicate.

Following the idea that the variance decreases when the sample size increases, three differently sized subsets of data were used to train the classifier: half, a quarter and an eighth of the sections were used to train the classifiers. For each classifier, B=100 bootstraps were run with random start point but with the same proportion of training data. The output of each bootstrap was two confusion matrices C_{ba} with classification probabilities $\hat{p}_{ijba}$. For each classifier the between variance of each $\hat{p}_{ijb}$ (V_{ij}) was estimated by using the formula for the between variance of a standard analysis of variance (ANOVA):

$$\mathrm{var}(\hat{p}_{\mathrm{ij}}) = \mathrm{V}_{\mathrm{ij}} = \frac{\sum_{b=1}^{B}(\hat{p}_{\mathrm{ijb}} - \bar{p}_{ij})^2}{\mathrm{B}-1}$$

Finally models were fitted to these between variances with the objective of being able to predict what the between variance be if all the data were used to train the classifier.

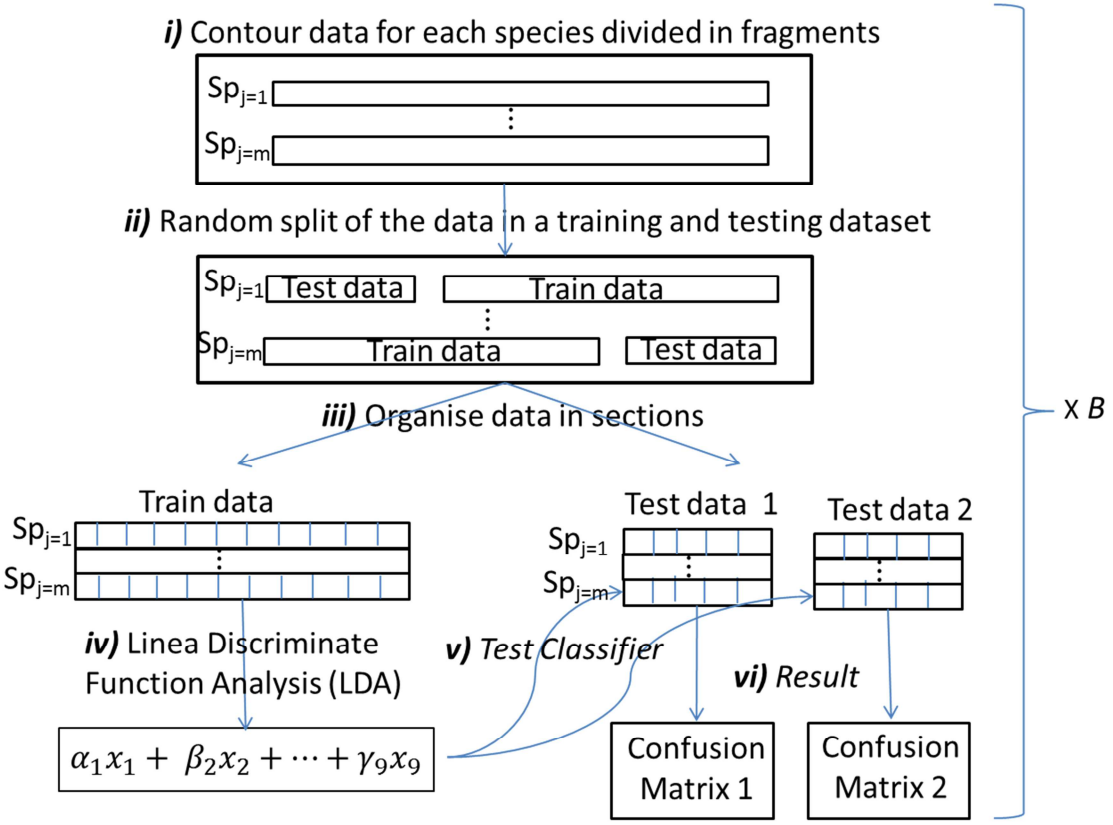


Figure 2-3: Training process of the modified PAMGUARD Whistle classifier. Note the testing dataset has been divided in two so it is possible to measure a between and within variance.

2.2.3. Models

2.2.3.a Underlying framework

It was assumed that, within the columns of C, the $\hat{p}_{ij}$ followed a Dirichlet distribution with parameters ($\alpha_{1j},\alpha_{ij},\ldots,\alpha_{mj}$). A confusion matrix of dimension $m \times m$ will have m Dirichlet distribution. A Dirichlet distribution (Royle and Dorazio, 2008) is a continuous multivariate distribution with concentration parameters ($\alpha_1,\alpha_i,\ldots,\alpha_m$) where for $\mathbf{x}$~Dir($\alpha_1,\alpha_i,\ldots,\alpha_m$), $x_i \in [0,1]$, $\sum_{i=1}^{m} x_i = 1$ and $E[x_i] = \frac{\alpha_i}{\sum_{i=1}^{m}\alpha_i}$.

Since in *C*, $\sum_{i=1}^{m} p_{ij} = 1$, this distribution seems a reasonable assumption. Consequently, we have

$$E[p_{ij}] = \frac{\sum_{b=1}^{B} \hat{p}_{ijb}}{B} = \frac{\alpha_{ij}}{\sum_{i=1}^{m} \alpha_{ij}}.$$

Then in theory, the variance V_{ij} for each $E[p_{ij}]$should be described by the variance of a Dirichlet distribution:

$$var(\hat{p}_{ij}) = V_{ij} = \frac{\alpha_{ij}(\alpha_0 - \alpha_{ij})}{{\alpha_0}^2(\alpha_0 + 1)} \qquad (2\text{-}1)$$

where $\alpha_0 = \sum_{i=1}^{m} \alpha_{ij}$

For real data, however, the variability was suspected to be different than for a true Dirichlet distribution due to the presence of the other sources of variability. Hence it was assumed that V_{ij} was proportional to the expression on the right hand side of Eq.2-1 To include these variance factors, two unknown parameters β_1 and β_2 were multiplied and added to the baseline Dirichlet variance formula:

$$V_{ij} = \beta_1 \frac{\alpha_{ij}(\alpha_0 - \alpha_{ij})}{{\alpha_0}^2(\alpha_0 + 1)} + \beta_2 \qquad (2\text{-}2)$$

Equation 2-2 for the variance suggested that the relationship between the concentration parameters (and indirectly the $\hat{p}_{ij}$) and the V_{ij} was not linear but quadratic. For this reason non-linear least square (NLS) models (Bates and Watts, 1988) of the form of Eq.2-2 (see below) were fitted to the V_{ij}. The models were fitted using the 'nls' library (Bates and Watts, 1988) implemented in the statistical software R (*R Development Core Team*, 2012).

2.2.3.b Models tested

From this underlying model, three different forms of the concentration parameters of the Dirichlet distribution were tested as a function of the cell-wise $\hat{p}_{ij}$'s and the sample size (number of section for each species used for training) to find which form fit best the data and thus will predict best the variance when all the data are used to train the classifier. The first form of the concentration parameters (Model1) was selected such that there was no dependency to the sampling size but only on the actual classification probabilities. For the

second (Model2) and third (Model 3) form, the concentration parameters were selected such that they were proportional to the sample size as well as to the classification probability.

Model1: Sample size independent Eq.2-3

In the simplest case, the concentration parameter was only dependent on the p_{ij}: $\alpha_{.j} = p_{.j}$.

$$\hat{V}_{ij} = \beta_1 \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{2} + \beta_2 \qquad (2\text{-}3)$$

Model2: Species sample size dependent Eq. 2-4

With acoustic data, there are different numbers of detections for each species in the training dataset, resulting in different numbers of sections (S_j) for each species, with for some species a large sample size (large number of training section) and for other a very small sample size.

The parameters α_{ij}in this model were choosen such as they were proportional to both $\hat{p}_{ij}$ and the sample size per species S_j : $\alpha_{.j} = S_j p_{.j}$.

The $\hat{V}_{ij}$ were then inversely proportional to S_j for each species.

$$\hat{V}_{ij} = \beta_1 * \frac{\hat{p}_{ij} * (1-\hat{p}_{ij})}{S_j + 1} + \beta_2 \qquad (2\text{-}4)$$

Consequently only 5 classification probabilities were associated with each sample size S_j. Due to this small sample size, the result of the model fitting process has limited validity and needed to be treated with caution.

Model3: All species sample size dependent Eq.2.5

In model 2 only 5 classification probabilities (as the classifier discriminate 5 species) were available for each sample size. Being aware that this small number of data can generate unreliable results, it was decided to explore what would be the consequences of having concentration parameters dependent on the total number of sections for the 5 species $\alpha_{.j} = Sp_{.j}$, with S ($S = \sum_j S_j$), being the total number of sections for the 5 species of the classifier.

$$\hat{V}_{ij} = \beta_1 \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S + 1} + \beta_2 \qquad (2\text{-}5)$$

Models comparison

Between these three modelling options, the model with the smallest AIC (Akaike, 1974) was selected. Predictions for when all the data were used, were derived from this model. To predict $\hat{V}_{ij}$, the average p_{ij} over the 3 classifiers (from using half, a quarter and an eighth of the original data for training) was used.

2.3. Results

2.3.1. Data description

The total number of sections used in the training dataset was unequal between species (Table 2-1). The majority of whistle contours in the data came from bottlenose and common dolphins. The number of sections for both Risso's and white beaked dolphin was very small: e.g., only four and 3 sections respectively when only an eighth of the sections were used to train the classifiers.

Table 2-1: Number of sections S_j for each species used to train the classifier. The number of sections is dependent on the proportion of the data used to train the classifier. The first classifier used half of all the sections, the second a quarter and the third an eighth, whereas for the prediction, 100% of the sections are used to train the classifier.

	S_j			
Proportion of training sections	**50%**	**25%**	**12.5%**	**100%**
Bottlenose dolphin	422	211	105	844
Common dolphin	595	297	148	1190
Risso's dolphin	17	8	4	34
White beaked dolphin	15	7	3	30
White sided dolphin	55	27	13	110
TOTAL	**1104**	**550**	**273**	**2208**

The variance of a Dirichlet distribution follows a bell shape curve moving from 0 to 1 with a maximum when $E[p_{ij}] = 0.5$ (Figure 2-4). The observed variances when half, a quarter and

an eight of the data section are used to train the classifier followed the same bell curve shape but with smaller values than the theoretical Dirichlet variances (Figure 2-4).

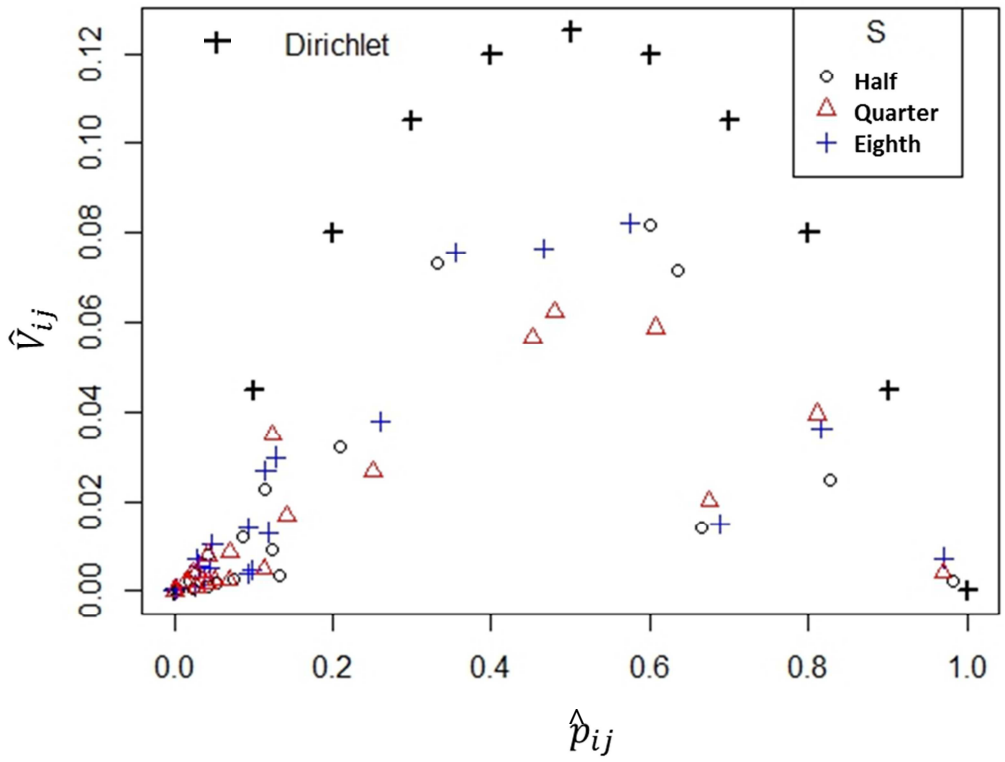


Figure 2-4: Variances of the classification probabilities ($\hat{V}_{ij}$) for a given classification probabilities ($\hat{p}_{ij}$) and a training sampling size (S). S is the proportion of the sections used to train the classifier: half of the sections used to train the classifier (black open circles), a quarter of the sections (red triangle) and an eighth of the sections (blue cross). Symbolised with a black cross are the variances as function of probabilities obtained from a Dirichlet distribution directly.

2.3.2. Model selection

Model 3 (variance dependent on the total number of section for all species, *S*) was the model with the smallest AIC and residual sum of squares (r^2) (Table 2-2). In this model the unknown parameter β_2 was not significantly different to zero (p>0.05) whereas β_1 was positively correlated to the V_{ij}'s.

$$V_{ij} = 70.19\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S+1} + 0.01$$

Table 2-2: Δ AIC, AIC and residual sum of squares for the three models

Model	Δ AIC	AIC	r^2
Model 1	18.38	-475.50	7.2×10^{-3}
Model 2	51.63	-442.25	11.1×10^{-3}
Model 3	0	-493.88	5.6×10^{-3}

Model 2 (for which the concentration parameters were associated with the number of sections for each species within the training dataset, S_j) was the model exhibiting the worst fit.

With Model3, the predictions of the variances if the classifier had been trained with all the sections available ranged from 0 (when $\hat{p}_{ij}$ =0) to 7.10^{-3}.

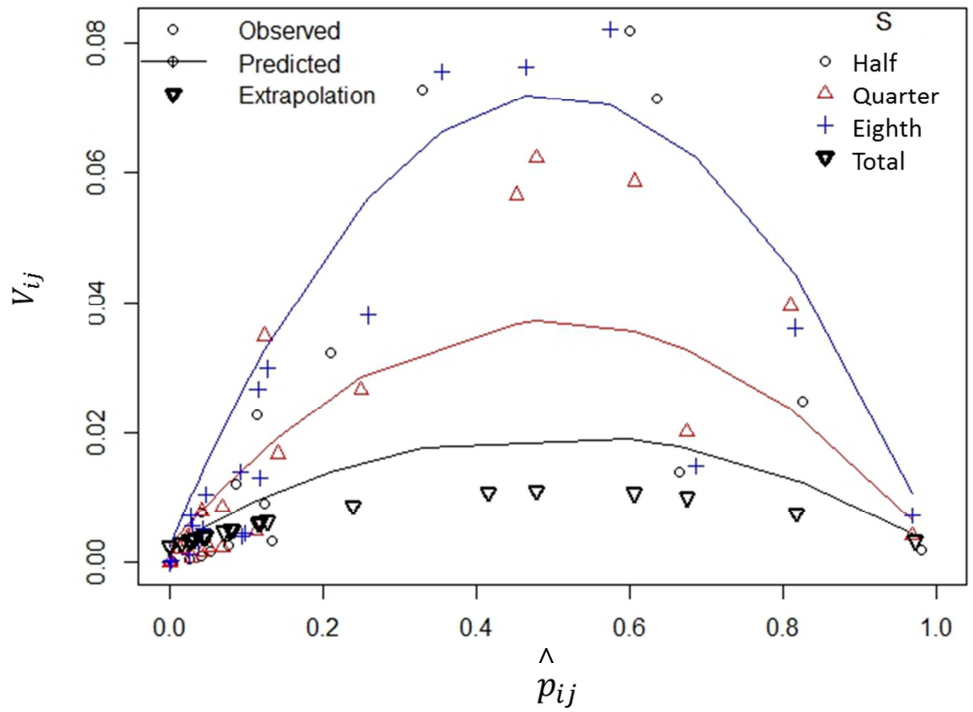


Figure 2-5 : Observed data (open symbols) versus predicted (lines) and extrapolation (bold black triangles) with full dataset. Each colour represents a sampling size as described in previous figure.

2.3.3. Comparison of the variance with the version of the PWC described in Gillespie et al. (2013)

Standard deviations were measured from these predicted variances and they were compared with the standard deviation measured with the original PWC (Table 2-3). The standard

deviation measured with the modified version of the whistle classifier was smaller than with the original version; the average standard deviation for all the confusion matrices was 3.9% (±3%), whereas the average standard deviation measured with the PWC of Gillespie et al., (2013), was 8.2% (±9%). Only for three classification probabilities the predicted variance is slightly larger (for white sided dolphin misclassified as bottlenose dolphins, bottlenose dolphins misclassified as Risso's dolphin and white beaked dolphins)

Table 2-3 Estimated standard deviation by the least squares model 3 if 100% of the data were used to train the classifier. Values in brackets show the measured standard deviation by the PWC of Gillespie et al., (2013) when 2/3 of the data are used to train the classifier. BND=Bottlenose dolphin COD=common dolphin, RSD =Risso's dolphin, WBD=white beaked dolphin and WSD= white sided dolphin

	True Species				
Standard deviation in %	**BND**	**COD**	**RSD**	**WBD**	**WSD**
BND	8.6 (26.7)	5.8(9.6)	2.3 (4.5)	2.8 (6.2)	1.8 (1.4)
COD	7.5 (18.0)	8.2(11.9)	0.0	8.7 (27.0)	5.5 (15.1)
RSD	2.5 (2.2)	0	2.4 (4.5)	0.0	0.0
WBD	3.3 (3.1)	5.5 (5.8)	0.0	8.8 (28.6)	3.4 (4.1)
WSD	4.7 (11.1)	4.6 (5.0)	0.0	4.3 (8.8)	6.7 (15.8)

In parallel to the least square method used, a Generalised Additive Model (GAM) was fitted to the data. These models gave a better fit of the data however the extrapolation to estimate what would have been the variance if 100% of the data were used to train the classifier appeared not to be realistic. For this reason only the result of non-linear least square models is presented here.

2.4. Discussion

With the Model 3 depending on the probabilities of classification and the total number of training sections of the classifier, the prediction of the data was the best obtained and seemed

reasonable. This model, selected because of its smaller AIC, tended to homogenise the variance between species. This homogenisation was a consequence of the denominator *S* (total number of training section) of the model. A more easily defended model is one where species with less data in the training data generated more variability. Model 2 should have captured this factor, because of the denominator of the model being directly dependent on the number of training section per species. The worse AIC value for Model2 than Model3 is perhaps a consequence of the fact that for this model the predictions were based on a small number of data. For each sample size only five data (one per species) were available. In theory, the model with the best diagnostics for fit is considered the 'best' statistically (Model 3 in this case) but biologically, another model (in this case Model 2) may be preferred. In this specific case the homogenisation of the variance generated by Model 3 will make the final precision of the estimate of the true number of detections less sensitive to the amount of detections for each species. Consequently the precision of the true number of detections for rare species will probably be lower and vice versa higher for the common species than if Model 2 was used.

In conclusion, this chapter proposed a new approach to try to measure the training variability of a whistle classifier. Other solutions may exist requiring a statistical approach more robust to small datasets and dealing with the complexity of the bootstrap method used by the PWC classifier. The following chapters show the importance of the quantity and quality of the training dataset to develop a reliable (low uncertainty) and accurate (high correct classification probability) classifier. Then the second part of this thesis will demonstrate how and why estimates of uncertainty in the performance of a whistle classifier should always be associated with the estimated confusion matrix if the acoustic data are to be used to estimate abundance of species.

Chapter 3: Classification of data from a reliable training dataset

3.1. Introduction

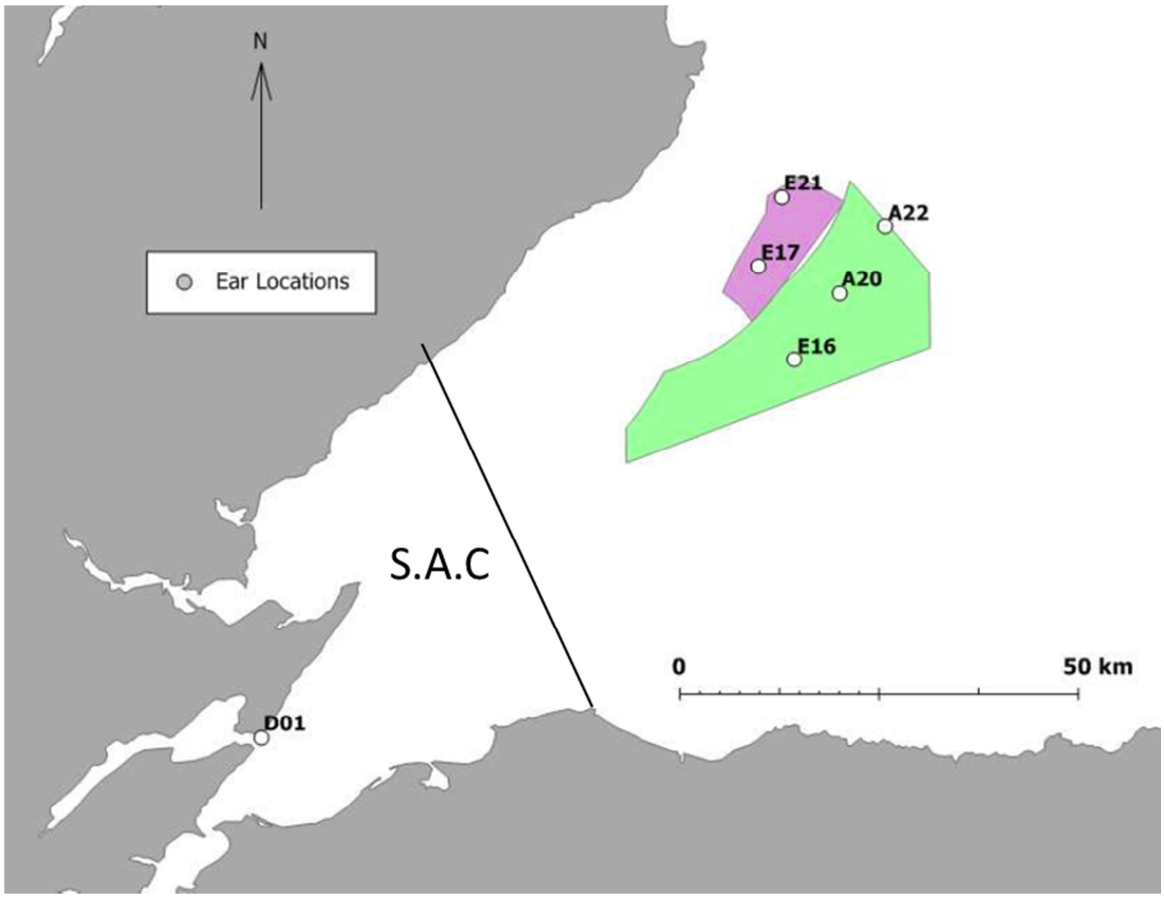
In certain circumstances, for example when vocalisation characteristics are easy to identify, it is possible to estimate the abundance of cetacean species using only passive acoustic devices. For example, Marques et al. (2011) obtained density estimates of the endangered North Pacific right whale (*Eubalaena japonica*) in the Bering Sea from fixed passive acoustic devices only. Martin et al. (2012) were able to estimate abundance of minke whales (*Balaenoptera acutorostrata*) in Hawaiian waters from 14 bottom-mounted hydrophones; and at present the SAMBAH[1] project aims to improve the management strategy for the conservation of the rare population of Harbour porpoises (*Phocoena phocoena*) in the Baltic Sea using acoustic data collected from a large array of C-POD hydrophones. Using solely passive acoustic data from fixed devices to verify presence and estimate abundance of species is cost-effective in the long term: once the hydrophones are installed the recordings can be collected remotely or can be retrieved by a small boat from the devices. Fixed hydrophones allow for large temporal coverage (as hydrophones can stay for months or years in the same place), but spatial coverage depends on the quantity and spatial extent of the installed devices. For this reason, environmental and governmental agencies are interested in passive acoustic methods to monitor and better understand the presence of cetacean species at a local scale.

This chapter presents the results of a study in which it was necessary to distinguish bottlenose dolphin (*Tursiops truncatus*), a protected species under Annex II of the EU Habitats Directive, from other species present at two major off-shore wind farm sites in the Moray Firth, called MORL[2] and BOWL[3] (Map 3-1). While mike whale, right whale, harbour porpoise have very distinctive vocalisations, bottlenose dolphins vocalisations are similar to those of the other species (common dolphin (*Delphinus delphis)*, white beaked dolphin (*Lagenorhynchus albirostris*), white sided dolphin (*Lagenorhynchus acutus*) and Risso's dolphin (*Grampus griseus*)) likely to be found in the same area (chapter 1 table1.1, p16). Hence, to be able to differentiate whistles from bottlenose dolphins accurately from those of

[1] Static Acoustic Monitoring of the Baltic Sea Harbour Porpoise. http://www.sambah.org/
[2] Moray Offshore Renewables Ltd
[3] Beatrice Offshore Windfarm Ltd

other species it is necessary to develop a reliable whistle classifier. A prerequisite to create such a classifier is the collection of data from already identified species (training data).

This chapter describes the development of two classifiers from the same, high-quality training dataset. The first classifier differentiates bottlenose dolphins from the four other species, where the latter are pooled into one group (two classification groups). The second classifier differentiates all five species (five classification groups). Then these classifiers were used to identify species within recordings made on the wind farm sites, for which no visual data were collected.



Map 3-1: Map of the North East coast of Scotland with the wind farm sites (in color) and the position of the EARs deployment (D01,E21,E17,E16,A20.A22) and the delimitation of the Special Area of Conservation (S.A.C) for bottlenose dolphins.

3.2. Methods

The classifiers in this chapter were created with the PAMGUARD Whistle Classifier (PWC) modules Gillespie et al. (2013) with the modification explained in the previous chapter (chapter 2: 2.2 p29). The classification of new data was done using the PWC module in a configuration to use the classifier to identify this data and not to create a classifier (Table 3-1).

3.2.1. Creation of the classifiers

When a classifier is created, the ultimate objective is to have a classifier algorithm as efficient as possible to discriminate the different species of interest and to create a confusion matrix illustrating the accuracy and precision of the classifier.

The creation of the classifiers with the PWC were made in several steps (Table 3-1 A.) described in the next sections (3.2.1a to 3.2.1.c)..

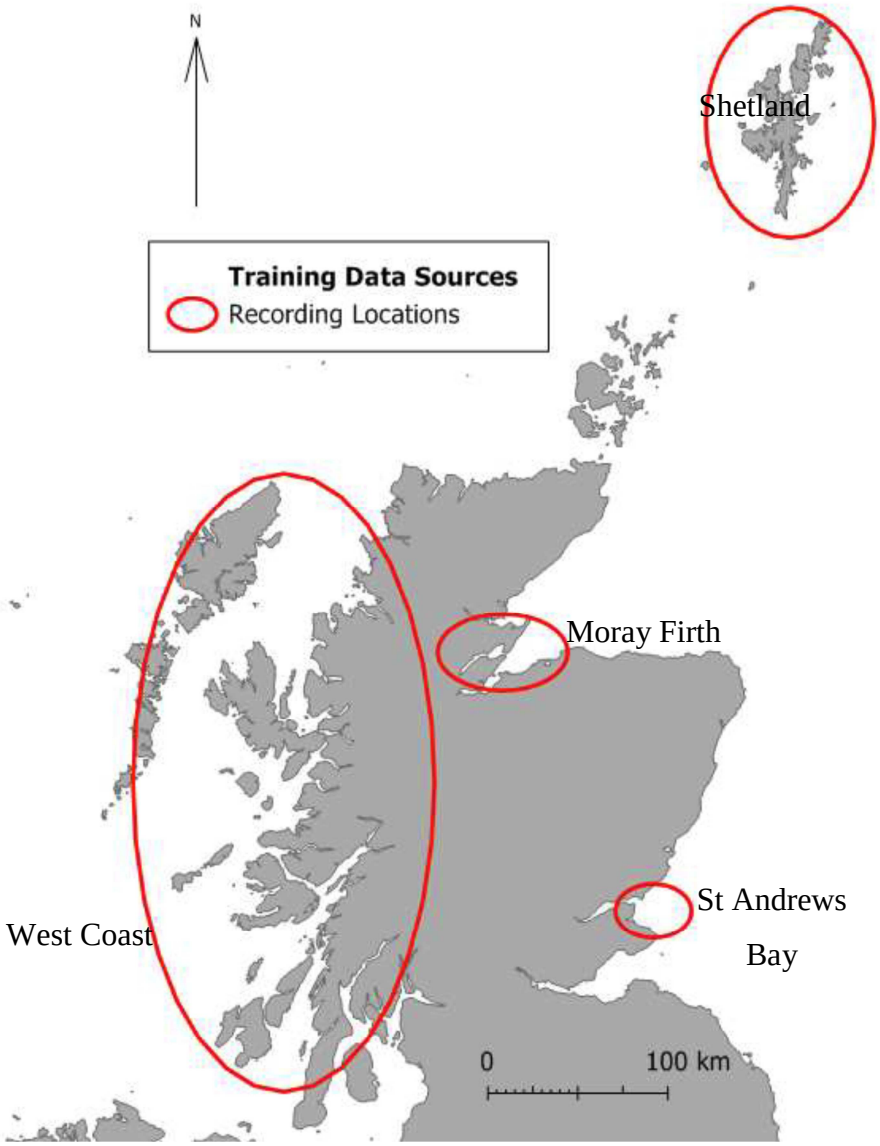
Table 3-1: Main stages to create a whistle classifier and to apply it on new data using the PWC.

A:Creation of a classifier with PWC	B: Classification of unidentified data with PWC
Data: time frequency contours from identified species	Data: time frequency contours from unidentified species organised in fragments and sections of optimal length measured in (A.1)
1. Selection of optimal fragment and section lengths (comparing quality coefficient, *Q*)	1. Classify sections
2. Creation of the confusion matrix: 2. Classification probabilities, p_{ij} 3. Variance for each p_{ij}	4. Organise sections in encounters and classify encounters (optional) 5. When it is possible, compare classification results with prior information

3.2.1.a Identified dataset

The identified data were used to create a classifier (Table 3-1 A). It was comprised of bottlenose dolphins, common dolphins, Risso's dolphins, white beaked and white sided

dolphins recordings collected by different research groups (Table 3-2) on different small surveys platforms (sailing boat, small motor boats) along the coast of Scotland (Map 3.2).



Map 3-2: Locations of the training dataset.

For all different recordings it was possible to identify the recorded species with high confidence due to the proximity of the animal to the visual observers.

The following data sources were used: Recordings of all the species, except for bottlenose dolphins, were collected from the quiet sailing boat of the HWDT[4] during small scale survey along the West coast of Scotland (Embling et al., 2010). Few additional recordings of Risso's dolphins came from the North of Scotland. All recordings of bottlenose dolphins were

[4] Hebridean Whale and Dolphin Trust

collected by scientists of the Sea Mammal Research Unit of St Andrews from a small motor boat in the North and in the East of Scotland for projects aiming to collect vocalisations to study social interaction or particularities in vocalisation patterns; e.g Janik, 2000; Quick et al., 2008) (Table 3-2). The sampling rates of the recordings varied from 48 kHz to 500 kHz.

Table 3-2: Training dataset and the general location and sources which collected them.

Species	Location	Sources
Bottlenose dolphin	Moray firth	St Andrews University
	St Andrews Bay	St Andrews University
	Shetland	St Andrews University
Common dolphin	West Coast	HWDT
White-beaked dolphin	West Coast	HWDT
White-sided dolphin	West Coast	HWDT
Risso's dolphin	West Coast	HWDT
	Shetland	St Andrews University

The first classifier (called *2Sp* classifier) classified acoustic detections as "BND" (for Bottlenose dolphins) or OTHER (for the four other species) (Table 3-3). The second classifier (called *5Sp* classifier) distinguished between all five species in classification groups called "BND", "COD" (common dolphin), "RSD" (Risso's dolphin), "WBD" (white Beaked dolphins) and "WSD" (white side dolphin).

Table 3-3: Groups of species classified for both classifiers. *2Sp* classifier discriminated Bottlenose dolphins from all other species pooled, whereas *5Sp* classifier discriminated between all five species.

Species	*2Sp*	*5Sp*
Bottlenose dolphin	BND	BND
Common dolphin	OTHER	COD
White-beaked dolphin	OTHER	WBD
White-sided dolphin	OTHER	WSD
Risso's dolphin	OTHER	RSD

To be comparable and usable by the PWC, all the recordings were decimated to 48 kHz. Any sounds over a defined threshold (8dB) were automatically detected using the PAMGUARD Whistle and Moan detection module (Gillespie et al., 2013). The output of the detector created a file for each recording, with the time-frequency contours of each sound detected (Figure 3-1). These contour files were then used in the PWC to train the classifier.

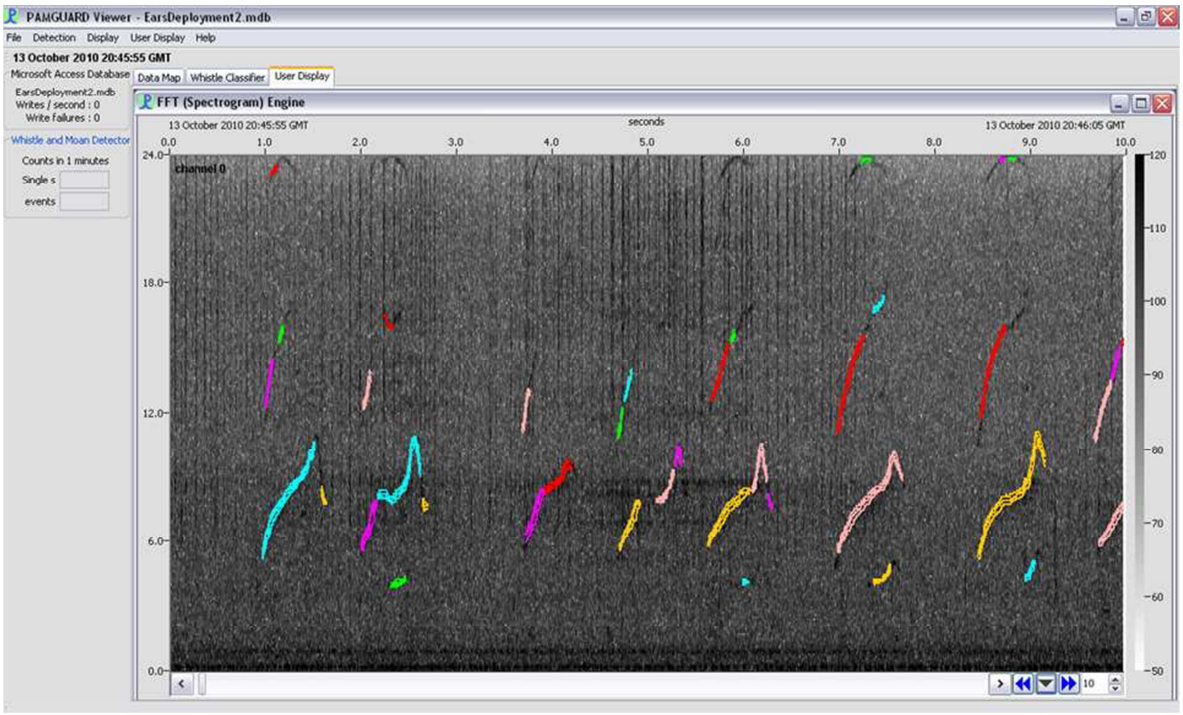


Figure 3-1 Example screen grab showing whistle contours extracted from recordings of bottlenose dolphins using the PAMGUARD Whistle and Moan detector module. Frequency (kHz) is on the y-axis and time (10 seconds) is on the x-axis). The different colours show the contours identified by the WMD (clicks are also visible above 6 kHz). (SMRU ltd et al., 2011)

3.2.1.b Selection of the optimal parameters

The PAMGUARD Whistle classifier works by comparing properties of a group of whistle contours and does not look at each contour individually. Indeed, the output of the detector is rarely a full whistle contour but a part of a whistle contour. Often, contours break into segments because of other transient noises masking the whistle for a very short period of time or because whistles are intersecting each other and it is difficult for the detector to recognise the full contour. To homogenise these contours, Gillespie et al., (2013) divided each contour into smaller uniformly sized units called fragments. Many consecutive (in time) fragments are then regrouped in sections, from which nine parameters are extracted to run the classifier (Gillespie et al., 2013; chapter 2). These parameters described the properties of each section. The length of these fragments and sections were expected to influence the quality of the classifier. Indeed, short fragments and sections are more likely to generate unstable measurement of parameters. Whereas long fragments and sections require many more whistles to obtain a classification result (Gillespie et al., 2013).

When a classifier is created, the effect of fragment and section lengths on the classification probabilities needs to be measured to select the optimum lengths. To do so, the whistle classifier process described in the previous chapter (Figure 2-3, *p30*) was applied to the identified data set, using 80% of the sections to generate the training data (of the classification process). One hundred bootstraps were run for each possible combination of fragment lengths ranging from 26ms to 187ms (equivalent to 5 to 35 bins) and section lengths ranging from 10 to 60 fragments. To select the optimum fragment and section length, a variable was introduced called quality coefficient (*Q*). For each species *j* and each combination of fragments and sections length, Q_j (Eq. 3-1) measured the quality of the classifier by subtracting the average correct classification probability (*T*) over the 100 bootstraps to the average false positives rates (*F*).

$$Q_j = \frac{\sum_{b=1}^{100} \mathrm{T}_{jb}}{100} - \frac{\sum_{b=1}^{100} F_{jb}}{100} \qquad \textbf{(3-1)}$$

A good classifier is characterised by a high correct classification probability and a low false positive classification probability so the higher Q_j, the better was the classifier.

3.2.1.c Creation of the confusion matrix

These optimal parameters were used to generate the final confusion matrix of both the *2Sp* and *5Sp* classifiers. The classification probabilities of these final confusion matrices were an average of 100 bootstraps run with the training section being 80% of the identified data and with the optimal fragment and section length.

To estimate the variability of the classification probabilities, each classifier were trained with a training dataset made of 12.5%, 25%, 50% and 80% of the identified data. The nonlinear Least Squares Model 3 of chapter 2 was used to predict the variance if all the identified data were used to train the classifier.

3.2.2. Classification of unidentified data

Once the optimal parameters were selected the final run of the classification process was made with 100% of the identified data. This final run generated the classifier algorithm. Once generated, this classifier algorithm was used to classify new data. If species identities of the new data are already known, then comparing the classification result with the reality allows

the user to confirm the reliability of the classifier. However, if species of the new data are unidentified, then only prior information concerning the classification groups (e.g abundance or density of the species classified) can be used to evaluate the reliability of the classifier.
The classification of unidentified data was done in several steps (Table 3-1B) using the PWC module to identify new data.

3.2.2.a Origin of the unidentified data

The unidentified data for this study were recordings collected from five (E16, A20, E17, E21) autonomous Ecological Acoustic Recorders (EARs, Lammers et al., 2008) positioned at the MORL and BOWL sites (Map 3-1) and one (D01) positioned in-shore within the Moray Firth Special Area of Conservation (S.A.C), which is one of the two UK areas of conservation for bottlenose dolphins (Cheney et al., 2012). The EARs recorded broadband sounds at 64 kHz sample rate discontinuously (30 minutes recording, followed by 30 minutes off) for periods ranging from 1 day to 25 days between July and October 2010 (Table 3-4).
To be used with the *2Sp* and *5Sp* classifiers the recordings were decimated to 48 kHz and processed with the PAMGUARD whistle and Moan detector prior to the classification.

Table 3-4: Details of EAR deployments from (SMRU ltd et al., 2011)

Site	Site	Deployment	Date Recovery	# Days
E16	MORL	22/09/2010	16/10/2010	24
A20	MORL	25/07/2010	15/08/2010	21
A22	MORL	22/09/2010	23/09/2010	1
E17	BOWL	24/07/2010	11/08/2010	18
E21	BOWL	16/08/2010	09/09/2010	24
D01	Sutors	07/10/2010	01/11/2010	25

3.2.2.b Classification process of the unidentified data

The PWC module used to identify new data works in real time or can process archived data. The recordings were processed with both the whistle detector and PWC modules activated. Each time a sound was detected, the sound frequency contours were divided in fragments of

the same length as used to create the classifier. These fragments were accumulated in sections until there was the same number of fragments as in the section used to create the classifier. The 9 parameters used in the PWC algorithm were then extracted and the classifier estimated the probability of the section to be one of the classification groups (chapter1, p26, stage *v*). The observed identification of the section was the species corresponding to the classification group with the larger probability. Then all the fragments were cleared and the PWC started accumulating new fragments. If there was less than five fragments within 10 minutes of recording then whatever the number of fragments within the section, this one was identified and a new section started when new fragments were detected. With this system some sections were classified despite not having the optimal number of fragments within it.

3.2.2.c Organisation of the sections in encounters

Only sections with the optimal length were used to analyse the classification result and short sections were discarded. When animals are passing close to hydrophones it is usual to get many whistles detected, as they are often travelling in group, prior to a gap without detections when the animals are too far to be detected. This period of high detections are commonly called *encounters*. By observing the classification result an encounter will be a period of time with many sections followed by a gap without sections. Grouping the sections in encounters and classifying these encounters allowed to be more accurate and to decrease the chance of misclassification. The identification of an encounter was the classification group with the higher average classification probability among all the complete sections of the encounter.
In this chapter given that recordings were made discontinuously every other 30 minutes, an encounter was defined by a succession of sections with less than 30 minutes between each of them.

3.2.2.d Analysis of the classification results

A manual verification was conducted by going through all the encounters to determine whether the contour classified were from dolphins or were false positive detections due to other noises. Classification results were then compared with data from previous visual studies.

3.3. Results

3.3.1. Training dataset

The number of whistle contours per species presented in the training is summarised in Table 3-5. Two species, bottlenose and common dolphins had more data than the others. Nevertheless, a reasonable amount of data was available for the other species that were included in the classifier.

Table 3-5: Number of whistle contours extracted for each species in the training data set.

Species	Number of whistle contours extracted
Bottlenose dolphin	61934
Common dolphin	69761
White-beaked dolphin	2554
White-sided dolphin	5505
Risso's dolphin	6358

3.3.2. Selection of the optimal fragment and section length

For both classifiers, the quality coefficient *Q* increased with fragment and section length and it reached a plateau at a fragment length of 25 bins (0.29 s) and a section length off 50 fragments (Figure 3-2,Figure 3-3). These parameters were close to the fragment lengths of 30 bins and section of 60 fragments measured in Gillespie et al. (2013).

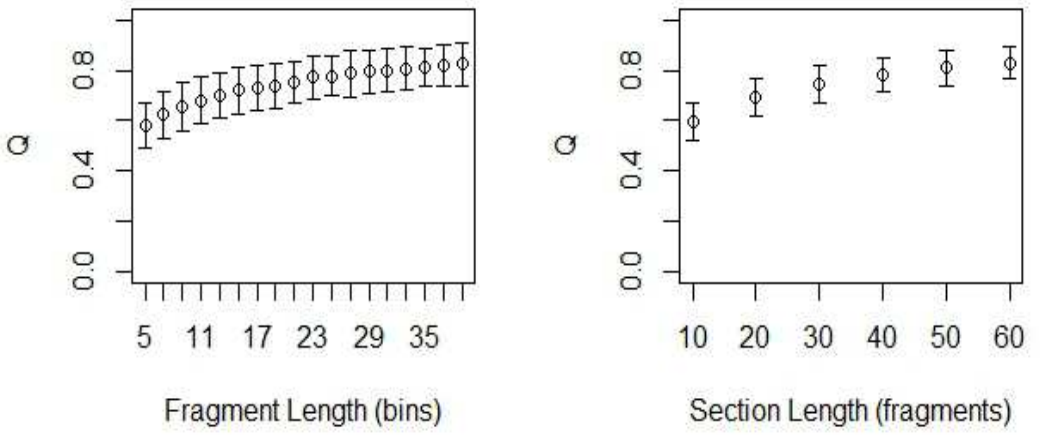


Figure 3-2: Quality coefficient *Q* of the *2Sp* classifier for varying fragment lengths (averaged over section lengths between 10 and 60 fragments) and varying section lengths (averaged over fragment lengths between 5 and 39 bins).

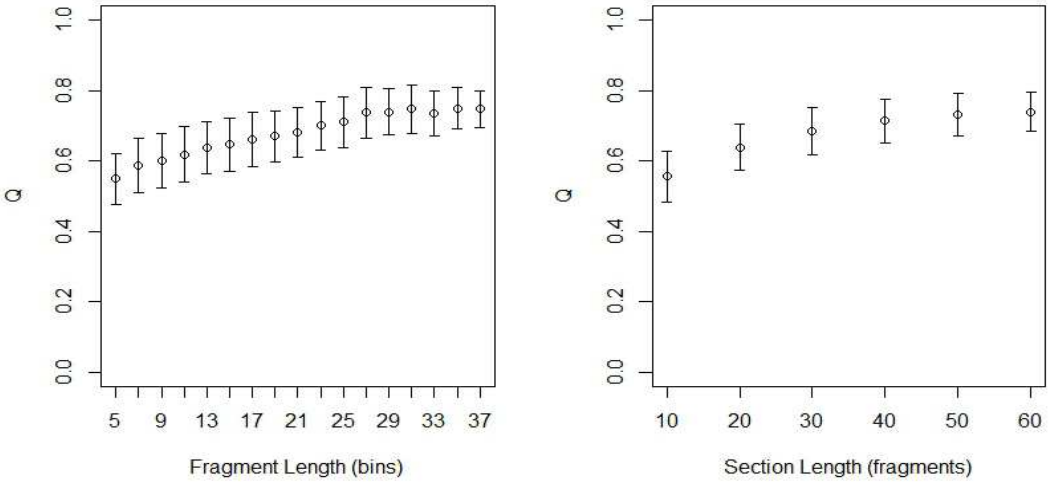


Figure 3-3: Quality coefficient *Q* of the *5Sp* classifier for varying fragment lengths (averaged over section lengths between 10 and 60 fragments) and varying section lengths (averaged over fragment lengths between 5 and 39 bins).

3.3.3. Training of the classifiers

3.3.3.a 2Sp classifier

The confusion matrix representing the classification probabilities of the 2Sp classifier is shown in Table 3-6. A t.test with an alternative hypothesis that the correct classification probabilities is not smaller than a random classification (>50%) proved that for both classification groups the null hypothesis failed to be accepted with a probability lower than 5%. For bottlenose dolphins, detections were correctly classified at 90.7% whereas 'Other' detections were correctly classified at 93.7%. The false positive classification probability was slightly higher for the 'Other' group (9.0%) than for the bottlenose dolphin group (6.5%).

Table 3-6: Confusion matrix of the 2Sp classifier. The classification probabilities are the probabilities observed when 80% of the training data are used to train the classifier. The standard deviation (in %, within the brackets) is an estimation if 100% of the data were used to train the classifier. BND=bottlenose dolphins, Other=all other species. *p* being the p-value of a t.test with the alternative hypothesis being the true difference in mean is not smaller than by chance.

	True Species			
Classified as %	**BND**	**Other**	**False Positive Classifications (%)**	***p***
BND	**90.7** (3.3)	6.3 (3.0)	6.5	$<2.10^{-16}$
Other	9.3 (3.3)	**93.7** (3.0)	9	$<2.10^{-16}$

3.3.3.b 5Sp classifier

The confusion matrix of the 5SP classifier is shown in Table 3-7. A similar t.test to the one applied on the 2Sp classifier, with a probability of being classified by chance of 20%, proved that the correct classification probabilities were significantly greater than a random classification and for four of the five species it was higher than 75%. Risso's dolphin's vocalisations seemed to be very distinctive from those of the other species with a correct classification probability close to 100% and both, false positive and false negative classification probability being very low, 1.6% and 0%, respectively. Bottlenose dolphin were still very well identified with a correct classification probability slightly smaller than with the *2Sp* classifier.

At the opposite end, white beaked dolphin was the species with the smallest correct classification probability of 59.8%, and classification events for this species were misclassified mostly (35.8%) as those of common dolphin.

The standard deviation of the correct classification probabilities for the five species was low, close to 10% of the correct classification probabilities. The standard deviations of the misclassification probabilities were often high relative to the estimated misclassification probabilities themselves.

Table 3-7: Confusion matrix for the 5Sp classifier. The classification probabilities were the probabilities observed when 80% of the training data were used to train the classifier. The standard deviation (in % within bracket) was an estimation if 100% of the data were used to train the classifier. BND=bottlenose dolphins, COD=common dolphins, RSD=Rissos' dolphins, WBD=white beaked dolphins, WSD= white side dolphins. *p* being the p-value of a t.test with the alternative hypothesis being the true difference in mean is not smaller than by chance.

	True Species						
Classified as %	**BND**	**COD**	**RSD**	**WBD**	**WSD**	**False Positive Classifications(%)**	***p***
BND	**86.6**(7.6)	3.3 (6.3)	0.0 (6.0)	2.0 (303)	0.0 (5.8)	5.8	$<2.10^{-16}$
COD	8.5 (6.9)	**77.3** (8.0)	0.0 (5.8)	35.8 (24.2)	18.6 (7.6)	44.9	$<2.10^{-16}$
RSD	1.6 (6.1)	0.0 (5.8)	**100** (5.9)	0.0 (5.8)	0.0 (5.8)	1.6	$<2.10^{-16}$
WBD	2.7 (6.2)	13.0 (7.2)	0.0 (5.8)	**59.8** (8.7)	4.1 (6.4)	24.9	$<2.10^{-16}$
WSD	0.6 (6.0)	6.4(6.7)	0.0 (5.8)	2.5 (6.3)	**77.3** (7.9)	11.0	$<2.10^{-16}$
False Negative Classifications	13.4	22.7	0.0	40.2	22.7		

3.3.3.c Classification of the EARs data with the 2Sp classifier

3.3.3.c.i Analysis of false detections

For all encounters, the spectrogram were investigated by eye to determine whether the encounter was correctly classified as dolphins or whether there had been any false detections (FD) due to artificial noise. An encounter was classified as FD if only all the contours within it were re classified as false detections. The majority of sounds identified as false detections were mechanical 'rubbing' sounds, potentially associated with a swivel on the mooring of the EARs deployment. These sounds generated an upsweeping tonal sound with several harmonics between 1.5-24KHz. (Figure 3-4).

On the 93 encounters detected from the EARs deployments, 40 were rejected as being false detections (Appendix A for details). The majority of them (80%) were detected at the E16 and A20 sites. Sites E17 and DO1 did not have any false detection and site E16 had only FD, so it was ignored for the rest of the analysis (Table 3-8).

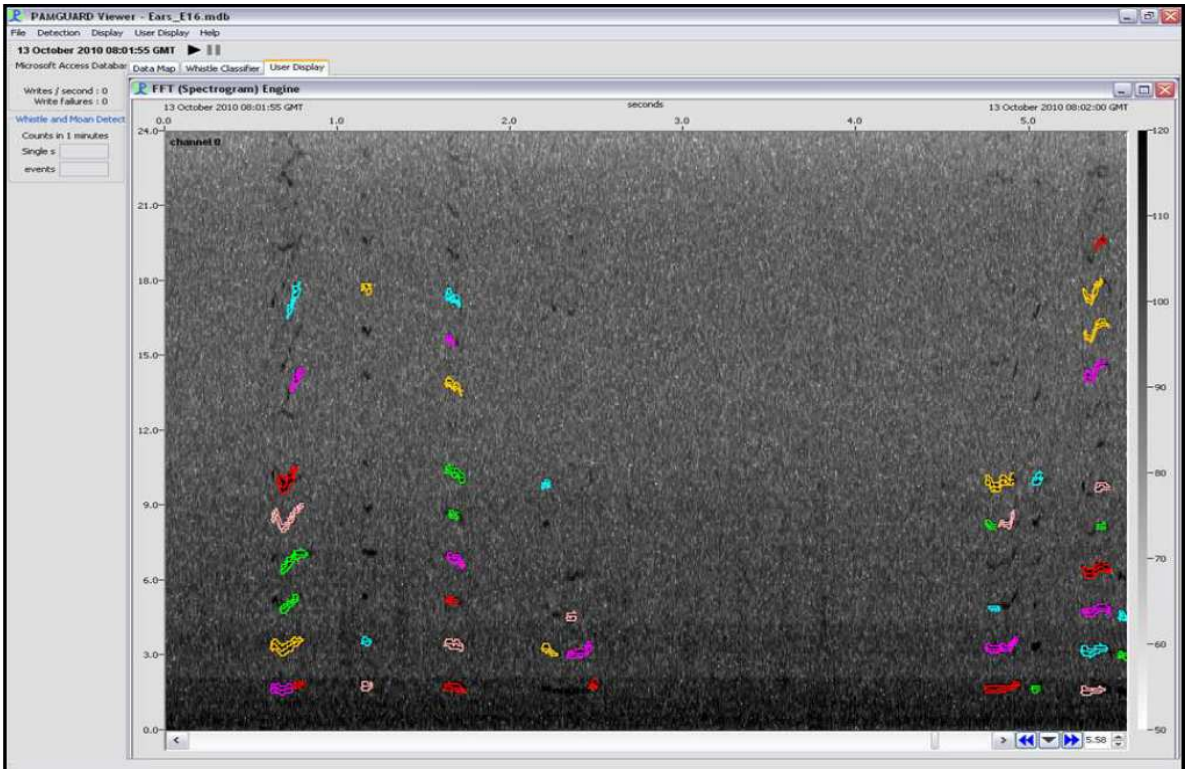
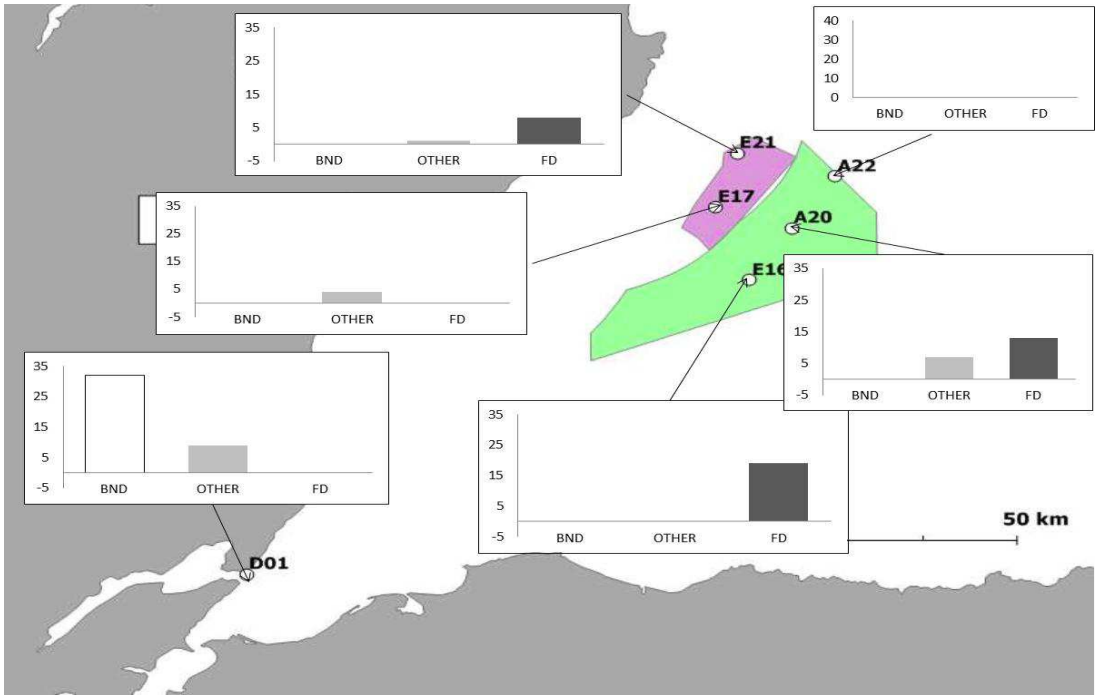


Figure 3-4: Screen capture from PWC of a "rubbing" false detection. Frequency is on the y-axis (0 to 24 kHz) and time (5.58seconds) is on the x-axis. The different colours show the contours generated by the PAMGAURD whistle detector. (SMRU ltd et al., 2011)

3.3.3.c.ii Analysis of the classification encounters

The summary of the classification by the 2Sp classifier is shown in Table 3-8 and in Map 3-3 (see Appendix A, Table A.1 for the full details of the classification). With the *2Sp* classifier, 32 encounters were identified as bottlenose dolphins, 21 as 'Other'. For the deployments E21 and E17 at the BOWL site, all the encounters not re-classified as FD (5) were classified as 'Other'.

At the MORL site, no detections were observed at the A22 deployment and seven encounters were classified as 'Other' at the A20 site. The EAR deployment at DO1 site was the only site with encounters (32) classified as BND, nine were classified as 'Other'.



Map 3-3 : Results of the classification of the EARs deployment using the 2Sp whistle classifier. Each bar represents the numbers of encounters classified as: bottlenose dolphins (BND) (white); 'other' dolphins species (OTHER, light grey); or as false detection (FD, dark grey).

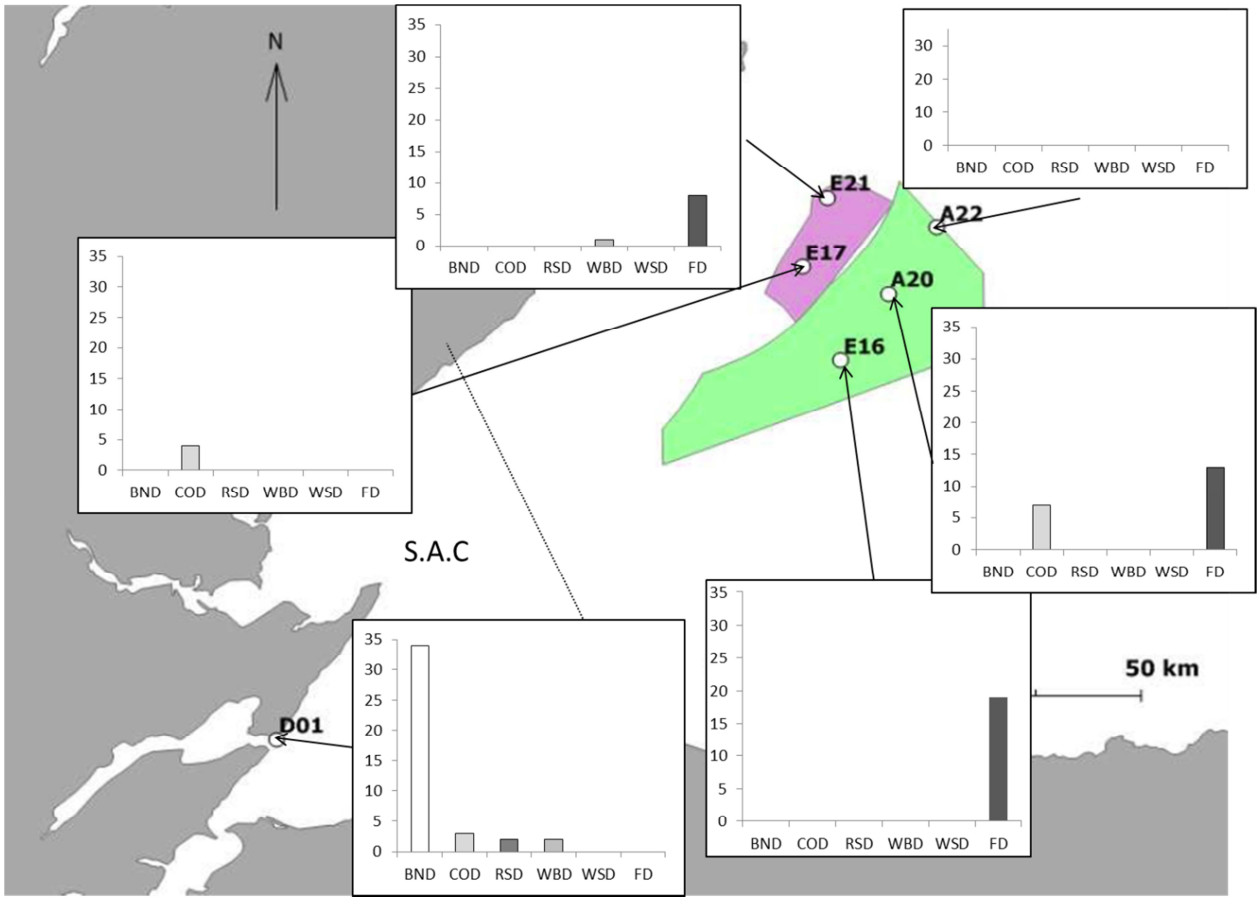
3.3.3.d Classification of the EARs data according to the 5Sp classifier

The four EARs deployments (E21, E17, A20, D01) for which some encounters have been classified as BND or 'Other' by the 2Sp classifier were subsequently classified using the 5Sp classifier. The summary of the classification result are show in Table 3-8 and Map 3-4, and the full detail are in the Appendix A, Table A.2. For the deployment at the wind farm sites (A20, E17, E21) no encounters were classified as BND and events classified as 'Other'(12 in total) were classified as COD (11) and WBD (1), respectively. As with the *2Sp* classifier the

encounters of the DO1 deployment were mostly classified as BND. One encounter classified as BND by the *2Sp* classifier was classified as RSD by the *5Sp* classifier. Three encounters classified as 'Other' by the *2Sp* classifier were now classified as BND by the *5Sp* classifier. The remaining 'Other' encounters were classified as COD, RSD and WBD .

Table 3-8: Comparison of the EARs recording classification by the *2Sp* and *5Sp* classifier. Only EARs deployments with dolphins encounters have been processed with the 5Sp classifier.

			***2Sp* classifier**		***5Sp* classifier**				
Site No.	**Nbs of Encounters**	**FD**	**BND**	**OTHER**	**BND**	**COD**	**RSD**	**WBD**	**WSD**
E16	19	19	0	0	0	0	0	0	0
A20	20	13	0	7	0	7	0	0	0
A22	0	0	0	0	0	0	0	0	0
E17	4	0	0	4	0	4	0	0	0
E21	9	8	0	1	0	0	0	1	0
DO1	41	0	32	9	34	3	2	2	0
TOTAL	**93**	**40**	**35**	**21**	**34**	**14**	**2**	**3**	**0**



Map 3-4: Results of the classification of the EARs deployment using the 5Sp whistle classifier. Each bar represents the numbers of encounters classified as: bottlenose dolphins (BND) (white); 'other' dolphins species (OTHER, light grey); false detection (FD, dark grey); common dolphins (COD); Risso's dolphins (RSD); white beaked dolphins (WBD); white sided dolphins (WSD).

3.4. Discussion

Regular visual aerial surveys have been conducted in the inner and outer Moray Firth since 2004. During these surveys common dolphins, white-beaked dolphins and Risso's dolphins were regularly sighted at the wind farm sites (Moray Offshore Renewables ltd, 2010) In contrast, in the S.A.C area (Map 3-1), the large majority of visual detections were of bottlenose dolphins, with very few reports of sightings of common dolphins and white beaked dolphins. The classification result supported these findings, at least during the sampling period of the study (July-October). None of the five EARs deployment within the wind farm areas recorded whistles encounters that could be attributed to bottlenose dolphins. However, encounters were classified as common dolphins, Risso's dolphins or white beaked dolphins by the *5Sp* classifier which is consistent with the visual survey data. The EAR deployment within the S.A.C was the deployment with most of the detections (41% of all the encounters)

and the only deployment with detections of bottlenose dolphins. Seventy eight per cent (78 %) of these detections were classified as bottlenose dolphins. The classification results estimated that three (7.3%) encounters of common dolphin and two (4.8%) of white beaked dolphin occurred in a 25 days period of time. Even though it is possible to observe common dolphin and white beaked dolphins within this area (Moray Offshore Renewables ltd, 2010), these encounters should be rare. Furthermore, the *5Sp* classifier predicted that 8.5% and 2.8% of bottlenose dolphin encounters should be misclassified as common and white beaked dolphin respectively. Hence, it is probable that the 3 (7.3%) encounters classified as common dolphin and the 2 (4.8%) of white beaked dolphin encounters were misclassifications by the *5Sp* classifier.

Rissos' dolphins have never been observed in the S.A.C, so these classifications were probably the result of misclassification by the classifier. The *5Sp* classifier predicted that on average 1.6% (sd=6.1) of the classification events for bottlenose dolphins should be misclassified as Rissos' dolphin. In the DO1 deployment the Risso's encounter represented 4.8% of all the encounters that is within the standard deviation of the expected misclassification probability.

The five species classified in this study have a large overlap in the frequency range of their sounds (chapter1 table1.1) given that one third of the whistle classifier parameters depends on the mean frequency, it is difficult to find an algorithm which will discriminate these species better using the mean frequency parameter. Increasing the amount of training data may improve the classifier by reducing the misclassification probabilities.

In this study the main objective was only to detect the presence of the bottlenose dolphins. Given the result and the clear difference in numbers of occurrence of classification events of bottlenose dolphins between the wind farm site and the S.A.C site, we can be confident that bottlenose dolphins were not frequent at the MORL and BOWL sites, at least between July and October 2010.

It is impossible to make a more accurate statement on the presence or absence of bottlenose dolphins in the area. Some missed-detections are to be expected due to the fact that the recordings are not continuous (30 minutes of recording, followed by 30 min off) and because of whistle rates being low or quiet whistles not reaching the detection threshold. Another important source of misclassification and/or missed-detection is the presence of high ambient noise. Depending on its frequency range, ambient noise can mask parts of or the totality of the signal of interest. The whistle detector is then not able to detect the whistles themselves.

For three of the five deployments at the wind farm sites, on average 91.3% of the encounters were false detections, and so irrelevant for the classification, because of the mooring structures (e.g. swivels, loose chains etc..). With the *2Sp* classifier, all of these false detections were misclassified as bottlenose dolphins, whereas with the *5Sp* classifier they were mainly classified as Risso's dolphins. In the case of this study, a control for misclassifications caused by noise was possible in form of a manual operator analysing all the classification events. However, for a bigger data set or during real time classification, this may not be feasible. Nevertheless, it may still be possible to re-analyse a sample of the data manually to detect any recurrent noise generating misclassifications and to set up some filters to remove these signals if they are outside the frequency range of the species of interest. For common noise sounds the classifier could be trained with this noise incorporated as an extra species.

Because this project focused on coastal species, it was relatively easy to build the training dataset of good/high quality based on local coastal surveys. This is not always possible.
Next chapter illustrates one possible way of developing a similar automatic classifier from acoustic data collected during a large scale offshore survey.

Chapter 4: Classification of data from a less reliable training dataset

4.1. Introduction

Large scale cetaceans surveys such as the North Atlantic Sightings Surveys (NASS) (Lockyer and Pike, 2009), the Southern Ocean Whale and Ecosystem Research Programme (SOWER) (Ensor et al., 2010) or the Small Cetaceans in the European Atlantic and North Sea survey (SCAN's), (SCANS-II, 2008), are encouraged by governmental and non-governmental agencies to estimate abundance of species and to detect changes in the distribution of the species. The information collected during these surveys are used to make management decisions.

These surveys often use a standardised survey protocol across several vessels and a large geographic area (Ensor et al., 2008; SCANS-II, 2008), and both visual and acoustic detection systems are commonly used to detect marine mammals. To be able to use the acoustic data, reliable classifier need to be developed to identify the species detected. As explained in the previous two chapters the classifier performance is dependent on the training dataset. An ideal training data set would consist of acoustic recordings made in the presence of visually identified species. These data could have been collected in a previous survey (as in chapter 3) or during the survey itself. Where the degree of intra species variation in whistles is high, the classifier performs better if trained with data collected in the same area as the survey. The offshore location, cost and geographic scale of some surveys often make it difficult and costly to organise pre-surveys with the sole objective of collecting an acoustic training data set. When the classifier training data set is collected at the same time as the survey it is necessary to associate visual detections (sightings) with acoustic detections to be able to assign species identity to acoustic recordings. Once a classifier is created, it can be used to identify detections made during the survey that are not associated with visual detections.

It is often the case that during combined visual and acoustics surveys, e.g. SCANS-II, CODA, hydrophone arrays are towed a few hundred meters behind the visual survey platform. This makes the task of associating visual detections with acoustic detections challenging, and requires numerous assumptions to be made. However, without the development of automated acoustic classifiers, acoustic data from most cetacean species cannot be used in any further analyses. Currently only those species with very distinctive

vocalisations such as the sperm whale (*Physeter macrocephalus*) (Wahlberg, 2002), harbour porpoise (*Phocoena phocoena*) (Goodson and Sturtivant, 1996) and some species of baleen whale (Gillespie, 2004; Mellinger and Clark, 1997) can be reliably detected and classified to species, and it is for these species that it is possible to estimate animal abundance using acoustic detection only (Barlow and Taylor, 2005; Gerrodette et al., 2011; Kyhn et al., 2012; Marques et al., 2011).

In July 2007 a large scale survey, Cetacean Offshore Distribution and Abundance in the European Atlantic (CODA) involving several vessels, was organised in European Atlantic waters beyond the continental shelf. The principal aims of this cooperative European project were to "(1) estimate the abundance of common dolphin (*Delphinus delphis*) and other cetacean species in offshore European Atlantic waters, (2) to assess the impact of by catch, and finally (3) to recommend safe by catch limits for the common dolphin" (CODA, 2009). During this survey both acoustic and visual data were collected.

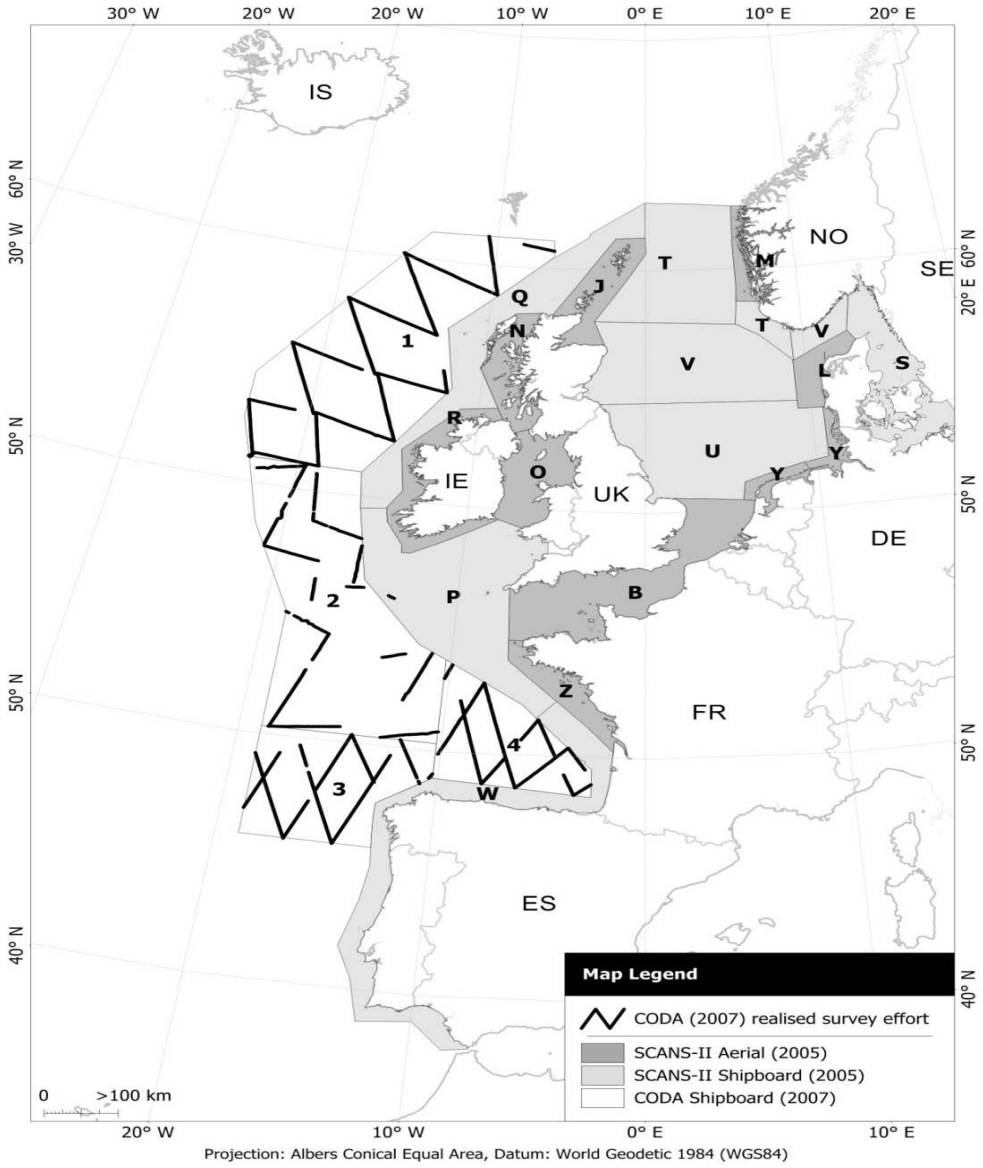
This chapter presents a method to create a classifier training dataset from the CODA visual and acoustic detections, and uses the classifier to identify acoustic detections not associated with a visual detection. Then the results are analysed to identify which parameters influence the quality of the classifier. The challenge of this chapter, contrary to the previous one, was that the acoustic data used to train the classifiers came from the survey itself and they were collected independently to the visual detections. The first part of this chapter describes the creation of a training dataset when acoustic detections were not identified in real time. This was done by relating visual identifications of sighted species to acoustic detections. Two training dataset were created, with data recorded in different area of the survey. Once the training datasets were created, the second part describes the creation of a classifier for each dataset with a similar approach than in the previous chapter. Each classifier was tested on the dataset not used to create it. Finally in a last part, these classifiers were used to identify acoustic detections for which no species identification was possible from the visual data. To evaluate the quality of the classifiers, these non-identified acoustic data were classified with a classifier created from a good training dataset independent of the CODA survey data. The classifiers were created using the same PAMGUARD whistle classifier module (Gillespie et al., 2013), as used in the previous chapters.

4.2. Datasets

During the CODA survey, five ships surveyed four offshore survey blocks that extended from the Faroe Islands in the North to the Portuguese EEZ in the South (Map 4-1). Each vessel sailed pre-designed transects and surveys were carried out using visual and acoustic methods. Data from four ships cruising in the blocks 2 to 4 were used for this analysis (two French vessels : A634 Rari and F735 Germinal, two Spanish vessels :RV. Investigador and RV. Cornide de Saavedra), data from the fifth ship (M/V Mars Chaser) surveying in block 1 was excluded from the analysis due to high levels of noise in the acoustic data. For clarity in this chapter data collected in block 2 are referred to as the French dataset while data collected in blocks 3 and 4 are referred to as the Spanish dataset.

4.2.1. Visual survey

Visual surveys were conducted using the survey methods developed and employed during the SCANS-II project (SCANS-II, 2008). A double platform of observers was used, with a “Primary” and a “Tracker” observer teams. The “Primary” team consisted of two observers searching with naked eyes an area ahead and at close distance to the vessel (out to 500m). The “Tracker” team was composed of two observers positioned on a second, higher, platform to scan an area far away from the ship using big eyes (10x25) or 7*50 binoculars (CODA, 2009). For each sighting (visual detection), information including the vessel's position, species identification, confidence level of this identification, radial distance and sighting angle relative to the vessel's heading to each group, behaviours and cues were recorded.



Map 4-1: CODA survey area and survey blocks (CODA, 2009). Block 2 was surveyed by French vessels in this chapter they are referred to the "French dataset". Blocks 3 and 4 were surveyed by Spanish vessels and they are referred to the "Spanish dataset".

4.2.2. Acoustic survey

4.2.2.a Description of the recording systems

The aim of the acoustic survey was to detect as many odontocete species as possible with a focus on sperm whales, beaked whales, oceanic dolphins and harbour porpoises (CODA, 2009). Two automated detection systems were used to record the wide range of frequencies emitted by these species:

1. A high frequency (sampling rate of 500 kHz) automatic click detector designed to detect harbour porpoise.

2. The second system recorded continuously at 192 kHz giving an effective system bandwidth of 2 kHz to 90 kHz making it sensitive to all other odontocete species (CODA, 2009).

A hydrophone array with two sensor sections was towed behind each survey vessel. The first sensor section consisted of 2 hydrophones at 200, 203m respectively from the dry end of the cable, while the second sensor section consisted of three 3 hydrophones at 400, 400.25 and 403m. Distance between elements was optimised for the localisation of harbour porpoise and sperm whale clicks. Hydrophone elements in the second sensor section were towed further behind the vessel to minimise the impact of the vessel noise on recordings. Only recordings coming from hydrophones in last sensor section were used for this analysis. Data were collected automatically during the day, using IFAW's Logger 2000- software (Gillespie et al., 2010) until it was switched off in the evening or until it crashed. The automatic recording system recorded continuously to hard disk using the *.wav format and recording were ranging from 1 seconds to 647 seconds with an average recording length of 427 seconds.
On shore each recording was re-processed with the PAMGUARD Whistle and Moan detector (Gillespie et al., 2013) using a high pass filter (1.5 KHz) to remove low frequency sounds generated by ambient noise. For each recording a "contour file" containing all the time frequency contours detected was created.

4.2.2.b False positive analysis

The automatic whistle and moan detector is not perfect and there are numerous sources of noise (electric, mechanical, sonar, echo sounders…) that can create false positive detections. These false positive contours can generate a non-negligible bias in the quality of the classifier. The main characteristic differences between a whistle contour and another non biological noise contour are the length and the regularity of occurrence of these noises. A false positive analysis was conducted to minimise the selection of these contours before the training process.

A false positive analysis consists on randomly selecting acoustic detection contours and checking visually on the spectrogram if the contour was made from a dolphin or not. Given the amount of acoustic data, to optimise the random selection of the contours, every recording

with acoustic detections was divided into one minute bins. The Total Contour Length per minute (L_m) was calculated by summing the length of all the whistle contours within the minute.

Sixty per cent and 80% (Figure 4-1) of the L_m were less than four seconds long for the French and Spanish dataset respectively.

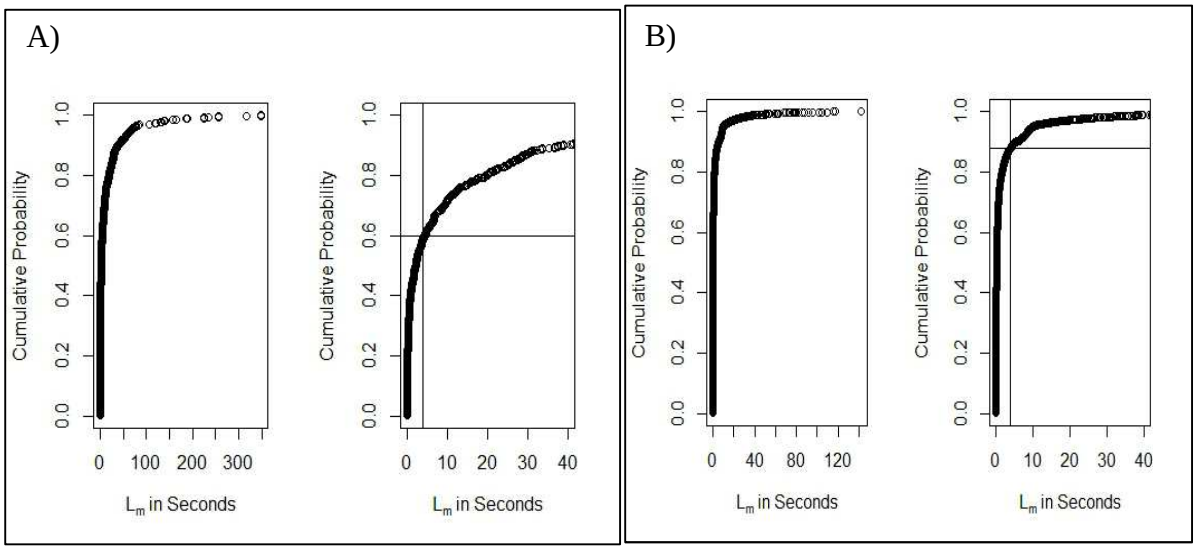


Figure 4-1: Total Contour Length per minutes for A) the French dataset and B) the Spanish data set. Figures on the right are zoomed to 40 s with the vertical line being placed at 4 s of contour length per minute.

From these results, the L_m was organised into seven categories. Because the length of false positive contours was expected to be small, categories reflect this expectation and smaller lengths were oversampled. The seven categories ranging from less than 0.1 second to more than four seconds were such:

1) $L_m \leq 0.1s$
2) $0.1s < L_m \leq 0.5s$
3) $0.5s < L_m \leq 1s$
4) $1s < L_m \leq 2s$
5) $2s < L_m \leq 3s$
6) $3s < L_m \leq 4s$
7) $L_m > 4s$

For each dataset and each of the seven categories, a maximum of 100 minutes were randomly selected. For categories with less than 100 minutes available, all the minutes were analysed. The spectrogram of each selected minute was visually inspected and each detected contour was classified as either false positive or whistle contour. For 98% of the minutes all the contours within the minute were either false positive or whistle contours, so the minutes were categorized as false detections (FD) contours or as whistle contours (W) otherwise. The 2% remaining minutes contained both false positive and whistle contours, the proportion of false positive contours was measured if this proportion was greater than 50% then the minute and so the contours within it were categorized as FD otherwise they were categorized as W.

Then the contour lengths of the contours within the FD minutes were compared with the contour lengths of the contours within the W minutes and the optimal contour length which discarded most of the FD contour in the same time as keeping most of the W contours was selected as a threshold. All the contours with a length under this threshold were discarded the longer contours were used for the rest of the analysis.

4.3. Methods

4.3.1. Creation of the training datasets

The inputs of the PWC are the time frequency contour files (one for each recording) extracted by the automatic whistle and moan detector. To train the classifier each recording needed to be associated with one visually identified species. This was done by linking recordings to sightings. This selection process was done in several stages described in a schematic diagram (Figure 4-2 *i* to *v*) and in the following paragraphs. The main stages were to (*i*) select the visual detections of interest, (*ii*) extract the acoustic data of interest (*iii*) link visual and acoustic detections, (*iv*) train the classifier, and (*v*) test it. This process was done individually for both the French and Spanish dataset.

4.3.1.a Selection of visual detections

During the survey seven whistling species were visually detected: bottlenose dolphin (*Tursiops truncatus*), common dolphin *(Delphinus delphis*), striped dolphin (*Stenella coeruleoalba*), killer whales (*Orcinus orca*), long finned pilot whale (*Globicephala. melas*), short finned pilot whale (*G. macrorhynchus*), and Risso's dolphin (*Grampus griseus*).

Common and striped dolphin were often observed together in large mixed groups, and in this situation the visual observer identified the groups as common and striped (C&S) .

The CODA visual survey protocol required observers to give the degree of confidence (High, Medium, Low) of species identification for each sighting (CODA, 2009). For quality assurance purpose, all primary and tracker sightings with high or assumed high (blank in the database) identification confidence were selected. *(*Figure 4-2, *i.a).*

4.3.1.b Link between visual and acoustic detection

4.3.1.b.i Time at hydrophones (Figure 4-2 iii.a)

As mentioned in the description of the data, the visual observers looked for animals ahead of the vessel, whereas the hydrophones, from which acoustic data were extracted, were towed up to 400m behind the vessel. Due to the distance between the visual platform and acoustic platform, the probability of simultaneously detecting the same animal both visually and acoustically was not optimal. Thus the following method was adopted for linking visual and acoustic detections; For each visual detection, the time when the hydrophones were at the perpendicular distance of the sighting (this variable will be called "abeam time": $\boldsymbol{T_{Ab}}$) was estimated using the formula below

$$T_{Ab} = \frac{cos(\hat{A})R + 400}{5.14} + T_V \qquad (4\text{-}1)$$

where $\hat{A}$ was the angle between the bearing of the vessel and the animal, radial distance (R) estimated by the visual observer. Then the distance between the visual team and the hydrophone was added (400m). This total distance was dived by the vessel speed 5.14 meters per seconds and added to the time of visual observation (T_V).

It was assumed that the animal did not move significantly between the visual detection and the time the hydrophones were abeam of the animals.

4.3.1.b.ii Acoustic selection (Figure 4-2 *iii.b)*

Each visual detection (Primary and Tracker) of species of interest with a high confidence level of identification was associated with the acoustic recordings corresponding to the "abeam time" of detection. To be sure not to miss any vocalisations, while at the same time

ensuring not to select recordings with two different species several rules were applied to be conservative on the choice of recordings:

- immediate recordings before and after the "abeam time" corresponding to the visual detection were selected;
- if within a selected recording more than one species was observed the recording was not selected for the analysis;
- if an adjacent recording contained a visual detection of a different species these adjacent recordings were not selected;
- the last two rules were not applied to the common (COD), striped (STD) and common/striped (C&S) detections. Indeed, during the visual survey an initial sighting would be made and then consecutive re-sightings were made during which the confidence of species identification went up. Common and striped dolphin were regularly observed in large mixed groups (C&S, common AND striped), within these mixed groups smaller, single species subgroups were observed (common OR striped; so that consecutive re-sightings separated by 5 minutes or less would alternate between groups consisting entirely of common dolphins and groups consisting entirely of striped dolphins. For this reason if any of these three groups (C&S, COD or STD) were sighted within the same or adjacent recordings, these recordings were selected and identified as CSD detections.

4.3.2. Creation of the classifiers

Four classifiers were trained and tested using the CODA data; two with the French dataset and two with the Spanish dataset. For each dataset a first classifier, called *2Sp French* classifier and *3Sp Spanish* classifier were trained with all the detections from COD, STD and C&S pooled in one unique classification group (CSD). This setup was a conservative approach which matched with the misidentification of these species by the visual teams. Then each dataset was used to train a classifier with the COD, STD and C&S detections representing a classification group each. They were called *4Sp French* classifier and *5Sp Spanish* classifier.

Finally a last classifier, called the *North Atlantic classifier*, has been trained using the data of Gillespie et al., (2013). This classifier was trained with the same species group as the *3Sp classifier* and with the optimal fragment and sections length measured by Gillespie et al.

(2013). This classifier was made using data recorded in different areas of the North Atlantic ocean generally from a small sailing research vessel in the vicinity of groups of dolphins or made while underway with dolphins close to the vessel (Gillespie et al., 2013)

The training was done following the method developed in the previous chapter (chapter 3 2.1 p 43). To identify the optimal fragment and section length, the quality coefficient (*Q*) was calculated on the pooled French and Spanish datasets. Fragments and sections ranging respectively from 5 to 15 bins (27ms to 80ms) and 10 to 30 fragments were tested.

Each classifier was represented by its confusion matrix when 80% of the training data were used to train the classifier. To estimate the precision of the classification probabilities if 100% of the training data were used to train it, each classifier was trained with different proportions of training data as described in (chapter 2). However the final algorithm of the classifier which was used to classify new data was created using 100% of the training data.

4.3.3. Classification of new data

To analyse the potential effect of acoustic differences between cetacean populations and the sensitivity of the classifier to the data, the acoustic detections of the French dataset were classified using the Spanish (*3Sp* and *5Sp*) classifier algorithms and the Spanish dataset were classified using French (*2Sp* and *4Sp*) classifier algorithms. Then both datasets were classified with the *North Atlantic* classifier.
Finally recordings without visual identifications were classified using the classifiers trained with data from the same detection area and the *North Atlantic* classifier.

The results of these classifications were presented in two different ways. First, as a confusion matrix, similar to the output from PAMGUARD, for the Spanish and French training data,
secondly, as in the previous chapter, sections were grouped in encounters. In this chapter the definition of an encounter is slightly different from the previous chapter 3. This difference is due to the type of hydrophones used and the recording pattern. In the previous chapter the hydrophones were bottom mounted with a discontinuous recording pattern, and the animals moved relative to them, whereas in this chapter the hydrophones recorded continuously and moved with the vessel, and the animals were assumed to be stationary with respect to the hydrophones; i.e. tow speed >> swim speed. So the interaction time between animals and

hydrophones were likely to have been shorter during this survey. For this reason and the observations of all the classification events, a gap of 10 minutes without any classification event was selected to define two encounters. The identification of an encounter was the classification group with the higher average classification probability among all the sections on the encounter (chapter 2: 1. *PAMGUARD whistle classi*fier stage v, *p2.1*).

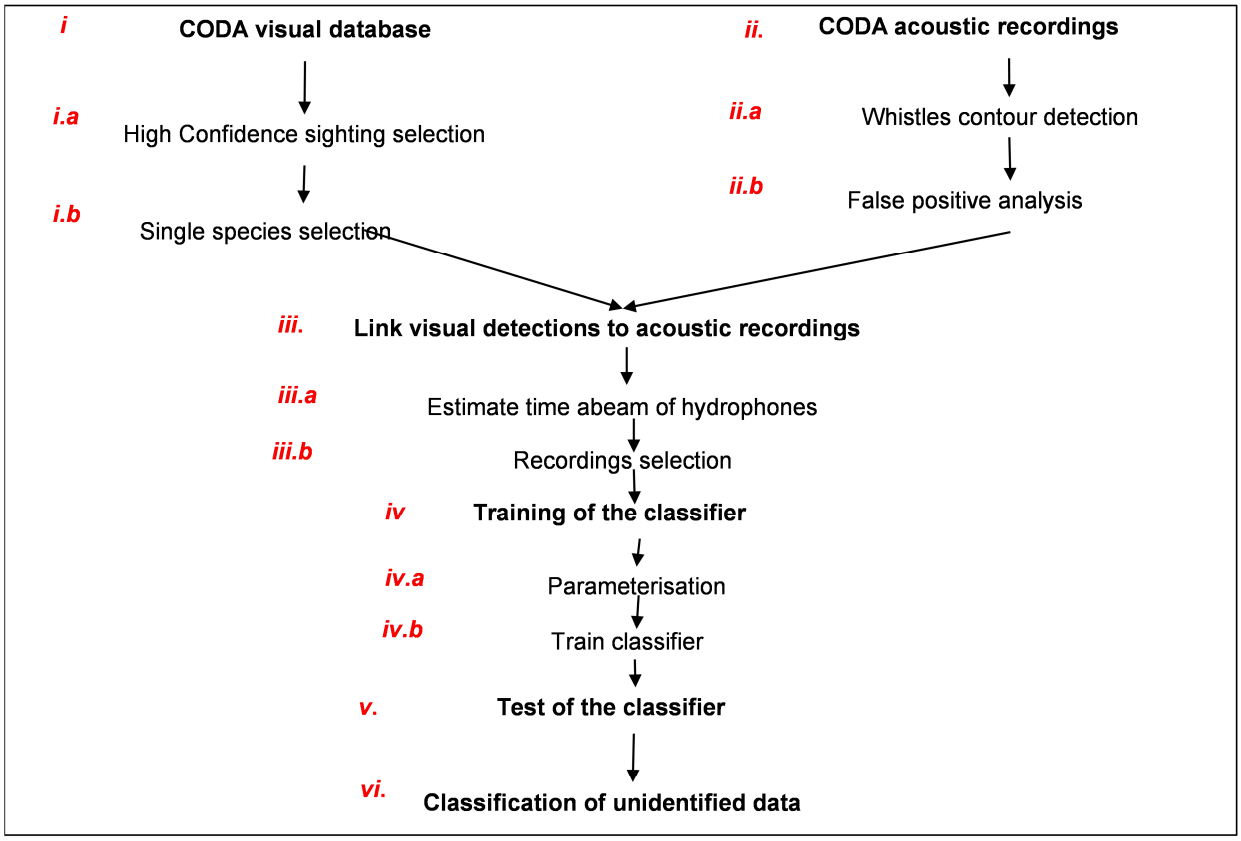


Figure 4-2: Schematic diagram of the data selection and decision process. (i) Selection of visual data with a high confidence of species identification. (ii) Detection and selection of whistles contour and discard of the false positive contours. (iii) Creation of the training dataset by assigning sightings to recordings. (iv) Training of the classifiers with the datasets. (v) Testing of the classifiers on identified data.(vi) use of the classifier to identify new data.

4.4. Results

4.4.1. Visual detection selection

Of the 1257 (782 for the Spanish data and 475 for the French data) primary and tracker visual detections between the four vessels, 443 (35.2%) were sightings of whistling species with 353 of this sightings identified with a high or assumed high confidents level and used in this chapter.

Eighty per cent of the selected sightings were of common and striped dolphin (CSD) species individually or together (Table 4-1). The other whistling species identified with confidence were bottlenose dolphin, pilot whale (both long finned pilot and short finned pilot whales) and Risso's dolphin.

More sightings, but fewer species were reported in the Spanish data set (five species for seven in the French dataset) (Table 4-1).

Table 4-1: Numbers of visual detections with a high or assumed high confidence level on the French and Spanish vessels.

Species	French data	Spanish data	TOTAL
Bottlenose dolphin	20	9	**29**
Common dolphin	37	119	**156**
Striped dolphin	10	39	**49**
Common and Striped dolphin	3	72	**75**
Long or short finned pilot whale	1	0	**1**
Long finned pilot whale	23	19	**42**
Risso's dolphin	1	0	**1**
TOTAL	**95**	**258**	**353**

4.4.2. Acoustic Data

4.4.2.a Quantitative description

The French and the Spanish acoustic datasets were made of 1367 (223.43.77 hours) and 2086 (250 hours) recordings respectively. Among them 51% of the French recordings and 92.3% of the Spanish recordings contained acoustic detections.

Table 4-2: Summary of the numbers (*n*) of recordings in total, with all the acoustic detections and when the false positive detections (FD) have been removed. Also summary of the total number of acoustic detections contours as well as the number of acoustic contours used for the rest of the analysis when the false positive detections were removed.

	French		Spanish	
	***n* Recordings**	***n* Contours**	***n* Recordings**	***n* Contours**
TOTAL	1367		2086	
With all Detections	697	92666	1925	77821
Without FD	102	23074	451	31676
With visual detections	34		135	
Without visual detections	68		316	

4.4.2.b False detection removal

Four hundred and seventy two (472) minutes and 558 minutes were analysed from the French and Spanish datasets respectively. For all minutes with a total contour length greater than 0.1 seconds the contour lengths from the FD minutes differed significantly from the contour lengths from the W minutes. The average contour length in the FD minutes was 0.07 seconds whereas the average contour length in the W minutes was 0.14 seconds (Figure 4-3).

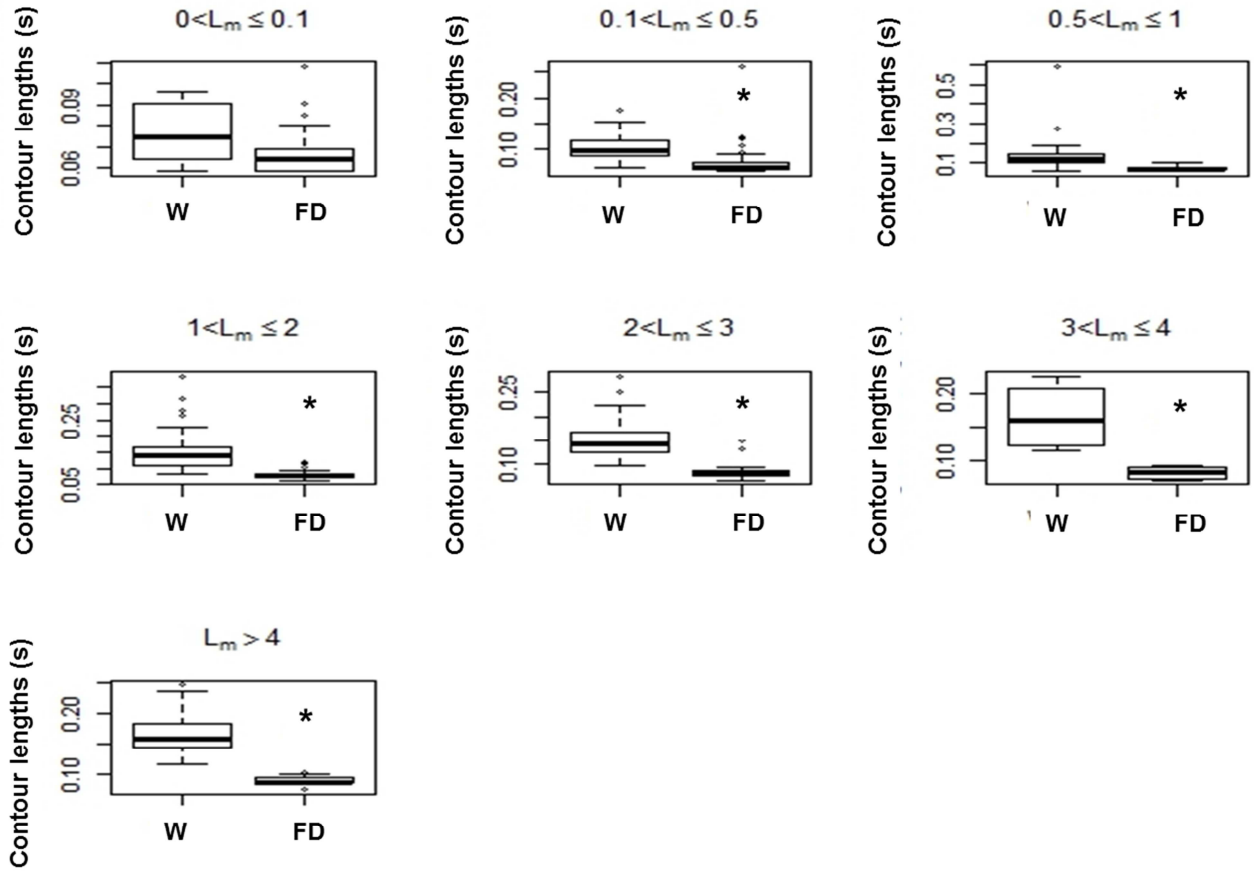


Figure 4-3: Distribution of the false positive detections (FD) and whistle (W) contour lengths for each category of L_m. The * indicates if the mean difference between the false detection contours and whistle contours was significant with a probability (p)<0.05.

The optimum contour length to discard the maximum of false positive meanwhile keeping the maximum number of whistle contours was 0.10 seconds. With a contour length of 0.10 s for both datasets 96% of the false detection contours were removed whereas 79% and 84% of respectively the French and Spanish whistle contours were kept (Table 4-3).

Table 4-3: Proportion (%) of detection contour lengths within the false detections (FD) and whistles (W) minutes below the contour length for each dataset.

		Contour length (s)				
	Contour class	**≤0.07**	**≤0.08**	**≤0.09**	**≤0.10**	**≤0.11**
French	**FD**	43%	76%	89%	96%	97%
	W	7%	11%	17%	21%	26%
Spanish	**FD**	47%	73%	90%	96%	98%
	W	4%	4%	9%	16%	25%

Once the false positive detection contours were removed only 40% and 25% of, respectively, the totality of the French and Spanish detection contours remained (Table 4-2) to be associated with the sightings of whistling species.

4.4.3. Link between Acoustic and Visual observations

The next stage in the creation of the classifier training dataset was to associate the 353 sightings (Table 4-1) of whistling species, for which the observer was highly confident on the species identification, to the 553 recordings with 54 750 whistle contours (Table 4-2).
In the French dataset 32% of the false detection free recordings were associated with at least one visual detection, and in the Spanish dataset 30% of these recordings were associated with at least one visual detection.

Finally Table 4-4 summarises the number of contours assigned to each species used to train the whistle classifiers. These contours have a minimum length of 0.10 seconds and a range in frequency from 1500Hz to 48000Hz. Contours above 48000 Hz were not selected as they may have contained other non-biological sounds and the frequency range for the species of interest was not higher than 24 kHz (chapter 1, table 1.1).

Table 4-4: Numbers of whistles contours used in the whistle classifier for each species and datasets.

Species	Abbreviation	French dataset	Spanish dataset	TOTAL
Bottlenose dolphin	BND	2	53	**55**
Common dolphin	COD	2164	18618	**20782**
Striped dolphin	STD	247	973	**1220**
Common/Striped dolphin	C&S	110	2917	**3027**
Long or short finned pilot whale	FPW	842	17	**859**
Risso's dolphin	RSD	3	0	**3**
TOTAL		**3368**	**22578**	**25946**

4.4.4. Parameter optimisation

The average *Q* over species and section length showed that a fragment length of 11bins (59 ms) gave the best classification result (Figure 4-4). *Q* increased slightly when the section length increased from 10 to 25 fragments per section. For some species not enough data were available to generate sections of 30 fragments of 11 to 15 bins long. This lack of data could explain the decrease of the average *Q* across all fragments when the section length reached 30 fragments. Even for sections of 25 fragments when the fragments length was of 13 or 15 bins some species did not have enough data to be part of the classifier.

So to insure to have enough data to train and test the classifiers for each species the optimal fragments length of 11 bins and section length of 20 fragments were selected to train the classifiers with the different datasets. While this very short fragment and section length (they were of respectively 25 and 60 in the previous chapter and in Gillespie et al., 2013), the very small number of contours assigned to Risso's dolphins made them unusable as it would not be possible to create a training and a testing section. For the same reason the bottlenose dolphin contours were excluded from the French whistle classifier while they were used in the Spanish whistle classifier. However, there were just enough pilot whale contours in the Spanish dataset to create at least one training and one testing section in the classifier.

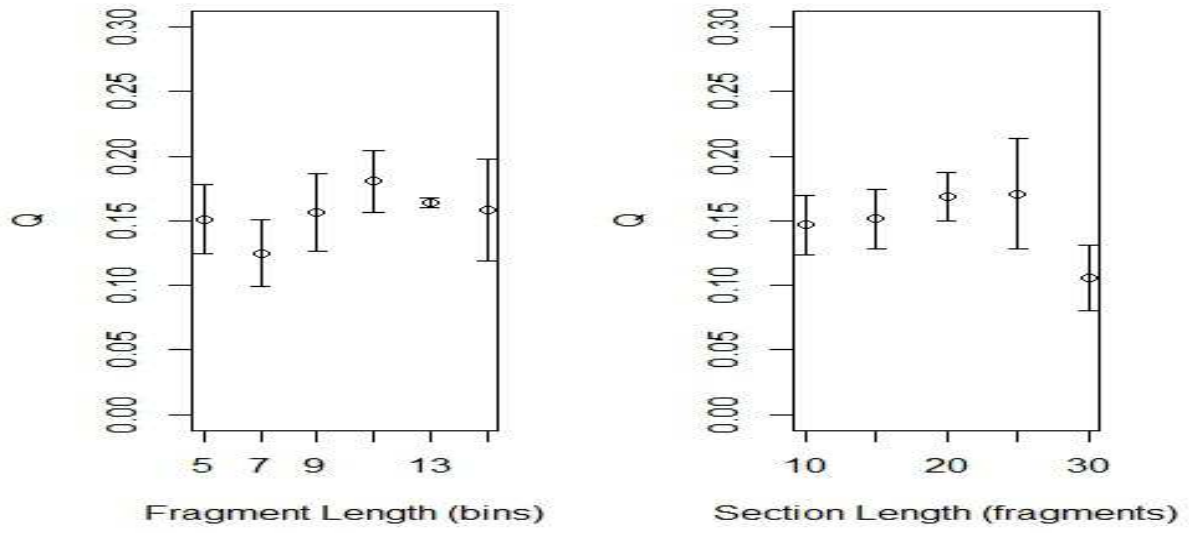


Figure 4-4: Quality coefficient *Q* for varying fragment lengths (averaged over section lengths between 10 and 30 fragments); and varying section lengths (averaged over fragment lengths between 5 and 15 bins) used to classify five groups of species (bottlenose dolphins, common dolphins, common/striped dolphins, pilot whales and striped dolphins) from both the French and Spanish datasets.

4.4.5. Classifier Training

4.4.5.a French classifiers

The *2Sp* French classifier classified the CSD detections better than by chance ($p<0.001$) with a correct classification probability of 65% (Table 4-5), however the pilot whale detections were classified at the same rate as if it was at random ($p=0.89$) (Table 4-5).

With the *4Sp* French classifier correct classification probabilities were low for all species with a maximum correct classification probability of 51% for the C&S group. Adding to this low correct classification probability the false positive misclassification probabilities were high for all species with a minimum of 55% for the pilot whale group. The STD detections were not classified better than by chance alone ($p=0.004$)

Forty five per cent, 37% and 66% of the COD, C&S and STD classification groups respectively were misclassified as one of these groups.

Table 4-5: *2Sp* French classifier with the classification probabilities when the classifier was trained with 80% of the French dataset. Standard deviations (%) are within the brackets. Species codes are the same as in table 1-4, with CSD = COD +STD + C&S pooled. P = p-value of a one-tailed t-test to test, the null hypothesis that the results were obtained purely by chance, $H_0 = 50\%$

Classified as %			False Positives (%)	p
	CSD	FPW		
CSD	**64.9 (*11.8*)**	50.3	**43.7**	**<0.001**
FPW	35.0	**49.7(*19.7*)**	**41.3**	**0.89**

Table 4-6: *4Sp* French classifier confusion matrix: Classification probabilities of the classifiers trained with 80% proportion of the French dataset. Standard deviations are within the brackets. p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H_0 = 25\%$

Classified as %	True Species				False Positives (%)	p
	COD	C&S	FPW	STD		
COD	**34.0 (*13.2*)**	8.8	23.2	34.6	**66.2**	**<0.001**
C&S	18.7	**51.3 (*13.3*)**	23.0	30.9	**58.6**	**<0.001**
FPW	21.4	11.7	37.5 (*13.3*)	13.9	**55.6**	**<0.001**
STD	25.9	28.2	**16.5**	**20.6 (*13.3*)**	**77.4**	**0.004**
False negatives	**66**	**48.7**	**62.5**	**79.6**		

4.4.5.b Spanish classifiers

With the *3Sp* Spanish classifier, both the BND and CSD classification groups were very well identified, with a correct classification probability greater than 90%. The false positive misclassification of BND was small (14%) whereas it reached 49% for the CSD classification group. This high rate was directly linked to the high misclassification of FPW detection as CSD (83%) consequently FPW detections were poorly classified with a correct classification probability of 6% different ($p<0.001$) and lower than by chance alone.

When the CSD classification group was divided in three groups (COD,C&S and STD) in the ***5Sp*** Spanish classifier, BND whistles contours were still very well discriminated (92%

correct classification probability) with a relatively low false positive misclassification probability of 20%. COD and STD had a probability of being correctly classified close to 40% with high false positive misclassification probabilities (70%). This high rate was largely due to the fact that 75% of both the FPW and C&S detections were classified as COD or STD (Table 4-8). Both the C&S and FPW detections were not classified better than by chance.

The confusion matrix of this Spanish classifiers differed significantly to the ***4Sp*** French confusion matrix for three (C&S, FPW, STD) of the four species in common.

Table 4-7: *3Sp* Spanish classifier confusion matrix with the classification probabilities when the classifier was trained with 70% of the Spanish dataset. Standard deviations (%) are within the brackets. Species codes are the same as in table 1-4, with CSD = COD +STD + C&S pooled. P = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H_0 = 33\%$

Classified as %				**False Positives (%)**	**p**
	BND	**CSD**	**FPW**		
BND	**91.5** *(13.6)*	2.9	11.5	**13.6**	**<0.001**
CSD	5.8	**91.1(***13.4***)**	83.0	**49.4**	**<0.001**
FPW	2.8	5.9	**5.5(***13.1***)**	**61.3**	**<0.001**
False Negatives	**8.5**	**8.9**	**94.5**		

Table 4-8: *5Sp* Spanish confusion matrix with the classifiers trained with 70%n of the Spanish dataset. Standard deviations are within the brackets. P = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H_0 = 20\%$.

Classified as %	True Species					False Positives (%)	p
	BND	COD	C&S	FPW	STD		
BND	91.5(11.4)	3.1	1.7	12.0	4.4	18.9	<0.001
COD	1.0	40.8 (12.4)	39.1	44.0	24.3	72.7	<0.001
C&S	1.2	23.1	17.0 (11.9)	7.0	26.2	77.2	0.004
FPW	3.8	5.0	6.3	6.0(11.0)	4.8	76.8	<0.001*
STD	2.5	28.0	35.9	31.0	40.3(12.5)	70.7	<0.001
False Negatives	8.5	59.2	83	94	59.7		

4.4.5.c North Atlantic classifier

From Gillespie et al., (2013), the optimal fragments and section length measured with their data were respectively 30 bins (160ms) and 60 fragments per section. The classification probabilities were high with the correct classification probabilities, greater than by chance for the three classification groups, ranging from 70% to 89% and a low false positive rates ranging between 4% to 28%.

Table 4-9: North Atlantic classifier confusion matrix with the classifiers trained with 80% of the dataset. Standard deviations are within the brackets. p = p-value of a one-tailed t-test to test the null hypothesis that the results were obtained purely by chance, $H_0 = 33\%$.

Classified as %				False Positives (%)	p
	BND	CSD	FPW		
BND	70.3 (10.8)	10.7	8.0	21	<0.001
CSD	26.9	88.8 (10.3)	5.7	27.5	<0.001
FPW	2.8	0.5	86.4 (10.2)	3.7	<0.001
False Negatives	29.7	11.2	13.7		

4.4.6. Classification of new data with the classifiers

4.4.6.a French dataset

4.4.6.a.i Classified with Spanish and North Atlantic classifiers

With the *North Atlantic* classifier, fewer sections were created and classified as the fragment and section length parameters were longer than in the French and Spanish classifiers.

As expected from the *5Sp* Spanish classifier confusion matrix both common and striped dolphin sections were correctly identified at a rate close to 40% (40% for COD and 32%. for STD) (Table 4-10). C&S dolphin sections were classified at a rate very different to the expected correct classification probability of the Spanish confusion matrix; 67% versus the 17% expected. The pilot whale sections were very poorly identified with the Spanish classifier with only 3% of the sections correctly identified whereas they were expected to be correctly identified at 38% with the *4Sp* French classifier.

Once classified with the *3Sp* Spanish classifier (Table 4-10), the identification of the CSD sections were much higher than the prediction with the French classifier and in the same order as the predictions of the Spanish *3Sp* classifier. The proportion of pilot whales detections correctly identified was better than expected from the confusion matrix of the Spanish classifier.

Finally classification of these data with the *North Atlantic* classifier (Table 4-10) gave on average a better correct classification probabilities. The main improvement was seen in the classification of FPW detections, but the proportion of FPW section correctly identified was still low at 34%.

Organising the sections into encounters reduced the amount of data available for classification. The 876 sections classified formed 16 encounters (Table 4-11) with a maximum of 403 sections per encounter for the common dolphin classification group (see Appendix B, Table B.1). Nevertheless the proportions of encounters correctly classified were slightly better than when the results were compared by sections. COD and C&S classification groups had a better classification probability when they were organised into encounters with up to 100% of correct classification when the *3Sp* classifier was used. However the four encounters of STD detections were never correctly identified with the *5Sp* classifier and they were misclassified as BND, COD or C&S. With the *3Sp* classifier three of them (75%) were correctly identified. Even if this classification probability was lower than the 91% observed

in the confusion matrix (Table 4-7), only one encounter out of a total of four was misclassified. However, having the results organised into encounters did not improved the identification of the French FPW sections, with both Spanish classifiers all FPW encounters were misidentified.

Table 4-10: Classification result of the French dataset classified with the Spanish and North Atlantic classifiers.

		True Species			
Classifier	**Classified as %**	**COD**	**C&S**	**FPW**	**STD**
5Sp	**BND**	2	0	30	3
	COD	40	27	23	30
	C&S	31	67	23	32
	FPW	2	0	3	3
	STD	25	7	22	32
2Sp	**BND**	3	0	28	5
	CSD	95	100	67	95
	FPW	2	0	5	0
North Atlantic	**BND**	13	0	5	0
	CSD	87	100	61	100
	FPW	0	0	34	0

Because of the bigger fragment and section lengths used in the *North Atlantic* classifier the French data were organised into only 11 encounters (Table 4-12, Appendix B, Table B.3). Only one of these encounters was misidentified (FPW encounter misclassified as common/striped).

Table 4-11: Encounters classification of the French acoustic dataset classified with the Spanish classifier.

		True Species			
	Number of Encounters Classified as	**COD**	**C&S**	**FPW**	**STD**
5Sp	**BND**	0	0	2	1
	COD	4	0	0	1
	C&S	1	1	1	2
	FPW	1	0	0	
	STD	2	0	0	
3Sp	**BD**	0	0	2	1
	CSD	8	1	1	3
	PW	0	0	0	0
	TOTAL	**8**	**1**	**3**	**4**
	Mean (*sd*) number of sections/encounters	74(*132*)	15(*0*)	78(*61*)	9(*9*)

Table 4-12: Encounters classification of the French acoustic dataset classified with the North Atlantic classifier (*N.Atlantic*)

		COD	**C&S**	**FPW**	**STD**
N. Atlantic	**BND**	0	0	0	0
	CSD	5	1	1	2
	FPW	0	0	2	0
	Number of Encounters	5	1	3	2
	Mean (sd) of number of sections/encounters	27.2(*33.0*)	2(*0*)	14.67(*13.6*)	3(*1.41*)

4.4.6.a.ii Unidentified dataset classified with French classifier

Fifty recordings, out of a total of the 68 selected recordings without visual detections, contained identified sections. These sections were grouped into 25 encounters with the French classifier and 8 with the *North Atlantic* classifier (Table 4-13, see Appendix B, Table B.3, for details). With the ***4Sp*** French classifier 5 (20%) of the encounters were classified as STD, 11 (44%) as COD, one (4%) as C&S and seven (28%) as FPW. One encounter had an equal probability to be classified as COD or STD. With the *2Sp* French classifier, 17 (68%) encounters were classified as CSD dolphins, 71% of them were in common with the *4Sp French* classifier. However only 3 of the 7 encounters classified as FPW were classified by the *4Sp French classifier* as FPW as well. Five encounters (62%) generated by the *North Atlantic classifier* were identified as CSD like they were with the ***2Sp*** French classifier.

These encounters were compared with the effort of the visual teams (see Appendix B, B.3). For 64% of these encounters the visual team was off effort. The remaining encounters were detected when the visual team was on effort. For four of these encounters no sightings of a whistling species was detected within 10 minutes of the whistles detections, for the remaining five, a visual detection of a whistling species was recorded.

Table 4-13: Encounter identification of the French acoustic detection, not associated with a visual detection, using the French and North Atlantic classifiers. COD=Common dolphin, C&S=Common and Striped dolphin, FPW=Pilot whale, STD=Striped dolphin, unidentified=when a section contain the same maximum classification probabilities between several species.

		French Classifiers	North Atlantic classifier
4Sp	**COD**	11	
	C&S	1	
	FPW	7	
	STD	5	
	Unidentified	1	
2Sp	**BD**		1
	CS	17	7
	PW	7	
	Unidentified	1	
TOTAL		**25**	**8**

4.4.6.b Spanish dataset

4.4.6.b.i Training dataset classified with French and North Atlantic classifiers

The Spanish data classified with the ***4Sp*** French classifier were on average correctly identified with almost the same probability as expected with the Spanish classifier itself (Table 4-14, and Appendix B, B2 for details). STD were slightly better identified with the French classifier (53.6% versus 40.3 with the 5Sp Spanish classifier) while COD detections were better identified with the Spanish classifier (40.8%) than with the French one (32.4%). C&S detections were largely misidentified as STD, the expected correct classification probability of this group with the *4Sp French* classifier was 51.3% and only 16.4 % of the Spanish C&S detections have been correctly identified. The Spanish data contained only one section of FPW detections which was misclassified as COD (Appendix B, B2).

The classification results with the *2Sp French* classifier (Table 4-14, and Appendix B, Tbale B3) were worse than what was expected from the confusion matrix. None of the three classification groups were correctly identified at a rate greater than 50%. The BND sections were not classified as the French classifiers did not have a bottlenose classification group (Table 4-4).

However, the *North Atlantic* classifier identified the COD, C&S and STD sections correctly more than 90% of the time (Table 4-14). But the only section of BND was misidentified as FPW and there were not enough sections of FPW to generate at least one section.

Table 4-14: Spanish data classified with the *3Sp, 2Sp* French classifiers and by the *North Atlantic* classifiers. CD=Common dolphin, CS=Common and Striped dolphin, PW=Pilot whale, SD=Striped dolphin.

		True Species				
Classifier	**Classified as %**	**BND**	**COD**	**C&S**	**FPW**	**STD**
4Sp	**COD**	0.0	32.4	24.7	100	28.0
	C&S	0.0	15.2	16.4	0.0	11.9
	FPW	0.0	15.4	13.2	0.0	6.5
	STD	0.0	37.0	46.1	0.0	53.6
2Sp	**CSD**	0.0	49	39	100	38
	FPW	0.0	51	61	0	62
North Atlantic	**BOD**	0.0	10	10	0.0	0.0
	CSD	0.0	90	90	0.0	100
	FPW	100	0.0	0.0	0.0	0.0

The 4602 sections of the Spanish data formed 48 encounters; 18 COD, 19 C&S, 1 FPW and 10 STD (Table 4-15, and Appendix B, Table B.2 for details). These encounters were made of 1 to 448 sections. Encounters of COD contained more sections (average of 216 sections per encounters), whereas the encounter of FPW was made of only one section. COD encounters were mostly (>50%) correctly classified by both classifiers. Encounters of other species were

largely misclassified as STD or COD when classified with the ***4Sp*** classifier and as FPW when classified with the French classifier. Fifty six per cent (56%) of the encounters were identified as the same species by the two classifiers.

Table 4-15: Encounters of the Spanish data classified with the *3Sp, 2Sp* French classifiers.

	Number of encounters classified as	**COD**	**C&S**	**FPW**	**STD**
4Sp	**COD**	10	4	1	4
	C&S	0	0	0	0
	FPW	1	2	0	1
	STD	5	12	0	5
	Unidentified	2	1	0	0
2Sp	**CSD**	12	6	1	4
	FPW	5	12	0	5
	Unidentified	1	1	0	1
	Total number of encounters	18	19	1	10
	Mean (*sd*) number of sections/encounters	216(*373*)	28(*62*)	1(0)	17(*23*)

With the *North Atlantic* classifier 97% of the 36 encounters were correctly classified, only the BND encounter, made up of a unique section, was misclassified as FPW (Table 4-16).

Table 4-16: Encounters of the Spanish data classified with the *North Atlantic* classifier

		BND	COD	C&S	FPW	STD
	Number of Encounters Classified as	1	17	13	0	5
	Mean (sd) of number of sections/encounters	1	42(74.12)	7(13.6)	0	6(5.5)
N.Atlantic	**BD**		0	0	0	0
	CS		17	13	0	5
	PW	1	0	0	0	0

4.4.6.b.ii Unidentified dataset classified with Spanish classifier

These detections were grouped into 62 encounters made of 1 to 370 sections (Appendix B. Table B.4 for details). The classification groups containing COD had a much higher number of sections per encounter (Table 4-17). Only one encounter of FPW was identified with only 1 section in it.

When this unknown dataset was classified with the *5Sp* classifier, 51 encounters were classified as COD, C&S and STD. With the *3Sp* classifier 54 were identified as CSD and only two of them were identified differently with the *5Sp* classifier. With the 5Sp classifier eight encounters were identified as BND seven of them were identified similarly with the *3Sp* classifier. The remaining BND encounter was classified as CSD with a probability just over the average (52%). The encounter identified as FPW with the *5Sp* classifier was classified as CSD with the *3Sp* classifier. With the *North Atlantic* classifier, 37 encounters were generated, with 32 of them being identified as CSD a similar number to the *3Sp* French classifier. Two were identified as FPW and one as BND.

The spectrograms of all the encounters were examined visually to detect any false positive detections (Appendix B, Table B.4). Eleven encounters (18%) were false positives detections, three were due to the presence of a sonar, producing a discontinuous long signal in the frequency range of the species of interest, and eight contained numerous electric noises.

Twenty seven (43%) encounters were detected when the visual team was off effort. For 19 encounters (31%) the visual team was on effort but did not detect any animal and for the

remaining five encounters (8%) a visual detection happen 10 minutes before or after the encounter time. Each of these encounter have been identified with the same species as the visual observations (Appendix B, Table B.4).

Table 4-17: Classification result of the Spanish acoustic dataset classified using the Spanish and North Atlantic classifiers. Number of sections classified with the corresponding classification probability (%).CD=Common dolphin, CS=Common and Striped dolphin, PW=Pilot whale, SD=Striped dolphin. The number in bracket is the number of sections classified similarly by the Spanish and the North Atlantic classifiers.

		Spanish Classifier	North Atlantic Classifier
5Sp	**BD**	8	
	CD	26	
	CS	11	
	PW	1	
	SD	14	
	Unidentified	2	
3Sp	**BD**	7	1
	CS	54	33 (32)
	PW	0	2
	Unidentified	1	1
TOTAL		**62**	**37**

4.5. Discussion

4.5.1. Parameters influencing the performance of a classifier

In this chapter, two sets of classifiers have been created using data collected during the CODA large scale survey. One of the main objective of the CODA project was to estimate the common dolphin abundance, and if possible the abundance of other cetaceans species. Visually Common and striped dolphins are hard to tell apart at large distances or are often found in mixed groups, hence the need for a C&S group. With the acoustic classifier the same result was achieved, indeed when these two species were pooled in one group, they were very well discriminated from bottlenose dolphin and pilot whale. However, like with visual detection, it was more challenging to tell apart acoustically common to striped dolphin. The results of this chapter were in accordance with previously published results (Gillespie et. al., 2013, Oswald, 2007) showing important misclassification between common dolphin and striped dolphins. The misclassification between these two species observed in those papers was smaller than the one observed in this thesis. This difference may be explained by the smaller size and the less accurate species identification of the training data which generated a new source of misclassification due to misclassification within the training dataset itself

4.5.2. Consequences of a lack of training data

The very large proportion of false positive detections (almost 80% between both datasets) reduced the amount of data available to build stable and reliable classifiers. The differences observed in the pilot whale correct classification probabilities between the French and Spanish classifiers (40% and only 5% respectively) may be explained by the lack of data in the Spanish dataset. Only 17 pilot whale contours were available to train the Spanish classifiers whereas 842 were available in the training data of the French classifiers. With this quantity of data it was difficult to create a stable classifier. This low ability to identify pilot whale detections explained the poor classification of the French pilot whale detections by the Spanish classifiers (maximum correct classification probability being 5%). Even so the French training dataset contained more pilot whale contours, the identification of pilot whale by the *2Sp French* classifier was not better than by chance. However, in the *North Atlantic classifier*, for which 20 times more data were used for this species, the expected correct classification probability reached 86%. So the classification probability observed for the *2Sp* French classifier may as well be due to a lack of data. Even if the French and Spanish pilot whale data classified with the *North Atlantic classifier* were not correctly identified with such

success (34%), of the five classifiers tested it gave the most accurate classification. This smaller correct classification probability relative to the expected one may be due to another source of misclassification; the difference of vocalisation characteristics between population or strongly related species. Indeed, the pilot whale detections of the French, Spanish and North Atlantic training data were a combination of both the long finned and short finned pilot whale species. Gillespie et al., (2013) shown that these two species can be discriminated well with a small misclassification probability between them (between 6 to 12%). In this chapter, the French pilot whale detections were in majority from long finned pilot whale whereas the majority of the pilot whale detections of the North Atlantic classifier were from short finned pilot whales (Gillespie et al., 2013).

4.5.3. Consequences of a lack of the accuracy of species identification

In this analysis, the species identification was done by associating a species sighted several hundred meters in front of the acoustic detection system. Given this distance, the method chosen to link the visual to the acoustic detection could be at the origin of a wrong species identification. The low classification probability of the C&S group within the *5Sp* Spanish classifier, despite being the group with the second largest number of contours, can be explained by the selection process for this group. During the selection only adjacent recordings with one species were selected. However, an exception was made for COD, STD and C&S groups (see section 3.1.b.ii for explanation). The Spanish CSD data contained more recordings associated with COD and STD sightings than the French CSD data. Maybe some assumptions were wrong and some significant differences were to be expected between these three groups. Each time the large CSD classification group was involved in the classification (either to train the classifier or when it is classified with the French classifier) the classification probability was low.

The difference between the correct classification probabilities for the STD group between the French and Spanish classifiers was less expected. In the French classifier the striped dolphin detections were classified randomly with an even misclassification between COD and C&S which means that the classification of the Spanish striped dolphin by the *4Sp* French classifier were similar to a random classification and so cannot be considered as reliable. When this source of mistake was removed by pooling the three classification groups as a unique one the confusion matrix for these classifiers showed an average correct classification probability of 60% versus the 38% when the three groups were run independently.

Finally the *North Atlantic classifier* which combined good large training dataset and reliable species identification for the training data and also one common group for the common and striped detections gave the best classification results for all the datasets.

4.5.4. Advantages of acoustic detections over visual detections

This chapter showed how difficult it is, using the current methodology to make a reliable link between visual and acoustic detections with survey method similar to the one used in the CODA survey. The main difficulty highlighted was the ability to create a large enough acoustic dataset despite weeks of acoustic survey. The presence of non-dolphin noise with the disproportionately high number of false positive detections, stems from the use of noisy survey platforms. Although it was possible to remove most false positive detections from the dataset, those that could not be removed shared characteristics similar to those of the whistles that were attempting to be classified.

The distance between the visual team and the acoustic detection was also a problem as it increased the chance of mistakes in the association of sightings to an acoustic detection. There is no doubt that acoustic detections are important to detect animals missed-detected visually (Table 4-18); the 61% of the classified encounters not associated to a visual detection proved it. Half of the encounters (50%) not associated with a visual detection were made when the observers were off effort. For the other half either no sighting was recorded or a sighting was recorded within 10 minutes but it was not selected due to the selection criteria used in the method of this analysis. However, with the current survey method and with the method used in this analysis, the classification results of the encounters not associated with a visual detection were not very reliable.

With some modification in the survey method it would be possible to create a training dataset of quality from the survey data themselves. Similarly to the selection of well trained and highly qualified visual observer to optimise the chance of getting reliable identifications, a good classifier training data will optimise the chance of getting reliable acoustic identification. The better classification results with the *North Atlantic classifier* illustrated this point. But given the observations of different whistles characteristics between populations, an ideal training dataset should be made of data collected in the same area of the

survey, which is not always possible. Nonetheless some modification of the survey actual method similar to what is regularly done by the Southwest Fisheries Science centre in the United State (Barlow and Forney, 2007) would be beneficial for both the visual and the acoustic detections. They are using a closing mode procedure during the survey that consisted of breaking the transect line to go closer to sightings. They can then, verify their species identification and group size estimates. This verification can be used to measure their probability of correct identification from the transect line. This method will be also beneficial for the acoustic detections as it will generate some recordings directly associated with a species. These recordings can then be used to train the classifier or to verify the quality of the classifier.

Table 4-18: Summary of the numbers of encounter classified per species for the training dataset and the number of encounters classified by the classifier but for which the species identification was not known.

	n visual detections	***n* encounters**	**Off Effort**	**On effort**	
				No visual detec.	**Visual detec.**
BND	29	0			
COD	156	26			
C&S	75	20			
FPW	43	4			
STD	49	14			
RSD	1	0			
No Id		87	43	28	11

If the closing mode is not possible then allowing for some visual detection when the animals pass abeam of the hydrophones and monitor the distance from the hydrophones should help also in the creation of a more reliable acoustic database of sound from which the species is identified.

In conclusion this chapter highlighted the need to improve and/or generalise the used of methods such as close up mode to accumulate a reliable training dataset to be used during the creation of an acoustic classifier. By developing more cost-effective methods to select a good

training dataset, it can be hoped that the part of the misclassification generated by misclassification within the training dataset will be diminished and only the misclassification generated by similarity between species will stay. Furthermore this misclassification can be decreased by grouping similar species as a unique classification group

Chapter 5: Classification: General Discussion

5.1. Introduction

One of the difficulties with acoustic data from cetaceans is correct identification of the detected sounds. Several classifiers (Datta and Sturtivant, 2002; Gillespie et al., 2013, 2011; Nanayakkara et al., 2007; Oswald et al., 2003; Soldevilla et al., 2008) have been developed based on different methods. With the exception of Gillespie et al., (2013), the confusion matrix is the only quantitative description for these classifiers. Gillespie et al. (2013) presented for the first time a measure of uncertainty of the confusion matrix by associating the standard deviation of the correct classification probabilities in the confusion matrix.

The objective of the first part of this thesis was not to develop yet another whistle classifier method but to determine the factors influencing the quality of acoustic classifiers. The analysis in these chapters were based on two case studies of surveys, which were organised to get more information about the distribution of cetacean species for conservation and management decisions. The prime objective of the first case study (MORL_BOWL, chapter 3) was to develop a classifier to discriminate bottlenose dolphin from other species present in the area of interest. This case study is similar to the numerous papers describing classifiers, in the respect that data to train the classifier have been carefully selected to optimise the classification result. The main objective of the second case study (CODA, chapter 4) was to detect the presence of cetacean species with particular focus on common dolphin species. For this case study the main analysis was done by visual observations but an acoustic detection system was added in the process to complement the visual survey. Classifiers were developed with the data collected during the survey itself.

5.2. Parameters influencing on the quality of the classifier

These two chapters demonstrated the importance of the quality of the dataset used to train classifiers. The quality of the training dataset was defined by the amount of data available to train the classifier, the reliability of the identification of species and the presence of false positive detections in the data.

5.2.1. Size of the training dataset

Both, the MORL-BOWL (chapter 3) and the North Atlantic (chapter 4) classifier were trained with data collected from quiet platforms with accurate visual confirmation of the recorded species. The average number of contours per species in the training dataset was 29222 for the MORL-BOWL classifier and 37000 for the North Atlantic classifier. Classifiers trained with these data were able to identify on average 83% (sd=12%) of the detections correctly. However, the French and Spanish classifiers from the CODA data were trained with data containing an average of 2680 contours per species and the species identification relied on a less accurate method than for the previous dataset. For these classifiers the average correct classification probability was 46% (sd=30%).

5.2.1. Reliability of the visual observation

An important assumption made in chapter 3 was the high confidence in the species visual identification of the acoustic detections used in the training dataset. At the opposite in chapter 4 it was suggested that one of the reason of the poor classification result was perhaps due to errors in the identification of the recordings occurring close to a visual detection. These possible wrong associations between visual and acoustic detections generated the creation of training datasets less reliable than for the previous chapter. This point highlights the problem of the accuracy of visual detection and its consequence during the classification process necessary to use most of the cetacean acoustic data. During the selection process of the training data in chapter 3, some of the initial data available from the west coast of Scotland were discarded due to misidentification of the species by the visual observer. While initially these data were included in the training dataset the output of the classifier was not good and raised suspicion. After a direct observation of these recording spectrograms it was clear that these acoustic detections were not from the species identified by the visual observer. They were then discarded form the training dataset and the classification result was largely improved. It was not possible to tell if the mistake come from a misidentification from the observer or from an error during the data transcription on the database.

This example illustrates perfectly a major problem encountered with cetacean visual detection which is the reliability and the lack of method to detect these misidentifications. At the opposite of acoustic detections, visual detections are most of the time not recorded such that a double verification is possible after the survey and so as soon as the animal is not visible any more the only information available is the species identification recorded by the

observer associate with its level of confidence. The method of double observer widely used in visual survey can help in dealing with this issue if the animals are detected by both observers. For terrestrial survey and mainly avian and anuran species the problem of misidentification, generating false positive detections, and its negatives consequences in abundance estimation is now recognise and analytical method are developed to correct it (McClintock et al., 2010a, 2010b).

5.2.2. Characteristics of the classification groups

The accuracy of the classifiers was dependent on the number and characteristics of the species groups used for classification. The high misclassification probabilities of the *4Sp French* and *5Sp Spanish* classifiers were explained by the similarity between common dolphins and striped dolphins. When these species were pooled the average correct classification probability of the classifiers increased. In contrast, the good result of the MORL-BOWL classifier was partly due to the large size of the dataset and partly due to two species, white side and white beaked dolphins, which are relatively easy to tell apart (Gillespie et al., 2013) from the other classification groups. Increasing the number of classification groups in the MORL-BOWL dataset slightly decreased the correct classification probability of the bottlenose dolphins (chapter 3).

5.2.3. False positive detections

The presence of false positive detections in the training data can be responsible for a bad classification result. Frequently in underwater acoustic surveys, there are numerous sources of noise with similar characteristics as the sound of interest. A high amount of broadband, short noises such as shrimp clicks, very short electric noises and echo sounders can easily been missed - detected as cetacean clicks. Other noises such as sonars, more persistent electric noises, and rubbing noises from mooring, can easily produce sounds with the same frequency range and length as whistles. Being able to develop a perfect detector that recognises all these natural and anthropogenic sources of noise will never be possible. However, a false positive analysis on the training data prior to training the classifier can reduce the impact of such false detections. Because the CODA data were known to be collected from noisy ships, a false positive detection analysis was conducted before training the classifier. Such an analysis permitted to identify a single parameter (contour length) which led to the removal of 80% of all the detections (of electric noise in the majority of these cases) leaving 20% good ones. It seems obvious that if these contours were used for the

classification, the classification results would have been worse despite being based on a larger dataset. This false positive detection analysis prior to training the classifier was not conducted on MORL-BOWL dataset because this dataset came from quiet platforms where operators were being more careful about the quality of the data that they were recording and hence the amount of noise was negligible.

5.3. Defining the robustness of a classifier

In addition to comparing the classification probabilities as a function of the quality of the training dataset, chapter 2 highlights the notion of uncertainty within the confusion matrix of a classifier. As explained in chapter 2 and in Caillat et al., (2013), whistles are highly variable within and between species, and hence the probability of obtaining exactly the same confusion matrix from two different samples of training data is very low. Due to this high variability, a classifier should be presented with a measure of uncertainty for each classification probability. In chapter 2 a method was proposed to measure and predict the variability of the classifier, along with a discussion of the limits of this method. However, by drawing a parallel between the measured variability of the classifiers in chapters 3 and 4, it can once again be seen that classifiers with a good training dataset contain less uncertainty: the average coefficient of variation of the correct classification probabilities for the MORL_BOWL and North Atlantic classifier was 10% while it was between 40% and 180% for some correct classification probabilities of the French and Spanish classifiers. This result will be essential for further analysis (Part II).

5.4. Specificity of the PAMGUARD Whistle Classifier

All the classifiers generated in this thesis were based on the automatic PAMGUARD Whistle Classifier and the whistle contours were detected by the automatic PAMGUARD Whistle and Moan detector. A disadvantage of such an automatic classifier is that it is more likely to include contours of false positive detections in the classification process than a classifier that is based on the selection of the whistles contours by an operator. A specific feature of the PAMGUARD Whistle classifier is the division of whistle contours into smaller parts. One can argue that by doing so, information on the overall shape of the contours is ignored and that classical parameters used to discriminate species, such as end frequency, start frequency, number of inflections, cannot be used. Consequently, important characteristics of the whistles are not taken into account. However, the good result of the identification of the EARs data in

chapter 3 showed that, with a good training data, this classifier correctly identifies unknown data. The advantage of automated classification on the other hand is the ability to process quickly a large amount of data from detection to classification and it is probably more consistent than a human operator.

To further improve our knowledge on the potential of this classifier it would be interesting to study its quality for a small, but accurate dataset and to analyse which of the nine parameters used in the discriminate function are the most useful for the classification.

5.5. Recommendations of creating a good whistle classifier

In conclusion, developing a classifier is a task which requires a training dataset of high quality to obtain accurate and good classification probabilities. In the case of marine mammals, it is often difficult and time consuming to get large training datasets. However, it is still possible to improve the methodology and to be rigorous in the collection of small datasets to assure their quality. For acoustic data, it is now relatively easy to collect a large amount of them (due to increase in computer storage capacity and improvement of technology), but if the operators are not careful about assuring the quality of the recordings, large datasets quickly become useless.

My recommendations to develop a classifier would be to:

1. ensure correct identification of species within the data used to train the classifier (see suggestion of methods in 4.5.4);
2. ensure that the training data does not contain false positive detections;
3. ensure that there is enough data for each species for the classifier to be reliable;
4. be careful about the selection of species to classify, in particular by selecting only species which are present in the area of interest. This avoids having too many species in the classifier which increase the probability of misclassification;
5. run a false positive detection analysis on a subsample of the data after classification;
6. measure the variability of the classification probabilities that is due to the sampling process;

Nevertheless, given the high variability of the whistles even with a perfect protocol, a very good data set and quiet acoustic system, the chance to create a classifier able to discriminate each species without misclassification is not possible. It is then important to find some methods which from the observed classification result calculate the true number of acoustic

detections for each species. The second part of this thesis demonstrates one method to solve this problem.

Part II. Misclassification

Chapter 6: A heuristic method to estimate the number of acoustic detections in the presence of species misclassification.[5]

6.1. Introduction

Over the last two decades, researchers and managers have become increasingly aware of the advantages of using passive acoustic monitoring over visual cues to detect marine mammals and so to potentially estimate their abundance. Many studies, in particular those processing large datasets from long-term fixed hydrophone deployments, rely on automatic detectors and species classifiers to decrease the time and cost of analysis. In the previous part of this thesis, it was demonstrated the importance of a good quality dataset to develop a reliable whistle classifier. It was also admitted than it will never be possible to develop a classifier able to discriminate species perfectly; hence there will always remain misclassification between species. However, in any management strategy, accurate and precise quantification of population size ("abundance") is crucial to develop appropriate management actions.

A standard method for estimating abundance based on acoustic detections is cue counting, where the cues are the vocalisations detected (Marques et al., 2011, 2009 and chapter 1.2.2.b.ii p9). The general formula to estimate a species' abundance from cues is given by

$$\widehat{N} = \frac{n(1-\hat{c})A}{aT\hat{P}\hat{r}} \qquad (6\text{-}1)$$

where n is the number of detected cues, $\hat{c}$ is the estimated proportion of false positives detected (calls classified as the species of interest which originated from other species or other sources of noise), a is the area in which cues can be detected, $\hat{P}$ is the estimated average probability of a cue being detected within this area during recording time T, $\hat{r}$ is the estimated cue production rate and A is the total study area (Marques et al., 2009). Apart from the fact that this formula requires knowledge of the cue production (i.e., vocalization) rate, which is unknown for many species, the abundance estimate in Eq. 6-1 only considers the presence of one species at a time in the area of interest.

[5] A slightly modified version of this chapter has been accepted in J. Acoust. Soc. Am. : Caillat, M., Thomas, L., Gillespie, D. (2013) The effects of acoustic misclassification on cetacean species abundance estimation.

In this part of the thesis (chapter 6 to 8), only issues on determining the true number of calls $\hat{v}$, which is equivalent to $\hat{v} = n(1-\hat{c})$ of Eq. 6-1, are addressed. Marques et al. (2009) estimated the proportion of false positive detections, $\hat{c}$, by visually examining 30 periods of 10 minutes from 6 days of recordings, a process which relied heavily on a human operator being able to distinguish between the sounds of interest and a range of other sound sources.

If the main source of false positive detections is the presence of other species with similar vocalisations in the study area, then the rate of false positive detections will be strongly related to the relative call densities from the different species. For example, if it is known that species A and B are often confused by the classifier, and that species B is much more common or more vocal than species A, then a high percentage of the detections attributed by the classifier to species A will in fact be false positive detections resulting from the presence of species B. If on the other hand, species B were extremely rare or very silent, then there would be few misclassifications assigned to species A from species B.

Since the interest is in estimating the density of multiple species within a given study area, it becomes necessary to replace the $(1-\hat{c})$ term with the more general equation

$$\hat{\boldsymbol{v}} = M(\boldsymbol{n}) \qquad (6\text{-}2)$$

Where $\hat{\boldsymbol{v}}$ and $\boldsymbol{n}$ are now vectors representing the true numbers of calls and the numbers of calls counted for each species after misclassification respectively, and M is a more general misclassification operator.

As described in the previous chapters, the level of misclassification between species can be described in terms of a confusion matrix formula 6-3 (e.g. chapters 3 and 4, Gillespie et al., 2013; Oswald et al., 2007), which summarises the probabilities for correct, false positive and false negative classifications of all species considered (Part I.1.2.2.b.iv, *p*11).

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{p_{11}} & \cdots & \boldsymbol{p_{1j}} & \cdots & \boldsymbol{p_{1m}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{p_{i1}} & \cdots & \boldsymbol{p_{ij}} & \cdots & \boldsymbol{p_{im}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{p_{m1}} & \cdots & \boldsymbol{p_{mj}} & \cdots & \boldsymbol{p_{mm}} \end{pmatrix}, \qquad (6\text{-}3)$$

where $\sum_j p_{ij} = 1 \; \forall 1 \leq j \leq m$.

The expected number of detected calls $E(\boldsymbol{n})$ for each species following misclassification is therefore given by

$$E(\boldsymbol{n}) = C.\boldsymbol{v}$$

and it follows that the true number of detections for each species can be estimated using

$$\widehat{\boldsymbol{v}} = C^{-1}.\boldsymbol{n} \qquad (6\text{-}4)$$

where C^{-1} is the inverse of the confusion matrix C.

Species classification is a stochastic process where each classification may be considered as an independent random event. In addition, it cannot be assumed that the confusion matrix is known precisely since it is typically derived from a finite sample of real data (chapter 2). Gillespie et al. (2013) and chapters 3 and 4 showed uncertainties, expressed as a measure of standard deviation, ranging from 0.04 to 0.48 for the probabilities of a typical confusion matrix. The stochastic nature of the classification process combined with the imperfect knowledge of the confusion matrix add to the uncertainty of any estimate of the true number of detected cues ($\widehat{\boldsymbol{v}}$) and consequently, to the uncertainty of estimated species abundance if misclassification is taken into account.

With this in mind, this chapter examines the bias and precision of the estimates of the true number of detected calls from multiple species which arise from the stochastic nature of the confusion process, as well as the uncertainty within the confusion matrix. This is achieved by looking at hypothetical confusion matrices and simulated data. After a brief description of the classification process in mathematical terms, which also serves as an introduction of notation, a simple model containing only the stochasticity within the classification process is analysed. This analysis is then extended by incorporating uncertainty in the rates of misclassification.

6.2. The classification process

Classification events are assumed independent of each other. Thus the classification for each species j can be described as the outcome of a multinomial process, where the vector of probabilities of the corresponding multinomial distribution is given by the probabilities of the j^{th} column of the confusion matrix. The numbers of trials in these multinomial distributions

are the true number of detections $\boldsymbol{v}$, i.e., v_j is the number of trials, or the true number of detections for species *j*.

The expected observed number of vocalisations of species *i* (n_i) is equal to the number of vocalisations of species *i* correctly classified as species *i* plus the false positive classifications when vocalisations of another species $j \neq i$ are misclassified as species *i*.

$$E[n_i] = \overbrace{p_{i=j} v_i}^{\text{Correct Classified}} + \sum_{j \neq i} \overbrace{p_{ij} v_j}^{\text{Misclassified species}} \qquad (6\text{-}5)$$

The following interpretation is useful when simulations are considered later on; since each column is identified with the probability vector of a multinomial distribution, it follows from Eq. 6-5 that the observed data for species *i* (n_i) are the sum of the output values of the i^{th} components of *m* multinomial distributions, i.e.,

$$n_i = \sum_{j=1}^{m} Multi(v_j, \boldsymbol{p}_{.j})[i] \qquad (6\text{-}6)$$

with the number of trials being the true number of detections v_j and the multinomial probability for species *j* being the j^{th} column $\boldsymbol{p}_{.j}$ of the confusion matrix, e.g., n_1 is the sum of the 1^{st} realized values of *m* multinomial distributions.

6.3. Methods

For this study, the effects of animal encounter rate (*v*) have not been considered, which can be an important source of uncertainty on animal abundance estimates, but would detract from the primary purpose of this chapter which was to examine the effects of misclassification. Therefore only the following two sources of uncertainty were considered:

1. the stochastic nature of the classification process;
2. Imperfect classifier performance (i.e., uncertainty on the values of the elements of the confusion matrix).

6.3.1. Models tested

First, only the stochastic nature of the classification process was considered, by assuming that the confusion matrix was known (i.e., no uncertainty). In a second step, additional uncertainty in the values of the confusion matrix itself was included.

The bias and variance on the estimates of the true number of detected calls was assessed using five different confusion matrices (Table 6-1) with increasing levels of misclassification. These included the identity matrix (i.e., no misclassification) and four others containing both low (Scenarios b or c) and high (Scenarios d or e) rates of misclassification with the misclassification being either the same (Scenarios b or d) or differing for each species (Scenarios c and e). For each confusion matrix the bias and variance using both equal data (i.e., same number of calls for each species, Scenario 1) and unequal data (i.e. differing numbers of calls per species, Scenario 2) were evaluated. All models were developed with four species. For equal data, the true number of calls was exactly 3000 for each species. For unequal data, values of 8000, 3000, 950 and 50 calls, respectively were selected. Thus the total number of calls was the same as the equal data, but with a 160-fold difference in the number of vocalisations between the most and the least abundant species.

The ten different scenarios (five confusion matrixes with equal and unequal data) are summarised in Table 6-2.

Table 6-1: The five different confusion matrixes (a - e) used during the simulation studies. Confusion matrix a is the identity matrix (no misclassification), b and c both have a high correct classification probabilities, but differ in that the misclassification probabilities of b are equal between species, whereas they are different in c. Confusion matrices d and e both have low rates of correct classification and again differ in that misclassification is equal between species in d, but varies in e.

		a) True species				b) True species				c) True species			
		SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Predicted species	SpA	1	0	0	0	0.85	0.05	0.05	0.05	0.85	0.08	0.02	0.01
	SpB	0	1	0	0	0.05	0.85	0.05	0.05	0.10	0.85	0.03	0.09
	SpC	0	0	1	0	0.05	0.05	0.85	0.05	0.03	0.05	0.85	0.05
	SpD	0	0	0	1	0.05	0.05	0.05	0.85	0.02	0.02	0.10	0.85
		Scenario x.a				Scenario x.b				Scenario x.c			

		d) True species				e) True species			
		SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Predicted species	SpA	0.52	0.16	0.16	0.16	0.52	0.04	0.20	0.20
	SpB	0.16	0.52	0.16	0.16	0.15	0.52	0.13	0.05
	SpC	0.16	0.16	0.52	0.16	0.10	0.14	0.52	0.23
	SpD	0.16	0.16	0.16	0.52	0.23	0.30	0.15	0.52
		Scenario x.d				Scenario x.e			

Table 6-2 Summary of the scenarios tested in the simulation study; similar misclassification probabilities means that elements of the confusion matrix outside the diagonal are the same between species (scenarios x.b and scenarios x.d), whereas for different misclassifications rates, they are different between species (scenarios x.c and scenarios x.e).

		Equal Data	**Unequal Data**
No misclassification		Scenario 1.a	Scenario 2.a
Low misclassification probabilities	**Similar misclassification probabilities**	Scenario 1.b	Scenario 2.b
	Different misclassification probabilities	Scenario 1.c	Scenario 2.c
High misclassification probabilities	**Similar misclassification probabilities**	Scenario 1.d	Scenario 2.d
	Different misclassification probabilities	Scenario 1.e	Scenario 2.e

6.3.2. Analytical approach

For the simple case, in which the variance within the values of the confusion matrix was assumed zero, an analytical solution for the bias and variance on the true number of detected calls (Appendix C.1) was derived. However, when uncertainty was added to the confusion matrix, the analytical approach became more complex, so bias and variance through

simulation were also explored. When variability in the values of the confusion matrix was added to the model, bias and precision were measured from simulation only.

6.3.3. Data simulation

6.3.3.a Stochastic nature of classification only

For each simulation (*b*), the numbers of misclassified, or observed, calls $\boldsymbol{n}_b$ were generated from the sum of four multinomial distributions with parameters $\boldsymbol{v}_b$ representing the true number of calls and *p*'s being the confusion matrix probabilities (Eq. 6-6). The estimated true number of calls $\hat{\boldsymbol{v}}_b$ was then estimated by multiplying the inverse of the confusion matrix by the number of misclassified (observed) calls (Eq. 6-7).

$$\hat{v}_b = C^{-1}\boldsymbol{n}_b \qquad (6\text{-}7)$$

For each scenario, this process was repeated 10 000 times and the mean (Equation C.4 in Appendix C.1) and variance (Equation C.10 in Appendix C.1) of the estimated $\hat{\boldsymbol{v}}$ calculated.

6.3.3.b Presence of uncertainty in the confusion matrix

When uncertainty in the confusion matrix was considered, the probabilities $\hat{p}_{.j}$ of the j^{th} column of the confusion matrix were viewed as realisations of a probability distribution. To meet the requirement that columns have to sum to one, this distribution was chosen to be a Dirichlet distribution (Part I.2.2.3.a,p 30)

For each of the 10 000 simulation trials, new values for the confusion matrix probabilities p_{ij} were generated from a Dirichlet distribution; these were then used in the same multinomial misclassification process as for the simpler situation. The true number of calls $\hat{\boldsymbol{v}}$ was again estimated using the inverse of the mean of the confusion matrix used to simulate the observed data (Eq.6-7). Simulations were run with two levels (low and high) of uncertainty on the confusion matrix. In both situations, the alpha parameters of the Dirichlet distribution were selected such that the means of the parameters were equal to the confusion matrix probabilities of the different scenarios (Table 6-3). To generate low uncertainty in the confusion matrix, the parameters were selected to have a variance equal to 0.01 on average. The parameters for the high level of uncertainty were selected to match a variance of 0.1 observed with real data in Gillespie et al. (2013).

Table 6-3: Examples of Dirichlet α parameters used for species A for each scenario. For the remaining species parameters α were the same but in different order to match the confusion matrices.

α for	Sc.x.a	Sc.x.b	Sc.x.c	Sc.x.d	Sc.x.e
Low uncertainty	100,0,0,0	85,5,5,5	85,10,3,2	52,16,16,16	52,15,10,23
High uncertainty	0.1,0,0,0	0.85,5,5,5	0.85,0.1,0.03,0.02	0.52,0.16,0.16,0.16	0.52,0.15,0.1,0.23

6.4. Results

Through this study the precision of the estimates was represented by the coefficient of variation (CV), which is the standard deviation of the estimate divided by the estimate, generally reported in per cent.

6.4.1. No uncertainty in the confusion matrix

When there was no uncertainty in the element of the confusion matrix, the analytical approach demonstrated that the means of $\hat{\boldsymbol{v}}$ were an unbiased estimate of the truth ($\boldsymbol{n}$), (Appendix C, Table C.1). The simulations verified this result (Appendix C, Table C.2); no significant difference between means and variances calculated analytically and estimated through simulation was observed.

As expected, without misclassification and despite the level of uncertainty, the estimates were unbiased and precise (CV=0). A decrease in the rate of correct classifications (scenarios b and c versus d and e) did not affect the $\hat{\boldsymbol{v}}$ estimate's means but it did significantly increase the variance and so the CV of these estimates (Figure 6-1).

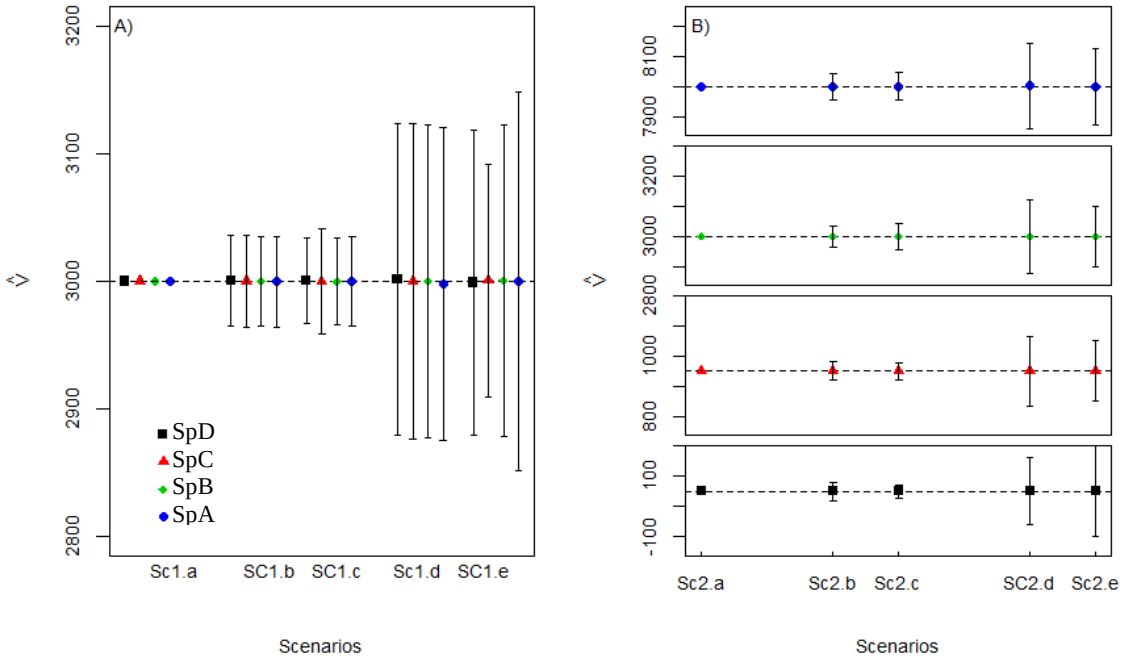


Figure 6-1: Expected true number of detections for each species, from simulation without uncertainty within the confusion matrix: for equal data scenarios Sc1a to Sc1e (A) and for unequal data scenarios Sc2.a to Sc2.e (B). Solid bars show the standard deviation and the dotted line the true number of detections.

Where there were different numbers of calls from the four species, unbiased estimates of the true numbers of calls were again obtained. The CV on the estimates of numbers of the more common species decreased (due to lower variance coming from misclassifications of the rarer species) but the CV of the estimates of the numbers of rare species calls rose significantly, reaching over 200% with confusion matrixes c and d (Figure 6-2 and Table C.1 & Table C.2).

6.4.2. Uncertainty in the confusion matrix

When uncertainty in the confusion matrix was included, the simulations again showed unbiased estimation of $\hat{v}$ for all the misclassification scenarios (Appendix C, Table C.3 and Table C.4). However, adding uncertainty to the confusion matrix generated a large increase in the CV due to an increase of the variance (Figure 6-3). With equal data the CV, across all scenarios, increased on average from 2% without uncertainty to 11.7% with low uncertainty and to 87.7% with high uncertainty (Figure 6-3A).

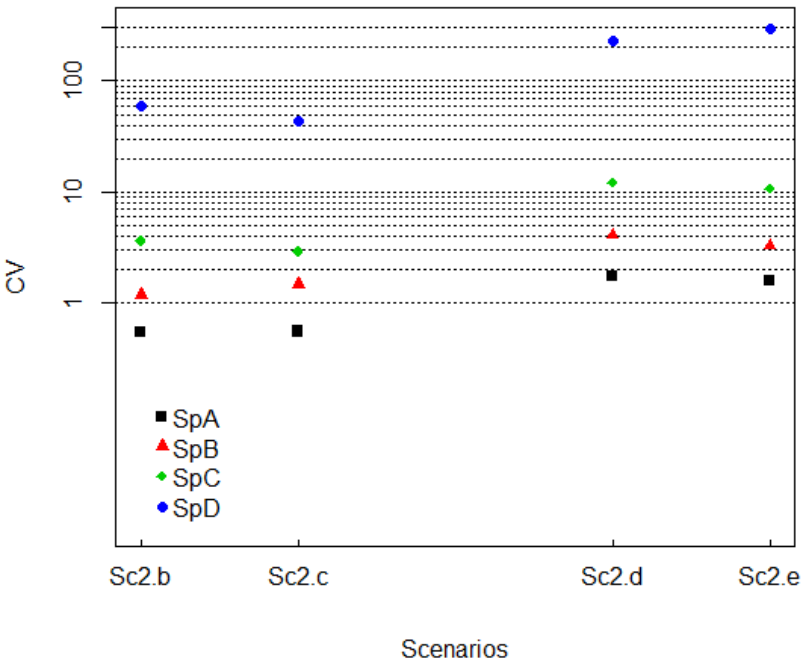


Figure 6-2: CV of the expected true number of detections for unequal data for each scenario (Sc2b to Sc2e), with different misclassification probabilities and no uncertainty in the confusion matrix. The y axis is on the log10 scale.

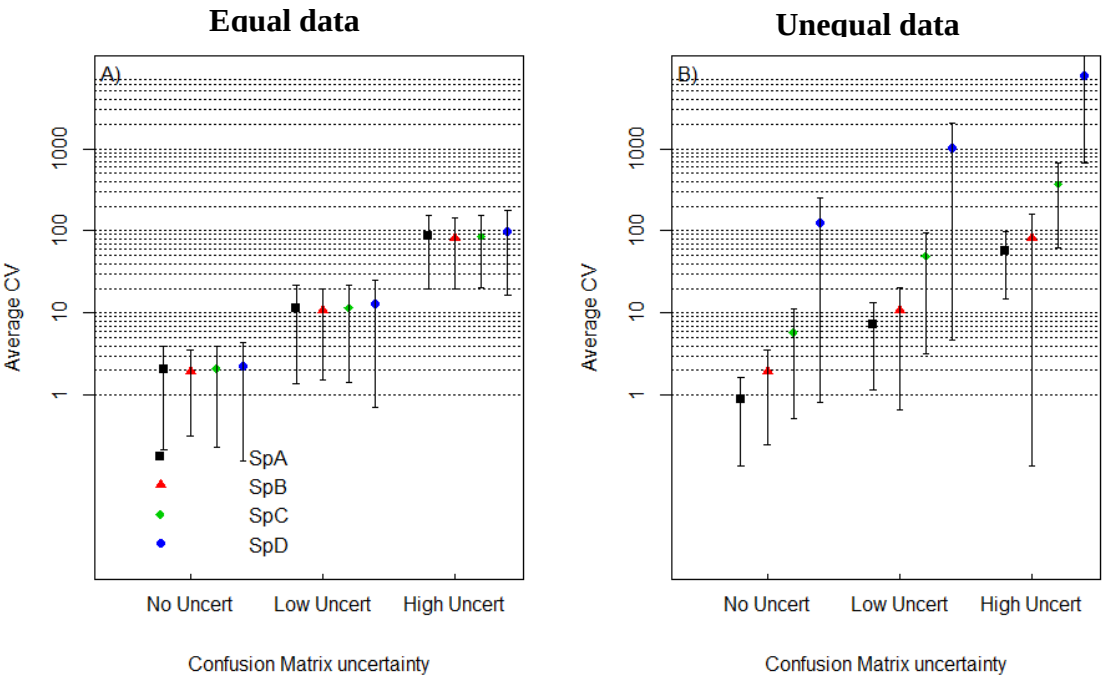


Figure 6-3: Mean of the CV of the expected true number of detections across the ten scenarios Sc1a to Sc1e (A) and Sc2a to Sc2e (B) for each species and each level of uncertainty of the confusion matrix values, no uncertainty, low uncertainty and high uncertainty. The y axis is on the log10 scale.

With the unequal data the average CV across all scenarios for the common species (species A and B) increased on average from 1.4% without uncertainty to 9% with low uncertainty to 69% with high uncertainty in the confusion matrix. For the rare species (species D) the average CV across the five scenarios was at 125% without uncertainty rising to 1 009% with a low level of uncertainty and 7 030% with a high level of uncertainty (Figure 6-3 B). With the high variability in the confusion matrix some individual simulation results gave some negative estimates of $\hat{\boldsymbol{v}}$, which is clearly not possible with real data.

The presence of uncertainties in the confusion matrix did not alter the fact that a confusion matrix with low misclassification will give a more precise estimation of $\hat{\boldsymbol{v}}$ than a confusion matrix with high misclassification probabilities (Appendix C, Tables C.3 and C.4).

6.5. Discussion

The results showed that it was possible to derive unbiased estimates of the true number of detections of each species from data containing misclassified acoustic detections. However the precision of the estimates was strongly related to the degree of misclassification (Figure 6-1) and the degree of uncertainty within the confusion matrix (Figure 6-3).

A low CV (<10%) on the estimated numbers of calls can be achieved in some situations, such as when there were similar numbers of calls between species, a low misclassification probability, and low uncertainty within the confusion matrix. In cases where there were large differences in the numbers of detected calls between species (scenarios 2.x), the CV was much higher on the estimates of the number of calls from the rarer species. In the more optimistic scenarios (low misclassification probability and low uncertainty within the confusion matrix), the CV for the common species A and B varied between 0.55% to almost 9%. However, the CV increased close to 100% for less common species (species C) in scenarios with a high rate of misclassification and low uncertainty for the values of the confusion matrix. For species with a very low encounter rate (Species D), even with a small level of uncertainty and low misclassification probability, the CV was higher than 400%, reaching the value of 2500% with a high misclassification probability. With uncertainties in the confusion matrix similar to those observed in real data (Gillespie et al., 2013), the CV

was higher than 50%, even for common species, and the estimate became totally uninformative for the rare species (CV>10000%).

For the rare species, some estimates of the true number of detections were not biologically possible as they were negatives. These negative predictions were a result of the mathematical characteristics of the inverse confusion matrices (containing negative values) associated with the stochastic process between the inverse of the confusion matrix and the observed number of detections. The inverse of all the confusion matrices used contained large negatives values. To obtain the true number of detections these inverse matrices were multiplied by a vector of observed data containing only positive values and obtained from a stochastic process from the confusion matrices (sum of multinomial distributions). Consequently some outputs could be negatives. For example if only 2 species A and B are considered with few detections of species A (10) observed and much more of species B (60) and species B is 40% of the time misclassified as species A then mathematically the true number of detections will be negative (see Box 1 for demonstration). However the average of the estimates was always positive and was unbiased.

Confusion matrix for species A and B: $\begin{matrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{matrix}$

Observed number of detections of species A and B: $\begin{matrix} 10 \\ 60 \end{matrix}$

True number of detections for species A and B: $\begin{matrix} a \\ b \end{matrix}$

Then:

$$0.8 \times a + 0.4 \times b = 10$$

$$0.2 \times a + 0.6 \times b = 60$$

So:

$$a = -45$$

$$b = 115$$

Box 1: Demonstration that in some situation it is possible to obtain negative estimates of the true number of detections *a* and *b* for respectively species A and B.

From the results it appeared that uncertainty in the confusion matrix was the parameter responsible of most of the variance of the estimates. Indeed the average CV, across all

species and all misclassification probabilities, was 70 times higher when a high level of uncertainty (average CV across 4 species = 1885) was assumed for the confusion matrix than where there was no uncertainty in the confusion matrix (average CV across 4 species = 27). Whereas the average variance, across all species and all levels of uncertainty within the confusion matrix, was only 29 times higher for models with a high misclassification probability (mean CV=13211) that for models with a low misclassification probability (mean CV=450).

A CV of 10% on a density estimate is considered as very good, a CV of 20% as reasonable and a CV of 100% near useless (Thomas and Marques, 2012). Particularly for rare species, CV's are often high, generally due to a low encounter rate. For example, Hammond et al. (2002) used visual line transect distance sampling methods to estimate the abundance of the relatively common European harbour porpoise, *Phocoena phocoena*, with a CV of 14%, but the abundance of the rarer common dolphin *Delphinus delphis* from the same survey, had a CV of 67%. (Gerrodette et al., 2011) estimated the abundance of the extremely rare Vaquita *Phocoena sinus* in the Gulf of California with a CV of 73%.

In this chapter, only uncertainty in estimates of the true number of detections due to misclassification has been considered. In practice, however, significant contributions to the overall CV can be expected from the estimate of detection range, the encounter rate, and the estimate of vocalisation (cue) rate which is unknown for many species. Thomas and Marques (2013) outline a number of methods for estimating both detection range and cue rate and the method chosen will be dependent on both the species and the study area. If we consider the species for which the true number of detection is estimated with a CV lower than 50% (for example, common species A and B), we can hope that, despite unavoidable misclassifications, acoustic detections provide useful information. However for the rare species, a small amount of misclassification from the more common species can render the acoustic data useless for all practical purposes.

Since the uncertainty on the estimate of each species is highly dependent on the presence of other species, incorporating information on the likely abundance of calls from other species will hopefully lead to more robust estimates. Therefore the next chapter presents the development of a Bayesian model which incorporates prior information on the relative abundance of calls from different species (based on previous survey work and information on call rates) as well as the uncertainty on the values in the confusion matrix. The Bayesian

approach will also resolve the problem encountered by your analytical method which can (incorrectly) produce point estimates that are negative: the Bayesian estimation has the stochastic nature of the observations built in.

Chapter 7: A Bayesian method to estimate the number of acoustic detections in the presence of species misclassification.

7.1. Introduction

The previous chapter proposed a non-Bayesian method to estimate the true number of acoustic detections, for several species at the same time, from simulated observed acoustic detections misclassified by an automatic classifier. The influence of the misclassification probabilities and of the uncertainty within the confusion matrix on the bias and precision of the estimates were compared. For all classification scenarios (Table 6-2,p103) the true number of acoustic detections could be estimated without bias. However, the precision of the estimates varied from a CV of few per cent to a CV of more than 1000% depending on the classification probabilities, the uncertainty within the confusion matrix and the encounter rate of a species. The method used in chapter 6 had two main limitations. It sometimes generated negative estimates (see Figure 6.1, Species D), which clearly would not be found in real data in the given context. In addition, a reliable measure of the precision of the estimates could only be obtained with a sufficiently large sample size. In practical applications, it may be unrealistic to collect data with a sufficiently large sample size to quantify the precision of the estimates.

However, for some acoustic detection studies the true number of detections and the amount of uncertainty in the confusion matrix might be available from prior surveys or analyses. Indeed the true number of detections depends mainly on the number of animals within the acoustic detection range, on the call rate and on the detection rate. If the PAMGUARD Whistle Classifier (PWC) is used to identify the species, then the true number of detections also depends on the fragment and section lengths of the whistle contours (see Part I). Previous surveys and also prior knowledge from other sources might be used to obtain some of this information. Given the development process of the PWC (chapter 2 and 3), it is possible to obtain measurements of the confusion matrix uncertainty when it is used for classification.

In this chapter it was assumed that such prior knowledge is often available. In the analysis of ecological data a Bayesian approach is frequently used since it allows us to include prior knowledge in the model (Eguchi and Gerrodette, 2009; King et al., 2010; Taylor et al., 1996;

Wade, 2000). In addition, it is possible to assess the impact of the precision of the prior knowledge on the final estimate by extracting summary statistics of the posterior distributions. Another advantage of the Bayesian framework is that prior distribution for the parameters may be chosen such that the estimate will have only positive values.

Similar to the previous chapter, this chapter discusses a simulation study. Here the performance of the estimation of the true number of acoustic detections for four species in a Bayesian framework is assessed. The same models as used in the previous chapter with different misclassification probabilities (Table 6-1), amount of uncertainty in the confusion matrix and level of knowledge about the true number of detections are considered.

7.2. Methods

7.2.1. Data

For each model a dataset was simulated for four species using the method described in the previous chapter (Part II.6.3.3.b, p104) where the confusion matrix and the (true) number of detections were the same as in the different scenarios listed in Table (Table 6-2, p103).

7.2.2. Overview

A summary of the notation that was already used in the previous chapter and is again needed for this chapter is provided in Table 7-1.

Table 7-1: Summary of the notation used in previous chapter.

Symbol	Description
$\boldsymbol{n}$	Vector containing the number of observed detection of species *i* with $n = (n_1, \dots, n_i, \dots, n_m)\ \forall \mathrm{i} = 1, \dots, \mathrm{m}$
$\boldsymbol{v}$	Vector containing the true number of detections: $v = (v_1, \dots, v_j, \dots, v_m)$ with $\forall \mathrm{j} = 1, \dots, \mathrm{m}$
$\hat{p}_{ij}$	Estimated classification probabilities: Estimated probability of classifying species *j* as species *i* from PAMGUARD Whistle classifier
y_{ij}	Number of acoustic detections classified as species *i* and made by species *j*
C	Confusion matrix: $m \times m$ matrix (6.2, pp100)
$f()$	Likelihood functions
$\rho()$	Prior distribution functions
$\pi()$	Posterior density functions

The Bayesian models used in the chapter are based on the model described in the previous chapter

$$E[\hat{v}] = E[\hat{C}^{-1}]E[n].$$

The parameters to estimate were the true number of detections $\hat{v}_j$ (Table 7-1) for each species *j=1,..,4*. The simulated data, ***n***, consisted of realisations from a stochastic process that depended on the classification probabilities p_{ij} in the confusion matrix *C* and on the vector ***v,*** which contained the true number of detections (Eq.7-1).

$$n_i \sim \sum_{j=1}^{m} Multi\left(v_j, \hat{\boldsymbol{p}}_{.j}\right)[i] \qquad (7\text{-}1)$$

As described in the classification process (Part II.6.2, p100) the observed data for species *i* (n_i) was the sum of the i^{th} realized values of the vector of m, multinomial distributions (the i^{th} realized values is symbolised by the $[i]$ in Eq. 7-1) where the number of trials being the true number of detections v_j and the multinomial probabilities for species *j* is given by the j^{th} row $\hat{p}_{.j}$ of the confusion matrix. If y_{ij} were the realized values of the j^{th} multinomial then

$$n_i = \sum_{j=1}^{m} y_{ij} \text{ with } y_{.j} \sim Multi\left(v_j, \hat{\boldsymbol{p}}_{.j}\right) \qquad (7\text{-}2)$$

According to Bayes' theorem and assuming $\boldsymbol{v}$ and $\boldsymbol{p}$ are independent, the joint posterior distribution for $\boldsymbol{v}$ is:

$$\pi(\boldsymbol{v}|\boldsymbol{n}, \boldsymbol{p}) \propto f(\boldsymbol{n}|\boldsymbol{p}, \boldsymbol{v})\rho(\boldsymbol{v})\rho(\boldsymbol{p}), \qquad (7\text{-}3)$$

where $f(\boldsymbol{n}|\boldsymbol{p}, \boldsymbol{v})$ is the likelihood and $\rho(\boldsymbol{v})$ and $\rho(\boldsymbol{p})$ are the prior distributions for $\boldsymbol{v}$ and $\boldsymbol{p}$, respectively. These prior distributions denote the probability of obtaining the acoustic detections $\boldsymbol{v}$ and the classification probabilities $\boldsymbol{p}$ before the data $\boldsymbol{n}$ have been observed. $\boldsymbol{v}$ and $\boldsymbol{p}$ are independent, as the classification probabilities in the confusion matrix are obtained independently of $\boldsymbol{v}$. For more details of Bayesian theory, see section Part I.1.4 (p17).

7.2.3. Likelihood functions

The simulated data derived as the sum of unobserved parameters $\boldsymbol{y}$ (Eq.7-2) hence

$$f(\boldsymbol{n}|\boldsymbol{p}, \boldsymbol{v}) = \prod_{j=1}^{m} f(\boldsymbol{y}_{.j}|\boldsymbol{p}, \boldsymbol{v})$$

with the likelihood being a product of multinomial distributions

$$f(\boldsymbol{y}_{.j}|\boldsymbol{p}, \boldsymbol{v}) = \frac{v_j!}{y_{1j}! \dots y_{ij}! \dots y_{mj}!} p_{1j}^{y_{1j}} \dots p_{ij}^{y_{ij}} \dots p_{mj}^{y_{mj}}$$

7.2.4. Prior distributions

7.2.4.a Prior distributions for the true number of detections *v*

The following was based in the assumption that some prior knowledge about some of the parameters (i.e r,P, see equation 6.1) driving the number of detections was available from previous studies. Indeed, depending on species, the CV of the abundance estimates may be very different. However, for whistling species, the CV of the abundance estimates frequently ranges from 20% to 60% (Barlow and Forney, 2007; CODA, 2009). For most species, the call rates are either completely unknown or known but with information on precision not available. Indeed call rates are dependent as it is a parameter that depends on various factors such as group size or behaviour (Buckastaff, 2004; Quick and Janik, 2008) hence they are difficult to measure. Hence the CV for this parameter was expected to be high. Finally, for the detection probability, two parameters are commonly measured to establish the performance of a detector, the precision, estimating the rate of correct detections, and the recall, measuring the detection efficiency(Gillespie et al., 2013).

To model these priors knowledge's, the prior distribution on the number of detections was assumed to follow a negative binomial distribution (Eq. 7-4) with parameters μ (mean) and σ^2(variance) to account for over dispersion in the data. Conventionally the parameters of a negative binomial function are the number of trials n and the probability of success for each trial *p*, respectively

$$\rho\left(v_j|\mu,\sigma^2\right) = \frac{\tau\left(v_j+n\right)}{\tau(n)v_j!}p^n(1-p)^{v_j} \qquad (7\text{-}4)$$

with $n = \frac{\mu^2}{\sigma^2-\mu}$ and $p = \frac{n}{n+\mu}$.

A prior sensitivity analysis was carried out with three different sets of priors (Table 7-2) and their impact on the estimate of true number of detections for each species was analysed. Each set of priors contained four prior distributions, one for each species in the classifier:

Prior V1: This set of prior parameters was chosen so that the CV of the prior distribution was equal to 40% (a common CV value found for abundance estimates of cetacean populations (Barlow and Forney, 2007; CODA, 2009; Forney et al., 1995)).

Prior V2: The second set of prior parameters, on the other hand, was selected to mimic a situation where the true number of detections was fairly well-known with a CV of 10%. In practice, such a situation is rather unrealistic, especially if the parameter $\boldsymbol{\nu}$ depends on other highly variable parameters such as the call rate. Despite the fact that this it is a rare situation, this CV was simulated to better understand the relative strength of the parameters' influence on the precision of an estimate. It is important to identify the main source(s) of uncertainty to target these specifically if one seeks to reduce uncertainty in the estimates of the true number of detections.

Prior V3: Since the previous chapter has indicated that rare species tend to be more sensitive to misclassification than common species, this chapter investigated the consequences of a prior with a large CV on rare species (as found in the literature (Gerrodette et al., 2011)) along with a small CV on the more common species for which more prior information are available as they are easier to detect. As a consequence, this last set of prior parameters was used only with unequal data (scenarios 2.x) and prior distribution parameters were chosen such as the CV's of the distribution were different for each species (Table 7-2). These parameters were chosen such as the CV for the rarest species (species D) was 60% for Models A (see section 7.2.4.b), but 40% for models B because a lack of convergence of the algorithm was noted when 60% was used.

Table 7-2: Prior parameters of the negative binomial prior distribution $\boldsymbol{V1}, \boldsymbol{V2}$ and $\boldsymbol{V3}$.

		Scenarios 1.x	**Scenarios 2.x**			
		All Species	SpA	SpB	SpC	SpD
Prior V1	mean	3000	8000	3000	950	50
CV=10%	variance	1.8×10^{5}	1.25×10^{6}	1.8×10^{5}	1.8×10^{4}	51
Prior V2	mean	3000	8000	3000	950	50
CV=40%	variance	1.4×10^{6}	10.2×10^{6}	1.4×10^{6}	1.44×10^{5}	400
Prior V3 (Models A)	mean		8000	3000	950	50
CV=10%,20%,40%,60%	variance		1.25×10^{6}	3.6×10^{5}	1.44×10^{5}	900
Prior V3 (Models B)	mean		8000	3000	950	50
CV=10%,20%,40%,40%	variance		1.25×10^{6}	3.6×10^{5}	1.44×10^{5}	400

7.2.4.b Prior distributions for the classification probabilities $p_{.j}$

In chapters 3 and 4 and also in Gillespie et al. (2013) the classification probabilities in the confusion matrix of each generated classifier were associated with a standard error. This variability was caused primarily by the sampling process used to develop the classifiers (chapter 2). The previous chapter (chapter 6) has demonstrated that uncertainty within the confusion matrix had more impact on the precision of the estimate of the true number of detections than the actual misclassification probabilities.

In the Bayesian model described previously, the prior distribution $\rho(p_{.j})$ reflects this uncertainty. Following the same reasoning as in the previous chapter, this prior followed a Dirichlet distribution, which took on similar values to mirror the entries of the confusion matrix (in particular the requirement that probabilities in each column of the confusion matrix have to sum to 1 (chapter6 Eq.6-3, p99) used in the different scenarios. Furthermore, the Dirichlet distribution is the conjugate prior distribution of a multinomial distribution (Gelman et al. 2004). The Dirichlet distribution is defined by a vector parameters $\alpha = (\alpha_1, \dots \alpha_i, \dots \alpha_m)$ with $\alpha_m > 0$ (Gelman et al., 2004) as

$$\rho(p_{.j}|\alpha_{.j}, \dots, \alpha_{ij}, \dots, \alpha_{mj}) = \frac{\Gamma(\alpha_{1j} + \cdots + \alpha_{ij} + \cdots \alpha_{mj})}{\Gamma(\alpha_{1j}) \dots \Gamma(\alpha_{ij}) \dots \Gamma(\alpha_{mj})} {p_{1j}}^{\alpha_{1j}-1} \dots {p_{1j}}^{\alpha_{ij}-1} \dots {p_{1j}}^{\alpha_{mj}-1} \quad (7\text{-}5)$$

with $p_{1j}, \dots p_{ij}, \dots p_{mj} \geq 0; \sum_{i=1}^{m} p_{ij} = 1; \alpha_0 = \sum_{i=1}^{m} \alpha_{ij}$

The results in chapter 6, showed that classification probabilities with high uncertainty reduced the precision of the estimate of the true number of detections. The consequences of different level of uncertainty on the classification probabilities are investigated in this chapter by a sensitivity analysis with four different Dirichlet parameters for each misclassification scenario (for the exact values of the parameters α, see Appendix D, Table D.1) and for each of the prior assumed for the parameters of the true number of detections.

Models A: P0

A first series of models were tested assuming the classification probabilities were not estimates but known values. These models (Models A, Eq.7-6) did not have a prior distribution (P0) on the *p* parameters.

$$\pi(v|n,P) \propto f(n|P,v)\rho(v) \quad (7\text{-}6)$$

Models B

Three sets of priors on the parameters *p* were tested, with each set containing four prior distributions on the parameters *p* for each species. The parameters α of the Dirichlet distributions were chosen such that $\frac{\alpha_{ij}}{\alpha_0} = E\left[\hat{p}_{ij}\right]$ and such that the CV's of the correct classification probabilities ($\hat{p}_{ii}$) were different between the priors tested. Comparing the CV of the confusion matrices generated in chapters 3, 4 and Gillespie et al (2013), it was observed that the CVs of the correct classification probabilities of a confusion matrix were influenced by the quality of the training data. In chapter 4 the quality of the training dataset was low and the CV of the correct classification probabilities ranged from 15% to 57%. In chapter 3 where the confusion matrix was obtained with a better data set the CVs ranged from 3.6% to 23%. Finally in Gillespie et al (2013) for which a very good quality training dataset was used to train the classifier the CVs of the correct classification probabilities ranged from 0.2% to 54%. The choice of the parameters of the Dirichlet distribution for each set of prior was made such that these ranges of CVs were represented.

Prior P1 was an informative prior with the parameters α selected for each species such that $\sqrt{\frac{var(p_{ii})}{E[p_{ii}]}} = 0.01$ to simulate a CV of 1% for the correct classification probabilities, with $E\left[\hat{p}_{ij}\right]$ being the classification probabilities of the confusion matrix used in Scenarios .x .

Prior P2: The second prior was less informative with a set of parameters selected such the CV of the correct classification probabilities was equal to 40% and with $E\left[\hat{p}_{ij}\right]$ being the classification probabilities of the confusion matrix used in Scenarios .x.

Prior P3: The third prior was selected to simulate a confusion matrix with random classification and hence all parameters α were equal to 1. With such parameters $E\left[\hat{p}_{ij}\right] = 0.25$ and the CV=77% for the correct classification probabilities.

7.2.5. Investigated scenarios

A total of 85 models were tested (Table 7-3) on simulated data.

7.2.5.a Models A: known p_{ij}

Twenty-five models were tested with a prior on the *v* parameters and no prior (*P0)* on the classification probabilities. The same confusion matrices as in the previous chapter (Scenario.a to Scenario.e) were used. Each of these five scenarios was associated with two (for equal data) and three (for unequal data) priors on the ***v*** parameters (Table 7-3).

7.2.5.b Models B: including prior on p_{ij} (P1 to P3)

Given the properties of the Dirichlet distribution it is not possible to choose parameters α such that $E[\hat{p}_{ii}] = 1$ and $var(p_{ii}) \neq 0$, consequently Scenario.a have not been tested with Models B. So Models B corresponded to 20 Models A for which priors on the p_{ij}'s were added to each model. The priors *P1*, *P2* and *P3* described in Section (7.2.4.b) were tested on each of the 20 models.

Table 7-3: Summary of all the investigated Bayesian models. Sc1.x and Sc2.x correspond to the scenarios of misclassification (Scx.a ,Scx.b, Scx.c, Scx.d, Scx,e) described in chapter 6. The prior parameters were described in the section 7.2.4. MH (Metropolis Hastings) and GS (Gibbs sampler) are the MCMC alogithms used in the models

		v priors			MCMC
		V1	V2	V3	Algorithms
p priors	P0	Sc1.x	Sc1.x		MH
		Sc2.x	Sc2.x	Sc2.x	
	P1	Sc1.x	Sc1.x		MH + GS
		Sc2.x	Sc2.x	Sc2.x	
	P2	Sc1.x	Sc1.x		
		Sc2.x	Sc2.x	Sc2.x	
	P3	Sc1.x	Sc1.x		
		Sc2.x	Sc2.x	Sc2.x	

7.2.6. Posterior inference

To obtain posterior inference on the parameters $\hat{\boldsymbol{v}}$, a Markov chain Monte Carle algortyhm was used (Part I.1.4.3.c, p19). For Models A, this was implemented using a Metropolis-Hastings (MH) sampling algorithm; for Models B a Gibbs sampling algorithm was added to update the parameters *p* (see below for details in both cases). All the algorithms were implemented in the statistical software (*R Development Core Team*, 2012)

7.2.6.a Metropolis-Hasting (MH) algorithm: the proposal density function

The MH (Hastings, 1970) algorithm was used to update the parameters ***y*** and ***v.***

Because the data were derived as a sum of the i^{th} elements of *m* multinomial distributions (Eq.7-2) the parameters were updated by blocks of *m* parameters. The proposal density function was a multinomial distribution such that:

$$y_{i\cdot} = \frac{n_i!}{y_{i1}! \dots y_{im}!}\left(\frac{v_1 P_{i1}}{\sum_{j=1}^{m} v_j P_{ij}}\right)^{y_{i1}} \dots \left(\frac{v_j P_{ij}}{\sum_{j=1}^{m} v_j P_{ij}}\right)^{y_{ij}}$$

Once a block of j y_{ij}'s was updated, the v_j were also updated (Eq. 7-7) and the current parameter values were accepted following the acceptance rules described in the introduction (Part I.1.4.3.c, p19).

$$v_j = \sum_{i=1}^{m} y_{ij} \qquad (7\text{-}7)$$

7.2.6.b Gibbs sampler

The Dirichlet distribution is the conjugate prior for the probability parameters of the multinomial distribution. Thus the conditional posterior distribution of the probability $p_{\cdot j}$ of observing $y_{\cdot j}$ was a Dirichlet distribution with parameters $(\alpha_1 + y_1, \dots, \alpha_{ij} + y_{ij}, \dots, \alpha_{mj} + y_{mj})$ (Gelman et al., 2004)**.**

7.2.6.c Convergence, burn-in and thinning

For each model, three MCMC chains (Part I.1.4.3.c, p19) with different initial values were run for up to 800 000 iterations. The initial values of one chain were the true values. The initial values of the two other chains were simulated from the prior distributions of the models and values at least 20% away from the true values were selected. For successive

sections of 10% of the iterations a convergence diagnostic was applied to detect the section in which the Markov chain had converged. Trace plots, auto-correlation plots (acf plot) and a BGR (Part I.1.4.5, p21) convergence test were used to determine if the model had reached convergence within this section. The section for which the convergence reduction factor was lower than 1.2 (Part I.1.4.5, p21) for the $\hat{v}_j$ parameters for each species was then identified as the section of convergence. The iterations before convergence were discarded as a burn-in. If convergence was not reached after 800 000 iterations, a second set of three chains was generated with one chain starting from the true value while the starting values of the two remaining chains were simulated similarly as in the first run, but with values being selected around 10% away from the true values. The same convergence diagnostic was applied. If after this second MCMC run, the chains still did not converge, the corresponding model was declared as non-converging and no further analysis was done for this model. Each converging model was replicated L=300 times. For all the replicates the initial values of the parameters have been chosen to be the true values used to simulate the data.

Due to high serial auto-correlation in the chains, only one in every four iterations ("thinning" (King et al., 2010)) were kept after burn in to sample from the posterior distributions of the $\hat{v}_j$ parameters. Summary statistics (mean and standard deviation) were extracted from these posterior distributions.

7.2.6.d Model performance and summary statistics

To analyse the impact of the different priors and confusion matrices on the estimates the relative errors (where "relative error" is defined as the difference between the mean of the posterior distribution and the expected true value divided by the expected true value) and posterior distribution coefficient of variation (CV) for each species were measured for each replicate. Based on the relative errors the relative biases (mean of relative errors) between models were compared to analyse the impact of the prior variances on the accuracy of the estimates. The means of the CVs for each model were compared to analyse the impact of the prior variances on the precision of the estimates.

7.2.7. Statistical versus biological significance

To compare the bias and precision between models ANOVA was used. In this thesis the null hypothesis compared if the relative bias or precision of the estimates of the true number of detections were significantly different between the priors used in the model. While most of the tests were statistically significant, the question of whether this marks a biologically significant difference has to be considered separately. The decision about whether or not such a difference is biologically significant has to take into account the specific context of the data. For example a difference of 50 detections between two estimates of a species for which the average number of detection is 3000 is not the same as a difference of 50 detections when the average number of detection for a species is 100 detections. In density estimation a CV of 10% of a density estimate is considered as very good, a CV of 20% is reasonable whereas a CV of 100% is nearly useless (Thomas and Marques, 2012). Given that this study was carried out within the larger picture of the whole abundance estimation process in mind, a relative bias or difference in CV that is lower than 10% will not be considered as biologically different in this discussion. A difference between 10 to 40% will be considered as biologically significant and one that is greater than 40% as highly significant.

7.3. Results

7.3.1.a Convergence and sensitivity with respect to the starting values

For Models A with known classification probabilities and equal data, convergence was always reached (Appendix D table D-2) after 30 000 iterations. The MCMC chain showed good mixing (Figure 7-1) and a relatively rapid decrease in auto-correlation (Figure 7-2).

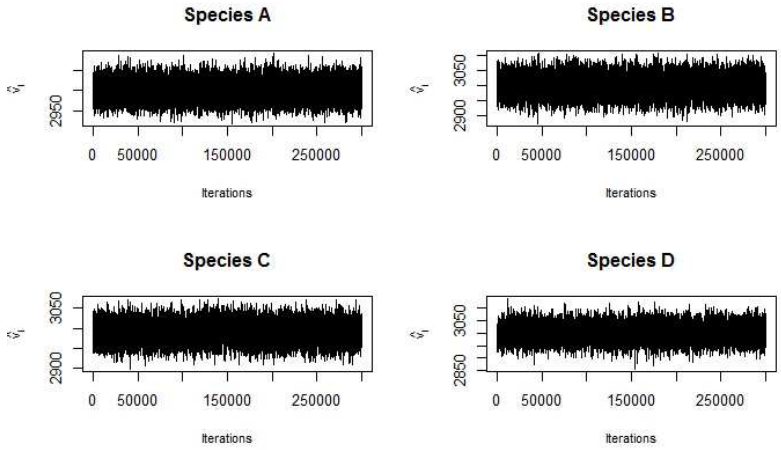


Figure 7-1: Trace plots showing MCMC sample values for parameters *v* (y-axis) vs. sample iteration (x-axis, after thinning), obtained with classification scenario Sc1.b and prior *V1*.

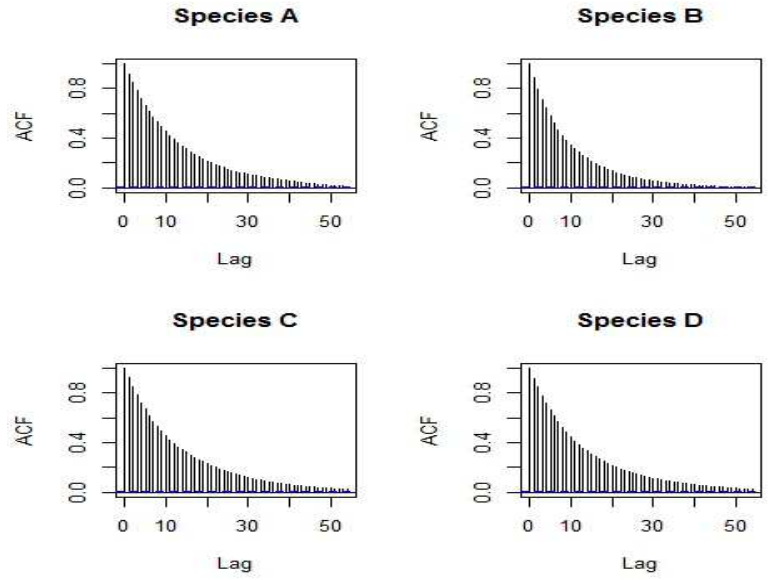


Figure 7-2: Auto-correlation plots of the posterior distributions of parameters v_j obtained from the model with classification scenario Sc1.b and prior *V1*. Note that samples were thinned, so that a lag of 1 corresponds to 4 MCMC sample iterations.

Nonetheless for models simulated with a CV of 40% for the prior distribution on the parameters $\boldsymbol{v}$ (prior V2) and high misclassification probabilities (Sc1.d and Sc1.e), convergence depended on the Markov chains' initial values. When these values were more than 20% away from the true values, despite apparently good mixing within each chain, the multiple chains did not converge (Figure 7-3, Appendix D Table D-2). Convergence was achieved for all models if the initial values were selected within 10% of the expected true values.

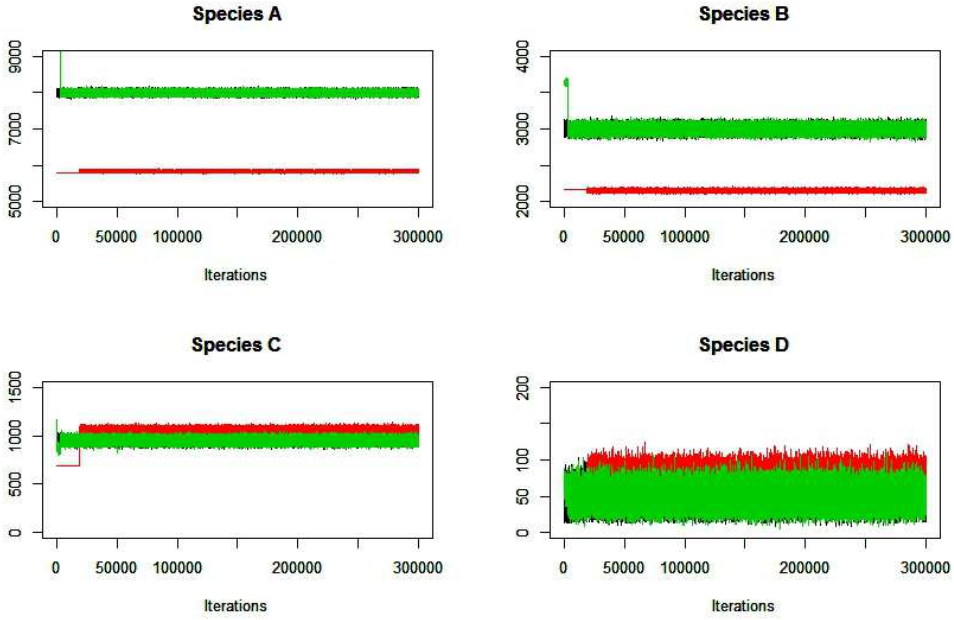


Figure 7-3: Trace plots for each chain obtaine in the analysis of model with scenario Sc.e and prior V2 for each species. Within-chain mixing is good but the chains arere not converging towards the same values.

In models were the true number of detections differed between species (classification scenarios Sc2.x), all models with a CV of 40% for the prior distribution of the parameters $\boldsymbol{v}$ (V2) were sensitive to the initial values of the Markov chain independent of the classification scenarios (Appendix D Table D-2). When the prior CV was different between species (V3), only models with high misclassification probabilities were sensitive to the initial values. However, it was not possible to estimate $\hat{v}_D$ for the rare species (Species D) in these models: after few iterations the chain stopped updating and no new parameter values were accepted. In these cases, more iterations (60 000) were necessary to achieve convergence in v for the other species.

Models B were also sensitive to the initial values of the Markov chains. An increase in the variability of the prior distributions as well as in the misclassification probabilities made the Markov chains more sensitive to the initials values (Appendix D Table D-3). The issue was aggravated with unequal data. With initial values, around 10% of the true values, as described for Models A convergence was reached for all models after up to 480 000 iterations.

7.3.1.b Sensitivity analysis and posterior inferences

Across all tested models, the Markov chain stopped updating for 0.02% and 4% of replicates of Models A and Models B, respectively. This concerned essentially estimates for species D. When the models were run with no misclassifications the estimates for all species were equal to the expected true values for each replicate.

7.3.1.b.i Bias in the estimated number of detections by species

Impact of the proportion of true number of detections between species

The accuracy of the estimates was influenced by the ratio in true numbers of detections between species. For each scenario with an equal true number of detections between species, the $\hat{v}_j$ estimates for each species were unbiased (Table 7-4, Figure 7-4).

For scenarios with unequal data between species, the relative bias was higher 0.3% and -3.51% in Models A and Models B respectively (Figure 7-4). Furthermore the standard deviations of the mean bias were higher than with equal data (Table 7-4).

Table 7-4 : Mean and standard deviation (in brackets) of the relative bias across all Models A and all Models B when the same number of detections for each species was simulated (Sc1,equal data) and when different number of detections between species were simulated (Sc2, unequal data).

	Models A	Models B
Sc1: Equal data	0% (0.06)	0% (0.1)
Sc2.: Unequal data	0.31% (4.6)	-3.6% (14.9)

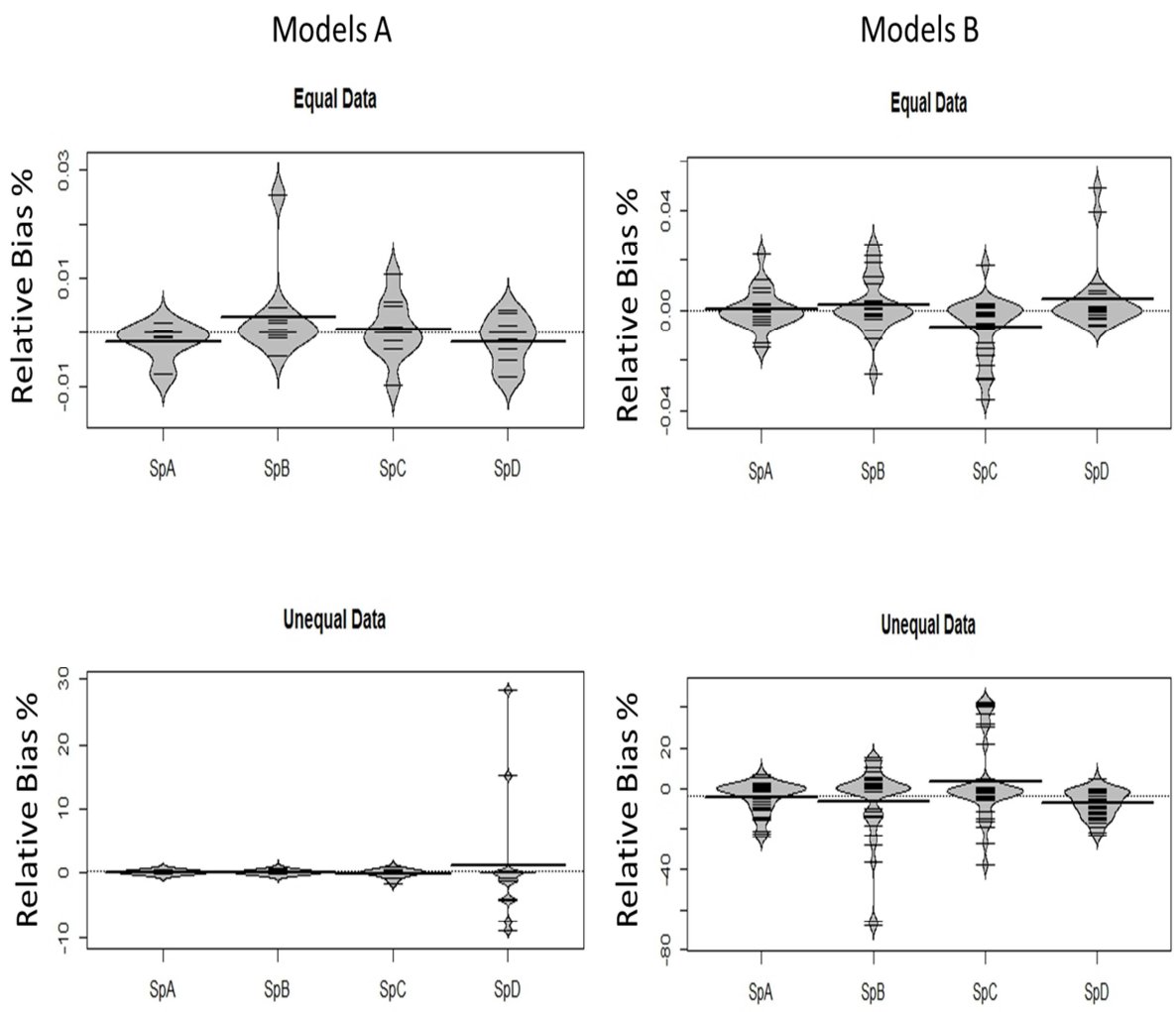


Figure 7-4: Relative bias (beanplots) and mean relative bias (bold lines) for each species for models where *p* is assumed known (Models A) or is estimated (Models B), and for models with equal and unequal data. The thin lines crossing the entire plot (close to zero) are the mean across the four species. For Models A and equal data each beanplot is computed from 10 values (the relative bias for 5 scenarios times 2 priors on v), for unequal data each beanplot is computed from 15 values (the relative bias for 5 scenarios times 2 priors on v). For models B each bean plot is computed from 24 and 36 values from equal and unequal data, respectively (relative bias for 4 scenarios times 3 priors on p times 2 or 3 prios on v, respectively).

Unequal data in Models A

In models A with unequal data, the impact of the prior variances and of the misclassification probabilities was different depending on the species. Nonetheless, decreasing the informativeness of the prior (by increasing its variance) on ***v*** increased the absolute values in the relative bias statistically significantly for all species ($p<0.001$) (Figure 7-5) but biologically insignificantly for all species except species D. For this rare species, this increase

was the most pronounced. On average across all misclassification scenarios the relative bias for species D ranged from -0.63 for V1 to 9.63 for V3. More precisely at a low misclassification probability level this bias reached 15% and 28% (Figure 7-5,) with prior V3 and misclassifications scenario b and c (not no data were available for scenario Sc2.d and e). For the other species the maximum bias difference observed between V1 to V3 was for Species C with a relative bias ranging from -0.45 to 0.42 whereas the maximum range of relative bias observed for the three other was for species C with a relative bias ranging from -1.63% to 0.88%.

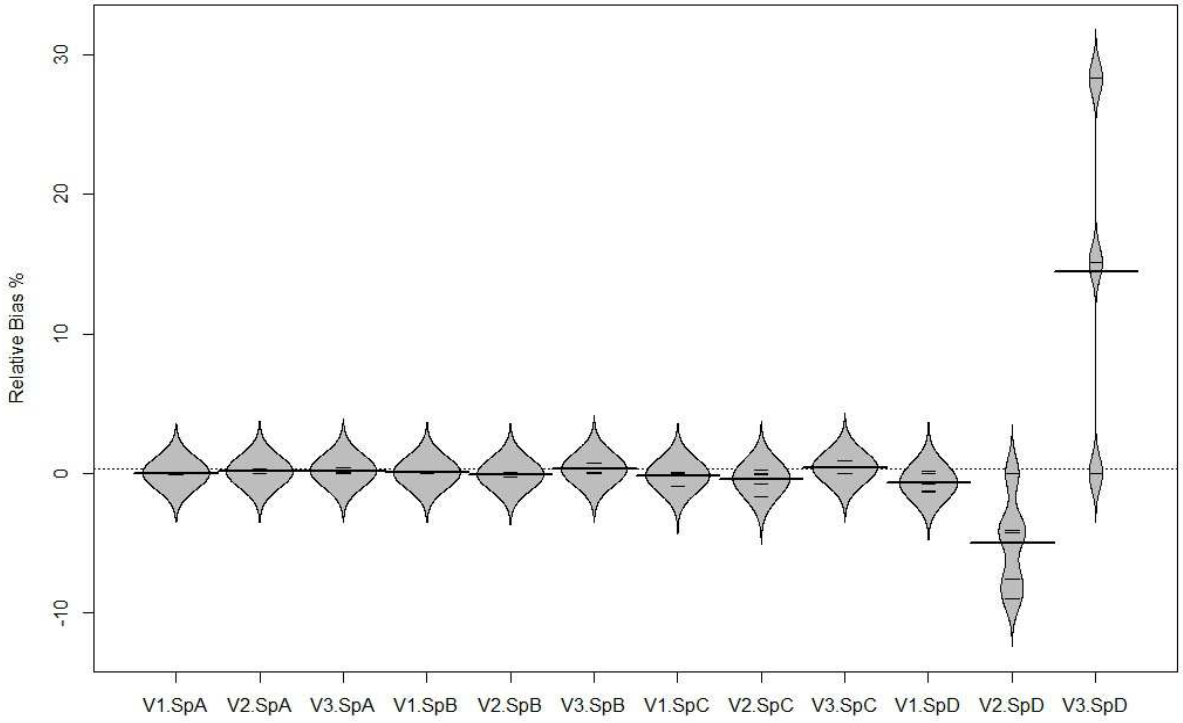


Figure 7-5: Beanplots of the relative bias of the estimates as function of the priors on the parameters *v* (V1 to V3) and species. The bold lines are the mean relative bias for each beanplot whereas the dotted line is the mean across all beanplots.

Increasing the misclassification probabilities increased the relative bias for all species. The standard deviations of the mean relative bias across priors on *v* increased when changing from low to high misclassification probabilities for species A to C whereas it decreased for species D (Table 7-2).

Table 7-5: Mean relative bias (%) and their standard deviation in brackets across the priors on *v* for each species. The different colours represent the different level of misclassification: No misclassification (white), low misclassification (light grey) and high misclassification (dark grey).

	SpA	SpB	SpC	SpD
Sc2.a	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Sc2.b	0.07 (0.5)	0.14 (1.18)	0.14 (3.32)	1.87 (27.01)
Sc2.c	0.03 (0.58)	-0.00 (1.55)	0.19 (2.91)	2.39 (25.11)
Sc2.d	0.16 (1.52)	0.26 (3.70)	-0.23 (9.94)	-4.37 (8.50)
Sc2.e	0.25 (1.25)	0.08 (3.04)	-0.42 (7.80)	-5.05 (7.35)

Unequal data in Models B

For Models A the Markov chains stopped updating for species D in models with classification scenarios Sc2.d and Sc2.e and with prior V3. For Models B this occurred for all four species but not for all replicates. In V1 andV2 in the models 30% of the replicates did not converge whereas in models with classification scenarios Sc2.e and prior P1 and 66% of the replicates did not converge. These replicates were not used in the rest of the analysis.

When the parameters *p* were estimated, rather than assumed known (i.e., Models B), the estimates of $\hat{\boldsymbol{v}}$ for all the species were significantly ($p<0.001$) influenced by the classification scenario, however with differences between the species (Figure 7-6). For species A and B, the largest bias occurred in scenario Sc2.e, which had high misclassification and asymmetric misclassification between species. The mean relative bias for species A for Sc2b,c,d together and Sc2.e was -1.30 (sd=0.71) and -16.0 (sd=6.63) respectively. For species B it was 1.95 (sd=1.25) versus -31.29 (sd=25.3) for Sc2.b,c,d all together and Sc2.e respectively. For species C and D differences between scenarios with high and low misclassification probabilities were less pronounced (Figure 7-6).

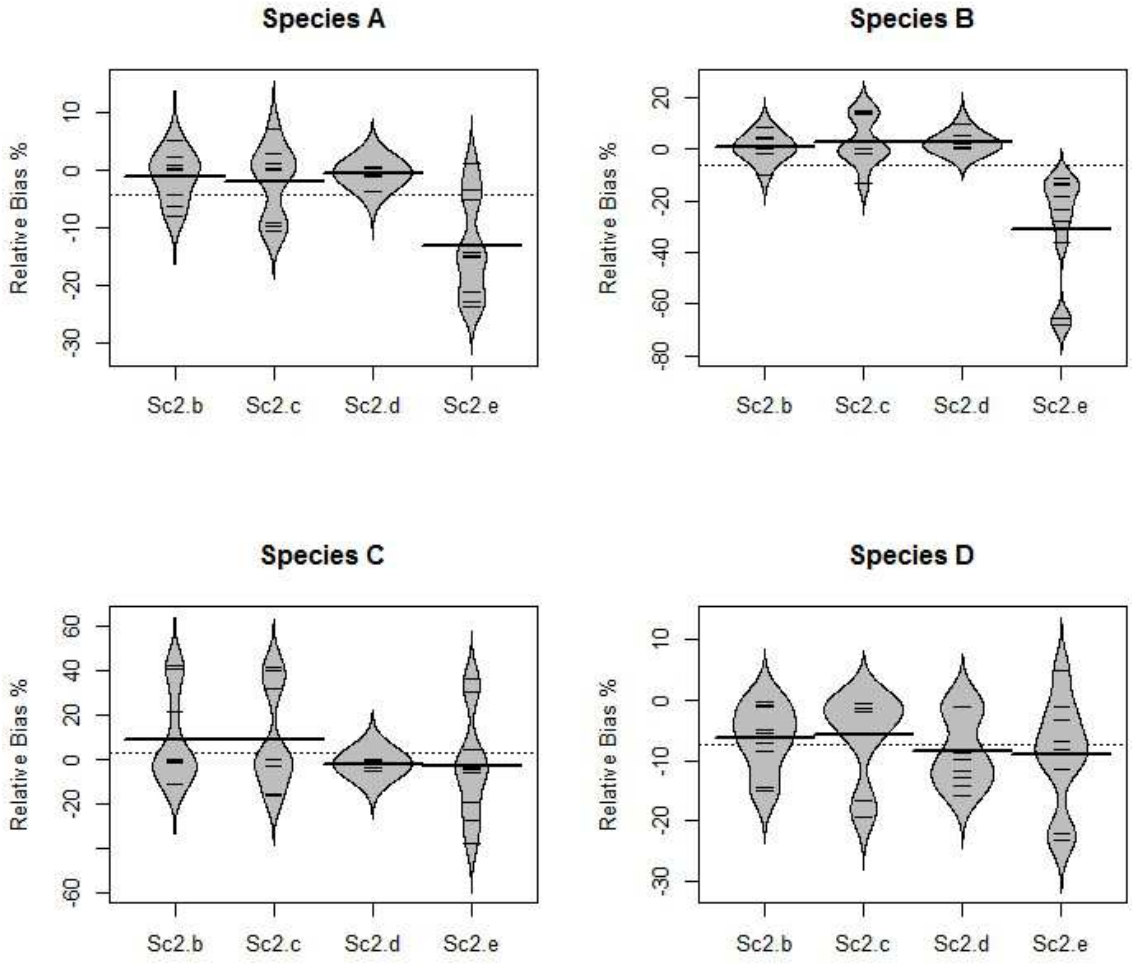


Figure 7-6: Beanplots of the relative bias for each species as function of the classification scenarios used in the Models B with unequal data. The bold lines are the mean of the relative bias for classification scenarios and the dotted lines are the average relative bias across the four classification scenarios. Each beanplot is computed from 9 values (three priors on p × three priors on v)

When Sc2.e was kept in the sensitivity analysis, no clear pattern was observable between species and the different priors (Appendix D fig D-1). However, once this scenario was removed, it was easier to detect the statistically significant ($p<0.001$) impact of the differences in prior variability on the accuracy of the estimates $\hat{v}_j$ for each species. Relative bias of the $\hat{v}_j$ appeared to be affected differently by the various prior variances, depending on the species (Figure 7-7): for species A to C, an informative prior on the classification probabilities (P1) decreased the effect of the priors on v, while for species D, the relative bias decreased with the informativeness of the prior on v from -0.82% to -7.50% with V1 and V2/V3 respectively independently of the informativeness of prior on the classification probabilities (P1).

When the priors on the parameters *p* were less informative (P2), the relative bias increased for species A to C and was influenced by to the variance of the prior on *v*. For species C the relative bias increased greatly and biologically significantly from -0.65% on average with P1 to 24% with P2, whereas for species D a less informative prior on *p* decreased slightly from -5% to -4% the overall relative bias between P1 and P2 respectively.

Finally with the vague prior P3, for species A to C the relative bias decreased in comparison to models with prior P2 and was sensitive to the *v* prior variability, whereas for the rarest species the relative bias doubled between model with prior P2 and models with prior P3.

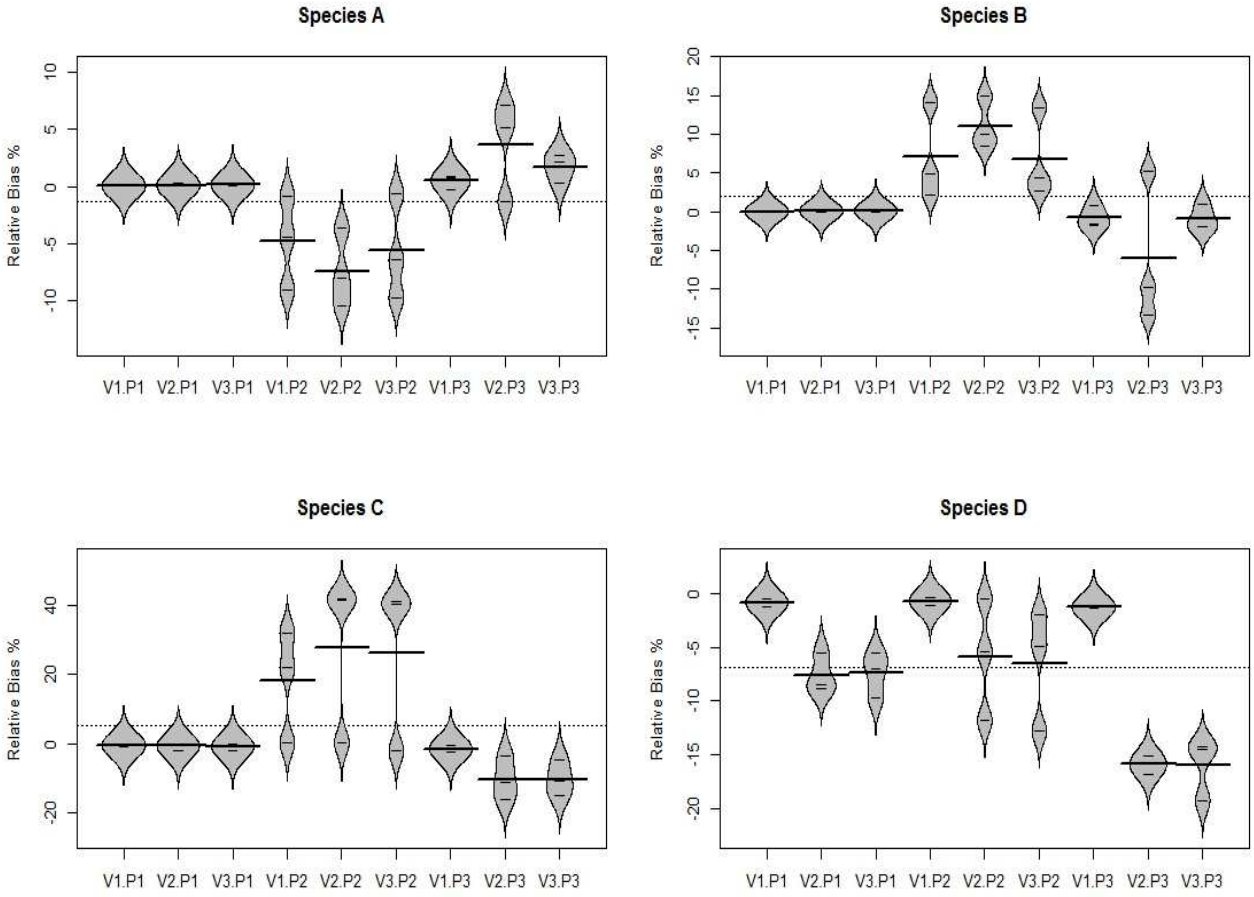


Figure 7-7: Beanplots of the relative bias distributions for Models B without classification scenarios Sc2.e with unequal data as a function of the *v* and *p* priors for each species. Each beanplot is made from 4 values (4 classification scenarios), the bold line being the mean relative bias of the bean plot. The dotted lines are the mean across all the prior combinations.

7.3.1.b.ii Precision of the estimated number of detections by species

The precision of the estimates was analysed by comparing the mean CV of the posterior distributions across replicate between models.

Equal data

The sensitivity analysis showed a statistically significant ($p<0.001$) impact of the degree of informativeness of the prior distributions on the precision of the estimates. For models with equal data and known *p* (Models A) or very informative prior on *p* (Models B, P1), statistically significant ($p<0.001$) but biologically insignificant differences (absolute difference of <2%), were observed between the mean of the CVs of models with informative (V1) or uninformative (V2) prior on *v* and between models with low or high misclassification probabilities (Table 7-6). When the variance of the prior on *p* increased (P2 and P3), the difference between models with informative (V1) and uninformative priors on *v* (V2) increased by a factor of 4. The difference between low and high misclassification probabilities were of a factor 8 for P2, whereas for P3 no difference were observed anymore between the mean CVs for the four classification scenarios (Table 7-6).

Table 7-6: Mean CV for models with scenarios Sc1.a to Sc1e with equal data priors V1 and V2 and Models A and Models B.

		Sc1.a		Sc1.b		Sc1.c		Sc1.d		Sc1.e	
		V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
Models A : P0		0	0	1.06%	1.06%	1.05%	1.05%	3.13%	3.34%	3.01%	3.18%
Models B	P1	0	0	1.89%	1.93%	1.87%	1.92%	3.76%	4.14%	3.58%	3.95%
	P2	0	0	1.16%	4.64%	1.45%	4.57%	8.72%	30.23%	8.72%	30.93%
	P3	0	0	8.75%	32.10%	8.73%	32.01%	8.77%	32.24%	8.78%	32.50%

Unequal data:

For unequal data, the same relationship was observed between the different priors and their effect on the CV of the estiamtes, but with the absolute values of the CVs being higher (23% versus 9% for equal data). The overall imprecision of the estimates $\hat{v}_j$ increased from 5.6% for Models A (P0) to 26% for Models B with P3. The precision was noticeably affected by the informativeness of the priors (Table 7-7). The differences on the mean CV between V1 and V2/V3 were <10% when no priors, prior P1 or prior P2 with low misclassification probabilities were used in the models. When the priors in the models were either P2 together with high misclassification probability or P3 these differences were greater than 20%.

The difference in mean CV between classification scenarios was small, when there was no prior on the *p* parameters (Models A) or prior P1 in Models B and was amplified with prior P2. Similar to the scenarios with equal data, these differences disappeared when prior P3 was included in the models (Table 7-7).

Table 7-7: Mean CV for unequal data for the four species for the different classification scenarios (Sc2.b to Sc2.c), priors on parameters *v* (V1 to V3) and no priors (Models A P0) or priors on parameters *p* (Models B P1 to P3).

	Sc2.b			Sc2.c		
	V1	**V2**	**V3**	**V1**	**V2**	**V3**
Models A : P0	4.42%	8.66%	7.06%	4.22%	7.45%	6.00%
P1	5.72%	12.25%	11.88%	5.61%	10.93%	10.98%
P2	10.09%	17.97%	16.51%	11.44%	15.60%	17.78%
P3	11.98%	36.64%	28.11%	12.11%	36.69%	34.58%
	Sc2.d			**Sc2.e**		
Models A : P0	6.70%	13.07%	4.53%	6.20%	12.52%	3.67%
P1	7.45%	13.90%	13.74%	2.24%	3.27%	11.77%
P2	11.65%	32.83%	27.23%	11.64%	32.48%	24.83%
P3	11.78%	39.52%	27.51%	12.01%	34.41%	26.57%

Comparing results across species, precision of the estimates was the lowest (higher CV) for species D, for all models (Figure 7-8).

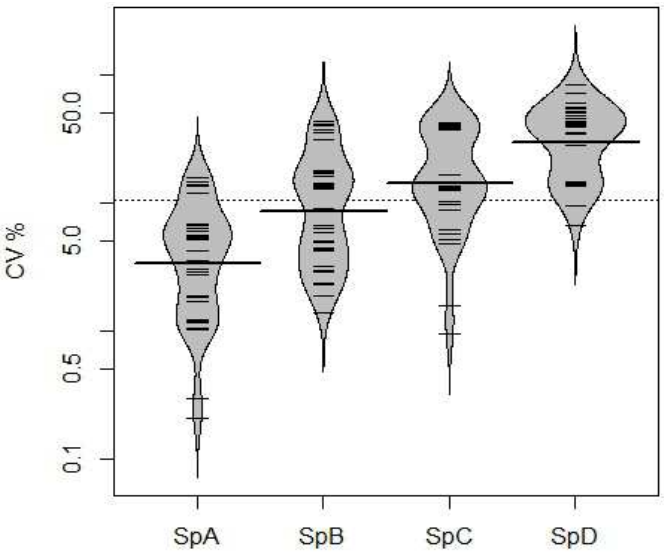


Figure 7-8: Bean plots of mean CV for all models as a function of species identities. The Y axis is displayed on the log scale.

7.4. Discussion

In this chapter, a Bayesian framework was developed to estimate the true number of detections from observed misclassified detections. The results showed that it was possible to estimate the true number of detections for each species and that bias and precision of the estimates depended on the prior information that was used to feed the models. This chapter highlighted that the uncertainty on the classification probability was the factor that generates most of the variability.

7.4.1. Consequences of unequal number of detections between species

This study showed that having an unequal number of detections between species generated bias and reduced the precision of the estimates. The estimates for species where less data were available were more sensitive to the prior variances and to the number of misclassifications.

7.4.2. Results of the prior sensitivity analysis

Different sources of variability were included in the models, namely different misclassification probabilities, uncertainty in the misclassification probabilities and uncertainty in the true number of detections between species. The sensitivity analysis demonstrated that among these different sources of variability, when each of them varied within the ranges observed with the real data of chapters 3 and 4, uncertainty within the classification probabilities was the most sensitive parameter. In the model without a prior for the parameters *p* (Models A) or when a very informative prior (P1) for *p* was chosen, the differences in the relative bias and in the mean CV between the two (for equal data) or three (for unequal data) priors on the parameters *v* were not biologically significant (Figure 7-5,Figure 7-7). However, if the prior variance on the *p* parameters was increased, the prior variance of the *v* parameters had a greater impact on the relative bias and the precision of the estimates. Models with prior V1 had a smaller absolute relative bias in comparison to models with V2 and V3. Model with prior V3 on the parameters *v* had a smaller absolute relative bias than models with prior V2. This result was unsurprising since in models with prior V3, the CV of the prior distributions were chosen to be different between species such that the total variance across the four species was smaller than for priors V2. The mean CV followed the same pattern.

The results for the relative bias of models with prior P3 on the parameters *p* were surprising. Estimates obtained with P3 were not the most biased despite prior P3 being the less informative prior. The parameters of this prior were such that the average classification probabilities were 25% for all species. The similar classification and misclassification probabilities explained why no differences in mean CV were observed between the classification scenarios, the source of variability being the same between species.

7.4.3. Impact of classification scenarios

The different classification scenarios had an impact on the relative bias of the estimates. However this overall result was most pronounced in classification scenario Sc.e which showed a large bias for all species, except species D, (Figure 7-6). This scenario simulated a confusion matrix with high and asymmetric misclassification probabilities. It is unclear why the relative bias and CV were so important for this scenario. Furthermore, more of the Markov chains for the replicates generated by the models under this scenario stopped

converging after several thousands of iterations than under other scenarios. To overcome this, a modification of the Metropolis Hasting algorithm in the analysis was introduced to try to propose new values for the Markov chain when it stopped converging. This modification did not solve the problem.

7.4.4. Rare species

Among the four species that were considered in the simulations, particular results were always derived for the rarest species D. The relative bias for the estimates of this species was more sensitive to the prior variances on the parameters *v* independent of the prior on the *p* parameters (Figure 7-7). The results for this species were also the most sensitive to the choice of the classification scenario (Table 7-5). Usually, a rare species will be difficult to observe and hence the prior knowledge on this species is likely to be vague (i.e. an uninformative prior should be chosen). Models A with prior V3 showed that when the prior CV was very high (60%) in addition to a high misclassification probability , it was not possible to derive an estimate with this Bayesian method as the chain converged towards zero and stopped updating. When the CV was reduced to 40% (Models B, V3), this problem was not observed. Nonetheless, for 30% of the replicates for this species, the Markov chains stopped converging.

7.4.5. Criticism of the model and conclusion

Overall, the results were highly sensitive to the initial values of the parameters and when these initial values were just over 20% away from the truth the models did not converge for some scenarios. This sensitivity can be explained by the fact that with this misclassification problem no unique set of solution exist. If there is not enough constraint on the priors and the initial values are far from the truth then the MCMC might converge but towards an estimate different to the reality (see next chapter). With such constraints it will not be possible to estimate the true number of detections if prior information on the true number of detections is totally absent and the prior on the classification rate is vague. This situation has more chance to happen for rare species for which it is difficult to collect any information. As the results shown, an important parameter which help in having more accurate and less bias results are the uncertainties around each classification rate. These measurements are mainly dependent on the data used to create the classifier and on the statistical methods used to measure them during the creation of the classifier and not on the abundance of the number of detections.

Finally, the parameters in this chapter were deliberately chosen with the specific intention to test the models under some extreme classification scenario or for cases of extreme prior variances. As such, they were not necessarily close to realistic values. Indeed, a CV of 40% was never observed in the classifiers developed in the previous chapter as long as the correct classification probability was sufficiently high. In the next chapter, the heuristic method of chapter 6 and the method of this chapter are used to estimate the true number of detections for the real data which is described in the first part of this thesis.

Chapter 8: Methods to estimate the number of acoustic detections in the presence of species misclassification applied to real data

8.1. Introduction

The output of the PAMGUARD whistle classifiers (PWC) used to identify the species in chapters 3 and 4 was a classification of acoustic detections organised in sections containing fragments of whistle contours. Due to the imperfection of the classifiers, some of the sections were misclassified. This chapter provide a demonstration of the three methods developed in chapters 6 and 7 with simulated data, and summarised in Table 8-1, with a selected subset of the real datasets and classifiers used in chapters 3 and 4 to estimate the true number of sections detected for each species.

Table 8-1: Summary of the methods used to estimate the true number of sections. For each method the type of confusion matrix (*C*) used in the models is described: PAMG. mean is the mean confusion matrix given by the PWC at the end of a classification process whereas PAMG. samples is the confusion matrices of each bootstrap of the classification process, Dirichlet dist. is the confusion matrices generated from a Dirichlet distribution. Initial values indicates whether the initial values are needed (Y) or not (N) for the method. prior on parameters *v* and *p* describe the parameters needed for the prior distributions on *v* and *p* in the Bayesian models.

Method name	Method description	*C*	Initial values	prior on parameters *v*	prior on parameters *p*
H1	Heuristic, known $\boldsymbol{p}$	PAMG. mean	N		
H2	Heuristic, estimated $\widehat{\boldsymbol{p}}$	PAMG. samples	N		
Models A	Bayesian known $\boldsymbol{p}$	PAMG. mean	Y	Mean, variance	
Models B	Bayesian, estimated $\widehat{\boldsymbol{p}}$	Dirichlet dist.	Y	Mean, variance	Alpha parameters

Among the five datasets of chapters 3 and 4, four classified with two classifiers are selected and analysed in this chapter

The first dataset is the French training dataset introduced in chapter 4. Since the data were used for classifier training, the true number of sections for each species is known and can be compared with the estimates provided by the heuristic and Bayesian methods. The second dataset is the MOR_BOWL training data introduced in chapter 3, which was used to identify the presence of bottlenose dolphins within wind farm sites. Similarly to the previous dataset, the true number of sections detected for each species is known (Table 4-4, *p*72), and it can be compared with the estimated true number of sections calculated with the different heuristic and Bayesian methods.

The third and fourth datasets are the EARs data also introduced in chapter 3, for which the species emitting the sounds detected is not known. These datasets were from the Moray Firth, which has been extensively surveyed for bottlenose dolphins (and other species). This meant that good information on the presence or absence of the classified species is known and thus could be used in the Bayesian models. Therefore, even though the results from these datasets could not be compared with truth, these datasets allow the consequences of misclassifying sections as species known to be rarely present in the study area to be investigated. The *5Sp* classifier, introduced in chapter 3, was used to generate the observed number of sections from these datasets.

The method section describes in detail the parameters needed for each of the four methods and the prior information available for each dataset. Estimates of the true number of sections are then obtained with the four methods, and a sensitivity analysis is conducted on the prior distributions of the true number of sections.

8.2. Methods

8.2.1. Heuristic methods

8.2.1.a Known classification probabilities (H1 method)

When the classification probabilities were known, the inputs needed to estimate the true number of sections $\hat{\boldsymbol{v}} = (\hat{v}_1, \dots, \hat{v}_j, \dots \hat{v}_m)$ for each species j were the observed number of sections $\boldsymbol{n} = (n_1, \dots, n_j, \dots n_m)$ for each species j and the mean confusion matrix ($\bar{C}$) of the classifier used to identify the sections:

$$\widehat{\boldsymbol{v}} = \bar{C}^{-1}\boldsymbol{n}$$ **(**chapter 6, Eq. 6-4, p100)

The confusion matrices used were the final confusion matrices given by the PWC (Table 8-9, Table 8-10).

8.2.1.b Estimated classification probabilities (H.2 method)

With the second heuristic model the classification probabilities were assumed to be estimated with some uncertainty. The inputs needed for this method were the observed number of sections $\boldsymbol{n}$ for each species *j* and several samples of the confusion matrix of the classifier used to identify the sections. In chapter 6, the confusion matrix samples were generated from a Dirichlet distribution. In this chapter, the data were classified with classifiers created with the PWC, so that at each bootstrap *b* of the classification process two confusion matrices were created (see chapter 2 for details of the bootstrap procedure). Each mean of these two confusion matrices ($\bar{C}_b$) were used as a confusion matrix sample. The estimate of the true number of sections was thus obtained by calculating the average of the observed data multiplied by the confusion matrix of bootstrap *b*, $\bar{C}_b^{-1}$:

$$\widehat{\boldsymbol{v}} = \frac{\sum_{b=1}^{B} \widehat{\boldsymbol{v}}_{\boldsymbol{b}}}{B} = \frac{\sum_{b=1}^{B} \bar{C}_b^{-1}\boldsymbol{n}}{B}$$

8.2.2. Bayesian methods

The parameters required to run both Bayesian methods were as follows.

1. observed number of sections for each species;
2. initial values of the Markov chains for all the model parameters;
3. parameters (mean and variance) for the prior distributions on the true number of sections *v*;
4. parameters (alpha) for the prior distributions on the classification probabilities *p* (for Models B only).

8.2.2.a Selection of initial values

Chapter 7 showed that the Bayesian models developed were sensitive to the initial values of the Markov chains. It was thus necessary to select an appropriate approach to generate these values such that they were not too far from the expected true values.

The initial values of the parameters $\boldsymbol{v}$ for Models A were obtained from the true number of sections estimated with the heuristic 2 method. However, the H2 method generated some negative estimates and it was thus not possible to use them as initial values. In this situation, the initial values of the species with negative estimates were set to the minimum value possible 1. The initial values of the parameters v in Models B were estimates of the true number of sections obtained with the Models A method. For Models B, the initial values of the parameters p were the classification probabilities of the confusion matrix of the classifier used to process the data.

8.2.2.b Parameters for the prior distribution on the *v* parameters

Given the real data used in this thesis, obtaining prior information on the expected true number of detected sections was difficult. Theoretically, the number of sections is dependent on the fragment and section length parameters of the classifier, on the average contour length and number of whistles per section, on the detection rate of the whistle detector, on the vocalisation rate and, finally, on the number of animals. For this chapter, different methods were used to give a value to the prior mean of the parameters v.

The prior means on the parameters v were estimated from prior knowledge obtained independently of the acoustic survey. Prior variances were then selected such that the coefficients of variation (CV) of the prior distributions were 40% and 10% similarly to the values used in chapter 7. The prior distributions in the Bayesian models were negative binomial distributions, consequently when the prior mean was small it was not possible to select a variance that allowed a corresponding CV of either 40% or 10%. When this situation happened, the variance of the prior distribution was selected such that the CV was the closest as possible to the desired CV. The sections below describe each dataset used in this chapter and how the prior means and variances have been selected.

8.2.2.b.i Data 1: French training dataset

For this dataset collected during the CODA survey, the abundances of all species identified by the classifier were estimated from visual detections (Table 8-2). From these estimated abundances the means of a first set of prior distributions for each species come from the proportion that each species contributed to the overall animal abundance. It was assumed that each species had similar vocalisation and detection rates.

Table 8-2: Abundances estimation from the CODA visual survey (CODA, 2009) for each species (BND (bottlenose dolphin), COD (common dolphin), C&S (common and stripped dolphin), STD (stripped dolphin), FPW (long and short finned pilot whale), and each classification group. The encounter proportion for each classification group is the abundance for that classification group as a proportion of the total abundance of the 5 species.

	BND	COD	C&S	STD	FPW	TOTAL
Abundance for each species	11536	56638	115398	33254	4857	84823
Abundance for each classification group	11536	68430			4857	84823
Encounter proportion	14%	80%			6%	100%

This data set was also used to train the French whistle classifiers (chapter 4) so the number of contours for each species before classification was known (Table 8-3). A second set of prior distributions were derived from these known numbers of contours. The number of contours for each species was converted to a proportion of the total of contours across all species, and these figures were applied to the observed number of section and they were used as the means of the prior distributions.

Table 8-3: Number of contours classified for each species and each classification group. The contour proprotion is the proportion of contours for a classification group relatively to the total number of contours of the 5 species.

	BND	COD	C&S	STD	FPW	TOTAL
Contours for each species	2	2164	110	247	842	3365
Number of contours for each classification group	2	2521			842	3365
Contour proportion	0.1%	74.9%			25.0%	100%

Then for each set of priors, two variance parameters were chosen such that the CV of the prior distribution was equal to 40% and to 10%. The prior parameters for these two set of prior distributions are summarised in Table 8-4.

Table 8-4: Mean and variance parameters (with associated CV in parentheses) of the prior distributions on parameters *v* for each species or classification group (CSD is common and striped dolphins). The number of observed sections *n* and the parameters α of the Dirichlet distribution for the prior distribution on the *p* parameters are also summarised.

		BND	CSD	FPW
Observed sections		**83**	**772**	**25**
prior from survey	mean	**123**	**704**	**53**
	Variance (40% CV)	2420.64	79298.56	432.64
	Variance (10% CV)	151.29	4956.16	27.04
prior from contour	mean	**1**	**659**	**220**
	Variance (40% CV)	1.01	69696	7744
	Variance (10% CV)	1.01	4356	484
Dirichlet parameters		2.93	0.10	0.115
		0.18	3.18	0.830
		0.09	0.21	0.055

8.2.2.b.ii Dataset 2: MORL_BOWL training data (chapter 3)

The training dataset of the ***5Sp*** classifier of chapter 3 was a concatenation of recordings made around the coast of Scotland during different independent surveys, so they were not associated with abundance estimates from visual detections. However, in chapter 3, the exact number of whistle contours for each species was measured. The number of contours for each species was converted to a proportion of the total contours across all species (Table 8-6), and these figures applied to the total number of observed sections were used as the means of the prior distributions (Table 8-5).

Table 8-5: Number of contours *n* classified for each classification group (bottlenose dolphin (BND), common dolphin (COD), Risso's dolphin (RSD), white beaked dolphin (WBD), white side dolphin (WSD). The contour rate is the proportion of contours for a classification group relatively to the total number of contours of the 5 species.

	BND	COD	RSD	WBD	WSD	TOTAL
***n* of whistles contours**	61934	69761	2554	5505	63584	146112
Contour proportion	42%	48%	2%	4%	4%	100%

Then for each set of priors, two variance parameters were chosen such that the CV of the prior distribution was equal to 40% and to 10% (Table 8-6).

Table 8-6: Mean and variance parameters (with associated CV in parentheses) of the prior distributions for each classification group (classification abbreviation similar to previous table) . Number of observed sections and the parameters α of the Dirichlet distribution for the prior distribution on the *p* parameters are also summarised.

		BND	COD	RSD	WBD	WSD
Observed *n* sections		**415**	**877**	**9**	**143**	**153**
prior parameters	mean	671	766	32	64	64
	Variance (40%CV)	*72038*	*93881*	*164*	*655*	*655*
	Variance (10%CV)	*4502*	*5867*	*32.01*	*64.01*	*64.01*
Dirichlet parameters		16.53	0.87	$2.5\ 10^{-3}$	0.62	0
		1.62	20.42	$2.5\ 10^{-3}$	11.01	5.04
		0.31	0.00	0.99	0.00	0.00
		0.52	3.43	$2.5\ 10^{-3}$	18.40	1.11
		0.12	1.69	$2.5\ 10^{-3}$	0.77	20.96

8.2.2.b.iii Dataset 3: Data recorded from the DO1 EAR in the Moray Firth S.A.C.

The DO1 deployment was in an area frequently visually surveyed to estimate abundance of bottlenose dolphins, harbour porpoises and harbour or common seals. Abundance estimates were available for these species. However, no abundance estimates were available for the other species used in the classifier. Only relative information such as the frequency of observations (i.e., seasonal, frequent or rare) were available (Moray Offshore Renewables ltd, 2010; Reid et al., 2003; Thompson et al., 2010). The prior distribution means were selected such that these observation frequencies were represented quantitatively. Within the S.A.C, bottlenose dolphins are common all year around, whereas sightings of common dolphins and white beaked dolphins are seasonal, and sightings of white sided dolphins and Risso's dolphins are rare (Hastie et al., 2003; Moray Offshore Renewables ltd, 2010). To match these observations, it was decided from the total number of observed sections that 90% of observed sections were bottlenose dolphins, 4% were common dolphins, 4% were white beaked dolphins, 1.5% were Risso's dolphins and 0.5% were white sided dolphins. These values were used as the means of the prior distributions for each of the species and two variance parameters were selected such that the CV of these prior distributions were as close as possible to 40% for one set of priors and as close as possible to 10% for the other set (Table 8-7).

Table 8-7: Number of observed sections detected by the DO1 deployment in the S.A.C, as well as the mean and variance parameters (with associated CV in parentheses) of the prior distributions on parameters *v* for each classification group (classification abbreviation similar to previous table).

		BND	COD	RSD	WBD	WSD
Observed sections		**94**	**15**	**5**	**16**	**1**
prior parameters	mean	118	5	5	2	1
	Variance (40%CV)	*2227.84*	*5.01*	*5.01*	*2.01*	*1.01*
	Variance (10%CV)	*139.24*	*5.01*	*5.01*	*2.01*	*1.01*

8.2.2.b.iv Dataset 4: Data recorded from EARs (E17,A20,A21) deployed in the MORL-BOWL wind farm sites

This dataset combined all the sections identified as vocalisations from dolphin species detected at the three EAR deployments E17, A20 and A21. In contrast to the Moray Firth S.A.C., bottlenose dolphins were rarely observed in the MORL_BOWL site whereas white beaked and common dolphins were the more frequent species visually detected. To match these observations, it was decided from the total number of observed sections that 0.5% of observed sections were bottlenose dolphins, 46% were common dolphins, 50% were white beaked dolphins, 2% were Risso's dolphins and 1.5% were white sided dolphins. These values were used as the mean of the prior distribution (Table 8-8) for each species, respectively. Two variance parameters were selected such that the CV of these prior distributions were as close as possible to 40% for one set of priors and as close as possible to 10% for the other set.

Table 8-8: Number of observed sections detected by the EARs deployed in the MORL_BOWL sites, as well as the mean and variance parameters (with associated CV in parentheses) of the prior distributions on the parameters *v* for each classification group (classification abbreviation similar to previous table).

		BND	COD	RSD	WBD	WSD
Observed sections		**15**	**351**	**1**	**72**	**35**
prior parameters	mean	2	218	9	237	8
	Variance (40%CV)	*2.01*	*7603.84*	*12.96*	*8987.04*	*10.24*
	Variance (10%CV)	*2.01*	*475.24*	*9.01*	*561.69*	*8.01*

8.2.2.c Parameters for the prior distribution on the parameters *p*

When all the datasets described above were used with the Models B method, the parameters of the prior distribution on *p* were selected such that they matched the classification probabilities and standard deviations of the confusion matrices from the given classifier used to classify the sections. For the first dataset the *3Sp* Spanish classifier was used to classify the French sections. The confusion matrix of the classifier is given in Table 8-9.

Table 8-9: *3Sp* Spanish confusion matrix, with the classification probabilities and their standard deviation (in brackets), discriminating bottlenose dolphins (BND), common and striped dolphins (CSD) and long and short finned pilot whales (FPW.)

	True Species		
Classified as %	**BND**	**CSD**	**FPW**
BND	**91.5** *(13.6)*	2.9	11.5
CSD	5.8	**91.1**(*13.4*)	83.0
FPW	2.8	5.9	**5.5**(*13.1*)

For the remaining three datasets the *5Sp* classifier of chapter 3 (Table 8-10) was used to classify the sections.

Table 8-10: Confusion matrix, with the classification probabilities and their standard deviation (in brackets), of the *5Sp* classifier discriminating bottlenose dolphin (BND), common dolphin (COD), Risso's dolphin (RSD), white beaked dolphin (WBD) and white sided dolphin (WSD).

	True Species				
Classified as %	**BND**	**COD**	**RSD**	**WBD**	**WSD**
BND	**86.6** (7.6)	3.3	0.0	2.0	0.0
COD	8.5	**77.3** (8.0)	0.0	35.8	18.6
RSD	1.6	0.0	**100** (5.9)	0.0	0.0
WBD	2.7	13.0	0.0	**59.8** (8.7)	4.1
WSD	0.6	6.4	0.0	2.5	**77.3** (7.9)

The alpha parameters of the prior Dirichlet distribution are in Table 8-4 and Table 8-6 for the 3Sp classifier and 5Sp classifier, respectively.

8.2.3. Description of the results

For each dataset, the following information is reported in a single table: (1) estimates of the true number of sections per species obtained with each method; (2) the prior parameters used;

(3) the mean and the CV of the estimates for the heuristic models; (4) the posterior means, CVs and 95% credible intervals for the Bayesian models.

8.3. Results

8.3.1. Dataset 1: French training dataset classified with the *3Sp* Spanish classifier

With the French dataset classified with the ***3Sp*** Spanish classifier, the first heuristic method (H1), estimated a very large number of sections for FPW and very large negative value for the other species (Table 8-11). By contrast, the H2 method estimates were all positive and relatively close to the truth (Table 8-11).

When the prior distribution means were chosen as a function of the species abundance estimated from visual surveys, the absolute relative error between the Bayesian estimates and the true number of sections ranged from 20% (for CSD) to more than 1000% (for BND). For these models, the presence of the prior distributions on *p* increased the CV of the estimates from 12.5 % to 13.5 % (Table 8-11). However, when the prior distribution means were based on the total number of contours, the absolute relative error for the estimates for each species decreased substantially, particularly for Models B, with an absolute relative error ranging from 0% for BND to 7% for FPW sections (Table 8-11). When no uncertainty in the confusion matrix was considered (Models A), the estimated number of sections attributed to CSD and FPW were significantly different ($p<0.05$) between prior distributions with a CV of 40% and a CV of 10%. In Models B, the 95% credible interval for the estimates overlapped between the two types of prior but the CV of the estimates was higher when the prior distribution CV was 40%.

Table 8-11: Mean, CV and 95% credible interval (CI) of the estimated true number of sections for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors on parameters *v* estimated from the visual survey (p. f.surv) or from the proportion of whistle contours per species (p. f. cont) with variance parameters such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species.

		BND			CSD			FPW		
		mean	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**
Truth		0			644			236		
Observed		**83**			**772**			25		
H1		-2144			-19824			22848		
H2		32	533.9		623	75.6		225	225.5	
A	**p. f surv. 40%**	60	8.7	50-70	773	2.4	730 - 803	47	42.4	16 - 94
A	**p. f surv. 10%**	71	6.3	62-80	759	1.0	743 - 774	50	14.2	37 - 65
B	**p. f surv. 40%**	89	21.0	52-131	741	3.5	684 - 785	50.0	38.1	20 - 94
B	**p. f surv. 10%**	115	9.6	96-139	712	1.8	685 - 735	53	12.9	40 - 67
A	**p.f cont 40%**	0			303	16.8	204 - 403	576	8.8	477 -676
A	**p.f cont 10%**	0			589	3.5	547 – 630	291	7.2	250 -333
B	**p.f cont 40%**	0			668	11.8	480 – 787	212	37.1	93 -399
B	**p.f cont 10%**	0		0	660	3.0	619 – 697	220	9.1	183-261

8.3.2. Dataset 2: Training data of 5Sp classifier

The estimates of the true number of sections for the five species with all the methods and models are summarised in Table 8-12. With the heuristic methods the estimate of the number of sections for the white beaked dolphin was negative.

With the Bayesian approach, when no uncertainty was considered in the confusion matrix (Models A), the true number of section estimates were close to the truth with an absolute relative error ranging from 0% to 2% for the species with most observed sections (BND and COD). However, for the rarest species, when the CV of the prior distributions was 40%, 0 sections of RSD and WBD were estimated and the Markov chain stopped updating (both CI and CV equalled 0). When the prior distribution was more informative, the number of estimated sections for RSD and WBD were 6 and 53 respectively.

When uncertainty in the confusion matrix was considered (Models B) and with the less informative prior (CV = 40%), the estimates of the true number of sections had a higher CV and absolute relative error values than the results from Models A. For all species, when the CV of the prior distribution was close to 10%, the Markov chains stopped updating after a few iterations. However, the mean of the posterior distributions for the two more common species (BND and COD) were very close to the truth before the chains stopped updating.

Table 8-12: Mean, CV and 95% credible interval (CI) of the estimated true number of sections for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species

		BND			COD			RSD			WBD			WSD		
		mean	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**
Truth		442			1031			4			22			98		
Observed		**415**			**877**			**9**			**143**			**153**		
H1		439			1069			2			-19			106		
H2		443	11.9		1105	23.3		1	6.0		-66	-3.5		114	76.4	
A	**Prior CV 40**	442	2.2	423 -461	1053	1.3	1025-1080	0		0	0			102	9.7	83-122
A	**Prior CV 10%**	442	2.2	424 -461	1011	1.3	985 -1037	6	21.0	4 -8	53	12.8	40-66	85	7.8	72-98
B	**Prior CV 40%**	452	13.1	343 -582	1004	6.9	860 -1135	9	63.4	0 -28	59	39.2	22-112	72	38.2	26-132
B	**Prior CV 10%**	445	0	445 -445	1031	0	1031-1031	6			54			82		

8.3.3. Dataset 3: Data recorded from the DO1 EAR deployment in the Moray Firth S.A.C

The estimates of the true number of sections for the five species of this dataset with all the methods and models are summarised in Table 8-13.

With the heuristic methods, the estimates for Risso's, white beaked and white sided dolphins were very imprecise with CVs ranging from 70% to 614%.

The Bayesian Models A estimated that 123 sections contained BND contours and the 8 remaining sections contained WBD contours. With this model, no other species were selected in the classification process. With the Bayesian Models B, all the sections were estimated to contain BND contours and, after moving from the initial values, the Markov chains stopped updating.

Table 8-13: Mean, CV and 95% credible interval (CI) of the estimated true number of sections detected by the DO1 deployment, for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species

		BND			COD			RSD			WBD			WSD		
		mean	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**
Truth		Unknown			Unknown			Unknown			Unknown			Unknown		
Observed		**94**			**15**			5			16			1		
H1		108			-3			3			23			0		
H2		110	10.1		9	7.16		3	70.4		10	590		-2	614	
A	**Prior CV 40%**	123	1.6	119-126	0		0-0	0		0-0	8	22.9	5-12	0		0-0
A	**Prior CV 10%**	131	0	131-131	0		0-0	0		0-0	9	22.2	5-12	0		0-0
B	**Prior CV 40%**	131	0	131-131	0		0-0	0		0-0	0		0-0	0		0-0
B	**Prior CV 10%**	131	0	131-131	0		0-0	0		0-0	0		0-0	0		0-0

8.3.4. Dataset 4: Data recorded from EARs (E17,A20,A21) deployed in the MORL-BOWL wind farm sites

The estimates of the true number of sections for the five species of this dataset with all the methods and models are summarised in Table 8-14.

With this dataset, the H2 method estimated a negative number of sections for the WBD species only, but the CVs for BND and WSD were very high due to numerous estimates with negatives values for these species as well.

With the Bayesian Models A, no sections were estimated to contain contours from BND, RSD or WSD. The estimates of the number of sections attributed to the remaining two species were significantly different ($p<0.01$) between Models A with a prior CV of 40% and Models A with a prior CV of 10%.

When a prior on the parameters *p* was added to the models (Models B), the estimate of the number of sections containing contours from RSD ranged between 1 and 13. Similarly to Models A, there was a significant difference between the estimates of the number of sections using priors on *v* with CV of 10% and CV of 40% for section attributed to both COD and WBD. For all Bayesian models, when the mean of the posterior distributions was zero, the Markov chain stopped updating.

Table 8-14: Mean, CV and 95% credible interval (CI)l of the estimated true number of sections detected by the EARs deployed in the MORL_BOWL sites, for the three species classified with the heuristic methods (H1 and H2) and with the Bayesian Models A (A) and Models B (B), with priors variance on parameters *v* such that the CV was 40% or 10%. The observed number of sections from the classifier results (Observed) and the true number of sections (Truth) from the training dataset are also reported for each species

		BND			COD			RSD			WBD			WSD		
		mean	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**	**mean**	**CV%**	**CI**
Truth		Unknown			Unknown			Unknown			Unknown			Unknown		
Observed		**15**			**351**			1			72			35		
H1		0			441			1			25			8		
H2		1	1795		483	31.5		1	40.3		-15	-948.6		3	1048	
A	**Prior CV 40%**	0		0-0	435	2.4	414-454	0		0-0	39	26.2	20-60	0		0-0
A	**Prior CV 10%**	0		0-0	353	2.7	334-372	0		0-0	121	8.0	102-140	0		0-0
B	**Prior CV 40%**	0		0-0	368	7.4	310-418	2	124.9	1-10	104	26.1	54-162	0		0-0
B	**Prior CV 10%**	0		0-0	258	6.2	227-290	4	99.1	1-13	212	7.6	180-244	0		0-0

8.4. Discussion

Applied to real data, the limitations of the heuristic methods were clearly demonstrated with negative estimates and/or unrealistic estimates e.g., 22848 sections were predicted for pilot whales in the French dataset, although only 880 sections were classified in total (Table 8-11). However, with the H2 method, when the estimates were not negative, they were relatively close to the truth. Of all the species classified (across the two classifiers) for which the truth was known, six had a relative error smaller than 10%, two had a high relative error due to either a small estimate (RSD) or because the true number of sections was zero (BND in French data) and so relative error was not measurable. As shown in the previous chapter, the Bayesian models were sensitive to the choice of the initial values of the Markov chains. When the initial values were too far from the truth, the Markov chains did not converge. Therefore, the decision of using the estimates from the H2 method was reasonable.

The results showed a clear negative impact on the estimates of a wrong or too uninformative prior. With the French dataset, the true number of sections estimated when the prior means were based on the abundance estimates of individuals was far from the truth. This difference mainly affected the estimates for bottlenose dolphins; in reality, no sections were from bottlenose dolphins but the Bayesian models estimated that between 60 (Models A, CV: 40%) and 115 (Models B, CV: 10%) sections were attributed to this species. On the other hand, when the prior means were based on the proportion of contours detected, the estimates of the true number of sections were closer to the truth when the uncertainty of the confusion matrix was included in the models (Models B). An explanation for the poor estimation using the first set of priors (based on abundance) is that unrealistic assumptions were made in order to link the number of individuals to the number of contours i,e., same vocalisation and detection rates between species were assumed. The prior means were probably too far from the truth to be able to give accurate estimates of the true number of sections. For the data from which the truth was known, the estimates generated by Models B were slightly less biased and also less precise than the estimates generated by Models A.

For the last two datasets for which the truth was unknown, the estimates reflected the expectation, modelled by the selection of the prior means, of the presence and absence of some species in the monitored area. Indeed, in the S.A.C. where the DO1 recording device was deployed, bottlenose dolphins were expected to be the predominant species. The Bayesian Models A estimated that few sections (8) were attributed to white beaked dolphins,

which were occasionally observed in the area. For the three other Bayesian models, the estimates predicted that all the sections detected were attributed to bottlenose dolphins.
In contrast, at the wind farm site, bottlenose dolphins were expected to be rare and none of the Bayesian models estimated that any section was produced by this species.

With regards to the results for the datasets where truth was known, it can be assumed that the estimates from Models B with a CV prior of 10% were probably the best estimates. However, it is important to keep in mind that these results are reliable only if the prior means on the parameters *v* were estimated accurately.

For all datasets, it was observed that each time the posterior mean was zero, the Markov chains stopped updating. This phenomenon was also observed in two other situations: under the Models B with the second dataset where the CV of prior distribution on *v* was close to 40%, and with the third dataset each time the estimates of the BND sections reached the true values. The stopping movement within the Markov chain was a consequence of the multinomial function used to update the parameters in the Metropolis Hasting (MH) function. The probability parameters of the i^{th} multinomial update function at iteration *t* of a Markov chains were dependent to the $y_{i.}$'s parameters (Part II.7.2.6.a, p121, Eq 7-7) of iteration *t-1*. If at iteration *t-1* one of the $y_{i.}$'s parameters was zero, consequently the probability of the multinomial distribution corresponding to this y_{ij} become also zero and so it can only propose new zero values at iteration *t*. The stopping of updating when the posterior mean was not zero was due to a very slow mixing, which can be due to an inappropriate proposal function. One potential solution is to use another proposal distribution such as a random walk when slow mixing was detected. Initially a random walk was used, but once priors on parameters *p* were added, the models were not updating.

The real data highlighted a limitation of the negative binomial distribution for the prior on the v parameters. When the true number of detections was predicted to be small, the variance needed to reach the CV wanted was smaller than the mean and so not possible to use it with the negative binomial prior. The implementation of a Conway-Maxwell-Poisson distribution (Conway and Maxwell, 1961) which allows for both under and over-dispersed data would have allowed to solve this issue.

In conclusion, this chapter demonstrated that the Bayesian models used to estimate the true number of sections were reliable when appropriate prior distribution means of the model parameters were used. Having an informative prior improved the precision of the estimates in comparison with the use of an uninformative prior, but, if the mean distribution of the prior on the true number of detections was completely inappropriate, the estimates of the true number of sections or detections will be unreliable even with an informative prior. This chapter shows that even if there is no prior information on the absolute abundance, relative abundance between species present in the area of interest and used in the classifier is good enough to be able to estimate the true number of detections which will can then be used to estimate absolute abundance when other parameters such as cue rates, detection rates will become available.

Chapter 9: Dealing with species misclassification: General discussion

It will never be possible to create the perfect classifier with the ability to identify whistles without error, so the next logical step is to develop methods able to estimate from the misclassified observations the true number of whistle detections for each species. The objectives of the last three chapters (6 to 8) were to: (1) find a reliable method to estimate this true number of detections in the presence of misclassification and (2) to identify those factors that most influenced the accuracy and precision of these estimates generated.

The heuristic methods used in chapter 6 were simple, intuitive but probably not optimal whereas the Bayesian methods of chapter 7 were more difficult to implement, less intuitive but gave better results. For these two chapters the data were simulated whereas in chapter 8 these two methods were applied to real data. With the heuristic methods, some estimates of the true number of detections for both the simulated and the real data were negative. Negative values are obviously not possible when trying to quantify a number of detections.
Both methods identified that the proportion of detections by species, the misclassification probabilities and the uncertainty of these misclassification probabilities had the greatest influence on the accuracy and precision of the estimates. However, the relative importance of these factors varied between methods.

9.1. Equal versus unequal detections between species

When the number of detections was high (3000) and similar between species (equal data) no bias was observed between the expected true number of detections and the estimated numbers, whatever the statistical approach and the parameters used in the models. In the heuristic models when the true number of detections was different between species (unequal data), no bias was observed between the estimates and the truth. However, in the Bayesian models relative biases ranging from 0.1% to 40% were observed when uncertainty in the confusion matrix was associated to unequal data.

With equal or unequal data the variance of the estimates was affected by the classification probabilities, uncertainty of the classification probabilities for all methods and prior

knowledge of the true number of detections for the Bayesian method. With the heuristic method, the CV of the estimates reached unreasonably high values (>400%) for rare species even with a low misclassification probability and a small uncertainty in the confusion matrix, whereas in the Bayesian model the highest CV observed for an estimate was 70% when a high misclassification probability and high level of uncertainty were simulated.

9.2. Prior sensitivity

In the Bayesian models, there were two random variables (the true number of detections *v* and the classification probabilities *p*) that required prior distributions. In general the estimates of the true number of detections were sensitive to the prior variances. When these variances increased, the precision of the estimates decreased: for example with both the simulated data (chapter 7) and the real data (chapter 8) the CV of the estimates were for the most part lower when the parameters of the prior distribution on the parameters *v* were such that the CV was 10% instead of 40%. In the scenario with low misclassification probabilities and with equal numbers of detections between species, both prior variances on *p* and *v* had a similar impact on the CV of the estimates. However when more misclassification was added to the models, and the number of detections between species were unequal, increasing the variance of the prior of *p* had a bigger impact that increasing the prior variance of *v* (Table 7-4, p126).

The prior variance on the parameters affected also the accuracy of the estimates. When this variance was equal to zero or small such that the prior CV was 10%, the bias for all species was zero (with heuristic methods) or small (Bayesian method) and insignificant for all practical purposes. The example of the French data in chapter 8 when the prior means were based on the abundance of a species showed that the mean of the prior was also a very important parameter to obtain unbiased estimates. With the simulated data the scenario where the prior means were intentionally different to the truth was not tested. However with the simulated data the situation where the prior means of the parameters *p* were far from the expected truth was tested. The results showed these priors had an impact on the accuracy of the estimates (Table 8.11).

9.3. Misclassification probabilities

Similarly to the influence of the prior on the *v* parameters, when the classification probabilities were considered as known (heuristic and Models A) or with small uncertainty

(Models B with prior P1) the different misclassification probabilities had no impact on the relative bias of the estimates (Figure 6-1 and Figure 7-7). On the other hand when larger uncertainty on the classification probabilities was simulated, then increasing the misclassification probabilities decreased the accuracy and precision of the estimates. Particularly high misclassification associated with asymmetric misclassifications (Sc2.e) between species generated the largest bias observed of all modes.

9.4. Grouping species, an alternative to decrease misclassification rates

Given the general availability of cetacean species and the cost and time necessary to obtain data, obtaining more precise information regarding the true number of detections can quickly become challenging and costly. Chapter 8 showed that using information of relative abundance is a good alternative to obtain reliable estimates of the true number of detections. However, improving the output of the classifier by decreasing the misclassification rates and their associated uncertainties depends on the training dataset quality and also on the method used to develop the classifier. As shown in chapters 3 and 4 grouping different species in one classification group can improve the general classification rates. Such grouping systems need to be used suitably and generally to answer a management or conservation concern. In chapter 3, species were grouped to answer a management problem question which was to identify the protected bottlenose dolphins from all the other species encountered in the same area. By grouping the species in two groups the classification results were greatly improved. Differently in chapter 3, common dolphin and spotted dolphins were grouped because of their very close acoustic characteristics generating a high level of misclassification between this two species. Given the objective of the CODA survey grouping these two species was not a problem and it decreased the level of misclassification in the classifier.

9.5. Rare species

For all methods and models, the estimates of the simulated rare species had a larger CV and a larger bias (with Bayesian models) than the other species. The data were simulated in such a way that it was not possible within the scope of this thesis to determine if these results were an artefact of unequal detections or just a consequence of a small number of detections. To distinguish between these hypotheses, models could be tested with 50 detections per species rather than the 3000 used here. However, it is more realistic to expect, in the real world, to encounter a situation similar to that simulated in this thesis. Given the results of these three

chapters, the benefice of using acoustic survey over the visual survey, for rare species will be mainly dependent on the vocal characteristics and vocalisation rates of the species. Indeed if the rare species vocalise regularly and can be clearly discriminated acoustically, such that it is possible to develop a reliable classifier with a low misclassification probability for this species then, it can be hoped that using acoustic detections will improve the accuracy of the abundance estimate for this species. On the other hand if the rare species is difficult to discriminate acoustically, as well as difficult to detect visually, then using acoustic detection may not be useful to improve its abundance estimation. The problem of rare species is recurrent for all detections method used. In several studies (McClintock et al., 2010a; Miller et al., 2011; Royle and Link, 2006) which tried to deal with species misidentification, a common conclusion was that when species misidentification is considered in the model the largest bias on the abundance estimate occurred when the occupancy probability is low.

9.6. Limitations of the methods

Both approaches showed their limits when the number of detections was small for a given species: with the heuristic method unrealistic estimates were predicted and with the Bayesian method the MCMC frequently stopped updating when the estimates of the true number of detections were zero. Furthermore the values of the confusion matrices had been selected such that it was possible to analyse the impact of the misclassifications rates and their uncertainty independently. The confusion matrices of the classifier created in the first part of this thesis as well as the confusion matrix of Gillespie et al., (2013), never had a high correct classification probability associated with a CV of 40% as it was simulated. In a further work, confusion matrices with different correct classification probabilities between species associated with a low CV for a high correct classification probability, and a high CV for a low correct classification probabilities can be tested. Given all the observed results more accurate and precise results are to be expected for the species with high correct classification probability and vice-versa.

Finally, estimating the true number of detections from misidentified data is not a problem specific to unidentified cetacean acoustic cues. The problem of species identification is also present with visual detections and with species other than cetaceans (McClintock et al., 2010b; Miller et al., 2011). Species misidentification (from visual survey) or misclassification (when identification via a classifier) generates false positive detections. In occupancy and

abundance estimation model the impact of false negative errors has been widely analysed and method to decrease the bias it can generate on the final estimate have been largely developed (Buckland et al., 2004; MacKenzie et al., 2002). However the problem of false positive detections due to misidentification has been ignored for a long time, despite demonstration that such errors occurred even with experimented observers (McClintock et al., 2010b; Simons et al., 2007). In their studies McClintock et al., (2010a), Miller et al., (2011), Royle and Link (2006) have demonstrated that false positives detections rapidly lead to misleading inferences. With cetacean surveys, false positive detection errors caused by misidentification from visual observations have always been ignored. Generally with cetacean acoustic, a parameter within the abundance formula includes false positives detections rates ((Marques et al., 2009; Thomas and Marques, 2012). To refer to the equation in this thesis, the misclassification parameter was called $\hat{c}$ in equation (6-1 (p98). Nonetheless this parameter is general and represents the probability that the detections are misclassified as another sound not species specific. It does not acknowledge the misclassification between species and its consequences on the misleading observed data. In anuran studies for which it is easier to detect false positive detections they are developing methods to measure the bias generated by such species misidentification on the final abundance estimation (McClintock et al., 2010a; Miller et al., 2011; Royle and Link , 2006). These studies focus either on misidentification between two species only or they were done in a very controlled system. The conclusions of this PhD with the consequences of misclassification with more than 2 species and a less controlled system are similar to the conclusions of the anurans studies. These similar conclusions being that the level of uncertainty of the species identification as well as the level of species concurrency played the major role on the bias and accuracy of the estimates.

9.7. Abundance estimation using misclassified observed detections

It is important to keep in mind that the true number of detections is only one variable in the process of estimating abundance from acoustic detections, and consequently it is not the only parameter responsible for the accuracy and precision of the abundance estimates. As expressed in Eq 9.1, at least two other parameters in the abundance equation need to be estimated: the cue rate ($\hat{r}_j$) and the detection probability ($\widehat{P}_j$).

$$\widehat{N}_j = \frac{\hat{v}_j}{aT\widehat{P}_j\hat{r}_j}A \quad (9.1)$$

Both these parameters are species dependent and can be challenging to estimate. As mentioned several times in this thesis, the cue rate is largely unknown for most of the whistling species and it is likely to be highly variable. The average probability of detection in itself is also dependent on numerous factors (such as distance from the hydrophone, directionality of the call, ambient noise, and detector performance).

If all these estimates are considered as independent then the precision of the abundance estimate ($\widehat{N}_j$) can be calculated by using the delta method (Gerrodette et al., 2011; Seber, 1982).

$$CV^2\left(\widehat{N}_j\right) = CV^2\left(\hat{v}_j\right) + CV^2\left(\widehat{P}_j\right) + CV^2\left(\hat{r}_j\right)$$

The CV of the true number of detections estimates is thus only one element of the overall CV of the abundance estimate. Its influence on the final abundance estimate can only be considered relative to the CV of the other estimates. Indeed if for example the CV of the estimated true number of detections is 70% (as the highest CV observed with simulated data in chapter 7) and the CVs of the cue rates and detection probability are 10% then the CV of the abundance estimate will be mainly influenced by $\hat{v}_j$. To improve these estimates this thesis showed that one solution is to improve the classification process, so that as the correct classification probability increases and the uncertainty around this rate decreases. Another solution is to have a robust method to estimate the true number of detections. While, if the contribution of the true number of detections CV is not important relative to the other parameters then more effort should be taken in improving the estimation of the cue rates and detection probabilities. However having biased estimates of the true number of detections is a more important problem than imprecise estimates, as in this situation the abundance estimate will also be biased and that can lead for example into inappropriate management, conservation decisions.

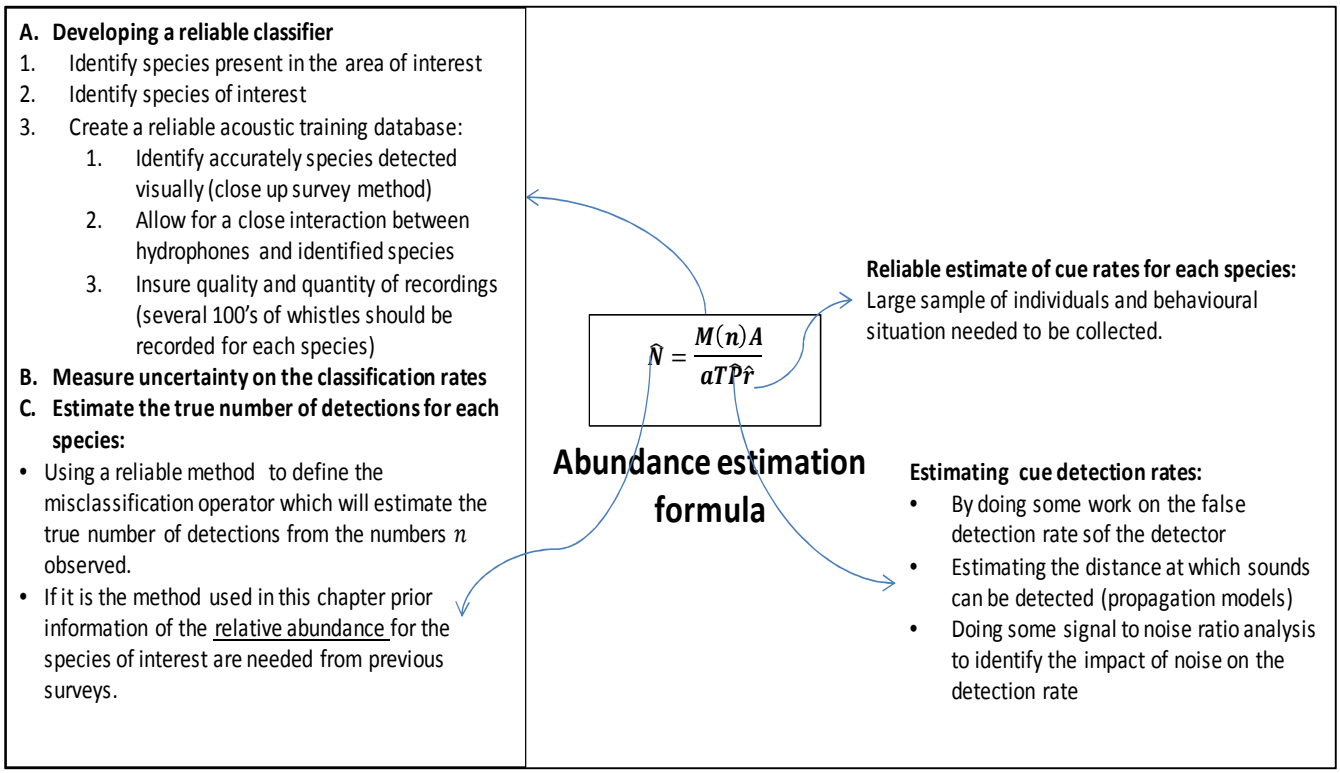
One advantage of the Bayesian framework developed in this thesis is that it was possible to quantify how much bias and variance the parameters used in the Bayesian models generated on the estimates. By incorporating the other parameters of the abundance equation in a Bayesian framework we can hope that it will be possible to identify for future surveys or projects those parameters of the abundance equation that generate most of the bias and

uncertainty, and once identified it will be easier to find solutions to improve the measurement. The Bayesian method developed in this thesis is a first approach and has its limitations. A priority will be to improve the method such that the Markov chains become less dependent on the initial values and the issue of lack of convergence should be solved.

9.8. Conclusion

In summary and conclusion this thesis highlighted more questions and problems to be solved than bringing complete solutions to estimate abundance of cetaceans solely from acoustic detections. Through the comparisons of the results of chapter 3 and 4 and from the results of the second part it is possible to suggest some methods which should help in the process of obtaining reliable abundance of cetacean using their acoustic signatures. A critical point is the correct identification of the sound detected which can be done by an automatic classifier. The creation of a reliable classifier with the quantification of the uncertainty for each classification rate has been shown to be very important. Box 3 summarises the important steps associated with some applied methods to reach this objective. The most important step is the availability of a reliable training dataset ideally without any misclassification. Using survey methods which allow a close interaction between the hydrophones, the animals and the observer should help to accumulate such a dataset. Then to be able to use the classifier outputs, having a measure of the uncertainty of this output is indispensable to be able to measure the bias and precision of the estimate of the true number of detections. Including in the classifier only species present in the area of interest and/or grouping, when possible, species with a high chance of misclassification in the same classification group will help to decrease the overall misclassification rate.

Nonetheless before being able to estimate abundance, parameters such as the cue rates and the cue detection rates needed to be estimated. The cue rate particularly is very difficult to obtain due to its high variability within and between individuals. Consequently a large sample size covering numerous different individuals and numerous behaviours is needed to obtain a reliable cue rate estimate. Cue detection rates can be estimated using propagation models, once the frequency and the source level of the sounds are known, associated with false positive detection analysis.



Box 2: Summary of the different parameters needed to estimate abundance form acoustic detection with suggestion of some method to obtain them.

Having uncertain estimates in itself is not a problem as given the complexity of biological models it will never be possible to have models representing a biological system without uncertainty. So ecologists often confront uncertainty and must try hard to identify the sources of uncertainty, how to quantify it and what are the consequences on the output of the model. Policy makers or environmental managers are now totally aware that it is impossible to ask for certain output, estimates and predictions. Large management programs such as the Revised Management Procedure (RMP) of the IWC have been developed to establish catch limit quotas to protect the stock of some species (Punt and Donovan, 2007). In this management procedure measurement of uncertainty is part of the models used to simulate the impact of the management decisions on the species stock of interest. More generally once uncertainty is identified and most importantly quantified, appropriate management options or policies can been established with more confidence (Ascough II et al., 2008; Harwood and Stokes, 2003). The managers or policy makers will be able to quantify the risk their decisions

create in a given situation and consequently to modify their strategy if this risk is not acceptable.

In the model used to estimate abundance from acoustic detections, this thesis only identifies and quantifies which parameters in the species identification process are responsible for most of the uncertainty of the estimate. These findings need to be implemented in the more complete and complex model of abundance estimation.

Finally, although this thesis focused only on whistling species, the problem can be easily extended for all species for which acoustic surveys are used to estimate abundance or for any problem of misclassification/ misidentification between species.

References

Akaike, H. (**1974**). “A new look at the statistical model identification,” IEEE Trans. Autom. Control, **19**, 716 – 723. doi:10.1109/TAC.1974.1100705

Akçakaya, H. R., Ferson, S., Burgman, M. A., Keith, D. A., Mace, G. M., and Todd, C. R. (**2000**). “Making Consisten IUCN classifications under Uncertainty,” Conserv. Biol., **14**, 1001–1013. doi:10.1046/j.1523-1739.2000.99125.x

Albert, J. (**2009**). *Bayesian Computation with R*, Springer, Second., 298 pages.

Ansmann, I. ., Goold, J. C., Evans, P. G. H., Simmonds, M., and Keith, S. G. (**2007**). “Variation in the whistles characteristics of short-beaked common dolphins, Delphinus delphis, at two locations around the British Isles,” J. Mar. Biol. Assoc. United Kingd., **87**, 19–26.

Ascough II, J. C., Maier, H. R., Ravalico, J. K., and Strudley, M. W. (**2008**). “Future research challenges for incorporation of uncertainty in environmental and ecological decision-making,” Ecol. Model., **219**, 383–399. doi:10.1016/j.ecolmodel.2008.07.015

Au, D., and Perryman, W. (**1982**). “Movement and speed of dolphins schools responding to an approaching vessel,” Fish Bull US, **80**, 371–379.

Barlow, J. (**1988**). “Harbor porpoise, Phocoena phocoena, abundance estimation for California, Oregon and Washington: I ship surveys,” Fish Bull US, **86**, 417–432.

Barlow, J., and Forney, K. A. (**2007**). “Abundance and population density of cetaceans in the California current ecosystem,” Fish. Bull.,.

Barlow, J., and Taylor, B. L. (**2005**). “Estimates of sperm whale abundance in the northeastern temperate pacific from a combined acoustic and visual survey,” Mar. Mammal Sci., **21**, 429.

Bates, D. M., and Watts, D. G. (**1988**). *Nonlinear Regression Analysis and Its Applications*, Wiley, 1st ed., 384 pages.

Baumgartner, M. F., Van Parijs, S. M., Wenzel, F. W., Tremblay, C. J., Esch, H. C., and Warde, A. M. (**2008**). “Low frequency vocalizations attributed to sei whales (Balaenoptera borealis),” J. Acoust. Soc. Am., **124**, 1339–1349. doi:10.1121/1.2945155

Bayes, M., and Price, M. (**1763**). “An Essay towards Solving a Problem in the Doctrine of Chances By the Late Rev Mr Bayes, F R S Communicated by Mr Price, in a Letter to John Canton, A M F R S,” Philos. Trans., **53**, 370–418. doi:10.1098/rstl.1763.0053

Boisseau, O., Gillespie, D., Leaper, R., and Moscrop, A. (**2008**). “Blue (Balaenoptera musculus) and fin (B physalus) whale vocalisations measured from northern latitudes of the Atlantic Ocean,” J Cetacean Res Manage, **10**, 23–30.

Borchers, D., Brewer, C., and Matthews, J. (**2007**). *Methods for estimating sperm whale abundance from passive acoustic line transect surveys* (Technical report No. 2007-3), . Retrieved from http://www.creem.st-and.ac.uk/len/onr/onr-sperm-whale.pdf

Borchers, D. L., Buckland, S. T., and Zucchini, W. (**2004**). *Estimating Animal Abundance*, Springer, 330 pages.

Borchers, D., and Samara, F. (**2007**). *Accomodating availability bias on line transect surveys using hidden Markov models* (No. CREEM technical report 2007-05), . Retrieved from http://www.creem.st-and.ac.uk/len/onr/onr-availability-bias.pdf

Boyd, I. L., Bowen, W. D., and Iverson, S. J. (**2010**). *Marine Mammal Ecology and Conservation: A Handbook of Techniques*, Oxford University Press, USA, 448 pages.

Brooks, S. P., and Gelman, A. (**1998**). “General Methods for Monitoring Convergence of Iterative Simulations,” J. Comput. Graph. Stat., **7**, 434–455. doi:10.1080/10618600.1998.10474787

Buckastaff, K. C. (**2004**). “Effects of Watercraft Noise on the Acoustic Behavior of Bottlenose Dolphins, Tursiops Truncatus, in Sarasota Bay, Florida,” Mar. Mammal Sci., **20**, 709 – 725.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (**2001**). *Introduction to Distance Sampling.*, Oxford University Press, Oxford, 432 pages.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (**2004**). *Advanced Distance Sampling: Estimating abundance of biological populations*, OUP Oxford, 448 pages.

Buckland, S. T., Plumptre, A. J., Thomas, L., and Rexstad, E. A. (**2010**). “Design and Analysis of Line Transect Surveys for Primates,” Int. J. Primatol., **31**, 833–847. doi:10.1007/s10764-010-9431-5

Buckland, S. T., and Turnock, B. J. (**1992**). “A Robust Line Transect Method,” Biometrics, **48**, 901–909. doi:10.2307/2532356

Caillat, M., Thomas, L., and Gillespie, D. (**2013**). “The effects of acoustic misclassification on cetacean species abundance estimation,” J Acoust Soc Am, **134**, 2469–2476.

Canadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E., and Hammond, P. (**2005**). “Habitat preference modelling as a conservation tool: proposals for marine protected areas for cetaceans in southern Spanish waters,” Aquat. Conserv. Mar. Freshw. Ecosyst.,.

Cheney, B., Corkrey, R., Quick, N. J., Janik, V. M., Islas-Villanueva, V., Hammond, P. S., and Thompson, P. M. (**2012**). *Site Condition Monitoring of bottlenose dolphins within the Moray Firth special Area of Conservation: 2008-2010* (No. 512), Scottish Natural Heritage Commissioned.

Clark, C. W., Charif, R., Mitchell, S., and Colby, J. (**1996**). “Distribution and behavior of the bowhead whale, Balaena mysticetus, based on analysis of acoustic data collected during the 1993 spring migration off Point Barrow, Alaska,” Rep. Int. Whal. Comm., **46**, 541–554.

CODA (**2009**). *Cetacean Offshore Distribution and Abundance in the European Atlantic (CODA)* (Final Report), Report available from SMRU, Gatty Marine Laboratory, University of St Andrews, fife KY16 8LB, UK.

Conway, R. W., and Maxwell, W. L. (**1961**). “A queuing model with state dependent service rates,” J. Ind. Eng., **12**, 132–136.

Cowles, M. K., and Carlin, B. P. (**1996**). “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” J. Am. Stat. Assoc., **91**, 883–904.

Datta, S., and Sturtivant, C. (**2002**). “Dolphin whistle classification for determining group identities,” Signal Process., **82**, 251–258.

Eguchi, T., and Gerrodette, T. (**2009**). “A Bayesian approach to line-transect analysis for estimating abundance,” Ecol. Model., **220**, 1620–1630. doi:10.1016/j.ecolmodel.2009.04.011

Ellison, A. M. (**1996**). “An Introduction to Bayesian Inference for Ecological Research and Environmental Decision-Making,” Ecol. Appl., **6**, 1036–1046. doi:10.2307/2269588

Embling, C. B., Gillibrand, P. A., Gordon, J., Shrimpton, J., Stevick, P. T., and Hammond, P. S. (**2010**). “Using habitat models to identify suitable sites for marine protected areas for harbour porpoises (Phocoena phocoena),” Biol. Conserv., **143**, 267–279. doi:10.1016/j.biocon.2009.09.005

Ensor, P., Komiya, H., Kumagai, S., Kuningas, S., Olson, P., and Tsuda, Y. (**2010**). *2008-2009 International Whaling Commission-Southern Ocean Whale and Ecosystem Research (IWC-SOWER) Cruise* IWC.

Ensor, P., Minami, K., Morse, L., Olson, P., and Sekiguchi, K. (**2008**). *2007-2008 International whaling Commission-southern Ocean Whale and Ecosystem Research (IWC-SOWER) Cruise* (No. IWC Paper SC/60/IA 1.), .

European Union (**1992**). "Council directive 92/43 EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora," Off. J. Eur. Communities, **206**, 7–15.

Fawcett, T. (**2006**). "An introduction to ROC analysis," Pattern Recogn Lett, **27**, 861–874. doi:10.1016/j.patrec.2005.10.010

Fienberg, S. E. (**1992**). "A Brief History of Statistics in Three and One-Half Chapters: A Review Essay," Stat. Sci., **7**, 208–225. doi:10.1214/ss/1177011360

Forney, K. A., Barlow, J., and Carretta, J. . (**1995**). "The abundance of cetaceans in California waters: PartII Aerial surveys in winter and spring of 1991 and 1992," Fish. Bull., **93**, 15–26.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (**2004**). *Bayesian Data Analysis*, Chapman & Hall/CRC, NW, 717 pages.

Geman, S., and Geman, D. (**1984**). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Trans. Pattern Anal. Mach. Intell., **PAMI-6**, 721 –741. doi:10.1109/TPAMI.1984.4767596

Gerrodette, T., Taylor, B. L., Swift, R., Rankin, S., Jaramillo_Legorreta, A. M., and Rojas-Bracho, L. (**2011**). "A combined visual and acoustic estimate of 2008 abundance, and change in abundance since 1997, for the vaquita, Phocoena sinus," Mar. Mammal Sci., **27**, E79–E100.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (Eds.) (**1995**). *Markov Chain Monte Carlo in Practice*, Chapman and Hall/CRC, 512 pages.

Gillespie, D. (**2004**). "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," Can. Acoust., **32**, 39–47.

Gillespie, D., Berggren, P., Brown, S., Kuklik, I., Lacey, C., Lewis, T., Matthews, J., et al. (**2005**). "Relative abundance of harbour porpoises (Phocoena phocoena) from acoustic and visual surveys of the Baltic Sea and adjacent waters during 2001 and 2002," J Cetacean Res Manage,.

Gillespie, D., and Caillat, M. (**2008**). "Statistical Classification of Odontocete clicks," Can. Acoust.,.

Gillespie, D., Caillat, M., Gordon, J., and White, P. (**2013**). "Automatic Detection and Classification of Odontocete Whistles," J Acoust Soc Am,.

Gillespie, D., and Chappell, O. (**2002**). "An automatic system for detecting and classifying the vocalisation of harbour porpoises," Bioacoustics, **13**, 37–61.

Gillespie, D., Leaper, R., Gordon, J., and MacLeod, K. (**2010**). "An integrated data collection system for line transect surveys," J Cetacean Res Manage, **11**, 217–227.

Gillespie, D., White, P., Caillat, M., and Gordon, J. (**2011**). *Development and implementation of automatic classification of odontocetes within PAMGUARD*. (SMRU Ltd.), . Retrieved from C:\Documents and Settings\Marjolaine\My Documents\PhD

Gilpin, M. E., and Soulé, M. E. (**1986**). "Minimum Viable Populations: the processes of species extinctions," Conserv. Biol. Sci. Scarcity Divers., Sinauer Associates, Sunderland, Mass, M.E. Soulé., pp. 13–34.

Goodson, A. D., and Sturtivant, C. R. (**1996**). "Sonar characteristics of the harbour porpoise (Phocoena phocoena): source levels and spectrum," J. Mar. Sci., **53**, 465–472.

Gordon, J. C. D., Matthews, J. N., Panigada, S., Gannier, A., Borsani, J. F., and Di Sciara, G. N. (**2000**). "Distribution and relative abundance of striped dolphins, and distribution

of sperm whales in the Ligurian Sea cetacean sanctuary: results from a collaboration using acoustic monitoring techniques," J. Cetacean Res. Manag., **2**, 27–36.

Hammond, P. S., Berggren, P., Benke, H., Borchers, D. l., Collet, A., Heide-Jørgensen, M. p., Heimlich, S., et al. (**2002**). "Abundance of harbour porpoise and other cetaceans in the North Sea and adjacent waters," J. Appl. Ecol., **39**, 361–376. doi:10.1046/j.1365-2664.2002.00713.x

Harris, D. (**2012**). *Estimating whale abundance using sparse hydrophones arrays.* University of St Andrews.

Harwood, J., and Stokes, K. (**2003**). "Coping with uncertainty in ecological advice: lessons from fisheries," Trends Ecol. Evol., **18**, 617–622. doi:10.1016/j.tree.2003.08.001

Hastie, G. D., Barton, T. R., Grellier, K., Hammond, P. S., Swift, R. J., Thompson, P. M., and Wilson, B. (**2003**). "Distribution of small cetaceans within a candidate Special Area of Conservation; implications for management," J. Cetacean Res. Manag., **5**, 261–266.

Hastings, W. K. (**1970**). "Monte Carlo sampling methods using Markov chains and their applications," Biometrika, **57**, 97–109. doi:10.1093/biomet/57.1.97

Hedley, S. L. (**2000**). *Modelling Heterogeneity in Cetacean Surveys* University of St Andrews, St Andrews, 132 pages.

Hedley, S. L., Buckland, S. T., and Bochers, D. (**1999**). "Spatial modelling from line transect data," J Cetacean Res Manage, **1**, 255–264.

Hiby, A. R. (**1985**). "An approach to estimating population densities of great whales from sighting surveys," IMA J. Math. Appl. Med. Biol., **2**, 201–220.

Hiby, A. R., and Hammond, P. (**1989**). "Survey techniques for estimating abundance of cetaceans," Rep.Int.Whal.Comm,.

Hiby, L. (**1999**). "The objective identification of duplicate sightings in aerial survey for porpoise," Mar. Mammal Surv. Assess. Methods, CRC Press, Baklema, Rotterdam, 1st ed., pp. 179–189.

Janik, V. M. (**2000**). "Source levels and the estimated active space of bottlenose dolphin (Tursiops truncatus) whistles in the Moray Firth, Scotland," J. Comp. Physiol. [A], **186**, 673–680. doi:10.1007/s003590000120

Johnson, M., Madsen, P. T., Zimmer, W. M. X., de Soto, N. A., and Tyack, P. L. (**2006**). "Foraging Blainville's beaked whales (Mesoplodon densirostris) produce distinct click types matched to different phases of echolocation," J Exp Biol, **209**, 5038–5050.

King, R., Morgan, B., and Gimenez, O. (**2010**). *Bayesian Analysis for Population Ecology*, Chapman & Hall/CRC Interdisciplinary Statistics, CRC Press, 456 pages.

Kyhn, L. A., Tougaard, J., Thomas, L., Rosager Duve, L., Stenback, J., Amundin, M., Desportes, G., et al. (**2012**). "From echolocation clicks to animal density-Acoustic sampling of harbor porpoises with static dataloggers," J Acoust Soc Am, **131**, 550–560.

Lammers, M. O., Brainard, R. E., Au, W. W. L., Mooney, T. A., and Wong, K. B. (**2008**). "An ecological acoustic recorder (EAR) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats," J. Acoust. Soc. Am., **123**, 1720–1728. doi:10.1121/1.2836780

Leaper, R., Burt, L., Gillespie, D., and MacLeod, K. (**2010**). "Comparisons of measured and estimated distances and angles from sightings surveys," J Cetacean Res Manage, **11**, 229–237.

Leaper, R., Gillespie, D., and Papastavrou, V. (**2000**). "Results of passive acoustic surveys for odontocetes in the Southern Ocean," J Cetacean Res Manage, **2**, 187–196.

Lewis, T., Gillespie, D., Lacey, C., Matthews, J., Danbolt, M., Leaper, R., McLanaghan, R., et al. (**2007**). "Sperm whale abundance estimates from acoustic surveys of the Ionian

Sea and Straits of Sicily in 2003," J. Mar. Biol. Assoc. United Kingd., **87**, 353–357. doi:10.1017/S0025315407054896

Linnenschmidt, M., Teilmann, J., Akamatsu, T., Dietz, R., and Miller, L. A. (**2013**). "Biosonar, dive, and foraging activity of satellite tracked harbor porpoises (Phocoena phocoena)," Mar. Mammal Sci., **29**, 77–97. doi:10.1111/j.1748-7692.2012.00592.x

Lockyer, C., and Pike, D. (**2009**). *North Atlantic Sightings Surveys: Counting Whales in the North Atlantic, 1987-2001*, Scientific Committee, North Atlantic Marine Mammal Commission, 244 pages.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (**2002**). "ESTIMATING SITE OCCUPANCY RATES WHEN DETECTION PROBABILITIES ARE LESS THAN ONE," Ecology, **83**, 2248–2255. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., et al. (**2013**). "Estimating animal population density using passive acoustics," Biol. Rev., **88**, 287–309. doi:10.1111/brv.12001

Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., and Tyack, P. L. (**2009**). "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales," J. Acoust. Soc. Am., **125**, 1982–1994. doi:10.1121/1.3089590

Marques, T., Munger, L., Thomas, L., Wiggins, S., and Hildebrand, J. (**2011**). "Estimating North Pacific right whale Eubalaena japonica density using passive acoustic cue counting," Endanger. Species Res., **13**, 163–172. doi:10.3354/esr00325

Martin, S. W., Marques, T. A., Thomas, L., Morrissey, R. P., Jarvis, S., DiMarzio, N., Moretti, D., et al. (**2012**). "Estimating minke whale (Balaenoptera acutorostrata) boing sound density using passive acoustic sensors," Mar. Mammal Sci., , doi: 10.1111/j.1748-7692.2011.00561.x. doi:10.1111/j.1748-7692.2011.00561.x

Matthews, J. ., Brown, S., Gillespie, D., Johnson, M., McLanaghan, R., Moscrop, A., Nowacek, D., et al. (**2001**). "Vocalisation rates of the North Atlantic right whale (Eubalaena glacialis)," J CETACEAN RES MANAGE, **3**, 271–282.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (**2010**). "Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections," Ecology, **91**, 2446–2454.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (**2010**). "Experimental Investigation of Observation Error in Anuran Call Surveys," J. Wildl. Manag., **74**, 1882–1893. doi:10.2193/2009-321

McDonald, M. A., and Fox, C. G. (**1999**). "Passive acoustic methods applied to fin whale population density estimation," J. Acoust. Soc. Am., **105**, 2643–2651. doi:10.1121/1.426880

Mellinger, D. (**2008**). "A neural network for classifying clicks of Blinville's beaked whales (Mesoplodon densirostris)," Can. Acoust., **36**, 55–59.

Mellinger, D. K., and Clark, C. W. (**1997**). "Methods for automatic detection of mysticete sounds," Mar. Freshw. Behav. Physiol., **29**, 163–181. doi:10.1080/10236249709379005

Mellinger, D. K., Martin, S. W., Morrissey, R. P., Thomas, L., and Yosco, J. J. (**2011**). "A method for detecting whistles, moans, and other frequency contour sounds," J. Acoust. Soc. Am., **129**, 4055. doi:10.1121/1.3531926

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (**1953**). "Equation of State Calculations by Fast Computing Machines," J. Chem. Phys., **21**, 1087–1092. doi:doi:10.1063/1.1699114

Miller, D. A., Nichols, J. D., Mcclintock, B. T., Grant, E. H. C., Bailey, L. L. L., and Weir, L. A. (**2011**). "Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification," Ecology, **92**, 1422–1428.

Moore, S. E., Dahlheim, M. E., Stafford, K. M., Fox, C. G., Braham, H. W., McDonald, M. A., and Thomason, J. (**1998**). "Acoustic and visual detection of large whales in the eastern North Pacific Ocean," Rep.Int.Whal.Comm,.

Moray Offshore Renewables ltd (**2010**). *Environmental Impact Assessment* (Scoping Report), Chapter 4.4,.

Moretti, D., Marques, T. A., Thomas, L., DiMarzio, N., Dilley, A., Morrissey, R., McCarthy, E., et al. (**2010**). "A dive counting density estimation method for Blainville's beaked whale (Mesoplodon densirostris) using a bottom-mounted hydrophone field as applied to a Mid-Frequency Active (MFA) sonar operation," Appl. Acoust., **71**, 1036–1042. doi:10.1016/j.apacoust.2010.04.011

Morgan, B. J. T. (**1984**). *Elements of Simulation*, Chapman and Hall/CRC, 1st ed., 368 pages.

Nanayakkara, S. C., Chitre, M., Ong, S. H., and Taylor, E. (**2007**). "Automatic classification of whistles produced by Indo-Pacific humpback dolphins (Sousa chinensis)," Oceans 2007 - Eur. Vols 1-3,.

Norris, T., Martin, S., Thomas, L., Yack, T., Oswald, J. N., Nosal, E. ., and Janik, V. (**2010**). "Acoustic ecology and behavior of minke whales in the Hawaiian and Marianas islands: localization,abundance estimation, and characterization of minke whale 'Boings'," Eff. Noise Ad Aquat. Life Adv. Exp. Med. Biol., Springer, Popper, A.N. and A. Hawkins., pp. 149–153.

Oswald, J. N., Barlow, J., and Norris, T. F. (**2003**). "Acoustic identification of nine delphinid species in the eastern tropical pacific ocean," Mar. Mammal Sci., **19**, 20–37.

Oswald, J., Shannon, R., Barlow, J., and Marc, O. L. (**2007**). "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am., **122**, 587–595.

Palka, D. L. (**1996**). "Effects of Beaufort Sea State on the Sightability of Harbour Porpoise in the Gulf of Main," Rep.Int.Whal.Comm, **46**, 575–582.

Potter, J. R., and Mellinger, D. (**1993**). "Application and comparison of neural nets for marine mammal call classification," J. Acoust. Soc. Am., **94**, 1822.

Punt, A. E. (**1992**). "Selecting management methodologies for marine resources, with an illustration for southern African hake," South Afr. J. Mar. Sci., **12**, 943–958. doi:10.2989/02577619209504754

Punt, A. E., and Donovan, G. P. (**2007**). "Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission," ICES J. Mar. Sci. J. Cons., **64**, 603–612. doi:10.1093/icesjms/fsm035

Quick, N. J., and Janik, V. M. (**2008**). "Whistle rates of wild bottlenose dolphins (Tursiops truncatus): Influences of group size and behavior," J. Comp. Psychol., **122**, 305–311. doi:10.1037/0735-7036.122.3.305

Quick, N. J., Rendell, L. E., and Janik, V. M. (**2008**). "A mobile acoustic localization system for the study of free-ranging dolphins during focal follows," Mar. Mammal Sci., **24**, 979–989. doi:10.1111/j.1748-7692.2008.00231.x

R Development Core Team, (**2012**). R: A language and environment for statistical computing, Vienna, Austria, R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org

Rasmussen, M. H., and Miller, L. A. (**2002**). "Whistles and clicks from white-beaked dolphins,Lagenorhynchus albirostris, recorded in Faxaflói Bay, Iceland," Aquat. Mamm., **28**, 78–89.

Reid, J. ., Evans, P. G. H., and Northridge, S. P. (**2003**). *Atlas of Cetacean Distribution in North-west European Waters.*, Joint Nature Conservation Committee, Peterborough, Eds., 76 pages.

Rendell, L., Matthews, J., and Macdonald (**1999**). "Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation," Zool. Soc. Lond., **243**, 403–410.

Richardson, W. J., Greene, C. H., Malme, C. I., and Thomson, D. H. (**1995**). *Marine mammals and noise*, Academic Pr, 575 pages.

Roch, M. A., Soldevilla, M. S., Burtenshaw, J. C., Henderson, E. E., and Hildebrand, J. A. (**2007**). "Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California," J. Acoust. Soc. Am., **121**, 1737–1748.

Rowe, W. D. (**1977**). *An anatomy of risk*, Wiley, 520 pages.

Royle, J. ., and Link, W. . (**2006**). "Generalized site occupancy models allowing for false positive and false negative errors," Ecology, **87**, 835–841.

Royle, J. A., and Dorazio, R. M. (**2008**). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*, Academic Press, Oxford, Elsevier., 464 pages.

SCANS-II (**2008**). *Small cetacean in the European Atlantic and North Sea* (Final Report to the European Commision No. LIFE04NAT/GB/000245), Gatty Marine Laboratory, University of St Andrews, st Andrews, fife, KY16 8LB,UK: SMRU.

Schipper, J., Chanson, J. S., Chiozza, F., Cox, N. A., Hoffmann, M., Katariya, V., Lamoreux, J., et al. (**2008**). "The Status of the World's Land and Marine Mammals: Diversity, Threat, and Knowledge," Science, **322**, 225–230. doi:10.1126/science.1165115

Seber, G. A. . (**1982**). *The estimation of animal abundance.*, Griffin, London, 2nd ed.

Simons, T. R., Alldredge, M. W., Pollock, K. H., Wettroth, J. M., and Dufty Jr, A. M. (**2007**). "Experimental analysis of the auditory detection process on avian point counts," The Auk, **124**, 986–999.

Sirovic, A., Hildebrand, J. A., and Wiggins, S. (**2007**). "Blue and fin whale call source levels and propagation range in the Southern Ocean," Retrieved from http://www.escholarship.org/uc/item/8mr3c6vn. Retrieved from http://www.escholarship.org/uc/item/8mr3c6vn

Skaug, H. J., and Schweder, T. (**1999**). "Hazard Models for Line Transect Surveys with Independent Observers," Biometrics, **55**, 29–36. doi:10.2307/2533892

Smith, A. F. M., and Gelfand, A. E. (**1992**). "Bayesian Statistics without Tears: A Sampling–Resampling Perspective," Am. Stat., **46**, 84–88. doi:10.1080/00031305.1992.10475856

SMRU ltd, Grellier, K., Booth, C., Caillat, M., and Gillespie, D. (**2011**). *Developing acoustic methods for determining the likelihood that dolphins using the MORL and BOWL wind farm sites are bottlenose dolphins* (No. 19.05.10.UOA), SMRU ltd.

Soldevilla, M. S., Henderson, E. E., Campbell, G. S., Wiggins, S. M., Hildebrand, J. A., and Roch, M. A. (**2008**). "Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks," J. Acoust. Soc. Am., **124**, 609–624.

Sousa-Lima, R. S., Norris, T. F., Oswald, J. N., and Fernandes, D. P. (**2013**). "A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals," Aquat. Mamm., **39**, 23–53.

Steiner, W. W. (**1981**). "Species-specific differences in pure tonal whistle vocalizations of five western North Atlantic dolphin species," Behav. Ecol. Sociobiol., **9**, 241–246.

Stuart, S. N., Chanson, J. S., Cox, N. A., Young, B. E., Rodrigues, A. S. L., Fischman, D. L., and Waller, R. W. (**2004**). "Status and Trends of Amphibian Declines and Extinctions Worldwide," Science, **306**, 1783–1786. doi:10.1126/science.1103538

Swartz, S. L., Cole, T., McDonald, M. A., Hildebrand, J. A., Oleson, E. M., Martinez, A., Clapham, P. J., et al. (**2003**). “Acoustic and visual survey of humpback whale (Megaptera novaeangliae) distribution in the eastern and southeastern Caribbean Sea,” Caribb. J. Sci., **39**, 195–208.

Taylor, B. L., Wade, P. R., Stehn, R. A., and Cochrane, J. F. (**1996**). “A Bayesian Approach to Classification Criteria for Spectacled Eiders,” Ecol. Appl., **6**, 1077–1089. doi:10.2307/2269592

Thode, A. M., Kim, K. H., Blackwell, S. B., Charles R. Greene, J., Nations, C. S., McDonald, T. L., and Macrander, A. M. (**2012**). “Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys,” J. Acoust. Soc. Am., **131**, 3726–3747. doi:10.1121/1.3699247

Thomas, J. A., Fisher, S. R., Ferm, L., and Holt, R. . (**1986**). “Acoustic detection of cetacean using a towed array of hydrophones,” Rep.Int.Whal.Comm,.

Thomas, L., and Marques, T. A. (**2012**). “Passive Acoustic Monitoring for Estimating Animal Density,” Acoust. Today, **8**, 35–44. doi:10.1121/1.4753915

Thompson, P. M., Cheney, K., Cândido, B., Bates, A., Richardson, H., and Barton, T. (**2010**). *Assessing the impact of seismic surveys on cetaceans in the Moray Firth.* (No. First year), DECC, Scottish Government, COWRIE and Oil & Gaz UK.

Tiemann, C. O., Porter, M., and Frazer, L. N. (**2004**). “Localization of marine mammals near Hawaii using an acoustic propagation model,” J. Acoust. Soc. Am., **115**, 2834–2843.

Vu, E. T., Risch, D., Clark, C. W., Gaylord, S., Hatch, L. T., Thompson, M. ., Wiley, D. N., et al. (**2012**). “Humpback whale song occurs extensively on feeding grounds in the western North Atlantic Ocean,” Aquat. Biol., **14**, 175–183.

Wade, P. R. (**2000**). “Bayesian Methods in Conservation Biology,” Conserv. Biol., **14**, 1308–1316.

Wahlberg, M. (**2002**). “The acoustic behaviour of diving sperm whales observed with a hydrophone array,” J. Exp. Mar. Biol. Ecol., **281**, 53–62.

Wiggins, S. M., McDonald, M. A., Munger, L. A., Moore, S. E., and Hildebrand, J. A. (**2004**). “Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea,” Can. Acoust., **32**, 146–154.

Zimmer, W. M. X., Harwood, J., Tyack, P. L., Johnson, M. P., and Madsen, P. T. (**2008**). “Passive acoustic detection of deep-diving beaked whales,” J. Acoust. Soc. Am., **124**, 2823–2832. doi:10.1121/1.2988277

Appendix A. Appendix for chapter 3

Table A-1 Classification result of the EAR data classified with the *2Sp* classifier: Encounters time: time of the first section of the encounter. n= total numbers of sections within each encounters of bottlenose dolphins (nBND) and other dolphins (nOTHER). p is the average probability of a section to be classified as bottlenose dolphins (pBND) or as other dolphins (pOTHER). Classified as: final classification of the encounter after observation by the manual observed. When all the contours within an encounter are false detections then the encounters was classified as a false detection (FD) encounters.

EAR: E21

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
18/08/2010 02:20:20	4	4	0	1.00	0.00	FD
18/08/2010 12:08:15	12	12	0	0.98	0.02	FD
20/08/2010 14:27:00	1	1	0	0.98	0.02	FD
20/08/2010 15:28:53	3	3	0	0.99	0.01	FD
22/08/2010 03:23:08	5	1	4	0.20	0.80	OTHER
22/08/2010 03:53:08	1	0	1	0.20	0.80	OTHER
24/08/2010 06:34:35	1	1	0	1.00	0.00	FD
05/09/2010 08:30:47	1	1	0	1.00	0.00	FD
07/09/2010 07:30:45	1	1	0	0.99	0.01	FD
07/09/2010 08:00:45	2	2	0	0.99	0.01	FD
07/09/2010 09:36:18	1	1	0	0.99	0.01	FD

EAR: 17

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
29/07/2010 09:22:19	15	0	15	0.09	0.91	OTHER
29/07/2010 13:22:47	40	0	40	0.12	0.88	OTHER
01/08/2010 23:31:19	57	0	57	0.05	0.95	OTHER
04/08/2010 21:26:29	8	0	8	0.08	0.92	OTHER

EAR: A20

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
22/07/2010 17:14:28	2	2	0	0.97	0.03	FD
24/07/2010 17:05:17	1	1	0	0.94	0.06	FD
26/07/2010 09:23:41	89	2	87	0.13	0.87	OTHER
26/07/2010 09:23:41	89	2	87	0.13	0.87	OTHER
29/07/2010 10:05:45	92	1	91	0.09	0.91	OTHER
29/07/2010 11:20:32	2	0	2	0.19	0.81	OTHER
29/07/2010 13:02:22	120	0	120	0.08	0.92	OTHER
31/07/2010 18:25:49	1	0	1	0.30	0.70	OTHER
01/08/2010 23:01:27	43		43	0.02	0.98	OTHER
03/08/2010 09:02:24	12	12	0	0.99	0.01	FD
04/08/2010 18:20:35	2	1	1	0.82	0.18	FD
04/08/2010 19:08:06	3	3	0	0.82	0.18	FD
05/08/2010 00:17:59	1	1	0	1.00	0.00	FD
06/08/2010 15:29:15	1	1	0	0.83	0.17	FD
07/08/2010 18:29:57	1	0	1	0.08	0.92	FD
09/08/2010 22:12:21	1	0	1	0.01	0.99	OTHER
11/08/2010 01:05:04	3	2	1	0.70	0.30	FD
12/08/2010 14:24:44	1	1	0	0.67	0.33	FD
13/08/2010 11:12:10	1	0	1	0.00	1.00	FD
13/08/2010 15:09:06	1	1	0	0.95	0.05	FD
13/08/2010 23:18:38	1	0	1	0.00	1.00	FD
14/08/2010 02:20:56	1	0	1	0.16	0.84	FD

EAR: E16

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
22/09/2010 23:23:22	1	1		0.83	0.17	FD
23/09/2010 10:19:58	1	1		0.91	0.09	FD
23/09/2010 19:25:05	2	2		0.72	0.28	FD
24/09/2010 10:21:22	2	2		0.76	0.24	FD
24/09/2010 13:16:04	1	1		0.8	0.2	FD
25/09/2010 10:16:34	1	1		0.84	0.16	FD
25/09/2010 12:03:50	2	2		0.78	0.22	FD
25/09/2010 23:16:22	2	2		0.91	0.09	FD
28/09/2010 08:11:12	2		2	0.09	0.91	FD
01/10/2010 15:23:38	2		2	0.25	0.75	FD
01/10/2010 17:22:16	2		2	0.34	0.66	FD
03/10/2010 16:13:02	1	1		0.99	0.01	FD
05/10/2010 05:24:23	1	1		0.59	0.41	FD
07/10/2010 12:14:32	1	1		0.99	0.01	FD
08/10/2010 13:03:57	2	2		1	0	FD
10/10/2010 06:24:58	1		1	0	1	FD
13/10/2010 06:20:44	87	87		0.98	0.02	FD
13/10/2010 15:14:31	1	1		0.99	0.01	FD
14/10/2010 22:12:31	1	1		1	0	FD

EAR: D01

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
08/10/2010 17:27:22	5	3	2	0.68	0.32	BND
09/10/2010 07:47:14	5	0	5	0.27	0.73	OTHER
09/10/2010 19:37:12	1	1	0	0.7	0.3	BND
09/10/2010 21:54:32	1	1	0	0.98	0.02	BND
09/10/2010 22:46:05	1	1	0	0.55	0.45	BND
10/10/2010 03:46:59	2	2	0	0.69	0.31	BND
10/10/2010 23:38:13	3	3	0	0.7	0.3	BND
11/10/2010 04:39:10	1	0	1	0.19	0.81	OTHER
11/10/2010 20:26:07	4	4		0.86	0.14	BND
11/10/2010 22:34:32	2	2		0.96	0.04	BND
12/10/2010 06:37:49	3	3		0.92	0.08	BND
12/10/2010 15:35:37	1	1		0.89	0.11	BND
13/10/2010 20:45:44	20	20		0.92	0.08	BND
14/10/2010 12:25:38	1		1	0.24	0.76	OTHER
14/10/2010 17:51:26	1	1		0.8	0.2	BND
15/10/2010 07:33:37	1	1		0.71	0.29	BND
15/10/2010 09:47:13	1	1		0.92	0.08	BND
15/10/2010 13:42:35	2		2	0	1	OTHER
15/10/2010 16:50:54	2	2		0.97	0.03	BND
16/10/2010 09:27:20	1	1		1	0	BND
16/10/2010 19:49:24	2	2		0.82	0.18	BND
18/10/2010 01:39:38	4	4		0.94	0.06	BND
20/10/2010 19:31:34	2	1	1	0.48	0.52	OTHER
20/10/2010 21:42:10	2	2	0	0.84	0.16	BND
20/10/2010 22:28:17	4	0	4	0.36	0.64	OTHER

EAR: D01

Encounters time	n	n		p		Classified as
		BND	OTHER	BND	OTHER	
21/10/2010 21:32:55	6	6		0.88	0.12	BND
21/10/2010 22:26:48	1	1		0.81	0.19	BND
22/10/2010 04:28:38	1		1	0.45	0.55	OTHER
22/10/2010 16:29:21	4	2	2	0.55	0.45	BND
22/10/2010 17:45:04	4	1	3	0.46	0.54	OTHER
22/10/2010 23:36:16	6	6		0.83	0.17	BND
23/10/2010 10:32:47	2	2		0.9	0.1	BND
23/10/2010 22:29:53	4	4		1	0	BND
24/10/2010 18:27:55	4	3	1	0.67	0.33	BND
24/10/2010 22:44:01	1	1		0.94	0.06	BND
25/10/2010 01:35:00	5	4	1	0.82	0.18	BND
25/10/2010 18:34:00	1	1		0.8	0.2	BND
25/10/2010 20:42:13	2	2	0	0.72	0.28	BND
26/10/2010 00:41:54	2	2		0.88	0.12	BND
27/10/2010 15:33:06	2		2	0.12	0.88	OTHER
28/10/2010 12:30:46	10	7	3	0.66	0.34	BND
30/10/2010 11:54:16	3	3	0	1	0	BND

Table A.2: Classification result of the EAR data classified with the *5Sp* classifier. The column headings are similar to the previous table with more species: COD=common dolphin, RSD=Risso's dolphin, WBD=white beaked dolphin, WSD=white sided dolphin. *5Sp* class as= classification result by the 5Sp classifier in comparison to the classification result by the *2Sp* classifier (*2Sp* class as) after the manual check of the spectrograms.

EAR: E21

Encounters time	n	n					p					*5Sp*	*2Sp*
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
18/08/2010 02:20:20	4			4		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
18/08/2010 12:08:15	12			12		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
20/08/2010 14:27:00	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
20/08/2010 15:28:53	3			3		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
22/08/2010 03:23:08	6	1	2	0	3	0	0.16	0.37	0.00	0.46	0.00	WBD	OTHER
24/08/2010 06:34:35	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
05/09/2010 08:30:47	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
07/09/2010 07:30:45	3			3		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
07/09/2010 09:36:18	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD

EAR: E17

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
29/07/2010 09:22:19	15		14		1	0	0.07	0.73	0.00	0.15	0.05	COD	OTHER
29/07/2010 13:22:47	40	1	25		14	0	0.07	0.57	0.00	0.33	0.03	COD	OTHER
01/08/2010 23:31:19	57	0	32	1	21	3	0.04	0.47	0.00	0.37	0.12	COD	OTHER
04/08/2010 21:26:29	8		6		2	0	0.04	0.67	0.00	0.23	0.06	COD	OTHER

EAR: A20

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
22/07/2010 17:14:28	2			2		0			1.00			RSD	FD
24/07/2010 17:05:17	1			1		0			1.00			RSD	FD
26/07/2010 09:23:41	89	7	68	0	12	2	0.13	0.64	0.00	0.18	0.05	COD	OTHER
29/07/2010 10:05:45	92	2	66	0	13	11	0.07	0.57	0.00	0.22	0.14	COD	OTHER

EAR: A20

Encounters time	n	n					p					5Sp Class as	2Sp Class. as
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD		
29/07/2010 11:20:32	2		2			0	0.28	0.52	0.19	0.00	0.01	COD	OTHER
29/07/2010 13:02:22	120	4	112		4	0	0.08	0.75	0.00	0.10	0.07	COD	OTHER
31/07/2010 18:25:49	1		1			0	0.10	0.82	0.00	0.00	0.08	COD	OTHER
01/08/2010 23:01:27	43	0	21	0	2	20	0.03	0.45	0.00	0.11	0.41	COD	OTHER
03/08/2010 09:02:24	12	0	0	12	0	0	0.00	0.00	1.00	0.00	0.00	RSD	FD
04/08/2010 18:20:35	5	1	1	3	0	0	0.21	0.08	0.62	0.08	0.01	RSD	FD
05/08/2010 00:17:59	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
06/08/2010 15:29:15	1				1	0	0.16	0.01	0.08	0.75	0.00	WBD	FD
07/08/2010 18:29:57	1				1	0	0.02	0.29	0.00	0.68	0.00	WBD	FD
09/08/2010 22:12:21	1		1			0	0.00	0.53	0.00	0.47	0.00	COD	OTHER
11/08/2010 01:05:04	3		2		1	0	0.00	0.67	0.33	0.00	0.00	RSD	FD
12/08/2010 14:24:44	1				1	WBD	0.09	0.01	0.02	0.87	0.00	WBD	FD
13/08/2010 11:12:10	1	1				COD	0.03	0.90	0.00	0.00	0.06	COD	FD

EAR: A20

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
13/08/2010 15:09:06	1			1		RSD	0.00	0.00	1.00	0.00	0.00	RSD	FD
13/08/2010 23:18:38	1			1		RSD	0.00	0.02	0.98	0.00	0.00	RSD	FD
14/08/2010 02:20:56	1				1	WBD	0.00	0.01	0.00	0.99	0.00	WBD	FD

EAR: E16

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
22/09/2010 23:23:22	1				1	0	0.00	0.00	0.00	1.00	0.00	WBD	FD
23/09/2010 10:19:58	1			1		0			1.00		0.00	RSD	FD
23/09/2010 19:25:05	2			2		0			1.00		0.00	RSD	FD
24/09/2010 10:21:22	2			1	1	0	0.03	0.00	0.72	0.24	0.01	RSD	FD
24/09/2010 13:16:04	1				1	0	0.10	0.00	0.09	0.80	0.00	WBD	FD
25/09/2010 10:16:34	1				1	0	0.34	0.01	0.05	0.60	0.00	WBD	FD
25/09/2010 12:03:50	2			2		0	0.00	0.00	1.00	0.00	0.00	RSD	FD

EAR: E16

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
25/09/2010 23:16:22	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
26/09/2010 00:14:41	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
28/09/2010 07:27:28	2			1	1	0	0.06	0.06	0.50	0.38	0.00	RSD	FD
01/10/2010 15:23:38	2			1	1	0	0.00	0.03	0.50	0.46	0.01	RSD	FD
01/10/2010 17:22:16	2				2	0	0.04	0.04	0.01	0.91	0.00	WBD	FD
03/10/2010 16:13:02	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
05/10/2010 05:24:23	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
07/10/2010 12:14:32	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
08/10/2010 13:03:57	2			2		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
13/10/2010 06:20:44	87			87		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
13/10/2010 15:14:31	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD
14/10/2010 22:12:31	1			1		0	0.00	0.00	1.00	0.00	0.00	RSD	FD

EAR: D01

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
08/10/2010 17:27:22	5	3	1		1	0	0.62	0.11	0.01	0.25	0.01	BND	BND
09/10/2010 07:47:14	5		2		3	0	0.11	0.29	0.00	0.59	0.01	WBD	OTHER
09/10/2010 19:37:12	1	1				0	0.79	0.20	0.00	0.02	0.00	BND	BND
09/10/2010 21:54:32	2	2				0	0.74	0.10	0.00	0.16	0.00	BND	BND
10/10/2010 03:46:59	2	2				0	0.64	0.18	0.00	0.18	0.00	BND	BND
10/10/2010 23:38:13	3	1			2	0	0.45	0.11	0.00	0.44	0.00	BND	BND
11/10/2010 04:39:10	1		1			0	0.30	0.64	0.00	0.06	0.00	COD	OTHER
11/10/2010 20:26:07	4	3			1	0	0.76	0.07	0.00	0.17	0.00	BND	BND
11/10/2010 22:34:32	2	2				0	0.94	0.06	0.00	0.00	0.00	BND	BND
12/10/2010 06:37:49	3	3				0	0.98	0.02	0.00	0.00	0.00	BND	BND
12/10/2010 15:35:37	1	1				0	0.95	0.04	0.00	0.01	0.00	BND	BND
13/10/2010 20:45:44	20	17		1	2	0	0.85	0.02	0.04	0.09	0.00	BND	BND
14/10/2010 12:25:38	1	1				0	0.56	0.44	0.00	0.00	0.00	BND	OTHER
14/10/2010 17:51:26	1	1				0	0.82	0.10	0.00	0.09	0.00	BND	BND

EAR: D01

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
15/10/2010 07:33:37	1	1				0	0.85	0.15	0.00	0.00	0.00	BND	BND
15/10/2010 09:47:13	1	1				0	0.98	0.02	0.01	0.00	0.00	BND	BND
15/10/2010 13:42:35	2			1	0	1	0.00	0.14	0.64	0.00	0.22	WSD	OTHER
15/10/2010 16:50:54	2	2				0	0.96	0.04	0.00	0.00	0.00	BND	BND
16/10/2010 09:27:20	1	1				0	1.00	0.00	0.00	0.00	0.00	BND	BND
16/10/2010 19:49:24	2	1		1		0	0.45	0.00	0.55	0.00	0.00	RSD	BND
18/10/2010 01:39:38	4	4				0	0.94	0.06	0.00	0.00	0.00	BND	BND
20/10/2010 19:31:34	2	1	1			0	0.46	0.41	0.00	0.12	0.01	BND	OTHER
20/10/2010 21:42:10	2	2				0	0.95	0.05	0.00	0.00	0.00	BND	BND
20/10/2010 22:28:17	4		4			0	0.06	0.73	0.00	0.13	0.08	COD	OTHER
21/10/2010 21:32:55	6	6				0	0.92	0.07	0.00	0.01	0.00	BND	BND
21/10/2010 22:26:48	1	1				0	0.64	0.21	0.00	0.15	0.00	BND	BND
22/10/2010 04:28:38	1		1			0	0.28	0.56	0.00	0.15	0.01	COD	OTHER
22/10/2010 16:29:21	4	2	2			0	0.60	0.32	0.00	0.08	0.00	BND	BND

EAR: D01

Encounters time	n	n					p					5Sp	2Sp
		BND	COD	RSD	WBD	WSD	BND	COD	RSD	WBD	WSD	Class as	Class. as
22/10/2010 17:45:04	4	2		1	1	0	0.41	0.11	0.25	0.23	0.00	BND	OTHER
22/10/2010 23:36:16	6	4		1	1	0	0.66	0.02	0.16	0.15	0.01	BND	BND
23/10/2010 10:32:47	2	1			1	0	0.73	0.02	0.00	0.25	0.00	BND	BND
23/10/2010 22:29:53	4	3			1	0	0.78	0.04	0.00	0.18	0.00	BND	BND
24/10/2010 18:27:55	4	3	1			0	0.71	0.26	0.00	0.03	0.00	BND	BND
24/10/2010 22:44:01	1	1				0	0.98	0.02	0.00	0.01	0.00	BND	BND
25/10/2010 01:35:00	5	4	1			0	0.84	0.11	0.05	0.00	0.00	BND	BND
25/10/2010 18:34:00	1	1				0	0.88	0.10	0.00	0.02	0.00	BND	BND
25/10/2010 20:42:13	2	1			1	0	0.51	0.04	0.00	0.45	0.00	BND	BND
26/10/2010 00:41:54	2	2				0	0.94	0.04	0.00	0.01	0.01	BND	BND
27/10/2010 15:33:06	2	2			1	0	0.26	0.21	0.00	0.53	0.00	WBD	OTHER
28/10/2010 12:30:46	10	9	1			0	0.76	0.17	0.00	0.04	0.03	BND	BND
30/10/2010 11:54:16	3	3				0	1.00	0.00	0.00	0.00	0.00	BND	BND

Appendix B. Appendix for chapter 4

Table B-1: Classification result of the French data classified with the 5Sp and 3Sp Spanish classifiers: BND=bottlenose dolphins, COD=common dolphins, C&S=common/striped dolphins, FPW=pilot whales, STD=Striped dolphins, n=number of sections per encounter, p= classification probabilities per classification group. Class as= classification result by the 5Sp classifier in comparison to the classification result by the 3Sp classifier

Truth: CD		5SP Spanish Classifier											3SP Spanish classifier						
Encounter Time	n	n					p					Class as	n			p			Class as
		BND	COD	C&S	FPW	STD	BND	COD	C&S	FPW	STD		BND	CSD	FPW	BND	CSD	FPW	
17/07/2007 15:53	26	0	11	6	0	9	0.02	0.39	0.28	0	0.31	CD	0	26	0	0.04	0.96	0	CS
20/07/2007 07:11	1	0	0	0	0	1	0.01	0.14	0.39	0	0.46	SD	0	1	0	0.06	0.94	0	CS
21/07/2007 15:21	143	3	45	51	1	43	0.02	0.32	0.34	0	0.32	CS	3	139	1	0.04	0.96	0	CS
21/07/2007 17:35	4	0	4	0	0	0	0.02	0.49	0.25	0	0.24	CD	0	4	0	0.04	0.93	0	CS
21/07/2007 18:55	3	0	0	1	2	0	0.01	0.12	0.16	0.57	0.14	PW	1	2	0	0.32	0.68	0	CS
24/07/2007 05:35	403	11	168	122	11	91	0.03	0.34	0.31	0.02	0.29	CD	12	379	12	0.06	0.91	0	CS
24/07/2007 07:32	4	0	0	1	0	3	0	0.28	0.33	0	0.39	SD	0	4	0	0	1	0	CS
25/07/2007 06:16	5	0	5	0	0	0	0	0.72	0.13	0.01	0.13	CD	0	5	0	0.01	0.98	0	CS

Truth: CSD		5SP Spanish Classifier											3SP Spanish classifier						
Encounter Time	n	n					p					Class as	n			p			Class as
		BND	COD	C&S	FPW	STD	BND	COD	C&S	FPW	STD		BND	CSD	FPW	BND	CSD	FPW	
09/07/2007 09:53	15	0	4	10	0	1	0.02	0.31	0.39	0	0.27	CS	0	15	0	0.05	0.94	0	CS

Truth: FPW		5SP Spanish Classifier											3SP Spanish classifier						
Encounter Time	n	n					p					Class as	n			p			Class as
		BND	COD	C&S	FPW	STD	BND	COD	C&S	FPW	STD		BND	CSD	FPW	BND	CSD	FPW	
18/07/2007 05:34	11	10	0	0	1	0	0.9	0	0	0.09	0	BD	9	0	2	0.82	0	0	BD
19/07/2007 10:23	66	57	3	1	4	1	0.86	0.04	0.03	0.05	0.02	BD	53	4	9	0.8	0.07	0	BD
21/07/2007 15:21	158	3	50	54	1	50	0.02	0.32	0.34	0	0.32	CS	3	154	1	0.03	0.96	0	CS

Truth: STD		5SP Spanish Classifier											3SP Spanish classifier						
Encounter Time	n	n					p					Class as	n			p			Class as
		BND	COD	C&S	FPW	STD	BND	COD	C&S	FPW	STD		BND	CSD	FPW	BND	CSD	FPW	
08/07/2007 11:30	24	0	5	9	0	10	0	0.3	0.35	0	0.34	CS	0	24	0	0.01	0.99	0	CS
09/07/2007 09:48	6	0	1	3	0	2	0.01	0.28	0.37	0	0.34	CS	0	6	0	0.02	0.98	0	CS
21/07/2007 17:35	4	0	4	0	0	0	0.02	0.49	0.25	0	0.24	CD	0	4	0	0.04	0.93	0	CS
24/07/2007 12:20	3	1	1	0	1	0	0.28	0.22	0.16	0.26	0.09	BD	2	1	0	0.53	0.44	0	BD

Table B-2: Classification results of the Spanish data classified with the 4Sp and 2Sp French classifiers with n, nCOD, nC&S, nFPW, nSTD, nCSD being respectively the total number of sections per encounters for all species and the number of section for common dolphins, common/stripped dolphins, pilot whales, striped dolphins and commons and striped together. pCOD, pC&D, pFPW, pSTD, pCSD being the classification probabilities per classification group and Class as is the classification result per encounter.

True Species : COD		4Sp French classifier									2Sp French Classifier				
EncounterTime	n	n				p				Class as	n		p		Class as
		COD	C&S	FPW	STD	COD	CSD	FPW	STD		COD	FPW	CSD	FPW	
05/07/2007 10:22	233	83	27	17	106	0.34	0.18	0.14	0.34	2	110	123	0.49	0.51	PW
06/07/2007 08:20	167	54	20	19	74	0.31	0.18	0.19	0.32	SD	74	93	0.51	0.49	CS
06/07/2007 11:20	46	25	6	13	2	0.41	0.14	0.27	0.18	CD	34	12	0.6	0.4	CS
07/07/2007 15:43	15	6	1	2	6	0.34	0.16	0.23	0.27	CD	8	7	0.53	0.47	CS
11/07/2007 05:28	43	16	3	1	23	0.31	0.15	0.17	0.37	SD	22	21	0.51	0.49	CS
11/07/2007 06:11	290	120	35	35	100	0.34	0.18	0.2	0.28	CD	162	128	0.53	0.47	CS
11/07/2007 08:47	428	123	74	63	168	0.3	0.19	0.2	0.31	SD	194	234	0.49	0.51	PW
11/07/2007 09:56	226	86	37	31	72	0.32	0.2	0.21	0.27	CD	132	94	0.53	0.47	CS
12/07/2007 12:46	448	125	100	85	138	0.29	0.22	0.22	0.28	CD	224	224	0.51	0.49	CS
12/07/2007 13:04	32	12	6	4	10	0.32	0.21	0.17	0.29	CD	18	14	0.53	0.47	CS
12/07/2007 14:49	162	47	32	45	38	0.3	0.22	0.24	0.24	CD	96	66	0.53	0.47	CS
12/07/2007 18:37	2	0	0	0	2	0.32	0.22	0.07	0.4	SD	1	1	0.45	0.55	PW
14/07/2007 07:26	100	44	16	5	35	0.33	0.21	0.16	0.3	CD	39	61	0.48	0.52	PW
14/07/2007 08:00	1642	483	229	272	658	0.3	0.18	0.21	0.3	2	765	877	0.5	0.5	2
18/07/2007 12:06	4	1	0	1	2	0.23	0.11	0.2	0.46	SD	3	1	0.54	0.46	CS
18/07/2007 13:46	41	31	3	1	6	0.48	0.18	0.08	0.26	CD	37	4	0.66	0.34	CS
18/07/2007 17:16	4	2	1	0	1	0.4	0.18	0.14	0.28	CD	3	1	0.53	0.47	CS
28/07/2007 17:29	12	3	3	5	1	0.27	0.26	0.32	0.15	PW	4	8	0.48	0.52	PW

True Species: CSD		4Sp French classifier									2Sp French Classifier				
EncounterTime	n	n				p				Class as	n		p		Class as
		COD	C&S	FPW	STD	COD	C&S	FPW	STD		CSD	FPW	CSD	FPW	
06/07/2007 12:56	45	14	11	3	17	0.26	0.27	0.14	0.32	SD	20	25	0.480	0.520	PW
07/07/2007 07:14	4	2	1	0	1	0.38	0.14	0.19	0.28	CD	3	1	0.570	0.430	CS
07/07/2007 09:58	1	0	0	0	1	0.06	0	0	0.94	SD	0	1	0.370	0.630	PW
11/07/2007 08:39	2	0	0	0	2	0.26	0.13	0.17	0.43	SD	0	2	0.300	0.700	PW
11/07/2007 12:32	1	0	0	0	1	0.15	0.17	0.3	0.38	SD	0	1	0.400	0.600	PW
11/07/2007 16:08	9	1	0	2	6	0.25	0.17	0.21	0.38	SD	4	5	0.470	0.530	PW
11/07/2007 16:32	271	55	52	37	127	0.27	0.22	0.19	0.31	SD	88	183	0.460	0.540	PW
12/07/2007 07:57	19	5	4	3	7	0.29	0.2	0.2	0.31	SD	10	9	0.510	0.490	CS
13/07/2007 06:09	7	3	0	0	4	0.4	0.11	0.09	0.4	2	3	4	0.490	0.510	PW
13/07/2007 14:54	1	0	0	1	0	0.17	0.13	0.38	0.32	PW	0	1	0.410	0.590	PW
13/07/2007 15:44	14	7	1	0	6	0.45	0.15	0.09	0.31	CD	6	8	0.480	0.520	PW
14/07/2007 07:20	6	1	0	0	5	0.26	0.14	0.13	0.46	SD	1	5	0.400	0.600	PW
14/07/2007 07:48	18	2	2	2	12	0.27	0.15	0.2	0.38	SD	5	13	0.450	0.550	PW
14/07/2007 08:54	109	32	13	14	50	0.3	0.19	0.19	0.32	SD	49	60	0.490	0.510	PW
18/07/2007 09:56	1	0	0	1	0	0.3	0.08	0.4	0.23	PW	1	0	0.680	0.320	CS
18/07/2007 11:58	12	6	0	2	4	0.36	0.14	0.21	0.29	CD	8	4	0.610	0.390	CS
18/07/2007 16:26	13	5	1	5	2	0.29	0.22	0.27	0.22	CD	7	6	0.500	0.500	2
20/07/2007 07:10	5	0	2	1	2	0.21	0.25	0.18	0.36	SD	3	2	0.540	0.460	CS
20/07/2007 08:04	0	0	1	0	1	0.25	0.23	0.13	0.39	SD	1	1	0.550	0.450	CS

True species: FPW		4Sp French classifier									2Sp French Classifier				
Encounter Time	n	n				p				Class as	n		p		Class as
		COD	C&S	FPW	STD	COD	C&S	FPW	STD		CSD	FPW	CSD	FPW	
08/07/2007 15:41	1	1	0	0	0	0.4	0.09	0.3	0.2	CD	1	0	0.67	0.33	CS

True species : STD		4Sp French classifier									2Sp French Classifier				
Encounter Time	n	n				p				Class as	n		p		Class as
		COD	C&S	FPW	STD	COD	C&S	FPW	STD		CSD	FPW	CSD	FPW	
06/07/2007 12:01	32	11	6	3	12	0.29	0.2	0.19	0.32	SD	15	17	0.49	0.51	PW
07/07/2007 08:17	21	9	5	2	5	0.34	0.22	0.18	0.26	CD	15	6	0.53	0.47	CS
08/07/2007 08:43	3	0	1	0	2	0.24	0.23	0.09	0.44	SD	0	3	0.39	0.61	PW
11/07/2007 16:38	76	12	7	0	57	0.28	0.16	0.15	0.41	SD	12	64	0.41	0.59	PW
12/07/2007 11:33	1	1	0	0	0	0.38	0.25	0.06	0.32	CD	1	0	0.67	0.33	CS
12/07/2007 14:57	1	1	0	0	0	0.44	0.06	0.13	0.37	CD	1	0	0.51	0.49	CS
13/07/2007 08:04	5	0	0	3	2	0.2	0.12	0.39	0.29	PW	2	3	0.46	0.54	PW
20/07/2007 07:56	3	0	0	0	3	0.28	0.19	0.17	0.36	SD	1	2	0.5	0.50	2
20/07/2007 12:38	1	0	0	0	1	0.14	0.36	0.13	0.37	SD	0	1	0.38	0.62	PW
27/07/2007 08:46	25	13	1	3	8	0.34	0.15	0.19	0.31	CD	17	8	0.55	0.45	CS

Table B-3: Classification of the French encounters, not associated with visual detections, with the 2Sp French classifier and the North Atlantic classifier. n=number of sections per encounter in total (n) and per classification groups (nCSD,nFPW etc..). p=classification probability per classification groups (pCSD,pFPW...).VisualDet=statute of the visual team during the encounters: On effort=visual team was on effort but they did not detect the animals, Off effort=the visual team was Off effort, sonar or electric = description of the sound generating false detections, species name at time=when a species has been observed by the visual team close to the encounter time.

	2Sp French classifier						North Atlantic classifier								
Encounter Time	n	n		p		Class as	n	n			p			Class as	VisualDet
		CSD	FPW	CSD	FPW			BND	CSD	FPW	BND	CSD	FPW		
08/07/2007 11:34	2	0	2	0.4	0.58	LF	3	0	3	0	0.2	0.8	0	CS	Off effort
08/07/2007 12:37	15	5	10	0.5	0.55	LF	6	0	6	0	0.03	0.97	0	CS	Off effort
10/07/2007 18:03	39	21	18	0.5	0.48	CS									Off effort
11/07/2007 06:33	245	152	93	0.5	0.47	CS	48	6	42	0	0.11	0.89	0	CS	Off effort
17/07/2007 06:25	3	3	0	0.8	0.25	CS									Off effort
17/07/2007 07:18	2	2	0	0.6	0.4	CS									Off effort
17/07/2007 19:17	2	2	0	0.6	0.4	CS									BND
18/07/2007 04:46	1	0	1	0.5	0.54	LF									Off effort
19/07/2007 13:06	11	6	5	0.5	0.47	CS	2	0	1	0	0.18	0.45	0.38	CS	On effort
19/07/2007 15:04	3	1	2	0.5	0.49	CS									Off effort
20/07/2007 04:43	39	23	16	0.5	0.49	CS	7	4	2	0	0.62	0.22	0.16	BD	Off effort
20/07/2007 05:00	3	1	2	0.5	0.5	2									Off effort
20/07/2007 06:00	1	0	1	0.5	0.55	LF									On effort
20/07/2007 19:44	3	3	0	0.6	0.36	CS									BND
21/07/2007 19:07	4	4	0	0.7	0.33	CS									BND/COD
23/07/2007 06:27	4	4	0	0.7	0.32	CS									COD
23/07/2007 07:28	5	5	0	0.6	0.42	CS									CS
23/07/2007 07:45	12	8	4	0.6	0.39	CS	1	0	1	0	0	1	0	CS	CS
23/07/2007 09:08	8	6	2	0.6	0.4	CS	1	0	1	0	0	1	0	CS	On effort
23/07/2007 10:32	1	1	0	0.6	0.45	CS									Off effort

	2Sp French classifier						North Atlantic classifier								
Encounter Time	n	n		p		Class as	n	n			p			Class as	VisualDet
		CSD	FPW	CSD	FPW			BND	CSD	FPW	BND	CSD	FPW		
23/07/2007 10:58	2	1	1	0.5	0.55	LF									Off effort
23/07/2007 14:07	3	1	2	0.4	0.63	LF									Off effort
23/07/2007 14:29	2	2	0	0.5	0.46	CS									Off effort
23/07/2007 14:42	1	0	1	0.4	0.6	LF									Off effort
24/07/2007 09:13	9	9	0	0.6	0.37	CS	1	0	1	0	0.04	0.95	0	CS	Off effort

Table B-4: Classification of the Spanish encounters (not associated with visual detections) with the 3Sp Spanish classifier and the North Atlantic classifier. n=number of sections per encounter in total (n) and per classification groups (nBND,nCSD etc..). p=classification probability per classification groups (pBND,pC&S…). VisualDet=statute of the visual team during the encounters: On effort=visual team was on effort but they did not detect the animals, Off effort=the visual team was Off effort, sonar or electric = description of the sound generating false detections, species name at (time)=when a species has been observed by the visual team close to the encounter time.

	3Sp Spanish classifier							North Atlantic classifier								
Encounter Time	n	n			p			Class as	n	n			p			Class as	VisualDet
		BND	CSD	FPW	BND	CSD	FPW			BND	CSD	FPW	BND	CSD	FPW		
30/06/2007 15:37	5	5	0	0	1	0	0	BD									Off effort
02/07/2007 15:24	2	2	0	0	0.76	0.04	0.2	BD									Off effort
04/07/2007 05:12	1	0	1	0	0	1	0	CS									Off effort
04/07/2007 05:29	1	0	1	0	0.1	0.9	0	CS									Off effort
04/07/2007 07:30	2	0	2	0	0	1	0	CS									On effort
04/07/2007 12:29	3	0	3	0	0	1	0	CS									On effort
05/07/2007 06:35	4	0	4	0	0.12	0.88	0	CS	1	0	1	0	0.01	0.99	0	CS	On effort
05/07/2007 07:27	5	4	1	0	0.7	0.2	0.1	BD	1	0	0	0	0	0	1	PL	sonar
05/07/2007 08:01	4	4	0	0	1	0	0	BD									sonar
05/07/2007 08:31	7	7	0	0	0.93	0.01	0.05	BD	1	0	0	0	0	0	1	PL	sonar

Encounter Time	3Sp Spanish classifier								North Atlantic classifier								
	n	n			p			Class as	n	n			p			Class as	VisualDet
		BND	CSD	FPW	BND	CSD	FPW			BND	CSD	FPW	BND	CSD	FPW		
05/07/2007 09:33	1	1	0	0	0.98	0	0.02	BD									sonar
06/07/2007 07:26	27	1	26	0	0.05	0.95	0	CS	7	0	7	0	0	1	0	CS	On effort
06/07/2007 08:09	2	0	2	0	0.22	0.78	0	CS	1	0	1	0	0	1	0	CS	On effort
06/07/2007 09:30	2	1	1	0	0.28	0.72	0	CS	1	0	1	0	0	1	0	CS	CD at 8:27
06/07/2007 10:47	12	0	12	0	0	1	0	CS	1	0	1	0	0.24	0.76	0	CS	On effort
06/07/2007 11:45	1	1	0	0	0.5	0.5	0	2									On effort
06/07/2007 12:19	231	8	217	6	0.06	0.9	0.03	CS	39	11	28	0	0.34	0.66	0	CS	SD at 12:10
07/07/2007 12:39	6	0	6	0	0	1	0	CS	1	0	1	0	0	1	0	CS	On effort
08/07/2007 07:21	1	0	1	0	0.07	0.93	0	CS									On effort
08/07/2007 09:48	1	0	1	0	0	1	0	CS									Off effort
08/07/2007 10:43	49	0	49	0	0.03	0.97	0	CS	7	0	7	0	0.05	0.95	0	CS	On effort
10/07/2007 06:02	1	0	1	0	0.12	0.8	0.07	CS									On effort
10/07/2007 10:57	12	0	12	0	0	1	0	CS	2	2	0	0	0.9	0.1	0	BD	Off effort
10/07/2007 11:30	4	2	2	0	0.48	0.52	0	CS	1	0	1	0	0	1	0	CS	Off effort
10/07/2007 11:59	205	17	185	3	0.14	0.85	0.02	CS	36	3	33	0	0.11	0.88	0.01	CS	Off effort
10/07/2007 16:57	370	9	360	1	0.07	0.93	0	CS	68	2	66	0	0.08	0.92	0	CS	Off effort
11/07/2007 05:23	3	0	3	0	0.05	0.95	0	CS									Off effort
11/07/2007 06:09	4	0	4	0	0.09	0.91	0	CS	1	0	1	0	0	1	0	CS	CD at 6:21
11/07/2007 06:51	3	0	3	0	0.01	0.99	0	CS									On effort
11/07/2007 07:50	2	0	2	0	0.29	0.71	0	CS									On effort
11/07/2007 08:11	27	0	27	0	0.02	0.98	0	CS	9	0	9	0	0	1	0	CS	On effort
11/07/2007 09:14	4	0	4	0	0.04	0.96	0	CS	1	0	1	0	0	1	0	CS	CD at 9:24
11/07/2007 15:05	2	0	2	0	0.02	0.98	0	CS									BND/FPW
11/07/2007 19:40	9	0	9	0	0	1	0	CS	2	0	2	0	0	1	0	CS	Off effort
12/07/2007 17:47	27	1	26	0	0.07	0.93	0	CS	5	0	5	0	0.08	0.92	0	CS	On effort
12/07/2007 19:30	5	0	5	0	0	1	0	CS	1	0	1	0	0	1	0	CS	Off effort
13/07/2007 13:24	4	0	4	0	0	1	0	CS									On effort

Encounter Time	3Sp Spanish classifier								North Atlantic classifier								
	n	n			p			Class as	n	n			p			Class as	VisualDet
		BND	CSD	FPW	BND	CSD	FPW			BND	CSD	FPW	BND	CSD	FPW		
13/07/2007 17:02	9	0	9	0	0.02	0.98	0	CS	2	0	2	0	0	1	0	CS	Off effort
13/07/2007 17:22	8	0	8	0	0	1	0	CS	1	0	1	0	0.16	0.84	0	CS	Off effort
13/07/2007 18:05	189	6	182	1	0.05	0.94	0.01	CS	33	2	31	0	0.05	0.95	0	CS	Off effort
13/07/2007 19:30	12	3	7	2	0.27	0.61	0.12	CS	2	1	0	0	0.5	0.01	0.5	2	Off effort
14/07/2007 12:14	68	4	60	4	0.09	0.84	0.07	CS	10	0	10	0	0.06	0.94	0	CS	Off effort
14/07/2007 17:17	2	0	2	0	0.01	0.99	0	CS									Off effort
14/07/2007 18:40	8	0	8	0	0.01	0.99	0	CS	1	0	1	0	0	1	0	CS	Off effort
18/07/2007 06:59	12	0	12	0	0.01	0.98	0.02	CS	2	1	1	0	0.36	0.64	0	CS	Off effort
18/07/2007 07:46	178	0	123	55	0	0.7	0.29	CS	3	0	3	0	0	1	0	CS	Electric
18/07/2007 08:06									11	0	11	0	0	1	0	CS	Electric
18/07/2007 09:21									8	0	8	0	0	1	0	CS	Electric
18/07/2007 19:01	1	0	1	0	0	1	0	CS									Off effort
20/07/2007 09:57	1	0	1	0	0	1	0	CS									Off effort
20/07/2007 17:15	1	0	1	0	0	1	0	CS									On effort
20/07/2007 17:41	2	0	2	0	0	1	0	CS									On effort
20/07/2007 19:11	1	1	0	0	0.84	0.04	0.13	BD									Off effort
21/07/2007 09:53	3	0	3	0	0	1	0	CS									Off effort
21/07/2007 13:59	9	0	9	0	0	1	0	CS	1	0	1	0	0	1	0	CS	On effort
25/07/2007 06:40	1	0	1	0	0	1	0	CS									On effort
25/07/2007 18:10	1	0	1	0	0.01	0.99	0	CS									Off effort
26/07/2007 14:44	6	0	6	0	0	1	0	CS	1	0	1	0	0	1	0	CS	On effort
27/07/2007 07:07	8	0	8	0	0.01	0.99	0	CS	1	0	1	0	0	1	0	CS	On effort
27/07/2007 09:39	1	0	1	0	0	1	0	CS	8	0	8	0	0	1	0	CS	On effort
27/07/2007 18:09	57	0	57	0	0	1	0	CS	1	0	1	0	0	1	0	CS	Off effort
28/07/2007 06:06	5	0	5	0	0	1	0	CS	1	0	1	0	0	1	0	CS	On effort
28/07/2007 06:41	11	0	10	1	0.02	0.94	0.05	CS	1	0	1	0	0	1	0	CS	On effort
29/07/2007 19:07	2	0	2	0	0	1	0	CS									Off effort

Appendix C. Appendix for chapter 6.

C.1 Analytic estimate of the bias and variance of the true number of detected calls when there is no uncertainty in the values of the confusion matrix.

The notations used in this Appendix are the same as the notations defined in the main body of the text in chapter 6.

The mean of a multinomially distributed random variable y~Multinom(v,p) is (Royle and Dorazio, 2008).

$$E[y_j] = vp_j \qquad \text{C.1}$$

with *v* being the numbers of trials and *p* the event probabilities.

The expected value of a sum is equal to the sum of the expected values

$$E\left[\sum_{j=1} Y_j\right] = \sum_{j=1} E(Y_j) \qquad \text{C.2}$$

In the following, these two expressions (C.1 and C.2) are used to derive the expected values of $\hat{v}$.

The model can be described as

$$E[\hat{\boldsymbol{v}}] = E[C^{-1}.\boldsymbol{n}] \qquad \text{C.3}$$

$$= \mathrm{C}^{-1}\mathrm{E}[\boldsymbol{n}]$$

With $\boldsymbol{v}$ being the true number of detections, $\boldsymbol{C}$ being a constant confusion matrix and $\boldsymbol{n}$ the observed detections.

Since **n** is a sum of several multinomial elements the latter is given by:

$$n_i = y_{i1} + y_{i2} + y_{i3} + y_{i4}$$

$$\text{With } y_{.i} \sim Multinom_j(v_j, \boldsymbol{p_j}) \qquad \text{C.4}$$

$$E[n_i] = \sum_{j=1}^{m} E(y_{ij}) = \sum_{j=1}^{m} v_j p_{ij}$$

The variance and covariance of a multinomial distribution are (Royle and Dorazio 2008):

$$Var(y_j) = vp_j(1 - p_j) \quad \text{C.5}$$

$$cov(y_{i,}y_j) = -vp_ip_j \quad \text{C.6}$$

In general, the variance/covariance of a matrix multiplying an uncorrelated random variable Z is:

$$cov(C\boldsymbol{Z}) = C.cov(\boldsymbol{Z}).C^T \quad \text{C.7}$$

With the model from equation C.3:

$$\begin{aligned} cov(\boldsymbol{v}) &= cov(C^{-1}.\boldsymbol{n}) \\ &= c^{-1}cov(\boldsymbol{n})C^{-1^T} \end{aligned} \quad \text{C.8}$$

Again identifying $\boldsymbol{n}$ as the sum of multinomial random variables:

$$cov(\boldsymbol{n}) \quad \text{C.9}$$
$$= \begin{bmatrix} var(n_i) & \cdots & cov(n_m, n_m) & \cdots & cov(n_1, n_m) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(n_i, n_1) & \cdots & var(n_j) & \cdots & cov(n_i, n_j) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(n_m, n_1) & \cdots & cov(n_m, n_j) & \cdots & var(n_m) \end{bmatrix}$$

with

$$var(n_i) = \sum_{j=1}^{m} var(y_{ij}) = \sum_{j=1}^{m} v_j p_{ij}(1 - p_{ij}) \quad \text{C.10}$$

and

$$cov(\boldsymbol{n_i}, \boldsymbol{n_k}) = \sum_j cov(y_{ij}, y_{kj}) = -\sum_j v_j p_{ij} p_{kj} \quad \text{C.11}$$

C.2: Tables

Table C-1: Analytically derived mean of the expected true number of calls, E[$\hat{v}$], and coefficient of variation (CV, expressed as a percentage).

Confusion Matrix	Scenario 1 (equal data)				Scenario 2 (unequal data)			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
a	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**8000** *(0%)*	**3000** *(0%)*	**950** *(0%)*	**50** *(0%)*
b	**3000** *(1.19%)*	**3000** *(1.19%)*	**3000** *(1.19%)*	**3000** *(1.19%)*	**8000** *(0.54%)*	**3000** *(1.19%)*	**950** *(3.34%)*	**50** *(59.9%)*
c	**3000** *(1.12%)*	**3000** *(1.36%)*	**3000** *(1.14%)*	**3000** *(1.17%)*	**8000** *(0.57%)*	**3000** *(1.48%)*	**950** *(2.91%)*	**50** *(43.85%)*
d	**3000** *(4.10%)*	**3000** *(4.10%)*	**3000** *(4.10%)*	**3000** *(4.10%)*	**8000** *(1.75%)*	**3000** *(4.10%)*	**950** *(12.13%)*	**50** *(223.51%)*
e	**3000** *(3.98%)*	**3000** *(3.00%)*	**3000** *(4.07%)*	**3000** *(4.96%)*	**8000** *(1.59%)*	**3000** *(3.29%)*	**950** *(10.66%)*	**50** *(299.92%)*

Table C-2: Simulation result, without uncertainty in the confusion matrix, of the mean of the estimates of the true number of calls E[$\hat{v}$] , and coefficient of variation (CV, expressed as a percentage).

	Scenario 1.				Scenario 2.			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Scenario x.a	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**8000** *(0%)*	**3000** *(0%)*	**950** *(0%)*	**50** *(0%)*
Scenario x.b	**2999.93** *(1.18%)*	**3000.12** *(1.18%)*	**3000.01** *(1.19%)*	**2999.94** *(1.18%)*	**8000.37** *(0.55%)*	**2999.47** *(1.19%)*	**950.14** *(3.67%)*	**50.02** *(59.89%)*
Scenario x.c	**3000.69** *(1.12%)*	**2998.99** *(1.36%)*	**3000.14** *(1.15%)*	**3000.18** *(1.17%)*	**7999.46** *(0.56%)*	**3000.40** *(1.49%)*	**949.95** *(2.94%)*	**50.19** *(43.7%)*
Scenario x.d	**2999.87** *(4.09%)*	**3001.49** *(4.14%)*	**2998.55** *(4.08%)*	**3000.09** *(4.12%)*	**8000.74** *(1.75%)*	**3000.72** *(4.08%)*	**949.64** *(12.14%)*	**48.90** *(229.82%)*
Scenario x.e	**2997.28** *(4.03%)*	**3002.00** *(2.98%)*	**3000.30** *(4.07%)*	**3000.41** *(4.92%)*	**7999.63** *(1.59%)*	**3000.88** *(3.27%)*	**948.58** *(10.69%)*	**50.92** *(295.94%)*

Table C-3: Simulation result, with a low level of uncertainty in the confusion matrix, of the mean of the estimates of the true number of calls E[$\hat{v}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1.				Scenario 2.			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Sc x.a	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**8000** *(0%)*	**3000** *(0%)*	**950** *(0%)*	**50** *(0%)*
Sc x.b	**3000.11** *(6.51%)*	**3000.58** *(6.58%)*	**2999.38** *(6.61%)*	**2999.92** *(6.54%)*	**8000.48** *(4.60%)*	**2999.24** *(8.57%)*	**949.98** *(24.85%)*	**50.30** *(467.87%)*
Sc x.c	**2999.72** *(6.68%)*	**2999.89** *(6.54%)*	**3000.13** *(6.57%)*	**3000.25** *(6.61%)*	**7999.70** *(4.60%)*	**3000.05** *(8.58%)*	**950.17** *(25.19%)*	**50.07** *(471.00%*
Sc x.d	**3002.12** *(22.90%)*	**2996.36** *(22.77%)*	**3001.90** *22.25*	**2999.35** *22.81*	**7998.41** *14.47*	**3000.28** *30.81*	**950.71** *92.89*	**50.60** *1722.71*
Sc x.e	**2999.25** *(21.00%)*	**2999.79** *(17.48%)*	**2999.06** *(21.97%)*	**3001.90** *(28.79%)*	**8001.65** *(13.42%)*	**2999.24** *(19.90%)*	**950.78** *(105.79%)*	**48.32** *2578.82%*

Table C-4: Simulation result, with a high level of uncertainty in the confusion matrix, of the means of the estimates of the true number of calls E[$\hat{v}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1.				Scenario 2.			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Sc x.a	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**3000** *(0%)*	**8000** *(0%)*	**3000** *(0%)*	**950** *(0%)*	**50** *(0%)*
Sc x.b	**2999.94** *(61.89%)*	**3000.14** *(61.28%)*	**3000.02** *(62.65%)*	**2999.90** *(61.70%)*	**8000.04** *(42.64%)*	**3000.03** *(80.85%)*	**949.97** *(226.46%)*	**49.94** *4485.69%*
Sc x.c	**2999.96** *(62.53%)*	**3000.00** *(60.69%)*	**3000.06** *(65.51%)*	**2999.98** *(62.27%)*	**7999.79** *(44.42%)*	**3000.11** *(83.19%)*	**950.01** *(236.21%)*	**50.08** *4490.55%*
Sc x.d	**3000.26** *(214.59%)*	**2999.97** *217.66%*	**2999.84** *212.69%*	**2999.94** *218.69%*	**8000.43** *101.44%*	**2999.42** *214.96%*	**949.61** *646.61%*	**50.53** *12788.65%*
Sc x.e	**2999.66** *195.02%*	**2999.97** *164.58%*	**3000.15** *200.83%*	**3000.22** *274.79%*	**8000.13** *93.18%*	**2999.67** *138.27%*	**949.83** *751.36%*	**50.37** *16944.28%*

Appendix D. Appendix for chapter 7

Table D-1: Values of the parameters α for the Dirichlet prior distributions, for each species, each scenario and each set of priors for p. The parameters α were selected such that: the CV of the correct classification probabilities (diagonal element) was equal to 1%, 40% and 77% and the means of the prior distribution were equal to the classification probabilities of the scenarios for P1 and P2 whereas for prior P3, the mean distribution was equal to 0.25 for each species.

Scenarios	α	Prior P1				Prior P2				Prior P3
Scx.b	α_1	1499.15	88.19	88.19	88.19	0.088	0.005	0.005	0.005	1
	α_2	88.19	1499.15	88.19	88.19	0.005	0.088	0.005	0.005	1
	α_3	88.19	88.19	1499.15	88.19	0.005	0.005	0.088	0.005	1
	α_4	88.19	88.19	88.19	1499.15	0.005	0.005	0.005	0.088	1
Scx.c	α_1	1499.15	141.10	35.27	17.64	$8.8\ 10^{-2}$	4.10^{-4}	01.10^{-4}	5.10^{-5}	1
	α_2	176.4	1499.15	52.91	158.73	5.10^{-4}	$8.8\ 10^{-2}$	$1.5.10^{-4}$	$4.5.10^{-4}$	1
	α_3	52.91	88.19	1499.15	88.19	1.510^{-4}	0.005	$8.8\ 10^{-2}$	0.005	1
	α_4	35.27	35.27	176.4	1499.15	1.10^{-4}	1.10^{-4}	5.10^{-4}	$8.8\ 10^{-2}$	1
Scx.d	α_1	4799.48	1476.76	1476.76	1476.76	2.48	0.76	0.76	0.76	1
	α_2	1476.76	4799.48	1476.76	1476.76	0.76	2.48	0.76	0.76	1
	α_3	1476.76	1476.76	4799.48	1476.76	0.76	0.76	2.48	0.76	1
	α_4	1476.76	1476.76	1476.76	4799.48	0.76	0.76	0.76	2.48	1

Scx.e	α_1	4799.48	369.2	1846.0	1846.0	2.48	0.19	0.95	0.20	1
	α_2	1384.5	4799.48	119.9	461.5	0.71	2.48	0.62	0.24	1
	α_3	923	1292.2	4799.48	2122.29	0.48	0.67	2.48	1.1	1
	α_4	2122.9	2769.0	1384.5	4799.48	1.1	1.43	0.71	2.48	1

Table D-2: Convergence test results for each model A and species. Y indicates that the chains for the corresponding species converged. The 0 value in Sc2.d and Sc2.e with prior V3 indicates that the posterior distribution for species D had stopped converging and the mean of this posterior distribution was 0. Grey cells indicate models that were found to be sensitive to the initial values of the Markov chains.

	Prior	**V1**	**V2**	**V3**
	Species	**ABCD**	**ABCD**	**ABCD**
Equal data	**Sc1.a**	YYYY	YYYY	
	Sc1.b	YYYY	YYYY	
	SC1.c	YYYY	YYYY	
	Sc1.d	YYYY	YYYY	
	Sc1.e	YYYY	YYYY	
Unequal data	**Sc2.a**	YYYY	YYYY	YYYY
	Sc2.b	YYYY	YYYY	YYYY
	Sc2.c	YYYY	YYYY	YYYY
	Sc2.d	YYYY	YYYY	YYY0
	Sc2.e	YYYY	YYYY	YYY0

Table D-3: Summary of the convergence test results for the Posterior distribution of the parameters *v* and *p* for all models B. Y indicates that the chains for the corresponding species converged whereas N indicates they did not converged.. Sc=Scenario for the different confusion matrices (Scx a to Scx e) and for the equal(Sc1.) and unequal dataset Sc2. The grey cells indicate models sensitive to the initial values of the Markov chains.

	Prior on *v*	**V1**			**V2**			**V3**		
	Prior on *p*	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**
	Parameters	***v p***	***v p***	***v p***	***v p***	***v p***	***v p***	***v p***	***v p***	***v p***
Equal data	**Sc1.a**									
	Sc1.b	YY	YY	YY	YY	YY	YY			
	Sc1.c	YY	YY	YY	YY	YY	YY			
	Sc1.d	YY	YY	YY	YY	YY	YY			
	Sc1.e	YY	YY	YY	YY	YN	YN			
Unequal data	**Sc2.a**									
	Sc2.b	YY	YY	YY	YY	YN	YY	YY	YY	YY
	Sc2.c	YY	YY	YY	YY	YY	YY	YY	YY	YY
	Sc2.d	YY	YY	YY	YY	YY	YY	YY	YY	YY
	Sc2.e	YY	YY	YY	YY	YY	YY	YY	YY	YY

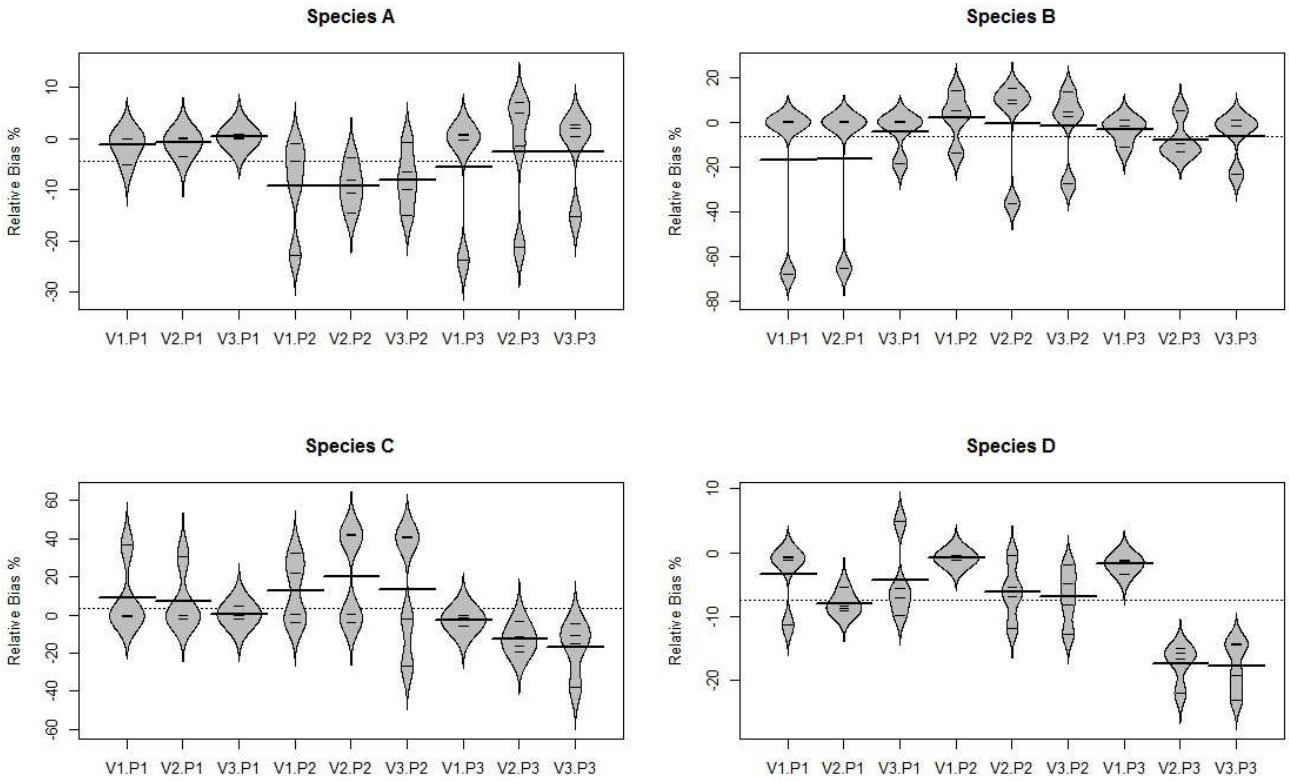


Figure D-1 : Beanplots of the estimates relative bias as a function of the priors on *v* and *p* in the models for each species. The bold lines are the mean relative bias for each beanplot whereas the dotted lines are the mean of the relative bias across all models for one species.

Appendix E. R Codes for models A without uncertainty on the classification rates:

```
library(MCMCpack)
#-----------------------------------------------------------------------------------------------------------------
#Equal Data priors and confusion matrices
#-----------------------------------------------------------------------------------------------------------------
#Prior parameters
prior.n1<- matrix(c(3000,90000,3000,90000,3000,90000,3000,90000),2,4)#variance such as CV =10%
prior.n2<-matrix(c(3000,1.3e6,3000,1.3e6,3000,1.3e6,3000,1.3e6),2,4)#variance such as CV=40%
#Confusion matrices
CM0 <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
CM1 <- matrix(c(0.85,0.05,0.05,0.05,0.05,0.85,0.05,0.05,0.05,0.05,0.85,0.05,0.05,0.05,0.05,0.85),4,4)
aCM1<- matrix(c(0.85,0.10,0.03,0.02,0.08,0.85,0.05,0.02,0.02,0.03,0.85,0.10,0.01,0.09,0.05,0.85),4,4)
CM2 <- matrix(c(0.52,0.16,0.16,0.16,0.16,0.52,0.16,0.16,0.16,0.16,0.52,0.16,0.16,0.16,0.16,0.52),4,4)
aCM2 <- matrix(c(0.52,0.15,0.10,0.23,0.04,0.52,0.14,0.30,0.20,0.13,0.52,0.15,0.20,0.05,0.23,0.52),4,4


#CM0-------------------------------------------------------------------------------------------------------------------
#initial parameters :
initparamY.SC0.1<- matrix(c(3000,1,1,1,3003,1,3000,1,1,3003,1,1,3000,1,3003,1,1,1,3000,3003),5,4)
initparamY.SC0.2<-
matrix(c(900,550,140,350,1940,400,1025,540,825,2790,450,450,2000,200,3100,200,200,200,2700,3300),5,4)
#models :
NoD_Sc0_1<- MH.Obs(c(3000,3000,3000,3000),CM0,500,100000,50000,initparamY.SC0.1,prior.n2,prior.n1)
#CM1-------------------------------------------------------------------------------------------------------------------
#initial parameters :
initparamY.SC1.1<-
matrix(c(2550,150,150,150,3000,150,2550,150,150,3000,150,150,2550,150,3000,150,150,150,2550,3000),5,4)
initparamY.SC1.2<-
matrix(c(900,550,140,350,1940,400,1025,540,825,2790,450,450,2000,200,3100,200,200,200,2700,3300),5,4)
#models :
NoD_Sc1_1<-
MH.Obs(c(3000,3000,3000,3000),CM1,500,200000,100000,initparamY.SC1.1,initparamY.SC1.2,prior.n2)
#aCM1-------------------------------------------------------------------------------------------------------------------
#initial parameters :
initparamY.SC2.1<-
matrix(c(2550,300,90,60,3000,240,2550,150,60,3000,50,90,2550,310,3000,30,270,150,2550,3000),5,4)
initparamY.SC2.2<-
matrix(c(2500,350,150,200,3200,400,1900,300,200,2800,300,100,2600,400,3400,280,320,470,1630,2700),5,4)
#models :
```

```
NoD_Sc2_1 <-
MH.Obs(c(3000,3000,3000,3000),aCM1,500,200000,100000,initparamY.SC2.1,initparamY.SC2.2,prior.n2)
#CM2----------
#initial parameters :
initparamY.SC3.1<-
matrix(c(1560,480,480,480,3000,480,1560,480,480,3000,480,480,1560,480,3000,480,480,480,1560,3000),5,4)
initparamY.SC3.2<-
matrix(c(1900,650,40,350,2940,400,1025,540,825,2790,450,450,2000,200,3100,200,200,200,2700,3300),5,4)
#models :
NoD_Sc3_1<-
MH.Obs(c(3000,3000,3000,3000),CM2,500,200000,100000,initparamY.SC3.1,initparamY.SC3.2,prior.n2)
#aCM2----------
# initial parameters :
initparamY.SC4.1<-
matrix(c(2550,300,90,60,3000,240,2550,150,60,3000,50,90,2550,300,3000,30,270,150,2550,3000),5,4)
 initparamY.SC4.2<-
matrix(c(2500,350,150,200,3200,400,1900,300,200,2800,300,100,2600,200,3200,280,320,470,1630,2700),5,4)
# models :
NoD_Sc4_1 <-
MH.Obs(c(3000,3000,3000,3000),aCM2,500,200000,100000,initparamY.SC4.1,initparamY.SC4.2,prior.n2)

#----------
#Unequal Data priors and confusion matrices
#----------
#Priors on the true number of detections:
prior.unBn1<- matrix(c(8000,1.25*10^6,3000,1.8*10^5,950,1.8*10^4,50,51),2,4)#var CV14% as if 10% for
speccies D var<mean and not possible with negbinomial
prior.unBn2<- matrix(c(8000,9.2*10^6,3000,1.3*10^6,950,1.31*10^5,50,361),2,4)#var CV40%
prior.unBn3<- matrix(c(8000,6.4*10^5,3000,3.6*10^5,950,1.31*10^5,50,900),2,4)#CV variable with
10%,20%,40% and 60% from common to rare species
prior.unBn4<- matrix(c(8000,6.4*10^5,3000,3.6*10^5,950,1.31*10^5,50,361),2,4)#CV variable with
10%,20%,40% and 40% from common to rare species
#CM0----------
#initial parameters :
initparamY.SC5.1<- matrix(c(8000,1,1,1,8003,1,3000,1,1,3003,1,1,950,1,953,1,1,1,50,53),5,4)
initparamY.SC5.2<- matrix(c(7800,50,50,100,8000,90,2500,175,135,3000,50,25,800,75,950,8,7,5,30,50),5,4)
#Model :
NoD_Sc5_1<-
MH.Obs(c(8000,3000,950,50),CM0,500,100000,50000,initparamY.SC5.2,initparamY.SC5.1,prior.unBn2)
```

```
#CM1----------
#initial parameters :
initparamY.SC6.1<-
matrix(c(6800,400,400,400,8000,2550,150,150,150,3000,47,48,807,48,950,2,3,2,43,50),5,4)
initparamY.SC6.2<-
matrix(c(6900,580,340,380,8200,400,1025,540,825,2790,50,125,800,75,1050,4,6,9,41,60),5,4)
#Model :
NoD_Sc6_1<-
MH.Obs(c(8000,3000,950,50),CM1,500,100000,50000,initparamY.SC6.2,initparamY.SC6.1,prior.unBn2)
#aCM1----------
#initial parameters :
initparamY.SC7.1<- matrix(c(6800,800,240,160,8000,240,2550,150,60,3000,19,28,808,95,950,1,4,2,43,50),5,4)
initparamY.SC7.2<-
matrix(c(6500,900,340,260,8000,210,2500,250,40,3000,30,58,700,162,950,0,10,15,25,50),5,4)
#Model :
NoD_Sc7_1<-
MH.Obs(c(8000,3000,950,50),aCM1,500,100000,50000,initparamY.SC7.2,initparamY.SC7.1,prior.unBn2)
#CM2----------
#initial parameters :
initparamY.SC8.1<-
matrix(c(4160,1280,1280,1280,8000,480,1560,480,480,3000,152,152,494,152,950,8,8,8,26,50),5,4)
initparamY.SC8.2<-
matrix(c(3900,1650,1100,1350,8000,610,1025,540,825,3000,125,50,675,100,950,3,4,1,42,50),5,4)
#Models :
NoD_Sc8_1<-
MH.Obs(c(8000,3000,950,50),CM2,500,100000,50000,initparamY.SC8.2,initparamY.SC8.1,prior.unBn2)
#aCM2----------
#initial parameters :
initparamY.SC9.1<-
matrix(c(4160,1200,800,1840,8000,240,2550,150,60,3000,190,123,494,143,950,10,2,12,26,50),5,4)
initparamY.SC9.2<-
matrix(c(4500,1150,950,1400,8000,400,2000,400,200,3000,210,100,500,140,950,4,6,10,30,50),5,4)
#Models :
NoD_Sc9_1<-
MH.Obs(c(8000,3000,950,50),aCM2,500,100000,50000,initparamY.SC9.2,initparamY.SC9.1,prior.unBn2)
#----------
#FUNCTIONS
#----------
#Routine to run Metropolis Hasting model
MH.Obs <- function(n.simul,#true number of detections for each species used to generate the data used
```

```
            CM,#CM use to simulate data
            data.its,# number of "bootstrap"
            nits,# number of iteration in the MCMC
            nburn,#burn in size
            init.param1,#initial values for chain1 for parameters y's
            init.param2,#initial values for chain 2 for parameters y's
            prior.param #prior parameters for y's parameters
)
{
  library(MCMCpack)
 library(coda)
  Allmean <- Allmean2 <- array(NA,c(data.its,4,4))#table to save the mean of the parameters for eahc chains
 Allsd <- Allsd2<- array(NA,c(data.its,4,4))
 AllAR <- AllAR2 <- matrix(NA,data.its,ncol(init.param1)) #contains all Acceptance Rate for each bootstrap
 colnames(Allmean)<-colnames(Allmean2)<-c("Mean","Sd","95Low","95High")
 printseq<-seq(1,data.its,100)

 if(data.its==1)
 {data.yest<-matrix(round(rowMeans(data.sim(n.simul,1000,CM)[[1]])),(length(n.simul)),1)}
 else {data.yest<-data.sim(n.simul,data.its,CM)[[1]]}
 #data.yest$Y<-round(apply(data.yest$Y,1,mean))#use this when only 1 simulation

 for (z in 1:data.its){ #for eahc simulated dataset
   Result <- met.hasObs (nits,data.yest[,z],nburn,prior.param, init.param1,CM )
   Result2 <- met.hasObs (nits,data.yest[,z],nburn,prior.param, init.param2,CM )
   ndraw1 <- mcmc(t(Result[[1]][(length(n.simul)+1),,]))
   ndraw2 <- mcmc(t(Result2[[1]][(length(n.simul)+1),,]))

   Allmean [z,1:2,] <- t(summary(ndraw1)$statistics[,1:2])
   Allmean [z,3:4,] <- t(summary(ndraw1)$quantiles[,c(1,5)])
   AllAR[z,] <- Result[[2]]

   Allmean2 [z,1:2,] <- t(summary(ndraw2)$statistics[,1:2])
   Allmean2 [z,3:4,] <- t(summary(ndraw2)$quantiles[,c(1,5)])
   AllAR2[z,] <- Result2[[2]]
   #print(date())
   print(z)
   return(list(PostMeans = Allmean,PostMeans2 = Allmean2,  AccepRate = AllAR, AccepRate2 = AllAR2,
initialValues = init.param1,priorParam = c(prior.param),CM = CM))
 }
```

```
}

#Function to simulate the data
data.sim<-function(n,#number of detection for each species
                   nits,#how many replicate of the data
                   CM#confusion matrix
)
{
 #simulate data (y's) from confusion matrix and true number of data (n's)
 s = dim(CM)[1]
 cont.n <- array(0,c(s))
 y.unknown <- array(0,c(s,nits,s))
 Y <- array(0,c(s,nits))
 nest.mean = array(0,c(s))
 for(j in 1:s){
   prob <- CM[1:s,j]
  y.unknown[,,j]<-rmultinom(nits,n[j],prob)
 }
 for (j in 1:s)
 { for (i in 1:nits)
  {Y[j,i] <- sum(y.unknown[j,i,1:s])
  }
 }
 return(list(Y=Y,n=n))
}

library(msm)
#Function to run each iteration of the MCMC
met.hasObs <- function(nits,# number of iteration in the MCMC
                       simul.y,#simulated true data
                       nburn,#burn in size
                       prior.param,#prior parameters for param y's
                       param,#initila value of one chian for y's parems
                       CM #p's values from CM used to simulate data
) {
 nSp <-ncol(param)
 Data = simul.y
 samples <-array(0,c(dim(param),nits))

 #Calculate likelihood or log(likelihood)
```

```
 likhood <- CalcObs(param,CM)

 ##measure the acceptance rate
 AcceRate <- matrix(,nits,nSp)

 #MCMC update
  for (t in 1:nits){
  #Update the parameters in the model using function "updateparam"
  output <- updateparamObs(nSp,param,CM,Data,likhood,prior.param)
  AcceRate[t,] <- t(output$accep.rate)
  param<-t(output$param)
  likhood <- output$likhood[1]
  samples [,,t] <- param
  }

#calculate the mean and standart deviation of the parameters following burn-in:
 subsample<-samples[,,(nburn+1):nits]
 AcceptanceRate <- colMeans(AcceRate)

 return(list(subsample,AcceptanceRate))
}

#function to calculate likelihood
CalcObs <- function(param,CM){
 nsp<-ncol(param)
 PartialLik <- numeric(nsp)
 for (sp in 1:nsp){
  PartialLik[sp] <- dmultinom(param[1:nsp,sp],param[(nsp+1),sp],CM[,sp],log=TRUE)
 }
 likhood <- sum(PartialLik)
 return(likhood)
}

#Function to update the parameters in the MCMC
updateparamObs <-function(nSp,# nbs of species
                          param,#inital values for the y's param
                          CM, #p values from CM used to simulate data
                          Data,#true data
                          likhood,#likelhood estiamted with ald param
                          prior.param# prior parameters for the y's param
```

```
){
   oldparam <- matrix(,2,nSp)
  accep.rate <-matrix(0,nSp,1)

 for (i in 1:nSp){
   #conserve old parameters
   oldparam[1,] <- param[i,]
   oldparam[2,] <- param[(nSp+1),]
   #Propose new parameters
   param[i,]<-(rmultinom(1,Data[i],(param[nSp+1,]*CM[i,])/sum(param[nSp+1,]*CM[i,])))
   if(sum(param[1:nSp,nSp])==0){param[i,nSp=1]}
   param[(nSp+1),] <- colSums(param[1:nSp,])

   #Calculate the new likelihood value for the proposed moved:

   newlikelihood<-CalcObs(param,CM)
   if(newlikelihood==0 || is.na(newlikelihood)==TRUE){print(param)
                                        print("Log lik not valid")}
   #Include the likelihood term in the acceptance probability
   num <- newlikelihood +
npriorObs(nSp,param[(nSp+1),],prior.param)+dmultinom(oldparam[1,],prob=((oldparam[2,]*CM[i,])/sum(oldp
aram[2,]*CM[i,])),log=TRUE)
   den <- likhood  +
npriorObs(nSp,oldparam[2,],prior.param)+dmultinom(param[i,],prob=((param[nSp+1,]*CM[i,])/sum(param[nS
p+1,]*CM[i,])),log=TRUE)

   #Acceptance probability:
   A<-min(1,exp(num-den))#if the difference is positive the min will be 1 so we will accept the move. If the
difference is negative, the min will be exp(num-den) so the move will be accepted in function of the uniform
distribution below.
   accep.rate[i,1]<-A

   # Simulate a random number in [0,1] and accept move with probability A;
   # else reject move and return parameter value to previous value
   u <- runif(1)
   if (u <= A) { likhood <- newlikelihood
   }
   else { param[i,] <- oldparam[1,]
        param[(nSp+1),] <- oldparam[2,]
   }
```

```
  }
  #set the values to be outputted from the function to be the
  #set of parameter values and log)likelihood) value:
  output <- list(param=t(param),likhood=likhood,accep.rate=accep.rate)
  #output the parameter values:
  output
}

#Function to generate prior on the true number of detections
npriorObs <- function(nSp,nparam,prior.param){
  #neg binomial prior
  prior <- numeric(nSp)
  for (m in 1:nSp){
    #prior[m] <- log((prior.param[1,m]+nparam[m]-1)param[m]) + nparam[m] * log(1/(prior.param[2,m]+1))
    alpha<-(prior.param[1,m])^2/(prior.param[2,m]-prior.param[1,m])
    pparam<-alpha/(alpha+prior.param[1,m])
    # prior[m] <- log(factorial(alpha+nparam[m]-1))-log(factorial(nparam[m]) + nparam[m] * log(1-pparam))
    prior[m]<-dnbinom(nparam[m],size=alpha,mu=prior.param[1,m],log=TRUE)
  }
  prior = sum(prior)
  return(prior)
}
```

Appendix F. R Codes for models B with uncertainty on the classification rates:

```
library(MCMCpack)
#----------------EQUAL DATA----------------------------------------------------------------------------------------
#Prior on true number of detections
prior.n1<- matrix(c(3000,90000,3000,90000,3000,90000,3000,90000),2,4)#variance such as CV =10%
prior.n2<-matrix(c(3000,1.3e6,3000,1.3e6,3000,1.3e6,3000,1.3e6),2,4)#variance such as CV=40%
#Priors on classification rates for each confusion matrices
CM0 <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
Sca.prior.p1 <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
initparamP.0<-CM0 #initial parameters

CM1 <- matrix(c(0.85,0.05,0.05,0.05,0.05,0.85,0.05,0.05,0.05,0.05,0.85,0.05,0.05,0.05,0.05,0.85),4,4)
Scb.prior.p1 <-
matrix(c(1499.15,88.19,88.19,88.19,88.19,1499.15,88.19,88.19,88.19,88.19,1499.15,88.19,88.19,88.19,88.19,1
499.15),4,4) #CV 1% for Correct classification Rates
Scb.prior.p2 <- matrix(c(85,5,5,5,5,85,5,5,5,5,85,5,5,5,5,85),4,4)#CV=4%
Scb.prior.p3 <-
matrix(c(0.088,0.005,0.005,0.005,0.005,0.088,0.005,0.005,0.005,0.005,0.088,0.005,0.005,0.005,0.005,0.088),4,
4)#CV 40% for correct classification rates
Scb.prior.p4<- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),4,4)
initparamP.1<-CM1

aCM1<- matrix(c(0.85,0.10,0.03,0.02,0.08,0.85,0.05,0.02,0.02,0.03,0.85,0.10,0.01,0.09,0.05,0.85),4,4)
Scc.prior.p1 <-
matrix(c(1499.15,176.37,52.91,35.27,141.1,1499.15,88.19,35.27,35.27,52.91,1499.15,176.37,17.64,158.73,88.1
9,1499.15),4,4)#CV 1% for Correct classification Rates
Scc.prior.p2 <- matrix(c(85,10,3,2,8,85,5,2,2,3,85,10,1,9,5,85),4,4)#CV=4%
Scc.prior.p3 <-
matrix(c(0.088,0.01,0.003,0.002,0.008,0.088,0.005,0.002,0.002,0.003,0.088,0.01,0.001,0.009,0.005,0.088),4,4)
#CV 40% for correct classification rates
Scc.prior.p4<- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),4,4)
initparamP.a1<-aCM1

CM2<-matrix(c(0.52,0.16,0.16,0.16,0.16,0.52,0.16,0.16,0.16,0.16,0.52,0.16,0.16,0.16,0.16,0.52),4,4)
Scd.prior.p1 <-
matrix(c(4799.48,1476.76,1476.76,1476.76,1476.76,4799.48,1476.76,1476.76,1476.76,1476.76,4799.48,1476.7
6,1476.76,1476.76,1476.76,4799.48),4,4)#CV 1% for Correct classification Rates
Scd.prior.p2 <- matrix(c(52,16,16,16,16,52,16,16,16,16,52,16,16,16,16,52),4,4)#CV=4%
```

```
Scd.prior.p3 <-
matrix(c(2.48,0.76,0.76,0.76,0.76,2.48,0.76,0.76,0.76,0.76,2.48,0.76,0.76,0.76,0.76,2.48),4,4)#CV 40% for
correct classification rates
Scd.prior.p4<- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),4,4)
initparamP.2<-CM2

aCM2<-matrix(c(0.52,0.15,0.10,0.23,0.04,0.52,0.14,0.30,0.20,0.13,0.52,0.15,0.20,0.05,0.23,0.52),4,4)
Sce.prior.p1 <-
matrix(c(4799.48,1385.81,923.88,2124.92,369.55,4799.48,1293.43,2271.63,1847.75,1201.04,4799.48,1385.81,
1847.75,461.94,2124.92,4799.48),4,4)#CV 1% for Correct classification Rates
Sce.prior.p2 <- matrix(c(52,15,10,23,4,52,14,30,20,13,52,15,20,5,23,52),4,4)#CV=4%
Sce.prior.p3 <-
matrix(c(2.48,0.71,0.48,1.09,0.19,2.48,0.67,1.43,0.95,0.62,2.48,0.71,0.95,0.24,1.09,2.48),4,4)#CV 40% for
correct classification rates
Sce.prior.p4<- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),4,4)
initparamP.a2<-aCM2

#initial parameters on the true number of detections
initparamY.1<-
matrix(c(2550,150,150,150,3000,150,2550,150,150,3000,150,150,2550,150,3000,150,150,150,2550,3000),5,4)
initparamY.2<-
matrix(c(1900,550,140,350,2940,400,1025,540,825,2790,450,450,1700,500,3100,200,200,200,2700,3300),5,4)
initparamY.3<-
matrix(c(800,50,40,50,940,1400,1525,540,825,4290,450,450,1700,500,3100,1200,700,700,1700,4300),5,4)

##RUN MODELS and play with different priors#####
#CM0----------
D_Sc0_nP <-
MH.ObsDir(c(3000,3000,3000,3000),CM0,300,300000,150000,initparamY.1,initparamP.0,prior.n1,Sca.prior.p1
)
#CM1----------
D_Sc1_nP1 <-
MH.ObsDir(c(3000,3000,3000,3000),CM1,300,300000,150000,initparamY.1,initparamP.1,prior.n1,Scb.prior.p
1)
#aCM1----------
D_Sc2_nP1 <-
MH.ObsDir(c(3000,3000,3000,3000),aCM1,300,300000,150000,initparamY.1,initparamP.a1,prior.n1,Scc.prior.
p1)
#CM2----------
```

```
D_Sc3_nP1 <-
MH.ObsDir(c(3000,3000,3000,3000),CM2,300,300000,150000,initparamY.1,initparamP.2,prior.n1,Scd.prior.p
1)
##aCM2----------
D_Sc4_nP1 <-
MH.ObsDir(c(3000,3000,3000,3000),aCM2,300,300000,150000,initparamY.1,initparamP.a2,prior.n1,Sce.prior.
p1)
#----------
#UNBALANCED DATA
#----------
```

##priors on the true number of detections

```
prior.unBn1<- matrix(c(8000,1.25*10^6,3000,1.8*10^5,950,1.8*10^4,50,51),2,4)#var CV14% as if 10% for
speccies D var<mean and not possible with negbinomial
prior.unBn2<- matrix(c(8000,9.2*10^6,3000,1.3*10^6,950,1.31*10^5,50,361),2,4)#var CV40%
prior.unBn3<- matrix(c(8000,6.4*10^5,3000,3.6*10^5,950,1.31*10^5,50,361),2,4)#CV variable with
10%,20%,40% and 40% from common to rare species
```

##Initial parameters on the true number of detections

```
initparamY.SC6.1<-
matrix(c(6800,400,400,400,8000,2550,150,150,150,3000,47,48,807,48,950,2,3,2,43,50),5,4)
initparamY.SC6.2<-
matrix(c(6900,580,340,380,8200,400,1025,540,825,2790,50,125,800,75,1050,4,6,9,41,60),5,4)

#CM0----------
D_Sc5_nP <-
MH.ObsDir(c(8000,3000,950,50),CM0,300,300000,150000,initparamY.SC6.1,initparamP.0,prior.unBn1,Sca.pr
ior.p1)
#CM1----------
D_Sc6_nP1 <-
MH.ObsDir(c(8000,3000,950,50),CM1,300,300000,150000,initparamY.SC6.1,initparamP.1,prior.unBn1,Scb.pr
ior.p1)
#aCM1----------
D_Sc7_nP1 <-
MH.ObsDir(c(8000,3000,950,50),aCM1,300,300000,150000,initparamY.SC6.1,initparamP.a1,prior.unBn1,Scc.
prior.p1)
#CM2----------
D_Sc8_nP1 <-
MH.ObsDir(c(8000,3000,950,50),CM2,300,300000,150000,initparamY.SC6.1,initparamP.2,prior.unBn1,Scd.pr
ior.p1)
#aCM2----------
```

```
D_Sc9_n1P1 <-
MH.ObsDir(c(8000,3000,950,50),aCM2,1,300000,295000,initparamY.SC6.1,initparamP.a2,prior.unBn1,Sce.prior.p1)
```

FUNCTIONS

```
#Routine to run Metropolis Hasting model
MH.ObsDir <- function(n.simul,#true number of detection for each species used to generate the data used in the MH
                      CM,#CM use to simulate data
                      data.its,# number of "bootstrap"
                      nits,# number of iteration in the MCMC
                      nburn,#burn in size
                      init.param1,#initial values for chain1 for parameters y's
                      init.pparam1,#initial values for parameters p's (only 1 chain)
                      prior.param,#prior parameters for y's parameters
                      pprior.param# prior parameters for p's parameters
                      )
{
  library(MCMCpack)
  library(coda)

  Allmean_n <- array(NA,c(data.its,5,4))#table to save the mean of the parameters for eahc chains
  Allmean_p <- array(NA,c(data.its,5,4))
  AllAR <-matrix(NA,data.its,length(n.simul)) #contains all Acceptance Rate for each bootstrap
  colnames(Allmean_n)<-c("Mean","Sd","95Low","Median","95High")

  printseq<-seq(1,data.its,5)
     #simulate Y data
     if(data.its==1)
     {data.yest<-matrix(round(rowMeans(data.sim(n.simul,1000,CM)[[1]])),(length(n.simul)),1)}
     else {data.yest<-data.sim(n.simul,data.its,CM)[[1]]}

       for (z in 1:data.its){
       Result <- met.hasObsDir (nits,data.yest[,z],nburn,prior.param ,pprior.param, init.param1,init.pparam1)
       #Result2 <- met.hasObsDir (nits,data.yest[,z],nburn,prior.param2 ,pprior.param, init.param1,init.pparam1)
       #Result3 <- met.hasObsDir (nits,data.yest[,z],nburn,prior.param3 ,pprior.param, init.param1,init.pparam1)

     #Extract inference from MCMC
      ndraw1 <- mcmc(t(Result[[1]][(length(n.simul)+1),,]),thin=100)
```

```
      #ndraw2 <- mcmc(t(Result2[[1]][(length(n.simul)+1),,]),thin=100)
      #ndraw3 <- mcmc(t(Result3[[1]][(length(n.simul)+1),,]),thin=100)
       pdraw1 <- mcmc(t(Result[[2]][1,,]),thin=100)
       #pdraw2 <- mcmc(t(Result2[[2]][1,,]),thin=100)
       #pdraw3 <- mcmc(t(Result3[[2]][1,,]),thin=100)
       Allmean_n [z,1:2,]<- t(summary(ndraw1)$statistics[,1:2])
       Allmean_n [z,3:5,]<-t(summary(ndraw1)$quantiles[,c(1,3,5)])
       AllAR[z,]<-Result[[3]]
       Allmean_p [z,1:2,]<- t(summary(pdraw1)$statistics[,1:2])
       Allmean_p [z,3:5,]<-t(summary(pdraw1)$quantiles[,c(1,3,5)])

       if(length(which(printseq==z))==1){
         windows()
         plot(ndraw1,main="n Post (Chain1)")
         windows()
         par(mfrow=c(2,4))
         autocorr.plot(ndraw1,main="n AutoCorr Param1")
       }
         if(z==1 ||z==100 || z==200 || z==250) {
              print(date())
              print(z)}
                    }

return(list(PostMeansN = Allmean_n,PostMeansP = Allmean_p,AccepRate = AllAR,priorParam =
list(prior.param),pprior.param=pprior.param))
  }

#MCMC function
met.hasObsDir <- function(nits,# number of iteration in the MCMC
                    simul.y,#simulated true data
                    nburn,#burn in size
                    prior.param,#prior parameters for param y's
                    pprior.param,#prior parameters of the p's paremeters
                    param,#initila value of one chian for y's parems
                    pparam#initial values for the p's param
                    ) {

  nSp <-ncol(param)
  # Data from simulation function
  Data = simul.y
```

```
 samples <-array(0,c(dim(param),nits))
 psamples <-array(0,c(dim(pparam),nits))
 #Calculate likelihood or log(likelihood)
 likhood <- CalcObsDir(param,pparam)
 #measure the acceptance rate
 AcceRate <- matrix(,nits,nSp)
 #MCMC update
 for (t in 1:nits){
 #Update the parameters in the model using function "updateparam"
 output <- updateparamObsDir(nSp,param,pparam,Data,likhood,prior.param)
 AcceRate[t,] <- t(output$accep.rate)
 param<-t(output$param)
 likhood <- output$likhood[1]

     for (n in 1:nSp){
     pparam[,n] <- rdirichlet(1,c(param[1:nSp,n]+pprior.param[1:nSp,n]))
                 }
  samples [,,t] <- param
  psamples[,,t]<-pparam
                 }
 #calculate the mean and standart deviation of the parameters following burn-in:
    subsample<-samples[,,seq((nburn+1),nits,4)]
    psubsample<-psamples[,,seq((nburn+1),nits,4)]

  AcceptanceRate <- colMeans(AcceRate)

  return(list(subsample,psubsample,AcceptanceRate))
                 }
#Function to update parameters
updateparamObsDir <-function(nSp,# nbs of species
                     param,#inital values for the y's param
                     pparam,#initial values for the p's parem
                     Data,#true data
                     likhood,#likelhood estiamted with ald param
                     prior.param# prior parameters for the y's param
                   ){

    oldparam <- matrix(,2,nSp)
    accep.rate <-matrix(0,nSp,1)
```

```
        for (i in 1:nSp){
          oldparam[1,] <- param[i,]
          oldparam[2,] <- param[(nSp+1),]
          param[i,]<-(rmultinom(1,Data[i],(param[nSp+1,]*pparam[i,])/sum(param[nSp+1,]*pparam[i,])))
          param[(nSp+1),] <- colSums(param[1:nSp,])

      #Calculate the new likelihood value 0for the proposed moved:
           newlikelihood <-CalcObsDir(param,pparam)
          if(is.na(newlikelihood)==TRUE){print(param)
                                  print(pparam)
                        print("Log lik is null")}

      #Include the likelihood term in the acceptance probability
          num <- newlikelihood +
npriorObs(nSp,param[(nSp+1),],prior.param)+dmultinom(oldparam[1,],prob=((oldparam[2,]*pparam[i,])/sum(o
ldparam[2,]*pparam[i,])),log=TRUE)
          den <- likhood  +
npriorObs(nSp,oldparam[2,],prior.param)+dmultinom(param[i,],prob=((param[nSp+1,]*pparam[i,])/sum(param[
nSp+1,]*pparam[i,])),log=TRUE)
      #Acceptance probability:
          A<-min(1,exp(num-den))#if the difference is positive the min will be 1 so we will accept the move. If
the difference is negative, the min will be exp(num-den) so the move will be accepted in function of the uniform
distribution below.

      # Simulate a random number in [0,1] and accept move with probability A;
      # else reject move and return parameter value to previous value
           u <- runif(1)
        # print(newlikelihood)
           if (u <= A) { likhood <- newlikelihood
                    accep.rate[i,1] <-1
                     }
                else { param[i,] <- oldparam[1,]
                   param[(nSp+1),] <- oldparam[2,]
                   accep.rate[i,1] <- 0
                     }
                   }
output <- list(param=t(param),likhood=likhood,accep.rate=accep.rate)
                  }
```

```
 #Function to calculate model likelihood
CalcObsDir <- function(param,pparam){
PartialLik <-vector(length=ncol(param))
   for (sp in 1:ncol(param)){
   PartialLik[sp] <- dmultinom(param[(1:ncol(param)),sp],param[(ncol(param)+1),sp],pparam[,sp],log=TRUE)
                    }
 likhood <- sum(PartialLik)
                     }
```