POLITECNICO
MILANO 1863

Politecnico di Milano

Dipartimento di Elettronica, Informazione e Bioingegneria

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING

A Species Independent Framework for Estimating Animal Wildlife Population Using Social Media Images Collections

Master of Science thesis of:
Matteo Foglio
Matricola 876457

Advisor:
Prof. Pier Luca Lanzi

Co-advisor:
Dott.ssa Tanya Berger-Wolf

Academic Year 2018/2019

Acknowledgments

Innanzitutto vorrei ringraziare la mia famiglia, a cui devo ogni successo e risultato. Grazie per le incredibili opportunità che mi avete dato. Grazie per supportarmi anche nella scelta di vivere lontano da voi.

Vorrei ringrazie anche il mio relatore, professor Pier Luca Lanzi, e la mia correlatrice, professoressa Tanya-Berger Wolf.

La lista di amici da ringraziare è infinita e di questo vi sono grato. Ringrazio in particolare quelli che mi sono stati più vicini: Guido, Riccardo, Paola e Jordie.

Grazie Franci.

MF

Abstract

We are losing biodiversity at an unprecedented scale and in many cases, we do not even know the basic data for the species. Traditional methods for wildlife monitoring are inadequate: they are expensive, time-consuming, and therefore unscalable. Development of new computer vision tools enables the use of images as the source of information about wildlife. Social media is the rich source of wildlife images, which come with a huge bias, thus thwarting traditional population size estimate approaches. Here, we present a new framework to take into account the social media bias when using this data source to provide wildlife population size estimates. We test the method on two different species: Grevy's zebra and Reticulated giraffe.

Our approach is composed of two steps. First, a regression model is trained to estimate the total number of animals, shared and not shared, photographed by a user, given the images collection the user has shared on social media. We show that this is a learnable and potentially solvable problem. Moreover, we explain how to create a dataset suitable for the training phase. In a second step, the trained regression model is applied on data scraped from social media. The regression output is than fed to a modified traditional wildlife estimator in order to provide an estimate for the entire population.

Finally, we show how to automatize a crucial part of the cleaning process: the identification and removal of captive animals from the dataset collected from social media.

Sommario

Stiamo perdendo biodiversità a un ritmo precedentemente mai visto. Per molte delle specie conosciute non disponiamo nemmeno delle più basiche informazioni. I metodi tradizionali usati per il monitoraggio della fauna sono inadeguati. Sono infatti costosi e richiedono un grande impiego di risorse e tempo: non sono dunque scalabili. Lo sviluppo di nuovi tool nell'ambito della computer vision permette l'utilizzo di immagini come fonte di informazioni sulla fauna. Una delle più ricche fonti di immagini di animali sono sicuramente i social media. Tuttavia, queste immagini portano con sé un bias che ne impedisce l'utilizzo nelle tradizionali tecniche per stimare le dimensioni di una popolazione animale. In questo lavoro presentiamo un nuovo framework capace di tener conto di tale bias quando si vogliano utilizzare i social media come fonte di dati per il monitoraggio della fauna. Il metodo è testato su due differenti specie: la Zebra di Grevy e la Giraffa Reticolata.

Il nostro approccio consta di due parti. Dapprima, viene sviluppato un modello di regressione per stimare il numero totale di animali fotografati da un utente, sia condivisi che non condivisi sui social media, sulla base delle foto da lui condivise. Mostreremo che questo problema è potenzialmente solvibile e può essere modellato dagli attuali modelli di machine learning. Inoltre, mostreremo come sia possibile creare un dataset adatto allo sviluppo di tale modelli. In una seconda fase, un modello di regressione generato sul precedente dataset viene applicato a dati scaricati dai social media. L'output del regressore viene poi usato in un tradizionale modello biologico al fine di fornire una stima per la dimensione dell'intera popolazione animale relativa ad una specie.

Infine, mostriamo come sia possibile automatizzare un processo fondamentale nella costruzione di un dataset per lo studio di una specie: l'identificazione ed eliminazione di foto provenienti da social media che mostrino animali detenuti in zoo.

Contents

List of Tables

List of Figures

Introduction 1

1.1 The problem

What is the number of endangered species? How many animals do they count? How long do we have before their extinctions? According to the IUCN Red List, assessing the number of threatened species is complicated because for some of them we have so little information that they have been classified as *Data Deficient* or *Not Evaluated* [17]. Moreover, even for some of the species that are considered assessed, we have very inaccurate data. For instance, the last global-scale genetic study on Whale Sharks (*Rhincodon typus*) have estimated the size of genetic effective population - the number of breeding adults - to be 103,572 with a standard error of 27,401–179,794. Furthermore, species assessment is rarely conducted with a yearly frequency. For instance, the Iberian Lynx (*Lynx pardinus*), an endangered species counting approximately 150 mature individuals, has been assessed for the last time in 2014 [16]. As at least other 28,000 species, Whale Sharks and Iberian Lynxes are threatened with extinction. Despite the efforts of different organizations, as of now we don't have resources nor tools to assess the status of these species as frequently and as accurately as needed. Finally, thousands of species have not been evaluated at all, as shown in Table 1.1

Conservation is an urgent problem: humanity has wiped out 60% of wildlife population just in the last 40 years. This is a threat not only to the wildlife population itself, but also to the human species which is part of this fragile and ill ecosystem. All of our economic activities

Table 1.1: SPECIES ASSESSMENT DATA FROM IUCN. FROM LEFT TO RIGHT, FOR EACH CATEGORY: NUMBER OF SPECIES CURRENTLY ASSESSED, NUMBER OF SPECIES THAT IUCN WANT TO ASSESS BY 2020, AND TOTAL NUMBER OF SPECIES DESCRIBED (INCLUDING NON ASSESSED ONES) BY IUCN.

Type	**Assessed**	**2020 Goal**	**Species**
Vertebrates	49688 - 71.1%	61635 - 88.3%	69788
Invertebrates	22311 - 1.6%	45344 - 3.3%	1359365
Animals, Plants, Fungi	105,700 - 5.5%	160,000 - 8.4%	1904587

ultimately depends on nature. [2] It is necessary to face this challenge seriously with all the means at our disposal.

1.2 Wildlife population estimation

Wildlife monitoring is not a solution per se, but it poses the basis for a deeper analysis of our Planet's health. If we could estimate the size of every species more frequently, we would be able to better understand what are the consequences of our actions on the environment, whether we wanted to study the effects of industrialization on a specific region or the reintroduction of a species into the territory from where it has disappeared. In fact, wildlife population monitoring can be used as an important index in answering several questions, such as defining the ecological health of an area, or in preventing potential pest species in causing agricultural, economic or natural resource damages that could lead to livestock disease or safety hazard [56]. Furthermore, having nearly real-time population estimates could be a way to get the attention of mass media and ultimately of public opinion, which can force governments to take actions in order to safeguard the environment and its biodiversity.

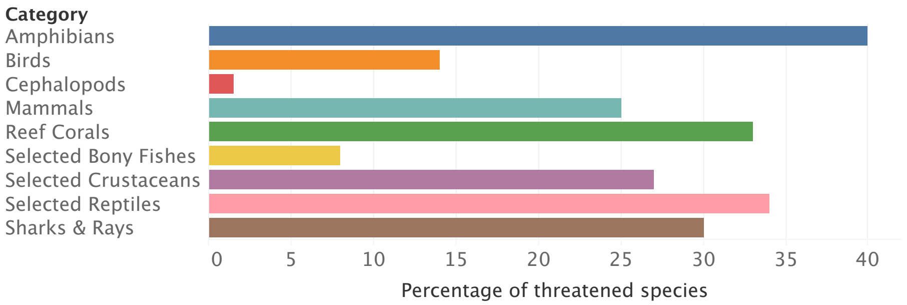


Figure 1.1: Percentage of threatened species for each category according the UICN Red List. Selected crustaceans include lobsters, freshwater crabs, freshwater crayfishes and freshwater shrimps. Selected reptiles include marine turtles, seasnakes, chameleons, crocodiles and alligators. Selected bony fishes include anchovies, angelfishes, billfishes, blennies, bonefishes, butterflyfishes, cornetfishes, groupers, parrotfishes, pufferfishes, sardines, sturgeons, surgeonfishes, tarpons, tunas, picarels, porgies, seahorses, seabreams, syngnathid fishes. Chephalopods include nautiluses, octopuses, squids. [17]

1.3 Limitations of traditional techniques

Assessing a great number of species is a complex and expensive task. Traditional methods used by biologist often rely on the capture-mark-recapture principle (see Section 2.1 and Section 5.6.2). These techniques require to capture and identify sets of individuals at different time intervals. Biologists need to scout natural areas in search of animals. Depending on the species, it might be sufficient to take pictures of the animals, or it might be required to capture individuals and tag them. In particular, the latter process can be dangerous for the biologists and animals themselves. As an example, in 2016 an orca died because of an infection caused by the application of a satellite tag [55]. Photos can be taken from inside the species habitat or from a plane, but in many case they must be reviewed manually by experts. As the number of photos growth, the task soon becomes unfeasible without the use of any dedicated computer vision tool. In brief, traditional methods for wildlife estimation require the mobilization of resources and biologists, making the techniques extremely costly and time-consuming, and therefore unscalable. Further considerations are presented in Section 2.1.

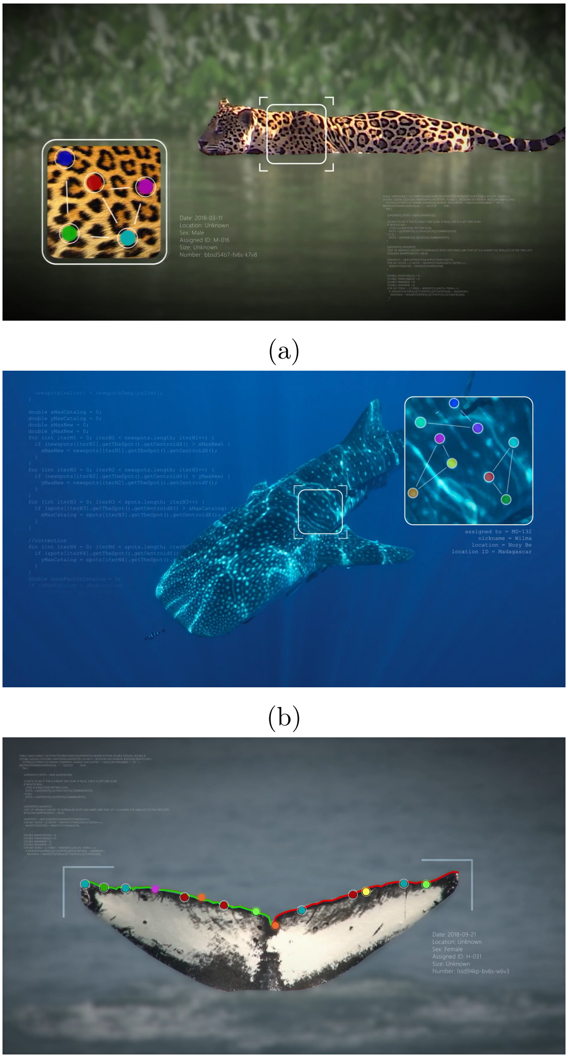
1.4 Social media: opportunities and challenges

1.4.1 A huge, free and updated source of data

Social media give us tremendous opportunities. As a third of the entire world population shares their lives on the Internet [29], an interesting number of wild animals has started to appear on social media. In 2013 more then 350 million photos were uploaded on Facebook every day [40] [1]. As this number keeps growing, social media will become an even bigger data source. It might be the case that the amount of data available on the Web may lead to more precise estimates than the ones computed with traditional data collection process. Moreover, the posting frequency may be sufficiently high to allow the detection of changes in wildlife population size with an almost real-time capability. Finally, this data source is available for free, overcoming the economical limitation posed by traditional data collection processes. In the last years, new computer vision tools, such as Wildbook [5], have enabled the use of large scale image datasets by providing an automatic solution for animal detection and identification in pictures. In fact, this software is not only capable of detecting several species in a photo, but it can also recapture the same individual across different pictures, allowing the application of capture-mark-recapture techniques on social media images.

1.4.2 The social media bias

Previous studies have warned against a naive usage of social media data [34]. In particular, it has been shown that the social media bias affects



Date: 2018-03-11
Location: Unknown
Sex: Male
Assigned ID: M-016
Size: Unknown
Number: bbsd54b7-fv6s-k7v8

(a)

assigned to = MD-132
nickname = Wilma
location = Nosy Be
location ID = Madagascar

(b)

Date: 2018-09-21
Location: Unknown
Sex: Female
Assigned ID: H-031
Size: Unknown
Number: lssd94kp-bv6s-w6v3

(c)

Figure 1.2: Different species can exhibit individual patterns that allow to recognize a single animal of the species. For instance jaguar (1.2a) and whale sharks (1.2b) can be identified using their spots while whales (1.2c) can be identified by the shape of their flukes. Images courtesy of Microsoft, AI for Earth initiative.

wildlife population assessment when using social media as a data source [30]. Understanding how and why this bias affects population estimates is the key to an useful utilization of social media data.

When estimating the number of animal of a given species, biologists are in charge of the data collection. This process is meticulously planned according theories and practices developed through a number of academic studies. Researchers know what information is needed, as well as when, where and how this information should be retrieved [56]. When it comes to the use of social media data, we cannot rely on the same effort from the typical social media user. People share images according to their tastes and personal experiences. Some people might share all of the images, others might decide to share just a few. It has already been shown that the problem of predicting the shareability of animal images - the likelihood for an image to be posted on social media - is a learnable problem [44] [30]. However, previous studies have been conducted to the level of single image and not images collections. In other words, we need to understand whether the shareability of an image is affected by other pictures taken by the same photographer. Moreover, while this study showed that we can predict which images will be shared, we lack of a research conducted on the possibility of estimating information about the images that have not been shared, given the ones that instead have been posted on a social media. In this research we will use regression algorithms to predict the number of animals photographed by an user, but not shared on social media, given the ones that the user has shared. The ultimate goal of this process is indeed to the create a framework which can be robust towards the social media bias.

1.4.3 Captive and wild animals

Finally, when using data from the Internet, the majority of animal pictures will show captive individuals. This is explicable by the higher chance for an animal in a zoo to be photographed. Including these photos in our dataset would lead to an extremely wrong assessment of a species population. As mentioned in the first part of this section, many techniques for species assessment are based on the number of captured and recaptured animal in different time instant. The use of data coming from zoo will greatly increase the number of recaptured animals. Therefore, we need a tool to classify zoo and non-zoo pictures in order to enable the use of big data coming from the Web. In this work we will present a pipeline to distinguish these categories of images.

1.5 Content of the work

The ultimate goal of this work is to provide a species independent framework to assess wildlife populations using social media data. This approach is conceived to solve the numerous issues related to the traditional techniques. These methods are expensive, time-consuming, and sometimes dangerous both for the animals and biologists [55] [22]. On the contrary, our approach will make use of free abundant data, constantly updated. This means that it will be able to provide estimate for animal species more frequently and at a much lower cost. Finally, the abundance of data may lead to more accurate results.

The work can be divided into three main parts.

The first is the development of a tool capable of handling the social media bias. In particular, we will design a pipeline to estimate the number of individual animals photographed but not shared by a user. This pipeline will make use of machine learning models which will be trained on an opportunely collected dataset. We will also explained how a suitable dataset can be created.

In the second part of the work, we will develop a framework to integrate the first part of the pipeline with traditional methods used by biologist to estimate the size of an entire animal species. We will scrape social media to collect animal images collections. Then, we will apply on them the machine learning models trained in the previous part. Finally, we will feed the output of the machine learning models to the adapted biological methods to provide the estimate for two endangered species: the Grevy's zebra (*Equus grevyi*) and the Reticulated giraffes (*Giraffa camelopardalis reticulata*). Moreover, by exploiting citizen science and crowdsourcing, we hope to engage people in wildlife preservation.

Finally, we face a third problem, which is the identification of images collections showing captive animals. They must be removed in order to provide correct estimate for any species. We will show how to train a classifier as a solution to avoid this time-consuming task.

The following chapters are divided as follows:

- Chapter 2: we discuss traditional techniques used to estimate animal wildlife population, challenges when conducting studies using social media as a data source, and previous works on estimating the number of wildlife animals of a given population using social media images.

- Chapter 3: we state the exact problems we aim to solve in the scope of this thesis.

- Chapter4: we describe the datasets used to train our machine learning models and the process of scraping a social media platform to estimate wildlife populations.

- Chapter 5: we discuss the theory behind the proposed solution.

- Chapter 6: we discuss the actual implementation of the solution proposed in Chapter 5.

- Chapter 7: we present the performance of the classification and regression models, and the results we obtained when estimating the number of Grevy's zebra and Reticulated giraffe using Flickr TM as data source.

- Chapter 8: we discuss the results and we propose future researches to improve the proposed framework

Related Work 2

In this chapter, we describe traditional methods used to estimate wildlife population. We then discuss the main issues related to the use of social data when making inference about real world data. Finally, we discuss the cutting-edge research in estimating wildlife population using social media images.

2.1 Traditional wildlife population estimation techniques

Several possible approaches could be taken when estimating the number of animals of wildlife population. Besides species-focused techniques, all general methods mainly rely on the probability of detection and the probability of sampling single individuals. Biologists have developed a number of estimators for these two quantities. These estimators exploit either direct or indirect observations of animals. In particular, population estimation methods can be divided into three broad categories depending on the percentage of individuals that can be observed: all, none or only a portion.[22]

When all individuals can be observed, it is possible to count their number directly. A particular case of this method occurs when the count is conducted only on a sub-sample of the whole environment inhabited by the species. Then, an estimate $\hat{N}$ for the population can be found assuming the density of the population over the area to be constant:

$$\hat{N} = \frac{\hat{N}'}{\alpha} \tag{2.1}$$

where α is the proportion of the total area to which $\hat{N}'$ pertains.

When none of the individuals can be observed, indirect indexes can be used. These methods do not rely on seeing directly the animals, but merely on observing signs and traces of their activities in the area. Indirect indexes are suitable for comparing activities in different areas, as they are also called "relative" or "activity" indexes. They are instead less appropriate when estimating the absolute number of animals, since animals traces are someway related to animals presence, but their numerical correlation is not always quantifiable.[56]

Finally, a third situation occurs when not all, but some of the individuals can be observed. This case is definitely the most common, as it is the most suitable for counting the number of animals in wildlife population. Two further sub-categories can be identified: removal methods and capture-mark-recapture methods.

Removal methods are old and can be used when animals are trapped and removed from the population [57]. Their attractiveness resides in the ease for the investigator to collect data, since other people, such as hunters, can provide it. These methods are certainly not suitable for endangered species.

Capture-mark-recapture (CMR) methods are based on the probability of recapturing animals in different time instants. [22] At time t_1 biologists capture a number n of animals. These individuals are marked so that biologists can recognize them in a successive time instance t_2. After being marked, the animals are then released and reintroduced in their habitat. Then, a certain amount of time is waited such that marked animals can mix with unmarked ones. The time waited should last enough to make legit to assume that the proportion of marked animals is constant over the entire population. Then, at time t_2 a number K of animals is captured. Among these K individuals, k of them will show the mark applied by biologists at time instance t_1. In other words, assuming that no animal has lost the mark, k is the number of animals captured both at time t_1 and time t_2. Under these assumptions, the number of animals of the species is:

$$\hat{N} = K * \frac{n}{k} \tag{2.2}$$

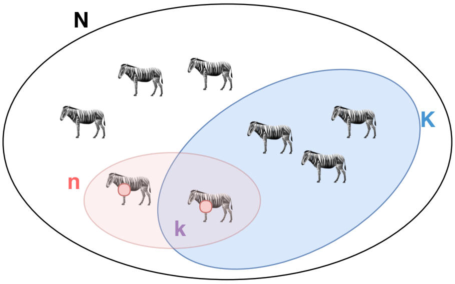


Figure 2.1: Schema of the mark-recapture method. N is the population for which we want to provide an estimate. n is the set of animals captured at t_1. K is the set of animals captured at time t_2. k is the set of animals captured at t_1 and recaptured at t_2.

The basic capture-mark-recapture idea has been extended by a variety of models. In our work we make use of the Jolly-Seber method [19] [45]. In Section 5.6 we will motivate our choice, describe the method and explain how we have adapted it to our use case.

Traditionally the application of marks to animals required the physical capture of wildlife. This procedure could be dangerous for the animal itself to the point that some statistical wildlife estimators allow taking into account the death of animals during their capture [21]. Nowadays, recent developments in computer vision techniques limit the necessity to use physical marks. Some species, in fact, are provided with some specific traits that are distinctive of each individual. In the same way, we can be identified by our fingerprints, zebras can be recognized by their stripes, giraffes by the pattern of their spots, and whale sharks by the shape of their fluke. As a consequence, we don't need to apply marks to individuals of these species. Instead, pictures themselves provide sufficient information for the purpose of applying capture-mark-recapture methods.

2.2 Challenges when using social media as a data source

Social media is an attractive source of data for many types of research. Not only it already provides a huge amount of information, but its growth rate is constantly increasing. From 2005 to 2015, the percentage of American adults that uses social media has grown from 7% to 65%, with a peak of 90% for adults aged between 18 and 29 [38]. In 2015 about 3 billion people were provided with Internet access, including 2 billion people from developing countries. At the same time, the percentage of users from developing countries has passed from being just a quarter of the users to being the majority of them [52]. The growing trend has not decreased, and as of June 2018, more than 4 billion people are provided with Internet access [48].

Companies, as well as academics, are trying to take advantage of this data. However, many warnings have been given on the naive usage of social media data. Well-known pitfalls, such as the Google Flu Trends [7], have proved that the quantity of data is not a valid substitute for the quality of data. In particular, many problems may arise when using social media data to understand phenomena that go beyond social media. This is because social media may presents biases that could lead researchers to wrong conclusions.

Biases could arise at many levels. They could be present in the data themselves, they could be generated during the data collection process, or they can be introduced by researchers during the preprocessing and the analysis phases. Furthermore, cautions must be taken when generalizing the results[34]. For instance, online users may not be a representative subset of the population we would like to generalize our findings to. Moreover, the same population may exhibit different behaviors on different platforms, according to the implicit social and behavioral norms of the service they are using[43]. Therefore, the validation process must take into accounts all these aspects and it should eventually limit the extents of the research conclusions.

Other challenges may come from fake or altered data. The recent spread of fake news shows that even for humans it can be difficult to discern between real and fake information. Moreover, by liking and sharing, social bots can augment the diffusion of this kind of data [23], whose algorithmic identification is still an animate research topic [36]. Likewise, when dealing with pictures, people fail to identify fake images on the web [20], even if algorithms appear to be provide better accuracy [26].

Finally, other biases may be introduced by the access policies of data and by their indexing on search engines. Privacy rules may impose to platforms not make some content public. Search engines may show biased results. While relevant contents may not be retrieved by search engines because they may lack appropriate tags or information, less relevant contents may appear among the first results because they are newer or cited by more websites. In general, search engine algorithms are industrial secrets. Some studies have been conducted to understand their mechanisms and biases, however, it is still impossible to get an exact description of their behaviors. Moreover, biases don't need to be explicitly hard-coded among the lines of these algorithms, but they may simply result in an undesired consequence of indexed data [54].

In the context of wildlife studies, particular attention is deserved by the presence of captive animals, altered images and replicated pictures. In our research we found that the first category is the one that can mainly affect a population estimate. However, as there are a few studies on assessing wildlife population using social media, there are no researches on the possibility to automatically identify captive and wild animals in pictures.

2.3 Social media and citizen science in ecological studies

Even though social media have been used by researchers in several fields, such as behavioral, political or seismological ones, only a few studies have made use of this data for ecological purposes.

2.3.1 Wildbook $^{\text{TM}}$

“Wildbook $^{\text{TM}}$ is an open source software framework to support collaborative mark-recapture, molecular ecology, and social ecology studies, especially where citizen science and artificial intelligence can help scale up projects” [28]. The core of the project is the Image Based Ecological Information System (IBEIS) [27], a Python-based application that facilitates detection and identification of animals in images. Wildbook $^{\text{TM}}$ is capable of scraping the Web for animal images, detecting several animal species, identifying individual animals, recognizing individuals among different pictures and tracking movements of these individuals using the data collected[5].

Wildbook $^{\text{TM}}$ has proved to be successful in analyzing thousands of images, accomplishing tasks unfeasible to even to the most expert biologist eyes. In fact, the detection and identification of thousands of animals would require a human to manually compare each animal of a given species to every other individual. This would not only be extensively time-consuming, but it is also an extremely complicated task for human sight since it is not easy for us to distinguish the complicated patterns that can identify a zebra or a giraffe.

2.3.1.1 Citizen science events

The power of the Wildbook ^{TM} framework allowed the creation of citizen science events dedicated to wildlife monitoring. The Great Zebra and Giraffe Count (GZGC) [42], the Great Grevy's Rally (GGR) [4] and the Great Grevy's Rally 2 (GGR2) [35] were all events organized to estimate the size of animal populations through the help of hundreds of volunteers. Scientists, together with passionate citizens, took thousands of pictures of zebras and giraffes. The pictures have then been analyzed using Wildbook ^{TM} and results have been used to provide an estimate of the Grevy's zebra population [37].

The data collected in GGR and GGR2 has been fundamental in our research. More information can be found in Chapter 4.

2.3.1.2 Animal wildlife population estimation using social media images

The thesis by the former UIC student, S. Menon, "Animal Wildlife Population Estimation Using Social Media Images" [30] [31], represents the first attempt to provide a way to use social media to estimate wildlife population. This previous work proved that the use of social media affects the final population estimate due to the presence of bias in the images shared online. Moreover, it was built a model to predict the likelihood for an image to be shared on social media, showing that this problem exhibits some learnable patterns. In this thesis, we aim to extend this work by considering two new elements.

First, the previous work analyzed the problem only at an image level. Instead, in this thesis we elevate the problem to the level of images collections. As mentioned in the introduction, we believe that the likelihood for an image to be shared cannot disregard other images taken by the same photographer.

Second, the previous research provides a model to predict the shareability likelihood of an image. In this thesis, instead, we aim at studying the reverse process. Given shared images, we want to estimate information regarding unshared images, such as the number of animals in unshared images. In other words, it has been already proved that there are some patterns that affect the way people share images. In this research, we want to understand how this pattern can be used to reconstruct the data that doesn't appear on social media due to the biases introduced by the way people share images.

2.3.1.3 Other works

Some other researches have made use of social media data to make ecological predictions. For instance, the potential of Twitter-mined information has been shown in the spatial and temporal monitoring of winged ant emergence, autumnal house spider sightings, and starling murmurations [13].

Problem Statement 3

The goal of this thesis is to provide an approach to estimate the number of wildlife animals of a given species using social media images collections. We believe that this problem can be instantiated into two main subproblems: the social media bias estimation and species population estimation. The proposed method should be species independent. It is not our ambition to create a unique model to be used on multiple species. Instead, we aim at providing a unique valid methodology to train and create different models for different species.

It is also worth noticing that, since it has been proved that the social media bias affects the population assessment when using social media data, our primary goal will be to design an effective framework to assess species health using these data, being the automation of the process only a secondary scope. However, we explored the possibility of automatizing the classification of images collections of wild and captive animals. Since this is the most time-consuming task of the data cleaning process, we believe it is useful to automatize this task in order to facilitate future data collections and eventually the creation of automatic tools to assess wildlife populations.

3.1 Population assessment

3.1.1 Social media bias estimation

Let us define C_i as the set of images belonging to the same images collection posted on a social media platform. For the species of interested, we require C_i to show at least a wild individual and none captive individuals. Let us define SD_i as the set of images in one or more SD cards from which the images in C_i have been taken.

First, given an images collection C_i shared on social media by a photographer P_i, we want to predict an estimate $\hat{N_i}$ for the real number of animals N_i present in the set SD_i of the SD cards of the photographer P_i. We model this problem as a regression model: given an images collection C_i, we map it to a set of features F_i which includes the number of individual animals in the collection C_i. Our regression problem is then to predict N_i from F_i.

3.1.2 Species population estimation

Finally, using a set of estimates N_i, we want to provide an estimate for the number of animals of the given species. In this thesis we will propose an approach to integrate traditional wildlife estimation methods with the regression problem proposed in the previous section.

3.2 Classification of captive and wild animals

Let us define S_i as the set of images belonging to the same images collection posted on a social media platform. We require S_i to show at least an individual of the species of interest. Considering the sole species of interest, we want to classify S_i as belonging to one of the two following groups: the set of images collections that show at least a captive individual (we will refer to this as the *positive class*), or the set of images collections showing none captive individuals (we will refer to this as the *negative class*).

Datasets 4

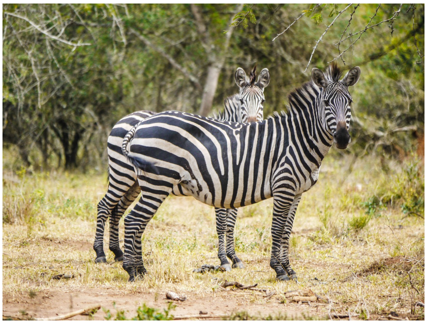
In this chapter, we will discuss the datasets used in our research. First, we will present the data we used to train our regression model. Then, we will discuss Flickr $^{\text{TM}}$ as a data source for wildlife population estimate and as a dataset to train a classifier for the identification of captive animals.

4.1 GGR and GGR2

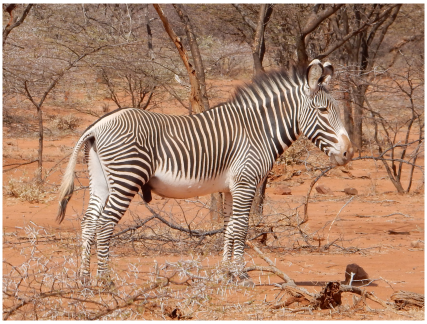
The Great Grevy's Rally is an event originally conceived as mean to census the Grevy's zebra (*Equus Grevyi*) population. Moreover, since the Grevy's zebra and the Reticulated giraffe (*Giraffa camelopardalis reticulata*) share a huge part of their habitat, during the second occurence of the GGR, the event organizers have decided to assess the health of both species in the territory.

4.1.1 Grevy's zebra

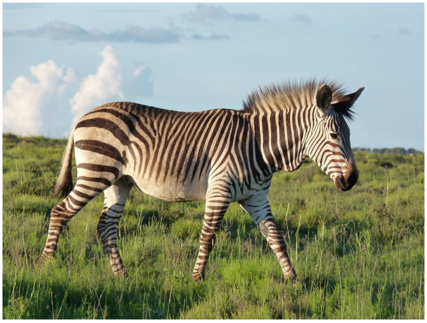
Zebras can be classified into three different species and further subspecies [58]. Among the species, the Grevy's represent the smaller population [15]. According to the last estimates [35], the population of Grevy's zebra counts 2,812 mature individuals that live in an area of 25000 kms^2 which spreads from the central part of Kenya to Southern Ethiopia. This dramatically small number qualifies the species to be listened in the IUCN Red List of Threatened Species [41].



(a) Plains zebras (*Equus quagga*) have wide stripes. They can be further classified into five subspecies: Burchell's, Chapman's, Crawshay's, Grant's and Selous zebras



(b) Grevy's zebras (*Equus grevyi*), also known as Imperial zebras, have the thinnest stripes.



(c) Mountain zebras (*Equus zebra*) have wide stripes similar to the ones of Plains zebras. However, Mountain zebras have a white belly. They can be classified into two subspecies: Hartmann's and Cape zebras.

Figure 4.1: Zebras can be classified into three species, shown above. From the most in danger to the less: Grevy's zebra is considered endangered, Mountain zebra vulnerable, and Plains zebra nearly threatened.

4.1.2 Reticulated giraffe

Until a few years ago, giraffes have been believed to be a single species with up to eleven subspecies [24]. However, in 2016 a new research found out that giraffes should be more appropriately distinguished into four species and an overall of seven subspecies [12]. Species distribution is shown in Figure 4.2 while differences in the aspect of different species are shown in Figure 4.3. Among the species, the Reticulated giraffes live in an area that covert part of Ethiopia, Kenya and Somalia, even though the majority of them live in Kenya [18]. Since they were just recently classified as a species, very few population estimates have been done in the past [18]. Since the species is considered to be endangered, it is now event more urgent to assess the health of this recently discovered species.

4.1.3 The event

Because of its classification as endangered species, the Grevy's zebra deserves a particular attention in the wildlife monitoring programs. Due to the difficulties encountered in the last aerial census, the Kenya Wildlife Service's Grevy's Zebra Technical Committee proposed the creation of a citizen science event where a large number of volunteers is asked to take pictures of the entire number of individuals of the species. Two events took place: the Great Grevy's Rally [4] (GGR) in 2016, and the Great Grevy's Rally 2 [35] (GGR2) in 2018. Scientists, together with volunteer citizens, were provided with cameras and jeeps to cover all the area inhabited by the species. Then, pictures have been gathered and processed by the Wildbook ^{TM} system. As a result, thousands of images have been collected in both the events. Each photo has been analyzed and animals of different species have been detected. Moreover, for Grevy's zebra (*Equus grevyi*) and Reticulated giraffe (*Giraffa camelopardalis reticulata*) each individual has been assigned a unique identifier. In fact, the Wildbook ^{TM} system has been capable of rec-

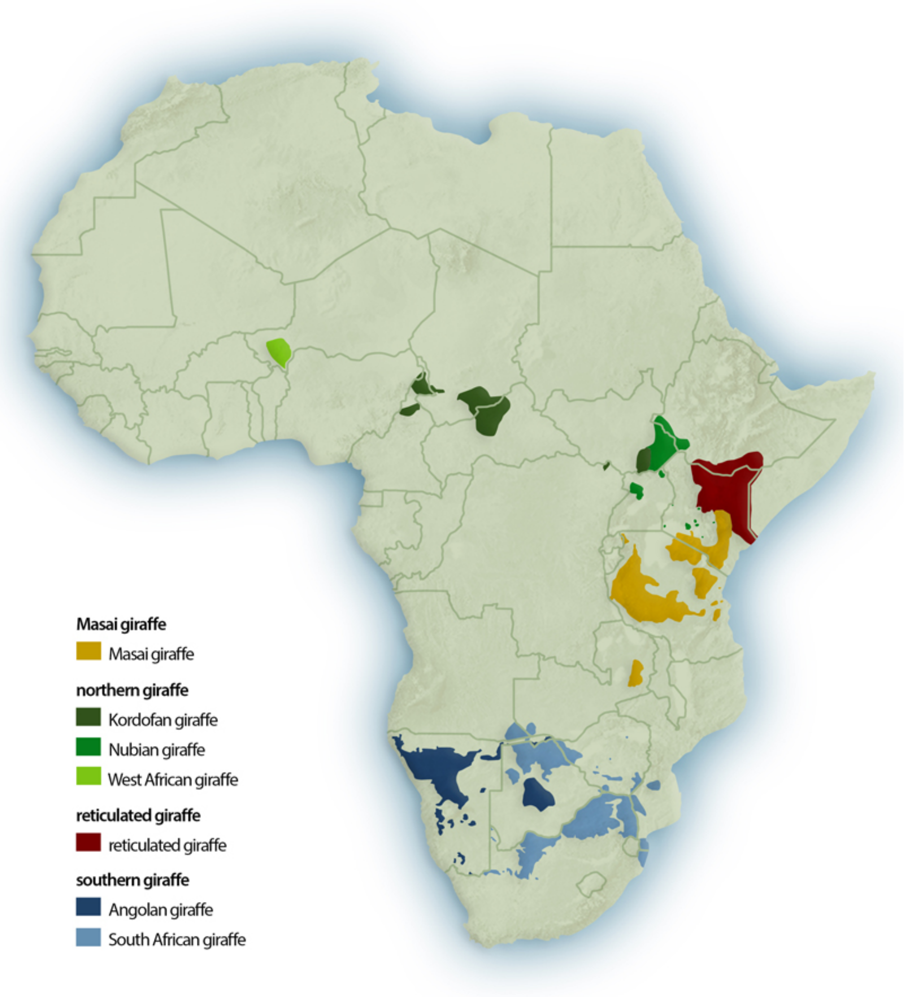
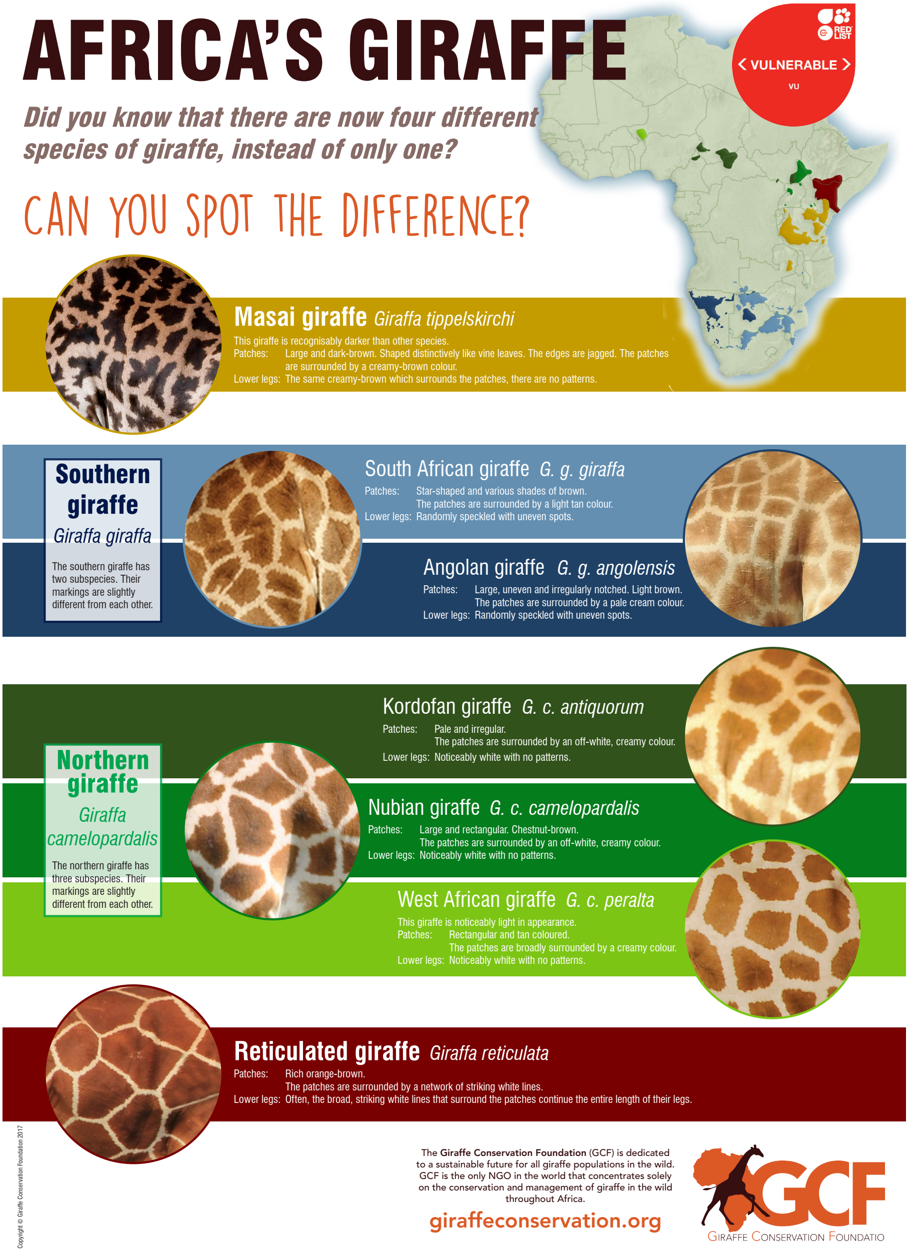


Figure 4.2: The map shows the territory inhabited by the different species and subspecies of giraffes. As shown in Figure 4.5, a big portion of the Reticulated giraffe's habitat has been covered by the GGR2. Image courtesy of Giraffe Conservation Foundation.



AFRICA'S GIRAFFE

Did you know that there are now four different species of giraffe, instead of only one?

CAN YOU SPOT THE DIFFERENCE?

Masai giraffe *Giraffa tippelskirchi*

This giraffe is recognisably darker than other species.
Patches: Large and dark-brown. Shaped distinctively like vine leaves. The edges are jagged. The patches are surrounded by a creamy-brown colour.
Lower legs: The same creamy-brown which surrounds the patches, there are no patterns.

Southern giraffe *Giraffa giraffa*

The southern giraffe has two subspecies. Their markings are slightly different from each other.

South African giraffe *G. g. giraffa*
Patches: Star-shaped and various shades of brown. The patches are surrounded by a light tan colour.
Lower legs: Randomly speckled with uneven spots.

Angolan giraffe *G. g. angolensis*
Patches: Large, uneven and irregularly notched. Light brown. The patches are surrounded by a pale cream colour.
Lower legs: Randomly speckled with uneven spots.

Northern giraffe *Giraffa camelopardalis*

The northern giraffe has three subspecies. Their markings are slightly different from each other.

Kordofan giraffe *G. c. antiquorum*
Patches: Pale and irregular. The patches are surrounded by an off-white, creamy colour.
Lower legs: Noticeably white with no patterns.

Nubian giraffe *G. c. camelopardalis*
Patches: Large and rectangular. Chestnut-brown. The patches are surrounded by an off-white, creamy colour.
Lower legs: Noticeably white with no patterns.

West African giraffe *G. c. peralta*
This giraffe is noticeably light in appearance.
Patches: Rectangular and tan coloured. The patches are broadly surrounded by a creamy colour.
Lower legs: Noticeably white with no patterns.

Reticulated giraffe *Giraffa reticulata*

Patches: Rich orange-brown. The patches are surrounded by a network of striking white lines.
Lower legs: Often, the broad, striking white lines that surround the patches continue the entire length of their legs.

The **Giraffe Conservation Foundation** (GCF) is dedicated to a sustainable future for all giraffe populations in the wild. GCF is the only NGO in the world that concentrates solely on the conservation and management of giraffe in the wild throughout Africa.

giraffeconservation.org

Figure 4.3: A genetic study in 2016 suggests that there are four different specie of giraffes. Each species is characterized by different spot and patterns. Moreover, each individual animal is characterized by its unique spots which allow to identify the individual among the entire species. Image courtesy of Giraffe Conservation Foundation

ognizing the same individual among multiple pictures. Experts have manually noted the values of different features for a portion of animals detected. Then, this data has been used to train machine learning models and to label all the unlabeled pictures. For each annotation of an animal of the previously cited species, the following features have been provided:

- Species of the animal
- Identifier of the individual
- Age of the animal
- Sex of the animal
- Viewpoint, intended as the side of the animal visible in the photo
- Geolocation of the annotation

The datasets resulting from GGR and GGR2 are unique in their kind. In fact, it is the first time that an entire species, the Grevy's zebra, has been cataloged in such a fine and granular way. Summarized information for these two datasets are shown in Table 4.1 while the areas covered by the events are shown in 4.4 and 4.5. In this thesis, we will use this data as the ground truth for the number of Grevy's zebras [4] [35], and, after an appropriate labeling process, to train a machine learning model, both for Grevy's zebra and Reticulated giraffe, with the goal of mitigating the biases related to the way people share images of animals on social media in the context of wildlife animal estimation.

4.2 Flickr TM

We chose Flickr TM as the source of images for social media data. The main reason behind this choice is that Flickr TM provides developers an

Table 4.1: AGGREGATED INFORMATION FOR GGR, GGR2, AND Flickr ^{TM}

	GGR	**GGR2**	**Flickr ^{TM}**
Images	40811	53194	123158
Photographers	162	212	837
Grevy's Zebra - Annotations	33151	54608	5138
Grevy's Zebra - Individuals	1942	2010	2931
Reticulated Giraffe - Annotations	0	30261	5482
Reticulated Giraffe - Individuals	0	1000	3473

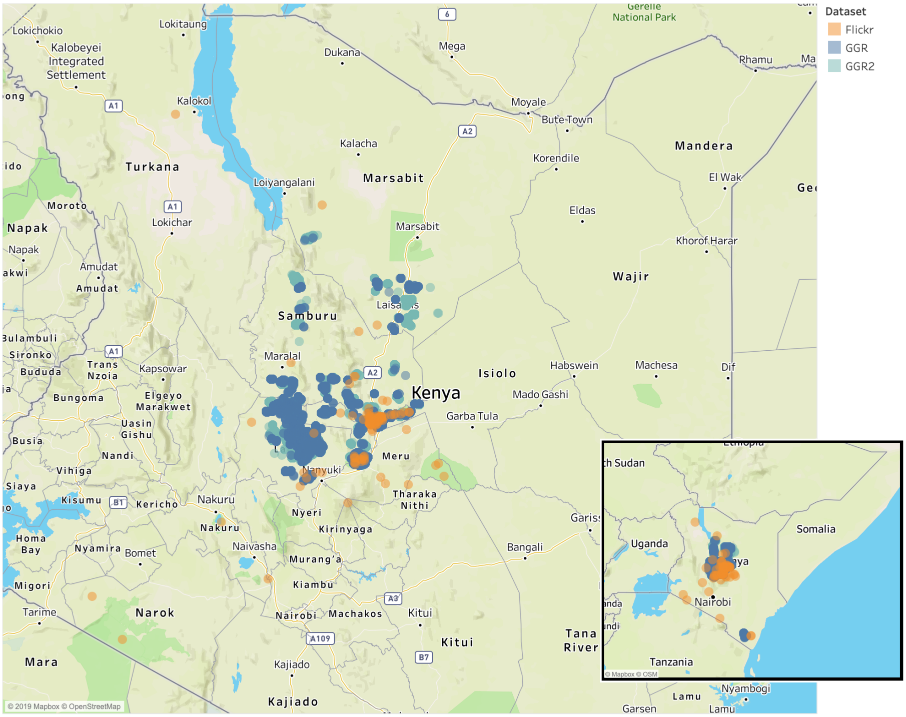


Figure 4.4: Map of GPS location for Grevy's zebra photos in Kenya.

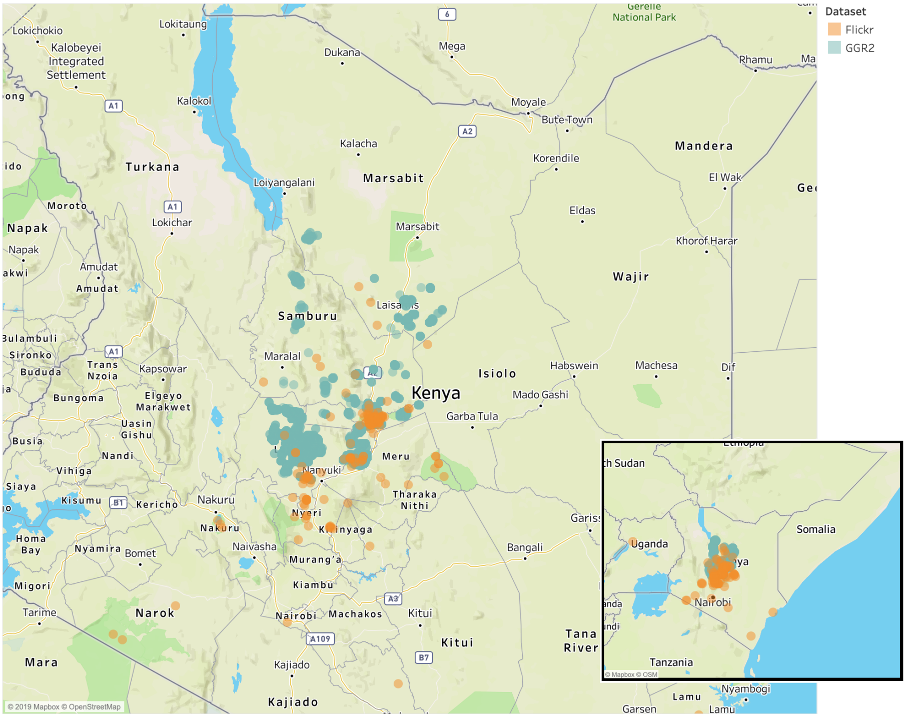


Figure 4.5: Map of GPS location for Reticulated giraffe photos in Kenya. Note: no picture of Reticulated giraffe was taken during the first GGR.

easy and unrestricted access through API [50], whereas other platforms are much more limited in the type and number of requests allowed. Of course private pictures cannot be retrieved because of privacy policies.

We scraped Flickr TM for Grevy's zebra and Reticulated giraffe images as described in this section. First, using the API, we searched for the keywords "grevy's zebra" and "reticulated giraffe" to retrieve a list of images. The search has been done using an implementation of the Flickr API [49] for Python and its function *flickrObj.photos.search* passing the mentioned keywords as *text* parameter one at a time. The function has been repeatedly called for each year from 2010 to 2018 setting *min_ taken* and *max_ taken* parameters.

The results of this search contained mainly images of the desired species. However, most of them were pictures taken in zoos, frequently in the United States or in England. Since our goal is to estimate the number of wildlife animals of a given species, we deleted those image. Their removal is particularly important, not only because these animals live in captivity, but also because they would introduce a systematic bias in the number of recaptured animals. In fact, these individuals are definitely more likely to be photographed and shared on social media than wild ones. Therefore, the use of these pictures would lead to underestimating the number of animals of the given species.

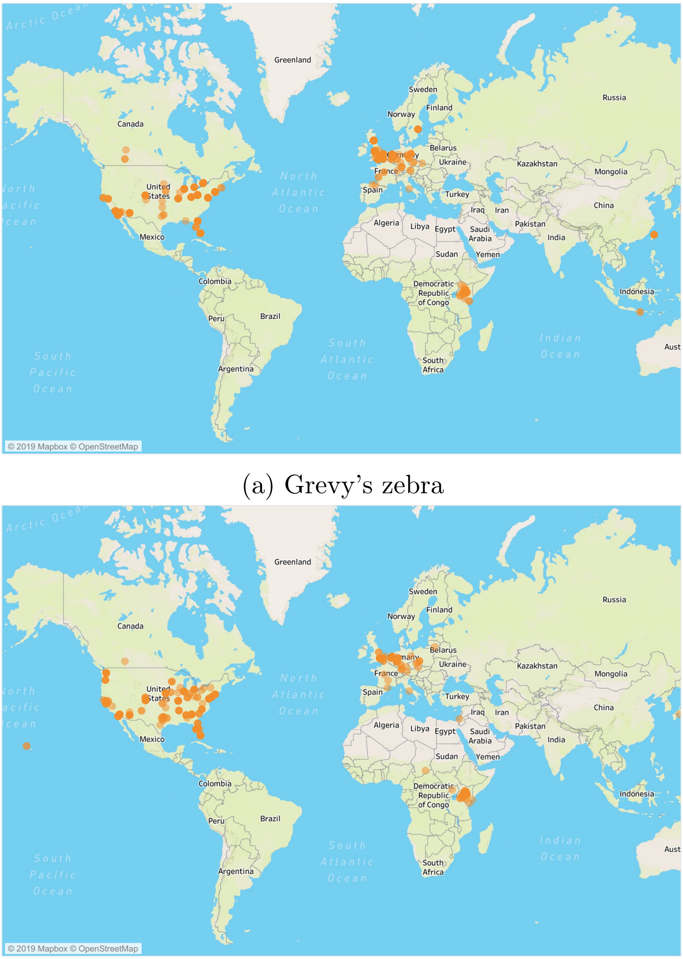
In order to filter those images, we deleted all the geotagged photographs whose location was outside Kenya and Ethiopia which represent the natural habitat of our species of interested. Moreover, we removed all the pictures that contained the keyword "zoo" in at least one among the their description, their album description or their album name. These filters drastically decreased the number of pictures available. Finally, to ensure the quality of our data, we manually controlled all the pictures and we deleted all the ones that were clearly taken in zoos. In fact, this kind of images can be recognized by the presence of fences that surround animals or by the presence of evident artificial elements, such as animal toys usually present in zoo enclosures. Finally, we deleted all the pictures that appear in the results of the search but that weren't showing neither Grevy's zebras or Reticulated giraffe. The result of this process was a set of pictures always showing at least an individual of our species of interested in their natural habitat.

We then took each of these pictures and, when available, we downloaded all the albums containing it. In fact, Flickr $^{\text{TM}}$ allows inserting the same photo in more than one collection. Downloading images collections lead to three main improvements. First, as explained in previous chapters, it allows studying the bias at a collection level instead of an image level. Second, it allows increasing the amount of images showing animals of the desired species. In fact, it is possible that not all zebras or giraffes present in a collection have been tagged, therefore some of them may have not been retrieved by the Flickr $^{\text{TM}}$ search. Third, by looking at albums, we discovered that some of the photographs we believed to be non-zoo images were instead taken in zoos: this conclusion originated in the presence of non-African animals in other pictures of the album. Moreover, we noticed a big portion of albums that were clearly artificial. We suppose they were created by animal enthusiasts who downloaded images from other websites to make their own collections. A portion of these albums can be identified by looking at the time period covered by the dates of the images: since images were most likely downloaded from the Internet, there is a high chance that, inside the same collection, images have been taken in different years. Of course, this is less likely to happen for images collections of vacations.

Aggregated data for this dataset is proposed in Table 4.1.

Table 4.2: PERCENTAGE OF Flickr $^{\text{TM}}$ IMAGES WITH GPS METADATA IN OUR DATASET. THIS NUMBER DO NOT EXACTLY MATCH THE NUMBERS AS IF WE RETRIEVE ALL THE IMAGES FROM Flickr $^{\text{TM}}$: IN FACT, AFTER WE RETRIEVED THE RESULT FROM THE SEARCH ENGINE, WE DOWNLOADED IMAGES COLLECTIONS ONLY FOR ANIMALS THAT WE BELIEVE TO BE IN THEIR NATURAL HABITAT, AND NOT FOR CAPTIVE ANIMALS. HOWEVER, CAPTIVE ANIMALS RETRIEVED FROM THE SEARCH ENGINE ARE INCLUDED IN THE COUNT, AS WELL AS THE IMAGES COLLECTIONS OF CAPTIVE ANIMALS WE DOWNLOADED BY ERRONEOUSLY IDENTIFYING SOME CAPTIVE ANIMALS AS WILD. NOTE: NO CAPTIVE ANIMAL WAS USED TO PROVIDE POPULATION ASSESSMENTS.

Species	**With GPS**	**Total**	**% With GPS**
Grevy's zebra	1214	3359	36.1%
Reticulated giraffe	1322	3766	35.1%



(a) Grevy's zebra

(b) Reticulated giraffe

Figure 4.6: Available location of Reticulated giraffe and Grevy's zebra photos retrieved from Flickr TM . As it can be understood from the locations, there are several images that do not show wildlife animals. In addition to those plotted on these maps, there were several photos without GPS metadata, more information in Table 4.2.

4.2.1 Classification of captive animals

In this work we also train a classifier model with the goal of identifying images collections where at least one of the animal of the species of interest is captive. This is done in order to avoid the need of manually distinguish collections of zoo or non-zoo images when scraping social media in the future.

For this task we will use all the images collection that we have downloaded from Flickr TM. It is worth noting that we haven't downloaded the images collection of every animal that we have retrieved from Flickr TM. Instead, we have done it only for those single pictures, retrieved using the Flickr TM API, that appeared to be in a wild context. However, as we found out that in many case these animals were actually in a zoo, we ended up with a dataset of several images collections, representing both positive and negative labels.

Table 4.3: STATISTICS ON THE IMAGES COLLECTIONS USED TO TRAIN THE CLASSIFIER FOR CAPTIVE ANIMALS IDENTIFICATION. WITH CAPTIVE ANIMALS WE REFER TO CAPTIVE ANIMALS OF THE SPECIES FOR WHICH WE WANT TO PROVIDE A POPULATION ESTIMATE.

Type	**Number**	**Percentage**
Images collections with captive animals	328	29.8%
Images collections without captive animals	773	70.2%

Methods 5

In this chapter, we will discuss the methods and tools used in this research. We will start by a brief introduction of the proposed solution.

5.1 Introduction to population estimation using social media images

Traditionally, wildlife population estimation is conducted by biologists who are in charge of both collecting and processing data. The quality of the data collection is assured directly by experts who have carefully planned the process. This procedure guarantees the correctness of the final estimate.

When using social media as a data source, we can't control the data collection process. Therefore, we need a method to estimate the biases that have been introduced in the data. This will allow us to reconstruct an unbiased dataset that can be used with traditional wildlife estimators such as the classical ones used by biologists. Estimating the number of animals using social media images collections is a non-trivial process. Therefore, it is necessary to decompose the problem into its parts.

We identify two main challenges:

1. Estimation of the bias related to the shareability of images collections
2. Estimation of the species population size using the estimates from the previous step

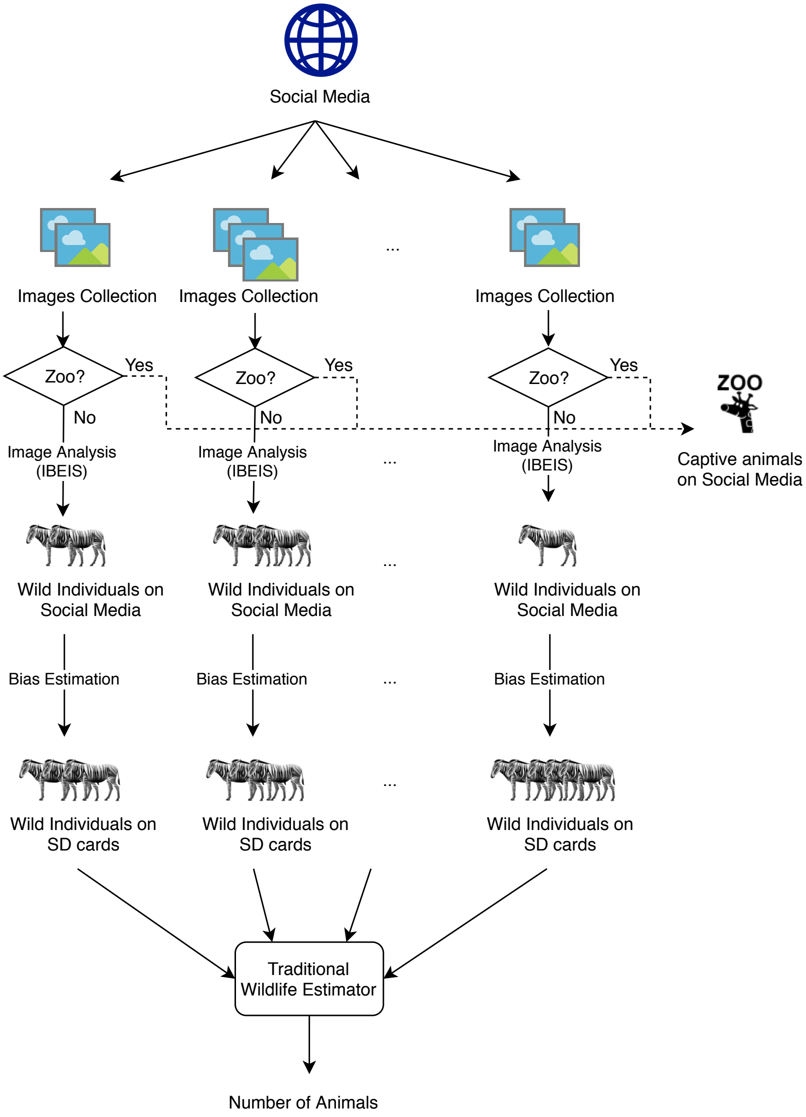

Social Media
Images Collection
Images Collection
...
Images Collection
Zoo?
Yes
No
Zoo?
Yes
No
Zoo?
Yes
No
ZOO
Captive animals on Social Media
Image Analysis (IBEIS)
Image Analysis (IBEIS)
...
Image Analysis (IBEIS)
Wild Individuals on Social Media
Wild Individuals on Social Media
...
Wild Individuals on Social Media
Bias Estimation
Bias Estimation
...
Bias Estimation
Wild Individuals on SD cards
Wild Individuals on SD cards
...
Wild Individuals on SD cards
Traditional Wildlife Estimator
Number of Animals

Figure 5.1: High level representation of the framework proposed to estimate wildlife animal population using social media images collections.

These two elements represent the main components of the framework we propose as solution, schematized in Figure 5.1.

As we explained in Chapter 3, a further problem is represented by the data collection and cleaning process. In Section 5.4 we will propose the use of a classifier to partially automatize this task. In all the other chapters we will assume the data retrieved from social medias have been properly filtered, as we manually did for our dataset as explained in Section 4.2.

The following sections provide an in-depth description of all the details of the framework proposed in Figure 5.1.

5.2 Overview of social media bias estimation

The process of estimating the social media bias can be decomposed in the identification and following reconstruction of several smaller biases.

In this chapter, we will assume that all of the images retrieved from social media have already been filtered as explained in Section 4.2. In briefs, we assume the images collections retrieved to show picture of wild animals. In fact, the problem of discarding captive animals can be separated from the issues related to the presence of other, more complex biases explained in this section: while the identification of captive animals can be modeled as a binary classification problem, we propose a solution based on a regression to face the other biases.

An in-depth decomposition of the biases is provided in Figure 5.2.

The first bias is the one relative to the portion of individual animals that can be observed by humans. This bias is almost impossible to be deleted, even for biologists. In fact, it represents the motivation for the birth of wildlife population estimation models. Whereas biologists usually try to cover all the area inhabited by the species, in the case of social media data we hope that the increase in touristic activities all around the world [53] could mitigate this bias. Moreover, it is worth noticing

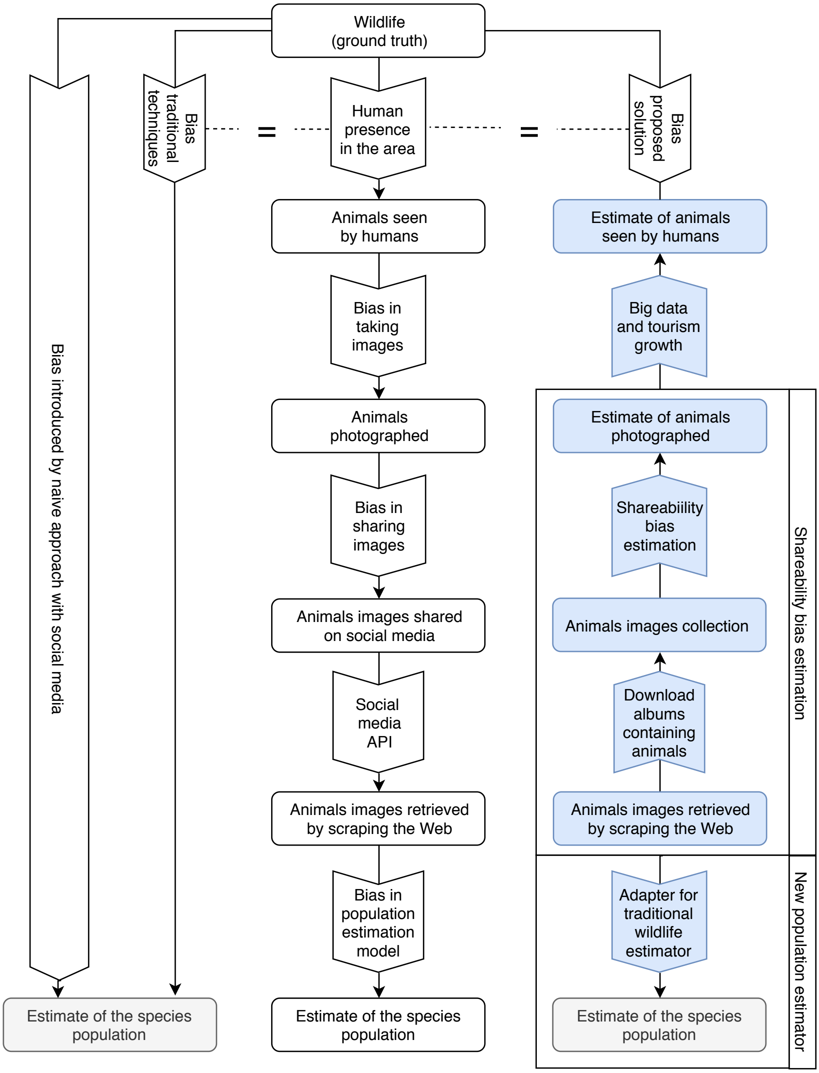

Wildlife (ground truth)
Bias introduced by naive approach with social media
Bias traditional techniques
=
Human presence in the area
=
Bias proposed solution
Animals seen by humans
Estimate of animals seen by humans
Bias in taking images
Big data and tourism growth
Animals photographed
Estimate of animals photographed
Bias in sharing images
Shareabiility bias estimation
Shareability bias estimation
Animals images shared on social media
Animals images collection
Social media API
Download albums containing animals
Animals images retrieved by scraping the Web
Animals images retrieved by scraping the Web
Bias in population estimation model
Adapter for traditional wildlife estimator
New population estimator
Estimate of the species population
Estimate of the species population
Estimate of the species population

Figure 5.2: In white, high-level representation of the biases that may affect wildlife population. In blue, proposed solution to estimate these biases.

that another mitigation may come by biologists themselves: they may be active social media users and it may be reasonable to suppose that they are likely to share pictures of the animals they study.

A second level bias accounts for the way people take images. When collecting data, biologists will try to do an exhaustive work, limiting the number of animals seen but not captured. Their goal will be of course to note all the animals seen. When using social media, we can't count on the same effort from every user. However, it is our hope that the volume of data available on social media may alleviate this issue. Individuals that haven't been photographed by a group of tourists may be later captured by some other group.

A third bias is introduced by the shareability of images. Even if big data may again mitigate this effect, we believe that there are opportunities for an in-depth study of this phenomena. Moreover, previous researches [30] [13] have highlighted the need of expanding our understanding of this bias. In our solution, we propose to train a machine learning model over a custom dataset generated for the species of interest. This dataset can be created by providing people different sets of pictures and asking interviewees which images they would share on a social media platforms. Then, we propose to train a machine learning model to predict information related to the photos people will not share, based on the photos that people would share. To generate this dataset we suggest to use crowdsourcing marketplace. More details can be found in Section 5.2.1.

A fourth bias is related to search algorithms used by social media platforms. Some relevant images may not appear among the results found by the search engines. These engines are not the only responsible for these phenomena. It could be the case that users haven't used appropriate tags or descriptions when uploading their content on social media platforms. As explained in Section 4.2, downloading albums associated with each picture retrieved by search engines may help to collect more data. Another improvement may reside in the use of multiple search engines: like every other website, social media platforms are indexed by external search engines such as Google$^{\text{TM}}$ or Bing$^{\text{TM}}$.

Finally, the model used to estimate the size of the species population may have some intrinsic biases. For instance, when using capture-mark-recapture (CMR) methods with pictures, a portion of recaptures may be lost because the same animal may show different sides in different pictures. In fact, in the case of zebras, the same animal may show its left side in a picture and its right side in another one. Due to the lack of symmetry in zebra stripes, computer vision tools, as well as biologists, may not be able to recognize the same individual. This issue may affect the final population estimate since CMR models are based on the hypothesis that marks are not lost, in the sense that an animal captured in a previous time instant can always be recognized in successive ones. Further considerations are made in Section 5.6.5.

In the following section, we will describe how crowdsourcing marketplace can be used to construct a dataset for the training of machine learning models to deal with shareability biases, as mentioned in this section. In particular, we will focus on Amazon Mechanical Turk$^{\text{TM}}$ since it has been our choice for the data collection.

5.2.1 Introduction to Amazon Mechanical Turk$^{\text{TM}}$

Amazon Mechanical Turk$^{\text{TM}}$ (MTurk) is a web service that provides manpower for tasks that require human intelligence, and cannot be accomplished by computer algorithms. These tasks are called Human Intelligence Task (HIT). Users on this platform are divided into two categories:

- **Requester** :

  Requesters post HITs to MTurk. They pay Amazon and optionally one or more workers to have their HITs completed.

- **Worker** :

  Workers complete HITs, usually in return for a money reward.

A HIT could be completed by one or more workers. An assignment is the set of answers given by a worker to a HIT. Since more than one worker could complete a HIT, a HIT could be linked to one or more assignments. In the last years, MTurk popularity has raised due to the possibility of recruiting workers rapidly and inexpensively [6].

5.2.2 Creation of a suitable machine learning dataset using MTurk

In our research, we used MTurk to create a dataset to be used in the training process of the machine learning models. In particular, we lacked the ground truth regarding preferences of people in sharing images. Using the images from GGR and GGR2 we created a number of surveys. Each survey contained a number of images, and for each of these pictures we asked whether the user would have shared it on social media or not. The intent was to simulate the process that a person would face when choosing which pictures to post online. The set of pictures we showed simulates the photographs on the SD card of a tourist after a hypothetic safari in Kenya. To prevent MTurk workers from choosing random pictures, we suggest to force them to answer a question for each image. At the bottom of the questionnaire, interviewees were able to review the pictures they selected and to eventually drop some of them. We believed that this final review could give workers a more complete overview of their choices, leading to the possible elimination of similar pictures. The review process has been also introduced to simulate the ones proposed by several social media platforms when uploading a set of pictures. Finally, a few more questions were proposed to interviewees to collect information about the population of participants and to evaluate the consistency of their results.

5.2.3 Ensuring data quality

We identify three main threats to the quality of the data collected through crowdsourcing marketplace. First, the presence of automatic bots among workers. [11] Second, the possibility that some workers could answer randomly to questions. Finally, even some well-intentioned workers may find themselves in the condition of answer randomly whenever technical issues, such as slow connection bandwidth, don't allow them to see all the images.

In order to deal with these risks, we propose a number of precautions that could be put in place. First, on popular crowdsourcing platforms, bots should be handled by the platform itself. These platforms usually provide anti-bot measures. On August 2018 panic arose about the possible presence of bots on MTurk [11]. However, further researches [32] showed that the quality drop in workers answers could be related to the presence of foreign workers. The study found no single evidence of bot, with not a single worker failing a re-captcha question. In addition, 95% of farmer workers (workers who submit most of their HITs from server farms) failed a basic English proficiency screener. Moreover, we believe that bots and bad workers could show some common traits, and given the unlikely presence of bots, we suggest to focus on the second threat as explained in this section. First, MTurk gives the chance to restrict the set of candidate workers to “master" workers. These are workers that “have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of requesters" [51]. The use of the “master" qualification requires a little increase in the cost per HIT but guarantees a consistent improvement in the quality of data. Moreover, one can set a threshold to filter results based on the time spent by workers to complete the surveys. We believe that a conscious choice should require at least a second for each image. In section 6.3 we will motivate this hypothesis.

Technicals issues could be dealt with in two different ways. First, since we didn't want to compromise the quality of pictures shown, we suggest mitigating the low-bandwidth issue by implementing client-side code to force the browser to load pictures in the order they will be seen by workers. Even though one could expect browsers to perform this task autonomously, we found that this was a common issue. Second, to take into account any other possible failures, we suggest asking workers directly whether they were able to see all the pictures. By explicitly stating that workers would have been paid regardless of their answer, we encouraged interviewee's honesty and preserve data quality.

5.3 Features

Different types of features have been extracted both to train machine learning models and to feed animal wildlife estimator methods. In this section, we provide an exhaustive list of the features we used and an explanation of the reasons behind these choices.

5.3.1 Features types

5.3.1.1 Biological features using Image Based Ecological Information System

Image-Based Ecological Information System is a cloud-based system capable of detecting several animal species. Moreover, for some animal species, the system provides tools to identify individual animals and to extract biological features.

Different categories of identifiers are used in Wildbook $^{\text{TM}}$ to identify images, photographers, animal annotations or animal individuals

Table 5.1: LIST OF IDENTIFIER TYPES USED BY Wildbook ^{TM}

ID Type	Meaning
GID	Image Identifier
AID	Annotation Identifier
NID	Individual Identifier
CID	Contributor Identifier

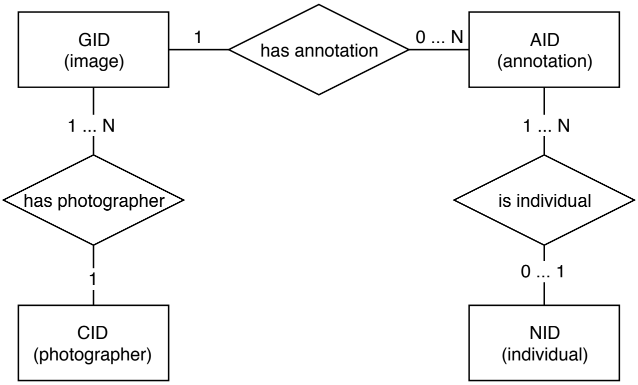


Figure 5.3: Cardinality relationships between different ID types on Wildbook ^{TM}.

as shown in Table 5.1. More precisely each image is identified with a unique GID. Each image may contain zero or more animal annotations identified by a unique AID. Finally, for some species (see Table 5.2), when the quality of the picture is good enough, each AID is matched to a NID denoting a unique individual of the given species. In other words, a NID is a unique name for an individual which may appear in different pictures. Each appearance is identified with an AID. Cardinalities of the cited relationships are summarized in Figure 5.3

Wildbook ^{TM} can be queried through API to gather different information about the identified animals. Table 5.2 lists features that can be extracted for Grevy's zebra, Plains zebra, Masai giraffe and Reticulated giraffe.

Table 5.2: LIST OF FEATURES PROVIDED BY Wildbook ^{TM} FOR GREVY'S ZEBRA, PLAINS ZEBRA, MASAI GIRAFFE, AND RETICULATED GIRAFFE

Attributes	**Description**	**Domain**
AID	Annotation ID	Integer
NID	Individual ID	Integer
Viewpoint	Side of the animal visible in the picture	back, backleft, backright, down, downfront, downleft, front, frontleft, frontright, left, right, up, upback, upfront, upleft, upright
Age	Age of the animal	Unknown to 2 Months, 3 Months to 5 Months, 6 Months to 11 Months, 12 Months to 23 Months, 24 Months to 35 Months, 36 Months to Unknown
Sex	Sex of the animal	Male, Female
Bounding box	Area of the picture delimiting the animal	Pixel coordinates
Geolocation	GPS coordinates of the picture	Latitude, longitude

5.3.1.2 Beauty features

We extracted a variety of features that have already been proved to be a useful indicator of the beauty of an image [44] [25] also for animals pictures [30] . These features are not correlated to any semantic aspect of an image, instead, they describe color and symmetries in the picture.

Any image described by RGB channels can be decomposed into HSV channel [47]. Let $\bar{V}$ and $\bar{S}$ be the mean values of the corresponding channels in the HSV representation, we extract the following features:

$$Arousal = 0.76\bar{V} + 0.32\bar{S} \tag{5.1}$$

$$Dominance = -0.31\bar{V} + 0.60\bar{S} \tag{5.2}$$

$$Pleasure = 0.69\bar{V} + 0.22\bar{S} \tag{5.3}$$

In order to model the variance in Hue, Saturation and Brightness we used histograms calculated by binning the HSV values into 12, 5 and 3 bins respectively. These particular numbers correspond to the number of pure colors for each channel in the Itten's color model. For each image we then extract the following features:

- Standard Deviation of HSV-itten histograms for Hue channel
- Standard Deviation of HSV-itten histograms for Saturation channel
- Standard Deviation of HSV-itten histograms for Brightness channel

Contrast has been modeled according to Weber's definition:

$$Contrast = \frac{Y_{max} - Y_{min}}{\bar{Y}} \tag{5.4}$$

where Y_{max}, Y_{min}, $\bar{Y}$ are maximum, minimum and mean of the luminance channel which can be calculated at pixel level as:

$$Y = 0.299 * R + 0.587 * G + 0.114 * B \tag{5.5}$$

where R, G, B are the values of the corresponding channels in the RGB representation. Finally, we use a gray scale representation of the image to compute indicator for its entropy and its symmetry [30].

5.3.1.3 EXIF features

The Exchangeable Image File Format is a standard for storing metadata in image files. Digital cameras save in EXIF data a variety of information regarding the camera used and its setting at the time of the capture. The features we extracted are:

- Date and time of the shot
- Geolocation

While the time of the shot can always be found, it is very common for geolocation to be missing on social media images. This can happen because the camera may have not been provided with GPS or because of the privacy policies adopted by social media users. However an approximate geolocation can be often inferred from tags and description of online pictures.

When the animal species studied lived in a delimited area, geolocation may be needed only to filter zoo images. The date of the photograph is instead always needed to estimate the number of animals since capture-mark-recapture methods relies on recapturing the same individuals in different time instants.

5.3.2 Modeling the images collection with time series

It is worth remembering that we are conducting a study on social media bias at the level of images collections, and not at the level of single images. In fact, we strongly believe the presence of other pictures in the SD card of a photographer may affect the shareability of individual pictures. Therefore, it is necessary to extrapolate meaningful information related to set of pictures. Since the features we have are computed on single images, we need methods to aggregate those values for an images collection. Instead of considering only simple aggregation variables, such as maximum, minimum or standard deviation, we decide to use time series to model the problem. Using biological and beauty features listed in the

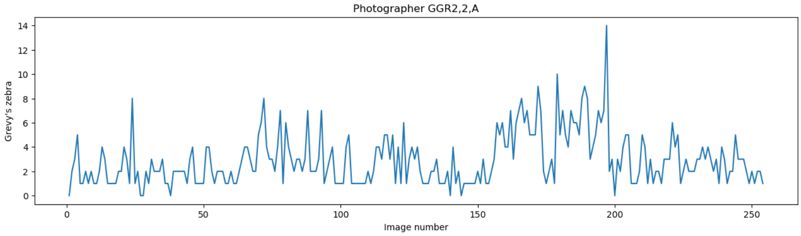


Figure 5.4: Time series showing the number of Grevy's zebra per image. Images are sorted by their date but then absolute time is not taken into account.

previous section, we create time-series aiming at modeling the order of images in an album. First, we sort the images in an album by the date they have been taken. Then we enumerate them, such that, given an album with N images, each image was assigned a number from 1 to N representing the relative temporal order of images in the album. Finally, for each feature, we create a time-series by modeling on the x-axis the relative temporal order images in the album, and on the y-axis the corresponding feature values, as shown in Image 5.4 . Therefore, instead of having proper time on the x-axis, we just have the order of images. This choice relies on the belief that, in this context, absolute time is less meaningful than the simple relative position of images. In particular, we consider that to be true for the labeling process on Amazon MTurk, where the interviewees had no notion of the time at which the photos were taken. Therefore, with respect to time series theory, we assume the sampling time between pictures to be constant. Our focus is indeed on the images collection, not on the time intervals between pictures. These time intervals can instead be modeled as a feature, as we do for all the beauty features. For instance, it wouldn't make sense to consider the arousal of images over absolute time. Instead, it may be more appropriate to consider it over the list of images, comparing the arousal of each image with its adjacent images.

We rely on the FRESH algorithm [9] to extract a huge number of features for time series. This algorithm has been shown to be particularly useful when domain knowledge is poor. The approach consists in extracting a huge number of features and then performing statistical tests to select only the robust ones.

FRESH performs a set of univariate features mapping, extracting more than 2000 features per time series. Then, for each feature, it deploys a singular statistical test against the null hypothesis:

$$H_0 = \{\text{The features are irrelevant for predicting the target}\} \tag{5.6}$$

In particular, since neither our target nor our features are binary, FRESH [8] [9] performs Kendal rank test to test whether two continuous variables are independent. Finally, FRESH performs the Benjamini-Yekutieli [3] procedure to control the False Extraction rate, that is:

$$FER = \mathbb{E}\left[\frac{\text{number of irrelevant extract feature}}{\text{number of all extracted features}}\right] \tag{5.7}$$

At the end of the procedure, FRESH returns a number of features which are guaranteed to be statistically correlated with the target values. However, since some of the features returned may be highly correlated between them, it is suggested [8], for certain models, to normalize the features and perform a Principal Component Analysis (PCA). PCA can be performed immediately after the features extraction, or at the end of the entire procedure. For the machine learning model who would benefit from normalization or PCA, we chose to adopt the latter option, which is known as *FRESH_PCAa(fter)*. This is because we want to be sure that PCA is performed on a set of robust features. The choice to use PCA, for the models that might benefit from it, reflects our preference in achieving a better accuracy over easier interpretability of the results.

There are however some hypothesis we wanted to test. Therefore, we developed dedicated experiments for them.

5.3.3 MTurk experiments to understand features importance

The bias related to certain features can greatly affect the population estimate. Therefore, we believe useful to test directly whether some factors may affect the way people share images.

We identified four main factors that could affect the final population estimate:

- Number of pictures in the SD cards
- Number of animals in the SD cards
- Presence of certain individual animals that may be considered more shareable than other ones
- Side shown by animals in pictures

The first three factors can greatly affect the estimate of an animal population because they are strictly related to the presence of a certain individuals on social media. The last factor instead can influcence the estimate because of the use of photos in our model: if the same animal shows different sides in different images, it won't be possible to recognize it as the same individual because certain species lack bilateral symmetry. A solution to this issue is proposed in Section 5.6.5.

For each of the factors, we have then identified different values in its domain, and test whether a different distribution of these values changes the percentage of images shared by a user. Specifically, we design each of the experiment in the following way:

1. Idenfity a factor of interest

2. Idenfity different V values v_i in the domain of the factor chosen

3. For each value v_i, generate a set of E experiments where the factor chosen has value v_i.

4. Deploy all the experiments on Amazon MTurk and retrieve the results

5. Define a sample S_e as an array of E elements whose components are the percentage of images shared in the experiments corresponding to the values v_i. If we have V different values than we have V different samples. If we have E experiments for each value v_i, then our samples are of length E.

6. For each value v_i, perform a one-way ANOVA test to test the null hypothesis H_0:

$$H_0 = \ \{ \ \mu_1 = ... = \mu_i = ... = \mu_V \ where \ 1 \leq i \leq V \ \} \tag{5.8}$$

   where μ_i is the mean of the sample S_e.

The last test, "Presence of certain individual animals that may be considered more shareable than other ones", slightly differs from the others because there are no explicit factors or values to be tested. To design the experiments, we propose to identify a set of individuals for which a sufficient number of images is available. Then, for each possible pair (A_i, A_j) of individuals in the set, it is possible to create an experiment in which half of the pictures are of the invidividual A_i and the other half are from the individual A_j.

5.4 Classification of captive and wild animals

Most of the biological methods used to estimate wildlife population rely on counting recaptured animals in different time intervals, as briefly explained in Section 1.3 and more precisely in Section 5.6.2. In particular, if the number of recaptured animals increases, the estimate of the population size will decrease, leading to an underestimation of the real number of animals.

When scraping social media for animal images, the quantity of images showing captive animals retrieved can represent the majority of the pictures. Interesting data are shown in Figure 4.6 and Table 4.2. This is likely to happen because of the popularity of zoo among people. Since for endangered species the number of recapture animals among different years can be very small, the presence of captive animals will greatly interfere in the population assessment.

In our experience, we found that many photos do not have any GPS metadata nor a textual description associated. Therefore, there is no other possibility than rely on the images themselves to classify them.

While the identification of captive animals can be done manually, we found this process extremely time-consuming. Therefore, we believe it could be useful to train a classifier capable of identifying captive animals. This can be an extremely complicated task, especially when trying to classify a single image showing an animal in the foreground and almost no background. However, for the purpose of our research, we need to classify images collections as a whole, instead of single pictures. This different setting makes the task more approachable, since it is likely to find at least some valuable features among the entire set of pictures. While some of the features might vary based on the species studied, we believe that this approach can at least be useful in accelerating the data cleaning process for future years. It also likely that the same classifier could at least be used for animals belonging to the same habitat.

First, images of the given species must be downloaded from social media platforms. Then, all or a portion of them should be appropriately labeled. A suitable set of features must be extracted, and finally a classifier is trained to distinguish images collection not showing captive individuals of the desired species from the other images collections. The classifier can then be used in the future to discard images collections that should not be included in the data source. Also, whenever it might be difficult for an human to label an images collection, the classifier might suggest the more appropriate tag.

We also suggest to keep note of the results of the classifier, and use them to create and update a comprehensive dataset of captive animals using Wildbook $^{\mathrm{TM}}$. This dataset can be matched against every individual animals retrieved from the Web in order to discard the undesired individuals.

5.4.1 Additional features

We believe that the classification problem can benefit from additional features with respect to the ones needed for the bias estimation problem. In fact, while in the latter we are somehow modeling the quality, the beauty, and attractiveness of photos, in the first problem we are evaluating more objective elements.

In particular, we believe that the key to succeed in this task is the identification of a large number of objects and animals. Crucial features might be the presence of urban elements in photos, as well as the presence of other animal species. The latter could be very useful when we can identify animals of species not pertaining to the habitat of the species of interested. In fact, it is often the case that a zoo will have animals of species living in different part of the worlds, such as tigers and zebras. Also, wildlife pictures may show animal species that are not popular in zoo. By automating the process of features extraction, we can train a machine learning classifier capable of analyzing all these elements.

We suggest to use the state-of-art object detection library to extract the largest possible quantity of features. In fact, the development of a general purpose detection framework falls outside the scope of this work.

We also recommend to make use of the other features listed in this work (see Section 5.3) for two reasons. In the first place, they include the detection of several animal species. Moreover, they include different features that can model colors and textures of the picture. While these features aim at modeling the beauty of an image when studying its shareability, it is likely that they will also be capable of modeling recurrent patterns in the habitat of the species of interest, such as the predominance of yellow tonality of savannah photos.

To conclude, aggregated features can be extracted according to the time series theory as explained in Section 5.3.2.

5.4.2 Evaluating the classifier performance

Let us define the define as positive class the set of images collections that shows at least a captive animal of the species of interest. As in any binary classification problem, there are two possible types of error: a type I error (false positive) and a type II error (false negative). Table 5.3 show the meaning of these two errors in our problem.

5.4.2.1 Type I Error - False positive

In this case a wild animal will not be included in the dataset. In this situation, it might happen that we will miss a recapture, with a strong impact on the final estimate of the population, which will be overestimated. However, since the number of recaptured animals, at least for the species of interest, has been noticed to be very small with respect to the population size, it is unlikely that we will miss a recapture.

5.4.2.2 Type II Error - False negative

In this case a captive animal will be included in the dataset. It is likely that the species size will be underestimate. This will happen if we encounter more false negatives relative to the same captive individual corresponding to different time intervals. This situation is not unlikely to happen. In fact, the number of captive animals, which correspond to the animal detained in zoos, is usually small with respect to the size of the entire wild population. However, a huge portion of social media photos show captive individuals. As a consequence, at least for certain captive individuals, there will be many photos on social media. Therefore, in case of a false negative, chances are high that the error will involve the same individual. Even in best scenario, this probability will not be negligible.

Table 5.3: CONFUSION MATRIX FOR CAPTIVE ANIMAL CLASSIFICATION.

	Actual Positive	**Actual Negative**
Predicted Positive	TP: captive animals correctly classified	FP: animals classified as captive but actually wild
Predicted Negative	FN: animals classified as wild but actually captive	TN: wild animals correctly classified

5.4.3 Metrics

Given that a false negative is more likely have an impact in the population estimate than a false positive, our goal will be to minimize the number of false negative. Therefore, our primary evaluation metric will be the Recall function:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5.9}$$

We will also provide the results for other common metrics: accuracy, precision and F1.

$$Accuracy = \frac{True\ Positive + \text{True Negative}}{True\ Pos. + True\ Neg. + False\ Pos. + False\ Neg.} \tag{5.10}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.11}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{5.12}$$

5.5 Social media bias estimation as a regression problem

We model the bias estimation as a regression problem. Informally, given an animal species and an images collection shared on social media, the goal is to estimate the number of individuals of the given species that have been photographed by the photographer. This number includes both the individuals shared on social media and the individuals that have been photographed but not shared on social media. In other words, we want to predict the total number of individuals present on the SD card of the photographer. Of course, the focus is on non-shared individuals, since the shared ones can be extracted using Wildbook $^{\text{TM}}$. Indeed, that latter will constitute parts of the features of our model.

More formally, we train a regression model to predict the *percentage bias* which is our target value. It is defined as:

$$Percentage\ Bias = \frac{N.\ individuals\ shared\ on\ social\ media}{N.\ individuals\ in\ SD\ card} \tag{5.13}$$

Then, we can compute the total number of individuals in the SD card by reversing the formula:

$$N.\ individuals\ in\ SD\ card = \frac{N.\ individuals\ shared\ on\ social\ media}{Percentage\ Bias} \tag{5.14}$$

The estimated number of individuals in the SD card is then used to feed traditional estimator models as explained in Section 5.6.

The model can be trained using the surveys deployed on MTurk (see Section 5.2.2) and it can then be used to predict the number of animals on real social media data. Since different species could be related to different shareability bias, we recommend repeating the data collection process and the model training for each species of interest.

5.5.1 Metrics for quantifying regression performance

In order to evaluate the performance of our models we used the Mean Squared Error (MSE) and the coefficient of determination, or R^2:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.15}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{5.16}$$

where $\hat{y}_i$ is the predicted value, y_i is the real value, and $\bar{y}_i$ is the mean value of the target values:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.17}$$

5.6 Population Estimation using the Jolly-Seber method

The machine learning model described in Section 5.5 can be used to predict the number of animal individuals captured by a photographer given an images collection shared on social media. However, this model is not sufficient to estimate the number of animals of an entire species. For the final estimate, we need to use a traditional wildlife estimator. Among the capture-mark-recapture methods, described in Section 2.1, we decide to use the Jolly-Seber method. In the following sections we will motivate our choice, describe the method, and explain how we propose to integrate it with our regression model.

5.6.1 Motivation behind the use of the Jolly-Seber method

The Jolly-Seber method extends the simple CMR method (see Equation 2.2) to open populations. A population is closed when it doesn't change in size during the period of observation. Therefore, a population can be usually assumed to be closed only for a short time of period. A population is open when it not closed. This assumption is more realistic and it takes into account the effect of births, deaths, and movements of animals. More specifically, given the area on which the study is being conducted, the assumption of open population allows animals to born and die in this area, as well as enter and exit its border. Moreover, with respect to the more basic types of capture-mark-recapture (CMR) methods, the Jolly-Seber method combines recaptures over multiple time instants. In fact, while the basic CMR formula (see Equation 2.2) makes use of only two time instants, the Jolly-Seber method extends this concept to provide an estimate for a time instant t_n by taking into account multiple time instants in the past and multiple time instants in the future. Quoting [21], the Jolly-Seber method assumes that:

1. "Every individual has the same probability of being caught in the *t-th* sample, whether it has been marked or it is still unmarked"

2. "Every marked individual has the same probability of surviving

from sample t_n to sample t_{n+1}"

3. "Individuals do not lose their marks and no mark is overlooked"

4. "Sampling time is negligible with respect to the time intervals between samples"

With respect to these assumptions, we believe that the first two cannot be assumed to be true a priori for every species. The truthfulness of these hypotheses are related to a number of complicated factors. When using social media as a data source, we suggest analyzing the connection between the species habitat and close touristic activities. Furthermore, the presence of research centers could affect the number of sightings of individuals. Finally, the veracity of the second hypothesis may be affected not only by human activities but also by other biological factors.

The third assumption instead may be assumed true only if mitigated using statistical methods. In fact, since the marks we use are biological features, such as stripes for zebras, they cannot be lost. However, the lack of bilateral symmetry for certain animal species may lead to the miss of some marks. For such situations, we propose a solution in Section Section 5.6.5. It should be also considered whether certain species may change their distinctive individual patterns over the course of the years.

The last assumption can hold because we will check for recapture among data coming from different years, and more precisely, as explained in 5.6.4, we will compare every images collection of a given year with every other images collection from other years. Since we can arbitrarily discard images collections that cover a long time period, we have control on the truthfulness of this assumption. However, it might be the case that considering multiple images collections as a unique sample for an entire year may lead to a violation of this assumption, even though by expanding the study to several year this issue can be mitigated.

5.6.2 The Jolly-Seber method: the model

In this section, we will describe how the original Jolly-Seber method allows estimating the number of animals of a given species. The definition are taken from [21]. Given a sample taken at time t, let:

- "m_t = number of marked animals caught in sample t "
- "u_t = number of unmarked animals caught in sample t "
- "n_t = total number of animals caught in sample t "
  $n_t = m_t + u_t$
- "s_t = total number of animals released after sample t "
  s_t = (n_t - accidental deaths or removals)
- "m_{rt} = number of marked animals caught in sample t last caught in sample r"
- "R_t = number of the s_t individuals released at sample t and caught again in some later sample"
- "Z_t = number of individuals marked before sample t, not caught in sample t, but caught in some sample after sample t"

Given a number of samples taken at different times, the proportion of animals marked is estimated as:

$$\hat{\alpha}_t = \frac{m_t + 1}{n_t + 1} \tag{5.18}$$

The size of the marked population is estimated as:

$$\hat{M}_t = \frac{(s_t + 1)Z_t}{R_t + 1} + m_t \tag{5.19}$$

Finally, the population size at time instant t is estimated as:

$$\hat{N}_t = \frac{\hat{M}_t}{\hat{\alpha}_t} \tag{5.20}$$

5.6.3 Integrating the Jolly-Seber method with the regression model

As explained in the previous section, the application of the Jolly-Seber method requires to know the number of marked animals caught in a sample s_t. In other words, given two samples s_1 and s_2 at time instants t_1 and t_2, we need to find the number of individuals that were present in both samples. However, using the regression model, we will not be able to reconstruct this information. In fact, given an images collection, the regression will return the estimated number of animals on the SD card of its photographer. Given this number, nothing can be said about the identifiers corresponding to the animals estimated to be on the SD card but not present on the social media platform. Therefore, we can get an estimate for the real size of the sample s_1 and s_2 but we don't have an estimate for their intersection. This situation is summarized in Figure 5.5. To conclude, we need to adapt the Jolly-Seber method to deal with the missing information.

With respect to Figure 5.5, let C_i be an images collection posted by photographer P_i on a social media platform:

$$s_i = \{\text{individuals in images collection } C_i\} \tag{5.21}$$

In the biological context, s_i represent a sample collected at time instant t_i. Let n_i be the cardinality of s_i:

$$n_i = |s_i| \tag{5.22}$$

Similarly, let SD_i be the set of pictures of the SD card of photographer P_i. We define S_i and N_i as:

$$S_i = \{\text{individuals in SD card } SD_i\} \tag{5.23}$$

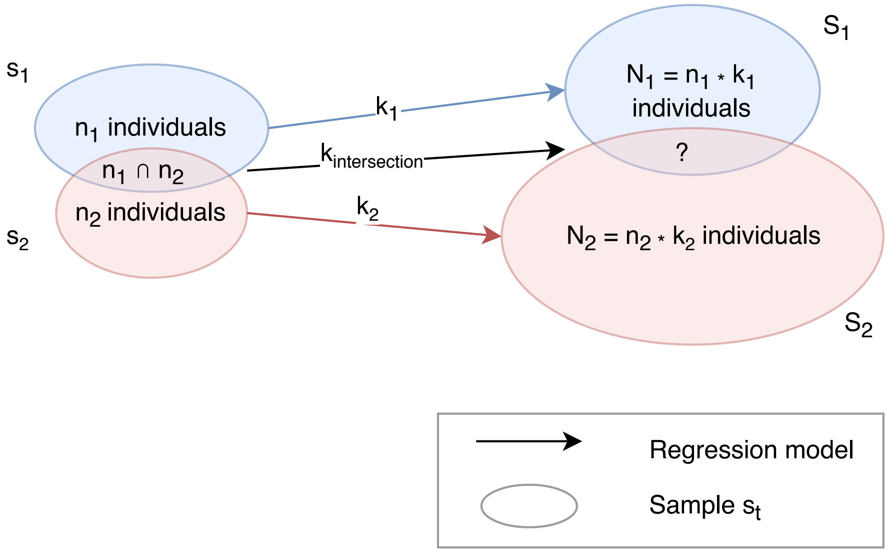


Figure 5.5: Using the regression model we will be able to estimate the size of S_1 and S_2 but not their single elements. Therefore, we won't be able to reconstruct directly $S_1 \cap S_2$.

$$N_i = |S_i| \tag{5.24}$$

Due to the presence of social media bias, n_i does not represent the real number of animals photographed by photographer P_i. The relation between individuals on a social media s_i and individuals on an SD card S_i is determined by the social media bias. Similarly, the relation between the number of individuals on social media n_i and the number of individuals on the SD card N_i is determined by the social bias. Using the regression model, we will not be able to reconstruct S_i from s_i, however, we can still make inference for N_i. In fact, given an images collection C_i, the real effect of the social media bias can be represented, by a coefficient k_i:

$$N_i = k_i n_i \tag{5.25}$$

Using the regression model described in Section 5.5, we get an estimate $\hat{k}_i$ for k_i as:

$$k_i = \frac{1}{Percentage\ Bias} \tag{5.26}$$

where *Percentage Bias* is the target value of the regression model. Let us consider two samples s_i and s_j taken at two different time instant t_i, t_j with $t_i \neq t_j$:

$$n_{s_i \cap s_j} = |s_i \cap s_j| \tag{5.27}$$

$$N_{s_i \cap s_j} = |S_i \cap S_j| \tag{5.28}$$

While n_i, n_j, and $n_{s_i \cap s_j}$ are known, using the regression model we can provide an estimate for N_i, N_j, but not for $N_{s_i \cap s_j}$. Our solution consists of providing an estimate $\hat{k}_{s_i \cap s_j}$ for $k_{s_i \cap s_j}$ using the weighted mean values of the coefficients over the number I_i and I_j of images in the images collections corresponding to s_i and s_j:

$$k_{s_i \cap s_j} = \frac{k_i * I_i + k_j * I_j}{I_i + I_j} \tag{5.29}$$

5.6.4 Extending the integration to more than two images collections

In the previous section, we provide a solution for estimating $N_{s_i \cap s_j}$ for two samples s_i and s_j when each of the samples corresponds to only a single images collection. However, when estimating the number of animals using social media images, it is often the case that each sample is composed of more than one images collection.

Given the time period for which we want to study a species, we retrieve from a social media platform all the images collections corresponding to that period. Then, we group albums by years. Formally, we create a partition of the set of the retrieved albums, such that in each subset all the albums contain pictures from the same year, and there are no different subsets containing albums from the same year. Since it may happen that an images collection contains images from two different years, we associate a time instant t_i to every images collection C_i and to its corresponding sample s_i :

$$t_i = \frac{t_{max} - t_{min}}{2} \tag{5.30}$$

where t_{max} and t_{min} are the maximum and minimum time instant over the set of time instants associated with the pictures of images collection C_i.

In this context, a biological sample s_{year_i} corresponds to the set of images collections from the same years.

$$s_{year_p} = \{s_i | t_i \in year_p\} \tag{5.31}$$

With respect to the situation described in the previous section, the only difference is that instead of having a single images collection for each time instant, now we have multiple images collections for each time instant, as shown in Figure 5.6 .

In this situation, the intersection between two samples, which we have previously indicated as $s_i \cap s_j$ cannot be computed as the intersection between two single images collection. Instead, given two samples s_{year_p} and s_{year_q}, their intersection $s_{year_p \cap year_q}$ can be computed as the union of all the intersections $s_i \cap s_j$ for all $s_i \in s_{year_p}$ and $s_j \in s_{year_q}$:

$$s_{year_p \cap year_q} = \bigcup_{\forall (i,j) \in \{(i,j) | t_i \in year_p \wedge t_j \in year_q\}} s_i \cap s_j \tag{5.32}$$

The intersection $s_{year_p \cap year_q}$ is the one we can obtain by analyzing the images taken from social media. However, we need to find a suitable way to estimate the real intersection $S_{year_p \cap year_q}$ representing the intersection

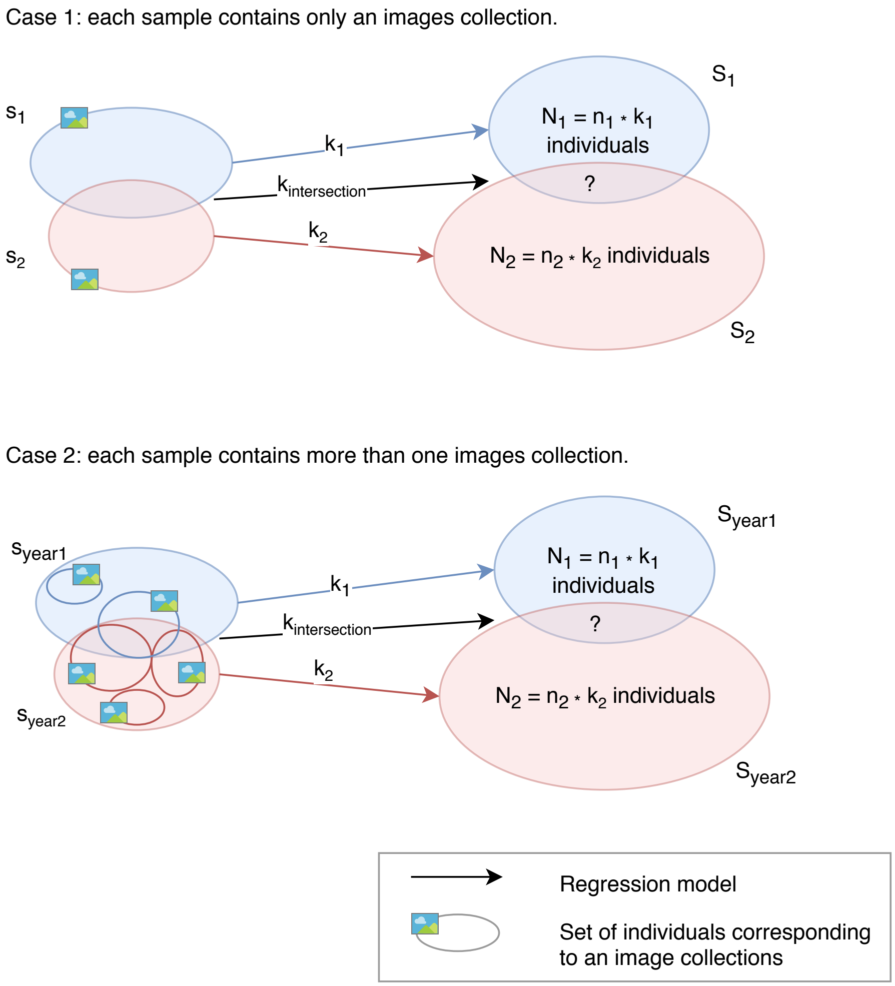


Figure 5.6: On the top, case 1 represents the trivial case where each sample corresponds to a single images collection. The intersection $S_1 \cap S_2$ is the estimated intersection of the two sets of images present on the SD cards SD_1 and SD_2. On the bottom, case 2 represents a more realistic situation where, for each year, we have a sample composed of more than one images collection. The intersection $S_{year_1} \cap S_{year_2}$ is the estimated intersection of more than two SD cards.

between the sets of individuals on the SD cards of photographers. In other words, we need to take into account the social media bias using our regression model.

Given the two set of images collections, corresponding to $year_p$ and $year_q$, we can estimate the number of recaptured animals by extending the trivial case showed in the previous section.

Given two years $year_p$ and $year_q$, we apply our regression model to each of the images collections corresponding to these years. We will then have a list of coefficient k_i of the same types of the one seen in the previous section. However, as in the trivial case we needed to estimate $k_{s_i \cap s_j}$, now we need to get an estimate for $k_{year_p \cap year_q}$. Again we propose to use the weighted average value over the number I_i of images in an images collections:

$$k_{year_p} = \frac{\sum_{i \in year_p} k_i * I_i}{\sum_{i \in year_p} I_i} \tag{5.33}$$

$$k_{year_1 \cap year_2} = \frac{\sum_{i \in year_1 \cap year_2} k_i * I_i}{\sum_{i \in year_1 \cap year_2} I_i} \tag{5.34}$$

5.6.5 Bias related to the lack of bilateral symmetry in animals

The proposed model relies on the use of computer vision tools to identify the same individual in multiple photos. Since not all animal species exhibit bilateral symmetry [39], it may be the case that an individual cannot be identified in every picture. For instance, the same zebra may appear in two different pictures. In one picture it may show its left side while in the other picture it may show its right side. If the patterns of the two sides of the zebras are not similar enough, computer vision algorithms will not be able to identify the individual in the two pictures. This will lead to underestimating the number of recaptured animals, with the consequence of overestimating the number of animals of the species. In this chapter we suggest how to deal with this issue.

Our solution is based on computing the probability of knowing the existence of a recapture given the existence of the recapture itself. These two probabilities may not coincide because of the lack of symmetry, as explained before. Suppose an animal can be modeled as a polyhedron with n sides $side_i$. For instance, such sides may correspond to right and left sides of an animal. Let us define:

$$N = \{i | 1 \leq i \leq n\} \tag{5.35}$$

$$P(side_{i-t_k}) = P(\text{animal shows } side_i \text{ at time instance } t_k) \tag{5.36}$$

The probability of knowing the existence of recapture of an animal actually captured at both t_1 and t_2 is the probability that the animal shows the same side in both time instants:

$$P(\text{knowing } \exists \text{ recapture} | \exists \text{ recapture}) = \sum_{i \in N} P(side_{i-t_{k_1}}) * P(side_{i-t_{k_2}}) \tag{5.37}$$

To conclude, when solving the formula for a given species, it must be remembered that it is likely that an animal shows more than one side in a single picture. For instance, this is the case of Grevy's zebras, as explained in Section 6.7

Once the probability has been computed, we can provide an estimate $\hat{R}$ for the real number of recaptures R given the number of known recaptures R_e:

$$\hat{R} = \frac{R_e}{P(\text{knowing } \exists \text{ recapture} | \exists \text{ recapture})} \tag{5.38}$$

Experimental Setup 6

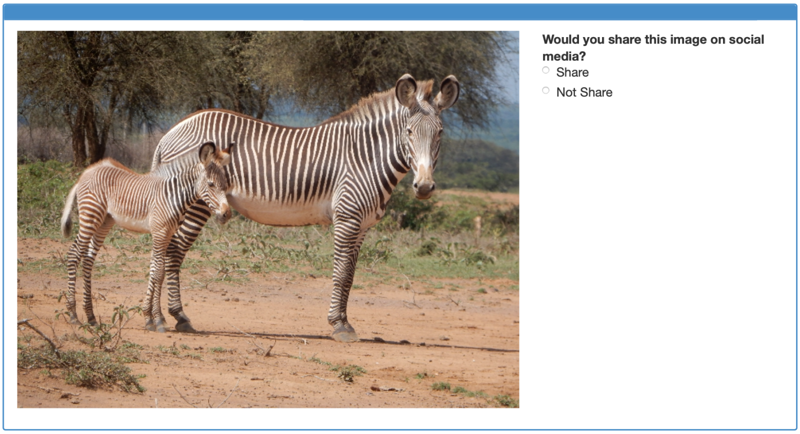
In this section, we will describe the implementation of the framework proposed in Chapter 5. Given the abundance of information coming from GGR and GGR2 and the power of Wildbook ^{TM}, we select Grevy's zebra and Reticulated giraffe as a specie to test our framework.

6.1 Synthesizing MTurk albums for machine learning training

In order to create a suitable machine learning model, we created a number of questionnaires to ask people which images they would like to shares on social media.

We took all the real SD cards of photographers from GGR and GGR2 and we created a questionnaire for each of the SD card. A questionnaire contains all the images from an SD card. Every questionnaire has been answered by one MTurk workers, with a few exceptions explained in Section 6.3.

To prevent users from selecting random images, we forced them to fill a checkbox for each image, as shown in 6.1. At the end of each questionnaire, workers had the opportunity to review the images they would have shared on a social media, as they could have done if they were uploading images on a real social media. A screen of this review process is shown in 6.2. To conclude, we synthesized 356 albums corresponding to 142 albums from GGR and 214 albums from GGR2. Some statistics for these albums are shown in Figure 6.3 and Figure 6.4.

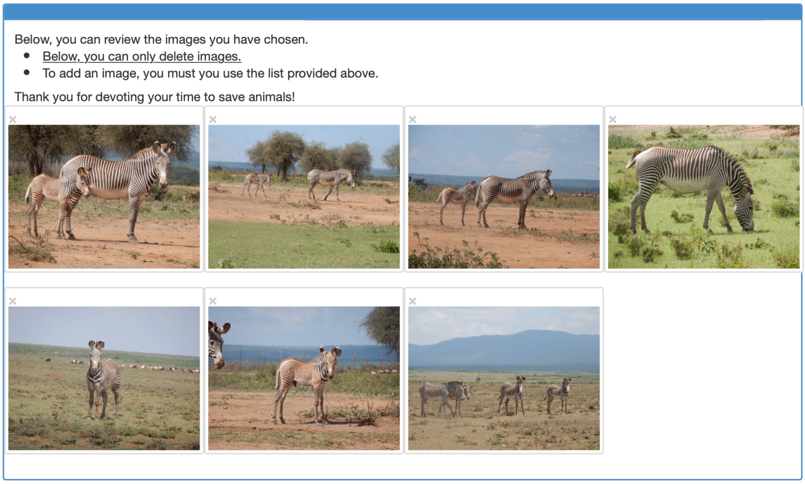


Would you share this image on social media?

○ Share

○ Not Share

Figure 6.1: Questionnaires are composed of multiple questions of this kind.



Below, you can review the images you have chosen.

- Below, you can only delete images.
- To add an image, you must you use the list provided above.

Thank you for devoting your time to save animals!

Figure 6.2: At the end of questionnaire users could review the pictures selected.

6.2 Synthesizing MTurk albums for hypothesis testing

A number of albums has been created to test some hypothesis, as explained in Section 5.3.3.

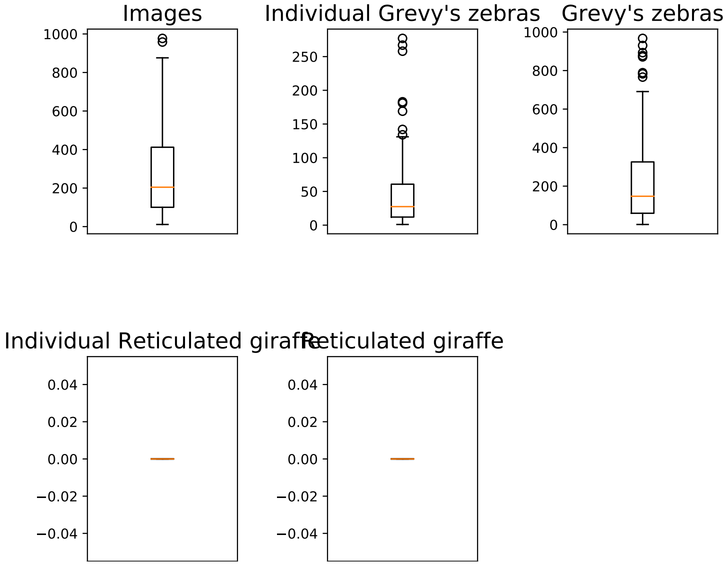


Figure 6.3: Box plots for the number of images, number of individual Grevy's zebras, number of Grevy's zebras, number of individual Reticulated giraffes and number of Reticulated giraffes for the SD cards of GGR. Note: no Reticulated giraffe has been photographed during the GGR.

6.2.1 Number of pictures in the SD cards

This experiment is designed to test whether the number of images in an SD card affects the percentage of images shared by a user. We generated 20 albums for each of the following album sizes:

$$values = \{25, 75, 150\} \tag{6.1}$$

We had each of the album labeled by an MTurk user. Then, for each value, we constructed an array whose elements are the percentage of images shared in each album. Each array is considered a sample in our experiment.

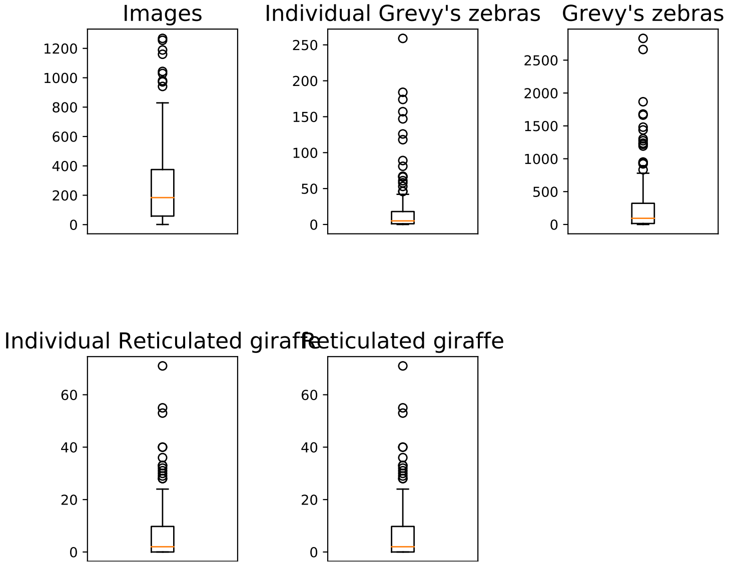


Figure 6.4: Box plots for the number of images, number of individual Grevy's zebras, number of Grevy's zebras, number of individual Reticulated giraffes and number of Reticulated giraffes for the SD cards of GGR2.

6.2.2 Number of animals in the SD cards

This experiment is designed to test whether the number of animals in an SD card affects the percentage of images shared by a user. We generated albums with the constraint of having 25 images each, and a given number of animals chosen from the following values:

$$values = \{25, 50, 100, 200\} \tag{6.2}$$

We had each of the album labeled by an MTurk user. Then, for each value, we constructed an array whose elements are the percentage of images shared in each album. Each array is considered a sample in our experiment.

6.2.3 Side of the animals visible in pictures

This experiment is designed to test whether the side shown by an animal in a picture affects the percentage of images shared by a user. We generated albums having images with the constraint of showing just one animal in each image. Given an album, all the animals have the constraint of showing the same side. The sides considered are the ones detected by Wildbook $^{\mathrm{TM}}$:

$$\begin{aligned} values = \{&left, frontright, right, backright, frontleft, backleft,\\ &front, up, back, down\} \end{aligned} \tag{6.3}$$

We had each of the album labeled by an MTurk user. Then, for each value, we constructed an array whose elements are the percentage of images shared in each album. Each array is considered a sample in our experiment.

6.2.4 Shareability of individuals

This experiment is designed to test whether certain individuals are more likely to be shared by a user on a social media. We generated albums having images with the constraint of showing just one animal in each image. Moreover, each album contains, in the same proportion, pictures of only two individual. In other words, half of the photos shows an individual and the other half another individual. We selected 11 individuals of Grevy's zebra. The criteria for the selection has been the number of pictures available for these individuals: we selected the ones with the biggest number. For each individual, we generated two albums for each of remaining 10 individuals, for an overall number of 110 albums. When selecting pictures of the same of individual, we avoided selecting pictures taken in short time intervals since they are likely to be very similar pictures. We had each of the albums labeled by an MTurk user. Then, for each individual, we constructed an array whose elements are the percentage of images shared of this individual over the overall number of images shared in each album in which the individual appears (i.e. both of this individual and the other one present in the album). Each array is considered a sample in our experiment.

6.2.5 Testing MTurk surveys

To ensure the quality of the data, the surveys have been tested on the following browsers:

- Internet Explorer - Version 11.345.17134.0 - Update Version 11.0.90 (KB4462949)
- Edge - Microsoft Edge 42.17134.1.0 - Microsoft EdgeHTML 17.17134
- Firefox Quantum - 62.0.3 (64-bit)
- Opera - Version:56.0.3051.43
- Safari - Version 12.0 (14606.1.36.1.9)
- Chrome - Version 69.0.3497.100 (Official Build) (64-bit)
- Chrome Android - Version 69.9.3497.100

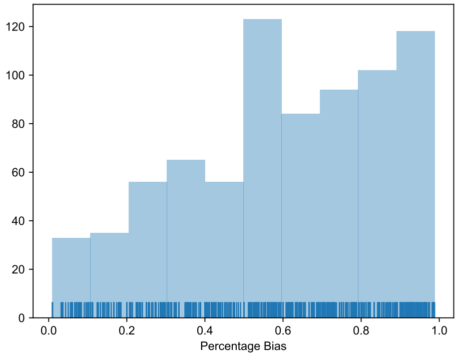
Moreover, we tested the code in a virtual environment with a single core and 2 Gigabytes of RAM to be sure that even low-end pc could manage the number of images provided in a questionnaire.

6.3 Quality of data collected

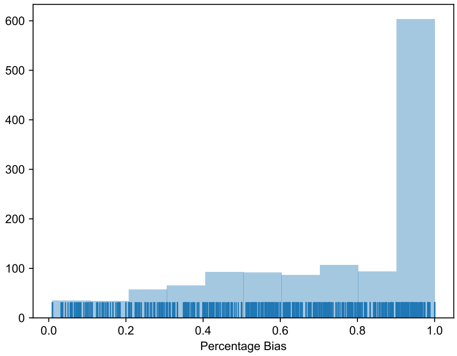
As explained in Section 5.2.3, there is no evidence of the presence of automatic bots on MTurk. Instead, the quality of the data collected could be affected by the presence of "lazy" workers who don't complete their tasks well. During the deployment of the first surveys, we realized that we were possibly facing this issue. The first symptom was the struggle for machine learning models to find patterns in the data. This hypothesis was confirmed by the lack of differences in the distribution of the extracted features, and in particular of the beauty features described in Section 5.3.1.2. Further analysis revealed the presence of several workers that were completing HITs very fast, choosing to share all the images proposed. Since these workers took less than a second per image, we decided to delete the results from these workers. In fact, they have actually spent even less time per image since they were also asked to read an introduction to the task, and to answer a few questions regarding their age, their country of provenience and their behavior on social networks. This issue was mainly solved by choosing to interview only "master" workers who are certified by Amazon to have a high percentage of history of tasks accepted by requesters.

In particular, we asked some questions to gather information about the data collection process and the quality of it. These questions are not intended to be evaluated quantitatively but rather qualitatively, to get some insights about the data collection process. In fact, these answers have not been used to train machine learning models. These questions were placed at the beginning of the questionnaire and they are shown in Figure 6.7.

One of the questions was "*Do you consider yourself an active user on social media?*" with multiple choice answer "Yes/No". Among the interviewees, 89.30% answer "yes" as shown in Figure 6.8. Even if the question was subjective, we can reasonably suppose that almost all the interviewees use social media. This hypothesis was confirmed by the answers to the question "*How often do you share images on social media?*" whose results are displayed in Figure 6.9. To understand if people

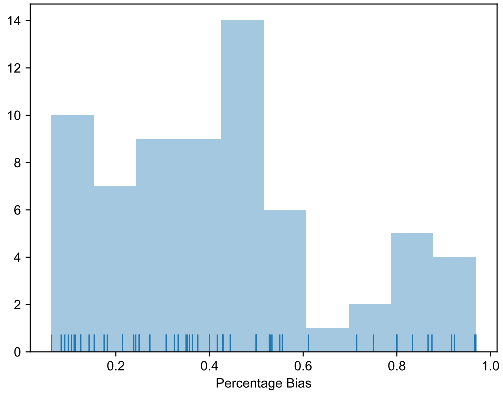


(a) Target values for Grevy's zebra when discarding target values equal to 1

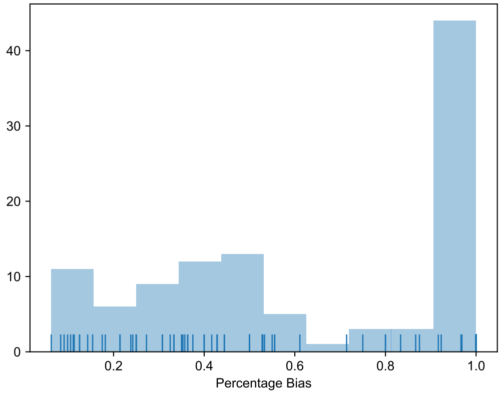


(b) Target values for Grevy's zebra

Figure 6.5: Regression target values for Grevy's zebra. These numbers include also the questionnaires developed for the MTurk experiments to test features importance.

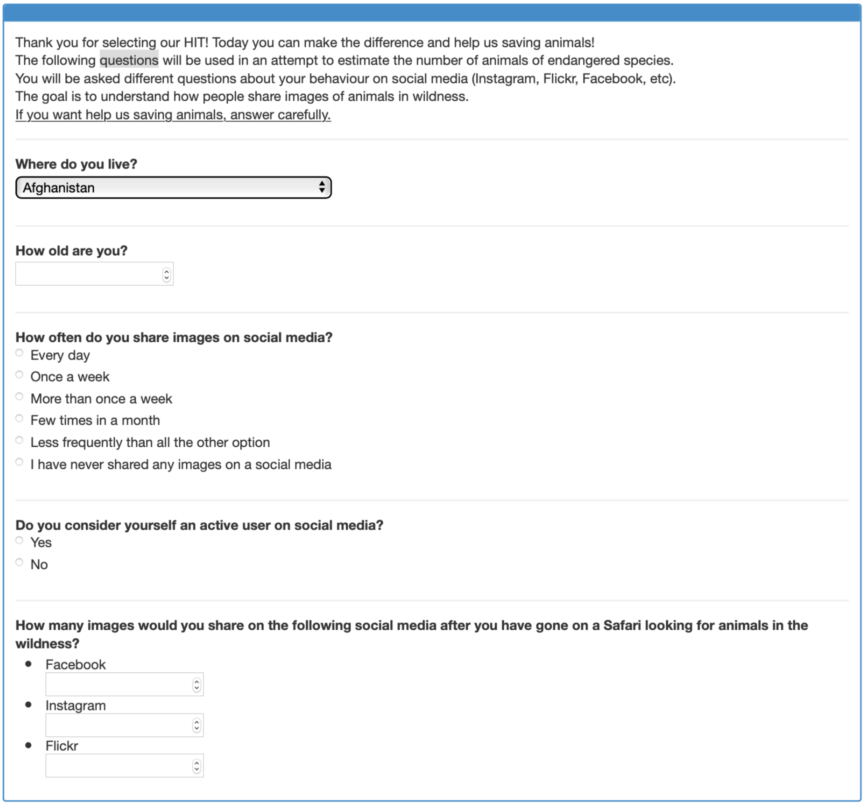


(a) Target values for Reticulated giraffe



(b) Target values for Reticulated giraffes when discarding target values equal to 1

Figure 6.6: Regression target values for Reticulated giraffe.



Thank you for selecting our HIT! Today you can make the difference and help us saving animals!
The following questions will be used in an attempt to estimate the number of animals of endangered species.
You will be asked different questions about your behaviour on social media (Instagram, Flickr, Facebook, etc).
The goal is to understand how people share images of animals in wildness.
If you want help us saving animals, answer carefully.

Where do you live?

Afghanistan

How old are you?

How often do you share images on social media?

○ Every day
○ Once a week
○ More than once a week
○ Few times in a month
○ Less frequently than all the other option
○ I have never shared any images on a social media

Do you consider yourself an active user on social media?

○ Yes
○ No

How many images would you share on the following social media after you have gone on a Safari looking for animals in the wildness?

- Facebook
- Instagram
- Flickr

Figure 6.7: Questions asked to each worker at the beginning of the questionnaire. These questions were intended to gather qualitative data about the quality of the data collection process and the population of interviewees.

were answering at random, we asked them "*How many images would you share on the following social media after you have gone on a safari looking for animals in the wildness?*". The question was asked for *Facebook*, *Instagram*, and *Flickr* ^{TM}. The computation of the Pearson correlations between the answers to these questions and the number of pictures actually chosen to be shared is shown in Table 6.1. These numbers reveal the existence of a correlation. This correlation cannot be considered strong, but this was not expected given the format of the question. In fact,

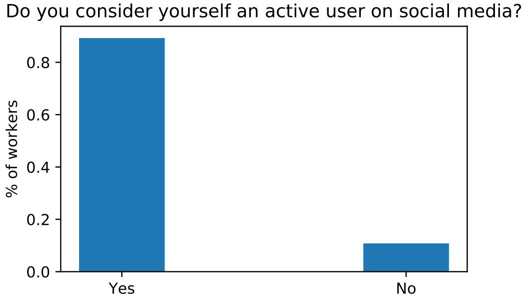


Figure 6.8: Answers to "Do you consider yourself an active user on social media?"

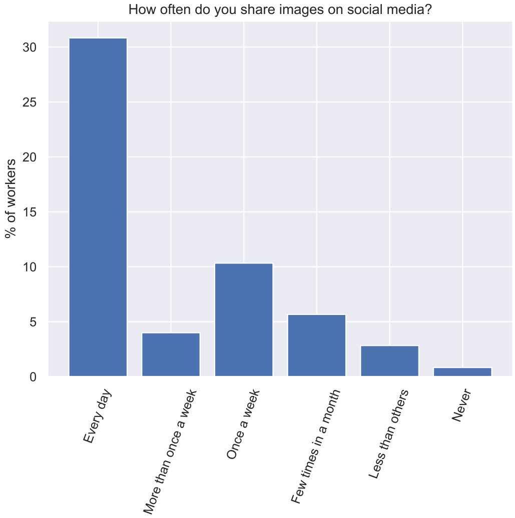


Figure 6.9: Answers to "How often do you share pictures on social media?"

it is reasonable to suppose that there is not a fixed number of photos that a person would share regardless of the photos themselves. Instead, the presence of a correlation, even though not extremely strong, proves the existence of some consistency in workers answers, showing that it is reasonable to suppose that workers were not answering randomly.

Table 6.1: PEARSON CORRELATIONS COMPUTED BETWEEN THE NUMBER OF PICTURES CHOSEN TO BE SHARED BY PEOPLE AND THE NUMBER OF PICTURES PEOPLE SAID THEY WOULD SHARE ON *FACEBOOK*, *INSTAGRAM*, AND *Flickr* TM WHEN ANSWERING TO THE QUESTION "*HOW MANY IMAGES WOULD YOU SHARE ON THE FOLLOWING SOCIAL MEDIA AFTER YOU GONE ON A SAFARE LOOKING FOR ANIMALS IN THE WILDNESS?*"

Social Media	Pearson Correlation
Facebook	0.37
Instagram	0.45
Flickr TM	0.32

6.3.1 Demographics of interviewed workers on MTurk

We asked workers two questions to collect information about the population of people we were interviewing. Results are shown in Figure 6.10 and Figure 6.11. These data tell us that the population of interviewees does not completely correspond to the population social media users. This may not be a problem but we suggest to investigate further how the shareability of images is affected by the different cultures. However,

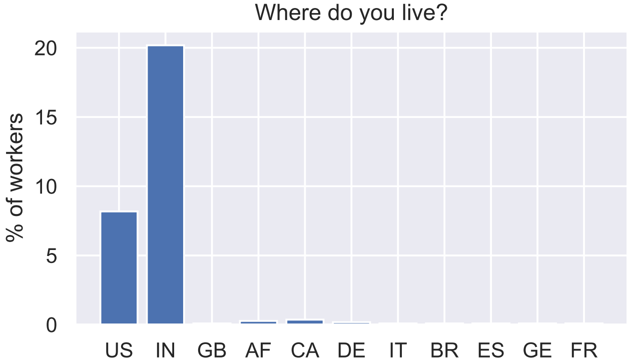


Figure 6.10: Countries where interviewed workers live.

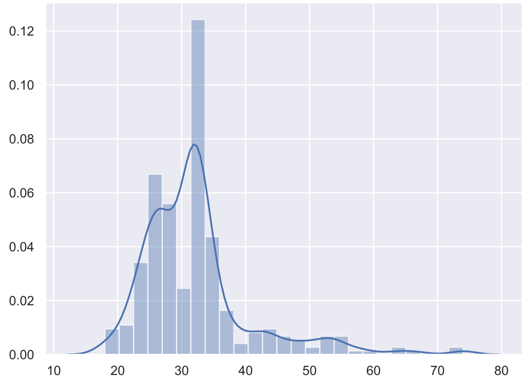


Figure 6.11: Distribution of the ages of the workers interviewed.

the population of interviewees seems to reflect the population of Amazon MTurk [10], giving another proof in supporting the hypothesis that people seem to have answered properly the questionnaire.

6.4 Classification models per identifying captive animals

When solving the problem of identifying images collections with captive animals, we chose as baselines the majority class. We tested different models, even though we expected the XGBoost classifier to outperform the others, as its strength has been proved by the number of competitions where winners used this model [33]. Moreover, having a huge number of features, we expect models capable of performing features selection, such as tree-based models, to perform well. Neural networks are another popular category of classifier used by many winners of Kaggle competitions, but given the size of the dataset we don't expect them to be able to generalize the problem as well as a gradient boosting algorithm. We also believe that the size of the dataset does not require to use any deep-learning techniques.

The models we tested are:

- Majority Class (Baseline)
- Gaussian Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- XGBoost Classifier

6.4.1 Additional features: object detection

When downloading images from social media there are very limited assumption on the content of the images retrieved. When scraping Flickr ^{TM}, initially we searched for pictures of the desired species. In that situation, most of the pictures were coherent with our search. However, after we downloaded the entire images collections of every photo retrieved, we found ourselves with photos of any kind. For this reason, we wanted to use a general purpose library for object identification.

For our test cases, we decided to use ImageAI [14], probably the most popular Python library for object detection. ImageAI provides trained models that support RetinaNet, YOLOv3 and TinyYOLOv3. We used the RetinaNet model because it is considered the most accurate one according to the ImageAI documentation. The version we have used is *resnet50_coco_best_v2.0.1.h5.*

The RetinaNet model is trained to detect 80 different objects. The full list is provided in Table 6.2.

6.5 Regression models

We chose as baselines two dummy regressors with the following strategies:

Table 6.2: ImageAI RESNET OBJECTS DETECTED.

Category	**Objects**
Animals	bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe
Food	banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, bottle, wine glass, cup, fork, knife, spoon, bowl
Means of transport	bicycle, car, motorcycle, airplane, bus, train, truck, boat
Urban elements	traffic light, fire hydrant, stop sign, parking meter, bench
Indoor objects	dining table, toilet, tv, microwave, oven, toaster, sink, refrigerator, couch, bed, hair dryer, toothbrush
Sport	frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket
Technology	laptop, mouse, remote, keyboard, cell phone
Others	person, chair, potted plant, book, clock, vase, scissors, teddy bear, backpack, umbrella, handbag, tie, suitcase,

- Always predicts the mean of the training set
- Always predicts the median of the training set

Given the large number of features we had, we expected trees ensembles to perform well, given their ability to perform features selection. Therefore we tried:

- XGBoost Regressor
- AdaBoost Regressor
- Bagging Regressor

- ExtraTrees Regressor
- Random Forest Regressor

Given their susceptibility to normalization, we apply it before training the following models:

- Lasso
- Elastic Net
- Support Vector Regressor

Finally, for the following models, we perform not only normalization but also Principal Component Analysis:

- Ridge
- Bayesian Ridge

6.6 Evaluating the performance of the models

Since the datasets available are not composed of many samples, we evaluated our models using a repeated N-fold cross-validation. In our implementation, we adopted a 10 times 10-fold cross-validation. This has been done both for the classification problem and regression problem.

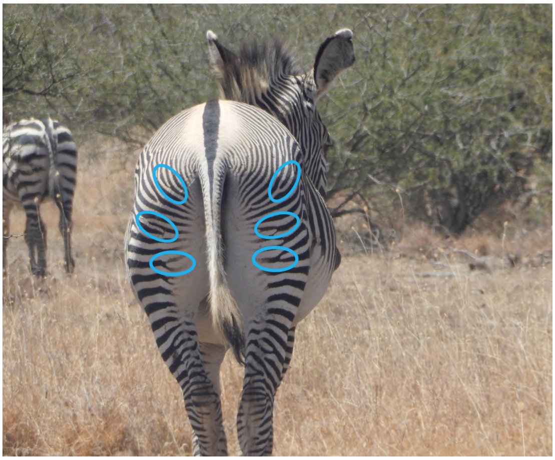
6.7 Correction for bias related to lack of bilateral symmetry

As better explained in Section 5.6.5, some animal species, such as Grevy's zebra and Reticulated giraffes, may lack of bilateral symmetry. An example of this situation is shown in Figure 6.12. Therefore, since the right and left sides of an animal may be different, computer vision algorithms, as well as biologists, may fail in identifying the same animal in multiple pictures when the animal shows different sides. This will lead to an underestimation of the number of recaptured animals. As a consequence, capture-mark-recapture methods, such the one used in this thesis, will overestimate the number of animals of the species studied. This problem may arise not only for the left and right sides of the animals but also for its front and back sides.

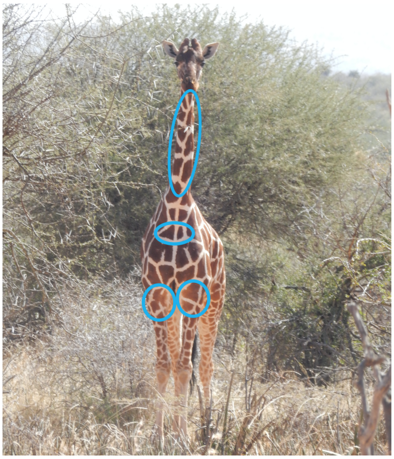
In Section 5.6.5, we suggest a viable approach to deal with this issue. In the case of Grevy's zebra and Reticulated giraffe, the Wildbook ^{TM} platform is capable of identifying the side that an animal shows in a picture. The side identified will be one among:

$$\begin{split} side \in \{&left, frontright, right, backright, frontleft, backleft, \\ &front, up, back, down\} \end{split} \tag{6.4}$$

According to the theory presented in Section 5.6.5, when calculating the number of recaptures between two samples, the probability we need to compute is:



(a) Grevy's zebra



(b) Reticulated giraffe

Figure 6.12: The pictures above highlight some asymmetry in animals that can lead to a missed recapture. In other words, the animal might be recognized as another one if seen from two different sides.

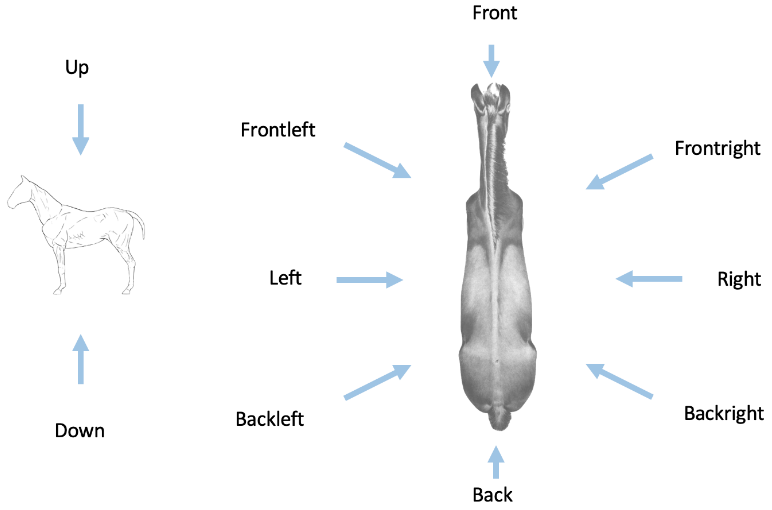


Figure 6.13: Sides detected by Wildbook $^{\mathrm{TM}}$ in the case of Grevy's zebra and Reticulated giraffe.

$$
\begin{aligned}
&P(\text{knowing } \exists \text{ recapture} | \ \exists \text{ recapture}) = \\
&\quad P(left_1) && * \ \Big(P(left_2) + P(backleft_2) + P(frontleft_2)\Big) + \\
&\quad P(right_1) && * \ \Big(P(right_2) + P(backright_2) + P(frontright_2)\Big) + \\
&\quad P(frontleft_1) && * \ \Big(P(left_2) + P(backleft_2) + P(frontleft_2) + \\
&&& \qquad P(front_2) + P(frontright_2)\Big) + \\
&\quad P(frontright_1) && * \ \Big(P(right_2) + P(backright_2) + \\
&&& \qquad P(frontright_2) + P(front_2) + P(frontleft_2)\Big) + \\
&\quad P(backleft_1) && * \ \Big(P(left_2) + P(backleft_2) + P(frontleft_2) + \\
&&& \qquad P(back_2) + P(backright_2)\Big) + \\
&\quad P(backright_1) && * \ \Big(P(right_2) + P(backright_2) + P(frontright_2) + \\
&&& \qquad P(back_2) + P(backleft_2)\Big) + \\
&\quad P(front_1) && * \ \Big(P(front_2) + P(frontleft_2) + P(frontright_2)\Big) + \\
&\quad P(back_1) && * \ \Big(P(back_2) + P(backleft_2) + P(backright_2)\Big) + \\
&\quad P(up_1) && * \ \Big(P(up_2)\Big) + \\
&\quad P(down_1) && * \ \Big(P(down_2)\Big)
\end{aligned}
\tag{6.5}
$$

Results 7

In this section, we provide results for the regression models and estimates for the population of Grevy's zebra and Reticulated giraffe.

7.1 MTurk experiments

In this section, we present the results of the experiments on Grevy's zebra shareability as described in Section 6.7. For each of the experiment we provide a box plot with samples group by values of the considered factor, and the p-value when testing the hypothesis that the samples have the same mean.

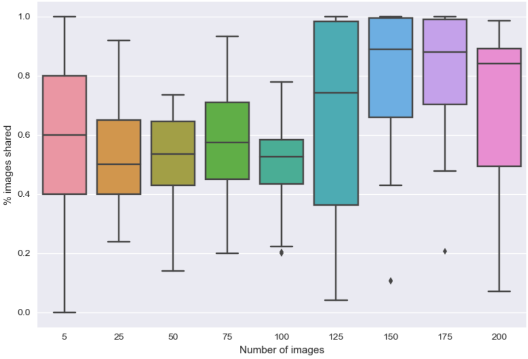


Figure 7.1: Box plots representing the samples relative to the experiment to test whether the number of pictures in the SD cards affects the percentage of images shared. A one-way ANOVA test, testing the null hypothesis that the means are equal, returns a p-value of 0,152.

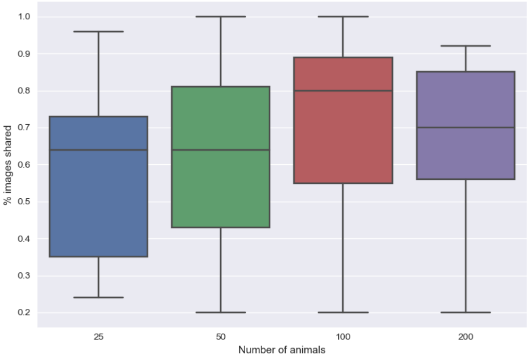


Figure 7.2: Box plots representing the samples relative to the experiment to test whether the number of animals in the SD cards affects the percentage of images shared. A one-way ANOVA test, testing the null hypothesis that the means are equal, returns a p-value of 0.330.

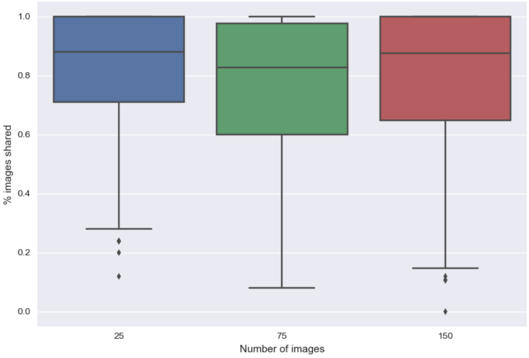


Figure 7.3: Box plots representing the samples relative to the experiment to test whether the side shown by animals in pictures affects the percentage of images shared. A one-way ANOVA test, testing the null hypothesis that the means are equal, returns a p-value of 0.100.

7.2 Classification model for captive animals

In this section, we will discuss the performance the classifier models listed in Section 6.4.

As explained in Section 5.4.2, we aim at maximizing the recall metrics because an higher recall will mean a smaller probability to commit

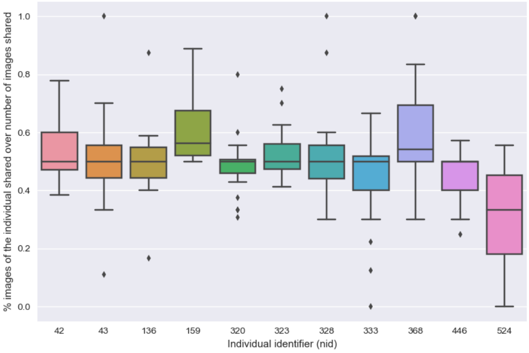


Figure 7.4: Box plots representing the samples relative to the experiment to test whether some individuals are more likely to be shared than other ones. A one-way ANOVA test, testing the null hypothesis that the means are equal, returns a p-value of $4.27 * 10^{-10}$.

an error that is likely to have an non-negligible impact on the species assessment. For further considerations, we suggest to read Section 5.4.2.

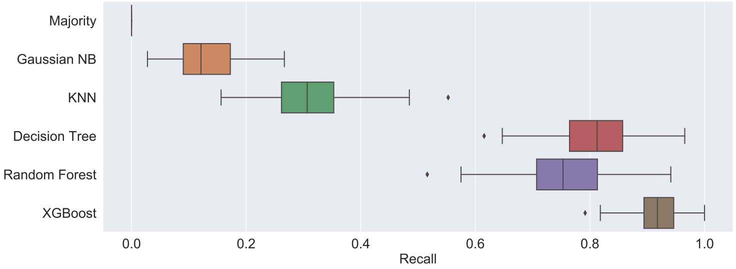


Figure 7.5: Box plots for the recall score evaluated over 10 times 10-fold cross-validation.

7.3 Regression model

In this section, we discuss the performance of the regression models mentioned in Section 6.5).

Figures 7.9 and 7.10 shows the box plots for R^2 and MSE measures.

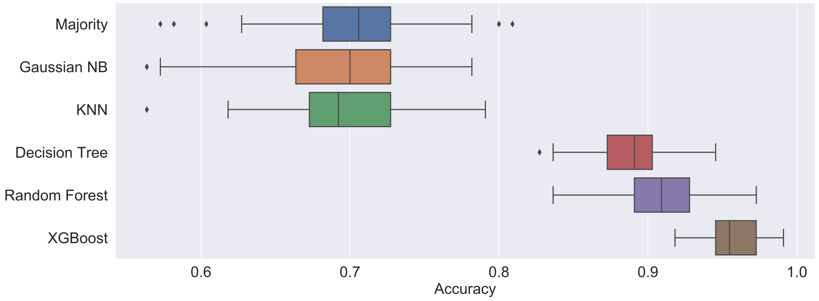


Figure 7.6: Box plots for the accuracy score evaluated over 10 times 10-fold cross-validation.

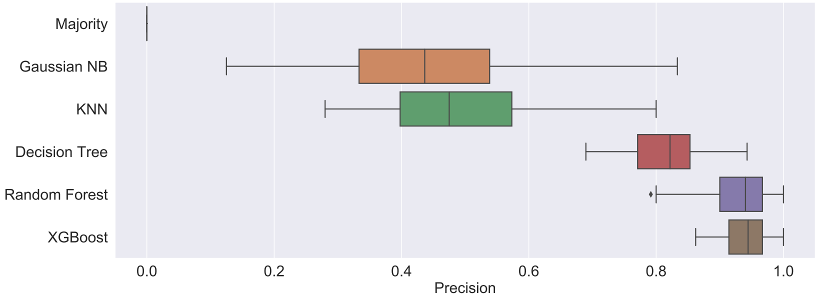


Figure 7.7: Box plots for the precision score evaluated over 10 times 10-fold cross-validation.

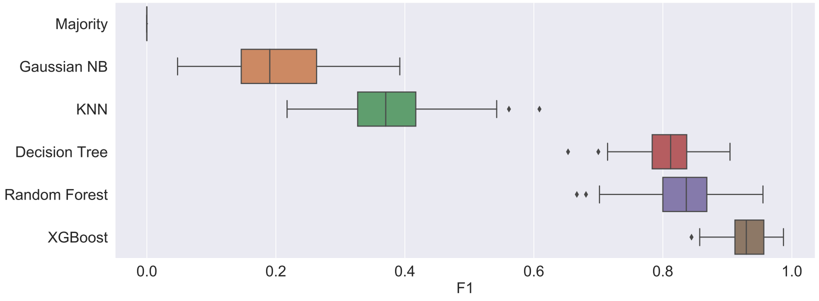


Figure 7.8: Box plots for the F-1 score evaluated over 10 times 10-fold cross-validation.

Even though results exhibit non-negligible variances, the implementations of the XGBoost regressor seems to overcome the other models in both R^2 and MSE for Grevy's zebra. On this basis, we decided to adopt it to provide estimates of both Grevy's zebra. According to Figure 7.10 and 7.11 it is difficult to identify the best model. We decided to use the Bayesian Ridge since it seems to perform better than other models, and it seems to have one of the smaller variances across R2 and MSE values. In addition, given the small size of the dataset of Reticulated giraffe, we believe it make sense to chose a simple model.

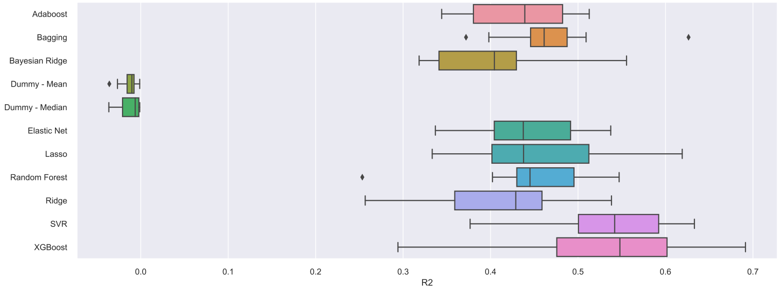


Figure 7.9: Grevy's Zebra - Scatter plots for R2 values evaluated over 10 times 10-fold cross-validation.

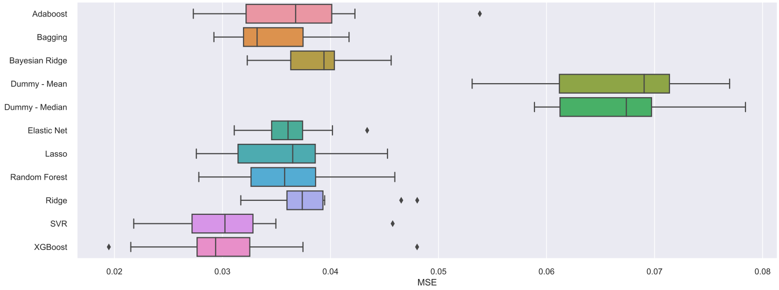


Figure 7.10: Grevy's Zebra -Scatter plots for MSE values evaluated over 10 times 10-fold cross-validation.

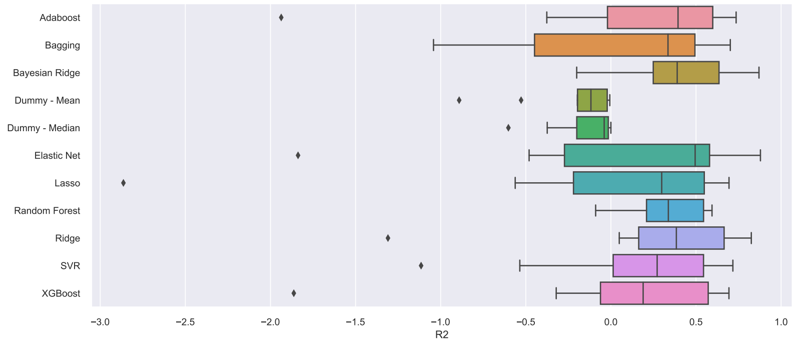


Figure 7.11: Reticulated giraffe - Scatter plots for MSE values evaluated over 10 times 10-fold cross-validation.

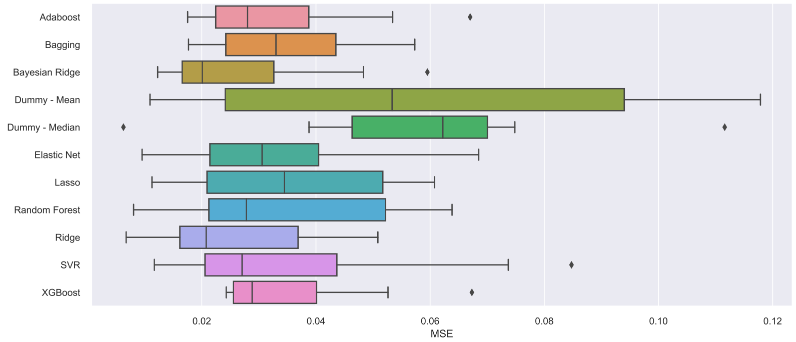


Figure 7.12: Reticulated giraffe - Scatter plots for MSE values evaluated over 10 times 10-fold cross-validation.

7.4 Population size estimates

Given the performance exhibited by the XGBoost regressor for Grevy's Zebra and the Bayesian Ridge for the Reticulated giraffe, we chose them to make the final estimate for the Grevy's zebra and Reticulated giraffe population. Therefore, we trained a model for each species on the dataset generated from the MTurk surveys. Then, we applied it to the images collections retrieved by Flickr ^{TM} to predict the *Percentage Bias* as explained in Section 5.5. Using the *Percentage Bias* we than estimate the number of real zebras and giraffes photographed by each Flickr ^{TM} user. Than, using our adapted Jolly-Seber method, we estimate the number of Grevy's zebras and Reticulated giraffes for each year from 2011 to 2018. As explain in Section 6.7 and in Section 5.6.5 we need a method to deal with the lack of bilateral symmetry of these species. In brief, since zebras and giraffes may be photographed from different sides, computer vision algorithms may fail in recognizing the same individual in multiple photos. As a consequence, we will underestimate the number of recaptured animals, and, therefore, we will overestimate the number of animals of the species.

7.4.1 Results

In Table 7.1 and Table 7.2 we propose different results. First, we show official estimates from IUCN [18] [15], GGR [4], GGR2 [35] and Kenya Wildlife Services [46] data. The estimates are not always available. Moreover, for certain years, only lower or upper boundary is available. In the tables, they are indicated respectively with symbol "+" and symbol "-". For each of the proposed results we computed the Root Mean Squared Error. However, given the lack of data for certain years, as well as the availability of the sole upper or lower boundaries for many others, the RMSE must be considered an approximation of the real RMSE. In addition, we must remember that also the official estimates are the output of statistical models and they are therefore subject to errors.

As a baseline for the results we provide the plain Jolly-Seber method without any of the theory developed in this thesis. In column "VP", we show estimates using the sole correction for the viewpoint bias, related to the bilateral asymmetry of animals. In column "ML", we show estimates using the sole correction for the social media bias (regression model). Finally, in column "VP + ML" we show results when using

Table 7.1: ESTIMATES FOR GREVY'S ZEBRA. COLUMN “OFFICAL" SHOWS THE ESTIMATE PROVIDED BY THE SCIENTIFIC COMMUNITY. COLUMN “BASELINE" SHOWS RESULTS COMPUTED USING THE TRADITIONAL JOLLY-SEBER METHODS. COLUMN “VP" SHOWS RESULTS COMPUTED USING JOLLY-SEBER WITH VIEWPOINT BIAS CORRECTION. COLUMN “ML" SHOWS RESULTS COMPUTED USING THE ADAPTED JOLLY-SEBER WITH OUR REGRESSION MODEL WITHOUT VIEWPOINT BIAS CORRECTION. COLUMN “VP + ML" SHOWS RESULTS COMPUTED USING THE ADAPTED JOLLY-SEBER WITH VIEWPOINT BIAS CORRECTION.

Year	**Official**	**Baseline**	**VP**	**ML**	**VP + ML**
2011	2827	1686	1270	1782	984
2012	1897+	1491	1298	1926	1005
2013	-	2360	1249	1933	1035
2014	-	1566	2613	3599	2150
2015	-	996	862	1235	755
2016	2250	93	1883	2296	1644
2017	1627+	6084	1927	2178	1806
2018	(2680)	42	76	64	82
Approximate RMSE	-	2570	878	591	1086
Viewpoint Correction			✓		✓
Regression Model				✓	✓

both corrections for social media bias and viewpoints.

7.4.2 Consideration on the results

As it can be seen from the tables, there is a systematic error for the estimates relative to the most recent year. Further experiments have shown that this issue always appears for the last year, even when cutting

Table 7.2: ESTIMATES FOR RETICULATED GIRAFFE. COLUMN “OFFICAL" SHOWS THE ESTIMATE PROVIDED BY THE SCIENTIFIC COMMUNITY. COLUMN “BASELINE" SHOWS RESULTS COMPUTED USING THE TRADITIONAL JOLLY-SEBER METHODS. COLUMN “VP" SHOWS RESULTS COMPUTED USING JOLLY-SEBER WITH VIEWPOINT BIAS CORRECTION. COLUMN “ML" SHOWS RESULTS COMPUTED USING THE ADAPTED JOLLY-SEBER WITH OUR REGRESSION MODEL WITHOUT VIEWPOINT BIAS CORRECTION. COLUMN “VP + ML" SHOWS RESULTS COMPUTED USING THE ADAPTED JOLLY-SEBER WITH VIEWPOINT BIAS CORRECTION.

Year	**Official**	**Baseline**	**VP**	**ML**	**VP + ML**
2011	5528+	9562	7537	10376	6255
2012	-	1931	2007	2682	1636
2013	6500-	8690	5435	8033	4312
2014	-	2935	2098	3033	1697
2015	8561+	9530	11110	14641	9305
2016	-	1651	1583	2286	1264
2017	-	6841	2703	3363	2343
2018	(15784)	140	202	198	204
Approximate RMSE	-	2708	1972	4576	1398
Viewpoint Correction			✓		✓
Regression Model				✓	✓

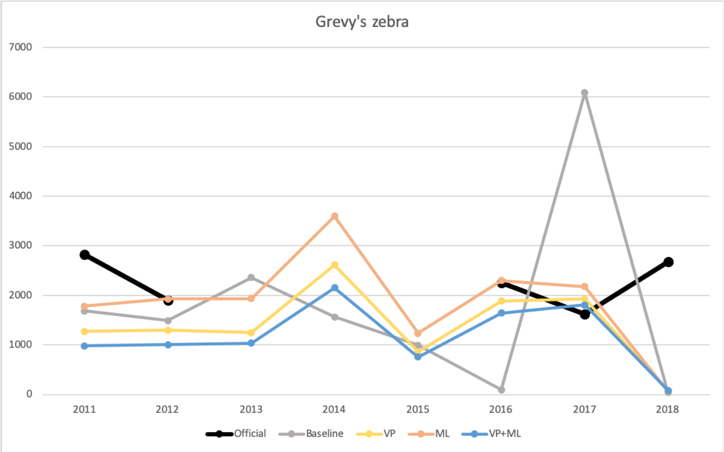


Figure 7.13: Estimates for Grevy's zebras computed using the proposed approaches on the Flickr TM dataset.

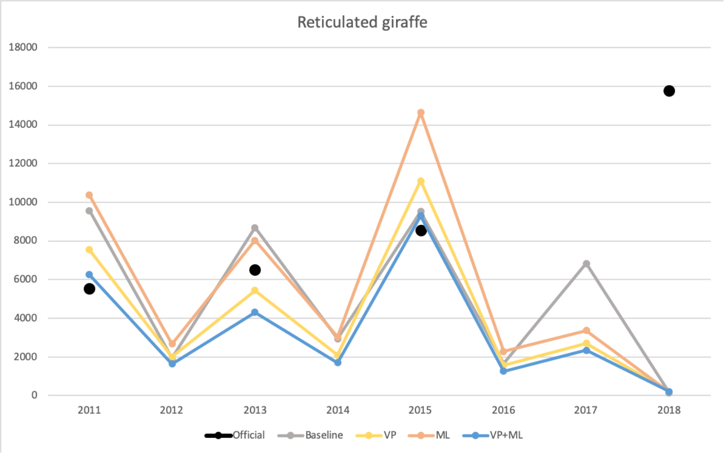


Figure 7.14: Estimates for Reticulated giraffes computed using the proposed approaches on the Flickr TM dataset.

the dataset to another time interval. In other words, given a year X, if we discard all the data related the interval of time that follows the year X, we will underestimate the number of animals for the year X. For this reason, we didn't include the data relative to the most recent year in the computation of the RMSE: the error would have been so big that it would have prevented the usefulness and interpretability of the RMSE, by prevailing over the errors relative to all the other years. We believe the issue to be intrinsic to the use of the Jolly-Seber methods which, when computing an estimate for a given year, uses both information relative to past years and the future years. Since in the case of the most recent year there are obviously no data relative to the future, the method might fail.

Finally, the results show that our framework was the best performing for the Reticulated giraffe species. In the case of Grevy's zebra, we can see that better results are achieved when correcting either the social media bias or viewpoint bias. In particular, the best estimate was achieved when correcting only the social media bias. However, the probability theory explained in Section 5.6.5 is against the theoretical correctness of this data. Indeed, this method was the worst one for Reticulated giraffe. With respect to the sole correction of the viewpoint bias, this method outperforms our framework in the case of Grevy's zebra but in the case of Reticulated giraffe. Since we don't have any official data for several years, and since the majority of the available official estimates represents either an upper or lower boundary, it is difficult to conclude which method is the best one.

Conclusion 8

In this thesis, we proposed a framework to assess the population of an animal species using social media images collections.

The framework is composed of two principal steps. The first step is the use of a regression model to estimate the number of animals photographed but not shared on social media. This model must be trained using a suitable dataset containing pictures of the species of interest. In Section 5.2.2 we suggest a viable way to construct such a dataset. The second part of the framework is the estimation of the number of animals of the given species using the predictions obtained through the regression model trained in the previous step. This estimate is computed using a Jolly-Seber method modified to take into account the viewpoint bias.

The results shown in Chapter 7 proved that the problem of predicting the number of animals photographed but not shared on social media, given an images collection shared, is a learnable problem.

Our research shows the presence of patterns when estimating the number of animals photographed by a user using the images shared on social media. Even though single factors, such as the ones analyzed with the MTurk experiments, are not sufficient to model this problem, we found more complicated features to be useful.

Finally, we proved that the task of identifying captive animals in images collections can be highly automatized.

In the following sections, we will suggest how the proposed approach could be improved and extended.

8.1 Future work

In this section, we propose further possible investigations that we believe could improve the goodness of the proposed framework.

8.1.1 Testing the framework on more species

The results shown for Grevy's zebra and Reticulated giraffes are promising but are not enough to state for certain that our complete framework outperform the sole correction of the viewpoint bias. Therefore, it is necessary to test it on more species. In particular, we believe to be of high interest the possibility of testing the framework on animals that live in different habitats, such as aquatic mammals. In particular, whale shark are already supported by Wildbook $^{\mathrm{TM}}$.

8.1.2 Improving most recent estimates

Our results show issues in assessing the size of the species for the most recent year. In our tests, we found that to be a systematic issue: whenever we restricted the dataset to a smaller time interval, the last estimates were always the less accurate. Since the most recent estimates are also the most important, we believe fundamental to continue the study in this direction.

8.1.3 Evaluation the bias associated to single image

Whenever a user publishes a single image, we have so little information that our regression model cannot be applied. In fact, the extraction of used features requires to have more than one image. In all these cases, since we weren't able to predict the *Percentage Bias*, we used instead the mean of the *Percentage Bias* that we observed in our training dataset. However, we suggest the investigation of more sophisticated approaches.

This would be required to include in the dataset also social media that do not allow user to publish images collections, but just single pictures.

8.1.4 Gaps in biases analysis

As explained in Section 5.2, there is a number of biases that may affect the entire process of estimating a population size using social media images. Not all of these biases are strictly related to social media. In this section, we are going to discuss the limits of our methods in account these biases.

With respect to Figure 5.2, we need to investigate whether touristic activities could be sufficient in covering the entire area corresponding to the habitat of species under study. However, we believe that the future growth of social media and tourism will lead to an improvement of estimates in future years. Moreover, a possible mitigation to the problem may come from scraping other social medias to retrieve more images of the species studied. This may lead to an improvement in the stability of results over the years, especially in the case of endangered species, for which the sighting of a single individual may affect the estimate for the entire species.

Moreover, for certain species, such as Grevy's zebra and Reticulated giraffe, further studies must be conducted on the bias related to the lack of bilateral symmetry. Even though in Section 5.6.5 we developed a solution to mitigate this issue, it may be the case that bilateral symmetry does not always prevent us to identify an animal. In other words, we need to quantify this bilateral symmetry for every species, taking into account its relation to the computer vision tools used to identify individuals.

8.1.5 Use of a wider range of pictures for MTurk experiments

When labeling pictures to train our machine learning model, we made use of sole animal pictures. Even though these photos are crucial for our estimates, it is unlikely that a tourist avoid taking pictures of landscapes, monuments, friends, or family. This is indeed shown by the presence of these types of photos on social media. Our study shows that the shareability of images exhibits some learnable patterns. Therefore, we suggest creating more complex albums containing a wider variety of subjects, not strictly related to animals. These would allow training machine learning models on data that will be more similar to the ones shared on social media, increasing the accuracy of the whole framework.

8.1.6 Conduct MTurk experiments on a different population

As explained in Section 6.3.1, the population on crowdsourcing marketplace, such as MTurk, may not be a good sample of social media users population. Therefore, we suggest to interview more social media users in order to train a better machine learning model.

8.1.7 Study of social interactions between individuals

For the species that live in a herd, we suggest creating networks models to represent herds. It is likely that a photo showing a few individuals could be related to the sighting of an entire herd. On the other hand, if a herd hasn't been photographed for a long time, it could be the case that the entire population has migrated to other regions, or it is suffering from a disease. This relation may provide information useful in improving the accuracy of machine learning models.

List of Abbreviations

CMR Capture-Mark-Recapture.

EXIF Exchangeable Image File Format.

GGR Great Grevy's Rally (2016).

GGR2 Great Grevy's Rally 2 (2018).

GZGC Great Zebra and Giraffe Count.

HIT Human Intelligence Task.

IBEIS Image Based Ecological Information System.

IUCN International Union for Conservation of Nature's Red List of Threatened Species.

MTurk Amazon Mechanical Turk.

UIC University of Illinois at Chicago.

Bibliography

[1] Omnicore Agency. *Facebook by the Numbers: Stats, Demographics and Fun Facts.* Available at `https://www.omnicoreagency.com/facebook-statistics/` (November 21, 2018).

[2] M. Barrett et al. *Living Planet Report 2018: Aiming Higher.* Glands, Switzerland, 2018. URL: `http://pure.iiasa.ac.at/id/eprint/15549/`.

[3] Yoav Benjamini and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *Annals of statistics* (2001), pp. 1165–1188.

[4] Tanya Berger-Wolf et al. *The Great Grevy’s Rally: The Need, Methods, Findings, Implications and Next Steps.* Tech. rep. Technical Report, Grevy’s Zebra Trust, Nairobi, Kenya, 2016.

[5] Tanya Y Berger-Wolf et al. “Wildbook: Crowdsourcing, computer vision, and data science for conservation”. In: *arXiv preprint arXiv: 1710.08880* (2017).

[6] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. “Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data?” In: *Perspectives on psychological science* 6.1 (2011), pp. 3–5.

[7] Declan Butler. “When Google got flu wrong”. In: *Nature* 494.7436 (2013), p. 155.

[8] Maximilian Christ, Andreas W Kempa-Liehr, and Michael Feindt. “Distributed and parallel time series feature extraction for industrial big data applications”. In: *arXiv preprint arXiv:1610.07717* (2016).

[9] Maximilian Christ et al. “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh–A Python package)”. In: *Neurocomputing* (2018).

[10] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. "Demographics and dynamics of mechanical turk workers". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 135–143.

[11] Emily Dreyfuss. *A Bot Panic Hits Amazon's Mechanical Turk.* Available at `https://www.wired.com/story/amazon-mechanical-turk-bot-panic/` (November 20, 2018).

[12] Julian Fennessy et al. "Multi-locus analyses reveal four giraffe species instead of one". In: *Current Biology* 26.18 (2016), pp. 2543–2549.

[13] Adam G Hart et al. "Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa". In: *Methods in Ecology and Evolution* (2018).

[14] ImageAI. *ImageAI.* Available at `https://github.com/OlafenwaMoses/ImageAI/tree/master/imageai/Detection` (November 15, 2019).

[15] IUCN. *Grevy's Zebras.* Available at `https://www.iucnredlist.org/species/7950/89624491` (November 20, 2018).

[16] IUCN. *IUCN Iberian Lynx.* Available at `https://www.iucnredlist.org/species/12520/50655794/` (November 10, 2019).

[17] IUCN. *IUCN Summary Statistics.* Available at `https://www.iucnredlist.org/resources/summary-statistics/` (November 10, 2019).

[18] IUCN. *Reticulated Giraffes.* Available at `https://www.iucnredlist.org/species/88420717/88420720` (November 15, 2019).

[19] George M Jolly. "Explicit estimates from capture-recapture data with both death and immigration-stochastic model". In: *Biometrika* 52.1/2 (1965), pp. 225–247.

[20] Mona Kasra, Cuihua Shen, and James F O'Brien. "Seeing Is Believing: How People Fail to Identify Fake Images on the Web". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM. 2018, LBW516.

[21] Charles J Krebs et al. *Ecological methodology.* Tech. rep. Harper & Row New York, 1989.

[22] Richard A Lancia et al. "Estimating the number of animals in wildlife populations". In: (2005).

[23] David M. J. Lazer et al. “The science of fake news”. In: *Science* 359.6380 (2018), pp. 1094–1096. ISSN: 0036-8075. DOI: `10.1126/science.aao2998`. eprint: `http://science.sciencemag.org/content/359/6380/1094.full.pdf`. URL: `http://science.sciencemag.org/content/359/6380/1094`.

[24] R Lydekker. “On the Subspecies of Giraffa camelopardalis.” In: *Proceedings of the Zoological Society of London*. Vol. 74. 1. Wiley Online Library. 1904, pp. 202–229.

[25] Jana Machajdik and Allan Hanbury. “Affective image classification using features inspired by psychology and art theory”. In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 83–92.

[26] Francesco Marra et al. “Detection of GAN-Generated Fake Images over Social Networks”. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 384–389.

[27] Wild Me. *Image Based Ecological Information System*. Available at `https://github.com/Erotemic/ibeis/tree/next` (November 21, 2018).

[28] Wild Me. *Wildbook: Software to Combat Extinction*. Available at `https://www.wildbook.org` (November 21, 2018).

[29] Mary Meeker and Liang Wu. “Internet trends 2017”. In: *Code Conference, Preso da: http://www. kpcb. com/internet-trends*. 2017.

[30] Sreejith Menon. “Animal Wildlife Population Estimation Using Social Media Images”. PhD thesis. 2017.

[31] Sreejith Menon et al. “Animal Population Estimation Using Flickr Images”. In: (2016).

[32] Litman L. Moss A. J. *After the bot scare: Understanding what’s been happening with data collection on MTurk and how to stop it.* Available at `https://bit.ly/2xAYf3j` (November 20, 2018).

[33] Didrik Nielsen. “Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?” MA thesis. NTNU, 2016.

[34] Alexandra Olteanu et al. “Social data: Biases, methodological pitfalls, and ethical boundaries”. In: (2016).

[35] Joy Omulupi Ontita et al. “The State of Kenya’s Grevy’s Zebras and Reticulated Giraffes: Results of the Great Grevy’s Rally 2018”. In: ().

[36] Ray Oshikawa, Jing Qian, and William Yang Wang. “A Survey on Natural Language Processing for Fake News Detection”. In: *arXiv preprint arXiv:1811.00770* (2018).

[37] Jason Parham et al. “Animal Population Censusing at Scale with Citizen Science and Photographic Identification”. In: *e AAAI 2017 Spring Symposium on AI for Social Good (AISOC)*. 2017.

[38] Andrew Perrin. “Social media usage”. In: *Pew research center* (2015), pp. 52–68.

[39] K Pavan Kumar Reddy and R Aravind. “Measurement of asymmetry of stripe patterns in animals”. In: *Signal Processing and Communications (SPCOM), 2012 International Conference on*. IEEE. 2012, pp. 1–5.

[40] Adi Robertson. *Facebook users have uploaded a quarter-trillion photos since the site’s launch.* Available at `https://www.theverge.com/2013/9/17/4741332/facebook-users-have-uploaded-a-quarter-trillion-photos-since-launch` (November 10, 2019).

[41] D Rubenstein et al. *Equus grevyi. The IUCN Red List of Threatened Species 2016: e. T7950A89624491.* 2017.

[42] DI Rubenstein et al. *The great zebra and giraffe count: The power and rewards of citizen science.* Tech. rep. Technical Report, Kenya Wildlife Service, Nairobi, Kenya, 2015.

[43] Derek Ruths and Jürgen Pfeffer. “Social media for large studies of behavior”. In: *Science* 346.6213 (2014), pp. 1063–1064. ISSN: 0036-8075. DOI: `10.1126/science.346.6213.1063`. eprint: `http://science.sciencemag.org/content/346/6213/1063.full.pdf`. URL: `http://science.sciencemag.org/content/346/6213/1063`.

[44] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. “An Image Is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures.” In: *ICWSM*. 2015, pp. 397–406.

[45] George AF Seber. “A note on the multiple-recapture census”. In: *Biometrika* 52.1/2 (1965), pp. 249–259.

[46] Kenya Wildlife Service. *Giraffe Conservation Status Report.* Available at https://giraffeconservation.org/wp-content/uploads/2016/03/Kenya-no-map.pdf (November 15, 2019).

[47] Alvy Ray Smith. "Color gamut transform pairs". In: *ACM Siggraph Computer Graphics* 12.3 (1978), pp. 12–19.

[48] Internet World Stats. *World Internet Users Statistics and 2018 World Population Stats.* Available at https://www.internetworldstats.com/stats.htm (November 20, 2018).

[49] Flickr (TM). *Flickr API Python implementation.* Available at http://stuvel.eu/projects/flickrapi (November 21, 2018).

[50] Flickr (TM). *Flickr Services.* Available at https://www.flickr.com/services/api/ (November 21, 2018).

[51] Amazon Mechanical Turk. *Amazon Mechanical Turk FAQs.* Available at https://www.mturk.com/worker/help/ (November 20, 2018).

[52] International Telecommunication Union. *The world in 2015.* Available at https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf (November 20, 2018).

[53] The World Tourism Organization (UNWTO). *2017 International Tourism Results: the highest in seven years.* Available at http://media.unwto.org/press-release/2018-01-15/2017-international-tourism-results-highest-seven-years (November 21, 2018).

[54] Liwen Vaughan and Mike Thelwall. "Search engine coverage bias: evidence and possible causes". In: *Information processing & management* 40.4 (2004), pp. 693–707.

[55] Craig Welch. *Orca Killed by Satellite Tag Leads to Criticism of Science Practices.* Available at https://www.nationalgeographic.com/news/2016/10/orca-killed-by-satellite-tag-159 (October 11, 2019).

[56] Gary W Witmer. "Wildlife population monitoring: some practical considerations". In: *Wildlife Research* 32.3 (2005), pp. 259–263.

[57] Calvin Zippin. "An evaluation of the removal method of estimating animal populations". In: *Biometrics* 12.2 (1956), pp. 163–189.

[58] San Diego Zoo. *Zebra: Equus zebra, E. quagga, E. grevyi.* Available at `https://animals.sandiegozoo.org/animals/zebra` (November 15, 2019).