12-31-2022

Photonic monitoring of atmospheric fauna

Adrien P. Genoud
New Jersey Institute of Technology, ap.genoud@gmail.com

ABSTRACT

PHOTONIC MONITORING OF ATMOSPHERIC FAUNA

by
Adrien P. Genoud

Insects play a quintessential role in the Earth's ecosystems and their recent decline in abundance and diversity is alarming. Monitoring their population is paramount to understand the causes of their decline, as well as to guide and evaluate the efficiency of conservation policies. Monitoring populations of flying insects is generally done using physical traps, but this method requires long and expensive laboratory analysis where each insect must be identified by qualified personnel. Lack of reliable data on insect populations is now considered a significant issue in the field of entomology, often referred to as a "data crisis" in the field. This doctoral work explores the potential of entomological photonic sensors to unlock some of the limitations of traditional methods. This work focuses on the development of optical instruments similar in essence to lidar systems, with the goal of counting and identifying flying insects from a distance in their natural habitat. Those systems rely on the interactions between the near-infrared laser light and insects flying through the laser beam. Each insect is characterized by retrieving its optical and morphological properties, such as wingbeat frequency, optical cross sections, or depolarization ratios. This project ran in parallel a series of laboratory and field experiments. In the laboratory, prototypes were tested and used to create a database of insects' properties. The data were used to train machine learning classifiers aiming at identifying insects from optical signals. In the case of mosquitoes, the sex and species of an unknown specimen was predicted with a 99% and 80% accuracy respectively. It also showed that the presence of eggs within the abdomen of a female mosquito could be detected from several meters away with 87% accuracy. In the field, instruments were deployed in real-world conditions for a

total of 520 days over three years. More than a million insects were observed, allowing to continuously monitor their aerial density over months with a temporal resolution down to the minute. While this approach remains very new, this work demonstrated that photonic sensors could become a powerful tool to tackle the current lack of data in the field of entomology.

PHOTONIC MONITORING OF ATMOSPHERIC FAUNA

by
Adrien P. Genoud

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Applied Physics

Department of Physics

December 2022

APPROVAL PAGE

PHOTONIC MONITORING OF ATMOSPHERIC FAUNA

Adrien P. Genoud

Dr. Benjamin Thomas, Dissertation Advisor — Date
Associate Professor of Physics, NJIT

Dr. Haimin Wang, Committee Member — Date
Distinguished Professor of Physics, NJIT

Dr. Camelia Prodan, Committee Member — Date
Associate Professor of Physics, NJIT

Dr. Tao Zhou, Committee Member — Date
Associate Professor of Physics, NJIT

Dr. Gregory Williams, Committee Member — Date
Adjunct Professor, Center for Vector Biology, Rutgers University, New Brunswick, NJ

BIOGRAPHICAL SKETCH

Author: Adrien P. Genoud

Degree: Doctor of Philosophy

Date: December 2022

Undergraduate and Graduate Education:

- Doctor of Philosophy in Applied Physics,
  New Jersey Institute of Technology, Newark, NJ, 2022
- Master of Science in Sciences of Ocean Atmosphere and Climate,
  Claude Bernard University, Lyon, France, co-accredited with Ecole Centrale, Lyon, France, 2017
- Bachelor of Science in Physics and Chemistry,
  Joseph Fourier — Grenoble 1 University, Grenoble, France, 2015

Major: Applied Physics

Presentations and Publications:

A.P. Genoud, Near-infrared optical sensors to monitor flying insects, *International Congress of Entomology*, Oral presentation, Helsinki, Finland 2022.

A.P. Genoud, Monitoring the aerial density and circadian rhythm of flying insects using a near-infrared stand-off optical sensor, *Mid-Atlantic Section of the American Physical Society*, Oral presentation, New Brunswick, NJ, 2021 .

A.P. Genoud, G.M. Williams, B.P. Thomas, Continuous monitoring of aerial density and circadian rhythms of flying insects in a semi- urban environment, *PLoS ONE*, vol. 16, 11, 2021.

A.P. Genoud, Tracking insect population densities and activities using an entomological stand-off optical sensor, *Mid-Atlantic Section of the American Physical Society*, Oral presentation, Newark, NJ, 2020 .

A.P. Genoud, J. Torsiello, M. Belson, B.P. Thomas, Entomological photonic sensors: Estimating insect population density, its uncertainty and temporal resolution from transit data, *Ecological Informatics*, vol. 61, 2020.

A.P. Genoud, Y. Gao, G.M. Williams, B.P. Thomas, A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals, *Ecological Informatics*, vol. 58, 2020.

A.P. Genoud, Y. Gao, G.M. Williams, B.P. Thomas, Identification of gravid mosquitoes from changes in spectral and polarimetric backscatter cross sections, *Journal of Biophotonics*, vol. 12, 10, 2019.

A.P. Genoud, Identification of gravid mosquitoes from changes in spectral and polarimetric backscatter cross sections, *29th International Laser Radar Conference*, Poster session, Hefei, China, 2019.

R. Basistyy, A.P. Genoud, A. Diaz, F. Moshary, B.P. Thomas, Active standoff mixing-ratio measurements of N2O from topographic targets using an open-path quantum cascade laser system, *Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE), 10779*, Lidar Remote Sensing for Environmental Monitoring XVI, Honolulu, HI, 2018.

A.P. Genoud, R. Basistyy, G.M. Williams, B.P. Thomas, Analysis of predictor variables for mosquito species identification from dual-wavelength polarization-sensitive lidar measurements, *Proceedings of SPIE, 10779*, Lidar Remote Sensing for Environmental Monitoring XVI, Honolulu, HI, 2018.

R. Basistyy, A.P. Genoud, B.P. Thomas, Backscattering properties of topographic targets in the visible, shortwave infrared, and mid-infrared spectral ranges for hard target lidars, *Applied Optics*, vol. 57, 24, 2018.

A.P. Genoud, R. Basistyy, G.M. Williams, B.P. Thomas, Optical remote sensing for monitoring flying mosquitoes, gender identification and discussion on species identification, *Applied Physics B*, vol. 134, 3, 2018.

A.P. Genoud, Tracking insect population densities and activities using an entomological stand-off optical sensor, *Mid-Atlantic Section of the American Physical Society*, Poster presentation, Newark, NJ, 2017.

A ma maman, partie trop tôt. J'espère que tu est fière de moi. Amour, ton Adriou.
To my mom, gone too soon. I hope you are proud of me. Love, your Adriou.

A mon papa, merci pour tout ton amour. Amour, ton fils.
To my dad, thank you for all your love. Love, your son.

ACKNOWLEDGMENT

First and foremost, I want to thanks my advisor, Professor Benjamin Thomas. Thank you for inviting me to work on this project. I will always remember our first interview together as one of the greatest opportunities of my life. Thank you for guiding me throughout all those years. You have always been available, reviewed and corrected so many of my drafts and taught me how to actually design and conduct lab experiments. I truly appreciate your ability to correct errors and mistakes with words of encouragement and support, it is a rare yet invaluable quality. Thank you for your unstoppable efforts in applying for grants and funding. This difficult, and sometimes mind-boggling, job has been a cornerstone of the success of our project. Working with you has been an honor and a privilege.

I would also like to thank all of my dissertation committee members. Professors, Haimin Wang, Camelia Prodan and Tao Zhou and last but not least Dr. Gregory Williams. Your contributions to the success of my PhD are valued and appreciated.

I should acknowledge the support of the NJIT Physics Department, which provided me with a teaching assistantship position. Thank you for your confidence and for providing me with an incredible experience, teaching younger minds about the wonders of Physics. Our research has also been supported by two grants from the National Institutes of Health (award numbers R03AI138133 and 1R21AI153732-01A1). I also received travel support from the National Aeronautics and Space Administration and the Research Foundation of The City University of New York, which has allowed me to present my research at the International Laser Radar Conference, an amazing experience that has made our new field better known to the scientific community.

I want to thank my peers and friends. To Professors Andrew Gerrard and Andrei Sirenko, thank you for all your work as Chair of the Physics Department. It

is an important responsibility that you have both fulfilled admirably. To Professors Haimin Wang and Bin Chen, thank you for being Doctoral Program Directors, your help in navigating my coursework every semester has been invaluable. I know how much time you have dedicated to all of us. To Sung Maeng and Xuecong Zhou thank you for all you have done to make teaching lab classes both interesting and fulfilling. To Professor Keun Hyuk Ahn thank you for overseeing so many interesting seminars, every graduate student in the Department has discovered new topics and interesting research thanks to your hard work. A special thanks to Professor Cristiano Dias, your interest in my research and our discussions on the topic have helped me more than you probably know. To Professors Trevor Tyson, Dale Gary, Slawomir Piatek and John Federici thank you for all you taught me while I was your student or your assistant. To my former colleague Yunpeng Gao, thank you for being a co-author on several publications. I want to express my thanks to Christine Oertel, Leslie Williams, Cindy Montalvo-Harden, and Donald James. Your help has been greatly appreciated. You were always here to help me, with quick and effective solutions to many of my problems. Everyone in the Physics Department has been wonderful and I feel fortunate for such a great workplace environment.

I also want to thanks the undergraduate students that have worked with me throughout my PhD. Christo, Leslie, Joseph, Michael, Jung and Rushi, thank you for your passion and dedication, it was a pleasure working with all of you. Our collaboration has both helped me make progress in my research and taught me how to better guide motivated students.

I also want to thanks my teachers and mentors, from Lyon University. Professors Patrick Rairoux and Alain Miffre, thank you both for your amazing commitment and for guiding me during my Master. Without you I would have never had the chance to work on such an interesting topic.

I would like to thank Ms. Clarisa Gonzalez-Lenahan and Dr. Sotirios Ziavras for all their help during the review of this document. This manuscript is certainly much better thanks to your hard work.

I am incredibly grateful for all the friends I made along my journey at NJIT. To Cyrus, you have been the best roommate I could have think of. I remember fondly about how we met during orientation and how we bonded around our shared interest of history. To Aida, thank you for being the root of many of my friendships. All the barbecue parties in your backyard have been wonderful. To Evaristo, thank you for all the fun we had going out and celebrating life together. Thanks to you, I was able to reconnect with my inner sportsman when you invited me to your soccer games. To Nina, you have grounded me in our shared European experience, Swiss and French, sharing cheese and bread knowledge together. To Bogdan, thank you for all of our Catan game nights and for your never-ending generosity. To Mojtaba, thank you for our movie nights and fun NYC outings. To Topu my friend and colleague, thank you for all our lunches, coffee breaks and field expeditions. I always had a great time talking and working with you. To Ssu-Ying my next lab buddy, thank you for all the amazing Korean BBQ, always happy to randomly cross path with you in the hallway. To so many of you thank you, Iria, Darwin, Jezabel, Juan, Sërije, Javid, Claudia, Poncho, Lara, Joey, Elham, Farbod, Sheri, Alina and everyone that I met.

I'll be remiss if I did not thank all my friends from back home. To Stan and Gus, my best friends in the whole world, our Skype talks throughout the years have been so meaningful to me, I would not miss those for anything. I have known you both since we were a few years old. It has been a privilege growing up with you and those two friendships are one of my greatest successes in life, may they continue forever. To Robin, Théo, Raph and Baptiste and Guilhem, I am fortunate to be able to count on you even after many years apart. My core group of friends from middle and high school, I feel so lucky to have you in my life, thank you for all the fun we have had

and will have. To the many other friends that I have known, Quentin, Mathilde, Clement, Jules, Oscar, Manon, Eloi, David, Lucile, Elias, Sylvain, Pia, Léo, Jérémie, Bais, Simon, Felix, Thibault, Marine, Bapt, Rémi, François, Pauline thank you for making my life better by your presence.

To my family, who has made it all possible. I have relied on you more than anyone. Thank you for all the wonderful things you have done for me. To my mom, Christine Lentheric, you have been the most exceptional mother that I could have imagined. Your constant love and support have made me the person that I am today. Without even my knowledge you gave me all the keys that I would ever need to grow and be happy. I love you so much, thank you. To my dad, Bernard Genoud, you are my hero and my model. Your selflessness, generosity and attention have nurtured my soul and my life. I love you so much, thank you. To Joséphine, my second mother by every definition, I have loved you ever since I can remember. You always have a special place in my heart. We may not be related by blood but we certainly are bound forever by love, thank you. To Gigi, your presence has always been a blessing. I love you and will always remember how much joy we shared whenever I would visit Montpellier. To Laetitia, Sven, Evan, Danielle and Mich, our side of the family is small in number but deep with love. To Hervé, Marieke, Gino and Marie you have been there for me and have supported me through some difficult times. I really consider you all as part of my family and send you all my love. To Sebastien, David, Hélène, Jean-mi, Mahina and Lisa, thank you for all the wonderful moments that we shared around Christmas. I am filled with love when I am around you. I have and will treasure every moment with you. To Michel, Romain, Manon, Cathy, Alain, Solange, Elodie, Sophie, and to everyone else in the family, thank you.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

Insects are everywhere, from the hottest deserts to the cold snows of Antarctica. Their diversity is astonishing with millions of identified species and even more yet to be discovered [1]. Insects are the most diverse group of animals [2] and represent one of the largest biomasses of terrestrial animals [3]. They play a quintessential part of terrestrial ecosystems due to their large number and variety of functions [4, 5]. For all of these reasons, the monitoring of insect is paramount. However, the monitoring of insects' activity faces serious challenges, resulting in little quantitative data about their population dynamics. Monitoring change in insect distribution, diversity and abundance poses a formidable challenge to entomologists. Long-term estimates of population trends among insect species are difficult to implement, they require qualified personnel to collect and identify insects captured in traps, often for years. The diversity, number, and size of insects are added difficulties making such endeavors both labor and capital intensive. As a result, studies reporting observations of actual changes in population remain rare [6, 7, 8, 9, 10, 11]. While immensely valuable, such studies are geographically limited and specific to just a few species of insects compared to the millions of discovered species. Furthermore, insects' spatial distribution and abundance can rapidly vary with global climate change [12, 13, 14, 15]. As a result, estimating trends in population of specific insect groups, both on a local or global scale, greatly suffers from our inability to collect entomological data. New and innovative optical methods have known recent improvements in the field of entomology [16], aiming for a better understanding of insect population dynamics without the need to capture or disturb the natural behavior of the studied specimens. The photonic monitoring of atmospheric fauna,

such as flying insects, could potentially unlock some of the current limitations and challenges faced by entomologists.

1.1 Impact of Insects on Ecosystems

Insects are a vital part of most, if not all, ecosystems on Earth. They provide ecological services that may often go unnoticed but are nonetheless fundamental. In their absence, many crops would go unpollinated, causing famine. Food chain would be greatly disturbed causing catastrophic loss of biodiversity. However, some species of insect also have negative impacts on human society. Pests, such as locust, can damage crops causing famine and forced migration. Disease vectors, such as mosquitoes, can spread many infectious diseases, malaria alone being responsible for over 600,000 death in 2020 [17]. For all these reasons, monitoring insect populations is paramount.

1.1.1 Pollination

Pollination is, in simple terms, the process of transferring pollen from the stamen of a plant to the stigma [18]. It is the process that later enables fertilization of the plant and allows for the production of seeds. As such, this process is fundamental for human societies to produce many of its crops. While several types of abiotic pollination exist, such as wind [19] or rain pollination [20], they only represent 20% of all flowering plant pollination [21]. Which means that 80% of angiosperm pollination is biotic, requiring the action of another organism, the pollinator. Biotic pollination can be conducted by several animal groups of which insects are dominant [22].

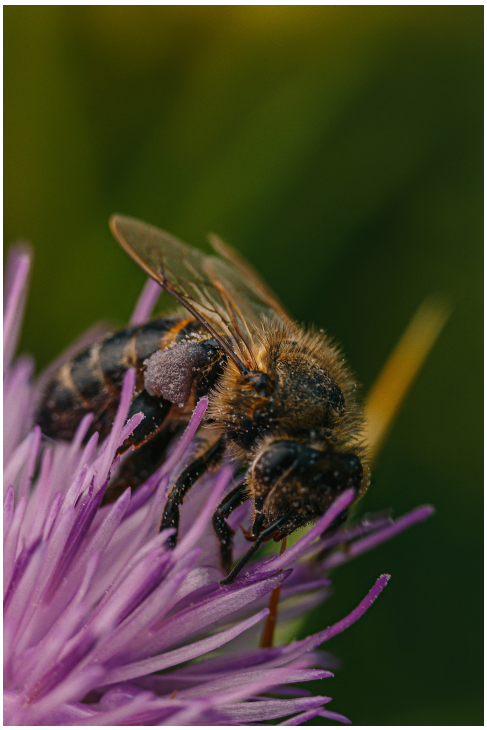

Figure 1.1 Illustration of biotic pollination by a bee.
Image credit: Raj Stevens.

Flying insects, and in particular bees, are the most important when it comes to crop pollination [23]. The economic value of pollination is estimated around 153 billion euros worldwide, which represents around 9.5% of the value of the world agricultural production [24]. This explains why the monitoring of flying insects, and in particular pollinators, is of grave importance when it comes to ensuring reliable food source for the growing human population. In the past decades, the decline of some key pollinator species has been concerning [25, 26, 27]. The now well-known colony collapse disorder has even been the focus of public interest [28]. A decline in pollinator abundance and diversity could be the cause for a severe economic crisis, food shortages and loss of plant diversity [29]. Yet the precise causes and consequences of such decline are difficult to rigorously identify and some authors are calling for additional monitoring campaigns and study [30, 31].

Figure 1.2 Insectivorous bird preying upon a flying insect.
Image credit: Rajukhan Pathan.

1.1.2 Food Source

Insects are an integral part of the food chain. They play a key role in the flow of nutrient from plants to larger animals that prey upon them. Therefore, insects are an important food source for several larger animals such as birds [32], bats [33], amphibians [34] and to a lesser extent humans [35]. In order to ensure the proper flow of nutrients through the ecosystems, it is necessary to protect the prey-predator relationship of insects with birds, bats and amphibians. Human activity and infrastructure, such as light pollution [36] or intensive agriculture [37] are known to negatively impact insect abundance, which can ripple through the food chain modifying the entire ecosystem dynamic. The disruption of the natural food source provided by insects, due to either their population decline, abrupt relocation, change in behavior or loss of diversity should consequently be well understood and closely monitored.

Figure 1.3 Group of *Lycorma delicatula* on an *Ailanthus altissima* tree. The dark liquid is tree sap pouring out of the wound caused by the pest feeding process. *Image credit: Lawrence Barringer.*

1.1.3 Pests

Insect pests are responsible for important damage in both agriculture and forests. It is estimated that around 20% of crop production is destroyed by insect pests worldwide [38]. Some food crops reach up to 40% loss despite an annual spending of 30,000 million of USD$ on insecticides [39]. Many crops require human intervention, such as insecticide [40] or physical barriers [41]. It is estimated that without protection from pests, the global potential loss of wheat and cotton, to name but a few important crops, would be respectively 50% and 80% annually [42].

Like food crops, trees are susceptible to insect pests. As an example, the spread of the spotted lanternfly (*Lycorma delicatula*) in the USA has put severe stress on several of its host plants. *Lycorma delicatula* damages trees by directly feeding upon them, leading to honey dew and sooty mold deposits [43, 44], which endangers the host tree.

Figure 1.4 *Aedes albopictus* female during the blood feeding process. This process can be responsible for spreading viruses, such as chikungunya or dengue, to the host. *Image credit: Anuj Prajith.*

1.1.4 Disease Vectors

Approximately 85% of disease-causing organisms are arthropods, with insects being of particular medical importance [45]. Through bite, sting or tissue infestation, insects can spread diseases and are a current health concern worldwide. Among the several species that can be disease vectors, such as midges or black flies, mosquitoes are by far the most concerning. Mosquito-borne diseases, that include malaria, dengue, yellow fever, chikungunya, Zika and West Nile still constitute a major health issue. In particular, considering that malaria is still impacting more than 200 million people per year and that Zika, chikungunya and dengue are causing international concern in view of their re-emergence [46, 47, 48, 49].

Several organizations and governmental agencies such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC, USA) have called for improved surveillance with respect to mosquito-borne diseases such as malaria or dengue fever [50, 51]. Vector control strategies can be used to mitigate the impact of mosquito borne disease on human and cattle populations. To be efficient, vector control strategies require accurate data on the fine-scale spatial

distribution of the relevant mosquito species. As stressed by Eisen et al. 2017 [30], lack of reliable data on the spatial distribution of key vectors has become a major limitation in the development of spatial epidemiologic and eco-epidemiologic models. Furthermore, the National Health Security Strategy and Implementation Plan 2015–2018 [52] has underlined that, with climate change, tropical and sub-tropical species can potentially reach new habitats and severely change the current distribution of disease vectors.

1.2 Insect Monitoring Techniques

Monitoring insect populations is paramount and has been evolving for many years. As such there exist multiple methods of study. The most common methods are centered around capturing, collecting and then studying wild insect specimens. While such methods allow for an extensive study of the captured specimens, they are often time consuming and thus expensive due to labor cost and microscopy or genetic laboratory analysis by trained professionals. Generally speaking, such approaches only provide results on long time scale, days to weeks, which may be insufficient for the proper understanding of short scale population dynamics, such as predator-prey interaction or light activated behavior. In order to complement those methods, and provide additional information, new remote sensing approaches were developed. Based on the indirect study of insects through the sounds that they emit or their interaction with light, those new methods can operate on a shorter time scale to capture important details of the insects' behavior. While such methods allow for the study of a considerable number of insects, even on short time scale, they suffer from the difficulty of remotely identifying the insect species that enter their sensing range. Therefore, more study and new techniques needs to be developed for such approaches to reach a more prominent status in the field of entomology.

Figure 1.5 Net sweeping process during wild specimen collection in Secaucus, NJ.

1.2.1 Sweep Net

Sweep netting is an insect sampling method that relies on capturing specimens with a net. This method is often used for its simplicity. The required equipment is lightweight, inexpensive and easy to use with little training. However, this method can be biased towards heavier and more active insects [53] which are more easily captured. Despite the simplicity of the process, rigorous and scientifically relevant procedures can be conducted. As an example, counting the number of captured insects after sweeping continuously over a predefined transect of a field for a specific amount of time can be used to estimate the relative insect abundance.

This process can be repeated multiple times over an entire monitoring campaign in order to study the change of insects' population. Although, due to the time-consuming process of manually capturing insects, abundance estimation is often done through alternative methods such as traps. Yet it can be useful as a screening process

to identify which species are present at the field location in order to decide which traps are most likely to be efficient or in order to capture live specimens for laboratory study.

1.2.2 Trap

Traps are the most widely used method for the study of insects. Unlike sweep-netting, they provide the advantage of continuously capturing insects without the need of direct human intervention. Many different designs of traps exist but they can be put into two main categories, non-attractant and baited. The first relies on the natural behavior of insects to trap them. The second uses bait such as light, pheromones, food, or CO_2 to attract nearby specimens into the trap [54, 55, 56, 57, 58, 59]. The non-attractant traps, such as the widely used malaise trap [60, 61], tend to capture fewer specimens than the baited traps but have the advantage of mitigating any bias that a bait may introduce. Indeed, baits used to attract insects are not evenly effective across species, age, and sex groups, resulting in systematic sampling imbalance between species present near the traps [62, 63, 64, 65]. Yet, in some cases, oversampling a specific species or sex group may be desirable. For example, capturing exclusively female mosquitoes is useful when investigating the risk of mosquito-borne disease transmission.

While traps provide a high accuracy for the identification of the family, genus, species and sex of the captured insects through expert identification or DNA barcoding [66], they do present some disadvantages. First and foremost, they require long and expensive laboratory analysis where each insect must be identified by qualified personnel. Traps also provide poor estimates of actual population size because the attractive range of the traps, and therefore the trap efficiency, is generally unknown and changes with weather conditions [67, 68, 69, 70]. Without an accurate estimate of population size, it is difficult to determine action thresholds or evaluate the effect of pollinator conservation strategies. Finally, traps suffer from poor time

Figure 1.6 Non-attractant malaise trap in the Marloth Nature Reserve, Western Cape Province.
Image credit: Frank Koch.

resolution: due to the time required to identify captured insects, traps are often used sporadically, e.g., one day per week in the six-year campaign reported in Kampen et al. 2017 [71]. Even while they operate, the time at which each insect has been captured is generally unknown, although physical traps with bottle rotator systems provide some information about the time of capture [72], the temporal resolution remains low. Despite recent improvements aiming toward automating some of the analysis tasks, such as smart traps [73, 74] or automated insect DNA barcoding [75], entomologists still face a lack of large scale and long term data on insect abundance.

1.2.3 Sound

Traditional methods of insect study that rely on the capture of specimen causes some limitations. The number of specimens sampled through those approaches are usually limited to a small subset of the total insect population. This limit emerges not solely due to the limited number of trapped insects but mostly from constraints due to the cost and man hours required for their formal identification. Hence, new alternative methods of insect sampling have been considered, including acoustic sensors. The

analysis of insect sound can be distinguished in two main categories. The first being the sounds deliberately emitted, such as the song of the cicadas. The second being the sounds involuntarily emitted, such as the characteristic buzzing sound of mosquitoes emitted as a byproduct of their flight. The biggest limitation in using sound to study insect is the range of detection, as sound intensity, which is relatively low to begin with, decreases with the square of the distance. Identifying insect sounds, outside of laboratory conditions, from a distance is thus made difficult by environmental noise, such as wind.

Acoustic study of insects has the potential to increase the availability of entomological data by automating part of the identification process. Identifying insects from their sounds has been made possible by the development of effective sound feature extraction and machine learning algorithms [76, 77, 78, 79]. A significant advantage of sound analysis is the availability of inexpensive acoustic sensors. This allows for the deployment of an array of acoustic sensors that can cover a significant area. In the specific case of mosquitoes, a recent initiative has shown that acoustic detection and classification of mosquitoes can harness the strength of citizen scientists using only low-cost smartphones that are widely available and used globally [80].

1.2.4 Radar

Radar technology has been used to monitor flying objects, and then later atmospheric conditions, for almost a hundred years. Briefly, radar technology consists of a transmitter that emits a radio wave and a receiver. As a target crosses the path of the emitted wave, some of it is scattered toward the receiver. Analysis of the scattered wave provides information about the target such as its position, heading, shape and cross section. Entomological radars [81, 82] have achieved significant results in the study of insects, in particular for high altitude migrations [83, 84, 85]. They benefit from observing large area and being able to operate over long periods of time

Figure 1.7 Shown beside Professor Jason Chapman is the entomological vertical-looking radar (VLR) device on the roof of the Plant and Invertebrate Ecology Division building at Rothamsted.
Image credit: Ian Woiwod.

with limited supervision, making them capable of observing extremely large number of insects, close to two million in the study by Hu et al. 2016 [86]. In addition, entomological radars have the ability to extract information about the observed insects. They can retrieve the insect trajectory, radar cross section, shape and in some cases even the wingbeat frequency [87, 88, 89].

However, large radar reflections from vegetation and ground features (clutter) prevent the radar methodology from operating at close proximity to the ground. Despite this limitation due to clutter, some studies of insect flight behavior at medium altitude have been achieved, as low as 15 m [90]. Alternative methods allow entomological radar to operate closer to the ground, in particular the harmonic scanning radar [91], but it requires the use of a transponder attached to the studied specimen and as such, is of particular use for the study of flight behavior but limited

for the study of insect abundance in their natural habitat. Moreover, the relatively large wavelength of radars make them ill-suited for the study of smaller insects such as mosquitoes, which are of significant importance for human health.

1.2.5 Lidar

Lidar methodology [92, 93], using laser beams instead of radio waves, has shown encouraging results. This remote sensing methodology is widely used for studying atmospheric processes, such as aerosol concentrations and formation [94, 95, 96, 97, 98, 99, 100], measurements of spatio-temporal distributions of trace gases such as, CO_2 [101], CH_4 [102, 103], O_3 [104], H_2O [105, 106], and also of volatile organic compounds [107].

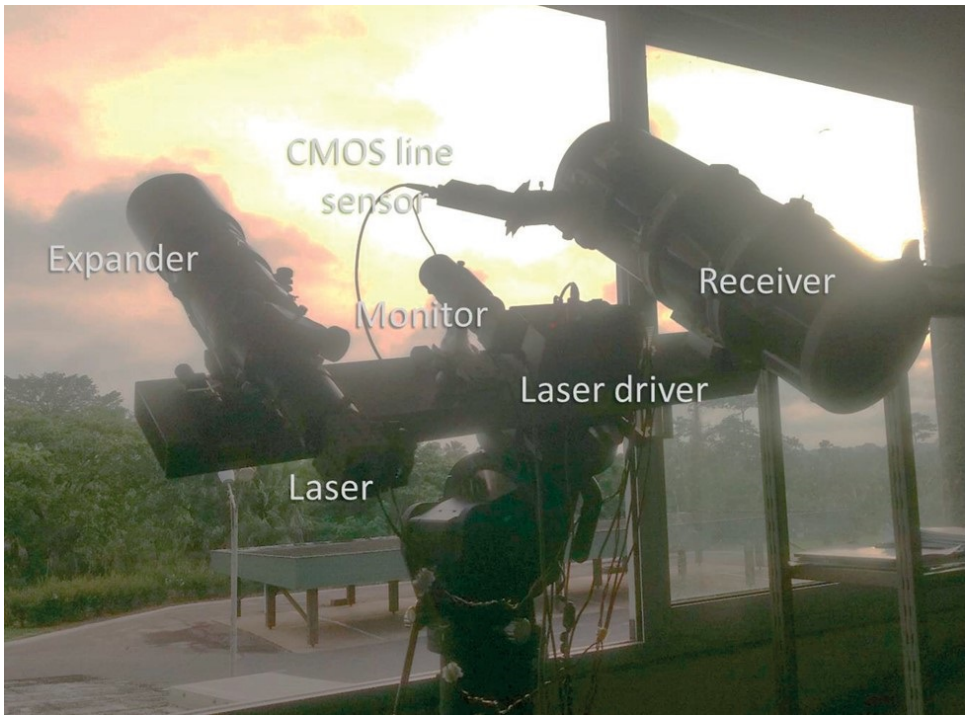

Figure 1.8 Picture of an entomological Scheimpflug lidar system.
Image credit: Samuel Jansson.

In the past few decades, a new and innovative usage of lidar technology has been emerging: the entomological lidar, directed toward the study of insects that cross its laser beam. This new approach aims at complementing traditional trapping methods

and to study insects that the radar technology still has difficulty studying. Indeed, the wavelength of lidar is by far small enough for the study of targets such as insects, even the smaller ones. Furthermore, lidars can operate near the ground without any issues from ground clutter. Infrared lidars in particular have gained significant traction in the field of entomology [108, 109, 110, 111, 112, 113]. With wavelengths usually between 800 and 1500 nm and up to hundreds of meters of range, they are particularly well equipped to study small, near-ground dwelling insects.

From studying the interaction between light and insects, entomological lidars can extract some of the specimen characteristics. It can evaluate the specimen wingbeat frequency [114], optical cross section [111] or velocity [113]. By evaluating the number of interactions between the instrument and insects, entomological lidars can also extract information about trends in insect populations. Using the time stamp of every interaction, they can also evaluate the period of insect activity, providing information on insect's circadian rhythms, with a much shorter time resolution than traps.

1.3 Interaction between Light and Biological Matter

Interaction between light and biological matter is at the heart of this manuscript. Unlike more traditionally studied light-matter interactions, the matter studied in this manuscript concerns targets that are a living, moving animal with all the complexity that this entails. It is therefore paramount to first understand light and several of its interactions with matter.

1.3.1 Description of Light

Light has been widely studied throughout history, since Democritus in the 5th century BCE to today. It is a complex phenomenon that can be interpreted and described in different ways.

One interpretation is to consider light as a collection of elementary particles: photons. Photons are a discrete quantum of the electromagnetic field. They have no charge nor mass. They are stable particles that travels at the speed of light in a vacuum, $c = 299792458\ m \cdot s^{-1}$. In this interpretation, photons have an energy E which is function of the Planck constant h, the speed of light in a vacuum c and their wavelength λ, as described by Equation (1.1). The photon interpretation of light is particularly helpful when considering phenomenon such as absorption or Compton scattering.

$$E = \frac{h \cdot c}{\lambda} \tag{1.1}$$

Another interpretation is to consider light as an electromagnetic wave. As developed by Huygens in the "Treatise on Light", considering light as a wave allows for a better understanding of the phenomenon of reflection and refraction, as latter put into mathematical equations by Fresnel. It also allows for the description of the phenomenon of interference as illustrated by the famous "double-slit experiment" designed by Young. A light wave can be described as a function $U(r, t)$ that is solution of the wave equation, see Equation (1.2), that can be derived from Maxwells' equations. Electromagnetic waves consist of the propagation of an oscillating electric and magnetic field, as illustrated by Figure 1.9.

$$\frac{1}{c^2} \cdot \frac{\partial^2 U}{\partial t^2} - \nabla^2 U = 0 \tag{1.2}$$

A simple, yet useful, solution of the wave equation is the monochromatic plane wave, as described by Equation (1.3). Plane waves can be used to derive properties of general waves as they often have an identical behavior while being mathematically simpler.

$$U(r,t) = U_0 \cdot e^{-i\left(\vec{k}\cdot\bar{r}-\omega\cdot t\right)} \tag{1.3}$$

Where U_0 is the wave magnitude, $\vec{k}$ the wave vector, $\bar{r}$ the direction of propagation, ω the angular frequency and t the time.

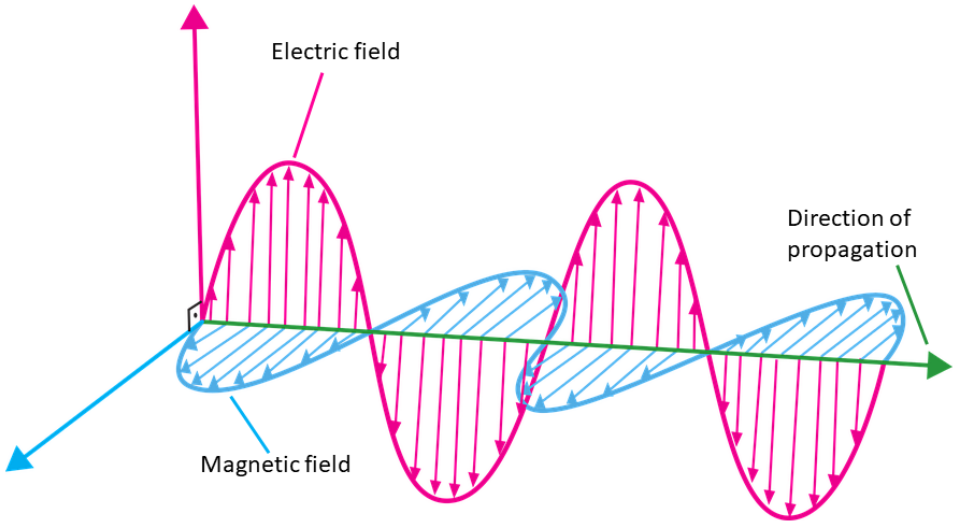

Figure 1.9 Illustration of the propagation of a linearly polarized transverse electromagnetic wave.

For both interpretations, the light, considered monochromatic throughout this manuscript as the presented work utilizes quasi-monochromatic laser sources, is characterized by a frequency, or a wavelength. Depending on the wavelength, the light interactions with an object change, which is of particular importance for this work. Table 1.1 briefly illustrates the different regions of the electromagnetic spectrum.

Table 1.1 Electromagnetic Spectrum

Type	Wavelength (m)
	$\leq 10^{-13}$
Gamma rays	10^{-12}
	10^{-11}
	10^{-10}
X rays	10^{-9}
	10^{-8}
Ultra violet	10^{-7}
Visible	10^{-6}
Infrared	10^{-5}
	10^{-4}
	10^{-3}
	10^{-2}
Microwave	10^{-1}
	1
Radio wave	10
	$\geq 10^{2}$

1.3.2 Polarization

One of the fundamental properties of transverse light waves is the polarization. In transverse waves, which are waves for which the coupled electric and magnetic fields are perpendicular to the direction of propagation, the direction of the polarization is defined as the orientation of the electric field oscillation. The polarization state of a light wave can be described by a set of four parameters $(S0, S1, S2, S3)$ which, when combined, form the Stokes vector $\vec{S}$. As defined by Equation (1.4) and the Poincare sphere Figure 1.10.

$$\vec{S} = \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix} = \begin{pmatrix} I \\ I \cdot P \cdot cos(2\Psi) \cdot cos(2x) \\ I \cdot P \cdot sin(2\Psi) \cdot cos(2x) \\ I \cdot P \cdot sin(2x) \end{pmatrix} \tag{1.4}$$

Where, I is the intensity of the wave, P the degree of polarization, x the ellipticity angle and Ψ the azimuthal angle, as detailed in Figure 1.10.

Light can be elliptically, circularly or linearly polarized. In this manuscript a special interest will be given to linearly polarized light as it has been used to study some of the properties of insects. In linearly polarized light, the oscillation of the electric field is bound to one dimension only. Light-matter interactions can change depending on the light polarization. Linear polarizers will transmit light with a polarization direction along their main axis and absorb all other directions. Polarization beam splitters will reflect s-polarized light, which is the light that has a polarization perpendicular to the plane of incidence, while transmitting p-polarized light, which is the light that has a polarization parallel to the plane of incidence.

Through interaction with a target, the polarization state of the light wave may be modified. Most commonly, some of the light will change its polarization, through multiple scattering, to become unpolarized, which is a state where the direction of

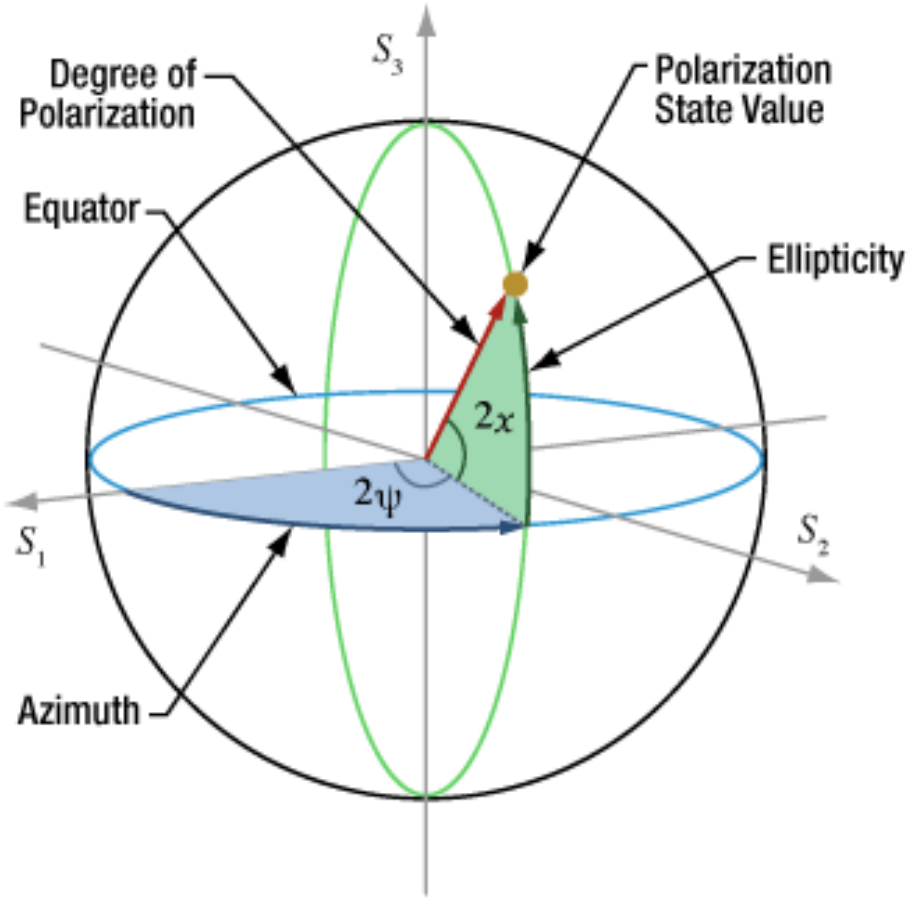

Figure 1.10 Poincare sphere that represents the instantaneous polarization state of an electromagnetic wave.

the electric field oscillation has no preferred direction. When considering the change in linear polarization for light-matter interactions, two directions of importance are to be considered. The initial direction of polarization, before the interaction, which is called the parallel polarization and the direction perpendicular to it, called the cross or perpendicular polarization.

An important target's property, for insects in particular, is the depolarization ratio, δ. This ratio describes how the interaction with a given target modifies the polarization of the incident light. As described by Equation (1.5), δ is defined as the ratio between the intensity of the light with a perpendicular polarization, $I_{\perp}$, by the intensity of the light with a parallel polarization, $I_{/\!/}$.

$$\delta = \frac{I_{\perp}}{I_{/\!/}} \tag{1.5}$$

1.3.3 Refraction

When crossing the interface between two mediums of different refractive indices, light undergoes the phenomenon of refraction. Its initial direction of propagation is modified at the interface between the two mediums, see Figure 1.11. According to Snell-Descartes law, Equation (1.7), the angle of refraction θ_t is function of the angle of incidence θ_i and the refractive indices of both medium n_1 and n_2. The refractive index of a medium n is given by the ratio of the speed of light in the vacuum c by the speed of light in the medium v, as displayed on Equation (1.6).

$$n = \frac{c}{v} \tag{1.6}$$

$$n_1 \cdot sin(\theta_i) = n_2 \cdot sin(\theta_t) \tag{1.7}$$

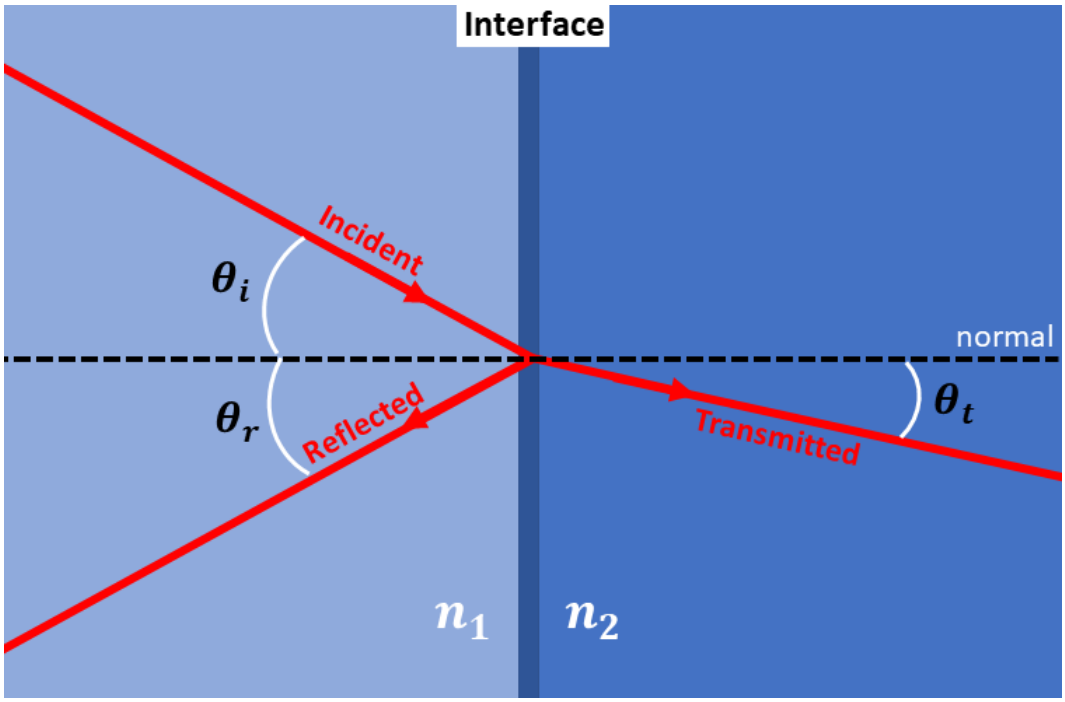

Figure 1.11 Refraction and reflection of light. An incident ray of light travels through a first medium of refractive index n_1 and reaches an interface with a second medium of refractive index n_2, with an angle of incidence θ_i. At the interface, part of the light is reflected with an angle of reflection $\theta_r = \theta_i$ and some is transmitted with an angle of transmission of θ_t. The amount that is reflected and transmitted is governed by Fresnel equations and depends on the polarization of the incident light.

This phenomenon is at the foundation of the construction of optical lenses. Lenses have been used in this contribution to collect and focus the light onto a detector and to expand and collimate a laser beam.

At the interface between two mediums, and in the absence of absorption, light either undergoes specular reflection or is transmitted. As described by the Fresnel equations, see Equations (1.8) and (1.9), the amount of reflected and transmitted light depends on the refraction index of the two mediums, the angle of incidence but also the polarization of the light. In those equations the index p corresponds to the p-polarized light and the index s corresponds to the s-polarized light. Coefficient R and T are respectively the reflection and transmission coefficient that describes the amount of reflected and transmitted light.

$$R_p = \left| \frac{n_1 \cdot cos(\theta_t) - n_2 \cdot cos(\theta_i)}{n_1 \cdot cos(\theta_t) + n_2 \cdot cos(\theta_i)} \right|^2, T_p = 1 - R_p \tag{1.8}$$

$$R_s = \left| \frac{n_1 \cdot cos(\theta_i) - n_2 \cdot cos(\theta_t)}{n_1 \cdot cos(\theta_i) + n_2 \cdot cos(\theta_t)} \right|^2, T_s = 1 - R_s \tag{1.9}$$

1.3.4 Scattering

As light encounters an object such as a particle, an aerosol or a rain drop, some of the light may undergo a phenomenon of scattering in which its direction of propagation is modified. Unlike specular reflection, where the reflected light propagates in a single new direction, the scattered light may be propagating in any new directions. Hence, the scattering phenomenon is often defined by a phase function that described the amount of light reflected in each direction, as illustrated in Figure 1.12. The theoretical form of this phase function varies in function of the light wavelength and the object properties.

Three main domains can be considered. If the wavelength is significantly smaller than the size of the object then Rayleigh scattering best describes the phase function. If the wavelength and object are of the same order of magnitude Mie scattering is used. Finally, if the wavelength is significantly larger than the object, we enter the domain of geometrical optics.

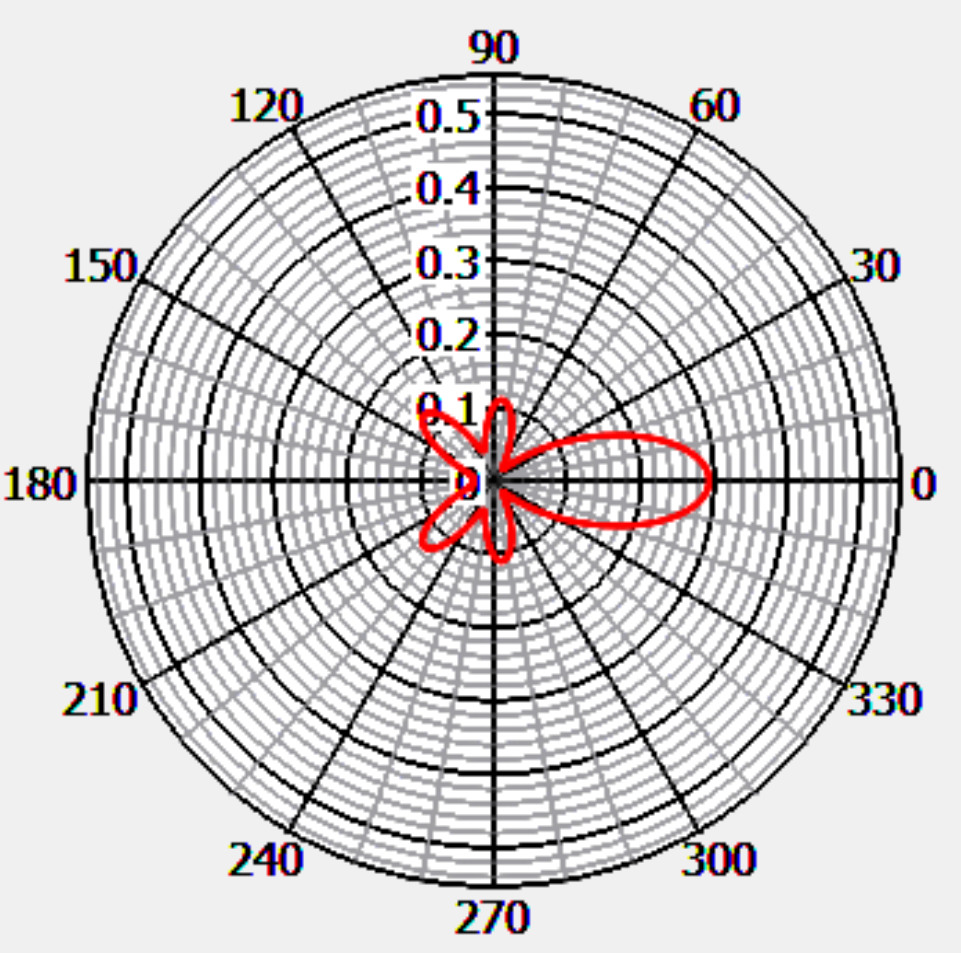

Figure 1.12 Polar representation of a phase function. The red line indicates the amount of light scattered in function of the direction, 0° being the forward direction, i.e., initial direction of the incident light. The phase function presented here is the result of the simulation of the scattering from a 400 nm radius sphere of melanin, for a wavelength of 940 nm. Results were obtained from the Mie Simulator: Virtual Photonics Technology Initiative of the Beckman Laser Institute, University of California Irvine, CA.

1.3.5 Absorption

Absorption is a phenomenon in which, as light propagates through a medium, some of the light energy is absorbed by matter and converted to another form. Using the wave interpretation of light, the optical absorption can be described by Equation (1.10). In an absorptive medium the wave number is complex, $\tilde{k} = k + i \cdot \alpha$. In this case, the monochromatic plane wave equation, Equation (1.3), can be rewritten to explicitly include the absorption coefficient which can also be defined from the complex refractive index of the medium through the Kramers-Kronig relations.

$$U(z,t) = U_0 \cdot e^{-\alpha \cdot z} \cdot e^{-i(k \cdot z - \omega \cdot t)} \tag{1.10}$$

Where α is the absorption coefficient, expressed in m^{-1}, z the distance travelled through the medium, expressed in m, and all other terms have been previously defined.

The photon interpretation of light is particularly useful to describe absorption. When the energy of the photon is equal to the energy of the transition between two electron states there is a non-zero probability of it being absorbed. As such the phenomenon of absorption heavily depends on the wavelength. This is of particular interest for this work as depending on the chosen wavelength the absorption due to the light interaction with an insect will change. In some cases, such as the backscattered configuration, see subsection 2.1.1, it may be beneficial to choose a wavelength that is not absorbed by the insect body, therefore increasing the amount of light being backscattered, increasing the signal to noise ratio (SNR). On the other hand, in the extinction configuration, subsection 2.1.2, it may be better to select a wavelength that is absorbed by the insect's wings in order to maximize their signal attenuation, also improving the SNR.

1.3.6 Backscattering and Extinction

Backscattering and extinction are special cases of general scattering and absorption. These two phenomena require special consideration as they are at the basis of the two types of entomological photonic sensor (EPS) used to study atmospheric fauna, see section 2.1.

As opposed to forward scattering, backscattering is the scattering of a light wave in the opposite direction of its initial (forward) direction of propagation. In Figure 1.12, the backscattering is indicated by the 180° direction. For an EPS in backscatter configuration, detailed in subsection 2.1.1 of this manuscript, the backscattered light is collected and studied to extract characteristics of the target. One of them being the backscatter cross section σ_{bck}, expressed in mm^2. The backscatter cross section is defined as the product of the target apparent geometrical cross section σ_G with the reflectance of the target R, see Equation (1.11). σ_G is defined as the surface equal to the intersection between the target and the laser beam which, in the case where the target is entirely within the laser beam, is equal to the projected area of the target in the photodetector plane of observation. The backscatter cross section can be understood as the cross-sectional area that a white Lambertian target would need to have in order to backscatter the same intensity as the considered target.

$$\sigma_{bck} = R \cdot \sigma_G \tag{1.11}$$

As light propagates through a medium or interacts with an object, some of it is absorbed and some is scattered away from the forward direction. The combination of these two effects is called optical extinction or attenuation. A simple description of this phenomenon is shown in Equation (1.12), which is the Beer-Lambert equation.

$$I = I_0 \cdot e^{-\mu_a \cdot l} \tag{1.12}$$

Where I is the intensity of the light (proportional to the squared amplitude of the electric field) that decreases as the wave propagates through the medium. I_0 is the initial intensity of the light, as it enters the medium, μ_a the attenuation coefficient of the medium and l the distance travelled through the medium.

The attenuation coefficient includes the effect due to both absorption and scattering. Consequently, it can be defined as the sum of the absorption coefficient α and the scattering coefficient μ_s, as sown in Equation (1.13).

$$\mu_a = \alpha + \mu_s \tag{1.13}$$

When considering the extinction due to a solid target, an insect in this contribution, the amount of reduction of the light intensity is function of the target geometrical cross section as well as the optical attenuation coefficient of the target. In extinction configuration, one of the key characteristics of the target is its extinction cross section σ_{ext}, expressed in mm^2. The extinction cross section is function of the target apparent geometrical cross section σ_G and the quasi-ballistic transmittance of the target T_{qb}, see Equation (1.14). Quasi-ballistic photons are the photons that are transmitted through the target and reach the active area of the photodetector. Those photons are the ones that are not absorbed, and that undergo a sufficiently small amount of scattering or diffraction so that their direction of propagation remains within the solid angle of detection of the photodetector.

$$\sigma_{ext} = (1 - T_{qb}) \cdot \sigma_G \tag{1.14}$$

For an EPS in the extinction configuration, detailed later in subsection 2.1.2 of this manuscript, the extinction cross section can be understood as the cross-sectional

area that an opaque (near 100% extinction) target would need to have, in order to induce an attenuation of the intensity equivalent to the one induced by the target.

CHAPTER 2

METHODS

2.1 Entomological Photonic Sensor (EPS)

As discussed earlier, there is a need for improved methods to collect entomological data. Optical sensors and lidars [92, 93, 94] using near-infrared light have the potential to observe a much larger portion of insect population in real-time with high accuracy and precision. Recent studies have made use of optical sensors for entomological research [108, 115, 116, 117, 118, 119]. The following subsections present the different optical systems used in laboratory conditions during the course of this project. It includes two main configurations, the backscatter and extinction configuration.

2.1.1 Entomological Stand-off Optical Sensor (ESOS)

The first type of system used to remotely study insects is the Entomological Stand-off Optical Sensor, referred to as ESOS throughout the manuscript. This system is based on the study of the light backscattered by the target. This system is similar to the well-known lidar system but without range resolution. The layout of the laboratory ESOS is presented in Figure 2.1.

The laboratory ESOS simultaneously records three different channels, from two continuous wave (CW) laser diodes in the short-wave infrared (SWIR) and near-infrared (NIR). The backscattering efficiency in the SWIR spectral range is rather unaffected by water and melanin, which is the most common chromophore in insect wings and body, while NIR backscattering efficiency will vary with the degree of melanization [118]. Thus, the two wavelengths were chosen in the NIR and SWIR spectral ranges in order to maximize the contrast between their respective backscattering intensities. Additionally, the 1320 nm and 940 nm wavelengths of the lasers have been chosen in order to ensure a measurable backscattered intensity

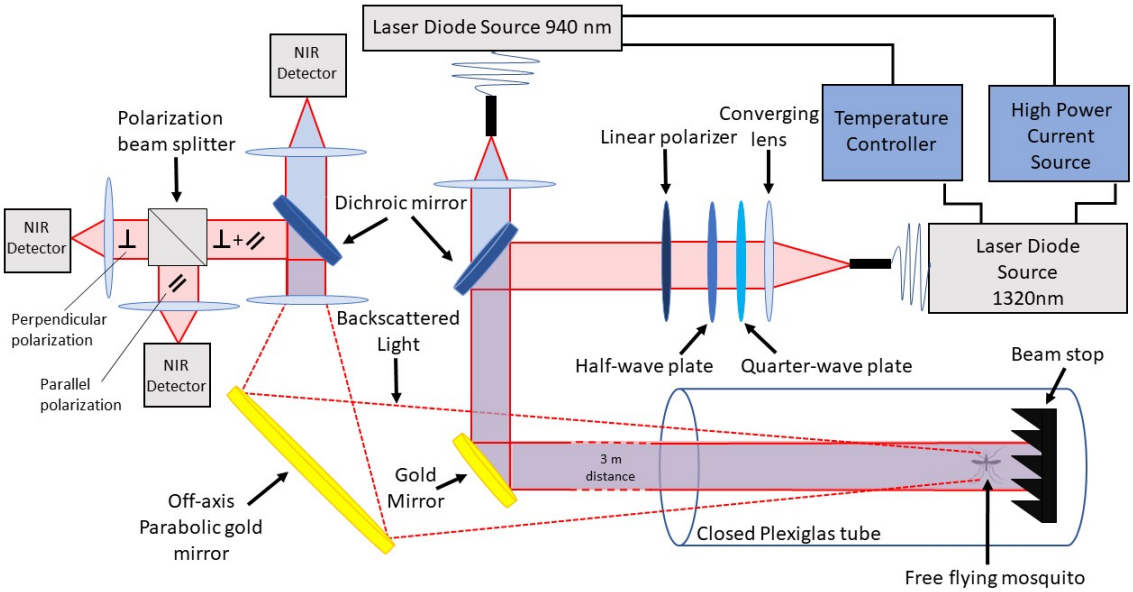

Figure 2.1 Optical layout of the laboratory ESOS system.

from the target while remaining outside of the visual perception of mosquitoes, which are the insect family of focus in this manuscript. Indeed, different wavelengths are known to influence *Cx. quinquefasciatus* mosquitoes [120]. The electro-retinogram of the female *Ae. aegypti* mosquito showed no response for wavelength above 700 *nm* [121]. The wavelength dependency for the absorbance of *An. gambiae* and *An. arabiensis* was measured in Mayagaya et al. 2009 [122]. This study allows for an educated guess on the range of wavelengths at which a non-negligible backscattering of the laser beam by the mosquitoes can be expected.

The first channel records the intensity of the light wave backscattered by insects from a 940 nm CW laser diode (L4-9891510-100M, Lumentum, Milpitas, CA, USA), denoted I_{940}. The second and third channels collect the backscattered light from a 1320 nm CW laser diode (4PN-116, SemiNex, Peabody, MA, USA). The 1320 *nm* wavelength is resolved in polarization. A quarter-wave plate is used to convert the elliptical polarization of the laser diode (3.6° of ellipticity) into a linear polarization. The direction of the linear polarization is then rotated by

a half-wave plate to coincide with the axis of the polarization beam splitter of the detection system. Finally, a linear polarizer is used to further enhance the linearity of the polarization. A polarizing beam-splitter cube separates parallel (///) and cross-polarized ($\perp$) backscattered signals, with respect to the laser polarization plane, resulting in channels $I_{1320,/\!/}$ and $I_{1320,\perp}$, which are respectively the second and third channels. As such, the system is rather similar to dual-wavelength polarization sensitive lidars commonly used to monitor and study atmospheric aerosols [98, 99]. Both laser diodes are operating with a maximal optical power of 10 W. For both laser diodes, laser beams are superimposed, using a dichroic mirror, to follow the same optical path and expanded to reach 2.54 cm full-width half-maximum diameter to increase the likeliness of an insect transiting through the laser beams while decreasing the beam divergence, which is kept below 1 $mrad$. The backscattered intensity on all three channels is measured by Indium Gallium Arsenide amplified photodetectors (InGaAs PDA20CS, Thorlabs, Newton, NJ, USA) with a 67 kHz bandwidth at 50 dB gain and recorded with a 16 bit 250 MS/s, 125 MHz bandwidth digitizer (M4i4420-x8, Spectrum, Stamford, CT, USA), data were pre-averaged and acquired at a sampling rate of 30,517 Hz by the acquisition software.

As the insect flies freely through the laser beam the intensity I of the backscattered light is recorded by the detector, as shown in Equation (2.1). This equation can be applied independently to any of the three channels of the laboratory ESOS.

$$I = \frac{K}{d^2} \cdot I_0 \cdot \sigma_{bck} + I_B \tag{2.1}$$

Where σ_{bck} is the backscatter cross section of the target expressed in mm^2, describing the interaction between light and the insect. d is the distance between the insect and the collection optics. K is a constant taking into account the size of the

off-axis parabolic mirror, quantum efficiency of the detector, and optical transmission or reflection coefficients of the transmitting and collecting optics. I_0 is the initial intensity of the laser beam while I_B is the unwanted light intensity measured by the detector from either scattering on the borosilicate glass, from gas molecules, particles in the probed volume of air, or from the beam stopper terminating the optical path. The optical transmission of the air is considered to be negligible over such distance. The contribution of I_B to the recorded signal I can be treated as constant, especially over a short period of time, the average transit time of a specimen through the laser beam being in the order a 100 ms. Therefore, this contribution can be removed by subtracting the average value of the background from the raw data.

Figure 2.2 shows an example of the backscattered signals from an insect transiting through the laser beams of the three channels ESOS. In the absence of a target in the laser beam, the recorded signal is a constant voltage equal to I_B. This baseline value can slowly vary from drift in the laser output power or change in optical backscattering from the air. Howbeit, such slow variations are essentially constant when considered over the short time scale of an insect flying through the laser beam. When an insect enters the laser beam, the recorded intensity increases as the target backscatters some light.

The signal displays an overall Gaussian envelope due in part from the insect progressively entering then exiting the beam and from the spatial profile of the laser beam, which is itself Gaussian. Moreover, sharp intensity peaks in the backscattered signals are visible due to the rapid movement of insects' wings causing the backscatter cross section of the insect to change. The backscatter cross section of the insect's wings periodically alternates between a maximum when the wing plane is normal to the laser optical axis, and a minimum when the wing plane is parallel to the laser optical axis.

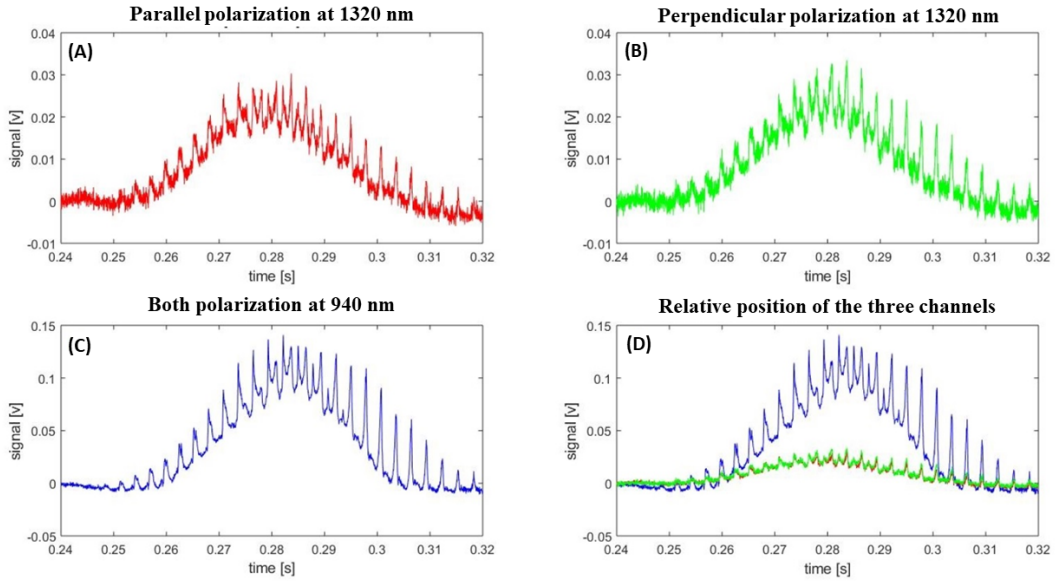

Figure 2.2 Example of an insect signal as it transits through the laser beams of the three channels ESOS system. The presented signals, in volts, are equal to the backscattered intensity I after subtraction of the baseline contribution I_B. (A), (B) and (C) respectively shows the signal recorded by the second, third and first channels. (D) shows all three signals simultaneously.

2.1.2 Entomological Bistatic Optical Sensor System (eBoss)

The second type of system used in this contribution is the Entomological Bistatic Optical Sensor System, referred to as eBoss throughout the manuscript. This system is based on the study of the light absorbed and scattered away from the detector by the target.

The eBoss relies on a low intensity continuous laser diode source (CPS980, Thorlabs, Newton, NJ, USA) with a peak optical power of 5 mW, operating at 980 nm. As shown in Figure 2.3, the laser beam emitted by the laser diode is expanded using a combination of converging lenses, leading to a beam diameter of 4.5 cm full width half maximum. The laser beam is transmitted through a Plexiglas tube (35 cm long, 12.5 cm diameter) located 50 cm away from the output lens. The laser beam enters the tube through a borosilicate glass tilted 5° downward at one end of the tube, and exits at other end of the tube through another borosilicate glass. In the laboratory, studied specimens are introduced in the tube through small apertures in

the tube, covered in mesh to prevent escapes. At the end of the optical path, the light is collected by a converging lens, goes through a bandpass filter and is then focused onto the active area of an amplified, switchable-gain, silicon detector (PDA36A2; Thorlabs, Newton, NJ, USA). The optical signal is recorded at a sampling frequency of 30,517 Hz by a 16-bit digitizer (M4i4420-x8; Spectrum, Stamford, CT, USA).

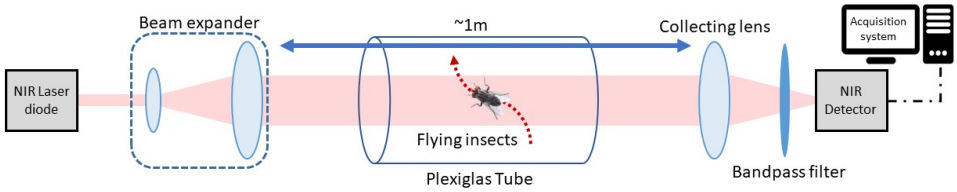

Figure 2.3 Optical layout of the laboratory eBoss.

As the insect flies freely through the laser beam the intensity I of the transmitted light is recorded by the detector, as shown in Equation (2.2).

$$I = I_0 \cdot K \cdot \left(1 - \frac{\sigma_{ext}}{\pi \cdot r^2}\right) + I_B \tag{2.2}$$

Where σ_{ext} is the extinction cross section of the insect, expressed mm^2. Extinction cross sections describe how the incident light is absorbed and scattered by a given target. K is a constant taking into account the size of the collecting lens, quantum efficiency of the detector, and optical transmission or reflection coefficients of the transmitting and collecting optics. r is the radius of the laser beam. I_0 is the initial intensity of the laser beam and I_B the intensity of the light from unwanted sources. The contribution of I_B to the recorded signal I can be treated as constant since there is little to no change in background infrared radiation over a measurement period. Consequently, this contribution can be removed to extract only the information related to the insect target.

Figure 2.4 shows an example of an extinction signal from an insect transiting through the laser beam of an eBoss. In the absence of any target in the laser beam, the recorded signal is a constant voltage corresponding to the constant flux of photons

received by the detector. This baseline value can change by a few % over the course of a few hours due to variation in unwanted light exposure, change in optical extinction of the probed air or small drift in laser power. Whenever an insect flies through the laser beam, the signal decreases as the target attenuates the optical intensity through scattering and absorption. Unlike the slow baseline variation mentioned above, these variations are happening on much shorter time scale, in the order of 20 ms to 1 s and can therefore be discriminated. For similar reasons as the ESOS signal described previously, the signal from an eBoss displays a Gaussian shape and sharp peaks.

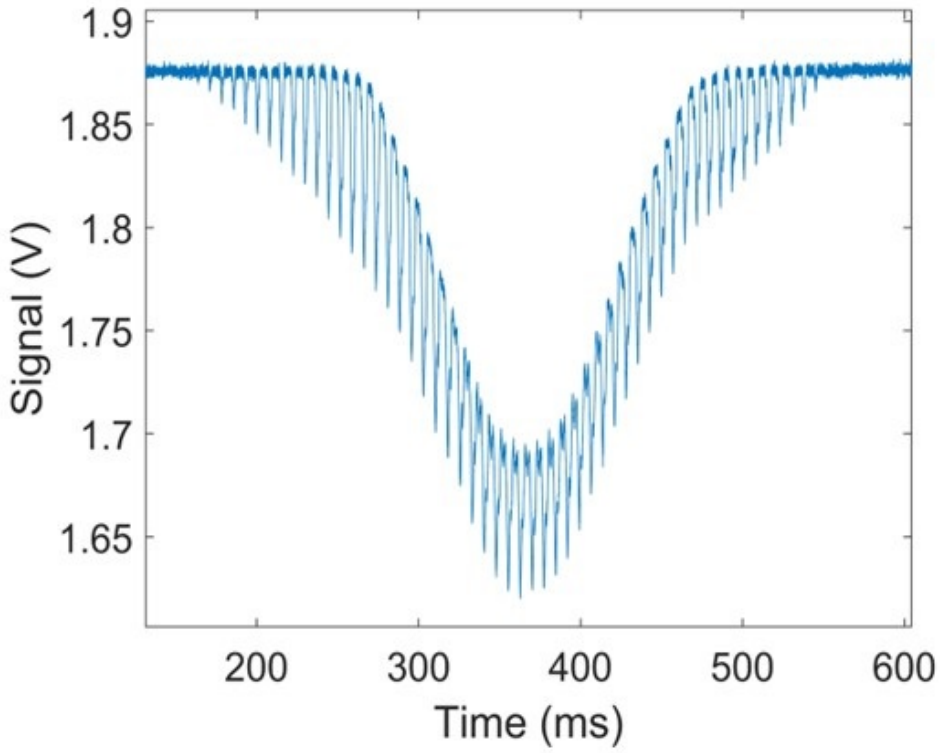

Figure 2.4 Example of an insect signal as it transits through the laser beam of the eBoss. The presented signal in V is equal to the transmitted intensity I.

2.1.3 Selection of the Laser Beam Radius

One of the most important parameters in designing an EPS is the cross section of its laser beam. This parameter will define the range of observable insects as well as the

size of the optics. As such, it is important to determine some objective criteria in order to select the most appropriate laser cross section.

So far, the detection efficiency has been implicitly considered as perfect, meaning that any insects crossing the laser beam would be detected and their characteristics estimated without any bias. In practice, the detection of insects is function of the insect size and laser beam cross section, respectively described by their radius r_I and r_L (in the simplifying assumption that both can be described by a disk in a two-dimensional representation). If the insect is too small compared to the laser cross section ($r_I \ll r_L$), its extinction cross section (or backscatter cross section) may be too low to be detectable by the eBoss (or ESOS, respectively), as the signal to noise ratio (SNR) goes below 1. Hence, specimens of this size would not be detected. A smaller laser cross section may allow for its detection as the energy density of the laser will be greater. On the other hand, if the insect size is comparable or larger than the laser beam cross section ($r_I \geq r_L$), it may be undetected or improperly characterized. In this case, the measured extinction or backscatter cross section of the insect may be underestimated as only a portion of the insect is fully within the laser beam. The signal could even be discarded entirely as a non-insect event due to the absence of amplitude modulation if the wings are not in the laser beam. As illustrated by Figure 2.5, the measured cross section of the insect σ_m may be different that its true cross section σ_I. This will happen if, while crossing the laser beam, the insect does not fully enters it, as shown in Figure 2.5, where only a portion of the insect is within the laser beam. In this case, the measured cross section is an underestimation of the true cross section of the specimen. Evidently, the laser cross section must be chosen so that it maximizes SNR while also allowing for the detection of large insects.

In total there are three cases to consider. First is the trivial case where the insect is completely outside of the laser beam. In this case, the insect is simply not

detectable by the system and its measured cross section is null, $\sigma_m = 0$. The second case is when the insect is completely within the laser beam. In this case, the measured cross section will be equal to the true cross section of the insect, $\sigma_m = \sigma_I$. The third case, which is the most subtle, is when the insect is only partially within the laser beam. In this case, the measured cross section will be smaller than the true cross section of the insect, $\sigma_m < \sigma_I$.

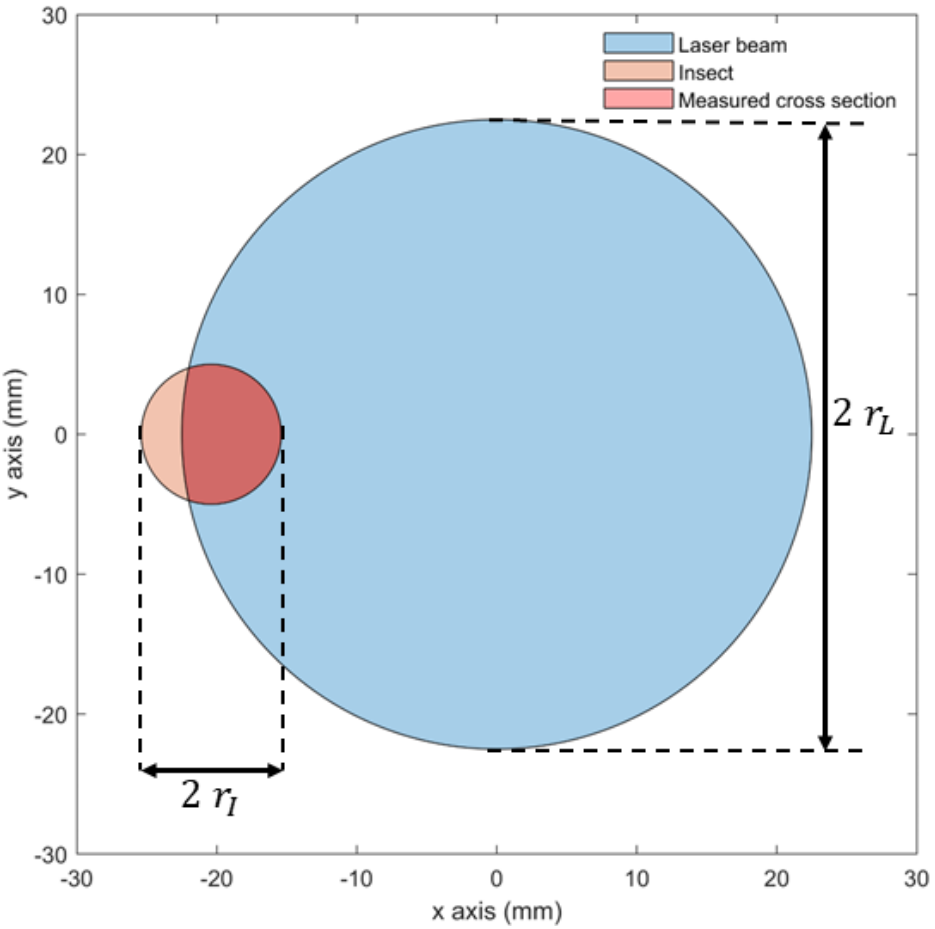

Figure 2.5 Illustration of the concept of measured cross section. In light blue is the surface occupied by the theoretical laser beam. In orange is the surface representing the theoretical insect and in red is the intersection of the two. For the entomological photonic system point of view the true cross section of the insect (orange) is inaccessible and only the measured cross section (red) can be evaluated.

The minimal detectable cross section, σ_{min}, is the cross section for which the SNR is equal to 1. Thus, σ_{min} is function of the noise of measured signal, N_S and the cross section of the laser beam, $\pi \cdot (r_L)^2$.

The measured cross section, σ_m, can be defined in general terms that applies to both ESOS and eBoss system, as shown in Equation (2.3).

$$\sigma_m = \frac{\Delta I}{I_{max}} \cdot \sigma_{max} \tag{2.3}$$

Where σ_{max} is the maximal measurable cross section, which is equal to the cross section of the laser beam, $\sigma_{max} = \pi \cdot (r_L)^2$, since an EPS system cannot measure a cross section that is larger than that of the laser beam. ΔI is the absolute variation of intensity due to the target of cross section σ_m. I_{max} is the maximal value of intensity that can be measured. In the case of the eBoss $\Delta I = I_{max}$ if the target absorbs and scatters away all of the energy of the laser beam. In the case of the ESOS $\Delta I = I_{max}$ if the target backscatters the entire laser beam back into the detector. Those are both extreme cases that, in practice, are very unlikely when observing insects.

The noise level, N_S, of the recorded signal can be evaluated by the standard deviation of the signal over a period with no target within the laser beam. Therefore, the minimum detectable cross section σ_{min} (SNR=1) can be defined by Equation (2.4), which is function of the noise of the system and radius of the laser beam.

$$\sigma_{min} = \frac{N_S}{I_{max}} \cdot \pi \cdot (r_L)^2 \tag{2.4}$$

From Equation (2.4), it is apparent that reducing the noise will lower the minimum detectable cross section. Using a detector with as little noise as possible, a laser with stable power output and reducing the source of parasitic noise should thus be encouraged. Once every and all improvements have been done on these parameters the only remaining degree of freedom is to reduce the size of the laser beam, effectively increasing the energy density, which reduces σ_{min}. Yet, reducing the size of the laser beam also has some detrimental effects. As discussed earlier,

if the size of the insects greatly exceeds the size of the laser beam, then any signal from the specimen is likely to be discarded as a non-insect event by the harmonicity algorithm, defined in subsection 2.2.2. Moreover, any insects larger than the laser beam, even if their signal is properly detected by the harmonicity algorithm, will inevitably see their cross section being underestimated since part of their body or wings will be outside of the laser beam.

Even for insects smaller than the laser beam, the probability of underestimation of their cross section is non zero, as an insect may fly on the edges of the laser beam. The likeliness of being partially within the laser beam is function of the ratio between the true cross section and the laser beam cross section, as shown in Figure 2.6. Consequently, there is a trade-off between a laser beam small enough to detect small insects and a laser beam large enough to reduce the likeliness of underestimation.

In order to formally evaluate the appropriate laser beam radius, a numerical simulation has been designed. In this simulation a laser beam of radius r_L and an insect of radius r_I were simulated, as in Figure 2.5. The position of the insect is randomly and uniformly chosen, with the condition that the center of the insect is within a distance of $r_L + r_I$ from the center of the laser beam. Meaning that if the center of the insect is exactly at this distance, the insect would be completely outside of the laser beam. In any other cases the insect will be either partially within the beam, i.e., underestimated, or completely within the laser, i.e., not underestimated. This process is repeated 1,000 times for several insect and laser beam sizes. For each repetition, the main simulation output is the value of the measured cross section, defined as the intersection between the laser beam and the insect. The final output of 1,000 values is then averaged to obtain the ratio between the average measured cross section and the true cross section $\frac{\sigma_m}{\sigma_I}$. If the insect is never underestimated, then this ratio will be equal to 1. On the contrary, the more the measured cross section is

underestimated (i.e., the more the insect is likely to be only partially within the laser beam) the lower the ratio.

A fit of the simulation results was conducted and the best fit is presented in Figure 2.6. Interestingly, the evolution of $\frac{\sigma_m}{\sigma_I}$ in function of $\frac{\sigma_I}{\pi \cdot (r_L)^2}$ follows an evolution efficiently described by an equation of the Hill–Langmuir form, as defined by Equation (2.5).

$$\frac{\sigma_m}{\sigma_I} = \left[1 + \left(\frac{\sigma_I}{l \cdot \pi \cdot (r_L)^2}\right)^n\right]^{-1} \tag{2.5}$$

Where l and n are parameters determined from a nonlinear least square analysis. The best fit is obtained for $l = 0.46$ and $n = 0.74$ with an $R^2 = 0.99$, and is displayed in Figure 2.6.

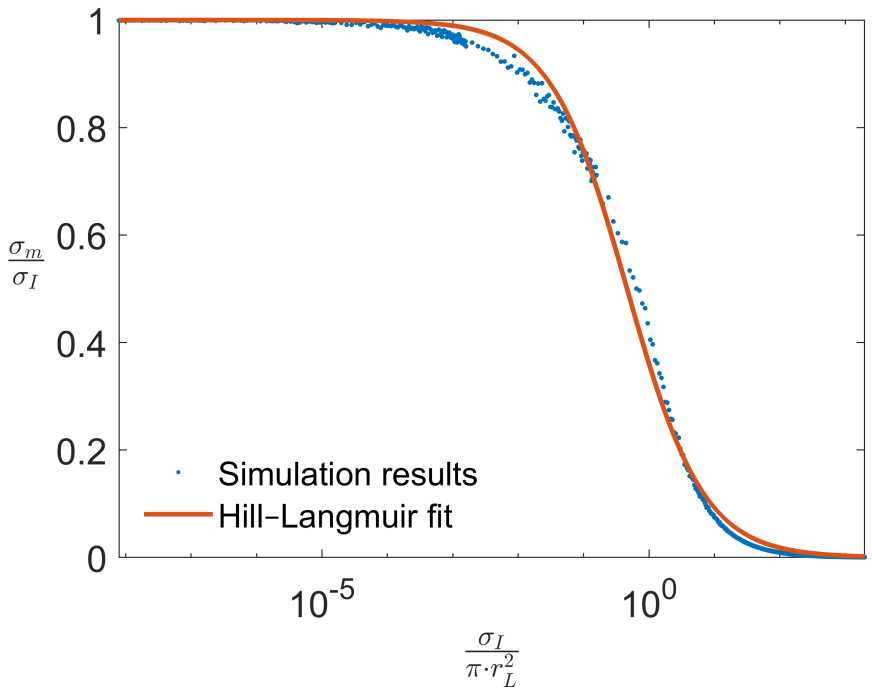

Figure 2.6 Results of the simulation (blue dots) and the best fit using Hill–Langmuir equation (orange line). This represents the value of the average (over 1,000 iteration) measured cross section divided by the true cross section of the insect $\frac{\sigma_m}{\sigma_I}$ in function of the ratio of the true cross section by the cross section of the laser beam $\frac{\sigma_I}{\pi \cdot (r_L)^2}$.

This shows that, the greater the insect size, the greater the underestimation of its cross section. Although, before this simulation result can be used to select the proper laser beam radius another consideration must be formalized. While the results of the fit of the Hill–Langmuir equation efficiently describes the limitation due to the underestimation of the cross section, it does not properly describe the limitation due to low SNR and the fundamental limit imposed by σ_{min}, see Equation (2.4). As such, another function, $H_{\sigma_{min}}$, is needed to describe the lower bound limit due to σ_{min}. $H_{\sigma_{min}}$ can be simply described by an Heaviside step function translated by a value of σ_{min} as defined by Equation (2.6). Which means that the measured cross section is 0 if the insect is below the minimum detectable cross section, which is a realistic description.

$$H_{\sigma_{min}}(\sigma_i) = \begin{cases} 1, & \text{if } \sigma_I > \sigma_{min} \\ 0, & \text{if } \sigma_I \leq \sigma_{min} \end{cases} \tag{2.6}$$

Combining the two functions together will provide an effective way to choose the laser beam size best suited for the type of insect that needs to be studied. Figure 2.7 presents three possible scenarios that differs only by the size of their laser beams. The noise level, used to estimate σ_{min}, is considered identical regardless of the size of the laser beam. Briefly, it is considered that the exact same laser source in used in all three cases but with different beam expanders that do not contribute to the noise. As for the value of the noise level, it was chosen to be equal to the noise level observed for the eBoss in laboratory conditions.

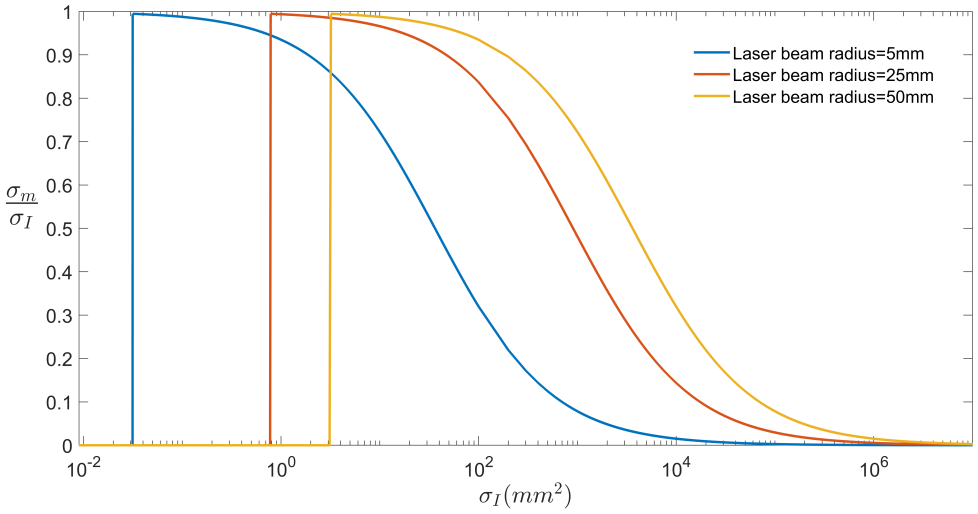

Figure 2.7 Evolution of the ratio of the averaged measured cross section by the insect cross section in function of the size of the insect. Three scenarios are being considered, where the radius of the laser beam is 5 mm (blue), 25 mm (orange) and 50 mm (yellow).

In the first case, $r_L = 5\ mm$, the laser beam is chosen as relatively small. This radius would allow for the study of the smallest insects, as small as 0.3 mm^2. Although, with such laser a beam, insects that are larger than 35 mm^2 would have their cross section underestimated by more than 50%, on average. As such, this size of laser beam would be suited if the study was limited to the smallest insects, such as mosquitoes, but would otherwise miss many other insect families.

In the second case, $r_L = 25\ mm$. With such radius, the minimal detectable cross section is a little below 0.8 mm^2. Thus, it is capable of detecting most smaller insects. In this configuration, insects with a size of up to 200 mm^2 can be studied with less than 25% of average underestimation and 50% average underestimation for insects up to 900 mm^2. For these reasons, this size of laser beam is well suited for the study of many of the most abundant insect species.

In the third case, $r_L = 50\ mm$, the laser beam is chosen as relatively large. In this case, only insects larger than 3.1 mm^2 can be detected. While this a minimal value is smaller than the cross section of many insects, including many mosquito species, theoretically allowing for their study, there is another limitation to consider.

In order for an insect event to be detected, the signal modulation due to the wing needs to be observable. Consequently, even if an insect may be large enough for its cross section to be above 3.1 mm^2, a small insect may not have a wing cross section sufficiently large for their influence on the signal to be properly detected, i.e., low SNR in the frequency spectrum, leading to higher probability of non-detection. On the other hand, this system allows for the study of much larger insects, which in some cases may be desirable. Insects with cross section of 3600 mm^2 can be studied with 50% of average underestimation, which includes most insects species.

While an average underestimation of 50% may seems as a significant bias, it is important to consider other sources of variation in insects' cross sections. The cross section of an insect can change depending on the insect orientation within the laser beam. Unlike in this simplified simulation, where insects are considered as a sphere in the 3D space and a disk in the 2D space, insects have complex morphological characteristics which have different true cross sections depending on the angle of observation, which will contribute to the uncertainty, or variability, of the measured cross section.

2.2 Signal Processing

In order to extract insect signals, also referred to as events, some analysis of the raw data is required. Regardless of the configuration, backscatter or extinction, insect events represent only a fraction of the total signal being recorded. Most of the time there is no insect flying through the laser beam, in which case the recorded signal is relatively constant. Thus, there is a need for the identification of the period of time during which an insect is interacting with the instruments, also called detection of insect signals. For both the ESOS and eBoss, the methodology of event detection is fairly identical. Unless explicitly mentioned otherwise, every step described in the rest of the manuscript applies to both systems.

2.2.1 Event Detection

Figure 2.8 displays an example of 400 s of raw data recorded by the eBoss, showing the noise and baseline variations. Figure 2.8 also shows a magnification of one of the insect events in the raw time series. Events are characterized by a sudden drop (or increase for ESOS) in signal that last a few milliseconds, followed by a sudden increase (or decrease for ESOS) returning the signal to its baseline value. Events from insects are automatically identified by a two-steps process, the first step consists in identifying a region of interest where an object is likely to have crossed the laser beam, while the second step is used to identify the amplitude modulation typical of an insect. The first step is based on a threshold detection algorithm. First, the signal is filtered by a 10-900 Hz digital band-pass filter that removes the contributions of slow baseline drift with a period above 100 ms and high frequency noise above 900 Hz. This also ensures that events identified by the threshold detection algorithm have frequency components in the frequency range matching those of flying insects. The threshold value is determined by a combination of the sliding average of the signal and the sliding standard variation as defined by the set of Equations (2.7), (2.8) and (2.9). Whenever the recorded signal crosses the threshold, the signal is detected and extracted for further analysis.

$$T_r(k) = m_0 \cdot S_{avg}(k) + s_0 \cdot S_{std}(k) \tag{2.7}$$

Where $T_r(k)$ is the value of the threshold at point k. $S_{avg}(k)$ is the sliding average centered around point k and $S_{std}(k)$ the sliding standard deviation centered around point k, respectively defined by Equations (2.8) and (2.9). m_0 and s_0 are respectively the weight of the sliding average and of the sliding standard deviation.

$$S_{avg}(k) = \frac{1}{N} \cdot \sum_{i=1+k-\frac{N}{2}}^{k+\frac{N}{2}} I(i) \tag{2.8}$$

$$S_{std}(k) = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1+k-\frac{N}{2}}^{k+\frac{N}{2}} \left(I(i) - S_{avg}(k)\right)^2} \tag{2.9}$$

$I(i)$ the signal intensity at point i and N is the number of points in the considered sliding window.

The threshold is fine-tuned by modifying the coefficients m_0, s_0 and N. Evaluating the efficiency of the threshold value was done manually by applying different combinations of the aforementioned coefficients until satisfactory results were achieved. The goal being to detect every true insect signal while including as few false insect signals as possible. For both systems, m_0 is usually chosen as equal, or close to one as the threshold should hover around the mean value of the recorded signal. For ESOS, s_0 is chosen positive in order to have the threshold slightly above the mean value. For the eBoss, s_0 is chosen negative so that the threshold will be slightly below the mean value. In both cases N is chosen large enough so that the threshold does not follow every short time variation of the noise but also small enough so that it does follow slower large scale baseline variations.

The first step of the event selection process is purposefully chosen to be permissive, meaning that it is designed not to miss any true events even if it means including false events in the process. A non-negligible number of false events (i.e., not caused by a flying insects) are expected to be extracted from the threshold detection algorithm. However, most of those false events will be discarded through a second step in the analysis procedure. This second step focuses on identifying periodic components in the 10-900 Hz range that correspond to the rapid movement of the insect's wings, more details is given in subsection 2.2.2.

In most cases, only a portion, local extremum, of the signal will cross the threshold. Hence, before being extracted for further analysis, the actual beginning and end of the event must be determined. This allows for the extraction of only the

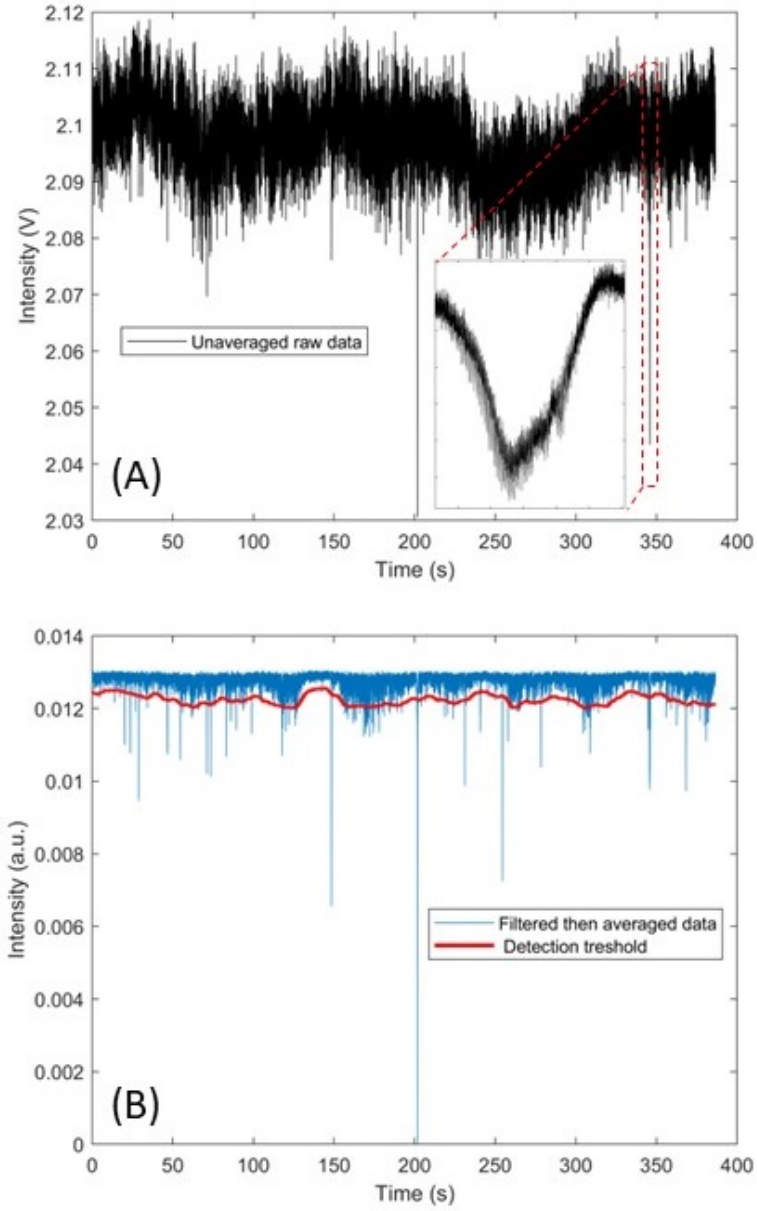

Figure 2.8 (A) shows 400 *s* of raw data, as recorded by the acquisition system with a sampling frequency of 30,517 Hz. A magnification of one of the identified insect events is displayed as illustration. (B) shows the same data after application of a digital bandpass filter [10 900 Hz] and averaging. The red line indicates the detection threshold. Any part of the filtered signal that crosses the threshold is identified as a region of interest and the corresponding raw signal is then extracted for further analysis.

relevant part of the signal and for the determination of the duration of the event, also referred to as transit time, which is the time it took for the target to crosses the laser beam. To find the beginning and end of an event, another threshold is defined with different m_0, s_0 and N parameters. Simply put, this transit time threshold is constructed so that it hovers very closely below (above for ESOS) the normal noise of the signal. As soon as the signal decreases (increases for ESOS) above the noise and crosses this threshold, the event beginning is found. The end of the event is found on identical consideration, when the signal return to noise levels. Beginning and end of an insect event are illustrated by the green and red dashed lines of Figure 2.16.

Every event that crosses the detection threshold have their beginning and end formally identified, defining their transit time, usually in the order of 100 ms. The raw data corresponding to the event, not filtered or averaged, are then saved in a separate individual file with a unique identifying number. Along with the raw data, some meta-data information about the event are saved. Those meta-data are the defining characteristics of the event, such as the transit time or time of observation, to name a few.

2.2.2 Wingbeat Frequency

Once a possible event has been detected by the previously defined event detection threshold algorithm, its frequency components can be studied. This process serves two purposes. First it allows for the removal of non-insect signals. Second it gives access to the wingbeat frequency of the insect specimen, which is one of the fundamental characteristics of atmospheric fauna.

To extract information on the frequency component, the possible insect signal is converted from the time domain to the frequency domain. This process is conducted by applying a Fast Fourier Transform (FFT). The FFT algorithm efficiently computes the discrete Fourier transform of the signal from which insect and non-insect

events can be discriminated. Indeed, well-defined periodic modulations of the signal amplitude are typical of signals originating from flying insects and not simply from a leaf, pollen or large aerosols passing through the laser beam. Although pollen or leaves are unlikely to cross the laser beam in laboratory conditions, this process has been designed with the future real-world experiment in mind. Figure 2.9 presents an example of two recorded events showing one from an insect with its typical periodic amplitude modulations [Figure 2.9 (A) and (C)] and one from a non-insect object [Figure 2.9 (B) and (D)].

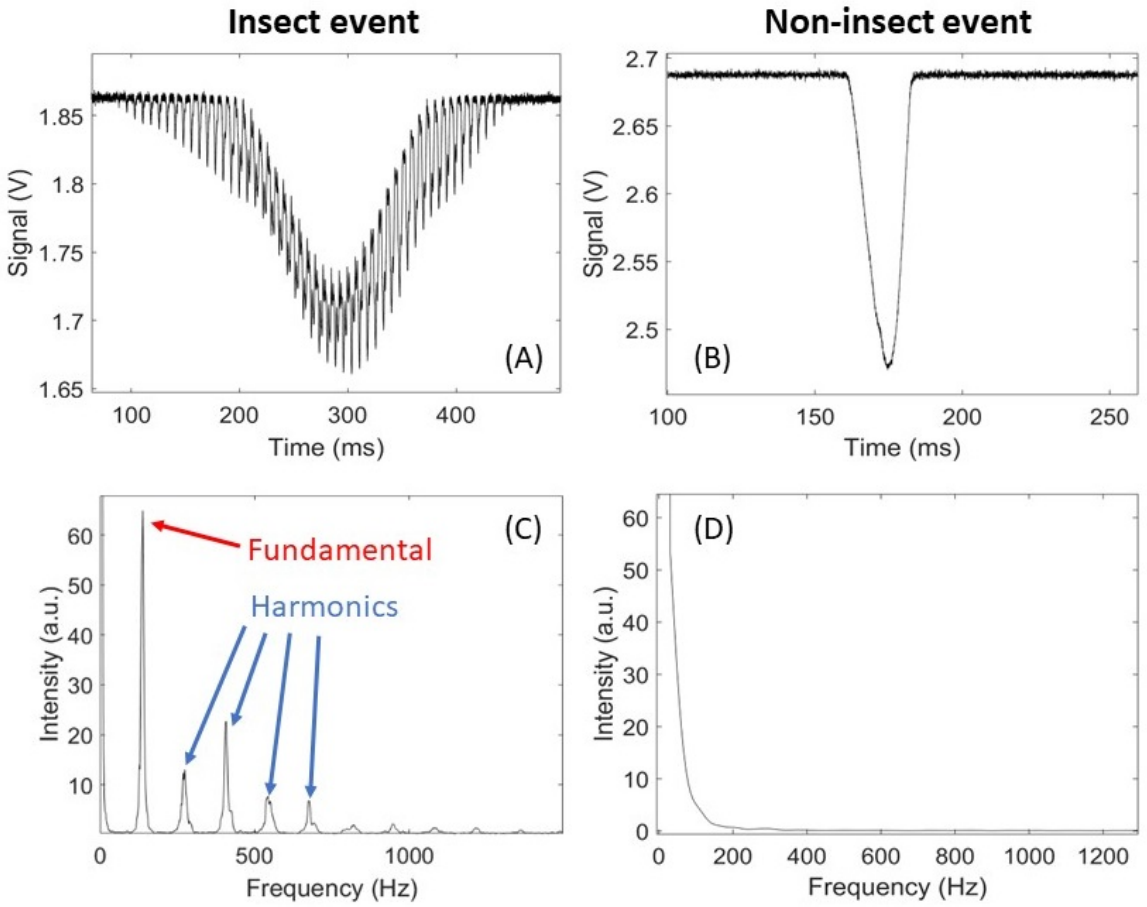

Figure 2.9 Example of two types of signals and their corresponding frequency spectrums. (A) is an example of signal caused by an insect showing clear periodic amplitude modulation. (C) shows its associated Fast Fourier Transform, from which the insect wingbeat frequency can be determined. On the contrary, (B) is an example of a signal from a non-insect target crossing the laser beam, which does not have any periodic amplitude modulation, as can be seem on its Fast Fourier Transform (D).

In order to discriminate insect and non-insect event, an analysis of the frequency spectrum is necessary. Insect events are characterized by the presence of a fundamental frequency along with one or more harmonics. Identifying the fundamental frequency from the frequency domain is not trivial. An algorithm was designed to determine the presence or absence of frequency components and to identify the fundamental, i.e., the wingbeat frequency. Referred to as the harmonicity algorithm, this process consists of three main steps. First, the frequency spectrum is pre-processed by applying a sliding average which reduces the noise while preserving the frequency resolution. As shown in Figures 2.10 and 2.11, this process greatly reduces the noise, allowing for an easier detection of the fundamental and harmonics. Second, the pre-processed frequency domain is scanned for peaks that exceed a user define predominance (i.e., identifying local maxima with sufficiently high values). From the local maxima, the central frequency of each peak is determined. Third, each of those peaks are temporarily considered as the potential fundamental and their possible harmonics are searched among every other selected peak(s) of greater frequency. Searching for harmonics is done by looking if any other selected peak is equal to an integer multiplier of the considered fundamental, within a reasonable error margin. This consists of finding if any other peak satisfies the harmonic condition, as defined by Equation (2.10).

$$f - f_0 = z \cdot f_0 \pm \epsilon \tag{2.10}$$

Where f_0 is the frequency of the peak being considered, f the frequency of the eventual harmonic it is being compared with, z is a positive integer and ϵ is a small positive value equal to a few percent of f_0.

If no peaks exceed the user defined predominance value, or if none of the peaks that pass this condition have harmonics, then the event is rejected as a

non-insect event. Otherwise, the peak that has the most harmonics is selected as the fundamental. The harmonicity algorithm is robust in the presence of outliers, as shown in Figure 2.10. It is also performing well even for low SNR event, as shown in Figure 2.11.

The harmonicity algorithm is capable of determining the appropriate fundamental frequency even if the frequency spectrum is polluted by an outlier, such as the 71 Hz peak shown in Figure 2.10 (A). In the example presented in Figure 2.10 (B), the 116 Hz peaks has two harmonics while the other peaks have none. Consequently, the event associated with this FFT will be classified as an insect of wingbeat frequency of 116 Hz. Despite the presence of a large peak at 71 Hz, the harmonicity algorithm still converges towards the appropriate fundamental. While the source of this 71 Hz peak has not been formally determined, it is likely originating from an unwanted source of infrared radiation in the vicinity of the instrument.

Insect events with low SNR are to be expected, in particular in real-world field experiments, and as such the harmonicity algorithm needs to be able to handle such cases. As shown in Figure 2.11 even when the SNR of the frequency spectrum is relatively low, the harmonicity algorithm still converges towards the correct fundamental frequency of 158 Hz. In this example, the peak at 158 Hz has three harmonics, the one at 323 Hz has one and every other has none. As the peak with the most detected harmonics, the peak at 158 Hz is selected as the fundamental.

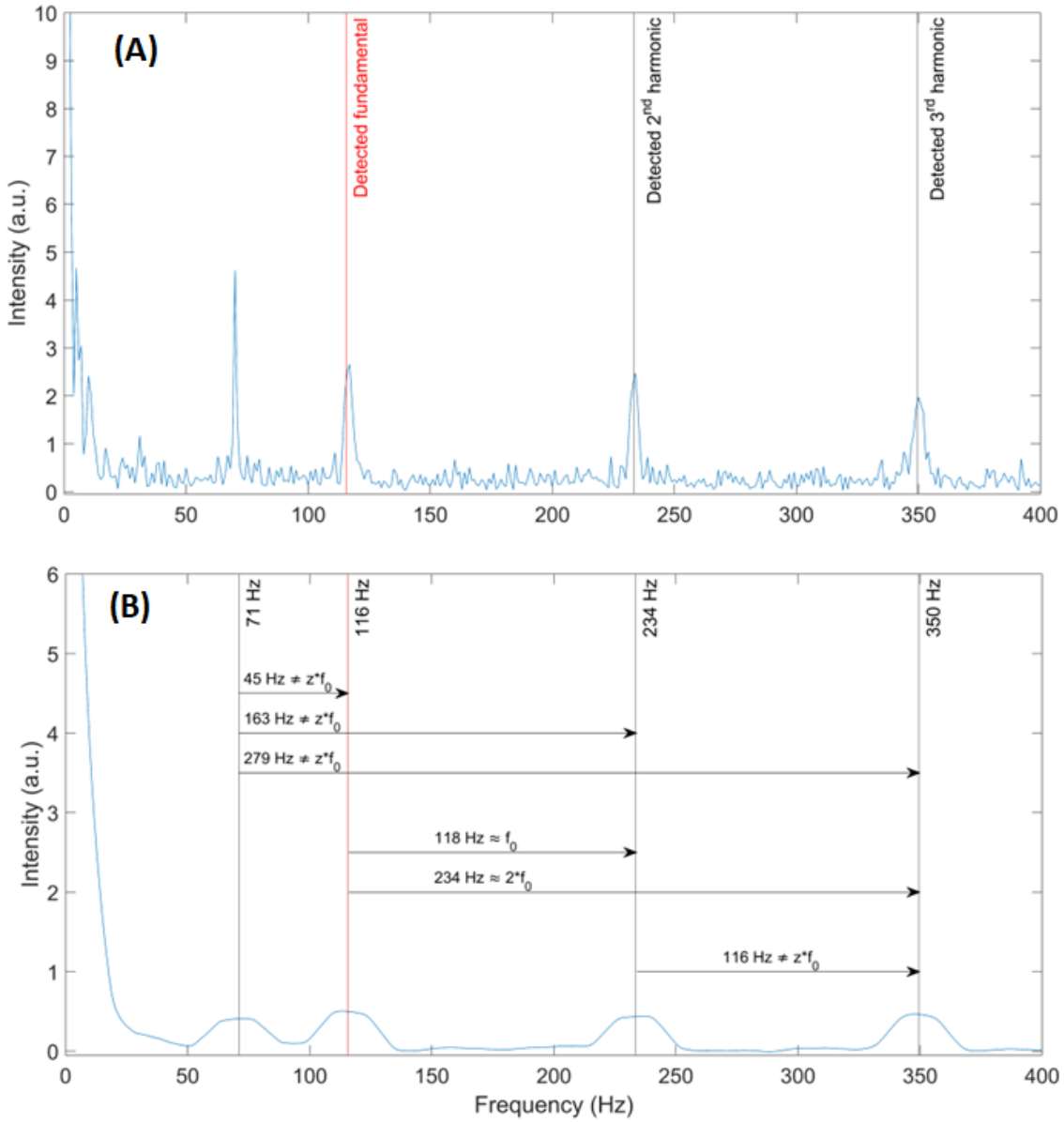

Figure 2.10 (A) shows the raw frequency spectrum of an insect event. (B) shows the same frequency spectrum after pre-processing. The red line indicates the position of the fundamental peak, selected by the harmonicity algorithm. For (A), black lines indicate the position of the detected harmonics. For (B), black lines indicate the position of the other peaks that have been considered as potential fundamental. If the difference between two peaks is close enough to an integer multiplier of the fundamental then a harmonic of the fundamental is found. As such, the peak at 116 Hz has two harmonics and all other peaks have no harmonics.

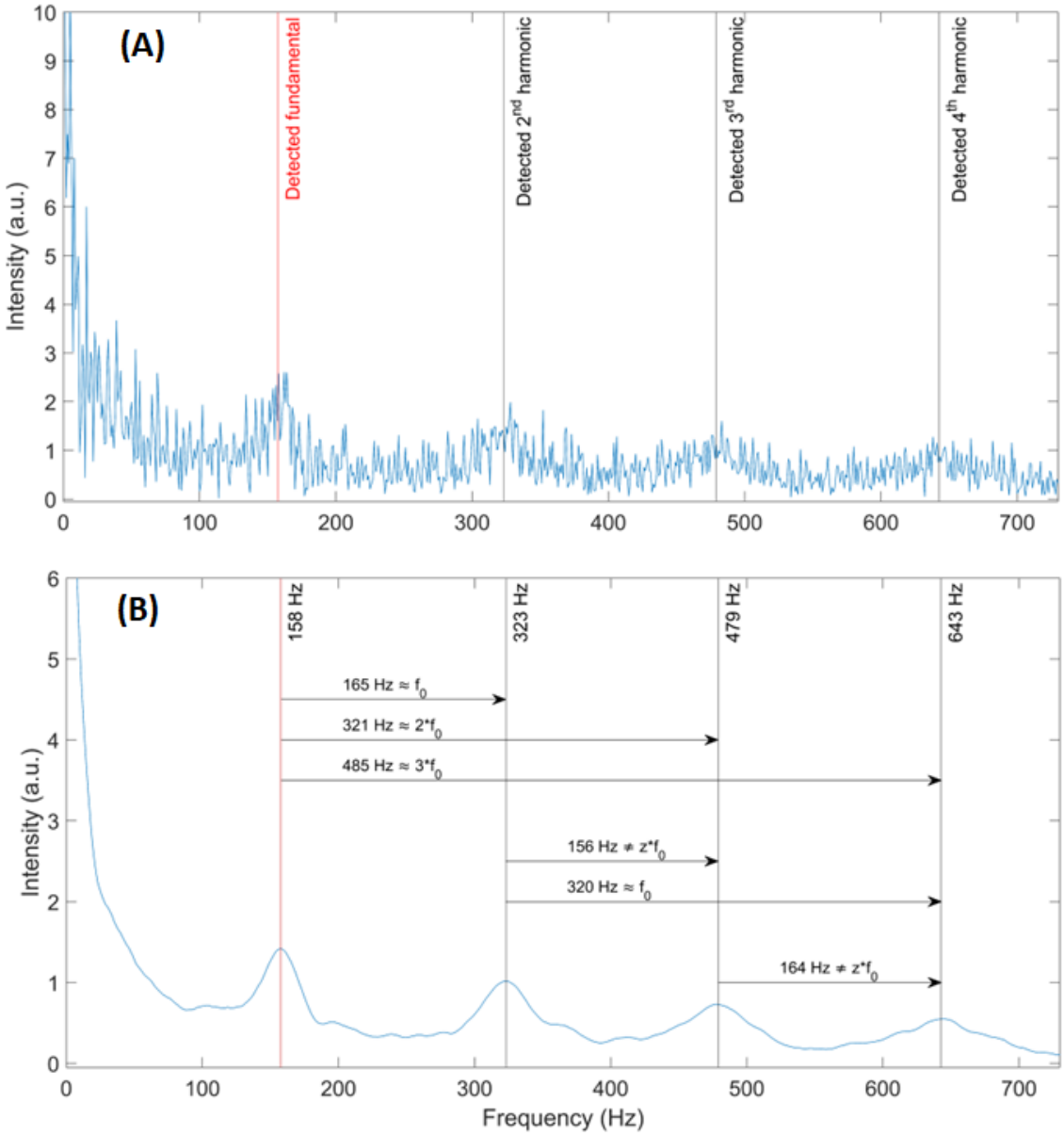

Figure 2.11 (A) shows the raw frequency spectrum of an insect event. (B) shows the same frequency spectrum after pre-processing. The red line indicates the position of the fundamental peak, selected by the harmonicity algorithm. For (A), black lines indicate the position of the detected harmonics. For (B), black lines indicate the position of the other peaks that have been considered as potential fundamental. If the difference between two peaks is close enough to an integer multiplier of the fundamental then a harmonic of the fundamental is found. As such, the peak at 158 Hz has three harmonics, the one at 323 Hz has one and all other peaks have no harmonics.

2.2.3 Mel Frequency Cepstral Coefficient (MFCC)

Mel frequency cepstral coefficients (MFCCs) are features that can be extracted from a signal and that describe its harmonic content. Originally intended for the analysis of sound signals, the MFCC methodology can be extended to optical signals such as the ones recorded by an EPS. The recorded intensity of a sound does not fundamentally differ from the recorded optical intensity from a photodetector. MFCC are commonly used in speech and voice recognition [123, 124] and are considered as particularly powerful. They can represent the frequency spectrum in a compact, low-dimension form (a few coefficients only).

Briefly, those coefficients are obtained from the discrete cosine transform of the log power spectrum from the nonlinear Mel scale, as illustrated by Figure 2.12. One of the significant advantages of the MFCC is that their wide spread use makes them particularly simple to implement, both Matlab and Python have dedicated script for the calculation of those coefficients.

While calculating MFCC, different bank filters could be considered. In this contribution the defaults bank, as shown in Figure 2.12, of the Matlab 'mfcc' function was chosen. In this bank, the first 10 triangular filters (10 with lowest frequency) are linearly spaced and the others are logarithmically spaced.

The fundamental idea in extracting the MFCCs of an insect signal is to find an efficient way to study the frequency spectrum. While the fundamental frequency (i.e., wingbeat frequency of the specimen) is an example of analysis of the frequency spectrum, it does not encapsulate every possible information contained within the spectrum. More information than the fundamental remains hidden within the complex frequency spectrum.

As an example, the A440 Hz (music note) played from a guitar and from a piano have the exact same fundamental frequency. Yet, it is clear for a human ear that the two sounds are different. It is because they have different timbre, which

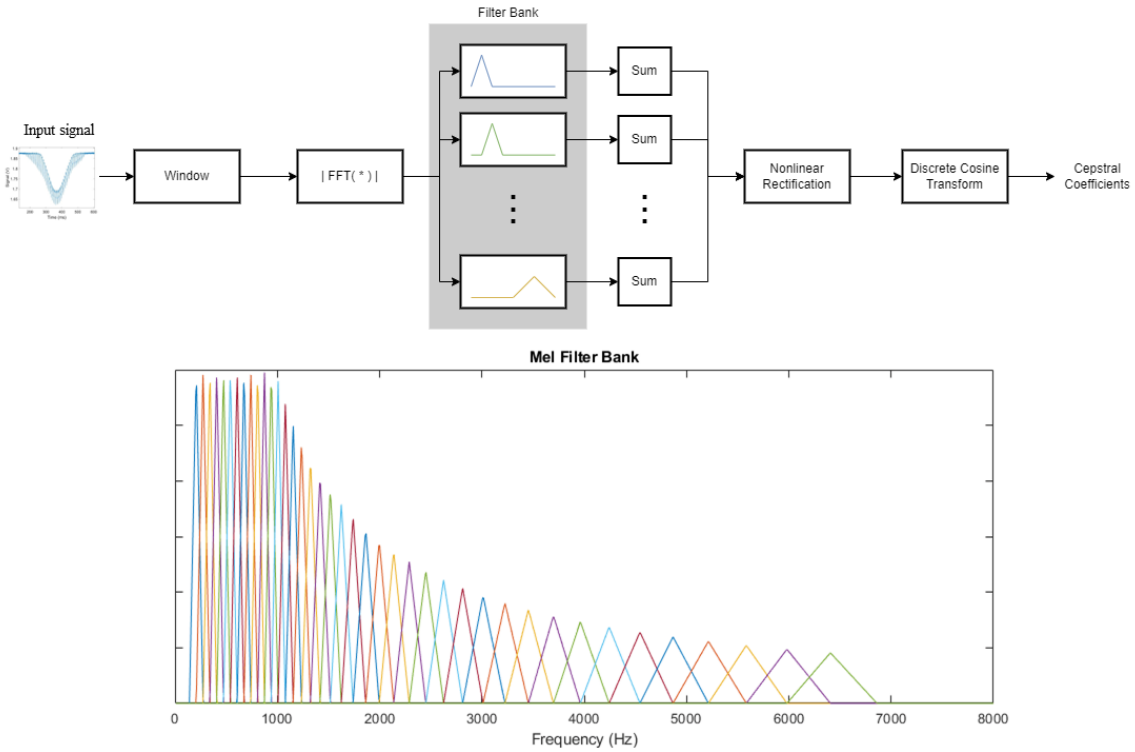

Figure 2.12 Simplified illustration of the method used to calculate the Mel Frequency Cepstral Coefficient, using the default Mel Filter Bank of the Matlab mfcc function. The input signal, which in this example is a signal from the eBoss, is windowed (in this application a single window including the entire event is used) then its FFT is calculated, the filter bank is applied. After summation and a nonlinear rectification, the discrete cosine transform is taken.

is a musician's way of naming a difference in the frequency spectrum of the two sounds. The relative strength of the different harmonics, in the frequency domain, can vary and the repeated pattern, in the time domain, can be different. MFCC are a condensed way to evaluate such differences with a reasonable number of parameters so that it can be used in machine learning algorithms.

As an insect flies through the laser beam, of either of the two presented EPS, the optical signal will display a periodic amplitude modulation that has been described in detail in subsections 2.1.1 and 2.1.2. The periodic modulation displays an overall peak-like shape, however this is but a simple description. When studied in more detail, the shape of the repeated pattern changes from one event to the next, as illustrated in Figure 2.13. Those differences induce changes in the frequency spectrum and can therefore be studied with the help of MFCC to retrieve information on the flying specimen.

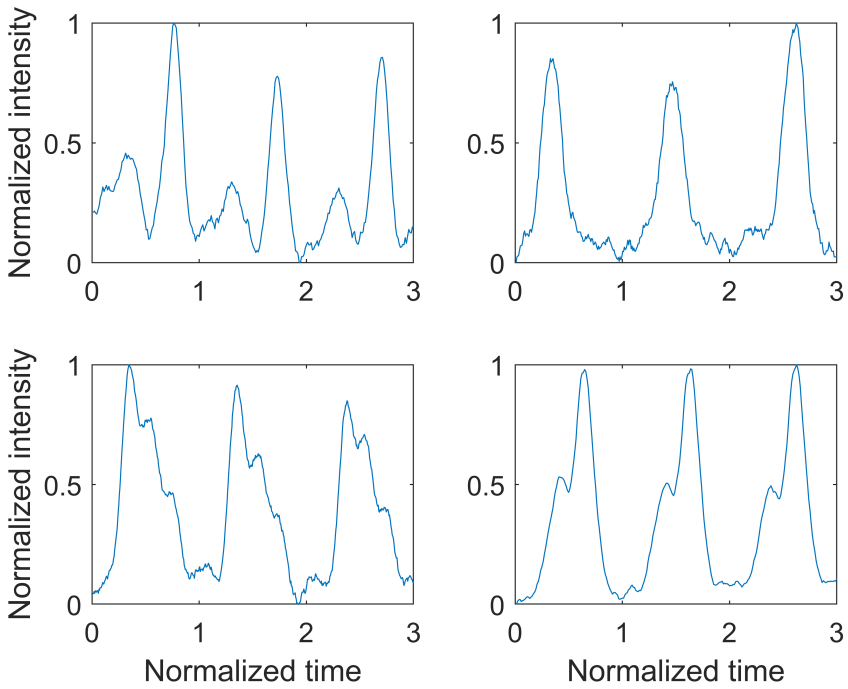

Figure 2.13 Examples of periodic amplitude modulation of the signal originating from different insect specimens. Each example is a magnification of the periodically repeated pattern within the wing contribution. Each figure presents three full wingbeat cycles. The difference in the shape of the repeated pattern can potentially provide information on the sex or species of the specimen through MFCC analysis.

2.2.4 Cross Section

Wingbeat frequency is the first and most fundamental insect characteristic that can be extracted from optical signals. The second characteristic that comes to mind is the size of the specimen. Indeed, insects size varies across species and is an interesting feature to consider.

The ESOS system, in its current form, is unable to retrieve such information outside of very specific laboratory conditions. Due to its lack of range resolution, the magnitude of the recorded intensity cannot be directly converted into backscatter cross section. As presented in Equation (2.1), the intensity of the ESOS depends on both the distance and the backscatter cross section of the insect specimen. As such, without range information it is impossible to discriminate between a small target at short distance versus a larger target at a longer distance. Designing a small enclosure allows, in laboratory conditions, to consider the distance as a known constant and resolve this problem. On the other hand, such measurement would be impossible in real-world experiments, reducing the interest of designing such laboratory experiment.

An alternative system, using continuous wave and capable of range resolved measurements does exist, the Scheimpflug lidar system [113, 125]. The principle of this system is described in Figure 2.14 extracted from Brydegaard et al. 2014 [126]. However, in order to achieve a sufficiently fine range resolution, the detector (CCD) needs to be composed of several individual channels. This imposes a technical limitation that limits the maximal duration of any field experiment. Due to the high number of channels and the requirements in terms of sampling frequency, the amount of data recorded can be in the order of terabyte per day. As a result, using a Scheimpflug lidar for extended periods of time, i.e., months or years, is prohibitively expensive. Furthermore, the system complexity is greater than either ESOS or eBoss systems, making its maintenance and operation more complex. Yet such a system

has shown great results for short term studies of insects activity, such as the behavior of flying insects during a solar eclipse [108].

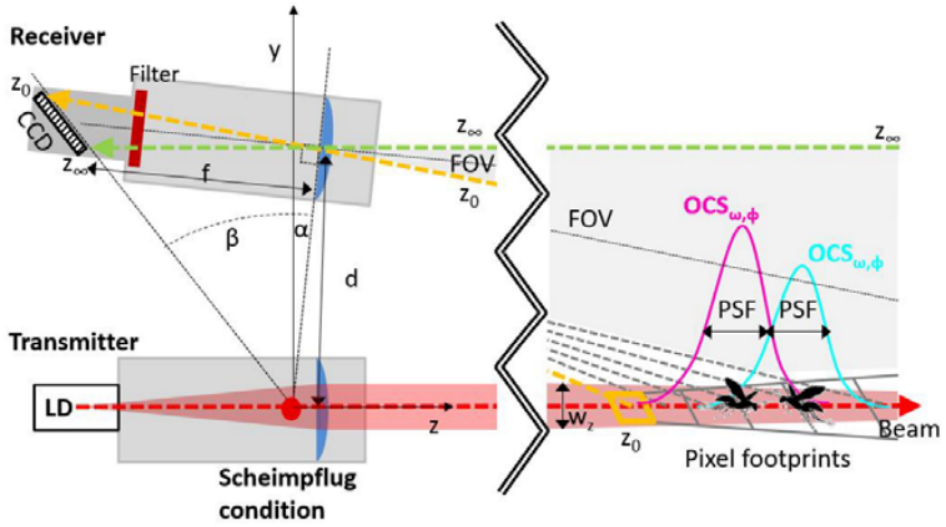

Figure 2.14 Illustration of the Scheimpflug lidar system. The plane of the CCD receiver is tilted. Therefore, the position where the light focuses is function of the target distance z. As such, every pixel of the CCD receiver corresponds to a given range of distance (i.e., pixel footprint), starting with the minimal observable distance z_0 up to infinity z_∞.
Image credit: Brydegaard et al. 2014 [126].

Unlike the ESOS, the eBoss is not limited by its inability to resolve range. Indeed, the eBoss with its extinction configuration does allow, with reasonable approximations, to determine the insect extinction cross section (ECS), even outside of laboratory conditions. ECS of a target crossing the laser beam is denoted as σ_{ext} and expressed in mm^2. The ECS is derived from the drop in voltage of the signal by a simple calibration procedure. The calibration constant is determined from the known diameter of the laser beam and confirmed using opaque (near 100% extinction) chrome steel spheres of known sizes dropped through the laser beam, which allows for the estimation of the relationship between intensity and ECS, see Equation (2.11).

$$\sigma_{ext} = E \cdot \frac{I_B - I_D}{I_B} \tag{2.11}$$

As shown in Figure 2.15, I_B is the intensity of the baseline, I_D is the value of the maximal signal decrease while σ_{ext} is the ECS of the target and E the aforementioned calibration constant (in mm^2).

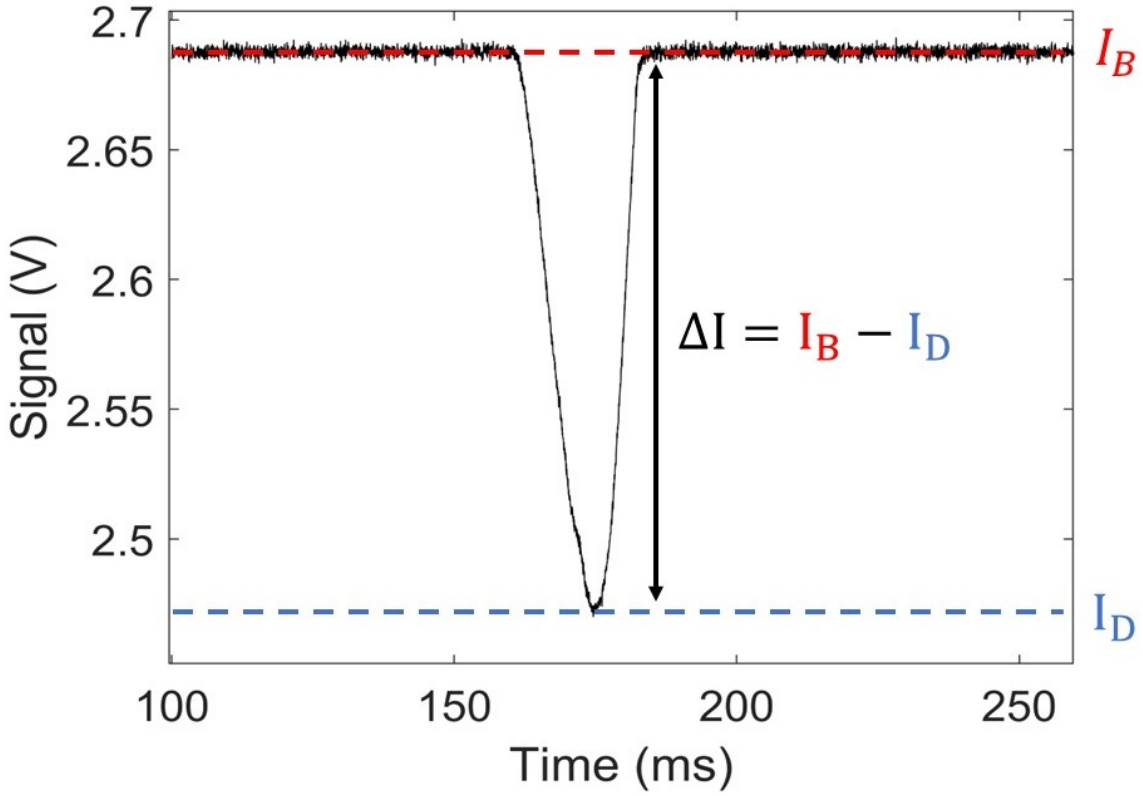

Figure 2.15 Signal due to an opaque chrome steel sphere of known size falling through the laser beam of the eBoss. The black arrow indicates the amplitude of the signal decrease that is due to the sphere. I_B is the value of the baseline and I_D is the value taken by the signal when the sphere is completely in the center of the laser beam.

2.2.5 Wing and Body Discrimination

As mentioned earlier, an insect signal is characterized by amplitude modulation due to the wing orientation rapidly changing overtime. In addition to providing an estimation of the wingbeat frequency, this feature can also be used to separate the contribution of the wings of the insect from the rest of the signal, referred to as body contribution. The term body contribution must be understood as every part of the

insect that does not move periodically (such as the wings) and not in the pure sense of the term, which would not include the insect legs or antennas.

For the ESOS, the total increase (or decrease for eBoss) in intensity due to the insect I is equal to the sum of the intensity due to the wing I_w and the intensity due to the body I_b, $I = I_w + I_b$. Indeed, the optical power measured by a detector can, within reasonable range of power (i.e., without any non-linear phenomenon), be considered additive. The following explanation applies to the ESOS, but an almost identical consideration can be done for the eBoss.

As detailed in Equation (2.1), the recorded intensity, after removal of the background contribution I_B, is proportional to the backscatter cross section of the insect, σ_{bck}. But, this simplified static equation is only applicable to static targets completely within the laser beam. When considering dynamic phenomenon (i.e., insect flying through the laser beam) this equation needs to be modified to include non-static considerations, see Equation (2.12).

$$I(t) = \frac{K}{d^2} \cdot I_0 \cdot [p_w(t) \cdot \sigma_{bck,w}(t) + p_b(t) \cdot \sigma_{bck,b}(t)] \tag{2.12}$$

Where K, d and I_0 have been previously defined. $I(t)$ is the recorded intensity in function of time t. p is the position coefficient that varies with the position of the insect within the laser beam cross section, i.e., $p = 0$ when the insect is outside the beam and $p = 1$ when the insect is at the center of the beam (maximal energy density for a Gaussian spatial profile). σ_{bck} is the backscatter cross section of the insect as observed from the detector. Index w and b are respectively referring to the aforementioned wing and body contributions.

To separate the wing and body contributions, a few assumptions are required. First, it is assumed that the position coefficient is identical for the wing and the body at every moment (i.e., $p_w(t) = p_b(t) = p(t)$). It is considered that insects are

small enough for the distance between wing and body to be negligible. Second, it is assumed that the backscatter cross section of the body is independent of time (i.e., $\sigma_{bck,b}(t) = \sigma_{bck,b} = \text{constant}$). Unlike the wings that move periodically and therefore change their backscatter cross section within the plane of the detector, it is assumed that the insect body does not change its orientation with time and that the body backscatter cross section within the plane of the detector does not vary. After taking into account those assumptions, Equation (2.12) can be simplified into Equation (2.13).

$$I(t) = \frac{K}{d^2} \cdot I_0 \cdot p(t) \cdot [\sigma_{bck,w}(t) + \sigma_{bck,b}] \tag{2.13}$$

The wing backscatter cross section is function of time due to the periodic change in orientation. As shown in Equation (1.11), the wing backscatter cross section is function of the wing projected area within the plane of the detector, as well as the wing backscattering efficiency that can be described by its reflectance. The wing reflectance is considered as independent from time. While deformation of the wing under constraints from muscle and air could modify its backscattering efficiency it is considered negligible when compared to the variation due to the projected area. As explained in subsection 2.1.1, the wing backscatter cross section varies between maximal and minimal value due to change in the projected area. When the wings are at their minimal projected area their contribution to the backscattered signal can be considered as negligible (i.e., $\sigma_{bck,w}(t) = 0$), as insects' wings can be as thin as only a few μm wide. For such a point in time, referred to as local minimums, Equation (2.13) becomes independent from the wing contribution. Those local minimums can be interpolated through a polynomial fit to evaluate the contribution of the body, as illustrated by Figure 2.16. Then by subtracting the body contribution from the total signal, the wing contribution can also be separated.

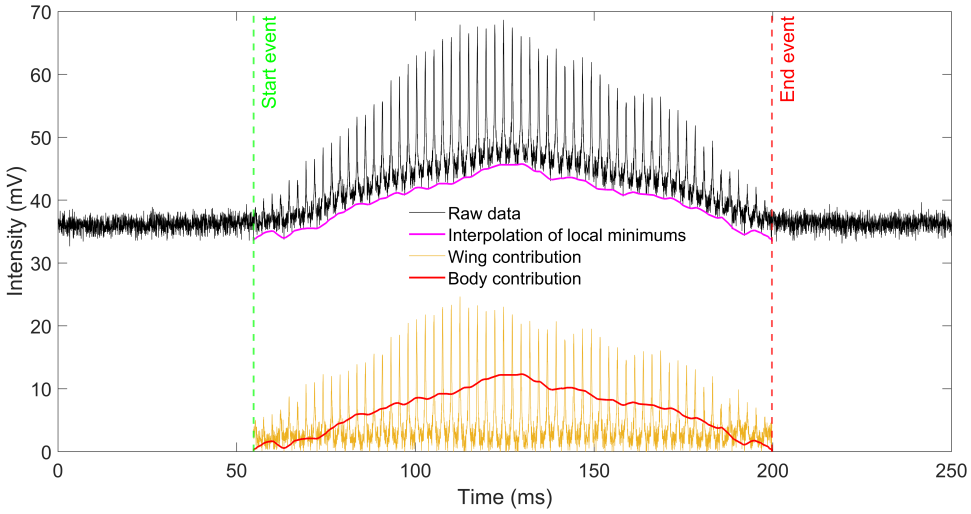

Figure 2.16 Example of wing and body discrimination. The raw data (black) are used in combination with the estimated wingbeat frequency to find local minimum. The interpolation of local minimums (magenta) is subtracted from the raw data to obtain the wing contribution (orange). The wing contribution and mean background value are then subtracted from the raw data to obtain the body contribution (red). The dashed green and red line represent the estimated beginning and end of the insect event respectively.

The described methodology is derived for the ESOS methodology (backscatter configuration). Using identical consideration, it could be derived for the eBoss methodology as well, i.e., extinction configuration. Starting from Equation (2.2) instead of Equation (2.1), a perfectly analogous argumentation could be conducted. As such, both ESOS and eBoss methodologies allow for the discrimination of the wing and body contributions.

From the wing and body contributions, the wing and body optical cross sections can be extracted, following the methodology described in subsection 2.2.4. As shown in Equation (2.13), the wing and body contributions both vary with the position coefficient $p(t)$. Thus, the wing and body contributions can only be equal to the wing and body cross sections when the position coefficient is equal to 1. While estimating the time t at which the position coefficient is exactly equal to unity (i.e., when the insect is within the area of maximal energy density) cannot be resolved, it can be approximated. Indeed, while the current methodology cannot ascertain the exact

position of the specimen within the laser beam, it is possible to estimate the most likely time when the insect was the closest to $p(t) = 1$. This corresponds to the moment when the contributions are at their highest (lowest for eBoss) values. In the simplifying assumption that the maximal (minimal for eBoss) values correspond to $p(t) = 1$ and $\sigma_w(t) = \max\{\sigma_w(t)\}$ (for the wing), both the wing and body cross sections can be retrieved. In practice, using the single maximal value may lead to an overestimation of the cross sections, since it is possible for a statistically unlikely high (low for eBoss) point to be included. Hence, in order to mitigate the influence that a few statistically rare extreme values may introduce, it was chosen to estimate the cross sections from the top 10% of the highest (lowest for eBoss) values within the contribution (both for the wing and the body). This specific number of points was chosen after careful consideration and multiple tries as it was providing the more consistent estimate that would perform efficiently for most, if not all, type of signals, from low to high SNR.

Studying wing and body cross sections (or their ratio for ESOS), refines the analysis possibilities. As an example, two insects could have the same overall cross section but very different wing and body cross sections. A butterfly and a bumble bee can have the same overall cross section. However, a butterfly would have a greater wing cross section than a bumble bee but a smaller body cross section, due to their morphological differences.

2.2.6 Depolarization and Spectral Ratios

Unlike the eBoss, ESOS allows for dual-wavelength and polarization sensitive measurements. As such, it can provide additional information on the studied insects, in particular: depolarization and spectral ratios for both the wing and the body. As shown on Figure 2.1, the ESOS records three different channels simultaneously. The first channel, not resolved in polarization, records the signal from a 940 nm

CW laser beam (index 940). The second channel records the signal from the parallel polarization of a 1320nm CW laser beam (index 1320,$/\!/$). The third channel records the signal from the perpendicular polarization of a 1320 nm CW laser beam (index $1320, \perp$). In addition, all of the three channels can be furthermore separated into the contributions of the wings (index w) and body (index b), resulting in six measured backscattered intensities: $I_{940,w}/I_{940,b}/I_{1320,/\!/,w}/I_{1320,/\!/,b}/I_{1320,\perp,w}/I_{1320,\perp,b}$.

From those six values, several of the insect's optical characteristics can be retrieved. The most important ones being the depolarization and spectral ratios of both the body and the wings. As shown on Equation (2.1), the measured intensities are function of the distance at which the insect crosses the laser beams.

Since this system uses CW diode lasers, the evaluation of the distance through time-of-flight methods, as in most lidar studies, is inapplicable. Yet, by considering the ratio between signals, the influence of the distance can be removed. Using a near Lambertian gold target, each detection channel was aligned so that the signal ratio between channels is constant with range, as shown in Figure 2.17. For instance, $\frac{I_{1320,/\!/}(d)}{I_{940}(d)}$ should be constant with the distance d varying from 3.2 to 4.4 m, which corresponds to the position of the insect enclosure. The consistency of these ratios with distance was determined and the worst variation was found to be at maximum 1.6% over the entire range. This ensures that, within a reasonable error margin, signal ratios are independent from the actual distance at which the insect transits through the laser beams.

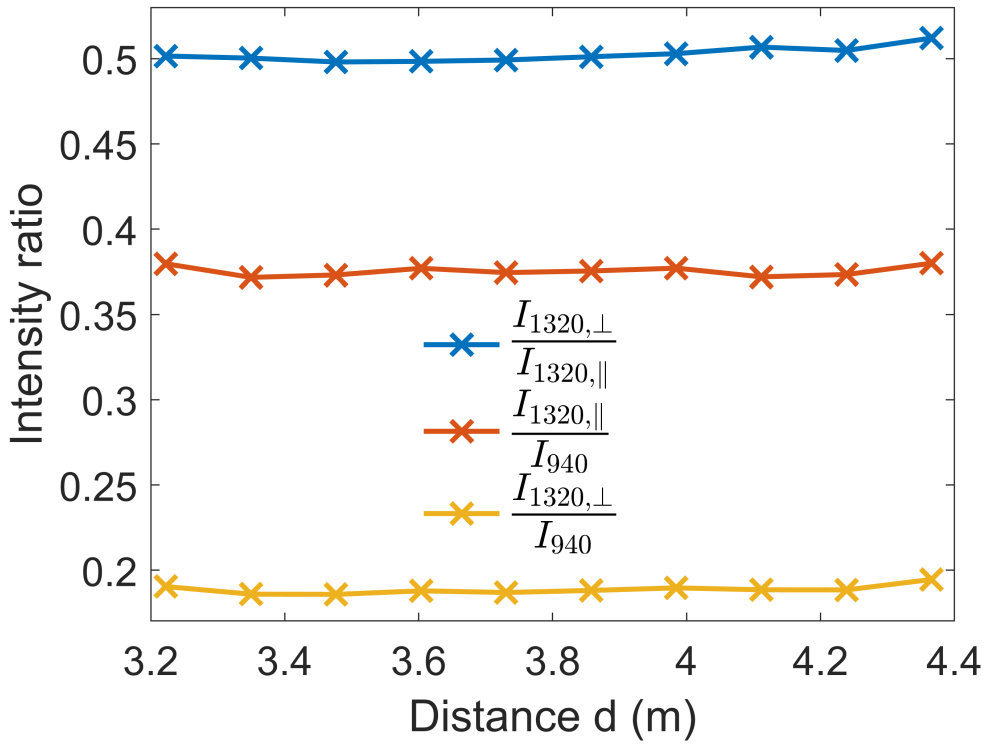

Figure 2.17 Ratios of the intensity measured by each of the three channels of the laboratory ESOS. Ratios are plotted in function of the distance d, which is the distance between the collection optics and the near Lambertian gold target.

In essence, signal ratios are dependent of the specificity of the optical setup, such as the gain on the detector, their quantum efficiency or the quality of the alignment. Therefore, in an effort to broaden the scope of this contribution, a calibration has been performed to retrieve backscatter cross section ratios (i.e., $\frac{\sigma_{940}}{\sigma_{1320,/\!/}}$) from the signal ratios (i.e., $\frac{I_{940}}{I_{1320,/\!/}}$), which are intrinsic characteristics of the specimen and independent from the measurement system. Equations (2.14), (2.15) and (2.16) describe the calibration of the measured intensity ratios into backscatter cross section ratios.

$$\frac{\sigma_{940}}{\sigma_{1320,/\!/}} = G_{\frac{/\!/}{940}} \cdot \frac{I_{940}}{I_{1320,/\!/}} \tag{2.14}$$

$$\frac{\sigma_{940}}{\sigma_{1320,\perp}} = G_{\frac{\perp}{940}} \cdot \frac{I_{940}}{I_{1320,\perp}} \tag{2.15}$$

$$\frac{\sigma_{1320,\perp}}{\sigma_{1320,/\!/}} = G \cdot \frac{I_{1320,\perp}}{I_{1320,/\!/}} \tag{2.16}$$

The following paragraphs describe the calibration for depolarization ratios and the calibration between the two considered spectral ranges. The calibration for the ratio $\frac{\sigma_{1320,\perp}}{\sigma_{1320,/\!/}}$ is a relatively usual situation in polarization sensitive lidar studies. A thorough and effective calibration method, described in Alvarez et al. 2006 [127], was used in this work. The advantages of this calibration method are that in addition to the retrieval of the calibration constant (G) between the parallel channel and the perpendicular channel, it also provides the depolarization ratio of the target (δ) as well as information on the possible misalignment between the polarization beam splitter and the parallel polarization of the laser beam. To briefly resume the methodology described fully in Alvarez et al. 2006 [127], the orientation of the emitted polarization is manually rotated away from its original alignment by the use of a half-wave plate. For each rotated state, characterized by the angle (γ), the ratio of the intensity of the channel $I_{1320,\perp}$ over channel $I_{1320,/\!/}$, noted δ^*, is measured and fitted using Equation (2.17). Figure 2.18 presents the measurements as well as the fit obtained, which provides the best suited values of depolarization ratios (δ), calibration constant (G) and reference angle (γ_0).

$$\delta^* = G \cdot \frac{\delta + tan\left(2 \cdot [\gamma - \gamma_0]\right)^2}{1 + \delta \cdot tan\left(2 \cdot [\gamma - \gamma_0]\right)^2} \tag{2.17}$$

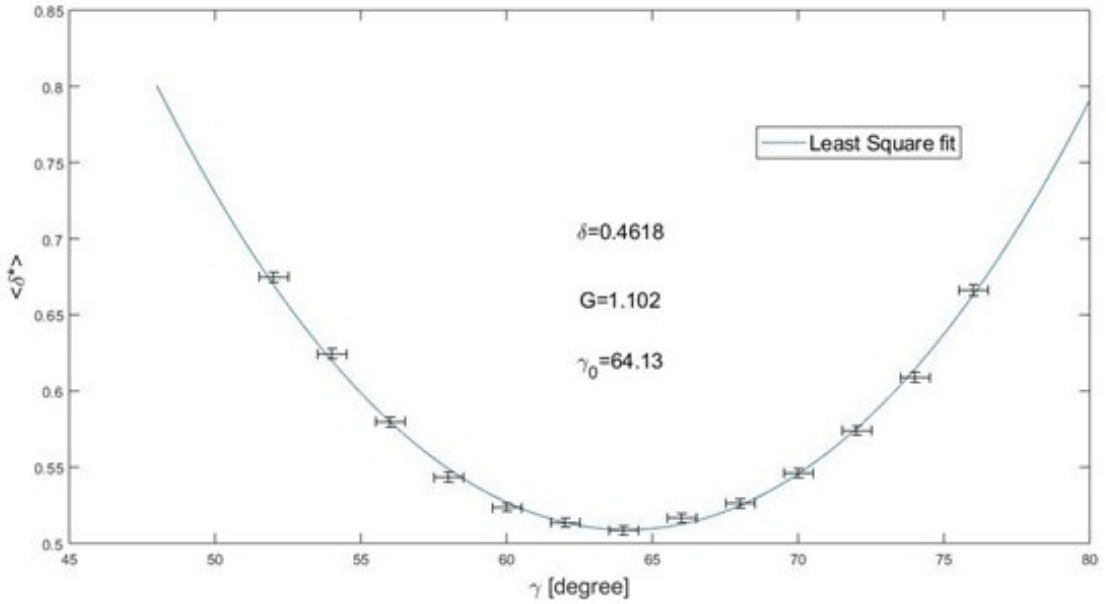

Figure 2.18 Least square fit based on Equation (2.17). This fit allows for the determination of the calibration constant G between channel $I_{1320,\perp}$ and $I_{1320,/\!/}$ and the depolarization ratio of the gold target used during calibration, δ.

The calibration constant between the 940 nm channel and both polarizations of the 1320 nm channel can be found by using the known reflectance of the gold target. In addition, the total backscatter cross section at 1320 nm is in turn equal to the sum of the backscatter cross section for the perpendicular and parallel polarization ($\sigma_{1320,/\!/}$ and $\sigma_{1320,\perp}$). The calibration constants, presented in Equations (2.18) and (2.19), $G_{\frac{/\!/}{940}}$ and $G_{\frac{\perp}{940}}$ can be found by measuring the intensity ratios between channels while knowing the depolarization ratio of the gold target ($\delta = \frac{\sigma_{1320,\perp}}{\sigma_{1320,/\!/}}$). Additionally, in the considered range of wavelengths, the gold target has the required properties of scattering light with the same efficiency regardless of the wavelength (i.e., $\sigma_{940} = \sigma_{1320}$).

$$G_{\frac{/\!/}{940}} = \frac{I_{1320,/\!/}}{I_{940}} \cdot \frac{\sigma_{940}}{\sigma_{1320,/\!/}} = \frac{I_{1320,/\!/}}{I_{940}} \cdot \frac{\sigma_{1320}}{\sigma_{1320,/\!/}} = \frac{I_{1320,/\!/}}{I_{940}} \cdot (1+\delta) \tag{2.18}$$

$$G_{\frac{\perp}{940}} = \frac{I_{1320,\perp}}{I_{940}} \cdot \frac{\sigma_{940}}{\sigma_{1320,\perp}} = \frac{I_{1320,\perp}}{I_{940}} \cdot \frac{\sigma_{1320}}{\sigma_{1320,\perp}} = \frac{I_{1320,\perp}}{I_{940}} \cdot (1+\frac{1}{\delta}) \tag{2.19}$$

2.2.7 Aerial Density

This section proposes a way to evaluate the aerial density of flying insects, expressed in number of flying insects per meter cube, from measurements obtained using any entomological photonic system (EPS). The term "entomological photonic sensor" is meant to represent any photonic device designed to monitor flying insects, generally using one or multiple electronic detectors to detect the electromagnetic radiations, from either an active or passive source, scattered or absorbed by flying insects in the device's field of view. This definition includes entomological lidars and radars as well as any sensors in either backscattered or extinction configurations using either an active light source, such as a laser or LED, or a passive light source, such as the sun. The sensor must meet only two requirements in order for this methodology to be applicable: 1) it needs to be resolved in time, i.e., the duration of the insect's transit Δt through the field of view of the instrument can be retrieved from the data collected by the instrument, and 2) the approximate volume of air ΔV probed by the instrument must be known.

Let's assume that a single insect specimen is randomly flying within a given volume V. Then, let's consider an EPS that probes a volume of air ΔV, such that $\Delta V \subset V$, for a total duration T. As the insect moves freely, it will randomly get in and out of the probed volume, these events are referred to as transits, and the duration of transit i through ΔV is denoted Δt_i. Finally, let's denote ΔT the total time that this single insect spent within ΔV, defined as the sum of its transit time Δt_i. Given that the initial position and movements of the insect within the considered volume V are random, it can be argued that, for a sufficiently long period of observation $T \to +\infty$, the ratio of time spent by the insect in the probed volume over the total time will tend to be equal to the ratio of the probed volume over the total volume, see Equation (2.20).

$$\frac{\Delta V}{V} = \lim_{T \to \infty} \frac{\Delta T}{T} \tag{2.20}$$

To illustrate this idea, one could imagine a situation where a single specimen is flying freely within a constraint volume V. If an EPS is probing 50% of this total volume (i.e., $\frac{\Delta V}{V} = 0.5$) then, assuming random motion, the specimen would spend 50% of its time in the FOV of the instrument (i.e., $\frac{\Delta T}{T} = 0.5$).

In the case where N_s insects are present in the volume V, it is assumed that each insect will spend on average a total time ΔT within ΔV. As a result, the sum of all transit times by all insects spent within ΔV will be equal to N_s times ΔT, as described by Equation (2.21). This is useful as in practice, the total number of insects N_s remains unknown, while it is easy to sum the duration Δt_i of all measured transits.

$$N_s \cdot \Delta T = \sum_i \Delta t_i \tag{2.21}$$

The absolute population density of flying insects ρ_s in the volume V is defined by Equation (2.22).

$$\rho_s = \frac{N_s}{V} \tag{2.22}$$

Finally, by combining Equations (2.20), (2.21) and (2.22), the retrieved aerial density of flying insects ρ_r can be expressed as a function of the sum of all transit times $\sum_i \Delta t_i$, the probed volume ΔV and the total duration of observation T, as described by Equation (2.23).

$$\rho_r = \frac{\sum_i \Delta t_i}{T \cdot \Delta V} \tag{2.23}$$

Depending on the type of instrument, ΔV can be described differently. Figure 2.19 presents the case where the probed volume ΔV is described as a truncated cone, which is a good approximation for most systems using an active light source, such as ESOS and eBoss. In this case, ΔV can be expressed as a function of the divergence θ of the laser beam, its initial radius R_0 and the maximal range of detection L.

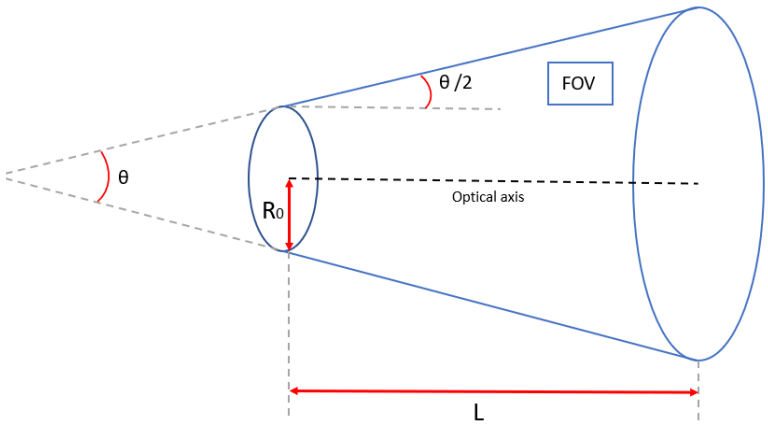

Figure 2.19 Illustration of the probed volume within the field of view (FOV) of an entomological photonic sensor. R_0 is the initial radius, L is the maximal range at which the sensor can detect a specimen and θ is the divergence of the laser beam.

As a result, the retrieved aerial density can be rewritten as presented in Equation (2.24), where all terms have been previously defined.

$$\rho_r = \frac{\sum_i \Delta t_i}{T \cdot \pi \cdot L \cdot \left[R_0^2 + R_0 \cdot L \cdot tan(\frac{\theta}{2}) + \frac{1}{3} \cdot L^2 \cdot tan^2(\frac{\theta}{2})\right]} \tag{2.24}$$

Theoretically, the retrieved aerial density derived from Equation (2.23) is an absolute value and is independent from the instrument or experiment characteristics. When using transit counts, which is the current norm, insects with high flying speed will result in greater transit counts than insects with lower flying speed independently of their actual population. Because it relies on the time of transit and not the number of transits, Equation (2.23) is independent of insect speed.

Due to the statistical nature of this methodology, Equation (2.23) only holds true for a sample of sufficiently large size, i.e., a large number of transits are observed

to be statistically relevant, upon which depends the uncertainty of the retrieved aerial density. The number of transits will increase with the number of insects N_s, the probed volume ΔV, and the duration of observation T. In practice, the number of insects is generally out of control of the experiment and the probed volume is defined by the instrument itself, leaving only the duration of observation as a practical way to influence the uncertainty on the retrieved aerial density.

When studying the retrieved aerial density over time, the duration of observation T becomes the minimum time required to retrieve the aerial density with a given uncertainty. In this case, T can be seen as the temporal resolution of the retrieved aerial density and, for this reason, T will be described as the duration of observation or the temporal resolution interchangeably throughout the rest of the manuscript. As with many experimental approaches, there is a trade-off between the temporal resolution and the uncertainty of the measurement: a better temporal resolution means a higher uncertainty, while more accurate measurements require the collection of data for longer period of time.

To test the veracity of Equation (2.23) a numerical simulation has been designed. This simulation is based on simulating random displacements of N_s flying specimens within a 3-dimension spatial domain of volume V. The position of each specimen at every instant is then compared to a subsection of the total spatial domain which simulates the volume probed by an EPS, referred to as ΔV. The model calculates the sum of all transit times and uses Equation (2.23) to evaluate the aerial density ρ_r, which is then compared to the true input of aerial density $\rho_s = \frac{N_s}{V}$.

One of the fundamental aspects of the simulation is to mimic the displacement of insects over time. Describing flying insect motion is not an easy task for several reasons. The motion of insects in 3D space can change depending on the insects goal (i.e., foraging, mating, exploration etc.. . .) [128]. Different species may also present various types of flight patterns. Yet, a stochastic approach such as a simple random

walk, which assumes that the insect is randomly moving in its environment, has been shown to be statistically meaningful for the study of insects' population density [129]. The process used in this contribution is a Gaussian random walk that does not require any assumption that would be overly specific to any given species. It can therefore be used, in respect to the presented methodology, for any flying insect species as long as corrected for the insect mean flight speed.

In order to simulate the movement of a specimen, its motion is decomposed into discrete iterations. Each iteration corresponds to a constant time interval of dt seconds during which the specimen is moving in a straight line from its previous position to its next. At a given iteration j, the position of the specimen is $P_j = (x_j, y_j, z_j)$. The position of the specimen at the next iteration, P_{j+1}, is defined as $P_{j+1} = P_j + \Delta P_j$ where $\Delta P_j = (\Delta x_j, \Delta y_j, \Delta z_j)$ is the displacement of the specimen at iteration j. The stochastic process describing the displacement of the specimens is performed in two steps. First, a direction is selected randomly with equal probabilities in all directions. This is done using the methodology described in Cook 1957 [130]. Briefly, this method randomly and uniformly selects a point on a sphere of unit radius. The associated unit vector, described by (r_{xj}, r_{yj}, r_{zj}), is the direction toward which the next displacement will take place. The second part of the stochastic displacement process is to select the distance traveled in the aforementioned random direction. This distance, D_j, is determined by the product of the flight speed with the time interval $D_j = V_j \cdot dt$. At every step, the flight speed is randomly selected from a normal distribution of mean value $\overline{v}$ and standard deviation σ'', $V_j = \mathcal{N}(V|\overline{v}, \sigma''^2)$. Overall, the iterative process of displacement is given by Equation (2.25).

$$P_{j+1} = P_j + V_j \cdot dt \cdot (r_{xj}, r_{yj}, r_{zj}) \tag{2.25}$$

The initial position P_0 of the specimen is randomly and uniformly selected within the considered spatial domain. An example of the displacement of a single specimen during this random walk process is illustrated by Figure 2.20.

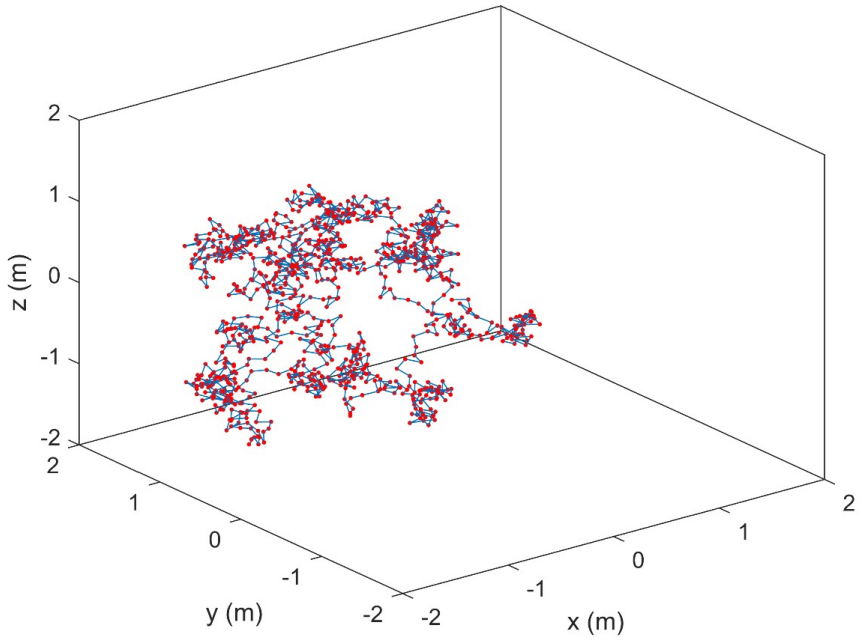

Figure 2.20 Illustration of the stochastic displacement of a single simulated specimen within a total cubic spatial domain of edges of 4 m. Blue lines are the linear displacement of the specimen between each position P_j represented by red dots. This figure shows a total of 1000 iterations with a time step $dt = 0.5\ s$ and an average flight speed $\overline{v} = 0.245\ m \cdot s^{-1}$.

Building up from the individual displacement, simulations are conducted for different numbers of specimens N_s. Each specimen trajectory is independent from the other. The trajectories are then used to calculate the total time that insects have been within the probed volume $\sum_i \Delta t_i$. The probed volume of the instruments is a truncated cone, described by a set of three parameters: L, R_0 and θ. Any specimen transiting through this volume will induce an increase in the total transit time.

For practical reasons, the simulation was restricted to a finite spatial domain which is a cube of edge equal to L, i.e., $\frac{-L}{2} \leq (x, y, z) \leq \frac{L}{2}$. The motivation was a compromise between including the entire probed volume of the system, limiting

the influence of any potential boundary effects and reducing the computational cost. The larger the spatial domain, the more specimens must be modeled to achieve a given aerial density. The boundaries of the spatial domain are to be considered as an absolute limitation for the specimen's position. If a specimen is located near the boundary at P_j then there is a possibility for P_{j+1} to be outside of the spatial domain, for example $x_{j+1} > \frac{L}{2}$. In such cases, the specimen position will be stopped at the boundary, $x_{j+1} = \frac{L}{2}$. Thus, the insect flux through the boundary is null, as if the same number of specimens would enter and exit the spatial domain in the case of random walk. This stems from the assumption that the insect population is homogeneous, without static cluster or a swarm of insects.

The input aerial density of flying specimen ρ_s within the model is known, as it is the ratio between the input values N_s and V. Using Equation (2.23) and the total transit time evaluated by the simulation, the retrieved aerial density ρ_r is calculated and its variability used to evaluate the relative standard uncertainty μ_r. Intrinsically, this model revolves around a stochastic process of displacement within the spatial domain. Consequently, stochastic fluctuations in the retrieved density are to be expected even between simulations with identical parameters. Hence, the simulation is repeated several times with identical input parameters to obtain statistics on the retrieved aerial density: its value distribution, average value and standard deviation, all three are represented in Figure 2.21 (A) and (B), for $T = 5\ min$ and $T = 20\ min$ respectively. From the standard deviation of the retrieved aerial density $\sigma''(\rho_r)$, the relative standard uncertainty μ_r is calculated and expressed in percent. The latter describes the uncertainty on the retrieved aerial density in the following manner: $\rho_r \pm \mu_r(\%)$. The relative standard uncertainty is plotted in Figure 2.21 (C) as a function of T, showing that the retrieved aerial density does converge toward the input aerial density for sufficiently high values of T, which further confirms the validity of Equation (2.23). It is important to note that the presented methodology

only considers insects that are flying, as all the simulated specimens were flying at all times. This means that the retrieved aerial density is of flying insects and does not consider resting or sleeping insects.

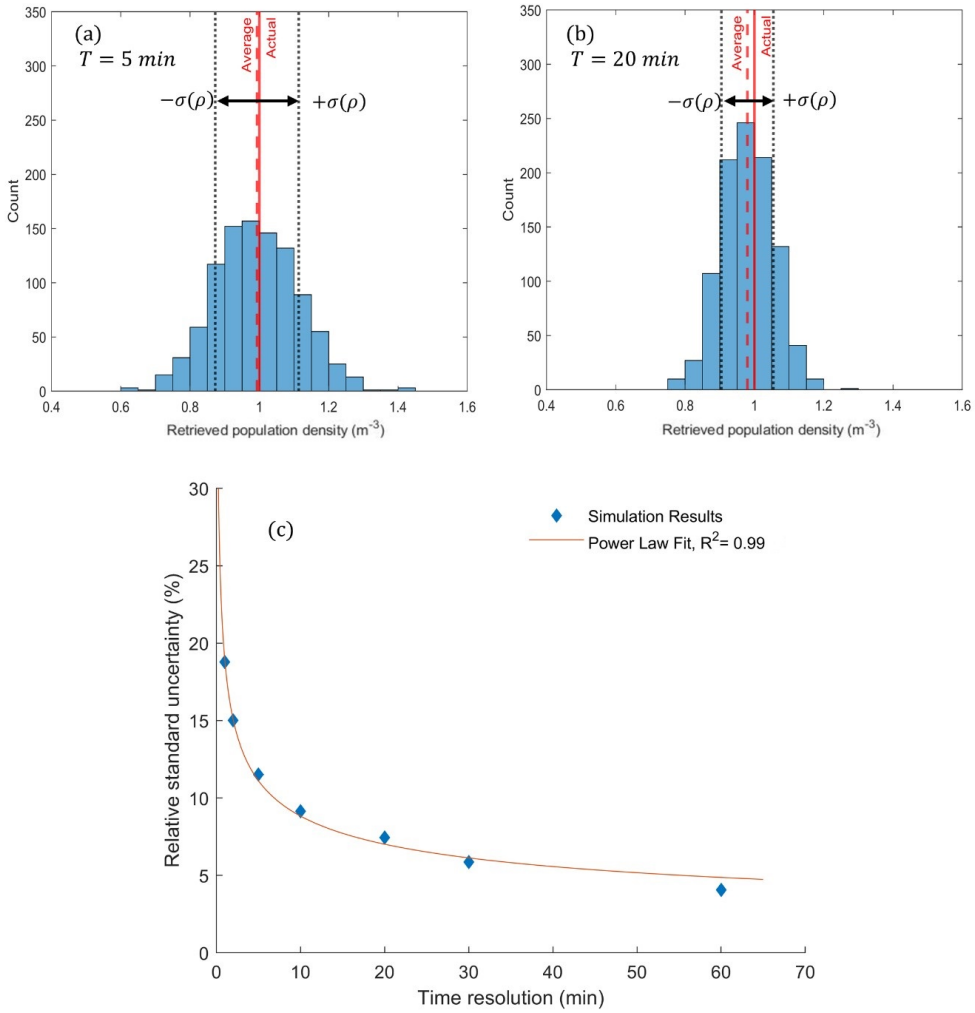

Figure 2.21 Histograms of the retrieved aerial density for 1000 repetitions. Simulation for $\Delta V = 15\ m^3, \rho_s = 1\ m^{-3}, T = 5\ min(A), 20\ min(B)$. Dotted lines represent the standard deviation. These two histograms are examples of the results used to plot the relative standard uncertainty in function of the time resolution (C).

2.2.8 Mass Estimation

Both in laboratory conditions and in real-world experiments, the eBoss allows for the retrieval of the extinction cross section of a specimen. Unlike the ESOS which can

only retrieve the backscatter cross section in limited laboratory conditions, that are inapplicable to the field conditions, due to the distance dependency as illustrated by Equation (2.1). This explains why the mass estimation, which uses cross sections, will only be discussed in the case of the eBoss.

When considering an insect i, the extinction cross section of its body $\sigma_{ext,b,i}$ is function of the geometrical cross section of the insect body $A_{b,i}$ and the quasi-ballistic transmittance of the specimen's body $T_{b,i}$ considered, in first approximation, as constant across insect species, see Equation (2.26).

$$\sigma_{ext,b,i} = (1 - T_{b,i}) \cdot A_{b,i} \tag{2.26}$$

In "surface area-volume ratios in insects" [131], Küshel et al. 2017 proposes and tests a relationship to estimate the volume V_i of an insect from its geometrical cross section $A_{b,i}$, without considering the wings sizes as they have a negligible contribution when considering the total mass of the insect. As shown in Equation (2.27), the volume is equal to the product of its geometrical cross section to the power 3/2 and a constant K_i. As discussed in more details in Küshel et al. 2017 [131], the proportionality factor K_i is estimated using beetles, flies, hymenopterans and butterflies and found to be in the $[0.027 : 0.034]$ interval, which is a relatively constant values across species.

$$V_i = K_i \cdot (A_{b,i})^{\frac{3}{2}} \tag{2.27}$$

The mass of an insect i is denoted m_i and is equal to its volume V_i times its volumetric mass density ρ_i, see Equation (2.28). The volumetric mass density of insects may vary from one species to the next. However, Küshel et al. 2017 [131] showed the relationship between an insect volume and its mass to be mainly linear,

$R^2 = 0.92$, across 113 insect species, indicating that considering the volumetric mass density of insect as constant across species is a reasonable approximation.

$$m_i = \rho_i \cdot V_i \tag{2.28}$$

Equation (2.29) is derived by combining Equations (2.26), (2.27) and (2.28) and by applying the aforementioned approximations (where $T_{b,i}$, K_i and ρ_i are considered constant across all insect species).

$$m_i = \rho \cdot K \cdot \left(\frac{\sigma_{ext,b,i}}{(1 - T_b)} \right)^{\frac{3}{2}} = \eta \cdot (\sigma_{ext,b,i})^{\frac{3}{2}} \tag{2.29}$$

The factor η takes into account the difference between geometrical and extinction cross sections, the relationship between surface and volume and the volumetric mass density of insect. While η is expressed in unit of mass per unit volume, it is important to note that it is not equal to the volumetric mass density of insects.

Using Equation (2.29), the mass of every flying insect that crosses the laser beam can be estimated. Combined with Equation (2.23), it allows for the determination of the mass of flying insects per unit of volume, referred to as biomass density throughout the manuscript, denoted ρ_b, which can be calculated for both wet and dry insect mass using Equation (2.30).

$$\rho_b = \frac{\sum_i^{N_i} \frac{\Delta t_i}{\Delta T} \cdot m_i}{\Delta V} \tag{2.30}$$

This equation is derived from a statistical approach detailed in subsection 2.2.7, where the aerial density (insects.m^{-3}) is determined, but modified to be expressed in terms of biomass density instead. N_i is the number of events observed during the time period ΔT. The mass m_i is retrieved using Equation (2.29), while the transit

time Δt_i represents the duration during which the insect was within the laser beam of the eBoss. The biomass density ρ_b is normalized by the volume of air probed by the instrument ΔV so that it is expressed in $mg.m^{-3}$.

The time resolution of the retrieved biomass density is defined by ΔT, which can be set to days, weeks or months to observe the long-term evolution of the biomass density or set to minutes or hours to observe daily changes in biomass density. The number of observed events by the eBoss, and thus the retrieved biomass density, is subject to stochastic fluctuations. For this reason, there is a trade-off between time resolution and uncertainty: small time scale, in the hour or minute range, may present significant statistical fluctuations on the retrieved biomass density while longer time resolution may provide more robust results.

2.3 Machine Learning Classification

As detailed in section 2.2, ESOS and eBoss can retrieved several of the insect characteristics such as wingbeat frequency or cross section. In laboratory conditions, the species of every insect studied is known but in real-world experiments this would not be the case. Incidentally, being able to identify the insects from the aforementioned characteristics would be an invaluable addition to the capabilities of entomological photonic sensors.

2.3.1 Predictor Variables

Every insect event is uniquely identified by an ID number and all the retrieved characteristics are associated with it in the form of a meta-data vector. This vector contains properties that are intrinsic to the insect, such as its wingbeat frequency. Those properties vary depending on the insect species. For example, mosquitoes have a wingbeat frequency in the 300-800 Hz range, depending on the species and

sex, which is higher than most flying insects. As such, this specific property can be used to differentiate mosquitoes from other insect families.

In this context, the vector containing the retrieved insect characteristics can be seen as a vector of predictor variables. By combining all those vectors together, a database of insect characteristics can be constructed. In laboratory conditions, the insect species, sex or even gravidity is known. Each event can therefore be labelled with the appropriate information and used for training and testing machine learning algorithms. Each unique label, also referred to as a class, consists of the species, sex and gravidity of the specimen. The goal of a machine learning classifier is to predict the correct class of an unknown event from its vector of predictor variables.

2.3.2 Training and Evaluation of Machine Learning Algorithm

All machine learning algorithms are trained and tested using leave one out cross-validation [132] as it is often the case when the training data are relatively scarce. In cross-validation the data are divided in training and testing sets. Events from each class are randomly separated into different subsets of equivalent sizes. All but one subset are used for training and the remaining subsets are used for testing. The process is repeated until every subset has been used as a testing set exactly once.

In some of the described classification tasks, see section 3.2, the number of samples in each class is imbalanced, sometimes by a 4:1 ratio between majority and minority classes. The imbalanced data set problem has been called out as a pervasive problem for the data mining community [133, 134]. Due to the unequal number of samples for each class, a data-level method has been used to prevent any bias caused by the imbalanced data set.

Within data-level methods, different re-sampling methods exist such as under or over-sampling. Briefly, oversampling consists of creating new samples in the minority class and undersampling of removing samples from the majority class,

both in order to balance the data set. Both methods have advantages and disadvantages. Undersampling may lead to the loss of information by removing potentially meaningful data while oversampling can lead to overfitting. The approach applied throughout this manuscript was undersampling, as it has been shown to perform better than oversampling when the imbalance is not severe [135]. Furthermore, losing potentially useful information was preferred to the risk of overfitting. Within undersampling, different approaches of elimination can be considered. Some undersampling methods eliminate samples that are redundant or at the boundary with minority classes [136]. An elimination based on a set of fixed rules could systematically eliminate meaningful features. In practice, the chosen approach was random undersampling, which is the random elimination of events from the majority class. This undersampling approach was applied until the number of events in every class was identical, ensuring no imbalance between classes.

To evaluate and compare different machine learning algorithms, evaluation metrics must be introduced. The first and foremost metric used in this contribution is the accuracy. The accuracy is defined as the number of correct prediction N_+, regardless of class, in respect to the total number of predictions N_{tot}, both correctly N_+ and incorrectly N_- predicted, as shown by Equation (2.31).

$$Accuracy(\%) = 100 \cdot \frac{N_+}{N_{tot}} = 100 \cdot \frac{N_+}{N_+ + N_-} \tag{2.31}$$

The other metrics used in this manuscript are class level recall and precision [137]. The class level recall of class C_j is defined as the number of events correctly predicted as class C_j, N_{+,C_j}, divided by the total number of events in class C_j, N_{tot,C_j}, as defined by Equation (2.32). For simplicity the class level recall is often referred to as recall.

$$Recall_{C_j}(\%) = 100 \cdot \frac{N_{+,C_j}}{N_{tot,C_j}} \tag{2.32}$$

The class level precision of class C_j is defined as the number of events correctly predicted as class C_j, N_{+,C_j}, divided by the total number of events predicted as class C_j, N_{pred,C_j}, as defined by Equation (2.33). For simplicity the class level precision is often referred to as precision.

$$Precision_{C_j}(\%) = 100 \cdot \frac{N_{+,C_j}}{N_{pred,C_j}} \tag{2.33}$$

Recall studies how likely an event from class C_j is to be correctly predicted while the precision studies how likely an event predicted of class C_j is to be from this class. Depending on the classification task, either the recall or the precision may be more important. In the case of disease detection, the recall should be maximized. For such classification tasks, not properly detecting a sick patient should be avoided. For a classification such as spam detection the precision should be maximized. In this classification task, predicting as spam an email that is not one should be prevented.

An extension of the recall and precision metrics can be found in their macro average [138]. The macro average recall/precision is defined as the unweighted average of all the class level recall/precision. Those metrics are used when the performance of the entire machine learning algorithm over all classes needs to be considered.

2.3.3 Considered Machine Learning Algorithms

For the identification of insects, several distinct machine learning algorithms can be used. Each algorithm has its own advantages and disadvantages. Depending on the classification tasks, different algorithms will perform differently. Therefore, it is important to explore multiple options in order to find the best suited machine learning algorithms for each of the desired classification tasks. During the course

of this project, five different machine learning algorithms have been considered and tested on a multitude of insects' classification problems.

Naïve Bayes (NB) Naïve Bayes is based on a probabilistic approach that relies on Bayes rule [139]. In such approach, performances are optimal when the predictor variables are independent. Although, even when the predictor variables are not independent, NB can still perform efficiently [140]. For every class, a probability density function is designed and used for prediction, as illustrated on Figure 2.22. Different distributions can be chosen and the one kept in this contribution is a Gaussian Kernel function. To make a prediction, the probability of each class is calculated for every predictor variable and the class with the higher score is chosen as the most likely. This approach was considered for its simplicity of implementation, which can be an advantage for practical use.

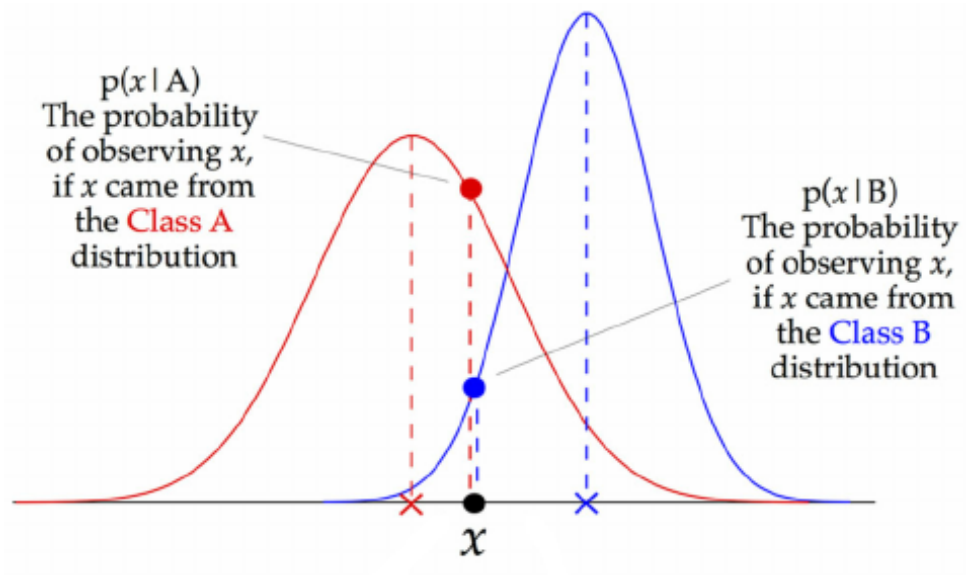

Figure 2.22 Illustration of Gaussian Naïve Bayes (NB) for a two classes, one dimension, classification problem. The red and blue lines represent the probability density functions (i.e., distributions) of a given variable from two distinct classes. For the observed value x, as indicated in the figure, the probability of belonging to class A and B are calculated, respectively $p(x \mid A)$ and $p(x \mid B)$. The class with the highest probability, class A in this example, will then be the predicted by the NB classifier as the most likely class.
Image credit: Prateek Majumder.

K-Nearest Neighbor (KNN) K-Nearest Neighbors is one of the most intuitive instance-based learning algorithms [141]. The KNN algorithm finds the k nearest instance of the unknown sample and predicts its class as the one that is the most frequent within the neighbors, as shown in Figure 2.23. The concept of proximity in this case can be defined by different distance metrics, such as Euclidean distance, as presented in Witten and Frank, 2005 [142]. To improve accuracy, weights on each neighbor can be added, such as a weight inversely proportional to the Euclidean distance, effectively giving a greater importance to neighbors that are closer. In this contribution, a squared inverse distance weight was chosen. This method was considered for its robustness against outliers which can be expected for entomological data.

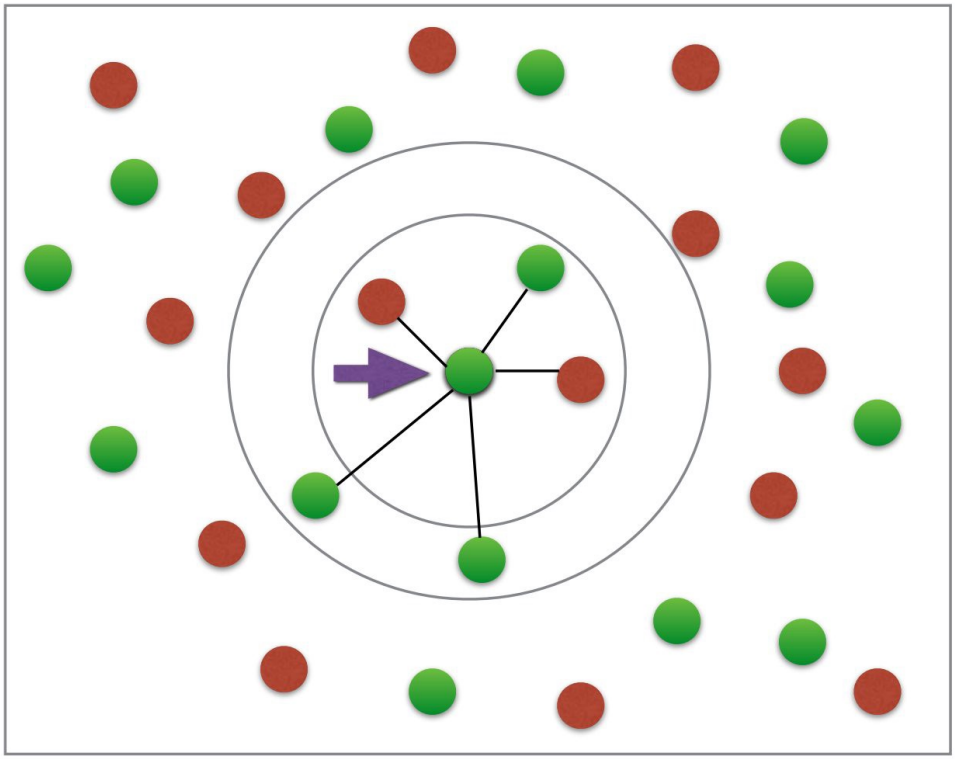

Figure 2.23 Illustration of k-Nearest Neighbor (KNN) for a two classes, two-dimension, classification problem. The red and green colors represent the class that each point belongs to. The purple arrow points towards the instance that is being considered for prediction. Both circles represent different values of k, which is the number of considered neighbor by the algorithm. The smaller circle is for $k = 3$ and the bigger circle for $k = 5$. In this example the distance from the neighbor to the considered instance are illustrated by straight black lines. For $k = 3$ two neighbors are from the red class and one from the green class, in the case of an unweighted KNN the predicted class would be red. For $k = 5$ two neighbors are red and three are green, in the case of an unweighted KNN the predicted class would be green. *Image credit: Eijaz Allibhai.*

Decision Tree (DT) Decision Trees classifiers are based on nodes and branches to separate instances based on feature values [143]. Each node is a feature that is used to discriminate between instances and each branch is a value that a given node can take, as displayed in Figure 2.24. After going through a different number of nodes/branches the algorithm will attribute a class for an instance based on the class label of the terminal nodes also called leaf node. There is a multitude of methods to build decision trees such as CART, C4.5 or Id3. In the absence of a single best method [143], the chosen building algorithm was CART [144] since it has the advantages of being able to identify the most significant predictor and to eliminate unnecessary ones while handling outliers [145].

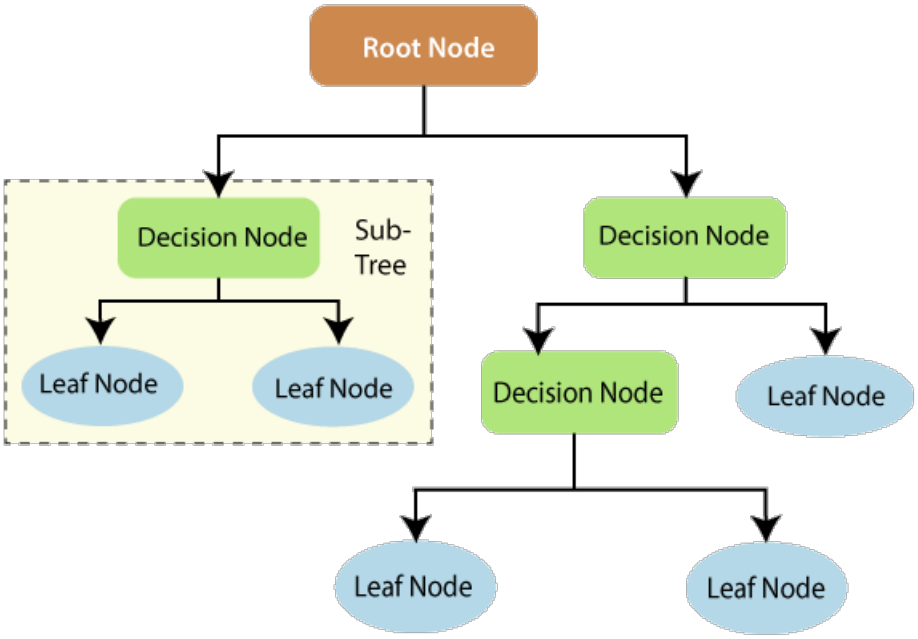

Figure 2.24 Illustration of a Decision Tree (DT) for a multiclass, multi dimension, classification problem. Each decision node represents a condition that separates events into different branches, represented by black arrow. Leaf node are terminal node that are not separated and correspond to a prediction of the model, i.e., equivalent to a class label.
Image credit: Onesmus Mbaabu.

Linear Discriminant Analysis (LDA) Linear Discriminant Analysis was first introduced for two classes taxonomic problems [146], also referred to as Fisher

Discriminant Analysis. The idea behind LDA is to find the linear combination of variables that maximizes the in-between class variance while minimizing the within class variance, as shown in Figure 2.25. LDA assumes Gaussian mixture model and, unlike its Quadratic version (QDA), similar covariance between classes. For the classification task that involves more than two classes the Fisher Linear Discriminant was extended to a multiple discriminant analysis [147]. LDA was selected as it performs well for some of the considered classification tasks, in particular gravidity identification, see subsection 3.2.4.

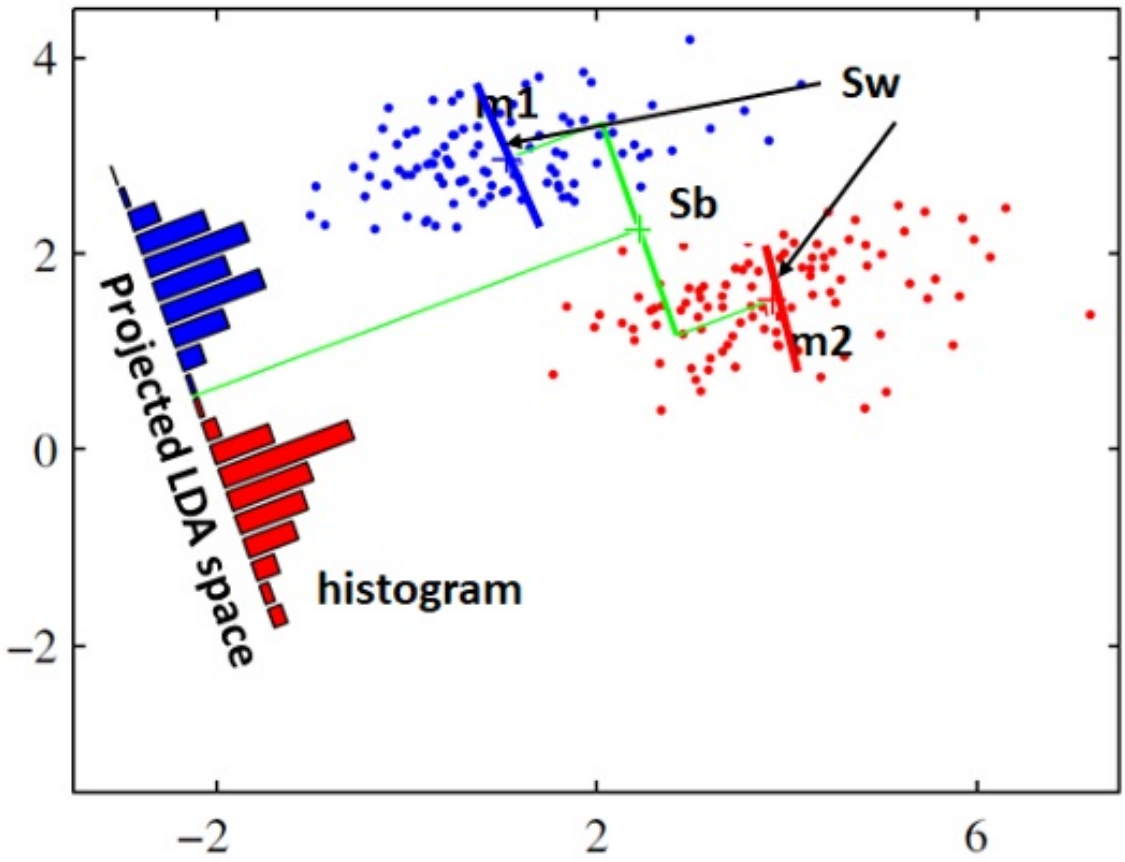

Figure 2.25 Illustration of a Linear Discriminant Analysis (LDA) for a two classes, two-dimension, classification problem. The red and blue colors represent the two classes. $m1$ and $m2$ correspond to the mean values of the blue and red class respectively. The bold blue and red lines illustrate the within class-covariance (S_w). The bold green line illustrates the between class-covariance (S_b). The projected LDA space is chosen in order to maximize S_b while minimizing S_w. Once the best projection has been found, classes are separated by a boundary that here would be the green line perpendicular to the projected LDA space. Every event that falls left of this boundary would be predicted as blue and any event that falls to the right would be predicted as red.
Image credit: Prachi Singh.

Support Vector Machine (SVM) Support Vector Machine is a binary classification method that separates classes by construction of an hyperplane that optimizes the distance between them [148, 149]. SVM will map the input space into a feature space of higher dimension where the hyperplane that maximizes the margin between the two classes can be found, as illustrated in Figure 2.26. In order to expand the linear SVM for a non-binary classification task, a one-vs-one majority vote was chosen [150]. This approach consists of creating all possible binary classifiers and predicting the class based on a majority vote. SVM has gained considerable traction within the past decades and has been tested for this reason.

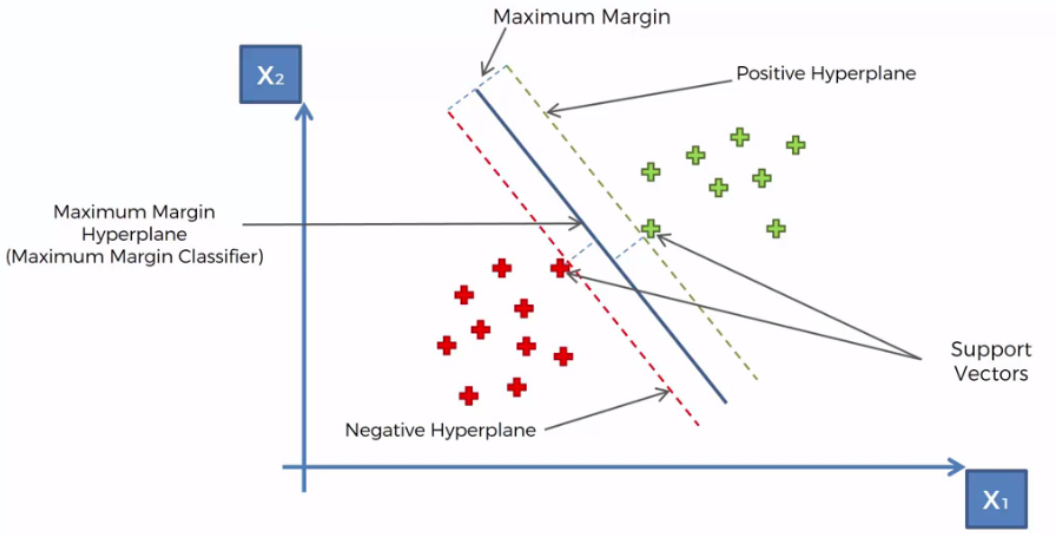

Figure 2.26 Illustration of a Support Vector Machine (SVM) of a two classes, two-dimension, classification problem. Green and red crosses represent the two classes. The SVM finds the extreme points, called support vectors and creates the positive and negative hyperplane. The maximum margin hyperplane is the boundary between the two classes. In this example any points below the boundary would be classified as red class and any points above as green class.
Image credit: Kushal Chakraborty.

CHAPTER 3

LABORATORY EXPERIMENTS AND RESULTS

The first step in the development of field ready Entomological Photonic Sensor (EPS) is to study insects in controlled environments where most environmental factors can be kept constant. In laboratory conditions the temperature, which is known to influence the wingbeat frequency [151], can be kept relatively constant and the exact nature (species, sex and gravidity) of every specimen is known.

3.1 Insect Rearing and Collection

In order to test the capabilities of EPS, whether they are in backscatter or extinction configuration, live insect specimens must be brought into the testing enclosure. Throughout this project several species of insects have been studied. *Culicidae* (mosquitoes) have been the object of a significant focus due in part to their important impact on human health. Among all the insect species studied, some have been reared by collaborators, brought from professional vendors or directly collected from the wild.

In total six species of mosquitoes have been studied in laboratory conditions:

- *Aedes albopictus*, also known as the Asian tiger mosquito, is an important epidemiological vector for infectious diseases such as yellow fever, dengue fever and chikungunya fever. It is originally found in tropical and sub-tropical areas of Southeast Asia, though this species has recently spread in Western countries.

- *Aedes aegypti*, largely known to be one of the main vectors of yellow fever, but also dengue fever, and Zika virus. This species originated in Africa but can now be found in tropical and sub-tropical regions. *Ae. aegypti* is responsible for the Zika virus outbreak in 2015-2016 in Brazil, and several regions of South and North America.

- *Anopheles quadrimaculatus*, one of the most important vectors of malaria. Medium size mosquito named after the four dark spots on its wings it is mostly present in the eastern United States.

- *Culex quinquefasciatus*, commonly known as the southern house mosquito, it can spread several arboviruses such as the West Nile or Zika virus. Most often found in tropical and subtropical regions, including America and Australia.

- *Culex* genus (mixed species of *C. pipiens* Linnaeus and *C. restuans* Theobald) from which several species are vectors of disease such as West Nile virus and multiple forms of encephalitis. *Culex* are widely geographically spread and one of the most encountered mosquito genera in North America. Those collected *Culex* mosquitoes were visually identified only to genus to avoid damage to the specimens.

Mosquitoes came from the Hudson Regional Health Commission, Mosquito and Vector Control. They were field collected as larvae from various locations around Hudson County, NJ, USA and reared in plastic trays (ca. 200 larvae/tray) in 1 L of deionized water with $0.3g$ of brewer's yeast provided on alternate days. After eclosion, adults were housed in 30 x 30 x 30 cm aluminum screen cages at 26°C, 75% relative humidity, with a 16:8 h Light:Dark photoperiod and provided a 10% sucrose solution. Specimens were sexed by visual inspection of the number of fibrils on their antennae, as illustrated by Figure 3.1. All species were studied from after they hatched until they died generally 12 to 18 days later. Gravid *Culex* mosquitoes were collected from Secaucus, NJ, USA with a Center's for Disease Control gravid trap [152] (John W. Hock Co., Gainesville, FL, USA) baited with a hay infusion [153]. Gravidity was visually confirmed by viewing egg masses through the cuticle of the mosquitoes.

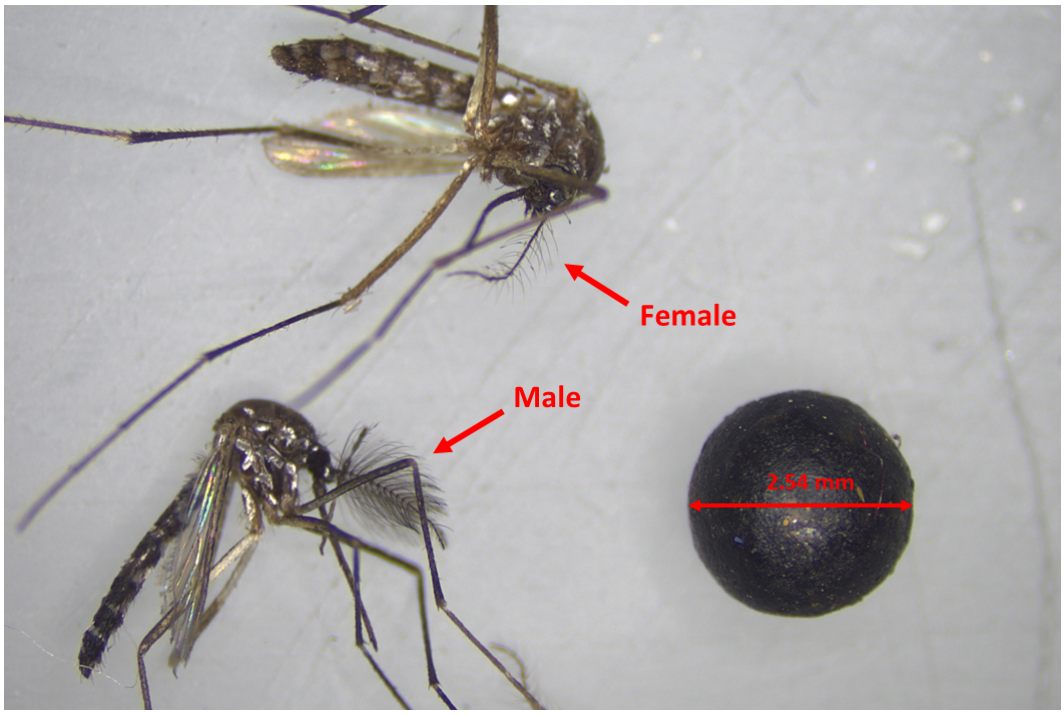

Figure 3.1 Microscope picture of two *Aedes albopictus* mosquitoes, after desiccation. The antennae of the male show a great number of fibrils while the female displays only a limited amount. This simple criterion has been used to visually discriminates males from females.

In addition to mosquitoes several other species of insect have been studied. Some have been obtained from professional vendors:

- *Musca domestica*, also known as house fly. It is one of the most common species of fly found in human habitations.
- *Osmia lignaria*, often named the orchard mason bee. This species is commonly used in Canada and in the USA for pollination.

The rest have been collected in Secaucus, NJ, through net-sweeping:

- *Hylephila phyleus*, or Fiery Skipper is a butterfly present in North and South America. It is not usually considered as a pest in its native range despite its larval state being known to damage Turfgrass.
- *Vespula maculifrons*, commonly referred to as the eastern yellow jacket is a wasp present in North America. They can be considered as pests when they nest in or near human buildings.

- *Bombus bimaculatus*, also known as the two-spotted bumble bee is present in the eastern half of the USA and some part of Canada. Both the queen and workers are known to play a part in the pollination of wild plants.

Once in the laboratory, and in compliance with the Arthropod Containment Level 1, all insects were stored in closed plastic containers. Those containers were placed inside a meshed cage and placed under a fume hood to prevent any unwanted escape. Specimens where carefully transferred between this enclosure and the EPS Plexiglas enclosure, see Figure 2.1 and 2.3, by using a meshed aspiration tube. All insects were fed daily with a 10% sucrose solution.

3.2 Insect Identification, Species, Sex and Gravidity

3.2.1 Mosquitoes Sex and Species Identification using One-Dimensional Naïve Bayes Classifier

As discussed earlier in this manuscript, EPS can extract some of the insect intrinsic characteristics and among them is the wingbeat frequency. The wingbeat frequency, which is the frequency at which an insect flaps its wings, is known to vary across flying insect species [73, 154]. In particular, male and female mosquitoes which, due to strong sexual dimorphism, have clearly distinct wingbeat frequencies [155, 156, 157, 158, 159, 160]. As shown in Table 3.1, *Ae. albopictus*, males have an average wingbeat frequency of 681 Hz while females have one of 456 Hz. Mosquitoes sometimes modulate their flight tone to communicate. For some species, male and female flight harmonics converge toward a common frequency while mating [161, 162]. However, the occurrence of such phenomenon is rather rare making wingbeat frequency a reliable means to differentiate male from female mosquitoes.

From transit signals, each insect event is characterized using a Naïve Bayes classifier. This classifier was chosen for its implementation simplicity and good performance with independent variables [163]. Two different scenarios were considered. First, the sex classification, aiming at discriminating male from female mosquitoes

regardless of their species, using the sole wingbeat frequency as a predictor variable. Then, a similar methodology using six classes is applied to identify the species and sex of mosquitoes, also referred to as the sex and species classification. As it is a laboratory study, mosquitoes' sex and species are known a-priori. Hence, the efficiency of the identification in both scenarios can be evaluated and discussed.

For this experiment the 1320 nm channel of the ESOS is used. Both polarizations, channels two and three, are merged together to form a single channel, not studied in polarization. By combining signals at both polarizations, the signal intensity will increase, improving the SNR.

The 1320 nm wavelength of the laser has been chosen in order to ensure a measurable backscattered intensity from the target while remaining outside of the visual perception of the mosquitoes, as discussed in subsection 2.1.1.

As detailed in subsection 2.2.2, for each mosquito event recorded, a wingbeat frequency f_w can be determined. For every class separately, both a mean μ and standard deviation σ_{SD} are calculated using Equations (3.1) and (3.2), respectively. Those two values are used to mathematically define the frequency distribution of the considered class.

$$\mu = \frac{\sum_{i=1}^{N_m} \frac{f_{wi}}{\sigma_i^2}}{\sum_{i=1}^{N_m} \frac{1}{\sigma_i^2}} \tag{3.1}$$

$$\sigma_{SD} = \sqrt{\frac{1}{N_m - 1} \cdot \sum_{i=1}^{N_m} (f_{wi} - \mu)^2} \tag{3.2}$$

Where N_m is the number of events in the evaluated class, f_{wi} and σ_i are respectively the retrieved wingbeat frequency and uncertainty of event i. The uncertainty σ_i is defined as half the full width half maximum (FWHM) of the fundamental frequency peak in the frequency domain, obtained from the harmonicity

algorithm. For each class, the mean value of the wingbeat frequency μ is a weighted mean. This will give a greater strength to the more precise values while reducing the influence of the more uncertain ones. This was done in regards of experimental uncertainty that fluctuates from around ± 2 Hz up to ± 50 Hz. This uncertainty is due to either a low signal to noise ratio or a change in the specimen wingbeat frequency during the time of transit, since mosquitoes can modulate their wingbeat frequency during flight [161, 162]. The results of mean value and standard deviation are presented in Table 3.1. In this table, the presented uncertainty of the average wingbeat frequency, for every case, is defined as the 95% confidence interval, using standard error. Meaning that if the experiment were to be conducted again in an identical manner, the average wingbeat frequency would be within this interval 95% of the time. The standard deviation presented in this table corresponds to the width of the frequency distribution, presented in Figure 3.2, which is an estimate of the variability of the wingbeat frequency across different specimens of the same sex and species.

Table 3.1 Average Wingbeat Frequency of Mosquitoes

	Average wingbeat frequency	Standard deviation		Average wingbeat frequency	Standard deviation
Male	617 ± 6 Hz	52 Hz	*Ae. albopictus*	681 ± 10 Hz	59 Hz
			Culex	541 ± 14 Hz	34 Hz
			Ae. aegypti	628 ± 12 Hz	54 Hz
Female	408 ± 10 Hz	64 Hz	*Ae. albopictus*	456 ± 12 Hz	41 Hz
			Culex	344 ± 6 Hz	20 Hz
			Ae. aegypti	425 ± 4 Hz	24 Hz

For each class, understood as species and/or sex, a probability density function is derived from μ and σ_{SD}. Those functions are then used to retrieve the likeliness that each measured wingbeat frequency has originated from a specific class. For each wingbeat frequency f_{wi}, the most likely class C_j will be determined through a Bayesian classifier. The probability that a measurement f_{wi} belong to the class C_j, $P(C_j|f_{wi})$, is described by Equation (3.3).

$$P(C_j|f_{wi}) = \frac{P(C_j) \cdot P(f_{wi}|C_j)}{P(f_{wi})} \tag{3.3}$$

Where $P(C_j)$ is the prior probability of the class C_j, $P(f_{wi}|C_j)$ is the probability of obtaining the value f_{wi} in the class C_j and $P(f_{wi})$ is the prior probability of the observed frequency f_{wi}. The prior probability, $P(C_j)$, will be chosen as equal for all classes, as it is often the case for field measurements, where it is difficult to know how likely one class is in comparison to another. Since $P(f_{wi})$ is independent of class C_j, and that the prior probabilities of every class are chosen equal, the only parameter that varies from one class to the next is $P(f_{wi}|C_j)$.

The wingbeat frequency distribution of mosquitoes from the same species and sex is known to follow a Gaussian distribution [164, 160, 165]. Thus, $P(f_{wi}|C_j)$ can be evaluated by using a Gaussian probability function also called probability density function as defined by Equation (3.4) and shown on Figure 3.2.

$$P(f_{wi}|C_j) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{C_j}} \cdot e^{-\frac{\left(f_{wi}-\mu_{C_j}\right)^2}{2\cdot\left(\sigma_{C_j}\right)^2}} \tag{3.4}$$

Where σ_{C_j} and μ_{C_j} are respectively the standard deviation and mean value of the class C_j.

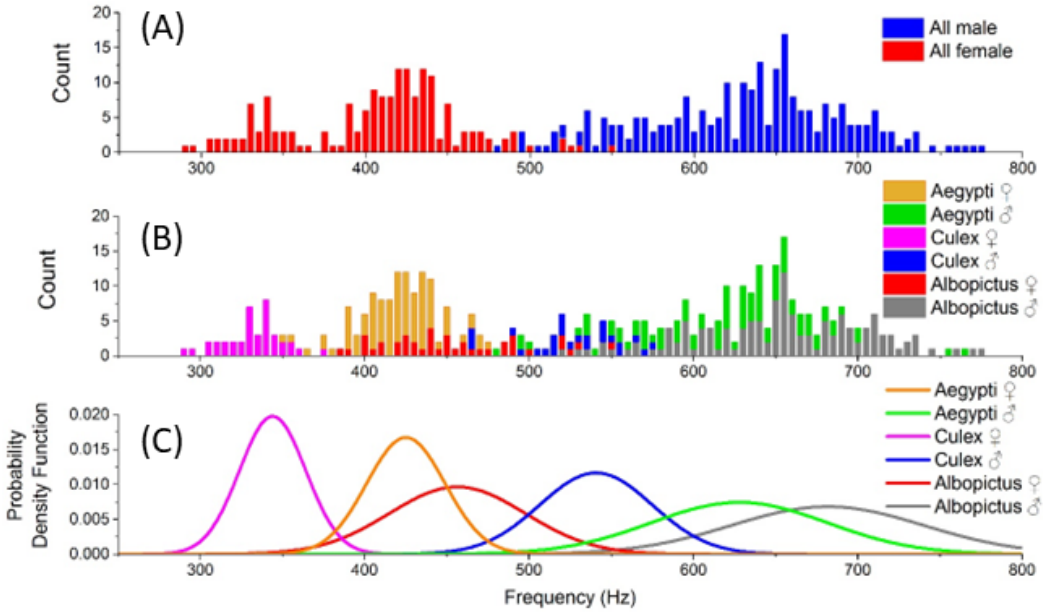

Figure 3.2 Histogram of the measured wingbeat frequencies for all females and males (A), all species and both sexes (B) and their respective probability density function (C). Every bar has a width of 5 Hz and its value is the number of events measured within this frequency window.

For every class a score, $S_{C_j}(f_{wi})$, defined by Equation (3.5) is attributed. This score can be considered as a normalized probability that f_{wi} belong to the class C_j.

$$S_{C_j}(f_{wi}) = \frac{P(C_j) \cdot P(f_{wi}|C_j)}{\sum_{C_k \in C} P(C_k) \cdot P(f_{wi}|C_k)} \tag{3.5}$$

The score is defined in a way that, for each frequency f_{wi}, the sum of the scores for all possible classes is equal to 1. In this regard, the score can be seen as the relative probability of the class C_j in comparison with all other possible classes. Likewise, the ratio of two scores allows for the evaluation of the relative likeliness of one class in comparison to the other. The attributed class (i.e., prediction) of any measurement f_{wi}: $A_{C_j}(f_{wi})$ will be the class C_j for which $S_{C_j}(f_{wi})$ is the greatest, since it is statistically the most likely, but it is not necessarily the correct one.

The only feature used as a predictor variable in this section is the wingbeat frequency. Every class has a distinct average wing beat frequency, which supports

that the wingbeat frequency can be used as a discriminatory factor between classes. Considering the standard deviation of the laboratory gathered data, some overlap between classes is unavoidable and will be the restricting factor of the discrimination attempt. A one-dimension Bayesian classifier, solely based on the wingbeat frequency, was applied using the data regrouped in Table 3.1 and presented in Figure 3.2. When applied to the male vs female classification, also referred to as the sex classification, the score for the male class is equal to a third of the sum of the score for male *Ae. albopictus*, male *Culex* and male *Ae. aegypti* classes. For the female class a perfectly identical consideration has been conducted. Once the Bayesian classification has been applied, the precision of the prediction for each class can be calculated. Data presented in this section are obtained in a controlled environment, thus the actual class of every measurement is known and the veracity of every class prediction can be easily evaluated. Table 3.2 shows the results of the class level precision using a Bayesian classifier, for both classification tasks.

Table 3.2 Precision in Function of the Classification Task

Sex classification		Sex and species classification	
Male	95.8%	*Ae. albopictus*	74.4%
		Culex	36.2%
		Ae. aegypti	45.5%
Female	97.6%	*Ae. albopictus*	45.2%
		Culex	88.6%
		Ae. aegypti	79.4%

For the sex classification, the wingbeat frequency alone allows for a precision of 95.8% and 97.6% for the male and female class respectively and an accuracy of 96.5%. Considering the relative simplicity of the classifier and the unicity of the discriminatory factor, this result demonstrates the effectiveness of the wingbeat

frequency for sex discrimination of the studied species. This is a promising result since the sex differentiation is paramount to the efficiency of the sterile insect technique, which is a cost-effective mitigation technique employed to reduce insect populations such as mosquitoes [166, 167, 168].

For the sex and species classification, the accuracy of the classification drops down to 62.3%. While three of the six classes still have a precision above 74% (female *Culex*, female *Ae. aegypti* and male *Ae. albopictus*), others drop to a lower value, down to 36.2%. The less accurate class still has more than twice the accuracy of a random prediction. This shows that wingbeat frequency is a valuable predictor variable for species identification. On the other hand, it also demonstrates that wingbeat frequency alone is insufficient to reliably classify mosquito species.

When employed in real-life experiments, more than three mosquito species are likely to be present, leading to more classes and larger overlaps between each wingbeat frequency distribution, resulting in lower accuracy. For this reason, the sole wingbeat frequency may be used as reliable predictor variable only in the case where a limited number of mosquito species might be encountered.

3.2.2 Mosquitoes Sex and Species Identification using Multi-Dimensional Naïve Bayes Classifier

The previous subsection was focused on one dimensional Naïve Bayes classifier, it showed that using only the wingbeat frequency was not sufficient for mosquito species identification. Hence, one solution is to explore other possible predictor variables. The goal of this study is to explore which predictor variables are useful for the sex and species classification of mosquitoes. In this experiment the species studied are the males and females *Ae. albopictus*, *Ae. vexans* and *Culex* genus (mixed species of *Cx. pipiens* Linnaeus and *Cx. restuans* Theobald).

Using the three channels of the ESOS systems, six different contributions can be extracted. The body and wing contributions for each of the three recorded channels. However, as discussed in subsection 2.2.6, only the ratios of those six contributions can be exploited as predictor variables, since on their own they would be function of the unknown distance between the insect specimen and the collecting optics. In total there are $\frac{6!}{2!(6-2)!} = 15$ different combinations that can be obtained from those six variables.

In addition to those 15 predictor variables, other characteristics of the mosquitoes can be considered. The most important one being the wingbeat frequency f_w, which was shown in the previous subsection to contribute positively to the sex and species classification. Moreover, mosquitoes' absorption is known to be wavelength dependent [122] which is what motivated the addition of another predictor variable: the ratios of the total backscatter cross section (body + wing) at each wavelength $\frac{\sigma_{1320,/\!/,b}+\sigma_{1320,/\!/,w}+\sigma_{1320,\perp,b}+\sigma_{1320,\perp,w}}{\sigma_{940,b}+\sigma_{940,w}}$, also referred to as $\frac{\sigma_{1320}}{\sigma_{940}}$ for simplicity. The wing to body size ratios can vary from one species to the next, due to species differences in morphological characteristics and composition (i.e., different materials, shapes and sizes), which motivated the use of the ratios of the wing contribution by the body contribution at both wavelengths $\frac{\sigma_{1320,/\!/,w}+\sigma_{1320,\perp,w}+\sigma_{940,w}}{\sigma_{1320,/\!/,b}+\sigma_{1320,\perp,b}+\sigma_{940,b}}$, also referred to as $\frac{\sigma_w}{\sigma_b}$ for simplicity. In total this gives access to 18 predictor variables.

As previously described, a probability density function is constructed for each class and every predictor variable. Those functions are then used to predict the most likely class for each event in the testing set. Unlike the one-dimensional Bayesian classifier the predicted class is now the class that has the highest total score, see Equation (3.6), which is the sum of the score for each predictor variable.

$$TS_{C_j}(E_i) = \sum_{pr=1}^{18} S_{C_j}(E_i)_{pr} \tag{3.6}$$

Where $S_{C_j}(E_i)_{pr}$ is the score of the class C_j for the predictor pr of the considered event E_i. Then the predicted class $A_{C_j}(E_i)$ is the class C_j for which the total score $TS_{C_j}(E_i)$ is the highest. This class is, for the Bayesian classifier, the most likely class but it is not necessarily the correct one since there is some overlap between the different probability density functions.

The number of events in each of the classes is imbalanced, some species have been more active during their laboratory study than others. Therefore, as discussed in subsection 2.3.2, the training set has been randomly undersampled so that when evaluating the accuracy, the number of events in each class is identical. In practice, this introduces uncertainties to the retrieved accuracy, some classes may perform better (or worse) simply because the randomly selected events where far (or close) from the class boundary decision. To mitigate the impacts of the random undersampling, every classification task has been repeated 10,000 times, each time with an equal number of randomly selected events for each class. This will provide 10,000 values of accuracy, one for each repetition, which can be averaged, improving the confidence that can be given to the results and their repeatability.

Using all of the 18 predictor variables, the accuracy of the sex and species classification is equal to 64.34%. The accuracy using only the wingbeat frequency is only of 55.58%, showing that adding more predictor variables has improved the accuracy of the classifier. Howbeit, Bayesian classifiers implicitly assume that all predictor variables are independent, which is not the case for all of the 18 predictors variables. As such, using only a subset of the predictor variables could improve the accuracy of the classification by either removing the noisiest ones or the ones with the greater overlaps between classes.

Selecting the best combination of predictor variables is not trivial. A natural instinct would be to test every possible combination. In practice, with 18 predictors variables, there are 262,143 possible combinations, assuming unweighted score for

each predictor. Coupled with the need for repetition mentioned earlier, testing every possible combination was discarded. Instead of trying every combination, a more practical approach was chosen. It starts from the accuracy results obtained with all the predictor variables $a(\sum_j pr_j)$ as a reference value. Then a given predictor pr_k was removed, leaving only the remaining 17 predictors. The accuracy using the remaining predictors $a(\sum_{j\neq k} pr_j)$ can thus be evaluated. Then the difference between the two $\Delta a(pr_k)$, also referred to as predictor contribution, is calculated as defined by Equation (3.7). Hence if removing the predictor pr_k decreases the accuracy then $\Delta a(pr_k)$ will be positive, on the contrary if removing the predictor pr_k increases the accuracy $\Delta a(pr_k)$ will be negative, and the predictor variable should be removed. The results of the predictors contributions are displayed in Table 3.3.

$$\Delta a(pr_k) = a\left(\sum_j pr_j\right) - a\left(\sum_{j\neq k} pr_j\right) \tag{3.7}$$

Of the 18 predictor variables, eight have a negative contribution, meaning that without them the classification accuracy increases. The remaining 10 predictors have a contribution ranging from 15.4% to 0.06%. As expected, the wingbeat frequency is the one with the highest impact on the classification and its importance far exceeds the one from other positive predictors. The knowledge about the importance of each predictor allows for the removal of the negative predictors from the classification, improving the accuracy of the classification. After removing every predictor variable with a negative contribution, the accuracy reaches 70.65% which is 6.31% more than with all the predictor variables and 15.07% more than with the wingbeat frequency alone.

Yet, one can note that the increase in accuracy is superior to the absolute sum of all the negative contribution. Similarly, it is observed that the predictor contribution varies depending on which subset of predictors is considered. In some cases, one

Table 3.3 Results of Predictor Variables Contributions

Predictor variable	f_w	$\frac{\sigma_{1320,/\!/,b}}{\sigma_{940,b}}$	$\frac{\sigma_{1320,\perp,b}}{\sigma_{940,b}}$
$\Delta a(\%)$	+15,4 %	-1,42 %	+0,27 %
Predictor variable	$\frac{\sigma_{1320,\perp,b}}{\sigma_{1320,/\!/,b}}$	$\frac{\sigma_{1320,/\!/,w}}{\sigma_{940,w}}$	$\frac{\sigma_{1320,\perp,w}}{\sigma_{940,w}}$
$\Delta a(\%)$	+1,48 %	+1,28 %	+1,81 %
Predictor variable	$\frac{\sigma_{1320,\perp,w}}{\sigma_{1320,/\!/,w}}$	$\frac{\sigma_{1320,/\!/,w}}{\sigma_{1320,/\!/,b}}$	$\frac{\sigma_{1320,\perp,w}}{\sigma_{1320,\perp,b}}$
$\Delta a(\%)$	+0,61 %	-0,06 %	-0,32 %
Predictor variable	$\frac{\sigma_{940,w}}{\sigma_{940,b}}$	$\frac{\sigma_{1320,/\!/,w}}{\sigma_{940,b}}$	$\frac{\sigma_{1320,\perp,w}}{\sigma_{940,b}}$
$\Delta a(\%)$	-0,23 %	+0,93 %	+1,09 %
Predictor variable	$\frac{\sigma_{1320,\perp,w}}{\sigma_{1320,/\!/,b}}$	$\frac{\sigma_{940,w}}{\sigma_{1320,/\!/,b}}$	$\frac{\sigma_{940,w}}{\sigma_{1320,\perp,b}}$
$\Delta a(\%)$	-0,1 %	-0,55 %	-0,81 %
Predictor variable	$\frac{\sigma_{1320,/\!/,w}}{\sigma_{1320,\perp,b}}$	$\frac{\sigma_w}{\sigma_b}$	$\frac{\sigma_{1320}}{\sigma_{940}}$
$\Delta a(\%)$	+0,06 %	-0,6 %	+0,83 %

predictor may have a negative contribution but a positive one in another context. This demonstrates the complex dynamics between predictor variables. While the predictor contribution is a useful indication of the quality of a predictor variable, it should be used as a guiding principle and not an absolute rule.

Bayesian classifiers implicitly assume independence between predictor variables. Although they have been shown to still perform well even when it is not the case. At most, there could only be a maximum of seven truly independent predictor variables, derived from the ratios of the wing and body backscatter cross section for each of the three channels plus the wingbeat frequency. This observation has motivated the exploration of the predictor variable combination with a lower subset of predictors. Starting from seven predictors, several combinations have been studied, using the predictor contributions to explore different branches of possibilities. After careful consideration the best set of predictor variables raises the accuracy to 74.95%, which

is 10.61% better than if all 18 predictor variables are included and 19.37% than with just the wingbeat frequency. The best predictor variables subset included only the six following parameters:

- f_w, which is the wingbeat frequency.
- $\frac{\sigma_{1320,\perp,b}}{\sigma_{1320,/\!/,b}}$, which is the depolarization ratio of the insect's body at 1320 nm.
- $\frac{\sigma_{1320,\perp,w}}{\sigma_{1320,/\!/,w}}$, which is the depolarization ratio of the insect's wing at 1320 nm.
- $\frac{\sigma_{1320}}{\sigma_{940}}$, which is the spectral ratio of the insect's total backscatter cross section.
- $\frac{\sigma_{1320,\perp,w}}{\sigma_{1320,\perp,b}}$, which is the wing to body ratio of the insect for the cross polarized light at 1320 nm.
- $\frac{\sigma_{1320,/\!/,w}}{\sigma_{940,b}}$, which is the ratio of the wing backscatter cross section of the insect for the parallel polarization at 1320 nm by the body backscatter cross section of the insect at 940 nm, which has no well-defined physical meaning.

While this subset corresponds to at least a local maximum of the predictor variables accuracy in the combination space, it has not been proven to be the absolute maximum. Consequently, it is possible that another subset could be even better. Yet, it shows a 19.37% increase in accuracy when compared to the one-dimensional Bayesian classifier using only the wingbeat frequency. This demonstrates the benefits of using dual-wavelength and polarization sensitive measurements for the species and sex identification of mosquitoes and likely, for insects in general. However, for many real-world applications, an accuracy around 75% could still be considered as unsatisfactory. A possible improvement may be found in finding additional predictor variables, but also by improving the classifier itself. Indeed, the Naïve Bayes classifier is not the only possibility, thus several other algorithms have been considered.

3.2.3 Comparison of Machine Learning Algorithms for Mosquitoes Sex and Species Identification

Naïve Bayes classification has the advantages of being relatively easy to implement but it is not always the best method for classification. Other machine learning algorithms have been developed throughout the years to tackle complex high dimensional problems. In this section, five common machine learning classifiers have been considered. The Naïve Bayes classifier (NB), as a reference, compared with Linear Discriminant Analysis (LDA), Decision Trees (DT), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), all previously described in subsection 2.3.3. In this comparative study, the species studied are the males and females *Ae. albopictus*, *Ae. vexans* and *Culex* genus (mixed species of *Cx. pipiens* Linnaeus and *Cx. restuans* Theobald).

As in the previous subsection, all classifiers were trained using leave-one out cross validation. The training set and testing set were balanced using random undersampling until the exact same number of events in every class was selected. This process was, as previously explained, repeated and the results averaged in order to mitigate the stochastic fluctuations introduced by the random undersampling.

While the score system of each of these classifiers varies, they all fundamentally share the same basic goal: to give the highest score to the most likely class of any unknown event, similarly to the description given in subsections 3.2.1 and 3.2.2. These five algorithms will be compared on three different classification tasks.

The first and most difficult task, due to its higher number of classes, consists of identifying the sex and the species of mosquitoes from their backscattered signals. This application is the one that would be the most suited for environmental monitoring where the goal is to survey mosquito populations when confronted with the presence of more than one species at a single location. The results of macro average recall and precision are presented in Table 3.4. Macro average recall and

precision are presented with their 95% confidence interval, for each machine learning algorithms.

Table 3.4 Results of Recall and Precision for the First Classification Task

	Macro average recall (%)	Macro average precision (%)
LDA	77.5 ± 10	78.1 ± 9.3
DT	71.6 ± 13	72.0 ± 13
SVM	**79.7 ± 6.8**	**80.5 ± 10**
KNN	65.6 ± 9.5	65.8 ± 8.8
NB	70.8 ± 8.6	72.8 ± 15

Among the five different machine learning algorithms, the one that perform best, highlighted in bold, is Support Vector Machine with a macro average recall of 79.7% and a macro average precision of 80.5%. Which are respectively 8.9% and 7.7% more than the Naïve Bayes classifier.

The second classification task is directed towards identifying the sex alone when different species are mixed, without trying to discriminate between species. With the reduction in the number of classes, there is a reduction in the overlap between them, and therefore an improvement in efficiency. This could be applied to a situation where the evaluation of the number of disease vectors (only females are vectors) is paramount. Indeed, in cases were knowing the exact species is irrelevant, sacrificing the species identification for a higher recall and precision in the identification of sex could be welcomed. The results of macro average recall and precision are presented in Table 3.5. Macro average recall and precision are presented with their 95% confidence interval, for each machine learning algorithms.

Among the five different machine learning algorithms, the two that performs best, highlighted in bold, are Support Vector Machine and Linear Discriminant Analysis. LDA, has a macro average recall of 91.3% and a macro average precision

Table 3.5 Results of Recall and Precision for the Second Classification Task

	Macro average recall (%)	Macro average precision (%)
LDA	**91.3 ± 3.0**	**91.4 ± 2.5**
DT	88.5 ± 0.04	88.5 ± 0.04
SVM	**91.0 ± 1.0**	**91.1 ± 0.08**
KNN	86.0 ± 0.5	86.1 ± 0.4
NB	84.3 ± 0.4	84.3 ± 0.3

of 91.4%, which are respectively 11.6 and 10.9% above the best performant machine learning algorithm of the previous task. It is also clearly an improvement when compared to the Naïve Bayes with respectively 7% and 7.1% more in terms of macro average recall and precision. With only a 0.3% difference in macro average recall and precision, which is within the uncertainty interval, SVM is performing just as well as LDA. As such they can both be considered as the best performing algorithm for this task.

The third task is the differentiation between male and female for a single species, or more accurately for closely related species that are *Culex pipiens* Linnaeus and *Culex restuans* Theobald. As mentioned earlier, the identification of sex for a single species is useful in a laboratory situation such as the sterile insect mitigation techniques. Where releasing sterile males is desired while releasing non-sterilized females should be avoided. The identification of sex by using fast and automated classification through machine learning algorithms could improve the efficiency of the sterile insect technique by combining it with an intelligent trap, as presented in Silva et al. 2016 [169]. The results of macro average recall and precision are presented in Table 3.6. Macro average recall and precision are presented with their 95% confidence interval, for each machine learning algorithms.

Table 3.6 Results of Recall and Precision for the Third Classification Task

	Macro average recall (%)	Macro average precision (%)
LDA	97.1 ± 2.8	97.2 ± 2.6
DT	**99.9 ± 0.05**	**99.9 ± 0.05**
SVM	98.5 ± 0.5	98.6 ± 0.4
KNN	92.1 ± 1.9	92.3 ± 1.6
NB	98.6 ± 0.8	98.6 ± 0.8

For the third task of identification of sex for a single species, DT algorithm is performing extremely well with an almost perfect classification, 99.9% for both macro average recall and precision. Although all other algorithms performed very well with metrics above 97%, expect for KNN. This makes the identification of sex for closely related species of mosquitoes through machine learning algorithms based on dual-wavelength and polarization sensitive backscattered signals a viable option that could be used in practical applications.

Depending on the classification task, the best performing machine learning algorithm changes and so do the performance metrics. The changes in performances, due to different classification tasks, were expected as increasing the number of classes or modifying the classes to be discriminated can impact the overlap in predictor variable distributions. However, the change in best performing algorithm could be seen as less obvious as it can be conceived that a well performing machine learning algorithm could outperform all others on all tasks.

When considering all three classification tasks, the best algorithms are SVM and LDA, while KNN and NB are never the best machine learning algorithms and, in most cases, are the least performing ones. For the KNN classifier, the most logical explanation for its poor performance is that for some of the predictor variables, there is little difference in value distribution making the search for neighboring instance

unreliable for prediction purposes. For the NB classifier, the lack of performance may be explained by an overreliance on the assumption of independence between predictors variables. In all cases, further investigation about predictor variable pre-processing using methods such as principal component analysis could prove to be beneficial by providing uncorrelated sets of lower dimensions to train the machine learning algorithms.

For the monitoring of sex and species of mosquitoes at a location where several species can be expected, Support Vector Machines seems to be the most efficient classification algorithm. This broad approach is the best suited for regular and systematic surveillance campaigns where time resolved activity of each species and sex can be monitored and compared. Decision Trees CART algorithm could then be applied to separate males and females more effectively, potentially providing information on the efficacy of sterile insect mitigation techniques.

3.2.4 Identification of Mosquitoes Gravidity

In the previous subsection, the species and sex identification of mosquitoes has been studied. In this subsection, the differentiation of gravid females (carrying eggs) from non-gravid females (not carrying eggs) mosquitoes is studied. This information can be used to identify possible breeding habitat, predict an increase in the mosquito population and provide a better overall understanding of the ecosystem dynamics. With the exception of a few autogenous species, in order to become gravid, a female mosquito must have had a blood meal in the previous days, as the complete digestion of the blood meal and the development of the ovaries takes between 69 and 97 hours [170]. Adult females lay their eggs in various ways depending on the species. For most species, females will lay around 100 to 200 eggs, in stagnant water. The eggs will then hatch as soon as the conditions, water and temperature, are favorable. It is during the blood meal process that most infectious diseases can be transmitted from

the host to the female. The infected female can then spread this infectious disease to any subsequent host that they will feed upon. Hence, monitoring the gravid female population together with atmospheric conditions may provide information on when newly emerged mosquitoes are likely to become active, as well as estimating the risk of exposure to infectious diseases.

As previously demonstrated, see subsection 3.2.2, dual-wavelength and polarization sensitive backscattered signals can be useful, in addition to the wing beat frequency, for sex and species identification of mosquitoes. This contribution also demonstrates that they can contribute to the identification of the gravidity of female mosquitoes from the *Culex* genus. As displayed in Figure 3.3, the results show that the depolarization ratio of the body is different, on average, between gravid females and non-gravid females, which to the best of my knowledge is the first study to demonstrates this effect. This may originate from the multiple scattering induced by the numerous eggs in the specimen's abdomen. In addition, a slight difference can also be observed in the wingbeat frequency. Indeed, the gravid females have a slightly higher average wingbeat frequency as they compensate for the increase in weight [171]. This difference on both predictor variables is then used to train and test different machine learning classifiers. Figure 3.3 encapsulates the most significant in-between class variation of the predictor variables, yet other variables are also considered for the gravidity identification task.

As in the previous subsection, all classifiers where trained using leave-one out cross validation and random undersampling. In this experiment the species studied are the male and female *Ae. albopictus*, *Ae. vexans* and *Culex* (mixed species of *Cx. pipiens* Linnaeus and *Cx. restuans* Theobald). Only specimens from the *Culex* genus were studied for gravidity. Unless explicitly mentioned otherwise, all classes should be considered as non-gravid.

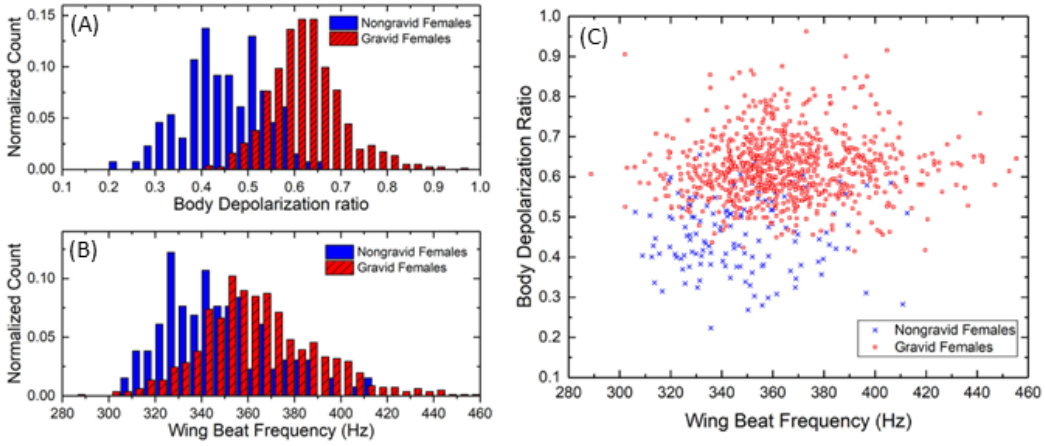

Figure 3.3 Normalized histogram for the depolarization ratio of the body (A) and the wingbeat frequency (B) for the gravid and non-gravid female of the *Culex* genus. Scatter plot for both aforementioned predictor variables and the same classes (C).

The first step in refining the identification of mosquitoes' gravidity is to explore which of the five machine learning algorithms performs best for the gravidity identification task. In this task only two classes are considered, gravid and non-gravid mosquitoes of the *Culex* genus. The results of macro average recall and precision are presented in Table 3.7. Macro average recall and precision are presented with their 95% confidence interval, for each machine learning algorithms.

Table 3.7 Results of Recall and Precision for the Gravidity Classification Task

	Macro average recall (%)	Macro average precision (%)
LDA	**86.6 ± 1.2**	**86.6 ± 0.7**
DT	82.8 ± 0.3	82.8 ± 0.2
SVM	**85.9 ± 3.7**	**86.0 ± 2.6**
KNN	82.2 ± 9.1	82.6 ± 6.0
NB	79.6 ± 18	81.7 ± 11

Among the five different machine learning algorithms, the ones that perform best, highlighted in bold, are Linear Discriminant Analysis with a macro average recall of 86.6% and a macro average precision of 86.6% and Support Vector Machine

with a macro average of 85.9% and a macro precision of 86.0%. As one of the two best classifiers for gravidity identification, LDA has been furthermore studied to obtain class level recall and precision, as presented in Table 3.8. Results are presented with their 95% confidence intervals.

Table 3.8 Results of the 2-class LDA classifier for the Gravidity Classification

	Recall (%)	Precision (%)
Culex female, non-gravid	87.1 ± 0. 16	85.9 ± 0.14
Culex female, gravid	**86.1 ± 0.13**	**87.3 ± 0.17**

These results show that predictor variables extracted from dual-wavelength polarization-sensitive measurements can be reliably used to train an LDA classifier for the purpose of differentiating gravid mosquitoes from non-gravid mosquitoes of the *Culex* genus. When only the wingbeat frequency is considered the recall and precision of the gravid class are lower, by 21.3 and 26.9% respectively. Meaning that the dual-wavelength polarization-sensitive measurements greatly improve the identification of gravidity of *Culex* mosquitoes.

The second task is also directed toward identifying gravid *Culex* mosquitoes but from among a broader pool of possible classes, as can be expected in real-world experiments. This is done by including male and female (all non-gravid) *Ae. albopictus*, *Ae. vexans* and *Culex* mosquitoes to the previously described two classes LDA classification. The results of class level recall and precision are presented in Table 3.9. Class level recall and precision are presented with their 95% confidence interval, for each machine learning algorithms.

These results show that even when the gravid females are to be identified within a larger pool of possible classes, namely other mosquito species and sex, they can still be separated with reasonable accuracy. Moreover, the classifier still performs efficiently for classes other than the gravid one with an accuracy of 75.7 ± 0.13%.

Table 3.9 Results of the 7-class LDA classifier for the Gravidity Classification

	Recall (%)	Precision (%)
albopictus male	84.6± 0.22	75.2± 0.30
albopictus female	77.1± 0.37	70.0± 0.38
Culex male	76.3± 0.37	64.6± 0.29
Culex female	79.3± 0.31	79.4± 0.36
vexans male	72.4± 0.24	77.2± 0.25
vexans female	65.9± 0.30	79.6± 0.35
Culex female gravid	**78.9± 0.31**	**83.7± 0.33**

When compared with the previous case, both the recall and precision of the gravid class have decreased. This was expected as this new task involves more classes and is thus more complex. Yet, even with this broader approach, closer to what can be expected during an actual field campaign, the recall and precision are still around 80%. Showing that the dual-wavelength and polarization sensitive backscattered signals from an ESOS, coupled with an LDA classifier, can identify the presence of eggs in the abdomen of a moving mosquito from several meters away with close to 80% accuracy.

3.3 Laboratory Study of Insects Characteristics from eBoss Data

The final purpose of entomological photonic sensors is to be able to study insects in their natural environment, i.e., outside of laboratory conditions. As demonstrated in the previous section, using the ESOS dual-wavelength and polarization sensitive measurements allows for the identification of mosquitoes' species, sex group and gravidity. However, this 3-channels ESOS has limitations in terms of field application. Operating such a finely tuned system outside of laboratory conditions introduces some challenges. The main issue being the need for long term autonomous operation in

the field. Insect monitoring campaigns usually last weeks to months. Maintaining a perfectly aligned dual-wavelength and polarization sensitive instrument for months at a time is difficult due to thermal expansion and vibrations from the environment. While it could be done with daily calibration and alignment procedures, it would defeat one the fundamental reason for the development of EPS: to monitor insect population with limited supervision, possibly replacing or at least complementing time consuming trapping studies.

To resolve this limitation, two approaches have been considered. First a simpler version of the 3-channels ESOS, see subsection 2.1.1, has been designed. This field ready ESOS operates with a single channel operating at 940nm, as described in more details in subsection 4.1.2. With a single channel, and a larger active area on the detector, the requirements in terms of alignment are greatly reduced and this ESOS can operate for extended periods of time with limited maintenance. Although, without the dual-wavelength and polarization sensitive measurement, and without ranging capabilities, the ability of this ESOS in terms of insect identification is limited. Only the wingbeat frequency and wing to body ratio at 940nm can be used as predictor variables. While that information is useful, and some of those results are used in Chapter 4, it does limit the identification accuracy of the system when compared to its laboratory version.

The second approach was to develop an EPS that would provide more predictor variables than the ESOS in field conditions. This is what has motivated the development of the eBoss, see subsection 4.1.3. With its extinction configuration, range is no longer an issue and this system can extract additional predictor variables, when compared to the single channel ESOS. In particular, the specimen mass along with its wing and body extinction cross sections. The eBoss can also operate for extended periods in the field with limited supervision.

In field conditions, being able to identify the species of the insects that cross the laser beam would be greatly beneficial. It would allow for the study of species-specific behavior and activity. As demonstrated from predictor variables extracted from ESOS signals, machine learning algorithms can be used for such tasks. Evidently, studies in a control environment, where the exact nature of every specimen crossing the beam is known, must first be conducted for the eBoss in order to train such classifiers. The single channel field ESOS does not require such additional studies since it is equivalent to the first channel of its parent design, the three channels laboratory ESOS, that has already been characterized.

3.3.1 Wingbeat Frequency Distribution

As mentioned earlier, the wingbeat frequency of insects is one of their most important characteristics for species and sex identification. The retrieval of wingbeat frequency from the amplitude modulation of an insect signal has been discussed in length in subsection 2.2.2. Using the eBoss, over 96,000 events from 12 different species and sex group of insects have been studied in the laboratory. Figure 3.4 shows the wingbeat frequency distribution retrieved from the laboratory eBoss.

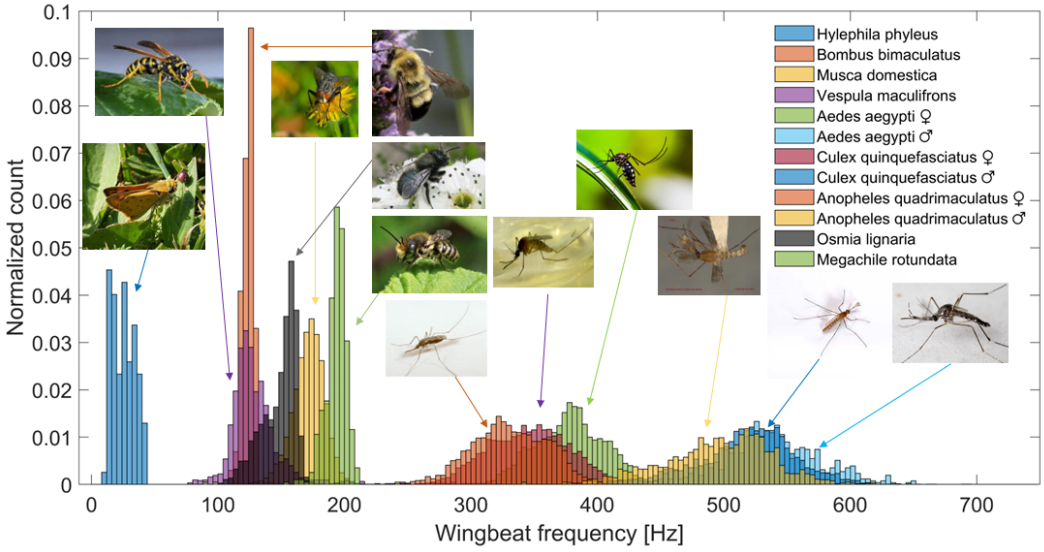

Figure 3.4 Wingbeat frequency distribution retrieved in laboratory conditions by the laboratory eBoss. The count is normalized so that every histogram may be clearly visible. Images, illustrating the morphology of each and every species, were added to help in visualizing the numerous species and sex groups. An arrow, of the same color as the bar plots, was added to indicate which species and sex group each picture corresponds to.

3.3.2 Wing and Body Extinction Cross Sections Distributions

The methodology described in subsections 2.2.4 and 2.2.5 was applied to the 96,000 events previously introduced. From such analyses, both the wing and body extinction cross sections have been retrieved. The results are presented in Figure 3.5.

As discussed in more details in subsection 2.1.3, one of the EPS limitations is the underestimation of the insects' cross section. Values of the body cross sections sometimes approach 0 which, for the bumble bee (*Bombus bimaculatus*) in particular, is too low to be a proper estimation of the cross section. This is due to the situations where one of the insect wings is in the beam, therefore providing its characteristics amplitude modulation, while most of the insect body is outside of the laser beam. In such situation, the event selection algorithm, described in subsection 2.2.1, will consider the event as an insect's event, possessing the periodic amplitude modulation, and thus will not discard it. While extremely low body cross section values could be filtered out, there exist a fundamental limit. This limit is defined by the extinction

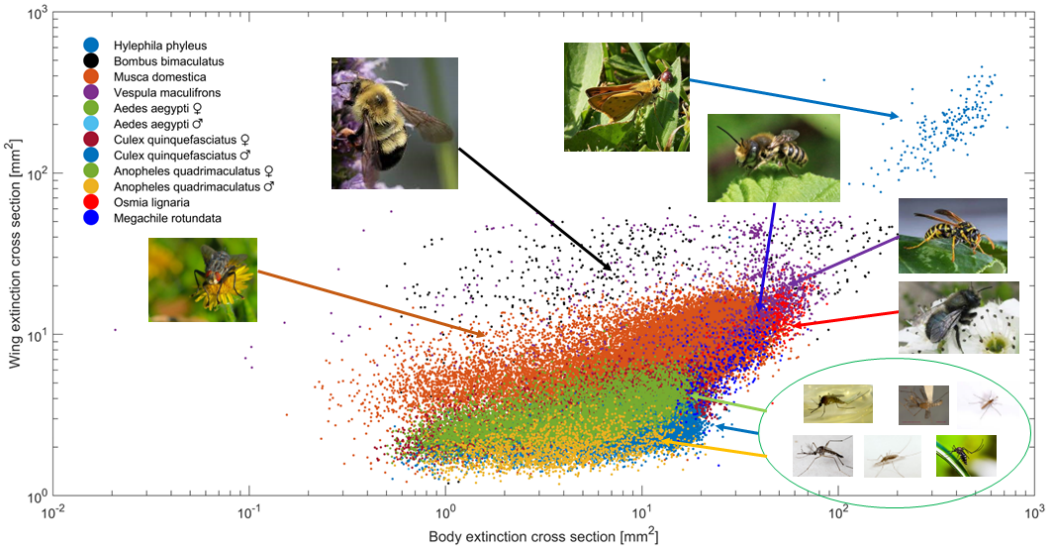

Figure 3.5 Scatter plot of the wing extinction cross section vs the body extinction cross section. Retrieved in laboratory conditions by the laboratory eBoss. Images were added to help in visualizing the different species and sex groups. The mosquito species are bundled together since the overlap is too great to clearly see each one of them separately.

cross section of the smallest insects (or the more transparent insect, or a combination of both) that can be measured, i.e., σ_{min}. Indeed, in its current form, the analysis cannot distinguish between the smallest insect completely within the laser beam and a larger insect partially within the beam. A similar issue arises when considering the wing extinction cross section. However, events for which the wings are not entirely within the laser beam are less likely to pass the harmonicity algorithm. The insects' wings are generally more transparent than their body, i.e., higher quasi-ballistic transmittance. Hence, if their contribution is negligible or not sufficiently periodic (i.e., wings only partially within the beam) the event is likely to be discarded as non-insect event. As such, the distribution of the extinction cross section will be spread towards lower values (left tail), increasing the variability of the measured extinction cross section, which is a predictor variable, reducing the accuracy of the identification task, but without preventing it entirely.

As expected, the wing and body extinction cross section of the three mosquito species overlaps greatly. Although, an unexpected result did arise from this study. It

appears that, for small insects such as mosquitoes, the retrieved extinction cross section is sometimes greater than the geometrical cross section, measured using microscope pictures. Meaning that they appear as if they were absorbing and scattering more photons than their physical size could possibly explain in geometrical optics. While this is still currently under consideration, two main areas of explanation are being considered.

The first possible explanation consists in investigating if this could be the consequence of deviation from the hypothesis of uniformity of the energy density of the laser beam. Indeed, one of the fundamental assumptions is that the light energy density is sufficiently close to uniformity for any variation to have only a negligible effect on the retrieved extinction cross section. If this assumption is incorrect then it is possible to overestimate the extinction cross section. While the insect travels through the area of higher energy density within the beam, the amount of energy absorbed and scattered away per unit of area is greater than anywhere else. As a consequence, it is interpreted as a larger insect traveling through a laser beam where the energy density is equal to the average value, i.e., overestimation of the cross section.

The second possible explanation relies on a more fundamental consideration. If the overestimation does not originate from an experimental imprecision or error, it could be due to an unexplained optical phenomenon. While yet to be fully understood, the phenomenon of measuring an extinction cross section higher than the geometrical cross section has been named, in some studies, as the extinction paradox [172, 173, 174]. When applied to targets such as mosquitoes, no satisfactory answer has yet been found to fundamentally explain the experimental observations. A new experimental protocol, aiming at studying this phenomenon, is currently under examination and theoretical consideration based on the Ewald-Oseen extinction wave [175] are being explored.

As such the extinction cross section of small insects can sometime be overestimated. Effectively, it will spread the distribution of the extinction cross section towards higher values (right tail), increasing the variability of the measured extinction cross section, which is a predictor variable, reducing the accuracy of the identification task.

3.3.3 Mass Calibration

As presented in subsection 2.2.8, Equation (2.29) allows for the retrieval of the mass of an individual insect using the measured extinction cross section of the insect body and the coefficient η. In order to evaluate coefficient η, the laboratory eBoss was used to retrieve the body ECS of insects of known weight. In total, five groups of insects were considered: in order of increasing mass 1) mosquitoes (both *Culex quinquefasciatus* and *Aedes aegypti*), 2) flies (*Musca domestica*), 3) bees (*Osmia lignaria*), 4) wasps (*Vespula maculifrons*), 5) bumble bees (*Bombus bimaculatus*). Insects were weighed immediately upon their death as well as post desiccation using a 0.1 mg precision scale, to determine their wet and dry mass.

A fit of Equation (2.29) provides an estimation of the insect mass m from their body ECS $\sigma_{ext,b}$, see Figure 3.6. In terms of percent error between predicted and actual mass, the worst prediction is for the dry mass of mosquitoes with a 113% percent relative error. However, the absolute error is only 1 mg. Meaning that the dry mass of mosquitoes tends to be overestimated by 1 mg. On the other hand, the best prediction in terms of percent error is for the dry mass of bumble bees with only 3.2% difference, due to an average overestimation of 2.4 mg.

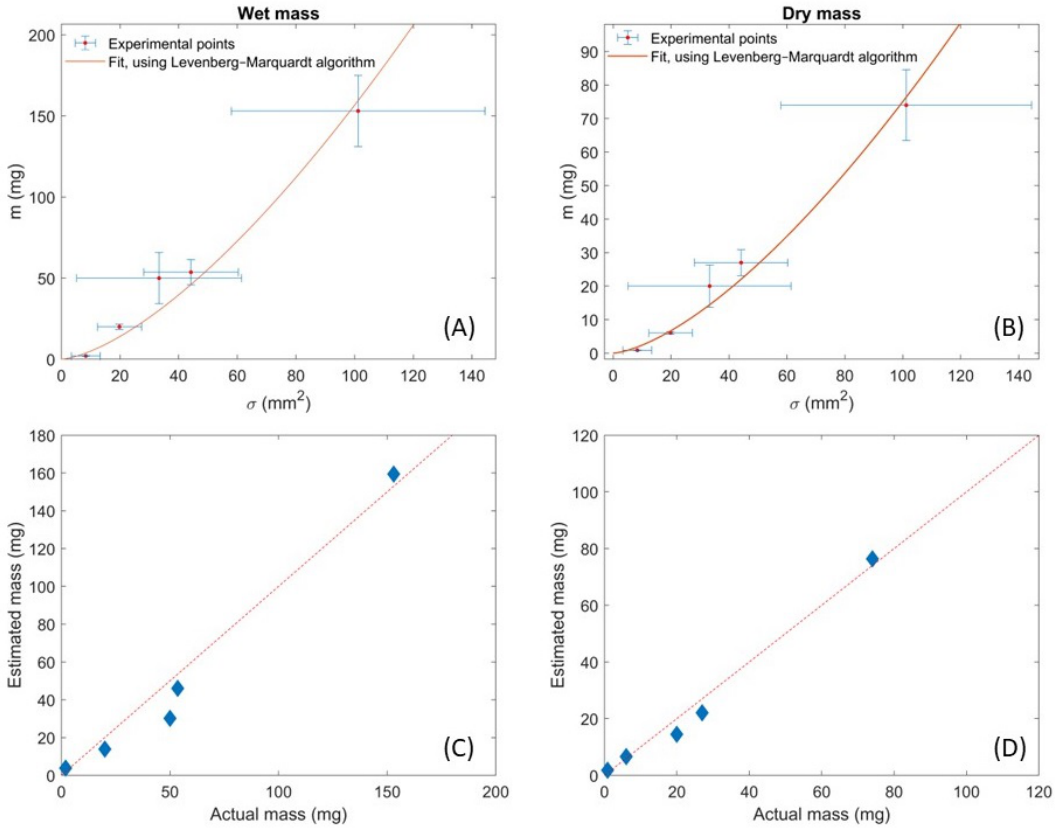

Figure 3.6 (A) and (B) display the result of the fit of Equation (2.29) for the wet and dry mass respectively. Error bars represent the standard deviation. (C) and (D) displays the results of the estimated mass (respectfully wet and dry) using the results of the previous fit and Equations (3.8) and (3.9), in function of the actual mass measured with a 0.1 mg precision scale. Blue diamonds represent the estimation for each of the five insect groups and the orange dotted line is the line of perfect estimation, for which the estimated mass is equal to the actual mass.

Using a Levenberg-Marquardt algorithm [176, 177] to fit Equation (2.29), the values of η can be estimated for both the wet and dry insect mass. This fit allows for the determination of the relation between mass and ECS of the insect body, see Equations (3.8) and (3.9).

$$m_{wet} = 0.157 \cdot (\sigma_{ext,b})^{\frac{3}{2}} \qquad R^2 = 0.96 \tag{3.8}$$

$$m_{dry} = 0.075 \cdot (\sigma_{ext,b})^{\frac{3}{2}} \qquad R^2 = 0.98 \tag{3.9}$$

Where m is the estimated mass in mg (either dry or wet), $\sigma_{ext,b}$ the insect body extinction cross section in mm^2.

3.4 Mel Frequency Cepstral Coefficients as Predictor Variables

3.4.1 Mel Frequency Cepstral Coefficients for Mosquitoes Sex and Species Identification

Analyzing the frequency spectrum derived from the Fourier transform of the amplitude modulation due to the insect wings provides information on the specimen. One piece of information extracted from the frequency spectrum is the wingbeat frequency of the specimen. While this information alone is sufficient for the discrimination between male and female mosquitoes, as detailed in subsection 3.2.1, in other cases it is not sufficient. As an example, for the discrimination between male *Culex quinquefasciatus* and male *Aedes aegypti* the wingbeat frequency distributions, presented in Figure 3.7, are so similar that the capability of the wingbeat frequency as a predictor variable is limited. Indeed, using an SVM classifier for the task of discrimination between males of the aforementioned species, the accuracy is of 53.6%, which is barely above random prediction.

Evidently, the frequency spectrum contains more information than just the wingbeat frequency. This information could be the key to discriminate species with a similar wingbeat frequency. Additional information could be retrieved by studying more subtle properties of the frequency spectrum. To do so, MFCCs are calculated, as descried in subsection 2.2.3. In Figure 3.8, the distribution of the first 12 MFFCs are presented, as an illustration of the information that they can provide. The overlap between distributions for each individual coefficient is comparable to the one between the wingbeat frequency distributions. Yet, when considered not individually but as features from a 12-dimensional space, they can discriminate between the males of the two considered species with a much higher accuracy than the wingbeat frequency alone. Using SVM, with the first 13 MFCCs as predictor variables instead of the

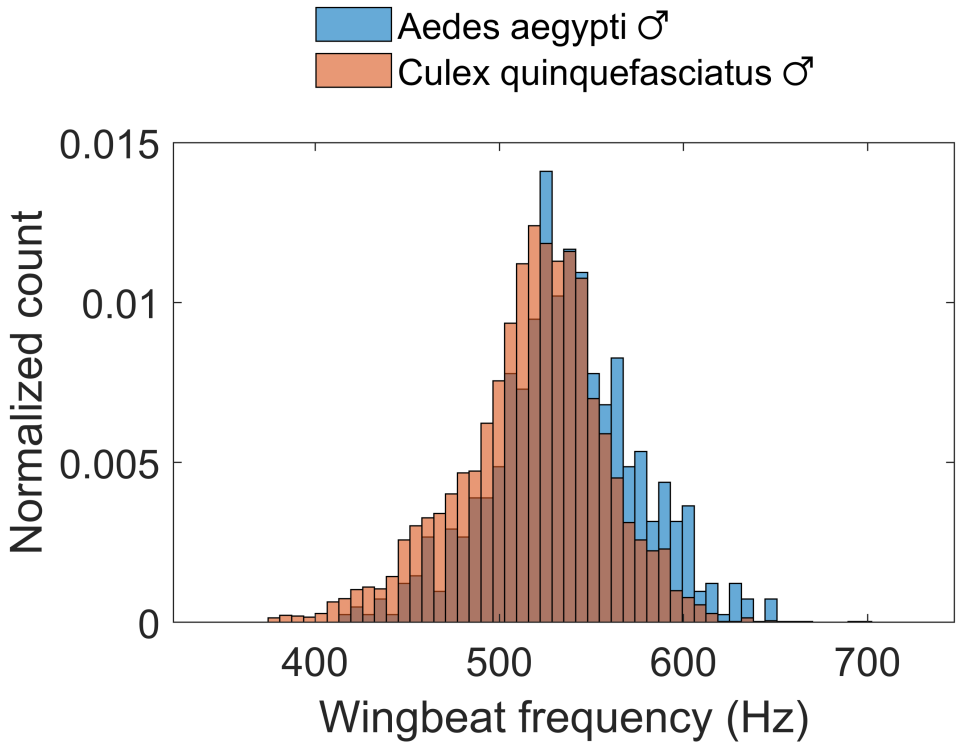

Figure 3.7 Wingbeat frequency distribution for male *Aedes aegypti* (blue) and male *Culex quinquefasciatus* (orange). The count is normalized so that the total area of both histogram is equal to one.

wingbeat frequency, the accuracy of the classification increases to 80.6%, which is 27% above the previous classification task. This clearly shows that the MFCCs provide additional information and that studying the timbre of mosquitoes is useful for the identification of closely related species. Additionally, if the MFCCs are used with the wingbeat frequency the accuracy reaches 83.4%, which shows that the fundamental frequency of the frequency spectrum does contain information that is not entirely described in the first 13 MFCCs.

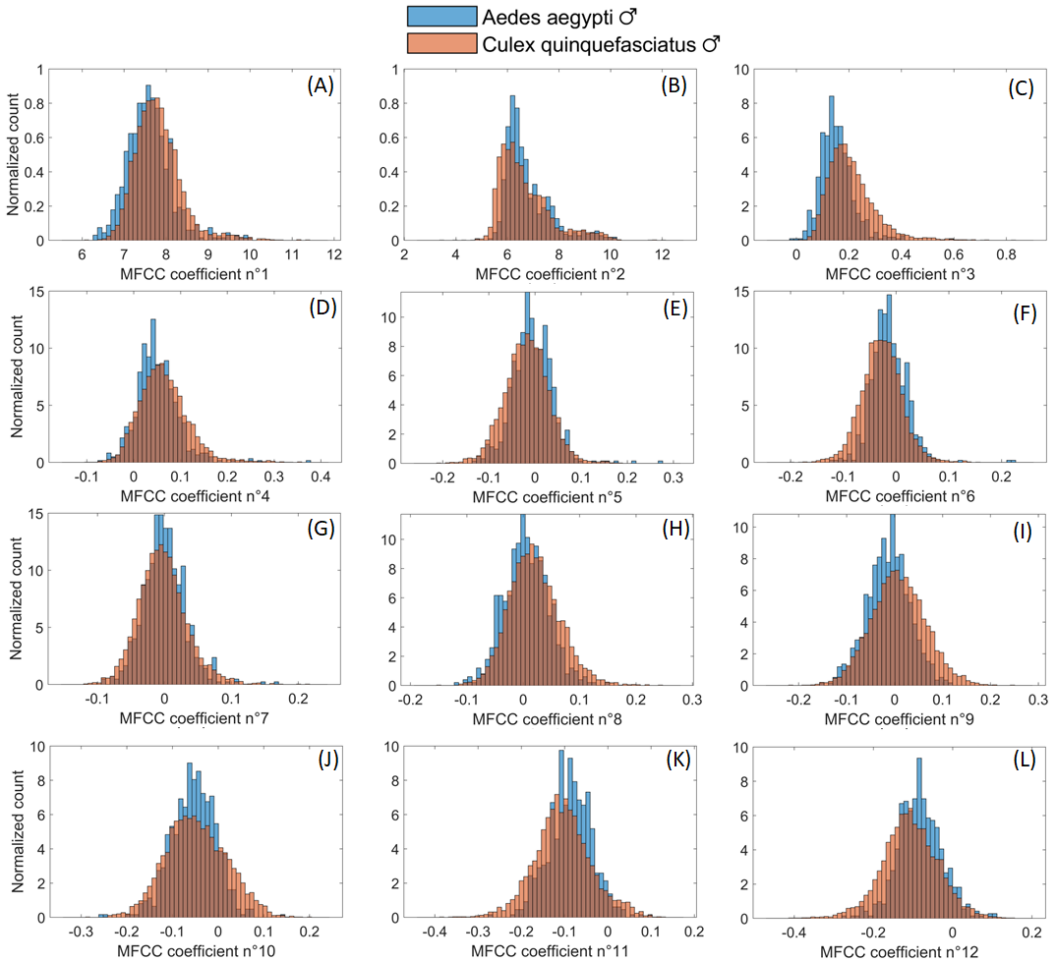

Figure 3.8 Distributions of the first 12 MFCCs for male *Aedes aegypti* (blue) and male *Culex quinquefasciatus* (orange).

3.4.2 Mel Frequency Cepstral Coefficients for the Discrimination between Wasps and Bees

Up to this point, MFCCs have been shown to contain more information than the wingbeat frequency alone but only for mosquito species. While this is an important result, mosquitoes' flight mechanism is significantly different than most other flying insects [178]. Indeed, most flying insects rely on translation-dominated aerodynamics but mosquitoes particularly rely on rotational-dominated aerodynamics. Therefore, in order to generalize the usefulness of MFCCs for insect identification, it is necessary to study more traditional fliers. Similar to the two mosquito species previously studied, the male *Osmia lignaria* (bees) and the *Vespula maculifrons* (wasps) have comparable wingbeat frequency distributions, as shown in Figure 3.9. Using only the wingbeat frequency with an SVM classifier, for the task of discrimination between the aforementioned species, the accuracy is of 69.7%.

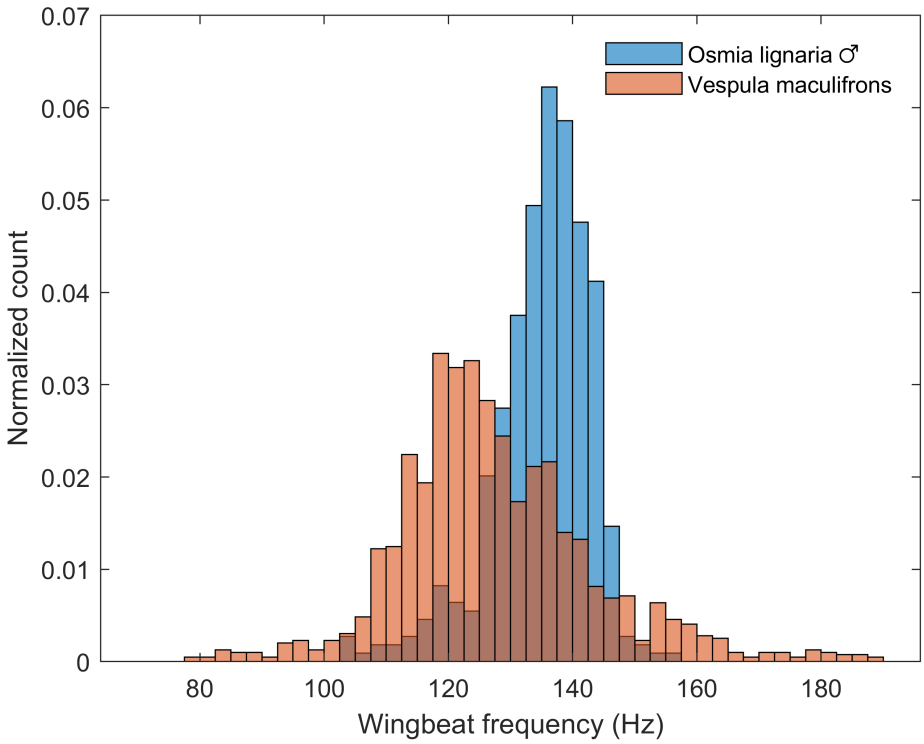

Figure 3.9 Wingbeat frequency distribution for male *Osmia lignaria* (blue) and undetermined sex of *Vespula maculifrons* (orange). The count is normalized so that the total area of both histogram is equal to one.

When using MFCCs, instead of the wingbeat frequency, with an SVM classifier, the accuracy of the classification rises by 21.1%, leading to an accuracy to 90.8%. This illustrates that, regardless of the type of fliers, the MFCCs contain information about the specimen that is not described in the wingbeat frequency alone. Thus, MFCCs are valuable predictor variables for insect identification.

The number of MFCCs was limited to the first 13 coefficients. However, this specific number of coefficients is not a necessary condition and more coefficients could be derived. While preliminary investigation suggests that employing the first 13 coefficients, as in the previous classification tasks, is providing good results, more studies are needed to find the best number and best combination of coefficients. Furthermore, 1^{st} and 2^{nd} order time derivatives of those coefficients can also be considered which adds another level of complexity to the selection process. This is currently under investigation and a new publication on the topic is being considered.

3.4.3 Specimen Orientation within the Laser Beam

The MFCCs effectively describes the influence of the repeated pattern within the insect's wing contribution on the frequency spectrum, as mentioned in subsection 2.2.3. The investigation on the influence of this pattern lead to a collaboration with scientific teams from the University of London and the University of Oxford. As presented in Bomphrey et al. 2017 [178], and illustrated in Figure 3.10, the mosquito model that has been developed by those teams uses a diffusion model combined with an overall mosquito shape, reproducing the outline of the mosquito morphology and its light scattering properties.

From this model, the wing contribution can be evaluated and its shape plotted in function of the specimen orientation within the laser beam, as shown in Figure 3.10. The orientation of observation is defined by two angles: Azimuthal and Polar. The Azimuthal angle, also referred to as Yaw, is equal to 0° when the mosquito is

observed from the anterior side (front) and 180° when observed from the posterior side (back). The Polar angle is equal −90° when observed from the ventral side (bottom) and +90° when observed from the dorsal side (top).

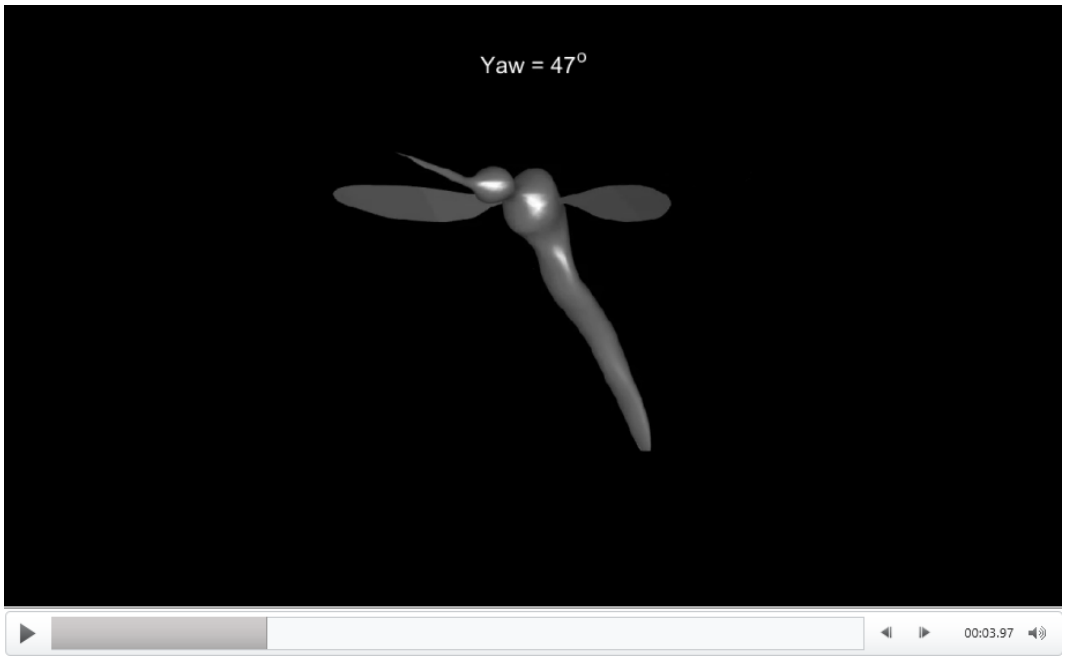

Figure 3.10 Image illustrating the mosquito model. The presented frame represents the mosquito model when observed with a Yaw angle of 47° and a polar angle of 0°. *Image credit: Bomphrey et al. 2017 [178]*

While such consideration cannot fully take into account the specular reflection or the interaction between light and the thin legs, hairs or fibrils of the specimen, it did confirm that from a diffuse standpoint the shape of the wing contribution (i.e., periodic amplitude modulation) to the signal will change depending on the orientation of the specimen within the laser beam, as shown in Figure 3.11.

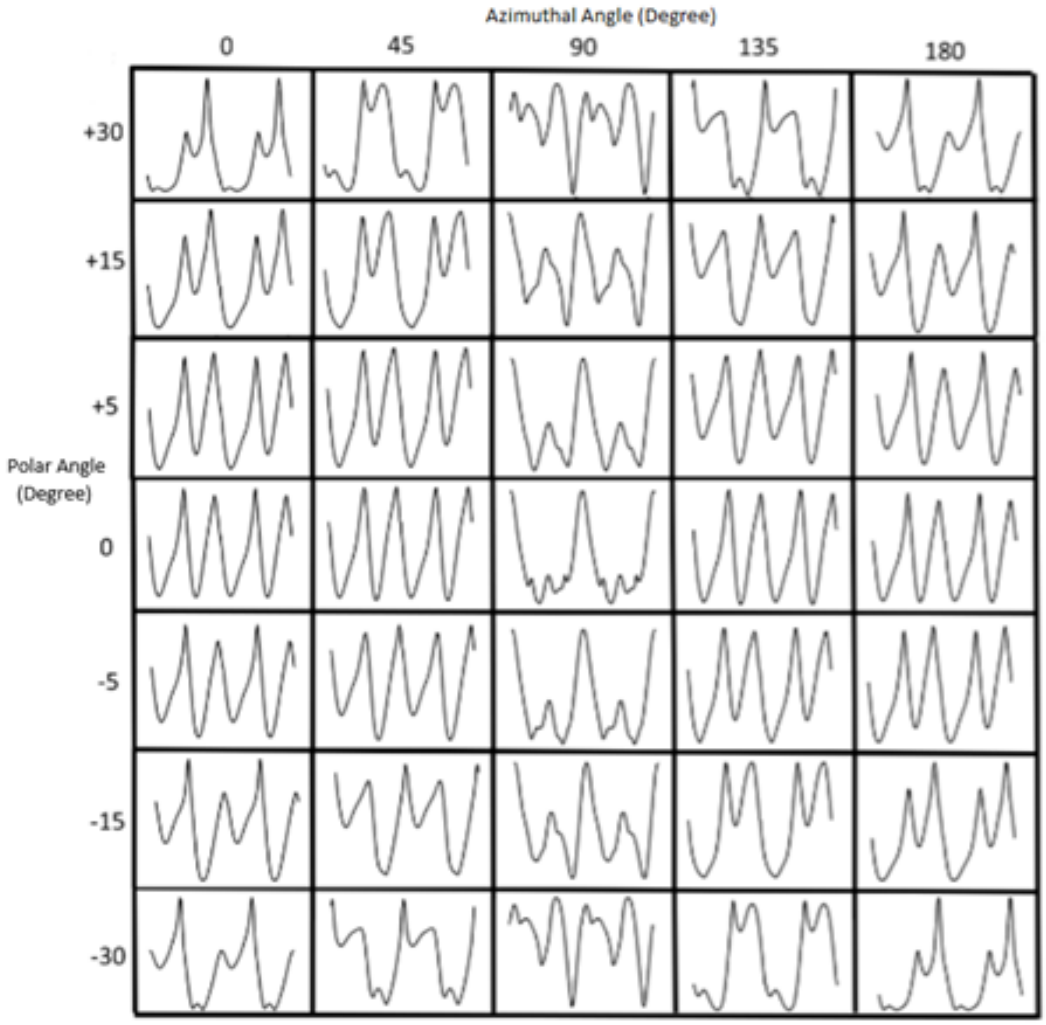

Figure 3.11 Theoretical periodic pattern of the wing contribution for different angles of observation, as predicted by the mosquito simulation. Every square contains two full wingbeat cycles.

The orientation is important to consider as it can influence the value distribution of the MFCCs. A given insect species observed from the side could provide an identical signal to another species when observed from the front. Consequently, the orientation of the specimen within the laser beam, which is unknown in real-world conditions, will lead to an increase of the variability of the MFCCs (and cross sections) distributions, effectively spreading their distributions. In turn, this will have the effect of increasing the overlap between classes, reducing the accuracy of the classification. Admittedly,

the fact that the orientation of the specimen within the laser beam is unknown is a current limitation of the methodology.

CHAPTER 4

FIELD CAMPAIGNS AND RESULTS

The capabilities of EPS in laboratory conditions have now been discussed but ultimately the two systems, ESOS and eBoss, are intended for real-world experiments. Referred to as field campaigns, the real-world experiments consist of deploying the EPS outside where wild flying insects are present in order to study their characteristics and behavior.

4.1 Field Setup and Location

Throughout the course of this project three field campaigns have been conducted, one per year starting in 2020. All three campaigns happened at the same location. The main differences between campaigns are changes in the type and number of systems deployed. The field ESOS, described in more details in subsection 4.1.2, was deployed with little to no modification in its design for all of the three campaigns. During the 2020 campaign only this field ESOS system was deployed. This campaign demonstrated that EPS could operate for extended periods of time at the chosen location. For the 2021 campaign, two more systems where added, the first design of the field eBoss, see subsection 4.1.3. Finally for the 2022 campaign, as in 2021, three systems were deployed but one of the first design eBoss was replaced by a second design eBoss, see subsection 4.1.3.

4.1.1 Field Location

The field campaigns took place in a semi-urban environment, a small patch of green within the city of Secaucus, Hudson County, NJ, USA, see Figure 4.1. Located in one of the world's largest megalopolises, the area has a relatively low density of insects when compared to the wetlands where insects studies are often conducted. The field

is approximately 40 x 10 m with tall grass bordered by a roughly 1 ha woodlot. A co-located portable weather station was operating next to the field where the campaigns took place, allowing for the recording of outdoor temperature, humidity, wind speed, wind direction, rainfall, and solar radiation from the all-in-one sensor.

Figure 4.1 Aerial view of the field location ($40°47'09.8''N\ 74°03'28.1''W$). Both green tents, used to protect the equipment from rain, are visible. The optical path of the laser beams, of both eBoss, is indicated by the orange arrow starting from the emitters and pointed toward the receivers. The optical path of the ESOS is indicated by the red double arrow following along the southern tree line. The weather station is located on top of a metal container located directly south of the field and is indicated by a blue circle.

This specific location has been chosen for several reasons: 1) The field being part of the regular operation sites of one of the project collaborators, no permit was required to access the field location, 2) The field is located within a protected campus, reducing the risk of unwanted human interaction with the instrument, 3) The field allows for an easy access to the electrical grid while being in a patch of green.

4.1.2 Optical Layout of the Field ESOS

The field ESOS system is a laser-based optical sensor that records the light backscattered by any targets that transit through its laser beam. Similar to a lidar system, the laser beam and the telescope optical axis are coaxial and are both pointing in the same direction. As displayed in Figure 4.2, the telescope focuses the light backscattered by insects on an optical detector after passing through an optical filter to reduce the sunlight contribution. The laser source is a 4 W continuous wave diode operating at 940 nm wavelength (L4-9891510-100M; Lumentum, Milpitas, CA, USA) with an initial beam diameter of 2.54 cm and 3 $mrad$ divergence. The telescope is a converging lens with a 25 cm focal length and anti-reflection NIR coating, the detector is an amplified, switchable-gain, Silicon detector (PDA36A2; Thorlabs, Newton, NJ, USA). The backscattered signal is recorded at a sampling frequency of $30,517$ Hz by a 16-bit digitizer (M4i4420-x8; Spectrum, Stamford, CT, USA). The optical path is titled $\sim 3°$ upward, starting about 25 cm above the ground near the emitter and finishing 36 m further in the forest approximately 2.5 m above the ground due to the slight inclination of the field. The instrument is powered by the electrical grid and uses less than 20 W, plus approximately 80 W for the computer used for the data acquisition.

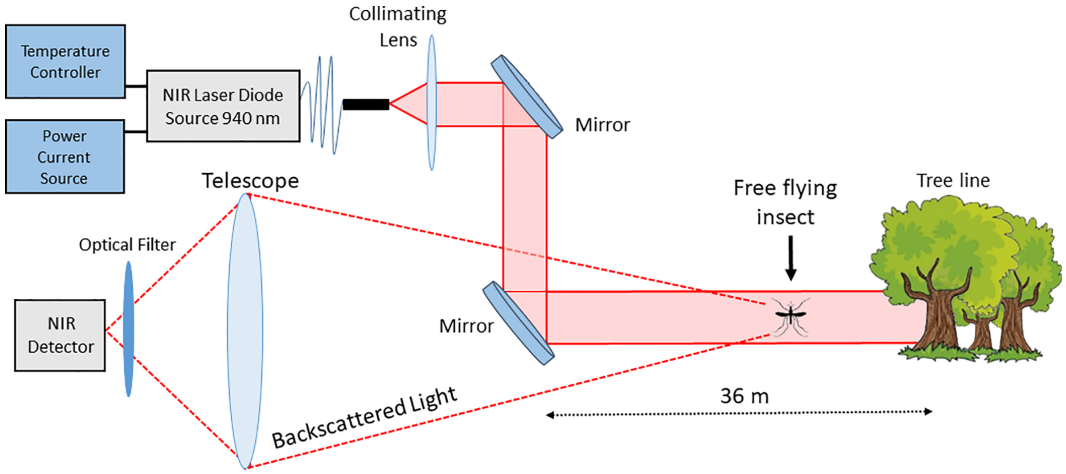

Figure 4.2 Optical layout of the field entomological stand-off optical sensor (ESOS). Divergence of the backscattered light has been exaggerated for ease of representation.

In essence the design of the field ESOS is similar to a single channel laboratory ESOS, but with two noticeable differences. The collecting optic of the field ESOS is larger, with a converging lens of 400 mm diameter, and the field detector has a larger active area ($\sim 13\ mm^2$) than the laboratory one. This allows the system to be less susceptible to misalignment, which reduces the need for re-alignment procedures throughout the campaign.

4.1.3 Optical Layout of the Field eBoss

The field eBoss (both designs) relies on a low intensity continuous laser diode source (CPS980, Thorlabs, Newton, NJ, USA) with a peak optical power of 5 mW, operating at 980 nm. As shown in Figure 4.3, for the first design, the laser beam emitted by the laser diode is expanded using a combination of converging lenses. For the second design the laser beam emitted by the laser diode is expanded using a combination of a convex mirror followed by a concave mirror. In both cases, the collimated laser beam (under 1 $mrad$ divergence) is propagating horizontally, 20 to 40 cm above vegetation, over a range of 36 m. At the end of the optical path, the light is collected by a converging lens, goes through a bandpass filter and is then focused onto the active area of an amplified, switchable-gain, silicon detector (PDA36A2; Thorlabs, Newton, NJ, USA). The optical signal is recorded at a sampling frequency of $30,517$ Hz by a 16-bit digitizer (M4i4420-x8; Spectrum, Stamford, CT, USA).

The second design, implemented for the third campaign, was aimed at improving the eBoss first design and reduce some of its limitation. First and foremost, the lenses forming the beam expander were replaced by mirrors. This was done in order to mitigate an experimental limitation of the first design. In the first design, the expanded beam profile was progressively modified as the beam propagated over the 36 m range. Briefly, some of the light energy was accumulating at the outer edge of the laser beam, inducing an outer ring of higher energy density on the receptor end.

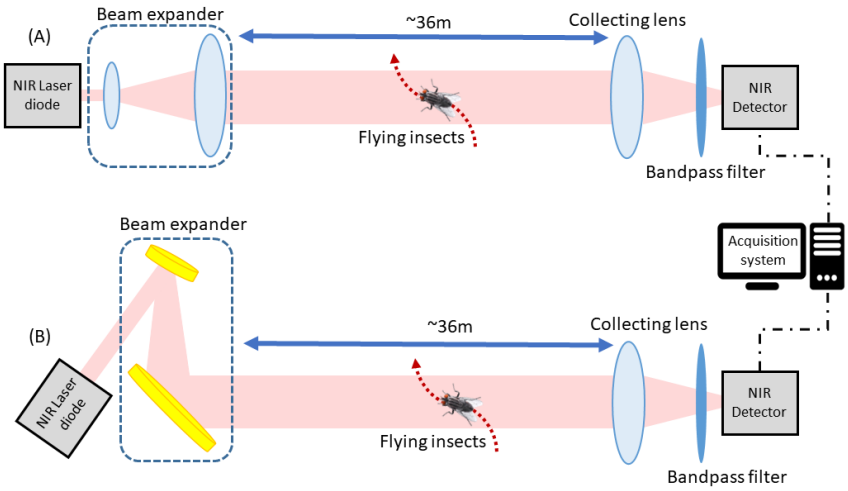

Figure 4.3 Experimental layout of the field Entomological Bistatic Optical Sensor System (eBoss). (A) illustrates the first design, using optical lenses for the beam expansion. (B) illustrates the second design, using mirrors for the beam expansion. Angles and divergence have been exaggerated for ease of representation.

While this issue was mitigated by using a well-placed iris, removing the ring from the field of view of the detector, it still increased the uncertainty of the retrieved extinction cross section by deviating from the beam energy density uniformity assumption. The use of mirrors has greatly reduced this effect making the system easier to operate and the retrieval of extinction cross sections more precise.

4.1.4 4G Network, Weather Station and CDC Light Traps

In order to provide the field with internet access, a 4G LTE router was installed. While not fundamentally necessary, internet access provides several useful services. Operating feedback, such as average signal value and standard deviation, can be sent to the internet data storage (or "cloud") where they can be remotely monitored by the user to ensure proper operation of the system without the need to physically visit the field location. Transferring the entirety of the raw data, and not only operating feedbacks, to the cloud was considered. However, each systems generates approximately 3GB of data par day, making the cloud storage of the raw data

relatively expensive. In addition, the location of the field did not allow for a sufficiently stable connectivity to reliably upload the data. Instead, a local storage of the raw data was chosen.

A portable, solar powered, weather station (WS-1002-WIFI, Ambient Weather, Chandler, AZ, USA) was installed at close proximity to the field, on top of a metal container located directly south of the field, see Figure 4.2. This station was used to monitor several key environmental parameters, such as UV radiation, temperature, wind direction and speed, relative humidity, rainfall and atmospheric pressure. Those measurements are saved both locally and in the cloud on the ambient weather network.

In addition, eight New Jersey light traps were operated across Hudson County, NJ throughout the study period, see Figure 4.4. The traps were deployed by the Hudson Regional Health Commission for monitoring mosquito populations. The traps used a 25-watt incandescent bulb to attract flying insects. A light sensor caused the trap to run every night from dusk to dawn. As insects flew near the bulb, a fan sucked them into a collection jar. The jar contained dichlorvos impregnated strips (Nuvan Prostrips, AMVAC, Los Angeles, CA, USA) to kill the insects. Specimens were collected two times a week and returned to the laboratory for identification. All *Culicidae* (mosquitoes) were retained and identified to the species level while other insect families were discarded. The relative counts of those mosquito traps can then be compared to the results obtained from entomological photonic sensors. As the EPS methodology is new, it calls for a validation by comparison with well tested and widely approved methodologies. The study of insects is often conducted with traps, which counts the number of captured specimens to study the abundance of insects.

Figure 4.4 Aerial view of the experiment location ($40°47'09.8''N$ $74°03'28.1''W$) and its surroundings, the U.S. Northeast megalopolis (Manhattan Island top right corner). The field location and the co-located weather station are symbolized by a black marker. The positions of the New Jersey light traps are indicated by a red marker.

Image credit: NASA's Earth Observatory, NASA/GSFC/MITI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team.

4.2 2020 Field Campaign

From the measurement conducted in real-world conditions, i.e., field measurements, important information can be extracted. First the wingbeat frequency of every observed insect can be retrieved in order to create and study the overall wingbeat frequency distribution. This can provide information on which species are present in the field, a priori unknown. Additionally, using Equation (2.23), the aerial density of flying insects can be studied on any user defined time resolution. A weekly to monthly time resolution would be best suited to study the seasonal trend of insect populations while a shorter, hourly or minute, time resolution would be useful to understand insects' circadian rhythms. Additionally, this information can be coupled with the environmental parameter recorded by the weather station, in particular the temperature and sunlight level (i.e., UV radiation) which both have a significant influence on insects' behavior.

The 2020 campaign was the first deployment of an EPS and as such served as a proof of concept and viability study. The measurement campaign took place from August 19^{th} to November 6^{th}, 2020. Over that period, the instrument was temporarily shut down every three to four days for checkups and routine verifications, resulting in an overall down time of 1.2%. Over 80 days the instrument detected a total of 72,975 transit signals, from which 59,105 (80.1%) presented acceptable harmonic series in the frequency domains and were retained as originating from insects for further analysis. From each of these transits, the wingbeat frequency, wing to body ratio and transit time were recorded together with the data collected by the weather station.

4.2.1 Temperature Correction

The wingbeat frequency of many insect species is known to change with air temperature due to the change in air density as well as the change in the insect's body temperature. While the exact relationship between ambient temperature and

wingbeat frequency is complex and beyond the scope of this contribution, it appears to be impacted by the size of the insects as well as the wingbeat frequency itself [151, 154]. With the field experiment starting in August and finishing in November, the air temperature observed during the experiment went from a minimum of $-2.2°C$ to a maximum of $35.9°C$. A local measurement of the air temperature may allow for the correction of the wingbeat frequency measurements, potentially avoiding any source of bias due to temperature. Figure 4.5 (A) presents a color plot of the measured wingbeat frequency as a function of the ambient temperature, while Figure 4.5 (B) and (C) are the wingbeat frequency distributions for two ranges of temperature, 22 to $26°C$ and 8 to $12°C$ respectively. Those histograms, and several other in the rest of the manuscript, are presented not in terms of count but in terms of aerial density per bin of frequency. Using Equation (2.23), the aerial density of the insects within each bin of frequency is calculated. With a temporal resolution, ΔT, over the entire period of measurement, from the first to the last second of recording of the considered system. Figure 4.5 (A) shows that the wingbeat frequency increases with the temperature and that the greater the wingbeat frequency, the greater the influence of the temperature. This is also illustrated by the change of the slope with f_0 in Equation (4.1).

From Figures 4.5 (B) and (C), the broadening of the insects' clusters, defined in more details in the following subsection, and the increase of the wingbeat frequency with the temperature can be observed. At every recorded temperature, the wingbeat frequency distribution presents four insect clusters, presented in Figure 4.6. The black lines in Figure 4.5 (A) show the linear fit of the local maximum for each of the four clusters at different temperatures. A correction in respect to temperature, applied to the measured wingbeat frequency, is derived from these linear fits and the resulting distribution of all wingbeat frequencies corrected in temperature is displayed in Figure 4.6 (A). The empirical correction is presented in Equation (4.1)

and was experimentally derived from a power law fit on the slope coefficients of the aforementioned linear fits of the maximums.

$$f_{corr} = f_0 - (T_0 - T_{ref}) \cdot \left(0.001189 \cdot f_0^{1.561}\right) \tag{4.1}$$

Where f_{corr} is the wingbeat frequency corrected in temperature, f_0 the original wingbeat frequency retrieved from the harmonicity algorithm, T_0 the temperature at the time of measurement and T_{ref} the reference temperature, chosen to be $20°C$. This correction term is in good agreement with other studies on the relationship between wingbeat frequency and ambient air temperature. In the study by Unwin et al. 1984 [154], a variation of the wingbeat frequency between 1.2 to 4.6% per degree was found for the stingless bee *Trigona jaty* (Hymenoptera: *Apidae*). Using Equation (4.1), the correction for insects with the same fundamental frequency as *T. jaty* would be 2.1% per degree, which is within the interval found in the aforementioned study. Similarly, in the study conducted by Oertli, 1989 [179], the correction for the click beetle *Agriotes sputator* (Coleoptera: *Elateridae*) and the eastern firefly *Photinus pyralis* (Coleoptera: *Lampyridae*) was found to be between 0.1 to 2.2% and 0.9 to 1.6% per degree, respectively, while Equation (4.1) yields a correction of 1.6 and 1.2%, respectively.

Unless otherwise specified, every subsequent value of wingbeat frequency is corrected in regards to temperature, using Equation (4.1).

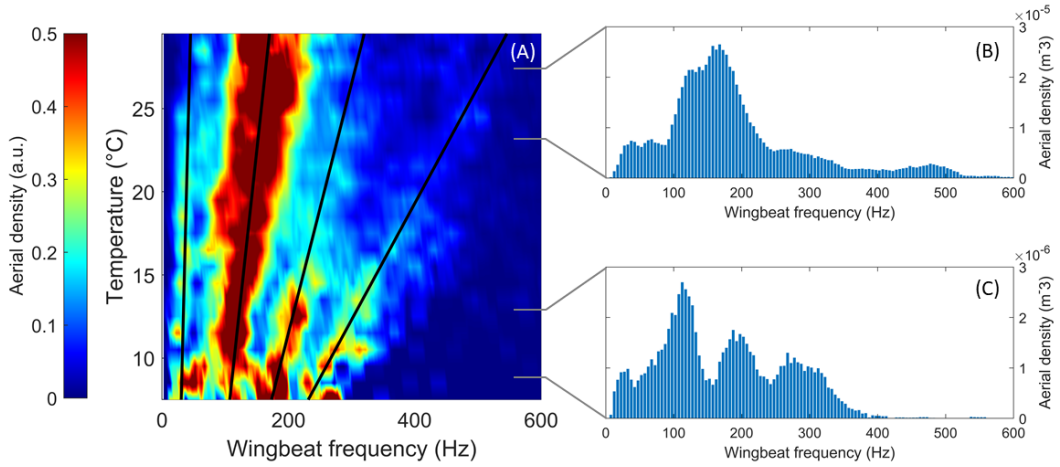

Figure 4.5 Relative aerial density per bin of temperature and wingbeat frequency (A) along with the wingbeat frequency distribution for two ranges of temperature, 22 to 26°C (B) and 8 to 12°C (C). On (A) the aerial density was rescaled from 0 to 1, for each temperature bin, by normalizing with the maximal value in order to emphasize local maximums. Black lines represent the linear fit of local maximal for each of the four insects cluster, illustrating the drift of the wingbeat frequency with the temperature.

4.2.2 Cluster Identification

Despite improvements in the last few years [74, 108, 156, 180, 181, 182, 183, 184, 185, 186], species level identification from infrared optical sensors has yet to reach practical reliability for field campaigns. Although, it is possible to identify different clusters of insects with shared characteristics, such as wingbeat frequency, ratio between the wing and body backscatter cross section (also referred to as the wing to body ratio) and activity patterns. Clusters still provide relevant information, especially for mosquito species since their high wingbeat frequency makes their identification from other insect families more reliable. Figure 4.6 displays the four distinct insect clusters C1, C2, C3 and C4 that have been identified from the field data.

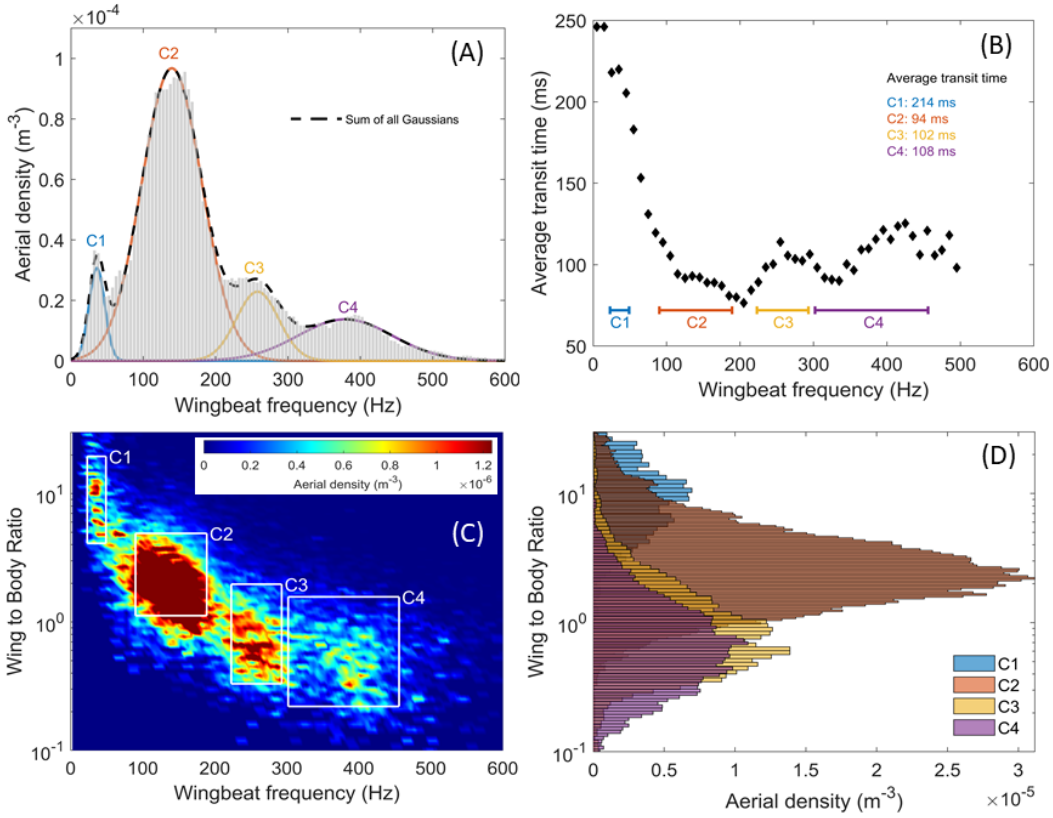

Figure 4.6 Distribution of the wingbeat frequency corrected in temperature (A) with four Gaussian fits corresponding to each insect cluster, and the total reconstructed Gaussian fit (dashed black line). Average transit time in function of the wingbeat frequency (B), the different cluster range are displayed, and their mean transit time indicated. Aerial density per bin of wing to body ratio and wingbeat frequency (C), the boundaries of the four insect clusters C1 to C4 are indicted by white lines. Distribution of the wing to body ratio separated per cluster (D).

Figure 4.6 (A) shows that the wingbeat frequency distribution of all events can be effectively fitted by four Gaussian distributions, which suggest the presence of a least four insect clusters. The boundaries of each cluster in terms of wingbeat frequency are defined by the full width at half maximum of each Gaussian fit. Figure 4.6 (D) displays the wing to body ratio of each cluster, showing a normal distribution. The boundaries of each cluster in terms of wing to body ratio are defined on the same model as the ones for the wingbeat frequency, i.e., full width at half maximum. It is interesting to note that the wing to body ratio distributions of each cluster do overlap, especially for cluster C3 and C4. This suggests morphological similarities between the insects present in both clusters. Figure 4.6 (C) shows the aerial density

as a function of the wing to body ratio and wingbeat frequency. The four clusters and their boundaries are illustrated by the white boxes and labeled C1 to C4. Finally, the average transit time as a function of the wingbeat frequency is presented in Figure 4.6 (B). Because insects may change directions multiple times while within the laser beam or fly along the optical axis, the path length is unknown and, thus, the flight velocity is unknown. However, once averaged, transit times as a function of wingbeat frequency indicate that some clusters include insects with different flight speed, with C1 having the highest transit time, suggesting slow insects and C2 with the lowest transit time, suggesting faster insects.

This analysis suggests the presence of at least four insect clusters that possess sufficiently distinct characteristics to be discriminated from one another. Within some of these clusters, it is likely for more than a single insect family to be present and a more refined analysis or experimental improvement are needed to discriminate down to a family or species level. While it is beyond the current capability of the ESOS instrument to identify with high certainty the family and species present in each cluster, the information available can be coupled with the information collected from insects captured in traps or through net sweeping. Based on the measured wingbeat frequency and the wing to body ratio, the following paragraphs offer suggestions as to which families/species could be found in each cluster:

- Cluster C1: The first cluster is composed of insects with low wingbeat frequency, between 23 and 49 Hz, large wing to body ratio between 4.1 and 19.5 and low velocity with an average transit time of 214 ms which is almost twice as every other clusters. This cluster may include crane flies (Diptera: Tipluidae), lacewings (Neuroptera: Chrysopidae), as well as moths and butterflies (Lepidoptera).

- Cluster C2: The second cluster is composed of insects with wingbeat frequency, between 90 and 189 Hz, medium wing to body ratio between 1.1 and 4.9 and the highest velocity with an average transit time of 94 ms, which is the lowest among all clusters. This cluster may include dragonflies and damselflies (Odonata),

wasps (Hymenoptera: Vespidae), beetles (Coleoptera) and flies (Diptera).

- Cluster C3: The third cluster is composed of insects with wingbeat frequency, between 223 and 293 Hz, small wing to body ratio between 0.33 and 2.0 and an average transit time of 102 ms. This cluster may include bees (Hymenoptera: Apoidea), leafhoppers (Homoptera: Cicadellidae) and flies (Diptera).

- Cluster C4: The fourth cluster is composed of insects with the highest wingbeat frequency, between 302 and 456 Hz, the smallest wing to body ratio between 0.15 and 1.3 and an average transit time of 108 ms. This cluster may include mosquitoes (Diptera: Culicidae) and midges (Diptera: Chironomidae).

4.2.3 Circadian Rhythm

The ESOS system benefits from a high temporal resolution when compared to traps, possibly down to a single minute resolution. It can retrieve the number of flying insects per meter cube of air, or aerial density, on a user defined temporal resolution. As discussed in subsection 2.2.7, the chosen temporal resolution directly impacts the uncertainty of the retrieved aerial density. As it relies on a statistical analysis, see Equation (2.23), the estimated aerial density is subject to stochastic fluctuations. These fluctuations, or uncertainties, are inversely proportional to the number of transit signals. Therefore, a smaller time resolution implies a higher uncertainty. Similarly, fewer flying insects and a smaller probed volume results in greater uncertainties on the aerial density. This entails that, for a one-minute resolution to be meaningful, the probed volume and/or the aerial density of insects must be sufficiently large. In this study, of a semi-urban environment with only 36 m of unobstructed range, the probed volume and insect abundance did not allow for such a resolution to be used with sufficient accuracy. Consequently, the one-minute resolution was averaged over a sliding window of one hour to mitigate the uncertainty by increasing the parameter T of Equation (2.23). Figure 4.7 shows the circadian rhythm of each of the four clusters, in the form of the aerial density over 24h, obtained

from the application of a one hour sliding window over two weeks of continuous recording.

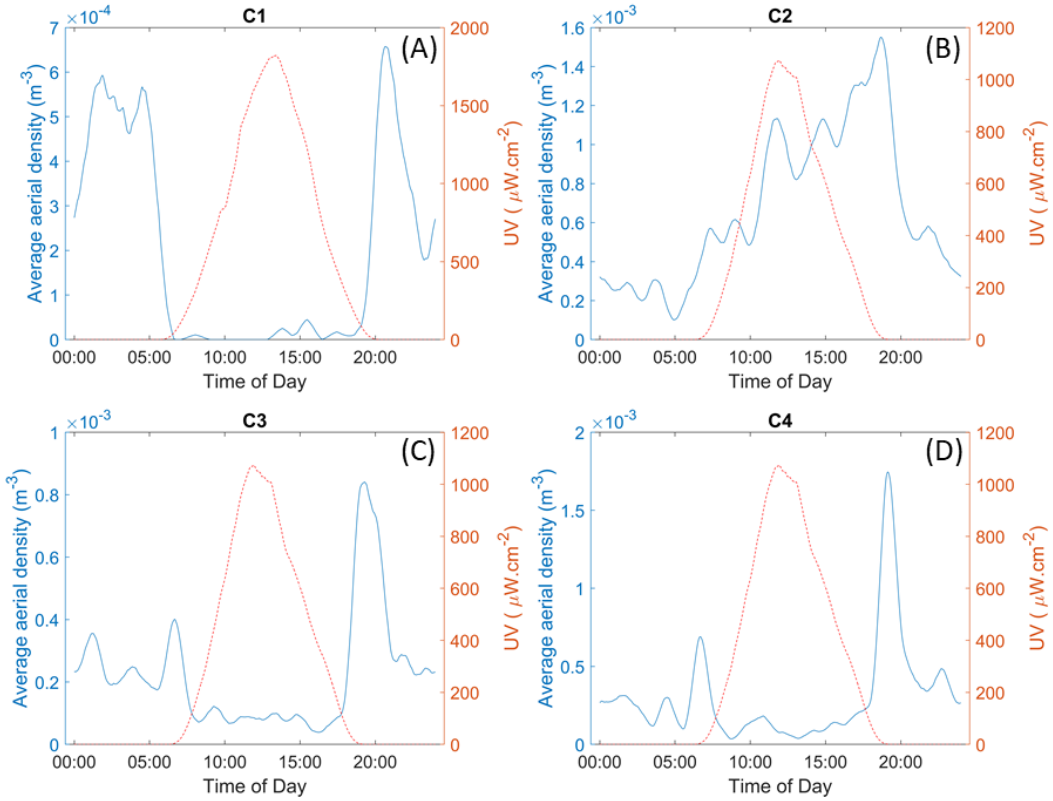

Figure 4.7 Typical circadian rhythm for cluster C1, C2, C3 and C4 (respectively A, B, C and D). The dotted line displays the UV radiation measured on the field which relates to the sun activity (day/night cycle). The average aerial density is the one hour sliding average over 14 consecutive days of measurements.

The circadian rhythm of the second cluster, Figure 4.7 (B), displays activity predominantly during the day, which provides additional information on cluster composition, as it must include mainly diurnal insects. Clusters C1 and C3 show activity mainly during the night with peaks of activity at sunset and sundown. Finally, cluster C4 displays activity almost entirely limited to dusk and dawn, which is consistent with the cluster being mainly composed of mosquito species.

Mosquitoes are of particular interest and have the significant advantage to be efficiently differentiated from other insects due to their relatively high wingbeat frequency. Thus, the cluster C4, henceforth referred to as the mosquito cluster, has been studied in more detail. As mentioned earlier, peaks of activity of mosquitoes

have been observed around sunrise and sunset time. A peak of activity is defined as a local maximal of aerial density (or biomass, see subsection 4.3.5), during which a statistically significant increase of insect activity can be observed, as illustrated in Figure 4.8. In order to study the influence of the sun on the activity of mosquito a simple algorithm was designed to identify peaks of activity within the raw data. As the goal is to track if the peaks of activity of mosquitoes follow sunrise and sunset times, the raw data were separated into two distinct time domains. One between 3am and 10pm (dawn time domain) and the other between 4pm and 11pm (dusk time domain). For every day separately, and for both time domains, the time of day at which the maximal aerial density occurs is determined. This gives access to the time at which the mosquitoes where the most active. Yet, not every local maximum can be considered as a peak of activity. Indeed, in some cases the local maxima is relatively close to the surrounding value of aerial density. In which case it cannot be said that mosquitoes are significantly more active at this time, and as such it would be misleading to consider this time as a period of increased activity. Instead, such cases are excluded and no peak of activity is found. This was done by applying the simple threshold condition on the prominence of the detected local maxima. Prominence being a metric evaluating how much a given peak exceeds its neighbor, i.e., how much does the peak stick out in comparison to its surroundings.

The peaks of activity of the mosquito cluster near sunset and sunrise are furthermore studied and the results presented in Figure 4.9. As shown in this figure, the peak of activity of this cluster follows the shift in time of sunset and sunrise throughout the season, showing that the behavior of the mosquito species present on the field is correlated with the sun activity, as expected. On average, the peak of activity of mosquitoes near sunset is 38 minutes after the actual sunset time, as provided by the National Oceanic and Atmospheric Administration Solar calculator, while the peak of activity near sunrise is on average 17 minutes before the actual

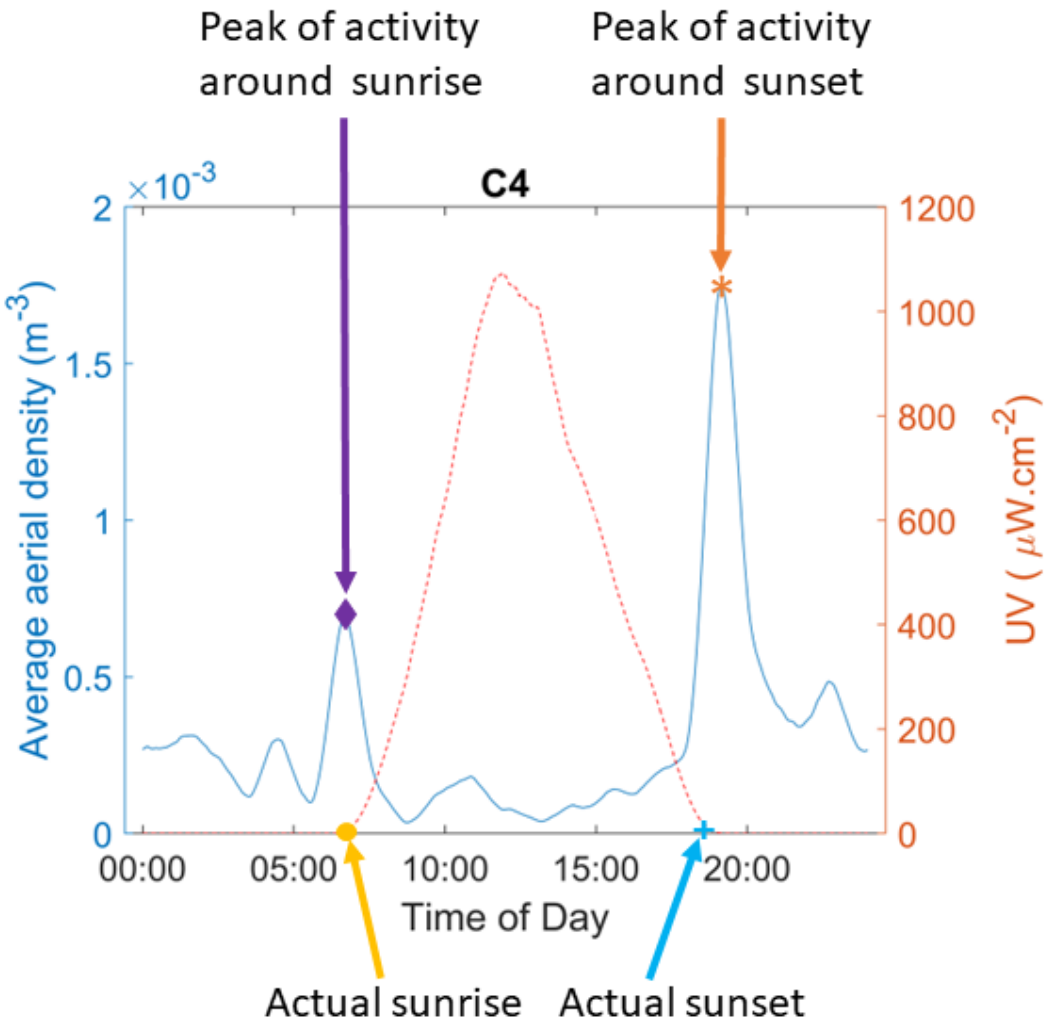

Figure 4.8 Illustration of the definition of a peak of activity. Orange star is the location of the detected peak of activity around sunset while the purple diamond illustrated the one around sunrise. The blue plus sign shows the actual sunset time while the yellow disk illustrates the actual sunrise time.

sunrise. Ultimately, using those circadian rhythms must be done with care. Without a species level identification, each cluster is likely to encompass several distinct species and as such can only display the overall circadian rhythm of the entire cluster. This amalgamated information could hide the specific behavior of the individual species present within the cluster.

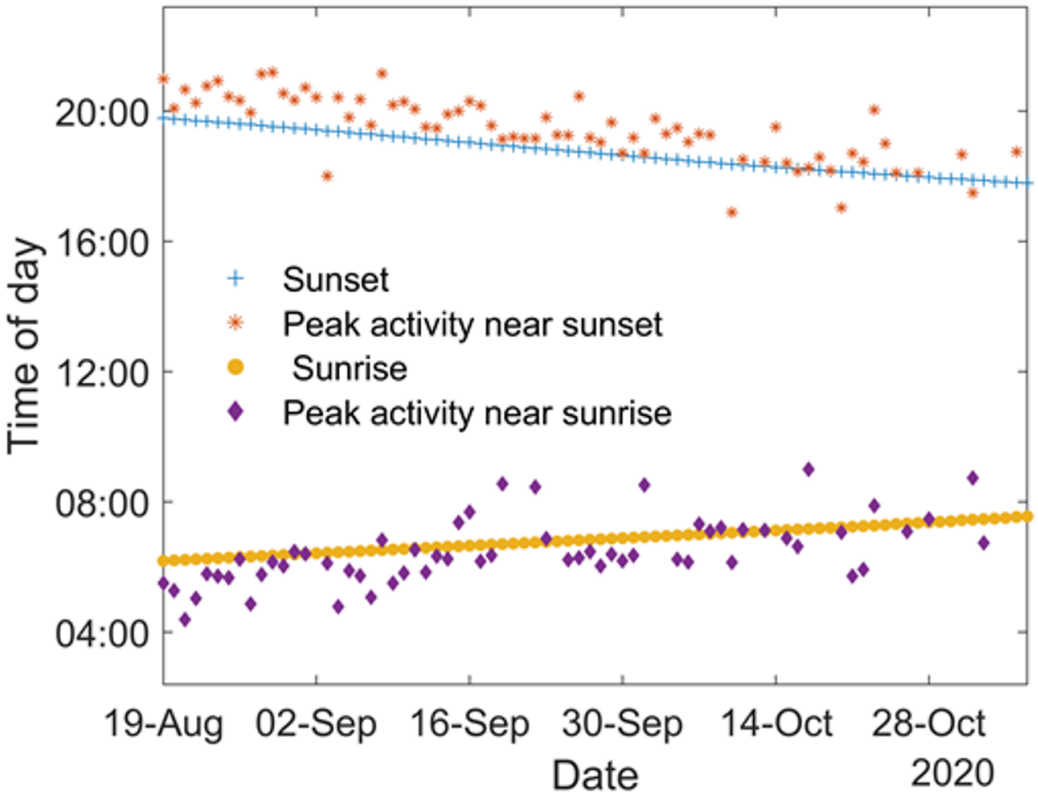

Figure 4.9 Peaks of mosquito activity (Cluster C4) near sunset (orange stars) and sunrise (purple diamonds) in relation with the actual sunset (blue plus) and sunrise (yellow disks) time.

4.2.4 Seasonal Variation of Aerial Density

Using the ESOS data from the 2020 campaign the aerial density for each day has been evaluated, by averaging the aerial density over 24h for every cluster, and the results are presented in Figure 4.10.

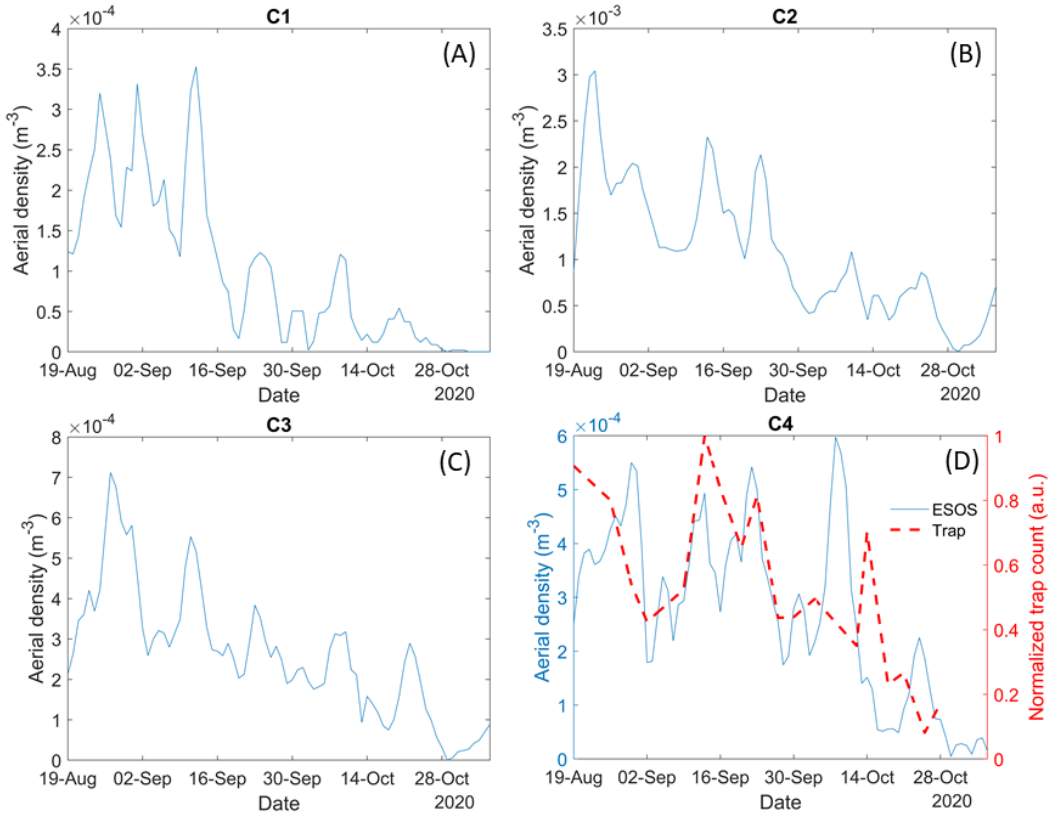

Figure 4.10 Evolution of the weekly rolling average of the daily aerial density for clusters C1, C2, C3 and C4, respectively (A), (B), (C) and (D) over the 80 days of the measurement campaign. (D) also displays the normalized trap count (red dashed line) of nearby traps for comparison purposes.

There is a clear overall decrease in insect aerial density over time for all clusters. This is likely due to the temperature decrease and day length reduction as winter approaches. In addition, cycles of increase and decrease of aerial density throughout the measurement campaign can be observed. The reasons for those cycles may be explained by the natural cyclical variation of population and environmental factors such as humidity and rainfall. These cycles may also correspond to the emergence and disappearance of different short lived and highly seasonal species. The second cluster is by far the most abundant, in terms of aerial density, with a density five to ten times greater than any other cluster. It is the cluster that most likely encompassed the greater number of distinct species, as almost all of the predominantly diurnal species are within this cluster, although its exact composition cannot be entirely defined by

the results of this methodology. The remaining three clusters all have comparable aerial density. However, cluster 4 only encompasses mosquito and, to a lesser extent midge species, with their typically high wingbeat frequency, while cluster 1 and 3 are likely to be much more diverse, in term of species.

As the EPS methodology is relatively new, it calls for a validation by comparison with a well-tested and widely approved methodology. The study of insects is often conducted with traps, which counts the number of captured specimens to study the population of insects. Every few days, traps (described in subsection 4.1.4) were emptied and the number of mosquitoes captured were counted by the Hudson Regional Health Commission. This gives access to the number of specimens retrieved from traps on different days. As presented Figure 4.10 (D), the number of mosquitoes captured at each trap location is normalized and the average is then displayed (red dashed line) with the aerial density of the mosquito cluster C4 retrieved by the ESOS system. While the captured mosquitoes have been identified at the species level, ESOS cannot discriminate at the species level, thus Figure 4.10 (D) presents the trap count for all species of mosquitoes combined, so that it can be compared to the ESOS data. This figure shows a good agreement between both methodologies, with a similar overall downward trend and cyclical variation between the two methods. This demonstrates that the aerial density retrieved by the ESOS system is coherent with the count of traps, which is currently one of the standard methods for monitoring mosquito populations.

4.3 2021 Field Campaign

This measurement campaign took place from June 4^{th} to December 7^{th} 2021. It involves one field ESOS and two first designs of the field eBoss, respectively named eBoss1 and eBoss2. Using three systems will allow for the comparison of their results,

which could help in exploring the differences in their insect detection capabilities and evaluate if they provide coherent estimate of aerial density.

While the ESOS system was in operation for the entire duration of the campaign, 187 consecutive days, the two eBoss were only introduced at later times. The first eBoss, eBoss1, was in operation from July 13^{th} to December 7^{th}, 148 consecutive days, and had a laser beam diameter of 7.2 cm. The second eBoss, eBoss2, was in operation from September 6^{th} to December 7^{th}, 93 consecutive days, and had a laser beam diameter of 5.5 cm. While a lower laser beam diameter implies a smaller probed volume and therefore a larger variability in aerial density estimation, see subsection 2.2.7, it also allows for a more efficient detection of smaller insects by increasing the signal to noise ratio. On the other hand, it also reduces the ability of the system to observe larger insects, as discussed in subsection 2.1.3.

During the entire campaign the down time of the instruments was below 6.5% due to temporary shutdown for alignment verification and routine maintenance. This value is about five times larger than for the 2020 campaign in part due to a few unvoluntary interruptions of the recoding that occurred due to the Ida storm of September 2^{nd} that flooded the area and from two power outages. Over the entire campaign, and all three instruments, a total of 508,602 insect events have been detected and kept for further analysis.

4.3.1 Wingbeat Frequency Distribution

From the half million events, the distribution of the wingbeat frequency can be studied. As shown in Figure 4.11 (D), the distributions of the ESOS, eBoss1 and eBoss2, respectively Figures 4.11 (A), (B) and (C), show clear differences. The wingbeat frequency distributions, recorded by each system, are function of the composition of the insects species present during the period of observation. Importantly, as indicated in Figure 4.11, the period of observation of the three systems

is not identical. As such, some of the differences can be explained by a change in insects' behavior along with the emergence and disappearance of seasonal species. Moreover, difference between ESOS and eBoss is furthermore affected by the altitude of observation.

For all three systems the wingbeat frequency distributions show distinct insect clusters. Two of them appear in the distribution of every system. One is centered around 280 Hz and the other 150 Hz. While the current methodology cannot formally confirm it, it is likely that those two clusters share strong similarities with the cluster C2 and C3 presented in subsection 4.2.2. A cluster similar to cluster C1 can also be observed on the ESOS distribution but is absent from the eBoss ones. This suggest that insect species present in cluster C1 either have a reduced activity from mid-July to December (when eBoss1 was active) or are predominantly present at altitude greater than 40 *cm* (maximal altitude of observation for both eBoss). The mosquito cluster, C4 in Figure 4.6, can be seen on both ESOS and eBoss1 distributions. In the eBoss2 distribution however, while not entirely absent, the cluster is less prominent, which suggest a diminution in mosquitoes' activity or abundance during September to December (when eBoss2 was active). This result is coherent with the known evolution of mosquito species present at the field location.

In order to further compare the results from the three systems, and to discuss more thoroughly any differences in detection efficiency between systems, it is necessary to limit the wingbeat frequency distributions to the time period where all systems were simultaneously active. Figures 4.12 (A), (B) and (C) respectively show the wingbeat frequency distributions for the ESOS, eBoss1 and eBoss2, for the period of time when they were all operating simultaneously.

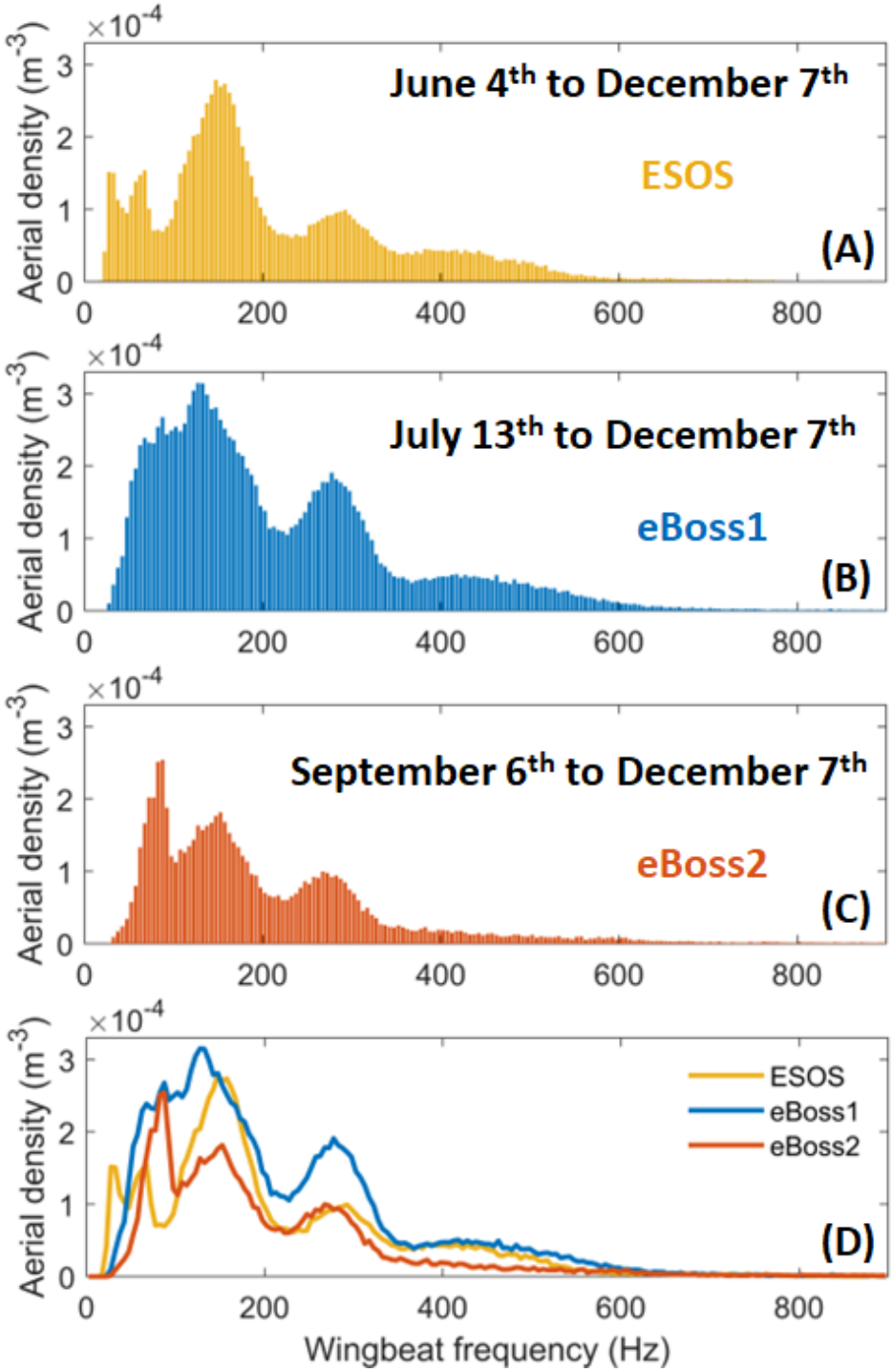

Figure 4.11 Wingbeat frequency distribution for the 2021 field campaign. (A), (B) and (C) display the wingbeat frequency distribution for the ESOS, eBoss1 and eBoss2 respectively. (D) shows the outline of the three distributions together. The considered time period is the entire period of operation of each system.

For all three systems, there is a noticeable peak of activity for insects with a wingbeat frequency around 150 Hz and 280 Hz, as for the previously discussed distributions. Those results confirm that, to some extent, all three systems do detect similar insect species or at the very least, insects with the same range of wingbeat frequency. Yet, some differences are also observed. Both eBoss systems shows a sharp peak of activity around 85 Hz which is not as clearly present on the ESOS distribution. While it does not necessarily mean that such insects are never observed by the ESOS it does means that there is a difference in the probability of observation for such insects between the field ESOS and both field eBoss. The most likely explanation for this difference lies within the range of altitude observed by the three systems. Indeed, the ESOS optical path is tilted upwards, and as such observes a range of altitudes from 25 cm to 2.5 m above ground while both eBoss have their optical path limited to 20 to 40 cm above ground. Hence, this observation may indicate that the considered insects, in the 85 Hz wingbeat frequency range, are near ground dweller that are more likely to interact with the low altitude eBoss than the ESOS. As previously discussed for the distributions presented in Figure 4.12, the insects of the C1 cluster are present in the ESOS distribution but not in the eBoss ones. This suggests that those low wingbeat frequency insects tend to fly at more than 40 cm above the ground. Moreover, by comparing the evolution of the aerial density of these insects between the two considered time periods, their population or at least their activity is decreasing as the season progress towards the colder months.

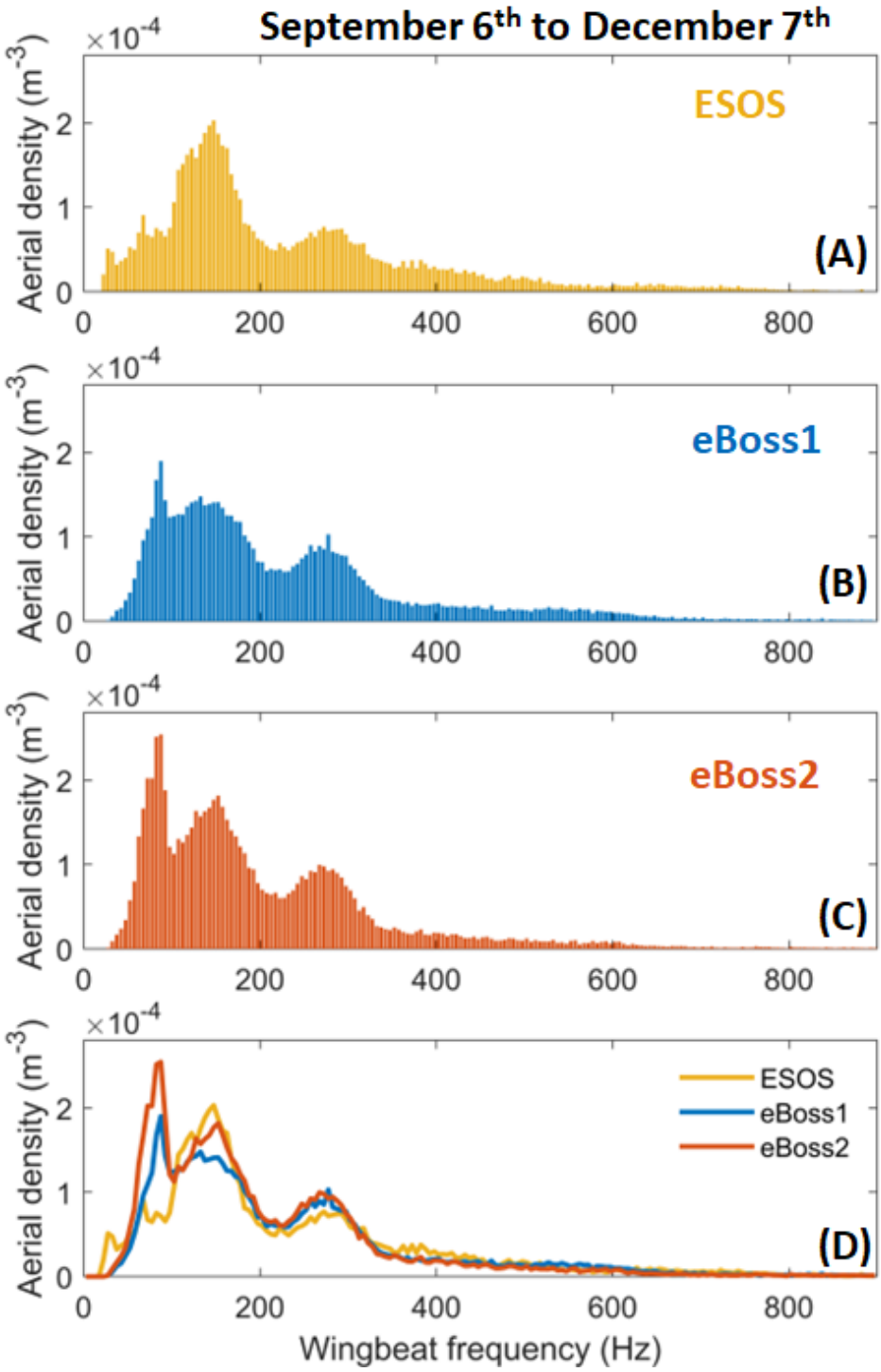

Figure 4.12 Wingbeat frequency distribution for the 2021 field campaign. (A), (B) and (C) display the wingbeat frequency distribution for the ESOS, eBoss1 and eBoss2 respectively. (D) shows the outline of the three distributions together. The considered time period is limited to events that occurred between September 6^{th} and December 7^{th} which is the only period of time where all three systems were active simultaneously.

4.3.2 Seasonal Variation of Aerial Density

As for the 2020 campaign, the aerial density for each of the three systems can be studied on a one-day time resolution. This provides information on the trend of the number of insects throughout the season. Figure 4.13 shows the aerial density of the three systems together.

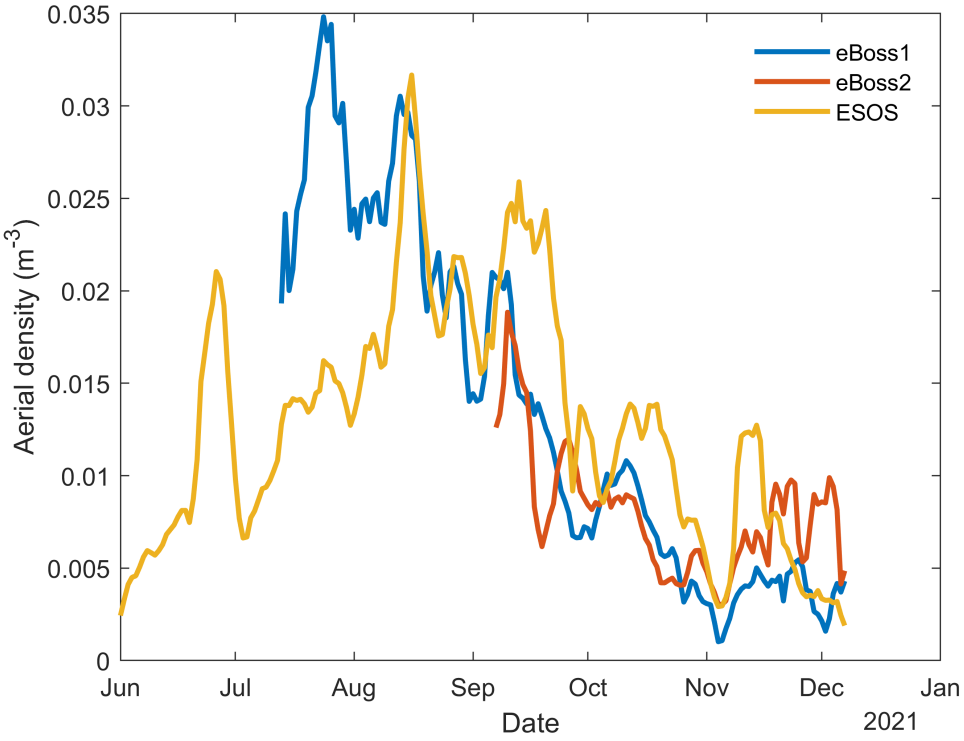

Figure 4.13 Evolution of the weekly rolling average of the daily aerial density for all three systems. Any missing days of recording are replaced by a null value and ignored by the rolling average.

From mid-August to December, all systems provide reasonably similar estimate of the insect aerial density. As could have been expected, a steady decrease in insect activity can be observed as winter approaches. However, from mid-July to mid-August, the ESOS and eBoss1 provide different estimate in terms of aerial density. Indeed, the eBoss1 measures a much higher insect activity than the ESOS, with eBoss aerial densities more than twice of the ESOS for some days. Moreover, around July 24^{th}, the eBoss1 observed a significant increase then decrease in insect aerial density

which is not seen by the ESOS. This suggests that both systems may be sensitive to different insects, at least during this period. Figure 4.14 displays the wingbeat frequency distribution for ESOS and eBoss1 for the period July 14^{th} to August 1^{st} which is the period during which the two systems show the greater difference in their total aerial density estimation.

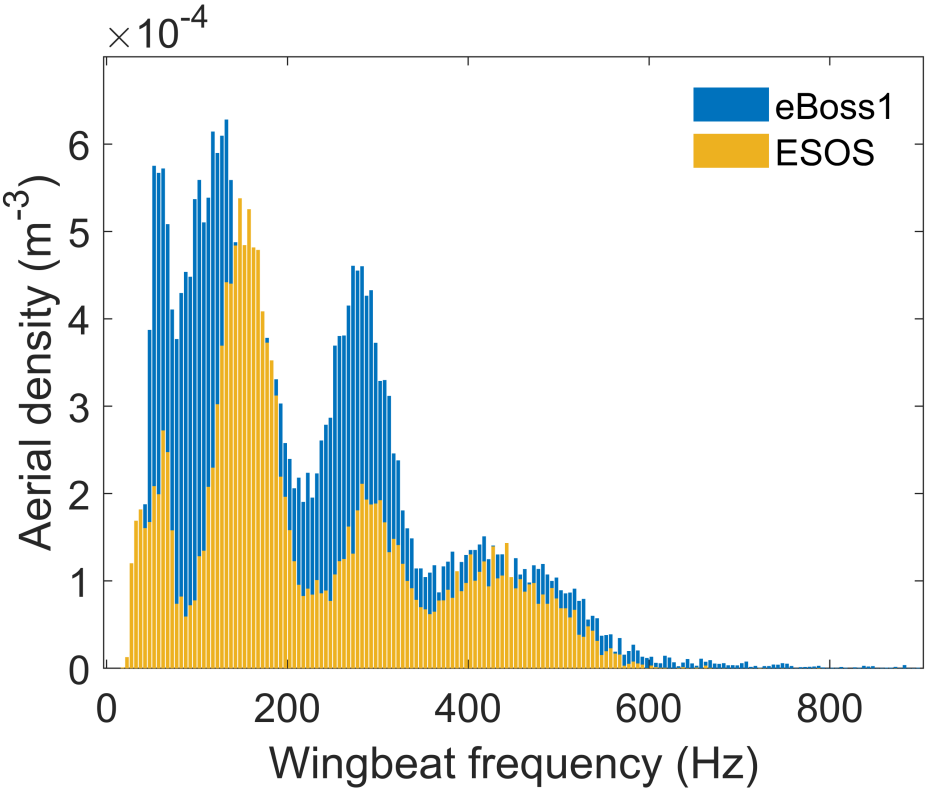

Figure 4.14 Wingbeat frequency distribution for ESOS and eBoss1. The considered time period is from July 14^{th} to August 1^{st} where the difference between the estimated aerial density between the two systems is the greatest.

The two wingbeat frequency distributions show clear differences. The eBoss records a greater aerial density for insects with a wingbeat frequency around 280 Hz than the ESOS. While both seem to observe those insects, it is likely that the eBoss is more sensitive to these types of insects than the ESOS. Similarly, in the 70 to 120 Hz range, the eBoss also seems to observe more insect activity than the ESOS. An additional consideration concerning the difference in the observed pool of insect species is the altitude of observation. Indeed, the laser beam of the ESOS is tilted

upward and as such is observing insects flying within the range of 25 cm to 2.5 m above ground level. In comparison, the eBoss only observes insects in the 20 to 40 cm altitude range. Hence, both systems could be observing different insects not only due to difference in sensitivity but also due to differences in altitude of observation. While both systems observing different subset of the overall populations of insects is currently one of the limitations, if understood and characterized, this difference could be used as a predictor variable. Insects that are efficiently detected by one system and not the other would effectively be easier to formally identify.

Comparing those distributions with those presented in Figure 4.15, which are the wingbeat frequency distributions for the period of August 15^{th} to September 6^{th} provides some additional information. During this period, both systems are in good agreement in terms of total aerial density, as shown in Figure 4.13.

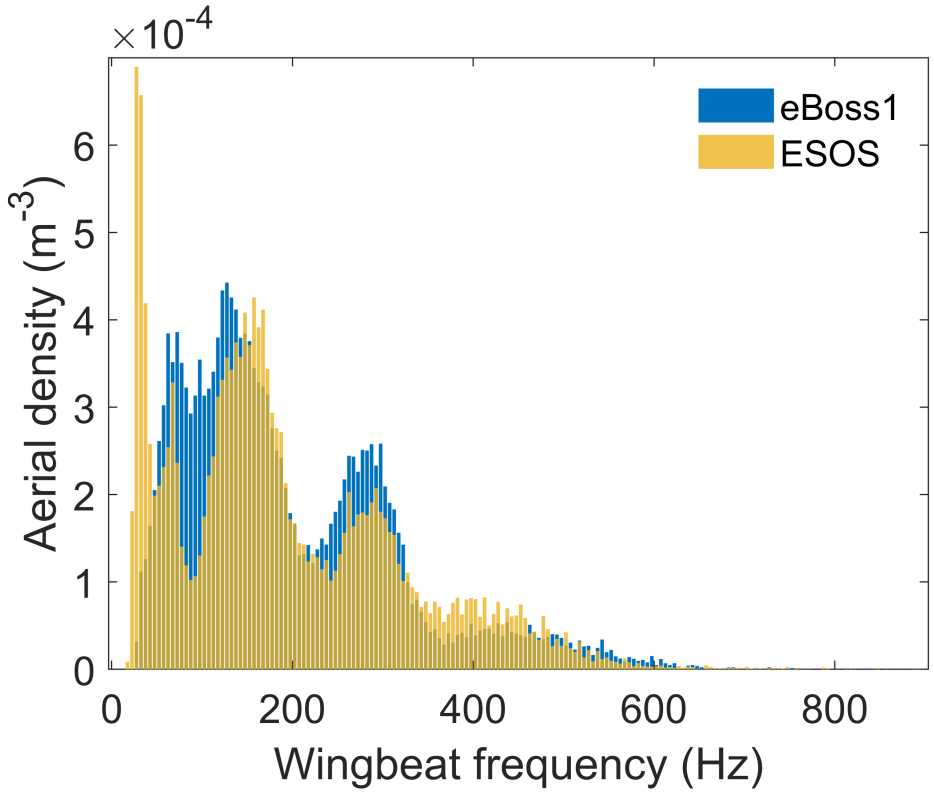

Figure 4.15 Wingbeat frequency distribution for ESOS and eBoss1. The considered time period is from August 15^{th} to September 6^{th}. Which is a period during which both systems are in good agreement in terms of total aerial density.

While some differences can still be observed, in particular for low wingbeat frequency insects that are better detected by the ESOS or only present at higher altitudes, both systems seem to observe a more similar pool of insect species than in the previously considered time period. This suggests that some of the species that were effectively underestimated by the ESOS are no longer active at the field location. This temporal variation of their population and/or activity could be an additional clue as to which species are impacted by the difference in detection efficiency between the two systems.

4.3.3 Species and Family level Clustering

Using the ESOS data of 2020, it was demonstrated that insects could be separated into different cluster as in Figure 4.6. With the exception of the mosquito cluster, the identified clusters are likely to include several distinct families of insects. While the study of these diverse clusters does provide additional information, a species or family level clustering would be even more beneficial. Indeed, if a single species, or closely related species, could be identified within the field data, the study of the daily and seasonal variation of their aerial density would provide important information on the species population and behavior.

While the field ESOS is limited in terms of species level clustering due to the dependency of its signal with distance, the field eBoss was designed to provide additional information that could be used for such prospect. As detailed in section 2.2, the wingbeat frequency, body extinction cross section, wing extinction cross section and aerial density can be retrieved from the eBoss data. Several species, known to be present at the field location, have been studied in laboratory conditions and the distribution of their wingbeat frequency and extinction cross sections determined, as shown in Figure 3.4 and 3.5.

From those distributions, three distinct SVM classifiers were trained to identify events originating from three group of insects: insects of the *Culicidae* family (mosquitoes), insects of the species *Musca domestica* (house flies) and insects from the species *Vespula maculifrons* (wasps).

To train each classifier, all the laboratory data were separated into two classes. The positive class corresponding to the group studied (i.e., mosquitoes, house fly or wasps) and the negative class containing every other events. Unlike in every other classifier training previously discussed, random undersampling was not performed, effectively positioning the positive class as a minority class, as can be expected in field conditions. This effectively creates a classifier that aims at identifying the few positive events from a large pool of negative events, i.e., a specific insect cluster. In this configuration, the SVM classifiers can be set to be more or less permissive in their detection of the positive class. They can be set to be extremely permissive, meaning that little true positive events are missed but at the cost of including several false positive events. Or they can be set to be extremely strict meaning that very little false positive events are detected but at the cost of missing several true positive events.

In practice, all three classifiers were set so that, when applied to the field data, the wingbeat frequency and extinction cross sections distributions of the positive class would be reasonably similar to the one found in laboratory condition. This was done by manually modifying the minimum score required for a positive detection. While this user defined threshold lacks an objective criterium and is based on human input, which limits the generalization of the process, it was found to provide promising results. Yet, more work will be needed in order to verify that no bias was introduced, to be able to more easily generalize the process and to verify the veracity of the prediction on the field data.

From the field data of the eBoss1, the wingbeat frequency and extinction cross sections of every event can be retrieved. Figures 4.16 (A) and (B) shows the aerial density per bin of wingbeat frequency and bin of wing or body extinction cross section, respectively.

As in the 2020 campaign, clear patches (i.e., clusters) of higher aerial density can be seen in the presented data, illustrated by black ellipses in Figure 4.16 (A). Similarly, to 2020, those four clusters could be defined, with a reasoning analog to the one that defined Cluster 1 to 4 in Figure 4.6. Each of these clusters is likely to encompass several distinct families of insects, except for the cluster with the highest wingbeat frequencies, which is presumably the mosquito cluster only including the *Culicidae* family. While an analysis of those clusters could be performed and would provide interesting information, this analysis would be redundant with the one performed on the 2020 campaign data. Instead, a more precise cluster analysis was performed using the aforementioned classifiers.

These classifiers allow for the extraction of the events that are predicted as belonging to the insect group that the classifiers have been trained for. From those events the seasonal variation of their aerial density can be retrieved, in order to study any possible differences between those subsets of the total insect population. Figure 4.17 shows the weekly rolling average of the daily aerial density for each of the three group of insects, along with the average ambient temperature.

For all three groups, the aerial density follows an overall downward trend. Yet, difference in aerial density between the insects' group can be observed. For example, in mid-August an increase in the *Musca domestica* aerial density can be observed while the aerial density of the *Culicidae* family is decreasing. This further supports that capability of the SVM classifiers to discriminate between insect groups. While it cannot be confirmed that the correct species or family are being studied, this

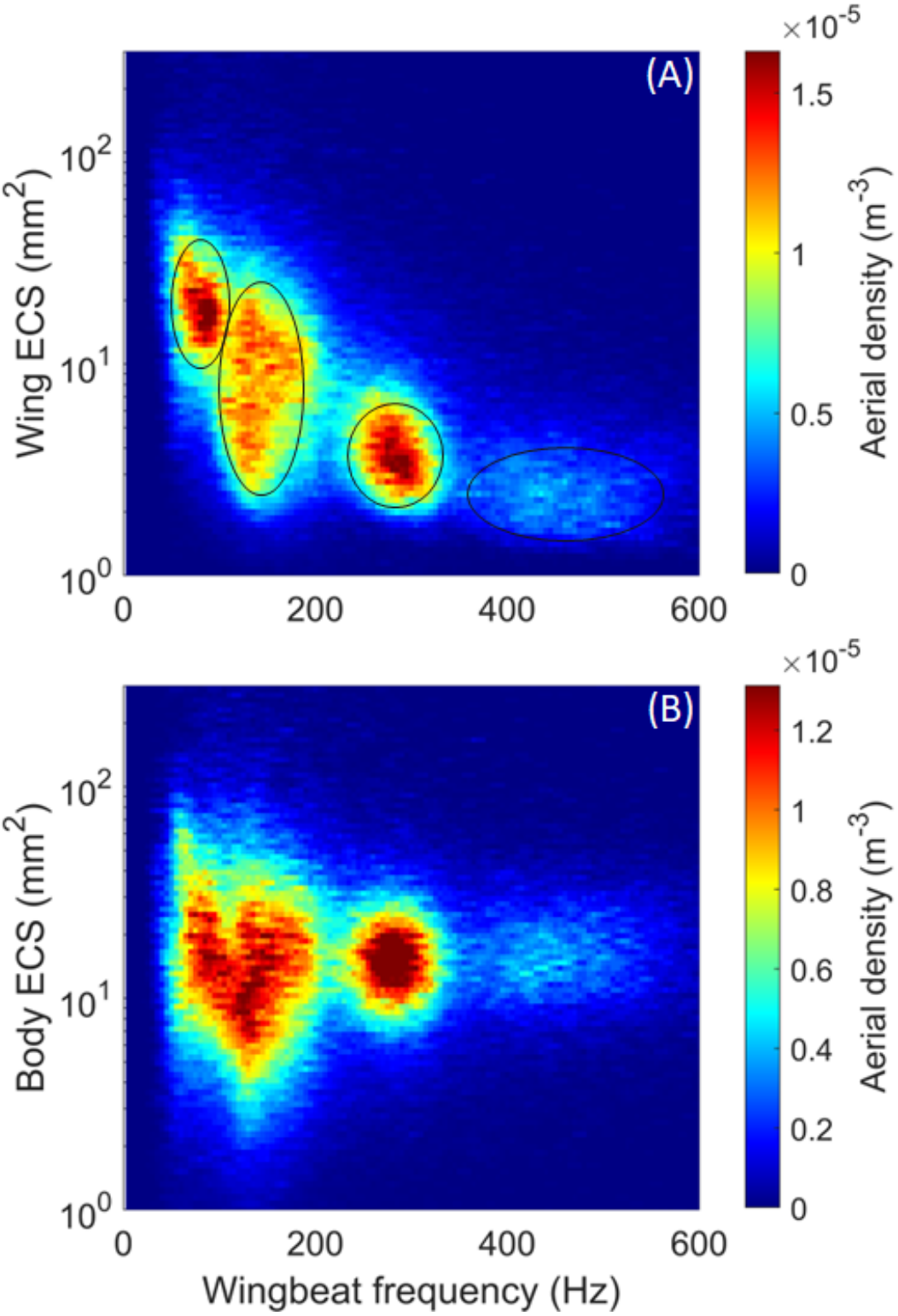

Figure 4.16 Color plots of aerial density per bin of wingbeat frequency and wing extinction cross section (A) or body extinction cross section (B). Results are obtained from the field data of the eBoss1. On (A), black ellipses illustrate the probable location of four distinct insects clusters.

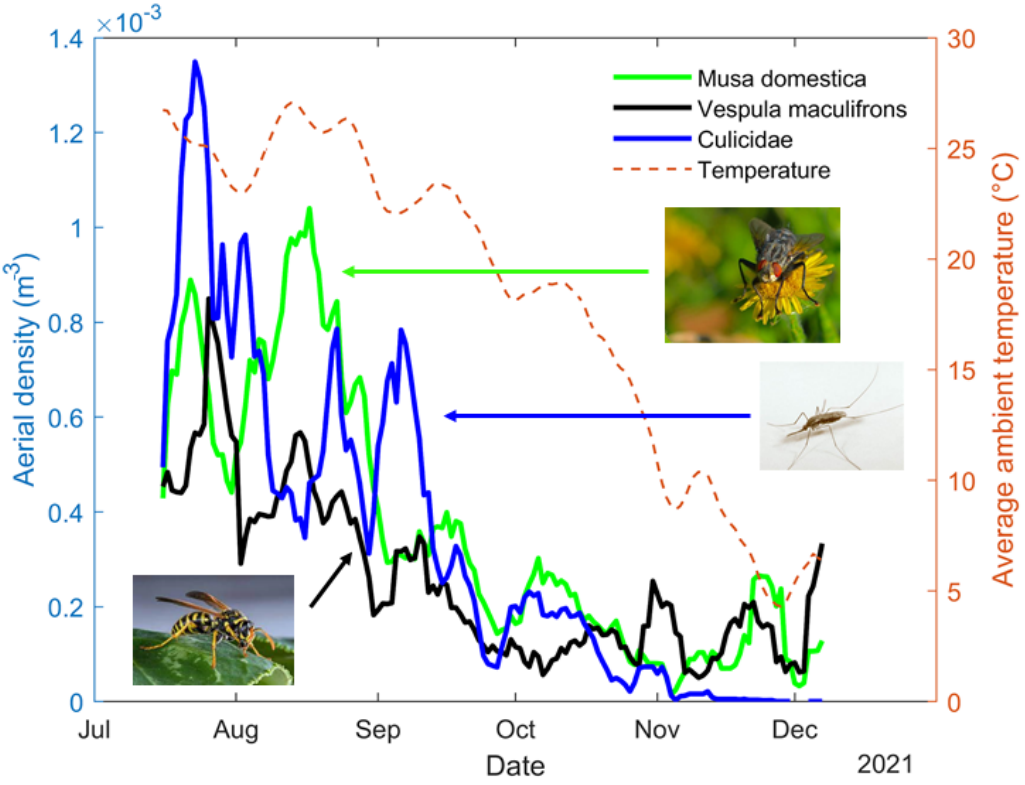

Figure 4.17 Weekly rolling average of the daily aerial density for the three considered group of insects, over the entire recording period of the eBoss1. The dashed orange line shows the evolution of the weekly rolling average of the ambient temperature.

observation supports that the different groups are at least composed of species with distinct population or behavior.

Additionally, the circadian rhythm of the three insect groups can be studied, as shown in Figure 4.18. Circadian rhythms, which are observed on a one hour rolling average of the one-minute resolution of the aerial density, are even more sensitive to the composition of the group than the seasonal aerial density. As such, it is interesting to study the circadian rhythms as well.

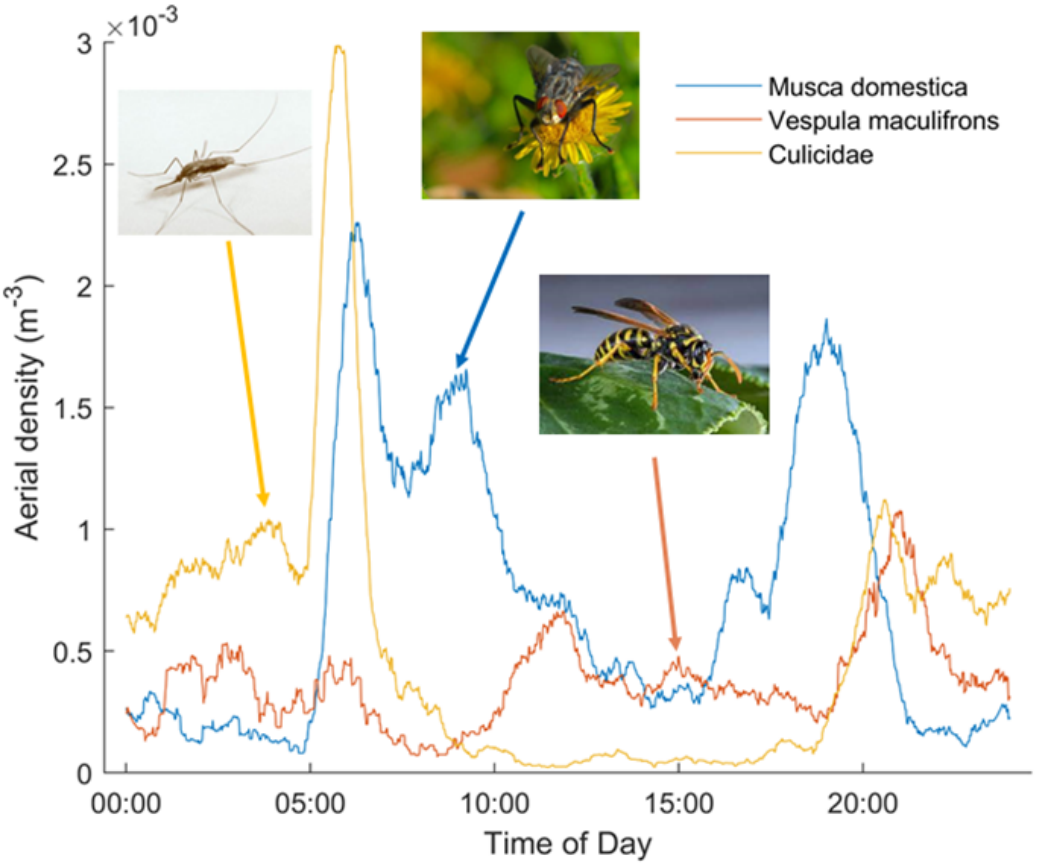

Figure 4.18 One hour rolling average of the one-minute resolution aerial density for the three insect groups. Those results illustrate a typical day during the month of August 2021.

As for the seasonal variation, the circadian rhythms of the three groups display clear differences. This is a particularly encouraging result since the circadian rhythm for each group correspond to their known behavior. The mosquitoes display more activity at night than during the day with peaks of activity around sunrise and sunset.

The house flies are more active during the day and the wasp shows an increase of activity around the warmer period of the day, as well as sunset.

Despite the encouraging results a strong disclaimer is warranted. While the three clusters are composed of insects with distinct behavior, they could be including other insect families than the one they are intended to represent, even possibly a completely different one. Additional experiments are needed in order to confirm or infirm if the SVM classifiers are effective when applied to field data. Consequently, the results presented in this section are to be considered as promising preliminary results only.

4.3.4 Seasonal Evolution of Flying Insect Biomass

As shown in Equation (2.30), the biomass density can be retrieved on a user defined time resolution. In Figure 4.19, the sliding weekly average over a one-day time resolution has been chosen to reduce uncertainties due to stochastic fluctuations. Only the results of dry biomass density will be presented, and not the results of wet biomass, see subsection 2.2.8 for more details. The dry biomass density was chosen over the wet biomass density for several reasons. The wet biomass includes the insect's water content which may not be relevant when evaluating the quantity of food availability for predators, one of the fundamental usages of insect biomass estimation. Moreover, the dry biomass is more easily compared with the gold standard results of insect traps, which often only provide dry insect mass from laboratory measurement. In addition, the dry mass estimation, from the described methodology, see subsection 3.3.3, is more accurate that the wet mass estimation.

For the study of biomass, only the results of eBoss1 are presented. As explained in subsection 2.2.8, only an eBoss can retrieve the mass of the specimen, which automatically disqualifies the ESOS. The first eBoss has been in operation for a longer period than the second eBoss and as such was chosen for this analysis. As

shown in Figure 4.19, the flying insect biomass density is following a downward trend as time passes. This result was expected as the population of several insect species are known to decrease with decreasing temperatures.

Figure 4.19 shows several fluctuations of the total biomass density which may be an indication of the emergence and disappearance of seasonal species or a shift in insect behavior. From mid-July to late August, the flying insect biomass density is at its highest values, oscillating around 0.4 $mg \cdot m^{-3}$. During this time, the maximum value of biomass density was around 0.6 $mg \cdot m^{-3}$ when considered over an entire day. Although, as can be observed in Figures 4.20 and 4.21, when considering a shorter time resolution, in the minute or hour range, the biomass density goes to higher values, above 1.1 $mg \cdot m^{-3}$, in particular around sunset and sunrise time. On the other hand, the minimum of flying insect activity takes place from the end of November to December when the biomass density varies around 0.1 $mg \cdot m^{-3}$ when considered over an entire day. Despite the average daily temperature dropping near $0°C$ during December, there was continuous flying insect activity.

In addition to the seasonal variation of the biomass density, the mass of each of the 262,870 insect events observed throughout the season by eBoss1 can be estimated. This provides information on the mass distribution of insects on the field during 2021. Between July and December 2021, the average dry mass of insects observed was 17.1 mg and the median value was 3.4 mg. This information can help in identifying the species present in the field by comparing those results to the recorded values of insects' dry mass made by entomologists.

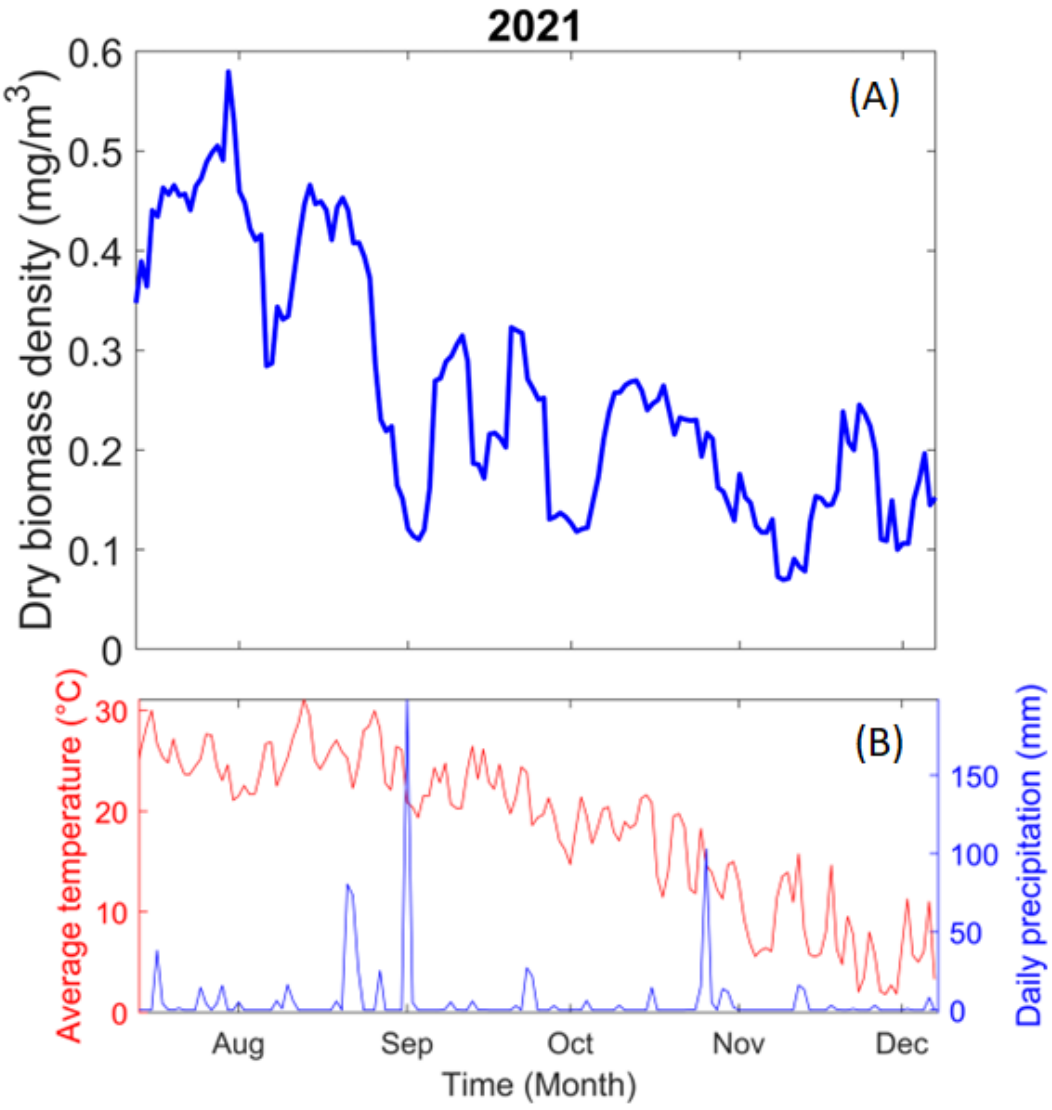

Figure 4.19 The blue solid line in (A) represents the weekly rolling average of the flying insect biomass density per meter cube of air. The biomass density, expressed in $mg \cdot m^{-3}$ is the dry biomass of insect retrieved from eBoss1. (B) shows the evolution of the daily average temperature (red) along with the daily precipitation (blue).

4.3.5 Daily Variation of Flying Insect Biomass

The eBoss allows for the study of biomass density on much shorter time resolution than the weekly rolling average, down to the minute range, see Figure 4.20. This allows for the refined study of insect activity and the observation of how the biomass density varies during a single day as well as seasonal variation.

The biomass density of flying insects varies throughout the day and is overall greater during sunrise and sunset where an increase in activity can be observed. The civil sunset and sunrise times are indicated by the semi-transparent lines in Figure 4.20. These results were expected as the behavior of several insect species is driven by light [187, 188, 189], furthermore it is coherent with previous observations from the 2020 campaign.

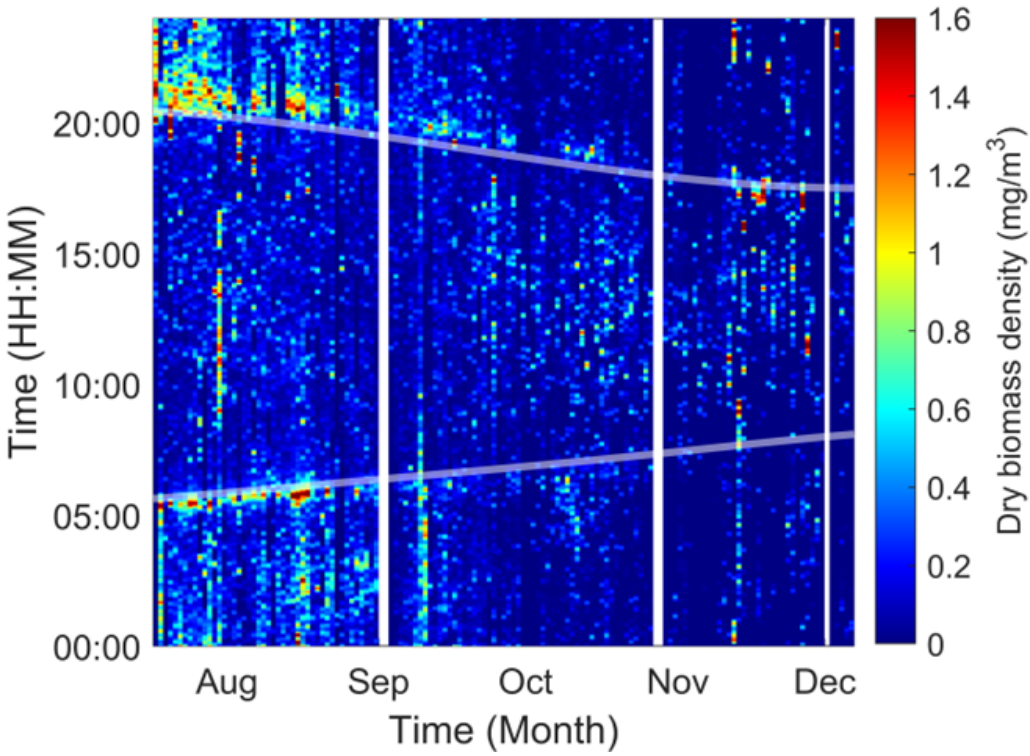

Figure 4.20 Dry biomass density estimation, with a one-minute resolution, in function of the time-of-day. Presented results are over the entire measurement campaign of the eBoss. Maximal values of density were artificially capped at 1.6 $mg \cdot m^{-3}$ in order to improve the color plot contrast. The vertical white lines indicate periods during which the system was offline. The semi-transparent lines indicate the civil sunrise and sunset time at the field location.

The daily biomass density variation over 24h is presented in Figure 4.21 for both the month of August (blue) and October (red). In August, there are two peaks of activity around 5:30am and 8:30pm, corresponding to sunrise and sunset time respectively. Furthermore, strong residual activity at night can be observed. On the contrary during the typical day of October, the dominant period of biomass density is no longer centered around sunset and sunrise but around noon with less activity at night. This can be the consequence of a change in insect populations or behavior between August and October. Insects being cold blooded, their mobility can be reduced with colder temperatures, favorizing activity during the day, as observed in October. Such information can be helpful for entomologists to infer the presence or absence of some insect species by comparing those results to known insect circadian rhythms. This can also be combined with the observed mass of the insects, and their wingbeat frequency, providing several parameters for species identification purposes.

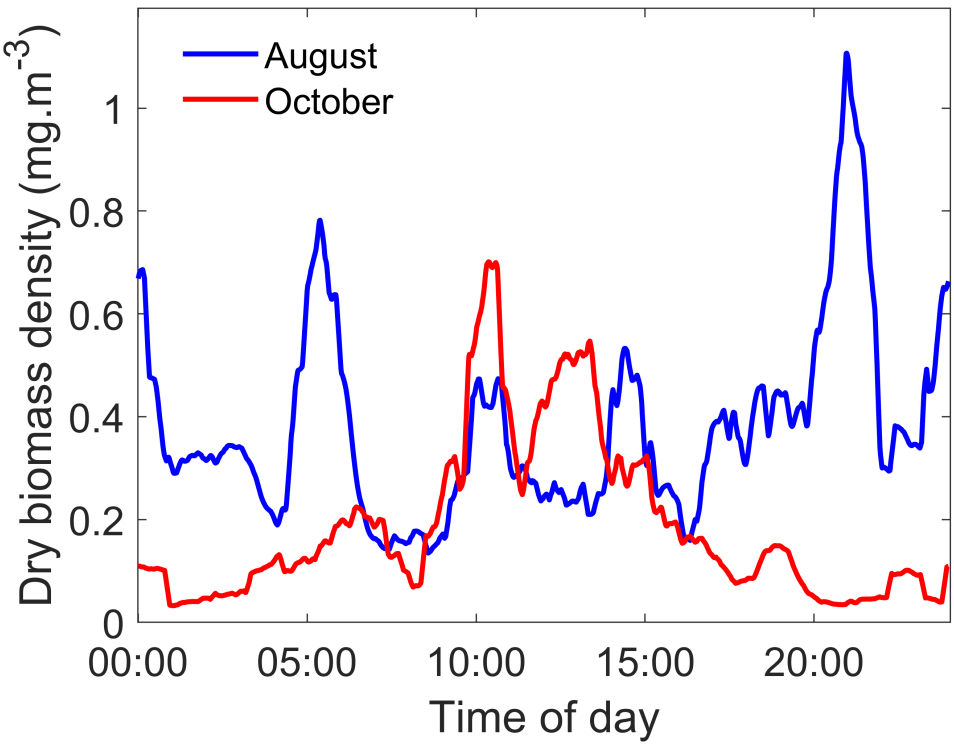

Figure 4.21 Dry biomass density over 24h. Blue and red line represent the average daily variation of the dry biomass density for the month of August and October respectively. The presented results are the 60 minutes rolling average of the biomass density, averaged over the entire month.

4.4 2022 Field Campaign

This measurement campaign started on March 23^{rd} 2022 and is currently still in operation. While the exact end of the campaign is not fixed and will depend on meteorological conditions, in particular the amount of snow precipitation and temperature, it is expected to end in early to mid-December 2022. During this campaign, three EPS have been deployed. One field ESOS, as in the previous two campaigns, and two field eBoss, one of the first design and one of the second design, respectively eBoss1 and eBoss2. Using both designs of the eBoss will allow for their comparison. The ESOS started its operation on March 23^{rd}, eBoss1 on April 6^{th} and eBoss2 on April 19^{th}.

As of the redaction of this document, only preliminary results have been gathered concerning the wingbeat frequency, extinction cross sections, aerial density and biomass of insects during the year 2022 in Secaucus, NJ, USA. The formal analysis of every event detected has not been achieved at the time of writing and is therefore not presented.

CHAPTER 5

SUMMARY AND CONCLUSION

The photonic monitoring of atmospheric fauna is a new and innovative field that aims at developing optical systems for the monitoring of flying insects. Entomological photonic sensors use the interaction between light and flying insects for the purpose of entomological study. This manuscript has addressed two main areas of interest: laboratory studies where insects' optical characteristics can be explored and field experiments where the behavior and abundance of wild insect species have been observed.

In the first chapter, the motivation for studying insects has been addressed along with the different monitoring methods employed to study their population. Some fundamental properties of light, necessary for the proper development of photonic sensors, have been presented. Insects can play a positive role in the Earth's ecosystem, both as pollinators and source of food for larger animals. These contributions are paramount and without them, famines and loss of diversity could be expected. On the other hand, insects also have negative effects. Some species are pests, reducing crop yields and some can transmit infectious diseases such as malaria. For all of these reasons, a fine scale understanding of insects' characteristics and behavior is paramount. Traditional methods for the monitoring of insects rely on the capture of insects through traps or net sweeping. While laboratory analysis of the captured specimen provides important information, those methods are labor and capital intensive. Hence, alternative approaches have been developed to complement those traditional methods. Based on acoustic or photonic sensors, those new approaches aim at the remote identification and count of insects. Among those new approaches, entomological photonic sensors have shown great potential. As this method relies on

the interaction between light and living matter, a brief description of light properties and possible interactions has been provided. This description covered several key concepts, polarization, refraction, absorption, scattering and absorption.

In the second chapter, the methodology and main concepts used for the development of entomological photonic sensors are presented. The optical layouts of the two main laboratory systems are detailed, along with the several steps involved in the processing of the optical signals. Finally, machine learning classification algorithms are presented. The optical layout of the two configurations, backscatter and extinction, respectively named Entomological Stand-off Optical Sensor (ESOS) and Entomological Bistatic Optical Sensor System (eBoss) have been detailed. Briefly, the backscatter configuration studies the light backscattered by the target while the extinction configuration studies the attenuation of the light due to the target. For both designs, the cross section of the laser beam will influence the range of observable insects. Thus, depending on the group of insects that are being studied, the ideal beam cross section will vary. In order to evaluate the best cross section, a numerical simulation has been developed and a laser beam with radius around 25 mm was shown to allow for a very versatile system capable of studying most flying insects, from the smallest to the largest. The recorded signal, from either configuration, contains information related to the characteristics of the target. The raw data recorded by the optical sensor are pre-processed through averaging and bandpass filters. Then, a threshold detection algorithm discriminates insect events from non-insect events. Each event, and its associated frequency spectrum, is then furthermore studied to extract characteristics of the specimen. The first and foremost insect characteristic that can be retrieved is the wingbeat frequency of the specimen. Briefly, the other characteristics that can be retrieved from the presented methodology are the optical cross sections of the insect's wings and body, the wing and body spectral and depolarization ratios as well as the mass of the specimen. Using the

labelled database of optical characteristics constructed from laboratory experiments, machine learning classifiers can be trained to identify the species, sex or gravidity of unknown events. Several algorithms have been considered and a brief description of the following algorithms is provided: Naïve Bayes, K-Nearest Neighbor, Decision Tree, Linear Discriminant Analysis and Support Vector Machine. Finally, when considering a large number of transit signals, additional information can be retrieved as well. In particular, the aerial density of flying insects, which is the number of flying insects per meter cube of air. Through a numerical simulation, it was shown that the aerial density of flying insects could be effectively retrieved from a statistical analysis of the results extracted from both ESOS or eBoss data.

In the third chapter, the main results from the laboratory experiments are presented. After introducing the methods used to obtain and maintain live insects in laboratory conditions, the performance of different machine learning classifiers is studied. The optical characteristics of several insect species retrieved from laboratory eBoss are presented and the potential of the Mel frequency cepstral coefficients for species level identification is discussed. Using a simple Naïve Bayes machine learning algorithm, the discrimination between males and females of three mosquito species mosquitoes was conducted. Using only the wingbeat frequency, the accuracy of the sex prediction reaches 95.8%. However, for the identification of both species and sex, resulting in six classes, the accuracy using only the wingbeat frequency drops down to 62.3%. This has motivated a search for additional predictor variables. Using the best combination of predictor variables that can be extracted from the laboratory ESOS, the accuracy of sex and species identification of mosquitoes increases to 74.95%, which 19.37% above the accuracy using only the wingbeat frequency. This demonstrates the benefits of using dual-wavelength and polarization-sensitive measurements for the species and sex identification of mosquitoes and likely, for insects in general. Yet, naïve Bayes algorithm is not the only possible algorithm that can be employed. As

such, other machine learning algorithms were considered. For the identification of species and sex of mosquitoes, the support vector machine algorithm showed the best performance with an accuracy of 79.7%. For the discrimination between male and female of a single mosquito species, the best performing algorithm is the decision tree with a near perfect accuracy of 99.9%. In addition to sex and species, the gravidity of mosquitoes has been studied. For this classification task, the linear discriminant algorithm shows the best performance, with 86.6% accuracy. Meaning that using dual wavelength and polarization sensitive measurements of the ESOS allow for the identification of the presence or absence of eggs within the abdomen of a moving female mosquitoes from several meters away. Several species of insects have also been studied with the eBoss methodology. The wingbeat frequency and extinction cross section distributions of 100,000 events from 12 species and sex groups of insects are shown. Using measurement of the extinction cross section of the insect body, both the dry and wet mass of a flying specimen can be evaluated from the eBoss data. This effectively allows for the non-destructive measurement of the mass of a flying insect from a distance. Finally, the Mel frequency cepstral coefficients, which are prominent features in the field of speech and voice recognition, were used in combination with a support vector machine algorithm. It was shown that those coefficients could discriminate males of two closely related species of mosquitoes with an accuracy of 80.6%, despite those two species having the very similar wingbeat frequency distributions, showing that support vector machine algorithm, coupled with the Mel frequency cepstral coefficient, can distinguish between the signals of male mosquitoes from two closely related species. Furthermore, it was demonstrated that the interest of Mel frequency cepstral coefficient for identification could be extended to other insect families as well.

In the fourth chapter, the main results from the field experiments are presented. First, the location of the field experiment, optical layout of the field systems and the

characteristics of the co-located weather station are described. Then the results of the three field campaigns are shown and discussed in details. The field location, which remained the same for all three campaigns, is situated in the city of Secaucus, New Jersey, USA. A co-located portable weather station was operating next to the field, allowing for the recording of environmental parameters such as outdoor temperature and solar radiation to name a few. The first campaign took place from August to November 2020 and used a single ESOS system. From the analysis of the field data, four insects clusters, with sufficiently distinct characteristics, were identified. While the exact composition of each cluster is yet unknown, the retrieved optical characteristics allowed for an educated guess of their composition. One cluster in particular, cluster C4, was identify with some degree of confidence as a cluster mainly composed of mosquito species. For each of the four clusters, the circadian rhythm (i.e., aerial density on a one-minute time resolution) was evaluated. This analysis showed that one of the clusters includes mostly diurnal species and more importantly that the mosquito cluster displayed strong peaks of activity around sunset and sunrise, which is a known behavior of the mosquito species present on the field. The peaks of activity of mosquitoes around sunset and sunrise were furthermore studied, and the observations show that the mosquito activity does indeed follow the change in sunset and sunrise time as season progress. Finally, the seasonal variation of the aerial density (i.e., aerial density with a one-day time resolution) of each cluster was evaluated. Each cluster showed a clear overall decrease in insect aerial density over time, which is likely due to the temperature decrease and day length reduction as winter approaches. Moreover, the variation of aerial density of the mosquito cluster was compared with the trap counts of mosquito traps, which is the gold standard when studying mosquito abundance. This comparison showed good agreement between both methodologies, with a similar overall downward trend and cyclical variation, which further confirmed the validity of the entomological photonic sensor methodology. The second campaign

took place from June to December 2021 and used one ESOS system along with two eBoss. The first result of the second campaign stems from the comparison of the wingbeat frequency distributions retrieved from the ESOS with the one retrieved from the eBoss. The differences between the two suggests that some insect species are predominantly active near the ground, especially those with a wingbeat frequency around 85 Hz. The seasonal variation of aerial density of the three systems are in good agreement, between mid-August to December, with similar downward trends. Yet, from mid-July to mid-August the ESOS and eBoss provide different estimate in terms of aerial density. This suggests that either the systems are sensitive to different insects during this period or that some low altitude insects were very active during this period, making them less likely to interact with the ESOS. Using laboratory eBoss data, three species level support vector machine classifiers were trained then used on the 2021 field data. The results obtained from this application are still in their early development stage and should therefore be considered as preliminary results only. Yet, some encouraging results were obtained. Three insects group were studied, house flies (*Musca domestica*), wasps (*Vespula maculifrons*) and mosquitoes (several species together). Both the seasonal variation and circadian rhythm of those specific groups were extracted and studied. The seasonal variation showed a similar downward trend for all three groups, but, in mid-August an increase in the *Musca domestica* aerial density can be observed while the aerial density of the *Culicidae* family is decreasing. In addition, the circadian rhythms of the three groups showed clear differences suggesting that the three groups are composed of different species or family of insects. Moreover, the circadian rhythm for each group corresponds to the known behavior of flies, wasp and mosquitoes. Using the results of the eBoss, the evolution of insect biomass was studied. As expected, the trend is similar to the evolution of the aerial density. Nonetheless, this analysis provides an important metric for the field of entomology: the available dry biomass of flying insect per meter

cube. This metric varied from 1.1 $mg \cdot m^{-3}$ around sunrise and sunset time in August down to 0.1 $mg \cdot m^{-3}$ in December. The variation of the insect biomass during a typical day in August compared with the one of a typical day of October showed a clear change in the insect species present on the field, or at least a change in their behavior. With predominantly nocturnal species with high activity peaks around sunrise and sunset in August to mostly diurnal activity centered around noon in during October. Finally, the 2022 campaign was introduced. This campaign started in March 2022 and is currently running. As such, no results could yet be presented.

In conclusion, this manuscript presents the work done on the development of entomological photonic sensors used for the study of flying insects. Two new designs of such sensors have been developed (ESOS and eBoss) and their capabilities characterized both in laboratory and real-word experiments. With the use of the two systems described in this manuscript, several morphological and optical properties of flying insects can be retrieved remotely in a non-destructive and non-intrusive way. The insect mass, wingbeat frequency, wing and body cross section (backscatter and extinction), as well as spectral and polarimetric ratios of both the wing and the body of the specimen which are used as predictor variables to identify insects crossing the laser beam. Furthermore, a methodology for the estimation of the aerial density ($insects \cdot m^{-3}$) of free flying insects has been designed, proven by simulation and tested in field conditions. This has allowed for the monitoring and study of wild insects' abundance and circadian rhythms over three years with multiple instruments. Finally, several machine learning algorithms trained for the automated identification of insects' species, sex and gravidity have been evaluated and discussed. This research demonstrates that an approach based on photonic instruments can provide rich entomological data, potentially unlocking the many limitations caused by a lack of data in the field of entomology. While the scientific community in this new emerging field remains very small, just a handful of groups in the world, the

trend seems to be rapidly increasing as more and more studies are published every year by new groups. The possibilities to improve entomological photonic sensors are still huge, many research avenues remains to be explored, either on the design of instrument themselves or on the processing of the data they provide. With the increasing concerns regarding the decline of insect populations, the prospect of a rapid evolution of the capabilities of entomological sensors may offer hope that solutions can be found.

REFERENCES

[1] N. E. Stork. How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology*, pages 31–45, 2017.

[2] A. D. Chapman. *Numbers of living species in Australia and the world*. Australian Biodiversity Information Services, Toowoomba, Australia, 2nd edition, 2009.

[3] Y. M. Bar-On, R. Phillips, and R. Milo. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), 2018.

[4] T. D. Schowalter, J. A. Noriega, and T. Tscharntke. Insect effects on ecosystem services—Introduction. *Basic and Applied Ecology*, 26:1–7, 2018.

[5] W. W. Weisser and E. Siemann. *The various effects of insects on ecosystem functioning*. Insects and Ecosystem Function. Ecological Studies (Analysis and Synthesis). Springer-Verlag, Berlin, Heidelberg, Germany, 2008.

[6] K. F. Conrad, M. S. Warren, Ri. Fox, M. S. Parsons, and I. P. Woiwod. Rapid declines of common, widespread British moths provide evidence of an insect biodiversity crisis. *Biological Conservation*, 132(3):279–291, 2006.

[7] C. A. Hallmann, M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan, W. Stenmans, A. Müller, H. Sumser, T. Hörren, D. Goulson, and H. De Kroon. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE*, 12(10), 2017.

[8] T. Jeppsson, A. Lindhe, U. Gärdenfors, and P. Forslund. The use of historical collections to estimate population trends: A case study using Swedish longhorn beetles (Coleoptera: Cerambycidae). *Biological Conservation*, 143(9):1940–1950, 2010.

[9] C. L. Boggs and D. W. Inouye. A single climate driver has direct and indirect effects on insect population dynamics. *Ecology Letters*, 15(5):502–508, 2012.

[10] R. Klink, D. E. Bowler, K. B. Gongalsky, A. B. Swengel, A. Gentile, and J. M. Chase. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science*, 368:417–420, 2020.

[11] M. S. Crossley, A. R. Meier, E. M. Baldwin, L. L. Berry, L. C. Crenshaw, G. L. Hartman, D. Lagos-Kutz, D. H. Nichols, K. Patel, S. Varriano, W. E. Snyder, and M. D. Moran. No net insect abundance and diversity declines across US Long Term Ecological Research sites. *Nature Ecology and Evolution*, 4(10):1368–1376, 2020.

[12] M. E. Cammell and J. D. Knight. Effects of climatic change on the population dynamics of crop pests. *Advances in Ecological Research*, 22(C):117–162, 1992.

[13] C. Robinet and A. Roques. Direct impacts of recent climate warming on insect populations. *Integrative Zoology*, 5(2):132–142, 2010.

[14] N. E. Rafferty. Effects of global change on insect pollinators: multiple drivers lead to novel communities. *Current Opinion in Insect Science*, 23:22–27, 2017.

[15] R. J. Wilson and R. Fox. Insect responses to global change offer signposts for biodiversity and conservation. *Ecological Entomology*, 2020.

[16] M. Brydegaard and S. Svanberg. Photonic monitoring of atmospheric and aquatic fauna. *Laser and Photonics Reviews*, 12:1800135, 2018.

[17] World Health Organization. World malaria report 2021. Technical report, Geneva, Switzerland, 2021.

[18] R. Fattorini and B. J. Glover. Molecular mechanisms of pollination biology. *Annual Review of Plant Biology*, 2020.

[19] T. M. Culley, S. G. Weller, and A. K. Sakai. The evolution of wind pollination in angiosperms. *Trends in Ecology and Evolution*, 17(8), 2002.

[20] X. Fan, S. C. H. Barrett, H. Lin, L. Chen, X. Zhou, and J. Gao. Rain pollination provides reproductive assurance in a deceptive orchid. *Annals of Botany*, 110:953–958, 2012.

[21] J. D. Ackerman. Abiotic pollen and pollination: ecological , functional , and evolutionary perspectives. *Plant Systematics and Evolution*, 222:167–185, 2000.

[22] C. W. Wardhaugh. How many species of arthropods visit flowers? *Arthropod-Plant Interactions*, 9:547–565, 2015.

[23] K. J. Hung, J. M. Kingston, M. Albrecht, D. A. Holway, and J. R. Kohn. The worldwide importance of honey bees as pollinators in natural habitats. *Proceedings of the Royal Society B: Biological Sciences*, 285, 2018.

[24] N. Gallai, J. M. Salles, J. Settele, and B. E. Vaissière. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics*, 68(3):810–821, 2009.

[25] G. Allen-wardell, P. Bernhardt, R. Bitner, A. Burquez, S. Buchmann, J. Cane, P. A. Cox, V. Dalton, P. Feinsinger, M. Ingram, D. Inouye, C. E. Jones, K. Kennedy, P. Kevan, R. Medellin, S. Medellin-morales, G. P. Nabhan, B. Pavlik, V. Tepedino, P. Torchio, and S. Walker. The Potential consequences of pollinator declines on the conservation of biodiversity and stability of food crop yields. *Conservation Biology*, 12(1), 1998.

[26] D. Goulson, E. Nicholls, C. Botias, and E. L. Rotheray. Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, 347(6229), 2015.

[27] J. C. Biesmeijer, S. P. M. Roberts, M. Reemer, R. Ohlemuller, M. Edwards, T. Peeters, A. P. Schaffers, S. G. Potts, R. Kleukers, C. D. Thomas, J. Settele, and W. E. Kunin. Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science*, 313, 2014.

[28] D. VanEngelsdorp, J. D. Evans, C. Saegerman, C. Mullin, E. Haubruge, B. Kim, M. Frazier, J. Frazier, D. Cox-foster, Y. Chen, R. Underwood, D. R. Tarpy, and J. S. Pettis. Colony collapse disorder: A descriptive study. *PLoS ONE*, 4(8), 2009.

[29] S. G. Potts, J. C. Biesmeijer, C. Kremen, P. Neumann, O. Schweiger, and W. E. Kunin. Global pollinator declines : trends , impacts and drivers. *Trends in Ecology and Evolution*, 25(6):345–353, 2010.

[30] L. Eisen and R. J. Eisen. Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases. *Emerging Infectious Diseases*, 13(12), 2007.

[31] C. J. Rhodes. Pollinator decline – an ecological calamity in the making? *Science Progress*, 101(2):121–160, 2018.

[32] M. Nyffeler, C. H. Sekercioglu, and C. J. Whelan. Insectivorous birds consume an estimated 400 – 500 million tons of prey annually. *The Science of Nature*, 105(47), 2018.

[33] E. Gould. The Feeding Efficiency of Insectivorous Bats. *Journal of Mammalogy*, 36(3):399–407, 1955.

[34] S. Livingston, S. R. Lavin, K. Sullivan, L. Attard, and E. V. Valdes. Challenges with effective nutrient supplementation for amphibians: A review of cricket studies. *Zoo Biology*, 33:565–576, 2014.

[35] D. Sun-waterhouse, G. I. N. Waterhouse, L. You, J. Zhang, Y. Liu, L. Ma, J. Gao, and Y. Dong. Transforming insect biomass into consumer wellness foods: A review. *Food Research International*, 89:129–151, 2016.

[36] A. Firebaugh and K. J. Haynes. Light pollution may create demographic traps for nocturnal insects. *Basic and Applied Ecology*, 34:118–125, 2019.

[37] F. Sánchez-Bayo and K. A. G. Wyckhuys. Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, 232(September 2018):8–27, 2019.

[38] S. Sharma, R. Kooner, and R. Arora. *Insect pests and crop losses*. Springer Nature, Singapore, 2017.

[39] S. Garcia-Lara and S. O. Serna-Saldivar. *Encyclopedia of food and health: Insect pests*. Academic Press, Oxford, England, 2016.

[40] I. H. Williams. *The major insect pests of oilseed rape in Europe and their management: An overview*. Springer, Dordrecht, Netherlands, 2010.

[41] P. G. Weintraub. *Physical control: An important tool in pest management programs*. Biorational Control of Arthropod Pests. Springer, Dordrecht, Netherlands, 2009.

[42] E. C. Oerke. Crop losses to pests. *Journal of Agricultural Science*, 144(1):31–43, 2006.

[43] S. K. Dara, L. Barringer, and S. P. Arthurs. Lycorma delicatula (Hemiptera : Fulgoridae): A new invasive pest in the United States. *Journal of Integrated Pest Management*, 6, 2015.

[44] D. Lee, Y. Park, and T. C. Leskey. A review of biology and management of Lycorma delicatula (Hemiptera: Fulgoridae), an emerging global invasive species. *Journal of Asia-Pacific Entomology*, 22:589–596, 2019.

[45] M. Sarwar. Insect vectors involving in mechanical transmission of human pathogens for serious diseases. *International Journal of Bioinformatics and Biomedical Engineering*, 1(3):300–306, 2015.

[46] L. H. V. Franklinos, K. E. Jones, D. W. Redding, and I. Abubakar. The effect of global change on mosquito-borne disease. *The Lancet Infectious Diseases*, 19:30161–30166, 2019.

[47] A. M. Kilpatrick and S. E. Randolph. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *The Lancet*, 380:1946–1955, 2012.

[48] E. S. Paixão, M. G. Teixeira, and L. C. Rodrigues. Zika, chikungunya and dengue: The causes and threats of new and reemerging arboviral diseases. *British Medical Journal Global Health*, 3, 2018.

[49] World Health Organization. Global malaria programme. World malaria report 2017. Technical report, World Health Organization, Luxembourg, 2017.

[50] USAID, CDC, and DHHS. The President's malaria initiative: 12th annual report to Congress. Technical Report United States Agency for International Development, Centers for Disease Control and Prevention (U.S.), Department of Health and Human Services (U.S.), Stephen B. Thacker CDC Library collection, 2018.

[51] World Health Organization. Global strategy for dengue prevention and control 2012-2020, Geneva, Switzerland. 2012.

[52] DHHS and ASPR. National Health Security Strategy and Implementation Plan 2015-2018, U.S. Department of Health and Human Services. Technical report, Department of Helath & Human Sercvices, Assistant Secretary for Preparedness and Response, United States, 2014.

[53] E. D. Doxon, C. A. Davis, and S. D. Fuhlendorf. Comparison of two methods for sampling invertebrates: Vacuum and sweep-net sampling. *Journal of Field Ornithology*, 82:60–67, 2011.

[54] T. J. Gibb and C. Oseto. *Insect collection and identification: Techniques for the field and laboratory*. Academic Press, Elsevier, London, United Kingdom, 2019.

[55] C. J. Mcmeniman, R. A. Corfas, B. J. Matthews, S. A. Ritchie, and L. B. Vosshall. Multimodal integration of carbon dioxide and other sensory cues drives mosquito attraction to humans. *Howard Hughes Medical Institute*, 156(5):1060–1071, 2014.

[56] H. Kawada, S. Honda, and M. Takagi. Comparative laboratory study on the reaction of Aedes aegypti and Aedes albopictus to different attractive cues in a mosquito trap. *Journal of Medical Entomology*, 44(3):427–432, 2007.

[57] R. C. Muirhead-Thomson. *Trap responses of flying insects*. Academic Press Inc., San Diego, CA, 1991.

[58] M. A. Mullen and A. K. Dowdy. A pheromone-baited trap for monitoring the Indian meal moth , Plodia interpunctella (Hubner) (Lepidoptera : Pyralidae). *Journal of Stored Products Research*, 37:231–235, 2001.

[59] D. L. Kline. Traps and trapping techniques for adult mosquito control. *Journal of the American Mosquito Control Association*, 22(3):490–496, 2006.

[60] R Malaise. A new insect trap. *Entomologisk Tidskrift*, 58:148–160, 1937.

[61] M. J. Skvarla, J. L. Larson, J. R. Fisher, and A. P. G. Dowling. A review of terrestrial and canopy malaise traps. *Annals of the Entomological Society of America*, 2020.

[62] W. H. Meeraus, J. S. Armistead, and J. R. Arias. Field comparison of novel and gold standard traps for collecting Aedes albopictus in northern Virginia. *Journal of the American Mosquito Control Association*, 24(2):244–248, 2008.

[63] A. Farajollahi, B. Kesavaraju, D. C. Price, G. M. Williams, S. P. Healy, R. Gaugler, and M. P. Nelder. Field efficacy of BG-Sentinel and industry-standard traps for Aedes albopictus (Diptera: Culicidae) and West Nile virus surveillance. *Journal of Medical Entomology*, 46(4):919–925, 2009.

[64] T. N. Crepeau, I. Unlu, S. P. Healy, A. Farajollahi, and M. Fonseca. Experiences with the large-scale operation of the Biogents Sentinel ™ trap. *Journal of the American Mosquito Control Association*, 29(2):177–180, 2013.

[65] N. C. Wipf, V. Guidi, M. Tonolla, M. Ruinelli, P. Müller, and O. Engler. Evaluation of honey-baited FTA cards in combination with different mosquito traps in an area of low arbovirus prevalence. *Parasites and Vectors*, 12, 2019.

[66] U. Jinbo, T. Kato, and M. Ito. Current progress in DNA barcoding and future implications for entomology. *Entomological Science*, 14:107–124, 2011.

[67] W. J. Mcgeachie. The effects of moonlight illuminance, temperature and wind speed on light-trap catches of moths. *Bulletin of Entomological Research*, 79:185–192, 1989.

[68] T. W. Sappington and D. W. Spurgeon. Variation in Boll Weevil (Coleoptera: Curculionidae) captures in pheromone traps arising from wind speed moderation by brush lines. *Environmental Entomology*, 29(4):807–814, 2000.

[69] R. Wedding, O. Anderbrant, and P. Jönsson. Influence of wind conditions and intertrap spacing on pheromone trap catches of male European pine sawfly, Neodiprion sertifer. *Entomologia Experimentalis et Applicata*, 77:223–232, 1995.

[70] B. V. Giordano, S. K. Bartlett, D. A. Falcon, R. P. Lucas, M. J. Tressler, and L. P. Campbell. Mosquito community composition, seasonal distributions, and trap bias in Northeastern Florida. *Journal of Medical Entomology*, 57(5), mar 2020.

[71] H. Kampen, A. Schuhbauer, and D. Walther. Emerging mosquito species in Germany—a synopsis after 6 years of mosquito monitoring (2011–2016). *Parasitology Research*, 116(12):3253–3263, dec 2017.

[72] R. L. Aldridge, S. C. Britch, M. Snelling, A. Gutierez, G. White, and K. J. Linthicum. Passive baited sequential filth fly trap. *Journal of the American Mosquito Control Association*, 31(3):278–282, 2015.

[73] A. Moore and R. H. Miller. Automated identification of optically sensed Aphid (Homoptera: Aphidae) wingbeat waveforms. *Annals of the Entomological Society of America*, 95(1):1–8, 2002.

[74] I. Potamitis, I. Rigakis, N. Vidakis, M. Petousis, and M. Weber. Affordable bimodal pptical sensors to spread the use of automated insect monitoring. *Journal of Sensors*, 2018:1–25, 2018.

[75] L. Wührl, C. Pylatiuk, M. Giersch, F. Lapp, T. V. Rintelen, M. Balke, S. Schmidt, P. Cerretti, and R. Meier. DiversityScanner: Robotic handling of small invertebrates with machine learning methods. *Molecular Ecology Ressources*, 22:1626–1638, 2021.

[76] S. A. A. Rahuman and D. J. Veerappan. Flying insect identification based on wing-beat frequency using modified SVM classifier. *International Journal of Research and Analytical Reviews*, 21(10):337–342, 2018.

[77] X. Dong, N. Yan, and Y. Wei. Insect sound recognition based on convolutional neural network. In *3rd IEEE International Conference on Image, Vision and Computing*, pages 855–859. IEEE, 2018.

[78] D. Vasconcelos, M. S. Yin, F. Wetjen, A. Herbst, T. Ziemer, A. Forster, T. Barkowsky, N. Nunes, and P. Haddawy. Counting mosquitoes in the wild: An internet of things approach. In *GoodIT 2021 -Conference on Information Technology for Social Good*, pages 43–48, 2021.

[79] I. Kiskin, B. P. Orozco, T. Windebank, D. Zilli, M. Sinka, K. Willis, and S. Roberts. Mosquito detection with neural networks: The buzz of deep learning. *arXiv preprint arXiv:1705.05180*, 2017.

[80] Y. Li, D. Zilli, H. Chan, I. Kiskin, M. Sinka, S. Roberts, and K. Willis. Mosquito detection with low-cost smartphones: data acquisition for malaria research. In *Neural Information Processing Systems, Workshop on Machine Learning for the Developing World. 2017. arXiv: 1711. 06346 v3.*, 2017.

[81] J. W. Chapman, V. A. Drake, and D. R. Reynolds. Recent insights from radar studies of insect flight. *Annual Review of Entomology*, 56:337–356, 2011.

[82] V. A. Drake and D. R. Reynolds. *Radar Entomology: Observing insect flight and migration.* CABI Publishing, Wallingford, United Kingdom, 2012.

[83] A. D. Smith, D. R. Reynolds, and J. R. Riley. The use of vertical-looking radar to continuously monitor the insect fauna flying at altitude over southern England. *Bulletin of Entomological Research*, 90(3):265–277, 2000.

[84] J. W. Chapman, D. R. Reynolds, and A. D. Smith. Vertical-looking radar: A new tool for monitoring high-altitude insect migration. *BioScience*, 53(5):503–511, 2003.

[85] V. A. Drake, H. K. Wang, and I. T. Harman. Insect monitoring radar: Remote and network operation. *Computers and Electronics in Agriculture*, 35(2-3):77–94, 2002.

[86] G. Hu, K. S. Lim, N. Horvitz, S. J. Clark, D. R. Reynolds, N. Sapir, and J. W. Chapman. Mass seasonal bioflows of high-flying insect migrants. *Science*, 354(6319):1584–1587, 2016.

[87] V. A. Drake. Distinguishing target classes in observations from vertically pointing entomological radars. *International Journal of Remote Sensing*, 37(16):3811–3835, 2016.

[88] R. Wang, C. Hu, X. Fu, T. Long, and T. Zeng. Micro-Doppler measurement of insect wing-beat frequencies with W-band coherent radar. *Scientific Reports*, 7(1396), 2017.

[89] Z. Hao, V. A. Drake, J. R. Taylor, and E. Warrant. Insect target classes discerned from entomological radar data. *Remote Sensing*, 12(673), 2020.

[90] G. M. Loper, W. W. Wolf, and O. R. Taylor Jr. Detection and monitoring of honeybee drone congregation areas by radar. *Apidologie*, 18(2):163–172, 1987.

[91] J. R. Riley and A. D. Smith. Design considerations for an harmonic radar to investigate the flight of insects at low altitude. *Computers and Electronics in Agriculture*, 35(2-3):151–169, 2002.

[92] R. M. Measures. *Laser remote sensing, fundamentals and applications*. Krieger, Malabar, FL, 1992.

[93] T. Fujii and T. Fukuchi. *Laser remote sensing*. CRC Press Ed., Boca Raton, FL, 2005.

[94] C. Weitkamp. *Lidar, range-resolved optical remote sensing of the atmosphere*. Springer Edition, New York, NY, 2005.

[95] U. Schumann, B. Weinzierl, O. Reitebuch, H. Schlager, A. Minikin, C. Forster, R. Baumann, T. Sailer, K. Graf, H. Mannstein, C. Voigt, S. Rahm, R. Simmet, M. Scheibe, M. Lichtenstern, P. Stock, H. Ruba, D. Schauble, A. Tafferner, M. Rautenhaus, T. Gerz, H. Ziereis, M. Krautstrunk, C. Mallaun, J. F. Gayet, K. Lieke, K. Kandler, M. Ebert, S. Weinbruch, A. Stohl, J. Gasteiger, S. Gross, V. Freudenthaler, M. Wiegner, A. Ansmann, M. Tesche, H. Olafsson, and K. Sturm. Airborne observations of the Eyjafjalla volcano ash cloud over Europe during air space closure in April and May 2010. *Atmosheric Chemistry and Physics*, 11:2245–2279, 2011.

[96] T. Nishizawa, N. Sugimoto, I. Matsui, and A. Shimizu. Transfer ground-based network observation using Mie – Raman lidars and multi-wavelength Raman lidars and algorithm to retrieve distributions of aerosol components. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 188:79–93, 2017.

[97] A. Miffre, G. David, B. Thomas, and P. Rairoux. Atmospheric non - spherical particles optical properties from UV - polarization lidar and scattering matrix. *Geophysical Research Letters*, 38, 2011.

[98] A. Miffre, G. David, B. Thomas, P. Rairoux, A. M. Fjaeraa, N. I. Kristiansen, and A. Stohl. Volcanic aerosol optical properties and phase partitioning behavior after long-range advection characterized by UV-Lidar measurements. *Atmospheric Environment*, 48:76–84, 2012.

[99] G. David, A. Miffre, B. Thomas, and P. Rairoux. Sensitive and accurate dual-wavelength UV-VIS polarization detector for optical remote sensing of tropospheric aerosols. *Applied Physics B: Lasers and Optics*, 108(1):197–216, 2012.

[100] Y. Dupart, S. M. King, B. Nekat, A. Nowak, A. Wiedensohler, H. Herrmann, G. David, B. Thomas, A. Miffre, P. Rairoux, B. D'Anna, and C. George. Mineral dust photochemistry induces nucleation events in the presence of SO2. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 109(51):20842–20847, 2012.

[101] J. B. Abshire, H. Riris, G. R. Allan, S. R. Kawa, X. Sun, W. E. Hasselbrack, C. Weaver, M. R. Rodriguez, and J. Mao. Measuring atmospheric CO2 for the NASA ascends mission: The CO2 laser sounder approach. In *IEEE International Geoscience and Remote Sensing Symposium*, 2012.

[102] G. Ehret, C. Kiemle, M. Wirth, A. Amediek, A. Fix, and S. Houweling. Space-borne remote sensing of CO2, CH4, and N2O by integrated path differential absorption lidar: a sensitivity analysis. *Applied Physics B: Lasers and Optics*, 90:593–608, 2008.

[103] B. Thomas, G. David, C. Anselmo, E. Coillet, K. Rieth, A. Miffre, J. Cariou, and P. Rairoux. Remote sensing of methane with broadband laser and optical correlation spectroscopy on the Q-branch of the 2V3 band. *Journal of Molecular Spectroscopy*, 291:3–8, 2013.

[104] S. Kuang, M. J. Newchurch, J. Burris, and X. Liu. Ground-based lidar for atmospheric boundary layer ozone measurements. *Applied Optics*, 52(15):3557–3566, 2013.

[105] D. N. Whiteman, D. Venable, and E. Landulfo. Comments on "Accuracy of Raman lidar water vapor calibration and its applicability to long-term measurements". *Applied Optics*, 50(15):2170–2176, 2011.

[106] B. Thomas, G. David, C. Anselmo, J. Cariou, A. Miffre, and P. Rairoux. Remote sensing of atmospheric gases with optical correlation spectroscopy and lidar: first experimental results on water vapor profile measurements. *Applied Physics B: Lasers and Optics*, 113:265–275, 2013.

[107] R. Robinson, T. Gardiner, F. Innocenti, P. Woods, and M. Coleman. Infrared differential absorption Lidar (DIAL) measurements of hydrocarbon emissions. *Journal of Environmental Monitoring*, 13:2213–2220, 2011.

[108] M. Brydegaard, S. Jansson, E. Malmqvist, Y. P. Mlacha, A. Gebru, F. Okumu, G. F. Killeen, and C. Kirkeby. Lidar reveals activity anomaly of malaria vectors during pan-African eclipse. *Science Advances*, 6(20), 2020.

[109] Z. Song, B. Zhang, H. Feng, S. Zhu, L. Hu, M. Brydegaard, Y. Li, S. Jansson, E. Malmqvist, K. Svanberg, G. Zhao, J. Bood, S. Svanberg, and D. Li. Application of lidar remote sensing of insects in agricultural entomology on the Chinese scene. *Journal of Applied Entomology*, 144(3):161–169, apr 2020.

[110] S. Jansson, E. Malmqvist, Y. Mlacha, R. Ignell, F. Okumu, G. Killeen, C. Kirkeby, and M. Brydegaard. Real-time dispersal of malaria vectors in rural Africa monitored with lidar. *PLoS ONE*, 16(3), 2021.

[111] M. Li, S. Jansson, A. Runemark, J. Peterson, C. T. Kirkeby, A. M. Jönsson, and M. Brydegaard. Bark beetles as lidar targets and prospects of photonic surveillance. *Journal of Biophotonics*, 14, 2021.

[112] D. S. Hoffman, A. R. Nehrir, K. S. Repasky, J. A. Shaw, and J. L. Carlsten. Range-resolved optical detection of honeybees by use of wing-beat modulation of scattered light for locating land mines. *Applied Optics*, 46(15):3007–3012, 2007.

[113] Y. Li, K. Wang, R. Quintero-Torres, R. Brick, A. Sokolov, and M. Scully. Insect flight velocity measurement with a CW near-IR Scheimpflug lidar system. *Optics Express*, 28(15):21891–21902, 2020.

[114] E. Malmqvist, S. Jansson, S. Török, and M. Brydegaard. Effective parameterization of laser radar observations of atmospheric fauna. *IEEE Journal on Selected Topics in Quantum Electronics*, 22(3), 2016.

[115] Z. Guan, M. Brydegaard, P. Lundin, M. Wellenreuther, A. Runemark, E. I. Svensson, and S. Svanberg. Insect monitoring with fluorescence lidar techniques: field experiments. *Applied Optics*, 49(27):5133–5142, 2010.

[116] L. Mei, Z. G. Guan, H. J. Zhou, J. Lv, Z. R. Zhu, J. A. Cheng, F. J. Chen, C. Löfstedt, S. Svanberg, and G. Somesfalean. Agricultural pest monitoring using fluorescence lidar techniques. *Applied Physics B: Lasers and Optics*, 106:733–740, 2012.

[117] M. Brydegaard. Towards quantitative optical cross sections in entomological laser radar - Potential of temporal and spherical parameterizations for identifying atmospheric fauna. *PLoS ONE*, 10(8), 2015.

[118] A. Gebru, M. Brydegaard, E. Rohwer, and P. Neethling. Probing insect backscatter cross section and melanization using kHz optical remote detection system. In *Proceedings SPIE, Remote Sensing and Modeling of Ecosystems for Sustainability XIII, 997504*, volume 9975, 2017.

[119] C. Kirkeby, M. Wellenreuther, and M. Brydegaard. Observations of movement dynamics of flying insects using high resolution lidar. *Scientific Reports*, 6(29083), 2016.

[120] Y. Wen, L. E. Muir, and B. H. Kay. Response of Culex quinquefasciatus to visual stimuli. *Journal of the American Mosquito Control Associatio*, 13(2):150, 1997.

[121] L. E. Muir, M. J. Thorne, and B. H. Kay. Aedes aegypti (Diptera: Culicidae) vision: spectral sensitivity and other perceptual parameters of the female eye. *Journal of Medical Entomology*, 29(2):278–281, 1992.

[122] V. S. Mayagaya, K. Michel, M. Q. Benedict, G. F. Killeen, R. A. Wirtz, H. M. Ferguson, and F. E. Dowell. Non-destructive determination of age and species of Anopheles gambiae s.l. using near-infrared spectroscopy. *American Journal of Tropical Medicine and Hygiene*, 81(4):622–630, 2009.

[123] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi. Mel-frequency cepstral coefficient analysis in speech recognition. In *2006 International Conference on Computing & Informatics*, Kuala Lumpur, Malaysia, 2006. IEEE.

[124] L. Muda, M. Begam, and I. Elamvazuthi. Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. *Journal of Computing*, 2(3):138–143, 2010.

[125] A. Y. Gbogbo, B. K. Kouakou, S. Dabo-Niang, and J. T. Zoueu. Predictive model for airborne insect abundance intercepted by a continuous wave Scheimpflug lidar in relation to meteorological parameters. *Ecological Informatics*, 68, 2021.

[126] M. Brydegaard, A. Gebru, and S. Svanberg. Super resolution laser radar with blinking atmospheric particles — application to interacting flying insects. *Progress in Electromagnetics Research*, 147:141–151, 2014.

[127] J. M. Alvarez, M. A. Vaughan, C. A. Hostetler, W. H. Hunt, and D. M. Winker. Calibration technique for polarization-sensitive lidars. *Journal of Atmospheric and Oceanic Technology*, 23(5):683–699, 2006.

[128] R. T. Carde. *Odour plumes and odour-mediated flight in insects*. Ciba Foundation Symposium 200, John Wiley & Sons, Chichester, United Kingdom, 1996.

[129] S. Petrovskii, D. Bearup, D. A. Ahmed, and R. Blackshaw. Estimating insect population density from trap counts. *Ecological Complexity*, 10:69–82, 2012.

[130] J. M. Cook. Rational formulae for the production of a spherically symmetric probability distribution. *Mathematics of Computation*, 11:81–82, 1957.

[131] S. Küshel, A. Bruckner, S. Schmelze, M. Heethoff, and N. Bluthgen. Surface area – volume ratios in insects. *Insect Science*, 24:829–841, 2017.

[132] M. Stone. Cross-validatory choice and assessment of statistical. *Journal of the Royal Statistical Society*, 36(2):111–147, 1974.

[133] N. V. Chawla, N. Japkowicz, and A. Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1), 2004.

[134] Y. Sun, A. Wong, and M. S. Kamel. Classification of imbalanced data : a review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

[135] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri. *The imbalanced training sample problem: Under or over sampling?*, volume 3138. Springer, Berlin, Germany, 2004.

[136] M. Kubat and S. Matwin. Addresing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186, 1997.

[137] M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Manadgement Process*, 5(2), 2015.

[138] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[139] S. Russell and P. Norvig. *Artificial intelligence a modern approach*. Pearson Education Inc., Upper Saddle River, NJ, 3rd edition, 2010.

[140] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[141] E. Fix and J. L. Hodges. *Discriminatory analysis. Nonparametric discrimination: Consistency properties*. USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[142] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2nd edition, 2005.

[143] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2:345–389, 1998.

[144] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Taylor & Francis, Ltd., Boca Raton, FL, 1984.

[145] S. Singh and M. Giri. Comparative study Id3, Cart And C4.5 decision tree algorithm: A survey. *International Journal of Advanced Information Science and Technology*, 3(7):47–52, 2014.

[146] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7:179–188, 1936.

[147] T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, 10(4):453–472, 2006.

[148] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[149] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, United Kingdom, 2000.

[150] C. Hsu and C. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 2002.

[151] S. M. Villarreal, O. Winokur, and L. Harrington. The impact of temperature and body size on fundamental flight tone variation in the mosquito vector Aedes aegypti (Diptera: Culicidae): Implications for acoustic lures. *Journal of Medical Entomology*, 54(5):1116–1121, 2017.

[152] P. Reiter. A portable, battery-powered trap for collecting gravid Culex mosquitos. *Mosquito News*, 43(4):496–498, 1983.

[153] J. J. Scott, S. C. Crans, and W. J. Crans. Use of an infusion-baited gravid trap to collect adult Ochlerotatus japonicus. *Journal of the American Mosquito Control Association*, 17:142–143, 2001.

[154] D. M. Unwin and S. A. Corbet. Wingbeat frequency, temperature and body size in bees and flies. *Physiological Entomology*, 9(1):115–121, 1984.

[155] I. Potamitis and I. Rigakis. Measuring the fundamental frequency and the harmonic properties of the wingbeat of a large number of mosquitoes in flight using 2D optoacoustic sensors. *Applied Acoustics*, 109:54–60, 2016.

[156] A. Gebru, S. Jansson, R. Ignell, C. Kirkeby, J. C. Prangsma, and M. Brydegaard. Multiband modulation spectroscopy for the determination of sex and species of mosquitoes in flight. *Journal of Biophotonics*, 11(8), 2018.

[157] D. Vasconcelos, N. Nunes, M. Ribeiro, C. Prandi, and A. Rogers. LOCOMOBIS: a low-cost acoustic-based sensing system to monitor and classify mosquitoes. In *16th IEEE Annual Consumer Communications & Networking Conference*. IEEE, 2019.

[158] M. I. G. Pérez, B. Faulhaber, M. Williams, J. Brosa, C. Aranda, N. Pujol, M. Verdún, P. Villalonga, J. Encarnação, N. Busquets, and S. Talavera. A novel optical sensor system for the automatic classification of mosquitoes by genus and sex with high levels of accuracy. *Parasites and Vectors*, 15, 2022.

[159] M. A. Caprio, J. X. Huang, M. K. Faver, and A. Moore. Characterization of male and female wingbeat frequencies in the Anopheles quadrimaculatus complex in Mississippi. *Journal of the American Mosquito Contol Association*, 17(3):186–189, 2001.

[160] J. Wang, S. Zhu, Y. Lin, S. Svanberg, and G. Zhao. Mosquito counting system based on optical sensing. *Applied Physics B: Lasers and Optics*, 126(2), 2020.

[161] B. J. Arthur, K. S. Emr, R. A. Wyttenbach, and R. R. Hoy. Mosquito (Aedes aegypti) flight tones: Frequency, harmonicity, spherical spreading, and phase relationships. *The Journal of the Acoustical Society of America*, 135(2):933–941, 2014.

[162] P. M. V. Simões, G. Gibson, and I. J. Russell. Pre-copula acoustic behaviour of males in the malarial mosquitoes Anopheles coluzzii and Anopheles gambiae s.s. does not contribute to reproductive isolation. *The Journal of Experimental Biology*, 220(3):379–385, 2016.

[163] D. J. Hand and K. Yu. Idiot's Bayes-not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

[164] G. E. A. P. A. Batista, Y. Hao, E. Keogh, and A. Mafra-neto. Towards automatic classification on flying insects using inexpensive sensors. In *10th International Conference on Machine Learning and Applications and Workshops (ICMLA), Honolulu, HI*, pages 364–369, 2011.

[165] Y. Chen, A. Why, G. E. A. P. A. Batista, A. Mafra-Neto, and E. Keogh. Flying insect classification with inexpensive sensors. *Journal of Insect Behavior*, 27(5):657–677, 2014.

[166] R. S. Lees, J. R. L. Gilles, J. Hendrichs, M. J. B. Vreysen, and K. Bourtzis. Back to the future: the sterile insect technique against mosquito disease vectors. *Current Opinion in Insect Science*, 10:156–162, 2015.

[167] L. Alphey, M. Benedict, R. Bellini, G. G. Clark, D. A. Dame, M. W. Service, and S. L. Dobson. Sterile-insect methods for control of mosquito-borne diseases: An analysis. *Vector-Borne and Zoonotic Diseases*, 10(3):295–311, 2010.

[168] M. Q. Benedict and A. S. Robinson. The first releases of transgenic mosquitoes: An argument for the sterile insect technique. *Trends in Parasitology*, 19(8):349–355, 2003.

[169] D. F. Silva, V. M. A.. De Souza, G. E. A. P. A. Batista, E. Keogh, and D. P. W. Ellis. Applying machine learning and audio analysis techniques to insect recognition in intelligent traps. *Proceedings - 12th International Conference on Machine Learning and Applications*, 1:99–104, 2013.

[170] T. S. Detinova, W. N. Beklemishev, and D. S. Bertram. *Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria.* World Health Organization monograph series, numero 47, Geneva, 1962.

[171] P. Belton. *Insect flight: Dispersal and mitigation, sounds of insects in flight.* Springer-Verlag Berlin Heidelberg, Germany, 1986.

[172] D. M. Le Vine. Comment on the extinction paradox. Technical report, NASA technical memorandum TM-85113, 1983.

[173] H. M. Lai, W. Y. Wong, and W. H. Wong. Extinction paradox and actual power scattered in light beam scattering: a two-dimensional study. *Journal of the Optical Society of America A*, 21(12), 2004.

[174] M. J. Berg, C. M. Sorensen, and A. Chakrabarti. A new explanation of the extinction paradox. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(7):1170–1181, 2011.

[175] M. J. Berg and C. M. Sorensen. A review and reassessment of diffraction, scattering, and shadows in electrodynamics. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 210:225–239, 2018.

[176] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

[177] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[178] R. J. Bomphrey, T. Nakata, N. Phillips, and Si. M. Walker. Smart wing rotation and trailing-edge vortices enable high frequency mosquito flight. *Nature*, 544(7648):92–95, 2017.

[179] J.J. Oertli. Relationship of wing beat frequency and temperature during take-off flight in temperate-zone beetles. *Journal of Experimental Biology*, 145:321–338, 1989.

[180] J. A. Shaw, N. L. Seldomridge, D. L. Dunkle, P. W. Nugent, L. H. Spangler, J. J. Bromenshenk, C. B. Henderson, J. H. Churnside, and J. J. Wilson. Polarization lidar measurements of honey bees in flight for locating land mines. *Optics Express*, 13(15):5853, 2005.

[181] A. P. Genoud, R. Basistyy, Gr. M. Williams, and B. P. Thomas. Optical remote sensing for monitoring flying mosquitoes, gender identification and discussion on species identification. *Applied Physics B: Lasers and Optics*, 124(3):46, 2018.

[182] A. P. Genoud, R. Basistyy, G. M. Williams, and B. P. Thomas. Analysis of predictor variables for mosquito species identification from dual-wavelength polarization-sensitive lidar measurements. *Proc. SPIE, 10779, Lidar Remote Sensing for Environmental Monitoring XVI, Honolulu, HI, USA*, 10779:24, 2018.

[183] A. P. Genoud, Y. Gao, G. M. Williams, and B. P. Thomas. Identification of gravid mosquitoes from changes in spectral and polarimetric backscatter cross sections. *Journal of Biophotonics*, 12(10), 2019.

[184] A. P. Genoud, J. Torsiello, M. Belson, and B. P. Thomas. Entomological photonic sensors: Estimating insect population density, its uncertainty and temporal resolution from transit data. *Ecological Informatics*, 61(November):101186, 2020.

[185] A. P. Genoud, Y. Gao, G. M. Williams, and B. P. Thomas. A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals. *Ecological Informatics*, 58, jul 2020.

[186] A. P. Genoud, G. M. Williams, and B. P. Thomas. Continuous monitoring of aerial density and circadian rhythms of flying insects in a semi-urban environment. *PLoS ONE*, 16(11), 2021.

[187] R. Wehner. Astronavigation in insects. *Annual Review of Entomology*, 29:277–298, 1984.

[188] R. Se. Pandian and M. K. Chandrashekaran. Rhythms in the biting behaviour of a mosquito Armigeres subalbatus. *Oecologia*, 47:89–95, 1980.

[189] A. C. S. Owens and S. M. Lewis. The impact of artificial light at night on nocturnal insects : A review and synthesis. *Ecology and Evolution*, 8:11337–11358, 2018.