# ENHANCING ACOUSTIC CLASSIFICATION USING META-DATA

**Lorène Jeantet & Emmanuel Dufourq**
African Institute for Mathematical Sciences
Stellenbosch University
National Institute for Theoretical and Computational Sciences
South Africa
{lorene,dufourq}@aims.ac.za

## ABSTRACT

Bioacoustics, the study of animal vocalizations and natural soundscapes, has proven to be a valuable source of data for wildlife monitoring. Just as a human would use contextual information to identify species calls from acoustic recordings, one unexplored way to improve deep learning classifier in bioacoustics is to provide the algorithm with contextual meta-data, such as time and location. We developed an algorithm to classify 22 bird songs for which the location can help to distinguish the different species. We explored different multi-branch convolutional neural networks, trained on both spectrograms and location information, as well as a geographical prior separately trained on location to estimate the probability that a species occurs at a given location. We compared the classification of the models to a baseline model without the spatial meta-data. Our findings revealed in each case an increase in the performance of the classification with the highest improvement obtained with the geographical prior (F1-score of 87.78%, compared to 61.02% for the baseline model). The methods based on multi-branch neural network proved to be efficient as well and simpler to use than the geographical prior as it requires a single model. Adding metadata to the acoustic classifier is a valuable source of information to improve classification performance, with room for further progress, and opens new opportunities for generalizing models.

## 1 INTRODUCTION

Wildlife monitoring has become even more important as biodiversity is declining at an unprecedented rate and effective protection measures are urgently needed (Almond et al., 2022). To this end, bioacoustics, the study of animal vocalizations and natural soundscapes, has proven to be a valuable source of data for understanding animal behaviors and biodiversity monitoring (Samuel et al., 2023). Therefore, it has become common practice to deploy multiple digital sound recorders in remote areas for an extended period of time to record and study populations of difficult to observe species (Gibb & Browning, 2019; Sugai et al., 2019). However, it generally results in long-term monitoring of high-resolution data that can be difficult and time consuming to verify manually. Processing automation has rapidly become a major issue in ecology in order to facilitate the analysis of these very large datasets. With its high ability to solve complex problems, deep learning has become the heart of the new challenges in ecology (Christin et al., 2019).

The use of deep learning in bioacoustic is still quite recent and not many studies relating its use can be found before 2017 (Stowell, 2022). The most common approach is to convert the raw acoustic recordings into a succession of spectrograms or mel-spectrograms – a visual representation of the frequency spectrum of sound over time – and treat the problem as an image classification. This representation generally gives a fixed image in time of the sounds and deprives them of their context. However, naturally, an expert in sound recognition would use a wide variety of contextual information when identifying species from acoustic recordings. Therefore, we hypothesis that one way to improve deep learning in species classification is to provide the algorithms with contextual information.

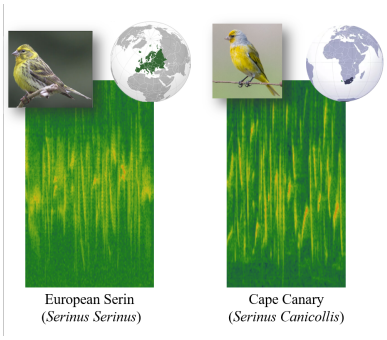


Figure 1: Example of spectogramms recorded for two species, European serin and Cape Canary, occuring in separate countries

Adding contextual information to deep learning classifiers in bioacoustics has not been widely explored, to the best our of knowledge only three studies have attempted this (Lostanlen et al., 2019; Madhusudhana et al., 2021; Roch et al., 2021). In these studies, the authors focused on the contextual information contained within the soundscape. However other meta-data, not directly contained in the soundscape but correlated, such as time/date and location, can provide relevant information. For example, two species of birds may have very similar songs but be found in different parts of the world (Figure 1). As it has become common practice to record the meta-data associated with each sound recording, it is then all the more interesting to valorize these data by incorporating them into deep learning classifiers to improve their performance. Therefore, the objective of this study is to enhance deep learning acoustic classifiers by adding spatial meta-data. To achieve this, we explored existing methods from image classification and applied them to a bird call classifier.

## 2 Materials & Methods

### 2.1 Dataset Generation

Xeno-canto is a well-known website created with the aim of sharing wildlife sound from all over the world (Xeno-Canto Foundation) and now stores a large dataset of bird songs recorded around the world (Vellinga & Planqué, 2015). In addition, each recording is associated with user-specified meta-data such as the country of recording, latitude and longitude and time of recording. To build our dataset from Xeno-canto, we firstly selected the ten most recorded families in the *Passeriformes* order, the most represented order in Xeno-canto database. From that, we sub-sampled the number of available species by retaining only the ten most recorded genus in each selected family. Then, based on the observation of the countries and the number of available recordings per species in one genus, we chose 22 species with similar songs but with different spatial distributions (Appendix A Table 2). The recordings were downloaded from the Xeno-canto database in .wav format and we manually labeled song occurrence within each audio file. We obtained 6537 occurrences of bird songs of various length from 967 audio files. We obtained on average 189 segments per species and per country with a large standard deviation (std =168 , min =24, max=1044). To balance our dataset, we set the number of segments per species and per country to 200 and processed data-augmentation or reduction as appropriate.

### 2.2 Preprocessing of the data

Each recording was downsampled to 22050 Hz and converted into a mel-spectrogram to be used as an input image to a 2-D Convolutional Neural Network (CNN). Mel-spectogram has the advantage of being smaller than linear frequency spectograms and gave better results in our case after comparison of the two formats. We performed a Hann analysis with a window size of 46 ms (1024 samples) with a hop size of 10.6 ms (256 samples) and 128 mel frequency. Similarly to Kahl et al. (2021), we restricted the frequency range of the spectrogram between 150 Hz and 15 kHz as most bird vocalizations occur between 250 Hz and 8.3 kHz (Hu & Cardoso, 2009). Using our manual labels, the songs were extracted from the spectrograms and divided into segments of ca. 3s using a sliding

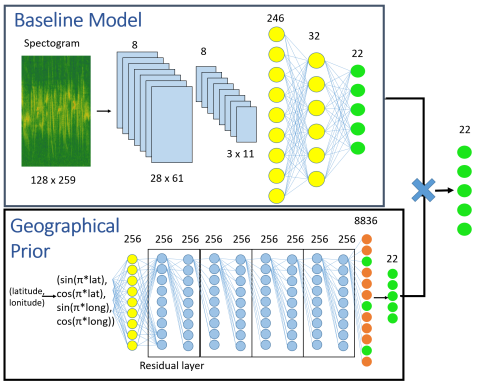


Figure 2: Architecture of the geographical prior trained separately to the baseline model. From the 8836 probabilities obtained (represented in orange), only the 22 species involved (represented in green) are retained and multiplied with the corresponding outputs of the baseline model.

window approach with an overlap of 1s. The resulting spectrograms each had a size of $128 \times 259$. Each spectrogram was paired with it's latitude, longitude and country of origin which was obtained via the meta-data associated with recording on Xeno-canto.

### 2.3 Model architecture

We explored four model architectures along with different ways of representing the meta-data to determine which approach would yield the best performance We implemented a baseline model without providing spatial meta-data (Case I). We tested two multi-branch neural networks with different meta-data pre-processing (Cases II and III). Finally we separately trained a geographical prior and combined its output with the baseline predictions (Case IV). The four classifiers were trained and tested as specified in Appendix B.

#### 2.3.1 Case I: Baseline model

As baseline, we used a simple CNN architecture consisting of two convolutional layers (8 kernels of size $16 \times 16$, and ReLU activations) followed by a max pooling layer ($4 \times 4$ kernel), a flattening layer, and 2 fully-connected layers (32 ReLU and 22 softmax units respectively). The softmax activation function was used to obtain the probability that the input signal corresponds to each species.

#### 2.3.2 Case II: One-hot encoding

In Case II, we tested a two-branch convolutional CNN (spectrogram input in one branch and meta-data input in the other). To achieve this, we assigned a unique number ($n = 28$) to each country used in this study and converted the number into one-hot encoded vector. The first branch used the spectrogram as input, and this branch had the same CNN architecture as Case I. The second branch used the one-hot encoded vector as input and was concatenated to the flattening layer in branch one.

#### 2.3.3 Case III: Meta-data embedding

Word embeddings are well known within the field of natural language processing. It consists of mapping words to continuous number vectors, and facilitates a dimensionality reduction of the categorical variables while keeping meaningful information in the transformed space. We used a multi-branch CNN, similar to Case II, but the input to the second branch was the names of the country of origin where the recordings took place. We added an embedding layer to map the country names into an 8-dimension transformed space. The output of the embedding was then concatenated with the flattened output of the first branch.

#### 2.3.4 CASE IV: GEOGRAPHICAL PRIOR

A geographical prior was developed by Aodha et al. (2019) to enhance image classifier with contextual information. Inspired from species distribution modeling, the model estimates the probability that the object category represented in the image occurs at a given location. It is trained and applied independently to the classifier using only the location and time data as input. Therefore, the probability that an object is represented in the image knowing its location can be estimated by multiplying the probabilities of the image classifier with those of the geographical prior (Aodha et al., 2019). In Case IV, we replicated this method and applied it to our bird song classifier. We used the entire dataset of Xeno-canto to train the geographical prior which contains calls from 8836 species. We took latitude and longitude as input, which we treated similarly to Aodha et al. (2019). For each latitude or longitude value $x$ we calculated $[\sin(\pi x), \cos(\pi x)]$, resulting in vector of dimension four for each input and preserving the continuity of the geographical coordinates all around the earth (Aodha et al., 2019). As output, we got the probability of presence (close to 1) or absence for each category. The architecture of the prior was exactly the same than proposed by Aodha et al. (2019), with a first dense layer of 256 hidden units followed by a ReLU, four residual layers (He et al., 2016), and a final dense layer of 8836 output units followed by a softmax activation function (Figure 2).

Table 1: Comparison between the different techniques. Incorporating geographical prior information directly into the CNN improves model performance.

| | **Baseline** | **One-hot encoding** | **Meta-data embedding** | **Geographical prior** |
|---|---|---|---|---|
| Accuracy | 97.62% | 98.52% | 98.43% | 99.19% |
| Sensitivity | 61.34% | 84.31% | 81.81% | 86.96% |
| Specificity | 98.72% | 99.23% | 99.17% | 99.57% |
| Precision | 70.21% | 80.20% | 78.69% | 91.06% |
| F1-score | 61.02% | 78.77% | 76.87% | 87.78% |

## 3 RESULTS AND DISCUSSION

The addition of contextual metadata to deep learning classifiers in bioacoustics has been poorly investigated although it can bring relevant improvements. We showed in this study that methods developed for image or natural language processing can be adapted to bioacoustics and used for this purpose. Therefore, by training a geographical prior on longitude and latitude data to predict the probability of species occurrence at a given location, we improved the F1-score of bird song classification of 22 species from 61.02% from the baseline to 87.78%.

The classification of bird songs is challenging, due to the fact that certain bird species possess rich and diverse song repertoires featuring highly complex sound sequences (Samotskaya et al., 2016). For the basic model, the main errors mostly concerned a multitude of species wrongly labelled as *Saxicola rubicola* and/or *Troglodytes aedon* (Figure 3). An analysis and visualization of the spectrograms did not provide an understanding of the origin of these errors, further studies are warranted to fully understand the nature of these errors. However, with the geographical prior, misclassifications between species from distinct regions were no longer observed (Figure 4), suggesting that location information was the missing piece of information contributing to the improvement.

The geographical prior has the advantage of being distinct from the classifier and can therefore be run independently. However, a representative spatial meta-data dataset must be available to train it which may be difficult to obtain for some species. We tested other methods based on multi-branch neural network that proved to be efficient as well and simpler to use than the geographical prior as it requires a single model.

We showed in this study that adding spatial meta-data can enhance a classification model. In addition to improving species identification and thus monitoring, they can also contribute to the generalization of species classifier. With the addition of meta-data, a single robust model can be generated instead of creating various species and geographical specific models. Furthermore, the present methods can also be used for species sound detection and become a valuable source of information to improve the precision of the detection.

## REFERENCES


Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (Software available from tensorflow.org), 2015. URL `http://arxiv.org/abs/1603.04467`.

REA Almond, M Grooten, D Juffe Bignoli, and T Petersen. Living planet report 2022–building a nature-positive society. *World Wildlife Fund*, 2022.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9595–9605, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00969.

François Chollet. Keras, 2015. URL `https://keras.io`.

Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019. ISSN 2041210X. doi: 10.1111/2041-210X.13256.

Rory Gibb and Ella Browning. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. 2019(September 2018):169–185, 2019. doi: 10.1111/2041-210X.13101.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90.

Yang Hu and Goncalo C. Cardoso. Are bird species that vocalize at higher frequencies preadapted to inhabit noisy urban areas? *Behavioral Ecology*, 20(6):1268–1273, 2009. ISSN 10452249. doi: 10.1093/beheco/arp131.

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61(December 2020):101236, 2021. ISSN 15749541. doi: 10.1016/j.ecoinf.2021.101236. URL `https://doi.org/10.1016/j.ecoinf.2021.101236`.

Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Pablo Bello. Robust sound event detection in bioacoustic sensor networks. *PLoS ONE*, 14(10):e0214168, 2019. doi: 10.1371/journal.pone.0214168.

Shyam Madhusudhana, Yu Shiu, Holger Klinck, Erica Fleishman, Xiaobai Liu, Eva Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, Ana Sirovic, and Marie A. Roch. Improve automatic detection of animal call sequences with temporal context. *Journal of the Royal Society Interface*, 18(180), 2021. ISSN 17425662. doi: 10.1098/rsif.2021.0297.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

Marie A. Roch, Scott Lindeneau, Gurisht Singh Aurora, Kaitlin E. Frasier, John A. Hildebrand, Herve Glotin, and Simone Baumann-Pickering. Using context to train time-domain echolocation click detectors. *The Journal of the Acoustical Society of America*, 149(5):3301–3310, 2021. ISSN 0001-4966. doi: 10.1121/10.0004992.

V V Samotskaya, A S Opaev, V V Ivanitskii, I M Marova, P V Kvartalnov, A S Opaev, V V Ivanitskii, I M Marova, and P V Kvartalnov. Syntax of complex bird song in the large-billed reed warbler ( Acrocephalus orinus ). *Bioacoustics*, 4622(January), 2016. doi: 10.1080/09524622.2015.1130648.

Ross Samuel, Robert Peter, James Orcid, Amandine Gasc, Jennifer N Phillips, Sarab S Sethi, Connor M Wood, and Zuzana Burivalova. Passive Acoustic Monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, 2023. doi: 10.1111/1365-2435.14275.

Dan Stowell. Computational bioacoustics with deep learning : a review and roadmap. *PeerJ*, (10:e13152), 2022. doi: 10.7717/peerj.13152.

Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, José Wagner Ribeiro, and Diego Llusia. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1):5–11, 2019. ISSN 15253244. doi: 10.1093/biosci/biy147.

Willem-Pier Vellinga and Robert Planqué. The Xeno-canto collection and its relation to sound recognition and classification. 2015. URL `www.xeno-canto.org`.

Xeno-Canto Foundation. Xeno-Canto : Sharing wildlife sounds from around the world. URL `https://xeno-canto.org/`.

## A Appendix A

Table 2 outlines the different genus, species and the total number of records which we used to create our dataset. The country for which the audio files were recorded in is also listed.

## B Appendix B Training and validation

To train and evaluate the efficiency of each method, the bird song dataset was split into a training and validation dataset. For each species and country, we randomly selected $70\%$ of the downloaded recordings for the training dataset and kept the remaining $30\%$ for the validation one. The data augmentation process was then applied only on the training dataset. In each case, the model was trained on 40 epochs with a batch size of eight segments and a learning rate of 0.001 using the Adam optimizer. The training process was performed ten times and evaluation metrics were calculated each time from predictions of the validation dataset. Therefor, each time the confusion matrix was build and the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) obtained. From that we calculated accuracy (TP+TN/TP+TN+FP+FN), sensitivity or recall (TP/TP+FN), specificity (TN/TN+FP), precision (TP/TP+FP) and F1-score (harmonic mean between precision and recall). For each metric, we calculated the average value of the ten trainings to compare the four methods. For Case IV, the geographical prior was trained once and the resulting probabilities were multiplied to the probabilities obtained from the baseline model. We therefore repeated this process ten times as well. Models were implemented in Python 3 using the TensorFlow and Keras libraries (Abadi et al., 2015; Chollet, 2015). Audio processing and spectrogram construction were performed using Librosa library (McFee et al., 2015).

## C Appendix C Confusion Matrix

Figures 3 and 4 present the confusion matrices for the baseline and geographical prior model respectively. The latter produced overall better classification results.

Table 2: Details regarding the dataset used in this study. The audio recordings were obtained from Xeno-canto.

| **Genus** | **Species** | **Total number of recordings** | **Average number of recordings per country** | **Country with records** |
|---|---|---|---|---|
| Muscicapidae | Saxicola gutturalis | 8 | 8 | Indonesia |
| | Saxicola rubetra | 87 | 11 | France, Germany, Ireland, Norway, Poland, Russian, Sweden, UK |
| | Saxicola rubicola | 70 | 10 | Belgium, France, Germany, Netherland, Poland, Portugal, Spain |
| | Saxicola tectes | 12 | 12 | France |
| | Saxicola tectes | 12 | 12 | France Saxicola torquatus88South Africa |
| Thamnophilidae | Hypocnemiscantator | 32 | 11 | Brazil, Suriname, Venezuela |
| | Hypocnemis hypoxantha | 29 | 10 | Brazil, Ecuador, Peru |
| | Hypocnemis peruviana | 33 | 11 | Brazil, Ecuador, Peru |
| | Hypocnemis striata | 11 | 11 | Brazil |
| Fringillidae | Serinus canicollis | 14 | 14 | South Africa |
| | Serinus serinus | 83 | 14 | France, Germany, Italy, Poland, Portugal, Spain, Netherland, Ireland |
| Turdidae | Catharus aurantiirostris | 68 | 14 | Colombia, Costa Rica, Hondura, Mexico, Panama |
| | Catharus bicknelli | 6 | 6 | USA |
| | Catharus fuscater | 29 | 7 | Colombia, Costa Rica, Ecuador, Panama |
| | Catharus fuscescens | 24 | 12 | Canada, USA |
| | Catharus guttatus | 37 | 12 | Canada, USA, Mexicos |
| | Catharus minimus | 16 | 16 | USA |
| | Catharus ustulatus | 33 | 16 | Canada, USA |
| Troglodytidae | Troglodytes aedon | 120 | 20 | Brazil, Chile, Colombia, Ecuador, Mexico, USA |
| | Troglodytes hiemalis | 39 | 19 | Canada, USA |
| | Troglodytes pacificus | 35 | 17 | Canada, USA |
| | Troglodytes troglodytes | 173 | 19 | Belgium, France, Germany, Ireland, Netherlands, Poland, Portugal, Spain, UK |

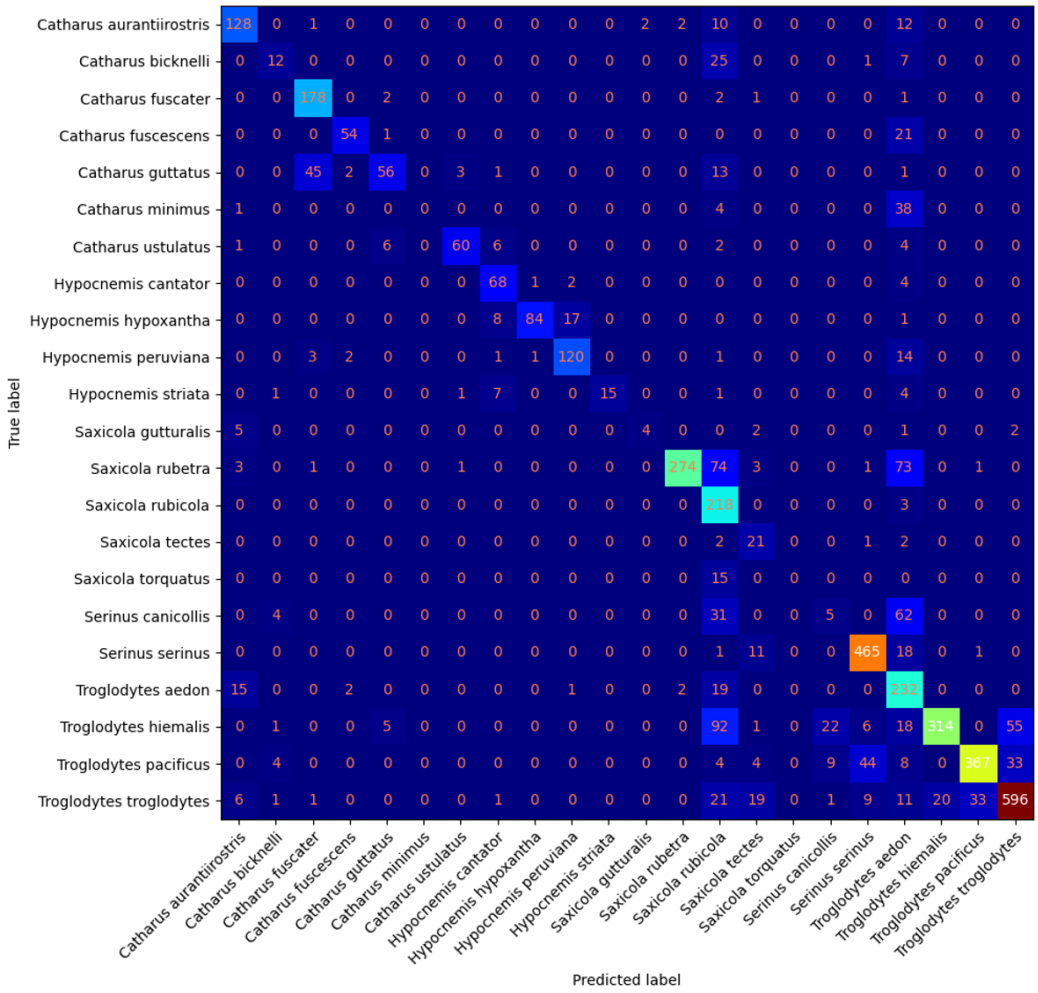


Figure 3: Confusion matrix associated to the predictions of the baseline model (Case I)

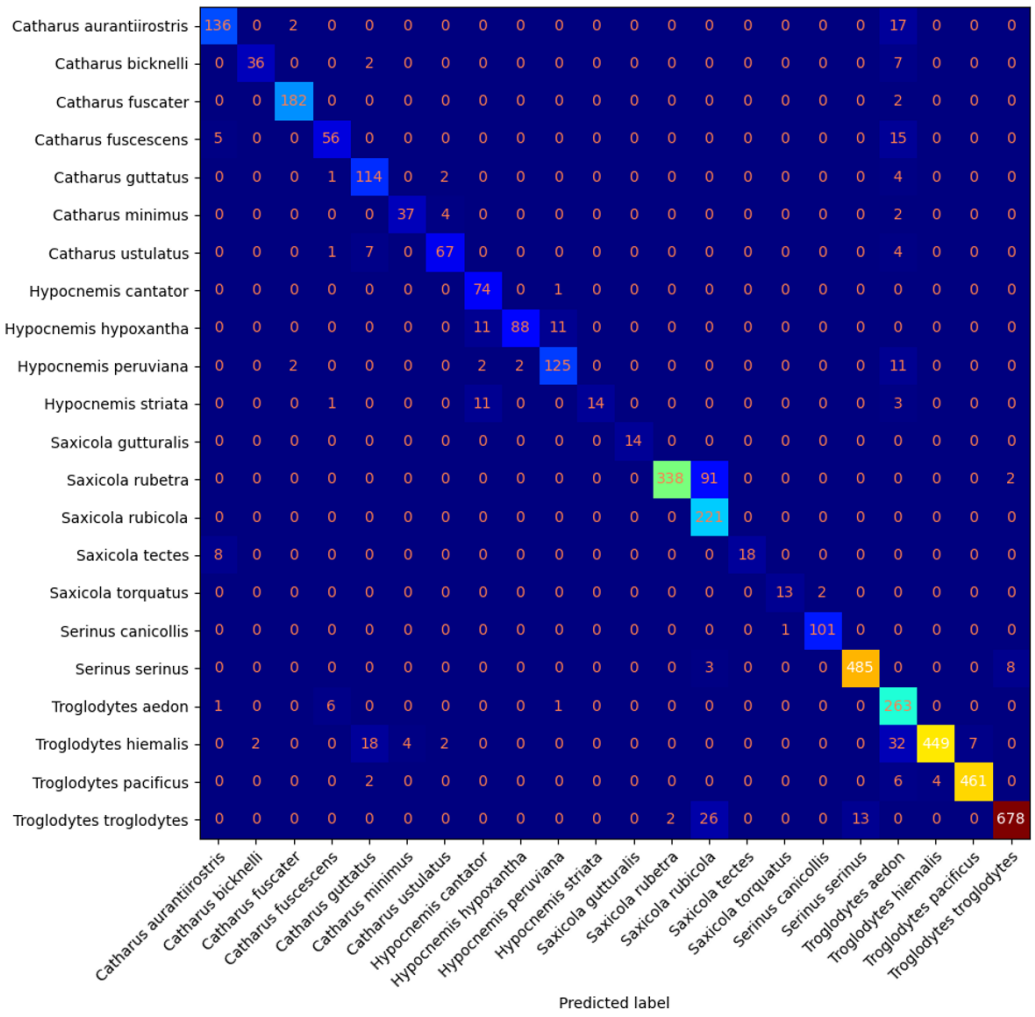


Figure 4: Confusion matrix associated to the predictions of the geographical prior (Case IV)