Aristotle University of Thessaloniki

Data-Driven Prediction for Mosquitoes Population Based on Earth Observation Data

Student: Argyro Tsantalidou (Student ID: 50)
Supervisors: Prof. Athena Vakali, Dr. George Arvanitakis

Elaboration of dissertation as part of Web and Data Science M.Sc. Programme
Faculty of Sciences
School of Informatics

March 2021
Statement Of Authorship

I, Argyro Tsantalidou, declare that the thesis with title, "Data-Driven Prediction for Mosquitoes Population Based on Earth Observation Data", and the work presented in it are a genuine work of mine. I ensure that:

• This work has taken place wholly or mainly during my candidacy for graduate studies at this university.

• Wherever any part of this dissertation has previously been used for the acquisition of a degree or other qualification in this or another university, this is clearly stated.

• Wherever I have consulted the third-party work published, this is correctly attributed.

• Wherever I have cited from third-party work, the source is always given. With the exception of these quotes, this thesis is entirely my personal work.

• I have cited all the utility sources.

• Wherever this thesis is based on collaborative work of my own and of others, I have made it clear which pieces have been made by others and how I contributed.

Signature: 

Date:
Abstract

Faculty of Sciences
School of Informatics
Data and Web Science Master’s Degree

by Argyro Tsantalidou

The set (volume) of environmental, meteorological, entomological and other spatio-temporal parameters affecting mosquito abundance makes the modeling and prediction tasks quite challenging. Up to now, mosquitoes abundance problems were addressed with ad-hoc area-specific and genus-tailored approaches. This thesis proposes and develops MAMOTH, a generic and accurate framework that predicts mosquito abundance for the upcoming period (the mean absolute error of the predictions do not deviate more than 14%). The designed machine learning model relies on satellite EO and geo-spatial data to tackle the problem. MAMOTH is independent of the application area or mosquito genus, and can be easily replicated, and applied to any case without any special requirement. The framework was applied to different mosquitoes genes (Culex - West Nile Virus, Anopheles - Malaria and Aedes - Zika / Chikungunya / Dengue virus) and different areas of interest (Italy, Serbia, France, Germany). The results show that it performs accurately and consistently for all cases of study. Additionally, the evaluation of different cases with the same framework using the same principles opens the opportunity for multi-case and multi-scope comparative studies.

More specifically, Chapter 1 outlines the problem addressed and the concept of the thesis, placing emphasis on the MBDs and their spread across Europe. Mosquito Borne Diseases (MBDs) is a problem that Europe is facing many years now, proving to be an insisted problem, with very complex solution. Preventive actions along with the epidemiological surveillance of MBDs and implementation of integrated mosquito control programmes have been employed to combat them, but an early warning system would provide great help in the mitigation of MBDs. Although the development of an early warning system have proven to be challenging so far due to the lack of well structured, consistent and reliable environmental, landscape and ecosystem data, this chapter delineates and offers a top-level description of the framework developed, MAMOTH, to overcome these challenges and help in MBDs minimization highlighting the open critical problems and presents the main contributions of this work.

Chapter 2 summarizes the literature of the most related research topics and themes. Emphasis is given on earlier work in the areas of remote sensing technologies, risk
factors contributing in the dissemination of the MBDs and the prediction of infectious diseases outbreaks. A comparative analysis is provided to highlight earlier approaches bottlenecks and to indicate the most crucial open issues which are addressed in this thesis. The approaches studied, can be splitted into two categories, those directly predicting the incidence of infections in an area, and those predicting the population of vectors which can be considered as a proxy for the upcoming human cases of an MBD. However a common characteristic among all the approaches is the use of any kind of environmental data, with temperature to be included in all of them. Finally, in all of the cases studied, the proposed approaches are tailored upon specific MBD and aim towards its minimization. The proposed framework of this thesis, fill the gap of a flexible model able to deliver predictions regardless the MBD of interest.

Chapter 3 covers the cases of study that will explore in the rest of the thesis. Each case is characterized by two main components i) the area of interest and ii) The mosquito genus with the corresponding transmitted disease. For each area of interest some basic information about its the topology and the climatological conditions are given. In addition, some basic information about mosquitoes genus and the corresponding diseases in terms of human cases, symptoms and treatment is given in an attempt to describe the content to which the framework was applied.

Chapter 4 presents an overview on earth observation data and describes the data used for the implementation of the framework. A brief introduction on remote sensing technologies is given in the beginning to introduce the technology behind earth observation data collection. After, the chapter focuses on the description of the dataset collected for the implementation of the framework. It describes the entomological data collected, and the earth observational data that are extracted out of images. After that, data preprocessing such as cleaning transformation and imputation of missing values is getting described to give a full overview of the data and their manipulation.

Chapter 5 is focusing on giving a description about the framework. The so called framework MAMOTH, is able to predict the upcoming mosquito abundance in form of risk classes from 1-10. The innovation of this framework is its capability for prediction in any area of any mosquito genus without modification of the features as it automatically wraps itself around any case. MAMOTH is consisting of 4 modules and this chapter elaborates on each one of these individual pieces and methods. It describes the process from start to end of how the MAMOTH is receiving a set of features and ends up with a model consisting of only a subset of them being able to deliver accurate predictions about the mosquito population of the next days. It also describes all the technical information such as the cost function used, and the hyper parameter tuning. Besides MAMOTH framework though this chapter also describes the evaluation methods and metrics under which the framework is assessed.

Chapter 6 discusses the results upon the cases of studies in which the MAMOTH was applied. It presents the results per case with plots about the error distribution and the error per class, while it also presents the features selected by the framework per case. Overall the framework performed consistently across all 5 cases predicting accurate the mosquito population. As mentioned before MAMOTH is using entomological
and earth observation environmental data to make predictions. However an evaluation of the framework with the use of only earth observation environmental data and not at all use of entomological data showed that is still able to predict the mosquito population with a small decrease in the overall performance.

Chapter 7 summarizes the results of the experimental process. The experimentation validates the claims for an reliable and accurate framework able to predict mosquito abundance efficiently with respect to the the area and mosquito genus. After studying the results obtained out of all 5 cases, max MAE is equal to 1.54, while minimum MAE is equal to 0.71. Even in the worst case, the magnitude of error is tolerable considering the fact the are 10 possible classes for each prediction. Furthermore accumulated rainfall from the beginning of the year and from the last two weeks seems to play an important in all of the cases. The only case that MAMOTH did not demonstrate high performance was in the Greek landscape, probably due to the extended irrigation systems against the rivery systems of the other cases. Finally experiments proved that MAMOTH is still able to predict mosquito population without any feature about the state of mosquito population in the past days, relying only in environmental data, causing although a small decrease in the performance.
Περίληψη

Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής
Π.Μ.Σ. Επιστήμης Δεδομένων και Παγκόσμιου Ιστού
της Αργυρώς Τσανταλίδου

Το σύνολο (όγκος) περιβαλλοντικών, μετεωρολογικών, εντομολογικών και άλλων χωροχρονικών παραμέτρων που επηρεάζουν την αφθονία των κουνουπιών καθιστά τις εργασίες μοντελοποίησης και πρόβλεψης αρκετά δύσκολες. Μέχρι τώρα, τα προβλήματα σχετικά με τον πληθυσμό των κουνουπιών αντιμετωπίζονταν με ειδικές προσεγγίσεις για συγκεκριμένες περιοχές και προσαρμοσμένες στο γένος. Αυτή η εργασία πroteίνει και αναπτύσσει το ΜΑΜΟΘ, ένα γενικό και ακριβές μοντέλο που προβλέπει τον πληθυσμό των κουνουπιών για την προσεχή περίοδο (το μέσο απόλυτο σφάλμα των προβλέψεων δεν αποκλίνει περισσότερο από 14%). Το μοντέλο μηχανικής μάθησης που αναπτύχθηκε βασίζεται σε δορυφορικά και γεω-χωρικά δεδομένα για την αντιμετώπιση του προβλήματος. Το ΜΑΜΟΘ είναι ανεξάρτητο από την περιοχή εφαρμογής ή το γένος κουνουπιών, και μπορεί εύκολα να ανα-παραχθεί και να εφαρμοστεί σε κάθε περίπτωση χωρίς καμία ειδική απαίτηση. Το μοντέλο εφαρμόστηκε σε διαφορετικά είδη κουνουπιών (Culex - Ιός του Δυτικού Νείλου, Anopheles - Ελονοσία και Aedes - Zika / Chikungunya / Δάγκειος ιός) και σε διαφορετικές περιοχές ενδιαφέροντος (Ιταλία, Σερβία, Γαλλία, Γερμανία). Τα αποτελέσματα δείχνουν ότι αποδίδει με ακρίβεια και συνέπεια για όλες τις περιπτώσεις μελέτης. Επιπλέον, η αξιολόγηση διαφορετικών περιπτώσεων με αυτό το μοντέλο χρησιμοποιούντας τις ίδιες αρχές ανοίγει την ευκαιρία για συγκριτικές μελέτες μεταξύ διαφορετικών περιπτώσεων.

Πιο συγκεκριμένα, το Κεφάλαιο 1 περιγράφει το πρόβλημα που αντιμετωπίζει αυτή η εργασία, δίνοντας έμφαση στις ασθένειες που μεταδίδονται μέσω των κουνουπιών και την εξάπλωσή τους σε ολόκληρη την Ευρώπη. Αυτές οι ασθένειες είναι ένα πρόβλημα που αντιμετωπίζει η Ευρώπη εδώ και πολλά χρόνια, αποδεικνύοντας ότι είναι ένα επίμονο πρόβλημα για πολλά περιπλοκές λύση. Προληπτικές δράσεις, μαζί με την επιδημιολογική παρακολούθηση των ασθενειών αυτών και την εφαρμογή συστημάτων προγραμμάτων καταπολέμησης κουνουπιών έχουν χρησιμοποιηθεί σε μια προσπάθεια καταπόλεμησής τους, αλλά ένα σύστημα έγκαιρης προειδοποίησης θα παρέχει μεγάλη βοήθεια στον μετριασμό τους. Παρόλο που η ανάπτυξη ενός συστήματος έγκαιρης προειδοποίησης έχει αποδειχθεί ότι είναι δύσκολη μέχρι στιγμής λόγω της έλλειψης καλά δομημένων, συνεπών και αξιόπιστων δεδομένων για το περιβάλλον, το τοπίο και τα οικοσυστήματα, αυτό το κεφάλαιο προσφέρει μια περιγραφή του πλαισίου που αναπτύχθηκε, ΜΑΜΟΘ, για να ξεπεραστούν αυτές τις προκλήσεις και να βοηθήσει στην ελαχιστοποίηση των ασθενειών που μεταδίδονται.
από τα κουνούπια, επισημαίνοντας τα ανοιχτά κρίσιμα προβλήματα και παρουσιάζει τις κύριες συνεισφορές αυτού του έργου.

Το Κεφάλαιο 2 συνοψίζει τη βιβλιογραφία των πιο σχετικών ερευνητικών θεμάτων. Έμφαση δίνεται σε προηγούμενες εργασίες στους τομείς των τεχνολογιών τηλεπισκόπησης, των παραγόντων κινδύνου που συμβάλλουν στη διάδοση των ασθενειών που μεταδίδονται από κουνούπια και στην πρόβλεψη των επιδημιών αυτών. Παρέχεται μια συγκριτική ανάλυση για να επισημανθούν αδύναμα σημεία των προσεγγίσεων αυτών και να υποδεχθούν τα πιο κρίσιμα ανοιχτά ζητήματα που αντιμετωπίζονται σε αυτή την εργασία. Οι προσεγγίσεις που μελετήθηκαν, μπορούν να χαρακτηριστούν σε δύο κατηγορίες, εκείνες που προβλέπουν άμεσα τη συχνότητα εμφάνισης λοίμωξης σε μια περιοχή και εκείνες που προβλέπουν τον πληθυσμό των φορέων που μπορούν να θεωρηθούν ως προσαρμοστικές προσπάθειες τέτοιων ασθενειών. Ωστόσο, ένα κοινό χαρακτηριστικό μεταξύ όλων των προσεγγίσεων είναι η χρήση διαφορικών περιβάλλοντικών δεδομένων, με την θερμοκρασία να έχει συμπεριληφθεί σε όλες αυτές τις προσπάθειες που μελετήθηκαν, οι προτεινόμενες προσεγγίσεις προσαρμόζονται σε μεμονωμένες ασθένειες στοχεύοντας στην εικονοποίησή τους. Το προτεινόμενο πλαίσιο αυτής της διατριβής, γεμίζει το κενό στην ύπαρξη ενός ευέλικτου μοντέλου που μπορεί να προσφέρει προβλέψεις ανεξάρτητα από την ασθένεια που μελετάται.

Το Κεφάλαιο 3 καλύπτει τις περιπτώσεις μελέτης που θα διερευνηθούν στο υπόλοιπο της έρευνας. Κάθε περίπτωση χαρακτηρίζεται από δύο κύρια συστατικά: i) την περιοχή ενδιαφέροντος και ii) το γένος κουνουπιών με την αντίστοιχη μεταδομένη ασθένεια. Για κάθε περιοχή ενδιαφέροντος δίνονται βασικές πληροφορίες σχετικά με την τοπολογία και τις κλιματολογικές συνθήκες. Επιπλέον, ορισμένες βασικές πληροφορίες σχετικά με το γένος των κουνουπιών και τις αντίστοιχες ασθένειες από άποψη προπέρας δόλωσης, συμπτωμάτων και θεραπείας σε μια προσπάθεια να περιγραφεί το περιεχόμενο στο οποίο εφαρμόστηκε το μοντέλο.

Το Κεφάλαιο 4 παρουσιάζει μια επισκόπηση σχετικά με τα δεδομένα παρατήρησης της γης και περιγράφει τα δεδομένα που χρησιμοποιήθηκαν για την εφαρμογή του ΜΑΜΟΤΗ. Μια σύντομη εισαγωγή για τις τεχνολογίες τηλεπισκόπησης δίνεται στην αρχή για να εισαχτεί την τεχνολογία πίσω από τη συλλογή δεδομένων παρατήρησης της γης. Στην συνέχεια, το κεφάλαιο εστιάζει στην περιγραφή του συνόλου δεδομένων που συλλέγονται, δηλαδή τα εντομολογικά δεδομένα και τα δεδομένα παρατήρησης της γης που εξάγονται από εικόνες. Μετά από αυτό, περιγράφεται η προεπεξεργασία δεδομένων όπως ο καθαρισμός του μετασχηματισμού και ο υπολογισμός των τιμών που λείπουν για να δώσει μια πλήρη επισκόπηση των δεδομένων και του χειρισμού τους.

Το Κεφάλαιο 5 εστιάζει στην περιγραφή του ΜΑΜΟΤΗ. Το ΜΑΜΟΤΗ, είναι σε θέση να προβλέψει τον μελλοντικό πληθυσμό κουνουπιών με τη μορφή κατηγοριών κινδύνου από το 1-10. Η καινοτομία αυτού του συστήματος είναι η ικανότητά του για πρόβλεψη σε οποιαδήποτε περιοχή και οποιοιδήποτε γένης κουνουπιών χωρίς τροποποίηση των εξαρτημένων μεταβλητών καθώς το σύστημα προσαρμόζεται αυτόματα σε κάθε περίπτωση. Το ΜΑΜΟΤΗ αποτελείται από 4 ενότητες και αυτό το κεφάλαιο επεξεργάζεται κάθε ένα από αυτά τα μεμονωμένα κομμάτια και τις μεθόδους που χρησιμοποιούν. Περιγράφει τη διαδικασία από την αρχή ώς το τέλος του τρόπο
με τον οποίο το ΜΑΜΟΘ λαμβάνει ένα σύνολο χαρακτηριστικών και καταλήγει σε ένα μοντέλο που αποτελείται από μόνο ένα υποσύνολο από αυτά που μπορούν να προσφέρουν ακριβείς προβλέψεις για τον πληθυσμό των κουνουπιών των επόμενων ημερών. Περιγράφει επίσης όλες τις τεχνικές πληροφορίες, όπως η συνάρτηση κόστους που χρησιμοποιείται και η επιλογή των υπερπαραμέτρων. Ωστόσο, εκτός από το ΜΑΜΟΘ, αυτό το κεφάλαιο περιγράφει επίσης τις μεθόδους και τις μετρικές αξιολόγησης βάσει των οποίων αξιολογείται το σύστημα.

Το Κεφάλαιο 6 περιγράφει τα αποτελέσματα των περιπτώσεων στις οποίες εφαρμόστηκε το ΜΑΜΟΘ. Παρουσιάζει τα αποτελέσματα με γραφήματα σχετικά με την κατανομή σφαλμάτων και το σφάλμα ανά κατηγορία πληθυσμού, ενώ παρουσιάζει και τις ανεξάρτητες μεταβλητές που επιλέχθηκαν από το σύστημα ανά περίπτωση. Συνολικά, το σύστημα συμπεριφέρθηκε με συνέπεια και στις 5 περιπτώσεις προβλέποντας τον πληθυσμό των κουνουπιών με μια σχετική ακρίβεια. Όπως αναφέρθηκε προηγουμένως, το ΜΑΜΟΘ χρησιμοποιεί εντομολογικά και περιβαλλοντικά δεδομένα με παρατήρηση της γης για να κάνει προβλέψεις. Ωστόσο, μια αξιολόγηση του συστήματος με τη χρήση μόνο περιβαλλοντικών δεδομένων παρατήρησης της γης και καθόλου χρήση εντομολογικών δεδομένων έδειξε ότι εξακολουθεί να είναι σε θέση να προβλέψει τον πληθυσμό των κουνουπιών με μικρή μείωση της συνολικής απόδοσης.

Το Κεφάλαιο 7 συνοψίζει τα αποτελέσματα της πειραματικής διαδικασίας. Ο πειραματισμός επικυρώνει τους ισχυρισμούς για ένα αξιόπιστο και ακριβές πλαίσιο ικανό να προβλέψει αποτελέσματα πληθυσμού των κουνουπιών συναρτήσεις με την περιοχή και το γένος των κουνουπιών. Αφού μελετήθηκε η ακρίβεια των αποτελεσμάτων από τις 5 περιπτώσεις, το μέγιστο MAE είναι ισο με 1,54, ενώ το ελάχιστο MAE είναι ισο με 0,71. Ακόμη και στη χειρότερη περίπτωση, το μέγεθος του σφάλματος είναι ανεκτό λαμβάνοντας υπόψη το γεγονός ότι είναι 10 πιθανές κατηγορίες για κάθε πρόβλεψη. Επιπλέον, οι συσσωρευμένες βροχοπτώσεις από τις αρχές του έτους και από τις δύο τελευταίες εβδομάδες φαίνεται να διαδραματίζουν σημαντικό ρόλο σε όλες τις περιπτώσεις στις οποίες εφαρμόστηκε το ΜΑΜΟΘ. Η μόνη περίπτωση που η ΜΑΜΟΘ δεν έδειξε υψηλή απόδοση ήταν στο ελληνικό τοπίο, πιθανώς λόγω των εκτεταμένων αρδευτικών συστημάτων και των ποταμιών των άλλων περιπτώσεων. Τέλος, τα πειράματα απεδείχναν ότι το ΜΑΜΟΘ εξακολουθεί να είναι σε θέση να προβλέψει τον πληθυσμό των κουνουπιών χωρίς να έχει στην διάθεσή του κανένα χαρακτηριστικό σχετικά με την κατάσταση του πληθυσμού των κουνουπιών της τελευταίας ημέρας, βασιζόμενο μόνο σε περιβαλλοντικά δεδομένα, με μια μικρή μείωση της απόδοσης.
Acknowledgements

The elaboration of the present research work presented many difficulties and was particularly time consuming. However, I would like to thank those helped me with practical ideas during the implementation and writing process and supported me the whole time.

First and foremost, I would like to thank my supervisor Pr. Athena Vakali, for the assignation of this thesis and the support she has shown me. Also, she provided me with great knowledge and skills throughout my bachelor’s and master’s studies.

Additionally, I would like to thank the Beyond team from the National Observatory of Athens and the director Dr. Charalampos Kontoes for giving me the chance to be part of their team during the implementation of this thesis. Through this experience I grew new skills and gain more knowledge.

Finally, I would like to thank Dr. George Arvanitakis, postdoctoral researcher in CS department of Aristotle University of Thessaloniki for explaining every ambiguity that was arising during the elaboration of the thesis and providing me with insightful ideas and solutions.
Thesis Achievements

This thesis took place from September 2020 to February 2021. Through that period, a number of achievements took place about the dissemination and recognition of the work.

- Part of this work have been reported in the Public Authorities of the Veneto Region, Italy for early warning report against West Nile Virus disease for the months September and October 2020.


- Grant for Exceptional Master Student Performance from the BEYOND Center for EO Research and Satellite Remote Sensing - National Observatory of Athens.

- This work was presented in the framework of the webinars organized by EYWA (Early Warning System for Mosquito-Borne Diseases) addressed three different target groups: 1) European, 2) Local and 3) non-European decision makers. Attended by 51 stakeholders representing indicatively, local (HNPHO), European (ECDC, JRC, UBA) and international (WHO, CSIR – Ghana, INPE – Brazil, Yale – USA, ICMR – India) organizations.
# Contents

Statement Of Authorship ........................................... i

Abstract ...................................................................... ii

Περίληψη ...................................................................... v

Acknowledgements ..................................................... viii

Thesis Achievements .................................................. ix

List of Tables ................................................................ xii

List of Figures ............................................................ xiii

Abbreviations ............................................................. xv

1 Introduction ............................................................. 1

1.1 Problem addressed ............................................... 3

1.2 Thesis Contribution ............................................... 4

1.3 Thesis Structure ................................................... 5

2 Literature Review ..................................................... 6

2.1 Remote sensing ..................................................... 6

2.2 Risk factors of MBDs ............................................. 7

2.3 Prediction of Infectious Disease Outbreaks ............... 10

2.4 Overview ........................................................... 13

3 Cases of Study ........................................................ 16

3.1 Mosquito - Disease .............................................. 16

  3.1.1 Culex Pipiens - West Nile Virus ....................... 16

  3.1.2 Anopheles - Malaria ....................................... 20

  3.1.3 Aedes - Dengue Fever / Zika / Chikungunya .... 23

3.2 Areas of Interest .................................................. 25

  3.2.1 Veneto, Italy .................................................. 25

  3.2.2 Vojvodina, Serbia .......................................... 27

  3.2.3 Grand Est and Corsica, France ......................... 28

  3.2.4 Baden-Württemberg, Germany ....................... 30

3.3 Overview .......................................................... 31

4 Data and Preprocessing ............................................ 32

  4.1 Remote sensing and Earth Observation Data .......... 32
4.2 Dataset .......................................................... 33
  4.2.1 Entomological network ................................. 33
  4.2.2 Earth Observation data ................................. 34
4.3 Data Cleaning, Transformation and Imputation ............ 36
4.4 Overview ...................................................... 37

5 System Description and Evaluation Method .......................... 39
  5.1 MAMOTH Framework ........................................ 39
    5.1.1 MAMOTH’s Cost function ............................. 40
    5.1.2 MAMOTH’s Feature space and solver .................. 41
    5.1.3 Feature extraction / engineering Module ............ 41
    5.1.4 Initialization Module .................................. 41
    5.1.5 Hyper-Grid Module ................................. 42
    5.1.6 Features Selection ................................. 42
  5.2 Training and Testing Method ................................ 42
    5.2.1 Random 10-Fold Validation ........................... 43
    5.2.2 Operational Validation ................................ 43
  5.3 Gradient Boosting ......................................... 44
  5.4 Evaluation Method ........................................ 45
  5.5 Features extracted ......................................... 46
    5.5.1 Mosquito Population Features ......................... 46
    5.5.2 Spatiotemporal Features ............................... 46
    5.5.3 PCA Features ......................................... 47
  5.6 Overview .................................................. 48

6 Experiments and Results ................................................. 49
  6.1 Italy, Culex genus Mosquito ............................... 49
    6.1.1 Performance without the Entomological features .... 53
  6.2 Serbia, Culex genus Mosquito ............................. 54
  6.3 Germany, Culex genus Mosquito ........................... 58
  6.4 France, Aedes genus Mosquito ............................ 62
  6.5 Italy, Anopheles genus Mosquito ......................... 65
  6.6 Overview .................................................. 69

7 Conclusions ......................................................... 70
  7.1 Results Overview and Discussion ......................... 70
  7.2 Future Work .............................................. 73

A Feature List .......................................................... 75

Bibliography - References .............................................. 78
List of Tables

2.1 MBDs Risk factor review summary ........................................... 9
2.2 MBDs prediction models summary ............................................. 12
2.3 MBDs prediction models data .................................................. 13
2.4 MBDs prediction models environmental data ............................ 14
3.1 WNV cases in EU countries [64] ............................................. 18
4.1 Final dataset of each case ....................................................... 37
6.1 Culex mosquito risk classes of Veneto region ............................. 50
6.2 Culex mosquito risk classes of Vojvodina region ....................... 55
6.3 Culex mosquito risk classes of Baden Wuerttemberg region ......... 59
6.4 Aedes Mosquito eggs risk classes of the Grand-Est and Corsica regions 63
6.5 Anopheles Mosquito Risk Classes of the Veneto region ............ 66
7.1 MAMOTH’s pre-operational applications and performance per case. 71
7.2 Most important features per case. ........................................... 72
# List of Figures

1.1 Machine Learning Process ........................................... 3

3.1 WNV Transmission Cycle ............................................. 17
3.2 WNV Distribution in 2018 ............................................. 18
3.3 WNV Seasonal Distribution ............................................ 19
3.4 WNV Age Distribution ................................................ 19
3.5 Distribution of the Culex pipiens in Europe ......................... 20
3.6 Malaria Distribution .................................................... 21
3.7 Malaria cases per age .................................................. 21
3.8 Malaria Seasonal Distribution ......................................... 22
3.9 Distribution of the Anopheles in Europe ............................. 22
3.10 Dengue Distribution ................................................... 23
3.11 Malaria cases per age .................................................. 24
3.12 Malaria Seasonal Distribution ......................................... 24
3.13 Distribution of the Aedes Albopictus in Europe ................... 25
3.14 Veneto Region ........................................................... 26
3.15 Veneto mean temperatures and precipitation ....................... 26
3.16 Vojvodina Region ....................................................... 27
3.17 Vojvodina mean temperatures and precipitation .................... 27
3.18 Grand - Est Region ..................................................... 28
3.19 Grand-Est mean temperatures and precipitation .................... 28
3.20 Corsica Region .......................................................... 29
3.21 Corsica mean temperatures and precipitation ....................... 29
3.22 Baden-Württemberg Region ........................................... 30
3.23 Baden-Württemberg mean temperatures and precipitation ........ 30

4.1 Entomological monitoring network in Veneto region. ............... 33
4.2 Example of a “false colour” image to calculate the vegetation of an area .................................................. 35

5.1 MAMOTH Description pipeline ........................................ 40
5.2 Random K-Fold cross validation ....................................... 43
5.3 Operational cross validation .......................................... 43
5.4 Cyclic Transformation .................................................. 47

6.1 Entomological network in Veneto ..................................... 49
6.2 CDF of the time difference between observations .................. 50
6.3 Feature importance of Culex - Veneto ............................... 51
6.4 Error distribution of Culex - Veneto .................................. 52
6.5 Error distribution per class of Culex - Veneto ...................... 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Error distribution per month of Culex - Veneto</td>
<td>53</td>
</tr>
<tr>
<td>6.7</td>
<td>Feature importance using only environmental data</td>
<td>54</td>
</tr>
<tr>
<td>6.8</td>
<td>Entomological network in Serbia</td>
<td>55</td>
</tr>
<tr>
<td>6.9</td>
<td>CDF of the time difference between observations Vojvodina region</td>
<td>56</td>
</tr>
<tr>
<td>6.10</td>
<td>Feature importance of Culex - Serbia</td>
<td>56</td>
</tr>
<tr>
<td>6.11</td>
<td>Error distribution of Culex - Vojvodina</td>
<td>57</td>
</tr>
<tr>
<td>6.12</td>
<td>Error distribution per class of Culex - Vojvodina</td>
<td>58</td>
</tr>
<tr>
<td>6.13</td>
<td>Entomological network in Baden Wuerttember</td>
<td>59</td>
</tr>
<tr>
<td>6.14</td>
<td>CDF of the time difference between observations in Baden Wuerttember</td>
<td>60</td>
</tr>
<tr>
<td>6.15</td>
<td>Feature importance of Culex - Germany</td>
<td>60</td>
</tr>
<tr>
<td>6.16</td>
<td>Error distribution of Culex - Germany</td>
<td>61</td>
</tr>
<tr>
<td>6.17</td>
<td>Error distribution per class of Culex - Germany</td>
<td>61</td>
</tr>
<tr>
<td>6.18</td>
<td>Entomological network in Grand-Est and Corsica regions</td>
<td>62</td>
</tr>
<tr>
<td>6.19</td>
<td>CDF of the time difference between observations in the Grand-Est and Corsica regions</td>
<td>63</td>
</tr>
<tr>
<td>6.20</td>
<td>Feature importance of Aedes - France</td>
<td>64</td>
</tr>
<tr>
<td>6.21</td>
<td>Error distribution of Aedes - France</td>
<td>64</td>
</tr>
<tr>
<td>6.22</td>
<td>Error distribution per class of Aedes - France</td>
<td>65</td>
</tr>
<tr>
<td>6.23</td>
<td>Entomological network in Veneto</td>
<td>65</td>
</tr>
<tr>
<td>6.24</td>
<td>CDF of the time difference between observations - Italy</td>
<td>66</td>
</tr>
<tr>
<td>6.25</td>
<td>Feature importance of Anopheles - Veneto</td>
<td>67</td>
</tr>
<tr>
<td>6.26</td>
<td>Error distribution of Anopheles - Veneto</td>
<td>68</td>
</tr>
<tr>
<td>6.27</td>
<td>Error distribution per class of Anopheles - Veneto</td>
<td>68</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLMS</td>
<td>Copernicus Service Land Monitoring Service</td>
</tr>
<tr>
<td>DEM</td>
<td>Digital Elevation Model</td>
</tr>
<tr>
<td>ECDC</td>
<td>European Center for Disease Prevention and Control</td>
</tr>
<tr>
<td>EO</td>
<td>Earth Observation</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>EVI</td>
<td>Enhanced Vegetation Index</td>
</tr>
<tr>
<td>GAM</td>
<td>Generalized Additive Model</td>
</tr>
<tr>
<td>GSD</td>
<td>Ground Sampling Distance</td>
</tr>
<tr>
<td>IMERG</td>
<td>Integrated Multi-satellite Retrievals for GPM</td>
</tr>
<tr>
<td>Km</td>
<td>Kilometer</td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbor</td>
</tr>
<tr>
<td>LST</td>
<td>Land Surface Temperature</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAMOTH</td>
<td>Mosquitoes Abundance Prediction Model autO-calibrated from features pleTHora</td>
</tr>
<tr>
<td>MBD</td>
<td>Mosquito-Borne Diseases</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>NDBI</td>
<td>Normalized Difference Build-up Index</td>
</tr>
<tr>
<td>NDMI</td>
<td>Normalized Difference Moisture Index</td>
</tr>
<tr>
<td>NDVI</td>
<td>Normalized Difference Vegetation Index</td>
</tr>
<tr>
<td>NDWI</td>
<td>Normalized Difference Water Index</td>
</tr>
<tr>
<td>NMSE</td>
<td>Normalized Mean Square Error</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>WNV</td>
<td>West Nile Virus</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Vectors are living organisms that can transmit infectious pathogens between humans, or from animals to humans (dead-end hosts). Such vectors are mosquitoes, which are one of the deadliest animals in the world. Vector-borne diseases (VBDs) account for more than 17% of all infectious diseases, causing more than 700,000 deaths annually [59]. In Europe, according to a European Centre for Disease Prevention and Control (ECDC) review [48], VBDs are widespread and are the best studied diseases associated with climate change. Despite the fact that many of these diseases have been effectively controlled, the last two decades the emergence of global trends is driving the re-emergence of mosquito-borne diseases (MBDs) [24]. Mosquito-Borne Diseases (MBDs) are diseases spread by the bite of infected mosquitoes. Tropical MBDs, such as West Nile Virus (WNV) transmitted with Culex mosquitoes, Malaria transmitted with Anopholes mosquitoes, and Chikungunya, Dengue and Zika transmitted with Aedes mosquitoes have posed challenges to national public health authorities in the European region [6]. The changing climatic and ecological conditions, global travel and trade, human behavior, as well as the rapid and unplanned urbanization, are key trends that influence the seasonal and geographic distribution of vectors’ population and therefore the transmission of the pathogens [19].

It is a wrong belief that MBDs are a problem that only the developing countries face. Europe has experienced many cases of MBDs outbreaks in the last 10 years. 2010 was a year with large outbreaks of West Nile Virus in Greece and Russia having 262 and 419 human cases respectively and a total of 1016 cases across all Europe [49]. WNV human infections have sharply increased in 2018 as compared to the previous years. According to the ECDC [63] the most affected countries include Italy (576 cases), Serbia (415 cases) and Greece (316 cases). The total number of cases in 2018 in European Union raised up to 1503 and 124 deaths [5]. Furthermore, according to ECDC, the number of confirmed Malaria cases reported in the EU from 2008 to 2012 ranged between approximately 5000 and 7000, whereas in 2019 they reached almost 9000. The number of cases shows that there is a need for preventive actions that will help in the decrease the dissemination of MBDs which may result in death. Those preventive actions include epidemiological surveillance of the diseases, timely implementation of appropriate integrated mosquito control programmes and implementation of personal protection measures against mosquito bites indoors and outdoors [5].
Although Public Health authorities in most European countries conduct on-going epidemiological surveillance of MBDs and implement integrated mosquito control programmes, an early warning system aiming to facilitate appropriate mosquito control programmes and take targeted preventive measures is of great importance in the combat against outbreaks. A recent scoping review [19] showed that despite the plethora of techniques developed, the implementation of them in different scale or areas is not possible yet. The different spatio-temporal parameters in which under the earth observational data are collected prevent their fusion, while ethics and personal data attributes make the epidemiological data either not available or not easily accessed at the appropriate spatial resolution.

The lack of well structured, consistent and reliable environmental, landscape and ecosystem data and their change in time is difficult to collect and this is one of the most important barriers. In-situ equipment that is necessary for environmental data collection is limiting the study area either because of the high cost of operation and maintenance or the inaccessibility of an area. Different spatial-temporal resolutions and ad hoc re-sampling and filtering techniques in limited areas increase a lot the complexity of comparative studies. However, nowadays with the advent of the big Earth Observation (EO) data and the plethora of satellite sensors e.g. Sentinel, Landsat, TERRA/AQUA (MODIS), etc allow shorter revisit times and larger coverage, enabling enhanced earth monitoring at global level, providing that way vast amounts of data which are accessible via open data platforms [38]. In addition, the revolution in data science and machine learning (ML) algorithms provides many opportunities for accurate and reliable data-driven solutions to the problem [28].

Under the realization that an early warning system against MBDs would provide valuable help in minimizing the spread of those diseases a generic framework was developed for mosquito abundance prediction. Taking advantage of the capabilities that the machine learning field provides, and integrating multi source Earth Observation data (EO) in order to extract environmental and topological information in a consistent, uniform and reliable way, we focused on developing an early warning system for the upcoming mosquito population. An architecture that is independent by the area of interest and the mosquito species was developed. This generic architecture is an auto-calibrated model that does not require any human action in selection of the features.

Our approach offers a generic and accurate architecture that is replicable on different areas and mosquitoes according to predefined criteria. Our model’s inference offers insights regarding the problem and the capability of easier comparison between different cases, “which characteristics were important in one case and which on the other?”.
1.1 Problem addressed

A lot of research work is focusing on predicting the upcoming MBD danger in order to successfully design preventive actions. The state of the art can be split into two directions, those works that aim to predict the upcoming human cases (epidemiological approach) and those that predict the mosquito population (entomological approach). As can be expected, the probability of human infection and mosquito population in a given area are strongly dependent variables [52]. This connection led to the development of a system which predicts the mosquito population abundance.

Given the input data, concerning geographical, climatological and other information that will be discussed later, an estimation by a machine learning model will be given, predicting a risk class of mosquito population ranging from 1 to 10. The higher the risk class, the higher the expected number of Culex population in about 15 days from the time of the prediction. This time window of 15 days can grant authorities time to take early preventive actions, such as activation of warning systems or use of mosquito control pesticides to minimize their population.

We transformed the mosquitoes population from regression to an ordinal regression problem because it offers multiple advantages in both technical domain and the dissemination of the results. Technically, this transformation makes our model more robust to outliers, once the error that an observation can contribute is limited and easier transferable to different cases with different population distributions. The conversion of the actual mosquito population to risk classes helps in the understanding of the results e.g. “The next two weeks the model expects risk population class 8 out of 10 for this region”, is more informative compared to “The next two weeks the model predicts on average 183 Culex mosquitoes for this region”.

Thus, taking advantage of the capabilities that machine learning techniques provide, and integrating multi source open Earth Observation (EO) data in order to extract as much as possible environmental and topological information, we built an early warning system of upcoming mosquito population aiming to facilitate appropriate mosquito control programmes and take targeted preventive measures. The main characteristic of the system is that the users do not have to specify the feature space of the observations, instead, is an auto-calibrated model that receives the entire pool
Chapter 1. Introduction

of features and selects which features to use in order to build the prediction model. In that way there is no human bias introduced over the features, rather the model selects the most important ones. The vision behind the system architecture was to develop a versatile and easily replicable framework that can be applied without any special requirement to any country without any modifications. The goal is to create a generic auto-adjustable system to predict the activity of mosquitoes’ population over space and time. The model’s capacity for inference on the features allows for further analytical study on those auto selected important features.

The framework was validated with data of different areas against different genus mosquitoes (Anopheles and Culex genus mosquito in Veneto region of Italy, Culex in Vojvodina region of Serbia, Culex in Baden Wuerttemberg region of Germany and Aedes in Grand-Est and Corsica regions of France), in order to establish its reliability and accuracy, defending our claims for an consisted auto calibrated and easy transferable system used against spread of mosquito borne diseases.

1.2 Thesis Contribution

Based on the above considerations, this thesis contributes in the battle against MBDs with a framework that it is:

- Data-driven relying on open EO global data, facilitating the data collection for any area of interest. EO data extracted from satellite images are independent of in situ equipment and are widely accessible allowing the framework to be applicable almost everywhere.

- Automated - self calibrated, which prevents the injection of human bias through automation in feature selection. In each case, the framework adjusts the features keeping only the important ones from a set of features, with no human intervention throughout the process. This way the framework wraps itself around the needs of the area in relation to the mosquito genus.

- Transferable and replicable to any area regardless the mosquito genus. Due to the capability of EO data to capture information in any area and the automation in the feature selection, the framework is able to be applied to any area and any mosquito genus (transferable) without any modification (replicable).

- Accurate and reliable as it was validated upon 5 different cases. The framework performed accurately and consistently in all of them, proving the above mentioned point for transferability and replicability.

- Replicable under the same architecture and the same mathematical principles, with extensive capability of comparative studies among different cases, responding to: “which characteristics seem important in one case and which on the other?”

According to the literature study, this is the first time that a single data-driven architecture has predicted mosquito abundance in a way that unifies all the cases from MBDs and is independent to the site of application thus presenting a high rate of transferability in different landscapes and climatic zones. Up to now, the proposed
solutions against the spread of MBDs are addressing the problem in an ad-hoc way tailoring the approach upon the area or the mosquito genus. In contrast, the proposed framework is applicable to all cases regardless the area of interest or the mosquito genus. It auto-adjusts itself in the case (area - mosquito genus) by selecting the most suitable features, predicting the mosquito abundance in an informative way by transforming the mosquito population of an area to risk classes making applicable to all cases regardless the area of interest or the mosquito genus.

Furthermore our model has already been operationally applied in the region of Veneto Italy providing useful warning in order to help in the MBD spread minimization. Part of the results which will be presented for the case of the Culex mosquitoes in Veneto, Italy were also real time reported to Italian authorities, so preventive measures against WNV could be taken on time.

1.3 Thesis Structure

The rest of the thesis is organized as follows, Chapter 2 presents the related literature and the methods on which the proposed models are grounded. In Chapter 3, mosquito genus and regions of interest in which the framework was applied are presented, while Chapter 4 elaborates on the data collected, their augmentation and their prepossessing. Chapter 5 presents the system in depth along with the evaluation methods and Chapter 6 demonstrates the experiments conducted and their results. Finally, chapter 7 introduces a discussion on the overall work and future research directions.
Chapter 2

Literature Review

Before getting into the details about the framework implementation, a literature review will be given to inform the reader about the up to date approaches against MBDs. This chapter reviews some of the existing studies focusing on remote sensing data and approaches on modeling MBDs outbreaks. As a first step, studies on how remote sensing technologies can be used along machine learning techniques are presented. Another issue under the microscope is about how EO data can be used in combat against infectious diseases, focusing on examining the conditions favoring the spread of MBDs, mainly the environmental and climatological ones, while others focus on predicting upcoming outbreaks.

2.1 Remote sensing

Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance usually from images taken from satellites. Remote sensing technologies are generating vast amounts of EO data. These data can capture different aspects of environmental variables such as the temperature or vegetation of an area. Measurements about Land Surface Temperature (LST) which is a proxy for air temperature, NDVI and NDWI which are proxies for vegetation and water bodies respectively [27], and evapotranspiration a proxy for humidity [68] are some of the most used products extracted from satellite images.

Remote sensing technologies offer some benefits but limitations too. It can provide regular data of large areas without the demand of any in situ equipment, accessing in that way remote places of earth with low cost. However the combination of high spatial resolution (less than 30m) with a high temporal resolution (less than one day) is difficult. The ability of capturing data over large areas on regular basis can produce big amount of data making difficult to store and process them [38]. A compromise between spatial and temporal resolution must be made in order to balance out the drawbacks. Some applications can demand higher spatial resolutions in order to capture greater detail of the area, while other applications are in need of less detailed but more frequent measurements to capture changes in time. [31] stated that remote sensing images could act as source of data to many machine learning applications and they stated out two applications. An automatic recognition of a crop type and an automatic recognition of dust sources. Crop type classification for example is independent of
high temporal resolution as there is no need of capturing changes over time, however a high spatial resolution could provide more accurate data about the crop. The same study elaborated on a machine learning technique to identify types of dust sources, a very important task in the protection of the environment, using multi spectral satellite data and an unsupervised multivariate non linear classification process. In this case a balance between spatial and temporal resolution may be a key to success, as detail and progress over time of the dust clouds can provide useful information for their recognition. So it becomes clear that remote sensing technologies are useful for a wide range of projects as they can provide vast amounts of data in the detail or the frequency needed.

Focusing in the epidemiological field, there are numerous studies on how remote sensing data can be used to applications to monitor and detect vector borne infectious diseases. Vector list includes mosquitoes, ticks and tsetse flies. A review stating those vectors and the diseases associated with them, as well as studies and technology focusing to map the environment of those vectors and the dissemination of these infectious diseases is discussed by [46]. EO data are capable of detecting the changes in the environment, which can facilitate disease outbreaks, through different indicators as mentioned above. However infections diseases transmission is affected not only by environmental factors, but also by interaction and connectivity between different regions through the human mobility and social activity, which can be also monitored through remote sensing technologies [57]. [21] added the human constructed and natural environment parameters as contribution factors on infectious disease outbreaks. [66] created an application which collects and stores to a database historical and real time available environmental data in earth science data archives, to facilitate and integrate the surveillance and forecasting of MBDs outbreaks as it can also collect epidemiological datasets.

2.2 Risk factors of MBDs

A variety of studies have shed light on those factors contributing in outbreaks of the infections transmitted through mosquitoes bites. Above normal temperatures contribute in new cases of WNV diseases and disperse into new areas of Europe. [53] found that temperature anomalies in July is a significant factor in the pathogen circulation. The overwintering infected mosquitoes and wild bird population are the reason behind the occurring cases in the same areas. Water bodies in early June also seem to play an important role. Large areas of water bodies can increase mosquito population, however limited number of water ponds may boost interaction between gathered wild birds and mosquitoes resulting in an increased transmission of WNV. [36] also linked the increased summer temperatures to increased incidence across Europe, as high temperatures can lead to increased mosquito bites and development rates. The same study identify as risk factors the increase in spring season precipitation and the presence of irrigated corp lands, as stagnant water pools contribute to mosquito larval development.

Abundance of mosquitoes was captured during summer months, and especially in July where there was a peak in West Virgina, United States. After empirical observation analysis [8] concluded that a rapid increase of LST was linked to an increase
of infected mosquitoes. Temperature is causing a shorting in the extrinsic incubation period of the virus. [33] conducted a spatio-temporal analysis to examine the association of environmental and topological factors with mosquito abundance. Using statistical analysis methods and data from 2007 in the area of Los Angeles they found that elevation is negatively associated to mosquito abundance and infection. The higher the elevation, the lower the mosquito abundance. Temperature was found to be positively associated, the higher the temperature the higher the population of the mosquitoes and the possibility of infection. Vegetation, contrary to what was expected, showed no correlation to mosquito abundance, however it was an important factor on another study examining the risk areas in Southern California [32] during the years of 2007-2009. Lower elevation was again linked to WNV incidence, as temperature did. It was also observed that deviations from the annual mean temperature were affecting the WNV incidence while giving emphasis and in the landscape diversity.

An early warning system [43] that predicts the start and length of the mosquito season and abundance was created using ten years of entomological and environmental data from Piedmont region in Italy through a mixed linear model. After aggregating the data in weekly periods to catch seasonal trends, they concluded that warm temperatures during early period resulted in early starting and prolonged mosquito period, while warm temperatures in late period extended the length season but decrease mosquito population as they are negatively affected by the high temperatures. Furthermore high temperatures and low precipitation leads in less mosquitoes as water ponds, which acts as breeding sites, are decreasing. Precipitation in early period, shortens the mosquito season but results in greater abundance due to many available breeding sites. Late period precipitation prolongs the season as it creates and sustains the mosquito breeding sites. Higher NDWI values also affect the mosquito season by early starting, while proximity to rice fields also indicates higher mosquito population.

In order to identify the the environmental conditions contributing in WNV dissemination in Greece, researchers on [55] employed the method of co-clustering. They examined each variable by applying space-time clustering on them to identify times and areas with and without WNV human cases. The best results were obtained when clustering in 2 temporal and 3 spatial groups. The variables that seemed to separate areas better was the number of summer days, annual average of mean and maximum LST, potential number of mosquito and virus cycles and mean temperature of the WNV transmission season (June-October), while in temporal clustering only annual mean LST as well as the annual average minimum LST manage to perform a more robust clustering of WNV years.

On the other hand, statistical analysis in Morocco during the years of 2003 and 2010 in which WNV cases were recorded, showed that there was no association with higher temperature. The most significant variables identified affecting the spread of the virus was the higher NDVI values and the extreme rainfall. In contrast with other areas, temperature is of minor significance possibly because of the warmer climate that prevails in the area of Africa [11]

On study conducted in France [50] during 2007 and 2008, researchers examined the
The relationship between the wetlands and infection rates, by studying the circulation of the virus across equines. More specifically using satellite images and statistical analysis, they studied how variations in areas of open water or flooded vegetation is affecting the WNV circulation across 3 time periods (November to February which is not a risk period, March to July which is an endemic period and August to October which is an epidemic period). Their findings showed that decrease of open water between winter and spring and spring to summer led to increased infection incidence, however these findings are based on study conducted in equines and not in humans but they can be still considered as reliable findings as horses are accidental hosts of the virus, just like humans.

Following a very different approach of the studies above who are statistical orientated, dealing with WNV from a different point of view, [29] used the SEIR (susceptible, exposed, infectious, removed) epidemiological model, a tool for mathematical modeling of infectious diseases based on differential equations, to assess the factors contributing in WNV amplification at Thessaloniki, Greece in 2010. They investigated the factors affecting the virus reproduction rate, the minimum infection rate and the number of infected individuals. They adjust the SEIR model’s parameters to include environmental variables and health status of mosquitoes and birds besides the human’s ones. The results showed that mosquito birth and death rate, as well as the rate of mosquito to bird bites are factors who affect the above mentioned quantities, with temperature being the driver behind those rates. However this study introduces a level of uncertainty on the results due to the lack of systematic surveillance upon mosquitoes and birds.

### Table 2.1: MBDs Risk factor review summary

<table>
<thead>
<tr>
<th>Reference</th>
<th>Data</th>
<th>Method</th>
<th>Number of Variables</th>
<th>Target Variable</th>
<th>Risk factors</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>[53]</td>
<td>2002-2013</td>
<td>Logistic Regression</td>
<td>12</td>
<td>WNV probability</td>
<td>Temperature anomalies in July, Water bodies in early June, cases of previous year</td>
<td>AUC: 2012: 0.81 2013: 0.85</td>
</tr>
<tr>
<td>[36]</td>
<td>2010-2012</td>
<td>Linear mixed-effects model</td>
<td>10</td>
<td>incidence per 100K residents</td>
<td>Increased summer temperatures, Increased precipitation during spring</td>
<td>R: 0.52</td>
</tr>
<tr>
<td>[41]</td>
<td>2001-2011</td>
<td>Linear mixed-effects model</td>
<td>5</td>
<td>Mosquito number on each trap</td>
<td>Warm temperatures in early / late periods, Increased precipitation in early / late period</td>
<td>R²: 0.46 - 0.49</td>
</tr>
<tr>
<td>[8]</td>
<td>2003-2012</td>
<td>Linear regression model</td>
<td>3</td>
<td>Probability of occurrence of infected mosquitoes</td>
<td>Rapid increase of temperatures</td>
<td>Pearson p-value: 0.64</td>
</tr>
<tr>
<td>[33]</td>
<td>2010</td>
<td>Statistical analysis</td>
<td>5</td>
<td>Mosquito Abundance infected mosquitoes</td>
<td>Increased temperature, Low elevation</td>
<td>-</td>
</tr>
<tr>
<td>[31]</td>
<td>2002-2010</td>
<td>Statistical analysis</td>
<td>4</td>
<td>-</td>
<td>Higher NDVI values, Extreme rainfall</td>
<td>Chi-square test: 0.64</td>
</tr>
<tr>
<td>[32]</td>
<td>2007-2009</td>
<td>Statistical analysis</td>
<td>5</td>
<td>-</td>
<td>Higher NDVI values, Higher Temperatures, Deviations from the mean annual temperatures, Lower elevation</td>
<td>-</td>
</tr>
<tr>
<td>[30]</td>
<td>2007-2008</td>
<td>Statistical analysis</td>
<td>12</td>
<td>infected positive horses</td>
<td>Open water deviations between winter and spring and from spring to summer</td>
<td>AUC: 0.88</td>
</tr>
<tr>
<td>[55]</td>
<td>2013-2016</td>
<td>Co-clustering</td>
<td>23</td>
<td>-</td>
<td>Number of summer days, annual averages of maximum and average LST</td>
<td>Mean and standard deviations of the clusters</td>
</tr>
<tr>
<td>[29]</td>
<td>2010</td>
<td>SEIR epidemiological model</td>
<td>25</td>
<td>Infected mosquitoes and humans</td>
<td>Mosquito birth and death rate, Rate of mosquito to bird bites, Temperature</td>
<td>-</td>
</tr>
</tbody>
</table>

The above Table 2.1 sums up all the studies focusing on identification of the conditions that contribute in circulation of MBDs. What it can be observed is that one of the most distinctive causes is the temperature no matter the area. The next section focuses on approaches of MBDs prediction modeling.
2.3 Prediction of Infectious Disease Outbreaks

There are a lot of studies focusing on predicting MBDs incidence in an area in an attempt to prevent outbreaks. [35] built a mosquito classifier in an attempt of automated mosquito identification which is the vector of transmission of all MBDs. This system could act as an extra method towards mosquito surveillance and their habitats.

Clustering was employed in [23] to classify areas of Greece as low, medium and high risk areas depending on potential of WNV cases. Using data of 2010 and 2011 performed two step cluster analysis to areas having reported human and wild birds cases of WNV using 37 environmental factors. Analysis resulted in 3 clusters. The most significant variables of clustering were the altitude and the distance from the water. 60% of human and bird positive WNV cases were clustered together in areas of low altitude and small distance from the water. The results showed that 4 out of 5 cases in 2012 were located in predicted high risk areas after classifying each area in one of the 3 clusters of the above analysis.

Decision trees where used in [67] to predict WNV across United States of America. Using environmental data from 2484 random counties as training set, they created several models including the data of all years between 2003 and 2008 and each year individually. After testing the models on the reaming 621 counties they came to really interesting conclusions. The results showed that there is not a standard model fitting all years as each year’s conditions are different, but a global model fitting all counties is feasible, performing better than many regional ones. They found that it is more difficult to predict incidence of WNV cases in small areas and areas of lower population due to other factors that are not covered by environmental variables. [14] created a WNV incidence of general additive model forecast model for the northern Great Plains of United States where the population density is lower than the rest of the United States of America, overcoming this problem by smoothing out the differences in the rate of WNV cases between neighboring regions. Environmental factors as independent variables were used aggregated in 8 days-blocks for predicting WNV cases. 15 models were fitted, each one responding to a date from April 22th to August 12th adding in each model 8 days extra data than the previous one. The model presented as expected higher performance as the time period considered was expanded. Periods of warm temperatures and high vegetation was once again linked to increased WNV cases and by this study despite the fact it was only based on 3 different variables.

In [4] tried to deal with the problem from another perspective. Boosting regression trees and Maximum entropy model were used to predict potential habitats of Culex pipiens in areas of Egypt and Lebanon in the Middle East North Africa, using environmental variables. Special focus was given on the impact of human population of an area on mosquito population. Both models produced similar results assigning greater probabilities of Culex pipiens habitats in areas with high human population as human activities such as agriculture or living conditions can provide a great habitat for mosquitoes. Quality and seasonality of vegetation was also found to be an important factor positively linked to mosquito population.

Aiming to capture the spatial and temporal aspect of data, [37] created a model to
predict the presence or not of WNV infected mosquitoes. The model was trained using environmental data of New York from 2001 to 2015 with respect to time and location. Except of the environmental data, that they are taken account by the majority of models studied, this one includes as a feature the age of the houses in the area, as the researchers claim that houses built before 1960 gather large population of mosquitoes. The model achieved high accuracy concluding that suburban and exurban areas are having a higher risk of WNV incidence. Positive correlation was found between temperature and WNV incidence, while precipitation and NDVI values was found to be negatively correlated. Areas of high vegetation are not Culex larvae friendly as they are getting targeted by other mosquitoes species, such Aedes, that thrive in areas of high vegetation, while heavy rainfall can flush out Culex breeding sites which are mostly found in artificial water pools.

SVM and Firefly algorithm was used to predict incidence of Malaria in the area of Jodhpur and Bikaner in India using 3 years of data \[12\]. Firefly algorithm is an algorithm used for optimizing the SVM hyper parameters. The model is pretty simple predicting monthly Malaria incidence using only 3 features, the mean monthly environmental data of temperature, precipitation and evapotranspiration as well. The combination of Firefly - SVM model is what differentiates it from other models, achieving better results over a neural network model and a time series model. Although the time series model failed in the case of India, it was applied in 12 districts of Ethiopia in forecasting Malaria cases with very interesting results. \[3\] included as parameters of the model the rainfall, LST, Enhanced Vegetation Index (EVI), and evapotranspiration and find that each of the environment variables can affect cases of Malaria depending on the time lag. LST was positively associated with Malaria cases one month late, while EVI was associated with Malaria cases within 3 months. The rainfall and the effect of evaporation in cases of Malaria ranged from one to two to three months. A very interesting finding, as each variable seems to be associated with a different time delay in predicting Malaria cases.

SVM was also used by \[41\] to forecast Dengue outbreaks. Using climatological weekly data and measuring the volume of the weekly search queries relevant to Dengue they created a model to predict the weekly cases of Dengue. The use of search queries allows to capture data that are not covered from entomological or environmental factors, and provide a good indication of potential cases. Year and week was also included as features to capture seasonality of the data. The SVR model outperformed other models in prediction upon the last 12 weeks of data but also in the 35th to 46th weeks of the year which were characterized as outbreak periods and the early warning of upcoming outbreak is crucial.

General additive model with and without boosting was used in forecasting Malaria incidence in Kenya one two and three months ahead \[34\]. Environmental data along with month and time in an attempt to capture seasonality such as in the previous case, were fed in the model as random variables. The results showed improved performance of general additive model with boosting. One month ahead predictions were generally close in the number of monthly cases. The two and three months predictions ahead underestimate the Malaria incidence indicating that the Malaria incidence of a month is strongly correlated to the cases of the previous month.
A not so common approach was adopted from researchers on [25]. Using multi-criteria decision analysis, implementing a very simple but effective system, areas of Rawalpindi in Pakistan were ranked in danger zones of Malaria risk. Experts assigned a weight on each of 4 environmental variables, calculating a final score for each area depending on the environmental conditions. Air temperature and relative humidity were the main factors, as they were assigned with greater weight, strongly affecting the score of Malaria risk of an area. The score practically indicates how favoring are the conditions for mosquito sustainability. The model that is not based on some complex mathematical functions or logic, was able to compete other more complex models classifying the areas of Rawalpindi quite successfully. However the assignation of the weights of each variable is an estimation based on the knowledge of experts, and not extracted out of an established scientific way, meaning that the model may be biased towards some variables.

Another very interesting study presented in [6] is using a set of association rules as classifier. A set of rules was produced by environmental weekly data. A subset of those rules that improved the total accuracy was selected to be included in the final set of rules forming a classifier. The predicted Dengue incidence in the right part of the rule, can be classified as either low or high, setting the threshold value of high incidence as +2 standard deviation from the mean incidence value. Three classifiers were constructed, two predicting the weekly Dengue incidence for 3 and 4 weeks ahead,
and one classifier predicting Dengue incidence for a 4-week period, from the 4\textsuperscript{th} to the 7\textsuperscript{th} week ahead which also includes socio-economic variables. [47] used a powerful but simple method. The authors used a KNN model to predict the population of eggs laid by Aedes mosquitoes in traps placed in outdoors of houses in western Argentina using also environmental data as features. The model which has a very simple logic of predicting the upcoming mosquito egg population based on its nearest neighbors, outperformed more complex models, performing more accurately. It is indirect way of predicting the spread of Dengue as the higher the vector population the higher the spread of a pathogen.

Table 2.2 gives a summary of the approaches of prediction models of MBDs outbreaks. More specifically from the above mentioned, and from the overview in the table, it can be extracted that there is a variety of ML models used following different tactics regarding the prediction target. Some approaches focus on predicting MBD vector parameters, while other focus in predicting the human cases.

### 2.4 Overview

By taking a closer look to the table 2.3 it is obvious that all the efforts of building systems that try to predict MBDs incidences can be grouped into two categories. The first category which includes the majority of the models studied is using epidemiological data, trying to predict the human incidence of infection, a direct approach in the combat against MBDs. However this approach may encounter obstacles regarding privacy concerns of patients or may face an underestimate of the actual cases as only 20% will encounter symptoms and even fewer will be diagnosed with WNV for example. The second group is using entomological data to predict MBDs incidence rate. An indirect approach which predicts mosquito population, with the results to be very indicative about the MBDs cases in humans in an attempt to overcome the problems mentioned above as the more the vectors the higher the spread. A link between those categories is the use of environmental data clearly stating that they are a significant predictors either way.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Epidemiological Data</th>
<th>Environmental Data</th>
<th>Entomological Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[67]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[12]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[41]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[14]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[25]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[6]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[47]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[34]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[37]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: MBDs prediction models data
Table 2.4 shows temperature seem to play a key role in MBDs dissemination as all of the studies using temperature as a predictor and most of them pointed out that higher studies were observed before or during outbreaks. [16] focused on further studying what is causing this behavior and they concluded that larger temperatures led in shorter extrinsic incubation periods. They discovered that mosquitoes kept in 30°C showed a minimum extrinsic incubation time of 4 days, with 90% of mosquitoes being able to spread the disease further in just after 12 days of their exposure to the virus. Opposed to that, only 30% of the mosquitoes kept in 18°C was considered able to spread the disease even after 32 days of their exposure to WNV. Based on what mentioned above, temperature is the main culprit behind increased fecundity of mosquitoes, increased mosquito bites and development, shorter extrinsic incubation period of WNV in mosquitoes and disperse of the virus in new areas.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Temperature</th>
<th>NDVI</th>
<th>NDWI</th>
<th>Precipitation</th>
<th>Evapotranspiration</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[67]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[41]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[14]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[25]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[6]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[47]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[34]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[37]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4: MBDs prediction models environmental data

Except from temperature, precipitation is also a driver factor in outbreaks and it is used in the majority of the predictive models. Based on the studies above precipitation can have both positive and negative effect on the mosquitoes and therefore on the dissemination of the MBDs. Extensive precipitation can flush out the breeding sites of the mosquitoes and reduce their future population and therefore the WNV circulation [37]. On the other hand drying conditions could increase the interaction of wild birds and mosquitoes around the remaining water pools, resulting in infections amplification [53].

Regarding the NDVI, a proxy for the vegetation of an area, the findings are contradictory. Some studies claim that increased NDVI values are positively correlated with mosquito population, while other studies indicate the opposite. It is expected based on our intuition that the higher the vegetation the higher the mosquito population. However there is a claim that in areas of high vegetation there are predators of Culex larvae resulting in a shorting of their population [37]. The evapotranspiration in all of the studies included showed that was positively correlated with the amplification of MBDs.

Except those basic driver environmental factors, the studies showed that the Culex larvae thrive in artificial water pools and water containers [36, 37], usually found
where there is human civilization and specially in areas with poorer population. Also the landscape can be a contributing factor as rural areas facilitate the interaction between birds and mosquitoes and then mosquitoes and humans [32, 36]. However a conclusion drawn from all the above is that spatiality is a parameter is that must be taken into consideration in all of the studies, as in each area are prevailing different conditions, leading in deviations from the general factors that are commonly observed. Moreover in all of the cases studied, limited selected environmental data were included, such as temperature and precipitation which were used as predictors along with other kinds of features depending on the case study. No extensive use of environmental features was observed throughout this literature review, sticking to the known basic ones.

Finally, each work presents a specific model or an architecture that focuses on a specific mosquito disease or area of interest, so all of those approaches are not directly comparable. Each conclusion drawn for each area and disease is resulting out of different method, thus there is no common ground for comparison among the factors affecting a MBD.

Based on the above, MAMOTH is filling the gap of a framework able to be applied in any case. It is based on environmental and entomological data predicting the mosquito population in accordance with Table 2.3, which could be consider as proxy for the MBDs human cases. A main difference between the approaches studied and the MAMOTH framework is that the features used from MAMOTH in each case are not fixed, rather they are auto - adjusted. It selects the most important features out of a greater set containing as many environmental variables possible, enabling to wrap itself around each case, delivering predictions no matter the area or the mosquito genus. This approach allows for comparison against the factors favoring a MBD outbreak through increase in the mosquito population, as the features selected in each case are based on the same architecture. The 5 cases that the framework was applied to support the claims of an auto adjusted and easy transferable and replicable framework from the Section 1.2 are given in the next Chapter.
Chapter 3

Cases of Study

This chapter reviews the cases of study that will explore in the rest of the Thesis. Each case is characterized by two main components i) the area of interest and ii) The mosquito genus with the corresponding disease.

3.1 Mosquito - Disease

3.1.1 Culex Pipiens - West Nile Virus

West Nile Virus is a MBD which is found in many counties around the world. WNV human infection cases have been reported in many countries of the world for over 50 years. It is a single-strand RNA virus of the genus Flavivirus, family Flaviviridae, which includes several other human pathogens such as dengue, Japanese encephalitis and yellow fever viruses [65].

Although the majority of human infection cases, around 80%, will not experience any symptoms the rest 20% of infected humans will go through minor symptoms including fever, headache, and tiredness. There is a minority of cases, less than 1% that will develop a severe illness affecting the central nervous system, like meningitis, and may lead to death. Serious illness can occur in people of any age, however people over the age of 50 and some immunocompromised persons are at highest risk for getting severely ill when infected with WNV. There is no vaccine available for humans yet, and only supportive treatment is available for those who will experience severe symptoms. The incubation period of the virus is calculated between to 3 and 14 days [65].

In nature the transmission cycle includes birds and mosquitoes. Birds are the reservoir hosts, harboring the pathogen and serving as a source of infection, while mosquitoes are the principal vectors contributing in the WNV circulation. Humans and horses are accidental and “dead-end” hosts, meaning that they can not spread the virus further to other humans despite being infected by the bites of the mosquitoes. So the main source of human infections are the mosquitoes and more specifically the genus Culex Pipiens which are able to carry and maintain the WNV through vertical transmission to their eggs [65].
The first human infection was discovered in Uganda in 1937 but it was until 1997 that it was considered as pathogenic, followed by a massive outbreak in 1999 at United States of America that lasted nearly 10 years [65]. Since then various large outbreaks have occurred in many areas such as Greece, Israel and Russia WNV appears to be expanding its geographical range in Europe and in the rest of the world, causing increasing number of outbreaks and escalating public health concern [44]. In Europe, human cases of WNV have been notified in almost all Eastern, Central, and Southern European countries [44] with hotspots in Italy since 2008 [9] and Greece since 2010 [39]. 2010 was a year with large outbreaks in Greece and Russia having 262 and 419 human cases respectively and a total of 1016 cases across all Europe [49]. Nowadays thousands of cases are reported every year, with most occurring in summer periods and more specifically August and September highlighting the need for preventive measures against the virus [63].

Table 3.1 shows all cases in countries of European Union (EU) by year. The number of cases is relative stable with an exception in the year of 2018 with cases rising up to 1503 and 124 deaths with Italy reporting one third of the cases. The increased number of cases is probably due to the unusual high temperatures in early spring and summer of that year [2].
Chapter 3. Cases of Study

<table>
<thead>
<tr>
<th>Year</th>
<th>EU</th>
<th>Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>128</td>
<td>14</td>
</tr>
<tr>
<td>2012</td>
<td>242</td>
<td>50</td>
</tr>
<tr>
<td>2013</td>
<td>228</td>
<td>69</td>
</tr>
<tr>
<td>2014</td>
<td>74</td>
<td>24</td>
</tr>
<tr>
<td>2015</td>
<td>108</td>
<td>60</td>
</tr>
<tr>
<td>2016</td>
<td>225</td>
<td>76</td>
</tr>
<tr>
<td>2017</td>
<td>204</td>
<td>57</td>
</tr>
<tr>
<td>2018</td>
<td>1503</td>
<td>576</td>
</tr>
<tr>
<td>2019</td>
<td>410</td>
<td>53</td>
</tr>
<tr>
<td>2020</td>
<td>308</td>
<td>66</td>
</tr>
</tbody>
</table>

Table 3.1: WNV cases in EU countries [64]

So far analyses revealed two different lineages of WNV strains. The WNV lineage 2 strain is behind the several outbreaks across Europe from Hungary in 2004 to Greece in 2010 and later in Italy, which previously had been detected only in birds [45].

The figure 3.2 shows that 2018 Italy was heavily struck by the WNV, with Hungary Greece and Romania reporting also a big number of cases. Also it can be perceived that countries of southern Europe reported many cases while countries of northern Europe reported zero cases. That may be a result of the warmer climate that prevails in those areas.

In the figure 3.3 below the distribution of WNV cases per month is presented, showing that cases are spread through the months of May to November. While throughout the years of 2014-2017 there was no significant deviation from the mean number of cases
across Europe, the year of 2018 was a complete different case as a stepper increase in the number of cases was observed, peaking in August, far above the mean number of cases in the previous years following although the seasonal patterns.

![Figure 3.3: WNV Seasonal Distribution [63]](image)

In the last figure 3.4 the rate of infections of year 2018 over different age groups is presented. The infections rates of men are higher than those of women in almost all group ages. Also the rates are increasing as the age goes up, peaking in the group age of over 65 years old, in which the biggest difference between the infections rates of men and women is observed.

![Figure 3.4: WNV Age Distribution [63]](image)

Due to absence of a vaccine, the only way to reduce infection in people is by raising awareness of the risk factors and developing a surveillance system capable of producing a warning when there are favorable conditions of WNV transmission, such as when there is expected a highly number of mosquitoes in an area.
Chapter 3. Cases of Study

Figure 3.5: Distribution of the Culex pipiens in Europe [15]

As pointed out in the figure 3.5 in Europe the Culex mosquito genus is thriving in many areas, as it can be found in almost all countries.

3.1.2 Anopheles - Malaria

Malaria is transmitted by the bites of Anopheles genus mosquitoes. It is a potentially life threatening disease caused by Plasmodium parasites however totally preventable and curable. Malaria is an acute febrile illness. Symptoms usually appear 10–15 days after the infective mosquito bite and include fever, headache, and chills and may be mild and difficult to recognize as Malaria. If not treated in time, it can progress to severe illness, often leading to death [61].

It is typically diagnosed by the microscopic examination of blood using blood films, or with antigen-based rapid diagnostic tests. The best available treatment, particularly for Malaria, is a combination of antimalarial medications that includes artemisinin. Up to date, the only vaccine available shows that it can significantly reduce malaria, and life-threatening severe malaria, only in young African children where it has been applied so far. Efforts to develop more effective vaccines are ongoing [61].

Only in 2019 the number of cases around the world was estimated to be 229 million while deaths were estimated to have reached the number of 409 thousands. Children of under 5 years old are in greater danger, as they accounted for the 65% of deaths, while Africa region was home to the 94% of cases and deaths. Europe has also been struck by Malaria outbreaks. In 2018, 8349 cases were confirmed in total, while 2840 were reported only in France [18]. The cases across the Europe in 2018 are presented in Figure 3.6, where it can be observed that the majority of cases were reported in France.
Males in total, are much more affected by Malaria across all age groups in Europe as figure shows 3.7.

Months from May to November are more prone to Malaria outbreaks as it is presented in Figure 3.8. 2018 cases did not deviate from the pattern with the peak of cases to be reported in September.
As mentioned above Anopheles genus mosquitoes are able to transmit the pathogen that causes Malaria. Anopheles mosquitoes lay their eggs in water, which hatch into larvae, eventually emerging as adult mosquitoes. Transmission also depends on climatic conditions that may affect the number and survival of mosquitoes, such as rainfall patterns, temperature and humidity. Transmission is more intense in places where the mosquito lifespan is longer (so that the parasite has time to complete its development inside the mosquito) and where it prefers to bite humans rather than other animals. So vector control population is one of the main measures that must be taken to mitigate the spread of the disease along with removal of stagnant water and personal protective measures against mosquitoes [61].
3.1.3 Aedes - Dengue Fever / Zika/ Chikungunya

Dengue fever is also a MBD, spread through infectious bites of Aedes Albopictus genus mosquitoes. The virus has five serotypes, thus infection with one type usually gives lifelong immunity to that one type. Symptoms typically begin three to fourteen days after infection, while recover is spanning from 2 to 7 days. The symptoms include high fever, headache, vomiting, muscle and joint pains, and a characteristic skin rash. In some rare cases some people will develop severe Dengue, which can be any number of complications associated with severe bleeding, organ impairment and/or plasma leakage. Severe Dengue has a higher risk of death when not managed appropriately [60].

A number of tests are available to confirm the diagnosis including detecting antibodies to the virus or its RNA. However there is not a specific treatment Fever reducers and pain killers are employed to treat flu-like symptoms. A vaccine for Dengue fever has been approved and is commercially available in a number of countries and it is recommended to people who have been previously infected, as they are in greater risk of severe Dengue in case of re-infection from another serotype of the virus [60].

Dengue virus is frequently transported from one place to another by infected travellers, when susceptible vectors are present in these new areas, there is the potential for local transmission to be established. It is estimated that infections are ranging between 100 to 400 million every year. In 2018, 2191 cases of Dengue were reported in Europe with the most of them occurring in Germany with 613 cases [60]. Figure 3.10 presents the cases around in Europe.

Contrary to the cases of WNV or Malaria, there is not a pattern related to the sex of the infected population, however from Figure 3.11 it can been observed that young and middle-aged adults are in higher risk.
Months from July to November are presenting with an increase in reported cases, however the overall pattern is more smooth as it can be seen in Figure 3.12 compared to WNV and Malaria where spikes indicated rapid increase of cases.

Aedes Albopictus mosquitoes are the vectors of the Dengue Virus. It has spread in more than 25 countries in the European Region, largely due to the international trade in used tyres (a breeding habitat) and other goods (e.g. lucky bamboo). Aedes Albopictus is highly adaptive. Its geographical spread is largely due to its tolerance of colder conditions, as an egg and adult. Transmission can occur from infected mosquito to humans but also from an infected human to a mosquito.
Figure 3.13 shows the Aedes Albopictus population around the Europe. Aedes is well established in Northern regions of Europe, while it is introduced in many regions of central Europe. Aedes is also vector to two other MBDs, Zika and Chikungunya. However both of these infectious diseases are not reporting many cases and will not be analyzed further as there are not spread out and thus do not pose any danger to public health.

### 3.2 Areas of Interest

#### 3.2.1 Veneto, Italy

Veneto region is located in north eastern Italy and the region’s capital is Venice. It is consisting of 7 provinces and 526 municipalities with nearly a population of 5 million, while 61% belongs to the group age of 18-64 [54].
The climate in Veneto can be characterized as more continental oriented than Mediterranean. It is governed by cold winters and hot summers. The average max temperature is about 30°C Celsius, while the average minimum temperature is about 0°C Celsius. The hottest months are July and August. There is a higher chance of precipitation during the months of April, October, December which peaks in November. Figure 3.15, show the weather conditions throughout the year [56]. The values have been calculated on data collection upon the last 30 years. Dashed lines show the average of the hottest day and coldest night temperature of each month.
3.2.2 Vojvodina, Serbia

Vojvodina is an autonomous province of Serbia, located in the northern part of the country. Novi Sad is the largest city and administrative center of Vojvodina and the second-largest city in Serbia. Vojvodina has a population of almost 2 million.

![Vojvodina Region, Serbia](image)

Figure 3.16: Vojvodina Region, Serbia

Vojvodina has a temperate climate which is characterized by four distinct seasons. The average annual air temperature in the city is 11 degrees Celsius. Summer months (June-September) have an average temperature above 20 °C. July is the month with highest precipitation. Figure 3.17, show the weather conditions throughout the year [56]. The values have been calculated on data collection upon the last 30 years.

![Veneto mean temperatures and precipitation](image)

Figure 3.17: Veneto mean temperatures and precipitation
Dashed lines show the average of the hottest day and coldest night temperature of each month.

### 3.2.3 Grand Est and Corsica, France

Grand - Est is an administrative region in Northeastern France and it superseded three former administrative regions, Alsace, Champagne-Ardenne and Lorraine. Population of the region is nearly 6 million inhabitants.

![Grand-Est Region, France](image1)

**Figure 3.18: Grand - Est Region, France**

![Grand-Est mean temperatures and precipitation](image2)

**Figure 3.19: Grand-Est mean temperatures and precipitation**

The climate in Grand - Est is characterized by cold winters with frequent days below the freezing point, and hot summers, with many days with temperatures up to 32°Celsius. Throughout the year there are no significant fluctuations on precipitation,
except in December where there is an increase. Figure 3.19, show the weather conditions throughout the year [56]. The values have been calculated on data collection upon the last 30 years. Dashed lines show the average of the hottest day and coldest night temperature of each month.

Corsica is a Mediterranean island and a territorial collectively of France, and a population of around 330 thousand people.

The island of Corsica has a typical Mediterranean climate with hot summers and humid winters. Throughout the year there are significant fluctuations on precipitation, with November having the highest precipitation and July the lowest. Figure 3.21, show the weather conditions throughout the year [56]. The values have been calculated on data collection upon the last 30 years. Dashed lines show the average of the hottest day and coldest night temperature of each month.
3.2.4 Baden-Württemberg, Germany

Baden-Württemberg is located in southwest Germany, east of the Rhine, which forms the southern part of Germany’s western border with France and it has a population of 11 million people.

![Figure 3.22: Baden-Württemberg, France](image)

![Figure 3.23: Baden-Württemberg mean temperatures and precipitation](image)

Baden-Württemberg with an average daily high temperature of 15 degrees Celsius. The climate offers varied seasons with deep winters and warm summers. Figure 3.23, show the weather conditions throughout the year [56]. The values have been calculated on data collection upon the last 30 years. Dashed lines show the average of the hottest day and coldest night temperature of each month.
3.3 Overview

This Chapter presented some basic information about the MBDs and the areas in which MAMOTH framework was applies to give the necessary content in order for the next chapter to introduce the earth observational and entomological extracted data upon these areas.

Our cases (5 in total) include three different mosquito genus and four different areas. A comparison of the same mosquito (Culex Pipiens) in three different areas can be performed, as well as a comparison between two different mosquitoes in the same area of interest. The model was applied for the Culex genus mosquitoes in the Region of Veneto (Italy), Vojvodina (Serbia) and Baden Wuerttemberg (Germany) and further to Anopheles in Veneto (Italy) and Aedes Albopictus Grand-Est and Corsica (France).
Chapter 4

Data and Preprocessing

The system is based upon environmental variables (geographical, climatological, and hydrological) derived from open and free satellite imagery for the period from 2010 to 2020. These variables were combined with entomological data in order to create the final dataset.

4.1 Remote sensing and Earth Observation Data

Information obtained via a remote sensing technologies is known as Earth Observation (EO) data. Earth observation is used to monitor and assess the status and changes in natural and built environments. Using remote sensing platforms such as satellites along with ground-based equipment when is possible, earth’s physical, chemical, and biological systems’ information can be obtained. EO data are vitally important because they can observe vast areas of the earth that are difficult to access even remote ones, without the use of any in situ equipment but with the use of satellites and processing of images, providing that way vast amounts of data which are accessible via open data platforms [38]. This data are usually in the form of digital images and they can be divided into two main categories.

- **Passive Imagery**: Sensors are designed to detect electromagnetic emissions from constituents of the Earth’s surface and atmosphere. Those emissions could be produced from the vegetation in the earth’s surface for example or it could be a result from the sunlight reflections. These kind of images may suffer from low quality because of weather conditions such as clouds or the absence of solar illumination.

- **Active Imagery**: A transmitter sends out electromagnetic signal of a specific wavelength and a sensor receives the result of this signal after bouncing on the earth’s surface. Those kind of systems are not weather conditions dependent as the Passive Imagery systems.

Those two categories can be divided further having its own advantages and limitations [20].

EO data are used in many useful applications of the modern world and are able to capture the geographical and the climatological conditions of a region. Land cover, weather and climatological condition monitoring, disaster management and agriculture are some indicative examples [30]. EO data can also be used for epidemiological
applications. They are able useful parameters which affect the transmission cycle of a virus. Variables such as rainfall, temperature and humidity affect the number and survival rate of mosquitoes and other vectors of diseases as seen on chapter 2. Satellite images are able to capture those variables, making them a great source of data.

4.2 Dataset

The first step to towards a development of a data driven system that predicts the population of the mosquitoes, is the collection of the data. There are two independent main sources for gathering the data. The entomological network out of which the mosquito population data is being extracted and satellites out of which the earth observational data are being extracted. These two main sources are used as data collection sources in all cases regardless the area or the virus of interest. Those two heterogeneous sets of data created must being harmonized in order to create one homogeneous dataset.

4.2.1 Entomological network

The entomological network set in each area is responsible for giving measurements upon mosquito population, so it is important to be collected with systematic way. CDC-CO₂ and light mosquito traps were utilized collect mosquito populations. Random mosquito pools out of the trapped mosquitoes went through laboratorial test to confirm existence of the virus. The traps are set from May to October and the revisit time to collect the mosquitoes trapped in them is set to 15 days with some deviations in some cases.

Figure 4.1: The entomological monitoring network of 140 traps of the Culex pipien in the Veneto region.
4.2.2 Earth Observation data

Various proxies were used in this study obtained from the Landsat-7 and 8 (30m GSD 16-day repeat cycle) and the Sentinel 2 satellites (10m GSD, 6-days revisit time) proxy of vegetation (Normalized Difference Vegetation Index - NDVI), proxy of changes in water content (Normalized Difference Water index - NDWI), proxy for the determination if vegetation water content (Normalized Difference Moisture index NDMI), for mapping urban built-up areas (Normalized Difference Build-up Difference Index). These indices have showed strong correlation with the mosquito’s behavior and biological cycle.

Temperature influences several processes that are related to the mosquito, the rate of the virus development inside the vector is linked to warmer temperatures [16]. The MODIS sensor of TERRA and AQUA has been used to estimate the Land Surface Temperature (LST) that is estimated from Top-of-Atmosphere brightness temperatures from the infrared bands of the satellite’s sensors. The product that has been integrated into the model is the MOD11A1, providing daily LST day and night and emissivity values with 1 kilometer (km) spatial resolution in a 1,200 by 1,200 km.

Rainfall can have a positive effect on larval carrying capacity but can negatively impact the reproductive cycle of the Culex pipiens by interrupting it through flushing out the aquatic stages from breeding sites [37]. The Integrated Multi-satellitE Retrievals for GPM (IMERG) gridded product with a resolution of 0.1° x 0.1° has been utilized for daily precipitation retrieval at each trap’s date of placement. The algorithm returns the cumulative rainfall of one week, two weeks and since January of the current year counting backwards from the date of placement.

High wind speed decreases the existence of infected mosquitoes, the opportunities for blood meals, and consequently human infections [51]. The ERA-5 Land Search Results Numerical Weather Prediction product has been utilized with a spatial resolution of 0.1° x 0.1°. The raw data represent the hourly u and v components indicating the direction of the wind. The daily magnitude of the vector has been computed, resulting in daily minimum, maximum and mean dominant wind in direction.

Topography has been indicated as a significant factor in the transmission of MBDs while it also affects the living conditions of several mosquito species and indicates the best suited breeding sites [19]. In order to model the topography of the study areas a Digital Elevation Model (DEM) was utilized and derivative variables like Slope and Aspect were extracted. The aforementioned products were collected from Copernicus Service Land Monitoring Service (CLMS) offered in a spatial resolution of 25 meters. For each trap station the mean elevation, slope and aspect were calculated within a buffer zone of 1 Km around each the trap. The buffer radius was determined upon the flight range of the most prominent mosquito species [13].

As mentioned before the raw EO data are in the form of images taken from satellites. They need thorough processing, to extract the values of different environmental parameters. Filters that combine various wavelength bands depending on the information to be extracted, are applied on images allowing to measure the amount of electromagnetic radiation reflected or emitted by the earth’s surface. An example of...
Figure 4.2: An example of a “false colour” image to calculate the vegetation of an area: Figure 4.2a is the original image and Figure 4.2b is the image after various wavelength bands are applied. Healthy vegetation creates chlorophyll which reflects near-infrared energy, and therefore appears in darker red on the image [20].

The extracted features from each image can be divided in three categories: satellite data, meteorological data and topological data.

- **Satellite data:**
  1. Normalized Difference indexes, $[-1, 1]$, for water, vegetation, moisture and build up area on the exact pixel of each mosquito trap location. Mean value and standard deviation are also calculated based on the neighboring pixels ($3 \times 3$ window) for these parameters.
  2. The accumulated precipitation from 1, 2 weeks before the day of measurement and from the 1st of January.

- **Meteorological data:**
  1. Minimum, maximum and mean wind magnitude daily calculated based on the hourly $u$ and $v$ components indicating the direction of the wind.

- **Topological data:**
  1. Mean Distance of sampling/trapping site within a buffer of 1000m from coastline
  2. Mean slope, within a buffer of 1000m around trapping sites
  3. Mean aspect, within a buffer of 1000m around trapping sites
  4. Mean elevation, within a buffer of 1000m around trapping sites
  5. Mean flow accumulation within a buffer of 1000m around trapping sites
6. Combination of breeding site length and water course of national hydrological data within a buffer zone of 1000m around each trapping site

7. Distance of combination of breeding site length and length of watercourses of national hydrological data within a buffer zone of 1000m around each trapping site

An extensive catalogue of the features can be found in Appendix chapter A.

4.3 Data Cleaning, Transformation and Imputation

The EO data have been pre-processed using geospatial python libraries (gdal, opencv, geopandas, rasterio) and have been resampled using bicubic/bilinear methods to 500m pixel resolution. To deal with the diverse revisit time of the satellites, the data have been temporally re-sampled following the biweekly circle of the entomological collection, by choosing the last available record.

Satellite images of the areas of each trapping site in the date of placement is searched in order to extract the corresponding features. However there are cases that are no satellite images available for this date or the feature extraction out of an image is unable due to cloud coverage. In those cases the closest to that date image is retrieved to extract the features, setting a threshold for the time difference between the date of placement and the date of the available image. This threshold is set to one week for the LST products and to one month for the indices products. If no data were found during this time window, the value was assigned as missing value.

Final datasets, formed after the integration of multi source data, suffered from vulnerabilities that had to be tackled. Duplicates of records were removed while a normalization from -1 to 1 was applied to the indexers, to ensure that all indexers will be treated equally from the learner.

Missing values were replaced by estimated values using an iterative imputing method.
Iterative imputation is a repeating process of predicting each missing value as a function based on all other features in a round-robin fashion, see Algorithm 1.

Algorithm 1: Data Iterative Imputation [42]

1. Data is separated into 2 sets. The first set includes all the rows that do not have any missing value. The second set includes all the rows with at least one missing value.
2. In the first iteration, for each column in the second set that has missing values, a regression model is constructed using the first set, where dependent variable is the column with the missing data, imputing in that way the missing values in set 2.
3. The missing values from the set 2 are getting imputed and the two sets are getting merged.
4. At the second iteration, new regression functions are defined for each column with missing values using the merged set to predict the missing values from the set 2.
5. if \( \frac{1}{n} \sum_{i=1}^{n} |v_{k}^{i} - \overline{v}_{k}^{i-1}| < \delta_{min} \) then
   stop;
else
   repeat process from the step 3;
end

where \( v_{k}^{i} \) is the \( i^{th} \) imputed missing value in the set 2 after iteration \( k \),

The predictions of the missing values in that case were based on a Bayesian Ridge Regression model.

<table>
<thead>
<tr>
<th>Area of interest - Mosquito</th>
<th>Year</th>
<th># of traps</th>
<th># of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy - Culex</td>
<td>2010 - 2020</td>
<td>140</td>
<td>4840</td>
</tr>
<tr>
<td>Serbia - Culex</td>
<td>2010 - 2019</td>
<td>124</td>
<td>926</td>
</tr>
<tr>
<td>Germany - Culex</td>
<td>2010 - 2019</td>
<td>86</td>
<td>3763</td>
</tr>
<tr>
<td>France - Aedes</td>
<td>2017 - 2019</td>
<td>81</td>
<td>1729</td>
</tr>
<tr>
<td>Italy - Anopheles</td>
<td>2010 - 2020</td>
<td>130</td>
<td>629</td>
</tr>
</tbody>
</table>

Table 4.1: Final dataset of each case

4.4 Overview

The final datasets formed along with some basic characteristics are given in Table 4.1. While in most of the cases of Culex mosquito in Italy and Germany the data collection period is quite extensive spanning to 9-10 years and the datasets present plethora of observations, in the case of Aedes mosquitoes the dataset collection was limited in 2 years. However those two years the collection of the data was consisted, due to the fact that the dataset is consisted of 1729 observations. Exactly the opposite of what happened in Anopheles case in Italy and Culex case in Serbia. Despite the fact that the collections are spreading from 2010 to 2019 - 2020, the datasets contains only 629 observations and 926 observations respectively, leading to the conclusion that there was no consistency in the way frequency that the data was collected.
Study cases (5 in total) include three different mosquito genus and four different areas. The particular datasets enable for a comparison of the same mosquito (Culex Pipiens) in three different areas can be performed, as well as a comparison between two different mosquitoes in the same area of interest. The model was applied for the Culex genus mosquitoes in the Region of Veneto (Italy), Vojvodina (Serbia) and Baden Wuerttemberg (Germany) and further to Anopheles in Veneto (Italy) and Aedes Albopictus Grand-Est and Corsica (France).

Furthermore, as it can been seen from the figures in the Chapter 3, which provides some basic details about the areas of interest, the areas are presenting a wide range of geographical distribution. Corsica is an island, while Veneto is a coastal area both of them locating in South Europe. In contrast, Grand-Est, Vojvodina and Baden-Württemberg are more central Europe located with no immediate access to the sea. This geographical distribution shows that the datasets are covering areas with different climatological conditions showing that MAMOTH can be applied in any area.

Those datasets were given as input to the framework analytically described in the next Chapter, while the results are discussed in Chapter 6.
Chapter 5

System Description and Evaluation
Method

In a supervised ML process, we assume an initial dataset X that is composed by a number of observations (rows) and a number of features (columns) named as feature-space. For each observation corresponds a label/target variable $y$ that should be estimated $\hat{y}$ from the ML model $f(\cdot)$ by observing the input information X and a set of learnable parameters $\vartheta$.

Machine learning field provides a lot of useful tools for the estimation of the mosquito population. So in this chapter the method of training, testing and evaluation will be described. There were several models developed before ending up in final model.

In this chapter, we present our framework named MAMOTH (Mosquitoes Abundance Prediction Model autO-calibrated from features pleTHora). As have been mentioned in the Introduction, MAMOTH’s main characteristic is that the user does not have to specify the feature space of the observations or models hyper-parameters. Instead, is an auto-calibrated model that receives the dataset in the initial feature space and self-tunes its hyper-parameters, as well as decides which features to use in order to build the prediction model. The vision behind the system’s architecture was to develop a versatile and easily replicated framework that can be applied without any special requirement to any area of interest.

5.1 MAMOTH Framework

The aforementioned approach raises three fundamental modeling questions that should be specified: i) Cost function - What we aim to solve? ii) Feature space selection - Which representation of the input is suitable for the optimization process? iii) Solver - How are we going to solve the optimization problem? That this section is going to discuss.

However, before describing the method used to approach the problem of mosquito population estimation, there is one very important issue. A model could be trained on predicting the exact number of mosquito population, but this kind of prediction may not communicate the seriousness of the situation. We transform mosquito populations from a regression to an ordinal classification problem, that offers multiple advantages in both technical domain and dissemination of the results. Technically,
this transformation makes our model more robust to outliers, once the error that an observation can contribute is limited. In terms of dissemination, it helps the better understanding of the results e.g. “the next two weeks model expects risk population class 8 out of 10 for this region”, is more informative compared to “the next two weeks the model predicts 183 Culex mosquitoes for this region”.

![Figure 5.1: MAMOTH Description pipeline](image)

Description of MAMOTH’s pipeline: As depicted in Figure 5.1 system’s architecture is consisted by 4 main modulus i) Feature Expansion / Engineering ii) Initiiaizer: a pre-process unit iii) Parameters Grid iv) Feature Selection. The main advantage of this architecture is that even if the final model is complex, each module, separately, is simple and its functionality is quite intuitive. This advantage is crucial for the implementation and the further evolution of the model.

### 5.1.1 MAMOTH’s Cost function

As already mentioned, the goal of MAMOTH is to estimate accurately the risk class of the mosquitoes’ populations (abundances). The prediction of mosquitoes’ populations can be valuable for a number of applications of ecosystem monitoring.

According to the aforementioned, the cost function that our model aims to minimize is the mean absolute error (MAE) between the real risk population classes and the predicted ones. Worth to be mentioned that the results obtained with mean absolute error criterion, for the cases that we studied so far, are similar with the results obtained with the mean square error criterion. Because of the mean square error criterion analytical properties, the training of the model is computationally much lighter than with mean absolute error criterion, so can be used if we need fast re-train of the models.

A reason behind using the MAE as cost function is that demonstrates higher tolerance against outliers. In this case as outliers should be consider cases of unexpectedly low or high population because of a factor that our features are not capturing. This may result in high deviations between the actual and predicted class. Mean Square Error (MSE) would assign greater weight to these errors and thus would prioritize to minimize them. However because such cases are highly unlikely to happen more than
once, we think of that cases as corrupted data. So the use of MAE as cost function makes the model more robust to these cases.

### 5.1.2 MAMOTH’s Feature space and solver

The extended feature space is received, and MAMOTH automatically decides the proper number of features and the which feature to use for every specific case (different mosquito or different area). The solver of the model is relying on gradient boosting machine learning technique for regression. Gradient boosting machines belong to a very powerful and popular family of machine learning techniques, the ensemble techniques that combine numerous weak learners in order to produce a powerful learner. A more extensive description of the gradient boosting algorithm will be given in Section 5.3. All the hyperparameters of the solver as well as the selection of the feature space are specified by MAMOTH as can be seen in the pipeline of Figure 5.1.

### 5.1.3 Feature extraction / engineering Module

The information that is already included in the dataset described in Chapter 4 can be used/restructured in order to generate new features that are informative regarding the target variable in a more algorithmic “friendly” way. This process requires a strong understanding of the physical problem and good knowledge of the related work in order to guide us to valuable features for the machine learning algorithms. This process includes various operations on the feature space such as i) non-linear transformations ii) linear and nonlinear combinations iii) time and space sifts iv) moving averages v) variables related with spatial clustering of the data vi) strong components of PCA vii) thresholds on variables. An more extensive description of these features is give in Section 5.5. The goal is to provide a more extended pool of features to the next modules that might be useful for the prediction. Should be mentioned that if there is evidence that a feature that is not included in the aforementioned list is informative regarding the prediction, can be added in the extended pool of features and the model will decide if it will use it or not.

### 5.1.4 Initialization Module

Takes as input the training set and two parameters in order to start the training process i) Number of risk classes: calculate the range of each risk population class and do balance handling if needed. The range of each risk class is selected in order all the classes to have equal probability of selection. In this paper we set up the number of risk population classes equal to 10 ii) Target set: the optimal time distance for prediction according to the training set is selected or proposed to the user e.g. predict the mosquitoes population for the next 15 days or 30 days. In order to decide the optimal time distance for the target setting, a cdf of the time distance of days between two consecutive observations was created and the minimum time distance that captures the as large as possible part of the data-set is selected.
5.1.5 Hyper-Grid Module

Hyper parameters tuning of the model uses only the most correlated features, based on Pearson correlation with the target variable in order to take an estimation of the hyper parameters (max_depth and number of estimators).

Pearson’s correlation is a popular method for calculating the importance of a feature in function with the target variable. It measures the strength of the association between the two variables. It was first introduced by the mathematician Karl Parson. Pearson’s correlation coefficient between two variables is described as the covariance of the two variables divided by the product of their standard deviations. The mathematical formula is the following:

\[
\rho_{X,Y} = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y}.
\]  

The correlation coefficient ranges from −1 to +1. A value equal to 0 implies no linear correlation between the two variables, while a value equal to +1 or −1 implies a high positive or negative correlation respectively. A positive correlation means that as X increases so does Y, and a negative correlation means that as X increases, Y decreases. [40]

5.1.6 Features Selection

Before feeding the data to the model, it is important to find a subset of features which can bring the best results in terms of performance. It would be ideal to find the subset that can bring the optimal results but it is impossible to try all the combinations of features due to their large number. So except from testing some combination of features that would make sense based on our knowledge and intuition, a more sophisticated way to extract the most useful features was adopted.

Given the parameters, the system will start with the entire set of features, as those specified in the feature extraction/engineering module. By performing recursive feature elimination and cross-validated selection will select both the optimal number of feature space and the features themselves. On the feature elimination process, the ranking of each feature is happening according to the usual relative importance score [26]. Finally, we use the coefficient of determination known as R² score in a 10-fold cross-validation set in order to select the model with optimal number of features.

5.2 Training and Testing Method

There were two methods deployed in order to assess the model’s performance. The first is a random K-Fold cross validation, and the second is an operational validation upon the data. All of the cases were evaluated using the K-Fold Validation, while in case of Veneto the operational validation also took place.
5.2.1 Random 10-Fold Validation

The usual method for training and testing is splitting the dataset into two parts. The first part usually about 70% - 75% of the observations of the dataset is used for training the model, and the remaining 30% - 25% is used in the testing phase. However to assure that the model’s performance is reliable, a 10-Fold validation was selected. The data are randomly shuffled and the splitted in 10 parts.

Each time the 9 parts are used as training set and the remaining one as test set. This process is repeated 9 more times, each time with a different part as test set. The final model’s performance is formed out of the average model’s performance in each of ten different test sets.

5.2.2 Operational Validation

Besides the classical K-Fold Validation, an operational validation was applied in one case. After sorting the observations of each trapping site in ascending order based on the date, a time series cross validation is applied to the data. In every iteration a batch of data until a certain date is used as training set and the model’s performance is validated upon a test set consisting of observations of the next month.
For example if we want to predict the number of mosquitoes in July of 2020, observations until July of 2020 will be used as training set, while data past of this date will be completely ignored. This method will be applied iteratively to cover all the available months of a year in cross validation fashion to assess the model’s performance. Once again the final model’s performance is formed out of the average model’s performance in each of the different test sets.

5.3 Gradient Boosting

As mentioned before, the system relies on a gradient boosting machine learning model to make the predictions. Gradient boosting machines belong to a very powerful and popular family of machine learning techniques, the ensemble techniques. The main idea behind the ensemble techniques is that combining numerous weak learners, a new powerful learner can be produced.

Ensembling methods can be splitted into two categories, bagging and boosting. Gradient boosting (see Algorithm 2) belongs to the boosting methods, as its name implies. Boosting is an ensemble technique consisting of multiple predictors, in which every predictor is not made independently, but sequentially based on the errors of the previous predictors so as to learn from them.

Algorithm 2: Gradient boosting algorithm [22]

1. $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma);$
2. for $m=1$ to $M$
   1. Compute the pseudo-residuals:
      \[
      r_{im} = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \bigg|_{F(x) = F_{m-1}(x)} \quad i = 1, \ldots, n; \\
      \]
   2. Fit a base learner $h_m(x)$ to pseudo-residuals, using as training set \[
      \{(x_i, r_{im})\}_{i=1}^{n}; \\
   \]
   3. Calculate multiplier $\gamma_m$:
      \[
      \gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)); \\
      \]
   4. Update the model:
      \[
      F_m(x) = F_{m-1}(x_i) + \gamma_m h_m(x_i); \\
      \]
3. Output: $F_M(x)$;

The above algorithm is the general description of a gradient boosting machine and the progress goes on iteratively until the loss of the model is less than a certain threshold or the model reaches the maximum structural complexity. It can be implemented with any supervised machine learning model. However in the most cases, and so as in this case of predicting the class of mosquito population, it is implemented using decision trees as base learners.

A simple description of the algorithm above is:

1. Train a decision tree.
2. Apply the decision tree just trained to predict.
3. Calculate the residual of this decision tree.
4. Use residual errors as the new target variable.
5. Repeat Step 1 (until the number of trees set to train is reached).
6. Make the final prediction by adding up the predictions of all trees.

The above algorithm as mentioned before, belongs to the ensemble methods. Each decision tree is constructed based on the error of the previous tree. Decision trees have a complexity of $O(mnd)$ [62], where $n$ is the number of observations in the training set, $m$ the number of features and $d$ the depth of the tree. If $M$ trees are grown as part of an ensemble learning model, then the total complexity is $O(M(mnd))$. So it becomes clear that the more the features used and the more the trees constructed with greater depth the higher the complexity of the algorithm. Assuming that depth of the tree is on the order of $\log n$, which is the standard rate of growth of a tree with $n$ leaves, provided that it remains “bushy” and does not degenerate into a few very long and stringy, branches the total complexity of the algorithm is $O(M(mn \log n))$. From the aforementioned it can be concluded that MAMOTH time complexity is directly affected by the number of features used and the hyper-parameters tuning (number of estimators and max depth of the trees). It is important to keep the model as simple as possible to avoid long training times.

5.4 Evaluation Method

There are several metrics to measure the quality of a model depending whether it is a regression or a classification model. A very popular metric for classification models is the accuracy metric, indicating how many many times the model predicted the right class. In the case of regression models there are metrics indicating the mean error of the predictions of a model, such as Mean Absolute Error (MAE), and Mean Squared Error (MSE).

As mentioned the goal of the model is to estimate the mosquito population in the future. To achieve that, the model’s predictions are assessed using the same metric as the cost function used during the training phase, MAE. This metric indicates the distance between the actual class and the predicted one, which gives an easy intuition of the prediction’s quality. Also another metric that can characterize the quality of the system is the the percentage of predictions with error equal or less to 3 classes. This metric quantifies the times that the predictions do not deviate too much from the actual ones,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} ,$$  \hspace{1cm} (5.2)

$$\text{Percentage of error } \leq 3 = \left( \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i| \leq 3}{n} \right) \times 100 .$$  \hspace{1cm} (5.3)
5.5 Features extracted

Based on the features mentioned above in Chapter 4 there are new features that can be extracted. Feature extraction aims on finding new features deriving from the existing ones, capable of capturing useful information and increasing the model’s accuracy. In this section the extracted features that expanded the initial dataset are described. All of these features are given in the feature selection module in order to auto select the most suitable features in each case.

5.5.1 Mosquito Population Features

The daily population of the mosquitoes is strongly connected to the population of the past days. So to capture this information, the sum mosquito population of the of the past 30 days, was calculated for each trapping site. Furthermore the sum of mosquito population of the same month of previous year was calculated for each trapping site, capturing historic information indicating the trend regarding the mosquito population in the area.

5.5.2 Spatiotemporal Features

The location of a mosquito trapping site is an important feature as mosquito population is affected by their environmental surroundings. Trapping sites surrounded by vegetation and high accumulation of water are more prone to high mosquito populations. So a feature capturing this spatial information can have great impact on an accurate estimation of the expected Culex population.

There are many ways to capture the spatial information. The simpler way is using the coordinates. However a transformation of the coordinates may has more to offer. One transformation is the assignment of each pair of coordinates to a cluster. However, each province can be considered as a cluster, so create one hot encoding features out of the provinces in which traps are located are created. Trapping sites in the same cluster belong to same areas and share common characteristics. Another way to extract the spatial information out of coordinates is to calculate the distance of each trapping site from a specific point of reference.

Except from the location of the trapping site, temporality is also a very important factor that affects the mosquito population. High temperature days are connected with high number of mosquitoes. Features such as the number of days within the year or the month with temperatures higher than 30°Celsius were calculated in order to be included in the feature list.

A temporal feature indicating the season may proven very useful towards a more accurate estimation. A date is not capable of communicated such information and must be transformed. A simple idea that captures the temporal information is to calculate the distance from a specific date regardless the year, or to extract the month and the season out of the date. Mosquitoes are closely associated with the summer months. Assigning a number of 0-3 to each season from autumn to summer respectively would result to a difference of 3 between the autumn and summer where the actual difference is only one. Extracting the month (for finer time granularity) or the season out of a
date, and using a sin-cosine transformation, converting the season or the months into a cyclic feature, resulted in a very useful feature as the difference between summer and autumn is calculated in a right way. Each value is mapped into a circle, where the lowest value appears right next to the largest value, and it can be described by two values much like a coordinates system.

The formula for transforming a variable to a cyclic feature [10] is:

\[
x \rightarrow [y_1, y_2] = \left[\sin\left(\frac{2\pi x}{\omega}\right), \cos\left(\frac{2\pi x}{\omega}\right)\right],
\]

(5.4)

where \(x\) is the variable to be transformed, and \(\omega\) is the period or the total number of available values of this variable.

5.5.3 PCA Features

Principal Components Analysis (PCA) is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information of the large set. The produced variables are called components and they are eigenvectors of the data’s covariance matrix that capture as much variance of the data as possible.

The initial dataset was shrink using PCA, in an attempt to capture all the information included in the several features. As features were used the \(n\) first components that capture at least 95\% of the initial data variance. In all of the cases this percentage was achieved using the first 3 components. These 3 new features extracted were given at the system as extra features along with the previous ones to form the final dataset given as input to the feature selection module. The total list of the features and their description is given in the Appendix A Chapter.
Chapter 5. System Description and Evaluation Method

5.6 Overview

This chapter reviewed the MAMOTH framework in detail, giving a full description on the individual modules that MAMOTH is consisted of, the validation methods and the metrics used to evaluate the framework’s performance and the extra features extracted upon the initial ones. The next chapter presents the experiments conducted and the results upon the cases of study described in Chapter 3.
Chapter 6

Experiments and Results

The system was applied in five different cases concerning different areas or different mosquito genus. So the first results presented are those for Veneto area, in which an operational application of the system took place, and after we present the results of our framework for the cases of Anopheles in Veneto region of Italy, Culex in Vojvodina region of Serbia, Culex in Baden Wuerttemberg region of Germany and Aedes in Grand-Est and Corsica regions of France. In all of those cases the results were promising and consistent.

6.1 Italy, Culex genus Mosquito

The study area is located in Northeast Italy, in the Veneto region.

Figure 6.1: The entomological monitoring network of 140 traps of the Culex pipien in the Veneto region
Chapter 6. Experiments and Results

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of mosquitoes</th>
<th>Probability of at least one mosquito positive to WNV</th>
<th>Risk assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-3</td>
<td>0.23%</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>3-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9-18</td>
<td>1.07%</td>
<td>medium</td>
</tr>
<tr>
<td>4</td>
<td>18-34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>34-58</td>
<td>2.82%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>58-100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>100-167</td>
<td>6.35%</td>
<td>high</td>
</tr>
<tr>
<td>8</td>
<td>167-293</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>293-568</td>
<td>8.01%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>568</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Culex mosquito risk classes of Veneto region

The entomological monitoring of the Culex pipien in the Veneto region has been effective from 2010 to 2020, gathering data from a network of 140 stations and resulting in a dataset of more than 4800 observations.

Table 6.1 presents class separation of the initialization module, the corresponding number of mosquitoes for each class as well as the probability of having at least one mosquito positive to WNV. It can be observed from that the probability is monotonically increasing as the number of mosquitoes increases, which supports the claim that the higher the mosquito population the higher the WNV circulation and thus dissemination in the community.

Nearly 80% of the observations had at most a 15 days time distance from the next observation as it can been seen from the cdf in Figure 6.2. So the target of prediction was set to 15 days.

![Figure 6.2: CDF of the time difference between two consecutive observations.](image)

Furthermore, the auto-calibration process was tuned to max_depth = 5, number of estimators = 23 and the system decided that the optimal number of features is 16.
The selected features with their corresponding importance are presented in Figure 6.3.

It is clear that the most important feature which affects mainly the prediction of the mosquito population class after 15 days is the current mosquito population. This claim is also supported by the fact that the accumulated mosquito population of the month seems to play an important role in the formation of the final prediction. Those two features are capturing the temporality in an indirect way, and seems to be important in all culex mosquito cases independent of the area of interest. Temporality is directly captured by the days distance from a certain date regardless of the year, indicating that the mosquito population is partly following a pattern. Besides the temporality and mosquito population though, presence of water is also a considerable factor as measurements on its different states are selected by the system by 3 different features (NDWI, two past weeks cumulative rainfall and cumulative from January rainfall). Temperature is also selected and represented by 2 features, however affecting much lower in the formation of the final prediction than expected by our intuition that temperature is a main contributor for the mosquito population. Spatiality expressed by the latitude and elevation of the trap site is also features that the system chose in order to make more accurate predictions.

Operational validation was used to assess the model’s performance in the area of Veneto in 2020 data. The MAE for all the predictions is 1.27. As mentioned in Chapter 5, the model is trained upon data until a specific date, checks the performance on next month’s data, while it ignores the rest of them. This process is repeated until all available 2020 data are used. The error distribution in Figure 6.4 shows that most
of the errors are spread across a small range, meaning that 97% of the predictions are less or equal to 3 classes away from the actual class. Those promising results shows that the system’s predictions are most of the time very close to the actual mosquito population that we aim to predict.

Figure 6.4: Error distribution of Culex mosquito system in Veneto, Italy. 97% of the predictions with error $\leq 3$.

Figure 6.5: Error distribution of Culex mosquito system per class in Veneto, Italy.

In the plot of error of each class in Figure 6.5, we can see that the model is performing similarly in all risk population classes, without any strong bias to low or to high risk classes.
Chapter 6. Experiments and Results

The prediction error of each month is relatively equal as it can be seen in Figure 6.6, the MAE in June is higher due to smaller size of dataset and the lack of data, before May of 2020, thus training the model only upon data of previous years and not in recent observations. Respectively the MAE of October is lower than the others, due to the training of the model in many more recent observations.

![Mean Absolute Error per month](image)

Figure 6.6: Error distribution of Culex mosquito system per month in Veneto, Italy

In order to validate the performance of the model except the operational application, the system was tested on random 10-fold validation using all the available data. The results showed slightly better behavior, in terms of MAE: 1.14, and similar performance in terms of percentage of error below 3 classes: 97%. This slight improvement can be explained by the fact that in the k-fold validation the samples selected for train and test process are selected uniformly from the entire dataset compared with the operational case where train and the test sets are totally separated in time. Those results are leading us to the conclusion that the performance of the model is stable according to train-test separation of the dataset.

6.1.1 Performance without the Entomological features

As depicted in Figure 6.3 the model relies a lot on the entomological features in order to predict the mosquito population for the upcoming period. The current number of Culex mosquitoes is the most important feature by far, while also the feature with the third highest relative importance score is the sum of Culex mosquitoes of the past 30 days and the fifth feature on the list is the mosquito population of the same month the previous year. The need of those entomological features could limit the wide use of the model, once this information is known only on the trap-site. Away of the trap-sites this information will not been know. Thus, the question that we like to answer is, could the model perform reliably if those important entomological features are missing from the feature space?
To test this hypothesis we removed all features relevant to entomological data and we re-training a new MAMOTH model using only EO data and features derived from them. The results showed that the model was still able to accurately predict the upcoming mosquito population with a small accuracy reduction compared with the model that used entomological features. The new MAMOTH model performed with 1.65 MAE and the percentage of errors below 3 classes was reduced to 92%. The wide applicability of a model that relies only on EO data, mark those results as promising for further research in that direction.

As we can see in Figure 6.7, the new model in order to fill the “room” that was created from the absence of the entomological features, increased the total amount of selected features to 34 (compare to the 16 of the model with the entomological features), as well as the importance of EO related features (rainfall, LST, ndwi, ndvi, ndbi) is significantly increased.

6.2 Serbia, Culex genus Mosquito

The study area is located in Northern Serbia, in the Vojvodina region.
### Table 6.2: Culex mosquito risk classes of Vojvodina region

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of mosquitoes</th>
<th>Risk assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 - 18</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>19 - 55</td>
<td>medium</td>
</tr>
<tr>
<td>3</td>
<td>56 - 115</td>
<td>high</td>
</tr>
<tr>
<td>4</td>
<td>104 - 197</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>198 - 202</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>203 - 313</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>314 - 482</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>483 - 731</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>732 - 1225</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&gt; 1225</td>
<td></td>
</tr>
</tbody>
</table>

The entomological monitoring of the Culex pipien in the Vojvodina region has been effective from 2010 to 2019, gathering data from a network of 124 stations and resulting in a dataset of more than 926 observations. Table 6.2 presents the corresponding number of mosquitoes for each class.

Nearly 70% of the observations had at most 25 days time difference from the next observation as it can be seen from the cdf in Figure 6.9. So the target of prediction was set to 25 days.
Figure 6.9: CDF of the time difference between two consecutive observations in Vojvodina region, Serbia. Nearly 70% of the observations have at most 25 days difference from the next observation, so target time was set to 25 days.

Furthermore, the auto-calibration process was tuned to max_depth = 4, number of estimators = 23 and the system decided that the optimal number of features is 12. The selected features with their corresponding importance are presented in Figure 6.10.

Figure 6.10: Feature importance of Culex mosquito system in Vojvodina, Serbia
The sum of the mosquito population the last 30 days is the most important feature while the distance of days from a certain date regardless of the year is the second most important feature showing that temporality is affecting the upcoming Culex mosquito population. Rainfall is also an important factor represented by 2 features followed by temperature and the maximum wind in the area. Elevation and slope of the area must also be considered in the case of Serbia. Comparing these features with those selected for the Veneto case in the previous section, it can be observed that there are similarities between these 2 cases due to the same mosquito genus, while the differences show that the framework modify the set of features to better fit the area of Vojvodina, Serbia.

Figure 6.11: Error distribution of Culex mosquito prediction system in Vojvodina, Serbia. 90% of the predictions with error $\leq 3$. 
Chapter 6. Experiments and Results

Figure 6.12: Error distribution of Culex mosquito system per class in Vojvodina, Serbia.

The MAE for all the predictions is 1.54. The error distribution in Figure 6.11 shows that the error is more spread out than the one in Veneto area, however still achieving 90% of the predictions to be less or equal to 3 classes away from the actual class.

In the plot of error of each class in Figure 6.12, we can see that the error is not distributed equally in all classes, rather it ranges from 0.75 to 2.87. This deviation among the error of each class is maybe a result of the small size of the dataset, as it is consisted of only 926 observations as it can been seen from the Table 4.1.

6.3 Germany, Culex genus Mosquito

The study area is located in Southwest Germany, in the Baden Wuerttemberg region.
Table 6.3: Culex mosquito risk classes of Baden Wuerttemberg region

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of mosquitoes</th>
<th>Risk assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 - 1</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>2 - 4</td>
<td>medium</td>
</tr>
<tr>
<td>3</td>
<td>5 - 7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8 - 10</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>11 - 15</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16 - 22</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>23 - 36</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>37 - 77</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>78 - 206</td>
<td>high</td>
</tr>
<tr>
<td>10</td>
<td>≥ 206</td>
<td></td>
</tr>
</tbody>
</table>

The entomological monitoring of the Culex pipien in the Baden Wuerttemberg region has been effective from 2010 to 2019, gathering data from a network of 86 stations and resulting in a dataset of more than 3700 observations.

Table 6.3 presents the corresponding number of mosquitoes for each class.

Nearly 80% of the observations had at most a 25 days time distance from the next observation as it can be seen from the cdf in Figure 6.14. So the target of prediction was set to 25 days.
Figure 6.14: CDF of the time difference between two consecutive observations in Baden Wuerttemberg region, Germany. Nearly 80% of the observations have at most 25 days difference from the next observation, so target time was set to 25 days.

Figure 6.15: Feature importance of Culex mosquito system in Baden Wuerttemberg, Germany.

Furthermore, the auto-calibration process was tuned to max_depth = 4, number of estimators = 23 and the system decided that the optimal number of features is 33. The selected features with their corresponding importance are presented in Figure 6.15.
In this case of study the framework chose a wider set of features in order to deliver more accurate predictions. However, just like the rest of the Culex cases distance of days remains an important feature along with rainfall and the current mosquito population.

![Error Distribution](image)

**Figure 6.16:** Error distribution of Culex mosquito system in Baden Wuerttemberg, Germany. 92% of the predictions with error $\leq 3$.

![MAE per Class](image)

**Figure 6.17:** Error distribution of Culex mosquito system per class in Baden Wuerttemberg, Germany.

The MAE for all the predictions is 0.97. The error distribution in Figure 6.16 shows that the error is ranging from its minimum possible value to the maximum possible value. However the majority of the error is accumulated in low values, resulting in 92% of the predictions to be less or equal to 3 classes away from the actual class.
In the plot of error of each class in Figure 6.17, we can see that the error is following an ascending pattern, with higher the actual class the higher the error of the prediction. This result is maybe because of the close margins of the classes, as well as the size of the dataset, which is not considered as big dataset, rather a dataset marginally sufficient to train a model of 10 classes with such close range of mosquito population between the classes.

6.4 France, Aedes genus Mosquito

The study area is located in northeastern France, in the Grand-Est region and Corsica region.

The entomological monitoring of the Aedes mosquito eggs in the Grand-Est and Corsica regions has been effective from 2017 to 2019, gathering data from a network of 81 stations and resulting in a dataset of more than 1700 observations.

Table 6.4 presents the corresponding number of mosquitoes for each class.

Nearly 80% of the observations had at most 20 days time difference from the next observation as it can been seen from the cdf in Figure 6.19. So the target of prediction was set to 25 days.
Chapter 6. Experiments and Results

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of mosquitoes</th>
<th>Risk assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>1 - 6</td>
<td>medium</td>
</tr>
<tr>
<td>3</td>
<td>7 - 15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16 - 29</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30 - 50</td>
<td>high</td>
</tr>
<tr>
<td>6</td>
<td>51 - 78</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>79 - 129</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>130 - 209</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>210 - 400</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&gt; 400</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: Aedes Mosquito eggs risk classes of the Grand-Est and Corse regions

Figure 6.19: CDF of the time difference between two consecutive observations in the Grand-Est and Corsica regions, France. Nearly 80% of the observations have at most 25 days difference from the next observation, so target time was set to 25 days.

Furthermore, the auto-calibration process was tuned to max_depth = 6, number of estimators = 15 and the system decided that the optimal number of features is 11. The selected features with their corresponding importance are presented in Figure 6.20. In this case of mosquito genus and area the framework chose temperature related features as well as water related features in order to achieve good prediction performance. The common characteristic of this case with the previous ones is the importance of the mosquito population (No of Ae. albopictus eggs and Aedes sum of population during the current month in the previous year).

The MAE for all the predictions is 0.71. The error distribution in Figure 6.21 shows that the majority of the predictions are accurate resulting in 0 error, while the rest of the errors are spread equally in the rest possible values. This results in 92% of the predictions to be less or equal to 3 classes away from the actual class.
Chapter 6. Experiments and Results

Figure 6.20: Feature importance of Aedes mosquito system in the Grand-Est and Corsica regions, France.

Figure 6.21: Error distribution of Aedes mosquito system in the Grand-Est and Corsica regions, France. 92% of the predictions with error $\leq 3$. 
In the plot of error of each class in Figure 6.22, we can see that the model is performing similarly with the model of Germany, with the error following an ascending pattern. The primary identified cause of this problem is once again the even smaller dataset size along with the close margins of the classes.

6.5 Italy, Anopheles genus Mosquito

The study area is located in Northeast Italy, in the Veneto region.
The entomological monitoring of the Anopheles in the Veneto region has been effective from 2010 to 2020, gathering data from a network of 130 stations and resulting in a dataset of more than 629 observations.

Table 6.5 presents the corresponding number of mosquitoes for each class.

Nearly 60% of the observations had at most a 30 days time distance from the next observation as it can be seen from the cdf in Figure 6.2. So the target of prediction was set to 30 days.

Furthermore, the auto-calibration process was tuned to max_depth = 4, number of estimators = 22 and the system decided that the optimal number of features is 54.
The selected features with their corresponding importance are presented in Figure 6.25. Due to the small size of dataset (629 observations) and thus the limited information provided by them, the framework selected an extended set of features for this case in order to accurately predict the Anopheles mosquito population in the framework in Veneto, Italy. Although this difference with the previous cases, once again is clear that the distance of days and the current mosquito population is the primary features that the framework takes into account to deliver the predictions.
Figure 6.26: Error distribution of Anopheles mosquito system in Veneto, Italy. 95% of the predictions with error $\leq 3$.

The MAE for all the predictions is 1.50. The error distribution in Figure 6.26 shows that most of the errors are spread across a small range, meaning that 95% of the predictions are less or equal to 3 classes away from the actual class, a very promising result considering the very small dataset size of 629 observation.

Figure 6.27: Error distribution of Anopheles mosquito system per class in Veneto, Italy.

In the plot of error of each class in Figure 6.27, we can see that the model has a worse performance in the extreme classes than the middle ones. While the MAE of the model is within acceptable values, in order to achieve that MAE the model had to
include 54 features presented in Figure 6.25. The high number of features is expected, as the model in order to make accurate predictions, has to overcome the problem of the very small size of the dataset along with the 30 days difference between the the current and the prediction date. The bigger the time interval between those two dates the harder for the model to make accurate predictions.

6.6 Overview

This Chapter went through the results over each case of study. It presented the features selected and the performance of the framework in each one of them accompanied by a short discussion. In the Appendix A Chapter an extensive catalogue of the features can be found for further reference over the available set of features used by the framework. Next chapter presents a more extensive discussion over the over results and provides the directions of the future work.
Chapter 7

Conclusions

In this thesis MAMOTH, an EO based prediction framework for the mosquitoes abundance was presented. MAMOTH is able to predict the upcoming mosquito population of any genus in any area, based on open EO data and entomological data. The main characteristic of the framework that discriminates it from the existing works is that MAMOTH decides what features to use without human interference. In that way no human bias is injected in the predictions, letting it to decide which features are important. This led to a generic auto adjustable framework that its architecture allows for a versatile and easily replicated framework that can be applied without any special requirement to any country without any modifications.

7.1 Results Overview and Discussion

MAMOTH prediction system was operationally successfully applied in the area of Veneto and validated upon 4 different other cases. The results presented showed that the model can perform reliable and accurately regardless the area or the mosquito of interest. Additionally, study of the features selected in each case is allows for further study of the factors affecting each mosquito population. The evaluation of different cases with a framework that uses the same principles gave the opportunity of comparative studies between them.

Table 7.1 presents an overview of the performance of MAMOTH to the aforementioned cases. The results make obvious that indeed the MAMOTH framework is generic and easily replicable to other cases. It is also shown that although the auto-tuned parameters are varying for the different use cases, the performance of the models remains stable and high with the maximum accuracy being returned in the case of Aedes in France where the MAE is surprisingly low.

The 13 most important features, selected by MAMOTH, and their corresponding importance for each case of interest are presented in the Table 7.2. By comparison between the different cases we can exclude some insights:

- The accumulated rainfall from the beginning of the year is important for all the cases, and for the cases of Culex, the accumulated rainfall of the last two weeks seems important as well.
- In all Culex cases, the rainfall and the water indices, NDWI, are more important than the temperature, LST.
Table 7.1: MAMOTH’s pre-operational applications and performance per case.

- Anopheles is the only mosquito that the most important feature is not the previous state of the mosquito population but the direct time distance.

- Aedes prediction is the only case where the direct time distance is not important for the model.

- NDWI metrics are very important of Aedes Albopictus prediction compared with the other mosquitoes.

- Mosquito population (current month or accumulated sums) along with the days distance are between the most important features across all cases.

As conclusion, this thesis proved that is feasible to develop a generic machine learning model that predicts the mosquito population class (1-10) of any mosquito genus in any area of interest. The results proved that this approach achieves accurate and reliable performance, using widely accessible satellite and entomological data.

By studying the features selected in each case, it is clear that the MAMOTH is able to adjust the features to achieve as high as possible performance. Among the most important features selected, are some of the ones constructed upon the initial ones pointing out the importance of feature engineering. The automated selection of the features prevents the injection of human bias in the framework, as the user does not specify the features based on their knowledge, rather they specify a list of potential features and the framework decides which of them to use. This capability of features auto-tuning allows the framework to be transferable and applicable to any case without any modification while users can always extend the potential features.

By getting a closer look to the results, the performance of the model over 5 different cases was accurate as maximum MAE is equal to 1.54, while minimum MAE is equal to 0.71. Even in the worst case of 1.54, the magnitude of error is tolerable.
### Chapter 7. Conclusions

<table>
<thead>
<tr>
<th>Aedes - France</th>
<th>Anopheles - Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature names</td>
<td>importance</td>
</tr>
<tr>
<td>mosq_now</td>
<td>0.501</td>
</tr>
<tr>
<td>lst_night</td>
<td>0.089</td>
</tr>
<tr>
<td>lst_day</td>
<td>0.079</td>
</tr>
<tr>
<td>ndwi_mean</td>
<td>0.073</td>
</tr>
<tr>
<td>mosq_month_previousYear</td>
<td>0.053</td>
</tr>
<tr>
<td>ndwi_std</td>
<td>0.043</td>
</tr>
<tr>
<td>acc_a infl fall jan</td>
<td>0.042</td>
</tr>
<tr>
<td>ndwi</td>
<td>0.041</td>
</tr>
<tr>
<td>PCA_2</td>
<td>0.029</td>
</tr>
<tr>
<td>PCA_3</td>
<td>0.027</td>
</tr>
<tr>
<td>mo_cos</td>
<td>0.023</td>
</tr>
<tr>
<td>LST_jan_mean</td>
<td>0.017</td>
</tr>
<tr>
<td>mosq_month_sum</td>
<td>0.014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Culex - Serbia</th>
<th>Culex - Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature names</td>
<td>importance</td>
</tr>
<tr>
<td>mosq_month_sum</td>
<td>0.265</td>
</tr>
<tr>
<td>days_distance</td>
<td>0.257</td>
</tr>
<tr>
<td>mosq_now</td>
<td>0.187</td>
</tr>
<tr>
<td>acc_rainfall_jan</td>
<td>0.083</td>
</tr>
<tr>
<td>LST_Mar_mean</td>
<td>0.039</td>
</tr>
<tr>
<td>DEM_000</td>
<td>0.036</td>
</tr>
<tr>
<td>acc_rainfall_2week2</td>
<td>0.036</td>
</tr>
<tr>
<td>Slope_1000</td>
<td>0.027</td>
</tr>
<tr>
<td>max_wind</td>
<td>0.022</td>
</tr>
<tr>
<td>mosq_month_previousYear</td>
<td>0.021</td>
</tr>
<tr>
<td>PCA_2</td>
<td>0.016</td>
</tr>
<tr>
<td>celsius</td>
<td>0.011</td>
</tr>
<tr>
<td>mosq_month_sum</td>
<td>0.004</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Culex - Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature names</td>
</tr>
<tr>
<td>mosq_now</td>
</tr>
<tr>
<td>days_distance</td>
</tr>
<tr>
<td>mosq_month_sum</td>
</tr>
<tr>
<td>DEM_1000</td>
</tr>
<tr>
<td>mosq_month_previousYear</td>
</tr>
<tr>
<td>acc_rainfall_jan</td>
</tr>
<tr>
<td>ndwi_mean</td>
</tr>
<tr>
<td>ndbi_mean</td>
</tr>
<tr>
<td>PCA_3</td>
</tr>
<tr>
<td>PCA_1</td>
</tr>
<tr>
<td>lst_day</td>
</tr>
<tr>
<td>y</td>
</tr>
<tr>
<td>ndwi</td>
</tr>
</tbody>
</table>

Table 7.2: Most important features per case.
considering the fact that are 10 possible classes for each prediction. These results proved that MAMOTH is a transferable and replicable framework over any case.

Strengthening the claims for an accurate and a reliable framework, the results of the MAMOTH model trained only upon environmental data, were not disappointing, as the performance was not deviated much from the initial model, proving that even in lack of entomological data, the system remains robust and able to predict mosquito population. This variation of the system offers a more flexible model applicable even to communities that have not implemented a mosquito surveillance program in order to collect entomological data. However the use of entomological data offers valuable information to the model enabling for more accurate predictions. An important difference between the two models, however, is the number of features selected by the model. In the second case where only EO data are used, the number of features is significantly larger, making the model much more complicated. Also the most important variables are those covering spatiotemporal.

Additionally the proposed framework, give the opportunity of comparative study between different areas or mosquitoes. The replicability of the framework under the same architecture and the same mathematical principles, lead to the extensive capability of comparative studies among different cases, responding to: “which characteristics seem important in one case and which on the other?”, which based on the literature review was not covered by any approach. A list of the most important features in each case is given in Table 7.2 for further study.

The only exception which MAMOTH did not demonstrate a good performance was the Greek landscape. The probable reason for this might be that the examined regions are characterized by the presence of extended irrigation systems and water bodies, rice fields and water demanding cultivations with a non-predictable anthropogenic water management, in contrast to the riverine systems of the other areas (Po river – Italy, Danube and Savvas river – Serbia, Rhine river – Germany, Herault and Aude – France). Thus, the causality behind the mosquito abundances fluctuations can not be captured by the generic earth observational data that are model used.

7.2 Future Work

There is a lot of work to be done, as this thesis is the first to address the problem, to the best of the author’s knowledge. Thus, some interesting future ideas for further investigation are proposed below:

1. Investigate the direction of non-entomological features. Explore other kind of data that can contribute to more accurate predictions.

2. Land-use features. Once the land-use features are a lot, research should happen in the direction of better representation of the land-use features (dimensionality reduction) in order to not explode the complexity.

3. Forecast away from the traps (extrapolate). Up until now, the framework has only been tested in trap locations where their past observations are used as
training set. As next step, the prediction of trap locations with no previous known data could be implemented.

4. Add diversity on the model, parallel training of multiple predictors in order to a) improve the accuracy b) estimate the uncertainty (or confidence) of the predictions.

5. Explore NN solutions. One consideration would be the creation of features that capture the neighboring trap sites status.

6. Use meteorological estimations to improve the model (this part except ML algorithms, needs development in order to “pull” meteorological estimations)

7. Connect with the human cases, risk analysis. Expand the framework to predict human cases of MBDs based on EO, entomological and other data.

Based on the above ideas MAMOTH framework will be further developed to facilitate MBDs human cases predictions expanding the pool of features and exploring new prediction ways towards higher accuracy performance. Also it is considered a potentially extension of its capability over prediction in areas with extended irrigation systems and water bodies, rice fields and water demanding cultivations with a non-predictable anthropogenic water management so as the framework to be applicable to a wider range of topological conditions.
## Appendix A

### Feature List

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dt_placeme</td>
<td>Date of the observation</td>
</tr>
<tr>
<td>stationid</td>
<td>Station ID</td>
</tr>
<tr>
<td>x</td>
<td>Longitude</td>
</tr>
<tr>
<td>y</td>
<td>Latitude</td>
</tr>
<tr>
<td>mosq_now</td>
<td>Mosquito population in trapping sites at the date of observation</td>
</tr>
<tr>
<td>NDVI</td>
<td>Proxy for the vegetation density and distribution.</td>
</tr>
<tr>
<td></td>
<td>Extracted pixel value of overlapping station ID coordinates</td>
</tr>
<tr>
<td>NDVI_mean</td>
<td>Proxy for the vegetation density and distribution.</td>
</tr>
<tr>
<td></td>
<td>Mean value of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDVI_std</td>
<td>Proxy for the vegetation density and distribution.</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDWI</td>
<td>Proxy for changes in water content</td>
</tr>
<tr>
<td></td>
<td>Extracted pixel value of overlapping station ID coordinates</td>
</tr>
<tr>
<td>NDWI_mean</td>
<td>Proxy for changes in water content</td>
</tr>
<tr>
<td></td>
<td>Mean value of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDWI_std</td>
<td>Proxy for changes in water content</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDMI</td>
<td>Proxy for determination of vegetation water content</td>
</tr>
<tr>
<td></td>
<td>Extracted pixel value of overlapping station ID coordinates</td>
</tr>
<tr>
<td>NDMI_mean</td>
<td>Proxy for determination of vegetation water content</td>
</tr>
<tr>
<td></td>
<td>Mean value of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDMI_std</td>
<td>Proxy for determination of vegetation water content</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDBI</td>
<td>Proxy for mapping urban built-up areas</td>
</tr>
<tr>
<td></td>
<td>Extracted pixel value of overlapping station ID coordinates</td>
</tr>
<tr>
<td>NDBI_mean</td>
<td>Proxy for mapping urban built-up areas</td>
</tr>
<tr>
<td></td>
<td>Mean value of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>NDBI_std</td>
<td>Proxy for mapping urban built-up areas</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of neighboring pixels (window of 3x3)</td>
</tr>
<tr>
<td>LST_day</td>
<td>Land surface temperature at day</td>
</tr>
<tr>
<td>LST_night</td>
<td>Land surface temperature at night</td>
</tr>
<tr>
<td>LST_Jan_mean</td>
<td>Mean temperature in January</td>
</tr>
<tr>
<td>LST_Feb_mean</td>
<td>Mean temperature in February</td>
</tr>
<tr>
<td>LST_Mar_mean</td>
<td>Mean temperature in March</td>
</tr>
<tr>
<td>LST_Apr_mean</td>
<td>Mean temperature in April</td>
</tr>
<tr>
<td>wind_max</td>
<td>Max magnitude of wind</td>
</tr>
<tr>
<td>wind_mean</td>
<td>Mean magnitude of wind hourly</td>
</tr>
<tr>
<td>wind_min</td>
<td>Min magnitude of wind</td>
</tr>
</tbody>
</table>
## Appendix A. Feature List

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc_rainfall_1week</td>
<td>Accumulated precipitation counting towards one week before the date of placement</td>
</tr>
<tr>
<td>acc_rainfall_2week2</td>
<td>Accumulated precipitation counting towards two weeks before the date of placement</td>
</tr>
<tr>
<td>acc_rainfall_jan</td>
<td>Accumulated precipitation counting from the 1st of January of each year</td>
</tr>
<tr>
<td>WC_L_1km</td>
<td>Combination of breeding site length and water course of national hydrological data within a buffer zone of 1000 m around each sampling/trapping site</td>
</tr>
<tr>
<td>DEM_1000</td>
<td>Mean elevation (resolution = 12.5 m), within a buffer of 1000m around trapping sites</td>
</tr>
<tr>
<td>Aspect_1000</td>
<td>Mean aspect (12.5 m), within a buffer of 1000 m around trapping sites</td>
</tr>
<tr>
<td>Slope_1000</td>
<td>Mean slope (12.5 m), within a buffer of 1000 m around trapping sites</td>
</tr>
<tr>
<td>Coast_dist_1000</td>
<td>Mean Distance of sampling/trapping site within a buffer of 1000m from coastline</td>
</tr>
<tr>
<td>WC_dist_1000</td>
<td>Distance of combination of breeding site length and length of watercourses of national hydrological data within a buffer zone of 1000m around each sampling/trapping site</td>
</tr>
<tr>
<td>Flow_acc_1000</td>
<td>Mean flow accumulation within a buffer of 1000 around trapping sites</td>
</tr>
<tr>
<td>mosq_month_sum</td>
<td>Cumulative mosquito population of the past 30 days</td>
</tr>
<tr>
<td>mosq_month_prevYear</td>
<td>Cumulative mosquito population of the month on previous year</td>
</tr>
<tr>
<td>mosq_bins</td>
<td>Mosquito bin based on the population on the date of observation</td>
</tr>
<tr>
<td>days_distance</td>
<td>Time difference in days between the date of placement and a specific date regardless the year</td>
</tr>
<tr>
<td>province</td>
<td>Province in which each trap is located</td>
</tr>
<tr>
<td>mo_cos</td>
<td>Cosine transformation of the month of date of placement</td>
</tr>
<tr>
<td>mo_sin</td>
<td>Sine transformation of the month of date of placement</td>
</tr>
<tr>
<td>celsius</td>
<td>LST_day to celsius conversion</td>
</tr>
<tr>
<td>summer_days_year</td>
<td>Days with over 30o celsius within the year</td>
</tr>
<tr>
<td>summer_days_month</td>
<td>Days with over 30o celsius within the month</td>
</tr>
<tr>
<td>PCA components</td>
<td>3 PCA components extracted from the whole dataset</td>
</tr>
</tbody>
</table>
Bibliography - References


