From sampling to models and maps: a Roadmap of best practice

AIMCOST Task 1.2, Deliverable 3.3 and Deliverable 3.4

Cedric Marsboom, Kamil Erguler, Miguel A. Miranda, Sophie Vanwambeke and William Wint

May 2023

info@aedescost.eu

i

ONE PAGE SUMMARY

At the request of the network members two summaries have been provided – this single page overview, a longer Executive Summary which follows.

Since 2018, *Aedes* Invasive Mosquito (AIM) COST Action (Project ID: CA17108) has promoted European networking and collaboration between the researchers, public health professional and the public so as to increase harmonisation, preparedness and capacity for surveillance and control of AIM vector species. Two of its defined outputs is *Integrating surveillance data analysis, spatial modelling & mapping to ensure the quality and applicability of future technical outputs at the European level (T1.2);* and *Guidelines of Customisation of mapped Outputs* (D3.4). This document addresses these by setting out a Roadmap of activities from: 1) surveillance of vector distributions to aid vector control to mitigate nuisance and disease risk through to 2) the modelling needed to fill the gaps in surveillance outputs to make the maps complete, and 3) the production of maps needed to communicate the outputs to a range of stakeholders

The Roadmap is not intended to be a comprehensive guide to all the many complex procedures and specialist methodologies that contribute to this chain of activities. Rather it is intended to be an extended aide memoire for a wide range of non-specialist professionals, providing recommendations for best practice for each of the major components.

The document summarises and illustrates activities contributing to four major Roadmap components:

Surveillance: The section covers sampling, data exploration, preparation and reporting. Overviews are given of the factors involved in identifying the places to sample, as well as when and how often the samples should be taken, and what parameters should be recorded. Some aspects of data processing and manipulation, standardisation and validation are also discussed, together with what elements of the recorded sample data should be included in published results to ensure consistency and usability by as wide a readership as possible.

Modelling: This section first discusses elements that need to be considered before modelling can be implemented, specifically issues of data scale, data aggregation and also the datasets other than the vector data, *i.e.* the predictor covariates that are needed to run a model. Two types of models are described: spatial models that usually produce static predicted vector distributions; and mathematical models that provide dynamic estimates of vector populations over time. The methods for both model families are set out and the constraints and prerequisites for effective modelling are illustrated.

Data visualisation and mapping: Both sampling and modelling are complex , so the results of which need to be communicated concisely, very often using the same outputs for a wide range of users. The key point emphasised throughout is that the map producer needs to have a clear concept of the message to be delivered. The section sets out the basic principles of cartography and mapping, links map design to the type of data the map contains, and illustrates some common mistakes to avoid.

Implementation and integration: The final section focuses on integration and implementation, with the specific example of *Aedes aegypti* considered as a use case. The content is derived from a series of workshops run during the Action Annual Meeting in Rome in February 2023, where participants were asked to devise strategies to promote integration at the operation level for each of the surveillance, mapping and modelling components, using this Road map for guidance.

The main activities identified were first to establish situation by confirming the true presence/absence of the vector in the region of first detection, and define sampling strategy needed to provide that information, including defining where to put traps using existing vector models or habitat suitability thresholds if available. The next step should be to use modelling to predict likely direction and extent of spread, using the predictions to define wider scale surveillance to monitor spread. Key to success is communication – namely providing maps of sampling results and distribution models, and predicted spread to stakeholders at the earliest opportunity .

EXECUTIVE SUMMARY

This document addresses several planned products for *Aedes* Invasive Mosquito (AIM) COST Action (Project ID: CA17108), namely:- *Task 1.2: Integrating surveillance data analysis, spatial modelling & mapping to ensure the quality and applicability of future technical outputs at the European level*; Deliverable 3.3: *Synthesised feedback with recommendation for customising surveillance, control and modelling guidelines*; and Deliverable 3.4: *Guidelines for customisation of mapped outputs.*

In order to effectively manage the nuisance and disease risk due to invasive mosquitos (and the pathogens they carry), we need to know where they are and anticipate where they will be and how prolific they will be where they are. This means we have to sample and map their locations, and refer to modelling techniques to patch up any gaps emerging in our knowledge. Here, we set out a high-level overview, a Roadmap, of activities comprising surveillance, modelling, data visualisation and mapping, and implementation and integration for *Aedes* invasive mosquito species. We aim to inform control but not to cover specific control methods or strategies. We also do not address training needs or provide detailed guidelines for sampling, or any other component. Instead, we aim to provide recommendations for best practice for each activity, from surveillance to implementation, for a wide range of non-specialist professionals.

SAMPLING

Defining a sampling strategy starts by precisely defining the objectives. Multiple decisions are required at this point to design an efficient and useful sampling effort: how often to sample, where to sample, which protocols to employ, and what parameters to record during sampling. Whether the objective is to detect presence, to perform control or assess control effectiveness, or to estimate disease risk, a key constraint is to ensure that any sampling is repeatable and is sufficiently standardised to produce results that are comparable between different locations and times. An example is provided by the AIMSURV initiative (Miranda et al., 2022), facilitated by AIM-COST to harmonize sampling activities and protocols used by European entomologists and implement data sharing and publication processes.

Fieldwork is usually designed in the form of cross-sectional sampling, where multiple locations are sampled only once, or longitudinal sampling, where the same locations are sampled repeatedly. Cross-sectional sampling covers a wider area with the same number of traps, and its output can be used for presence and absence modelling, *i.e.* to project species distribution. For this, sampling locations are assigned to different geographical zones (or strata) related to the biology of the vector, *e.g.* vegetation category. Seasonality is an important concern in determining sampling times. For instance, when detecting the presence of a target species, you should aim to sample at the height of activity. Longitudinal sampling, on the other hand, monitors populations over time, which makes the results suitable for mathematical modelling (also called dynamic or mechanistic modelling). Mathematical modelling concerns projecting population dynamics (abundance as well as activity), and it requires data capturing change in the environment and its effects on the population. As additional evidence, you may include samples of different life stages collected from different habitats and at times beyond the limits of vector activity.

When designing a sampling strategy for the first time either for a species, or an area, carry out exploratory sampling. Determine the locations to cover plausible niches, and the sample sizes to balance labour and statistical rigor as well as possible. Useful information can be gathered by simply examining the results, or visualising the data as maps.

Two keys to good sampling deserve highlighting: one is matching the sampling to the objective, as we outlined above, and the other is the reproducibility of the methods by using standardized protocols. In order to make sample reports comparable from study to study, recorded numbers need to be standardised according to sampling effort, and the samples need to be adequately georeferenced and time stamped. The minimum requirements for reporting can be summarised as follows: trap type/sample methods, number of traps/sample events, start and end collection dates, trap/sample geographical coordinates and location

descriptors, number of specimens caught by species (including zeros), and method of species identification. It is just as important to know where a species in not found as it is to record its presence or abundance. Particularly, the beginning or end of a season, or the edge of a species known range, bears valuable information to train a good model.

MODELLING

Species distribution and abundance are determined by biotic factors, such as predation, nutrient availability, and growth rate, and abiotic factors, such as temperature, rainfall, and humidity. Spatial (or statistical) modelling establishes statistical relationships between these factors (covariates) and species presence/absence or abundance data (observations) collected from a set of locations. It then applies these relationships to predict presence/absence or abundance at other locations. Statistical relationships between covariates and observations are only as robust as the set of observations allows. How large the set should be? depends on the complexity of the relationships you wish to validate. The outputs of species distribution models usually do not reflect temporal or seasonal changes, and thus, are visualised as static maps.

Due to the extent and complexity of many environmental datasets, data reduction techniques can be applied. These yield summary statistics, such as minima, maxima, or temperature of the warmest quarter. Correlations between covariates, such as vegetation and temperature (or rainfall), interfere with the modelling process, and should be removed in advance. Advanced statistical methods, such as temporal Fourier analysis, can be used to remove potential correlations and to summarise long time series data.

Machine learning models are widely used in spatial modelling. These use algorithms to find patterns in the data, classify observations, and project future distributions. Machine learning is data-driven, and useful when dealing with "wide data", where the number of covariates exceeds the number of observations. It masters complex relationships, but cannot isolate the effect of a single covariate. Also, it is ideal when you are not interested in (directly) estimating uncertainty.

In contrast to spatial modelling, mathematical modelling is process based, or mechanistic. It concerns creating a mathematical representation of the biological/ecological processes and the covariates driving their dynamics and interactions. Constructing such complex relationships requires extensive data collection and expert consultation. Any observation is good for model calibration and validation, as long as it is comparable with a particular model output. These include (i) number of eggs in an ovitrap, (ii) number of adults caught in an adult trap, and (iii) number of bites reported by a person. However, longitudinal observations are preferred, as they capture the dynamics of populations and covariates. Mathematical models may represent spatial and temporal dynamics of vector populations, including adult activity and abundance, duration of the peak season, and the first emergence of adults. Such indicators can be summarised into future projections (for the dynamics) and maps (for the distribution) of risk.

Regardless of their type, models are constructed, calibrated (trained or configured), and validated (tested). Model construction relies on a particular mathematical method and expert knowledge (for picking up the right covariates, performing supervised machine learning, or composing mathematical representations). Once constructed, we perform training and testing by splitting the data in 2 or 3 independent subsets. The largest subset (70–80%), the training set, is used to calibrate your model (to configure internal parameters). An optional intermediate subset (10–15%) can be used for validation and finetuning. The last 10–15% constitutes the test set, which is used to evaluate model performance. This way, you use independent data sets at each step and avoid a self-fulfilling prophecy. If you use the same data to train and test your model, you risk becoming too good when predicting your own data, but farcical when presented with an out-of-sample case.

DATA VISUALISATION AND MAPPING

Maps are efficient ways of rapidly displaying complex information about a single parameter. They can be used to validate data entry and to locate outliers or anomalous values, which may represent errors. Mapping can be helpful for typical situations, such as planning fieldwork for surveillance, planning monitoring for interventions or management of decision making. Simply mapping sampling locations will provide

information about the sampling effort, spatial imbalances, and gaps. They are most used however as ways to convey information to a wide range of technical and lay audiences.

There is no single way of mapping: the cartographer bears the responsibility of proposing something that allows the reader to draw relevant conclusions. Whatever the audience, you must first identify the message you want your map to convey. Then you must identify how the map can help you get that message across. The map needs to do that by itself, *i.e.* be clear and easy to read. As a rule of thumb, a map is only necessary if we consider that spatial heterogeneity matters. If one trap or model, or the aggregated (or maximum value across places) observation or prediction provides the information needed, then a map is unnecessary. A graph following the progress of values may be more useful.

Spatial data, *e.g.* trapping locations, have associated numerical or categorical attributes which may include their geographical coordinates, dates of trapping, trap type, person trapping, and number of mosquitoes collected. Any parameter, including model outputs, can be an attribute as long as it is linked to the spatial entity by identifiers or geographical coordinates. Variables representing absolute numbers, *e.g.* total population, number of schools, number of disease cases, are mapped using symbols of varying size. Choropleth maps - maps in which an entire area, typically an administrative unit, is given a colour - are more appropriate for proportions, *e.g.* population densities, schools per inhabitant, and disease incidence rate.

Any flat map distorts the actual (spherical) distance between two locations. To try to minimise such distortions maps can be drawn using different projections, where the distance on the ground is represented differently on a map. Maps can only be combined if they have the same projection. The map scale results from the combination of the scales of the data presented and the mapping choices. It is important to be consistent and to harmonise by using the same amount of detail across a map. The background should be relevant to the information to be communicated but also not be too cluttered.

Spatial displays can be summarised into coarser units through spatial aggregation. Administrative units have the advantage of being widely recognised on a map as well as summarising mapping results in terms of operational structures for decision-making. Attributes can also be aggregated by categories or by thresholds (for continuous variables) leading to different class memberships and so different outcome maps.

Aggregating data also helps to protect confidentiality, especially for health-related data and endangered species observations. The use of a big symbol is insufficient as the location can be easily deduced. To overcome this, some software packages are able to "jitter" positions (slightly reposition). What may also help is to remove identifying elements.

It is difficult to represent confidence intervals on the same map as the average values, and so they are rarely mapped. To find ways to integrate them when they matter is a growing interest. Possible solutions may involve adding two supplementary maps that show the low and high confidence bounds, or greying out areas that fall outside the confidence interval.

IMPLEMENTATION AND INTEGRATION

The final section focuses on integration and implementation, with the specific example of *Aedes aegypti* considered as a use case. The content is derived from a series of workshops run during the Action Annual Meeting in Rome in February 2023, where participants were asked to devise strategies to promote integration at the operation level for each of the surveillance, mapping and modelling components, using this Road map for guidance.

The main activities identified were first to establish situation by confirming the true presence/absence of the vector in the region of first detection, and to define sampling strategy needed to provide that information, including defining where to put traps using existing vector models or habitat suitability thresholds if available. This strategy is likely to be affected by the abundance of the target population, the weather condition during sampling and the sensitivity of the sampling methods used. The next step should be to use modelling to predict likely direction and extent of spread, using the predictions to define wider scale surveillance needed to monitor spread. Key to implementation success is communication – namely providing customised maps, charts and interpretations of sampling results, distribution models and predicted spread to local and regional stakeholders at the earliest opportunity.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Aedes Invasive Mosquito (AIM) COST Action (Project ID: CA17108), initiated in 2018, has three major objectives:

- To develop pan-European networking and collaboration in monitoring and surveillance of invasive *Aedes* species

- To increase preparedness and capacity to fight against invasive *Aedes* species by triggering optimization and innovation in surveillance and control strategies

- To disseminate, customize and communicate the AIM-COST Action outcomes.

This document is one of the outputs of *Task 1.2: Integrating surveillance data analysis, spatial modelling & mapping to ensure the quality and applicability of future technical outputs at the European level*. It therefore aims to contribute to both the first and third Action objective. Given that it involved direct input and feedback from the Network members it also addresses two of the Action Deliverables, namely Deliverable 3: *Synthesised feedback with recommendation for customising surveillance, control and modelling guidelines*; and Deliverable 4: *Guidelines for customisation of mapped outputs.*

In order to understand the impact invasive mosquitoes (and the pathogens they carry) have in terms of disease risk or nuisance, and to facilitate control and mitigation in the event of a new introduction we need to know where they do or might occur. This means we first of all have to sample and map where they are, especially when they first appear, and if they are already widespread, then they need to be sampled and mapped throughout as much of their range as possible. Likely changes in distribution caused by climate change means we should keep updating these maps.

Figure 1: Road Map Overview
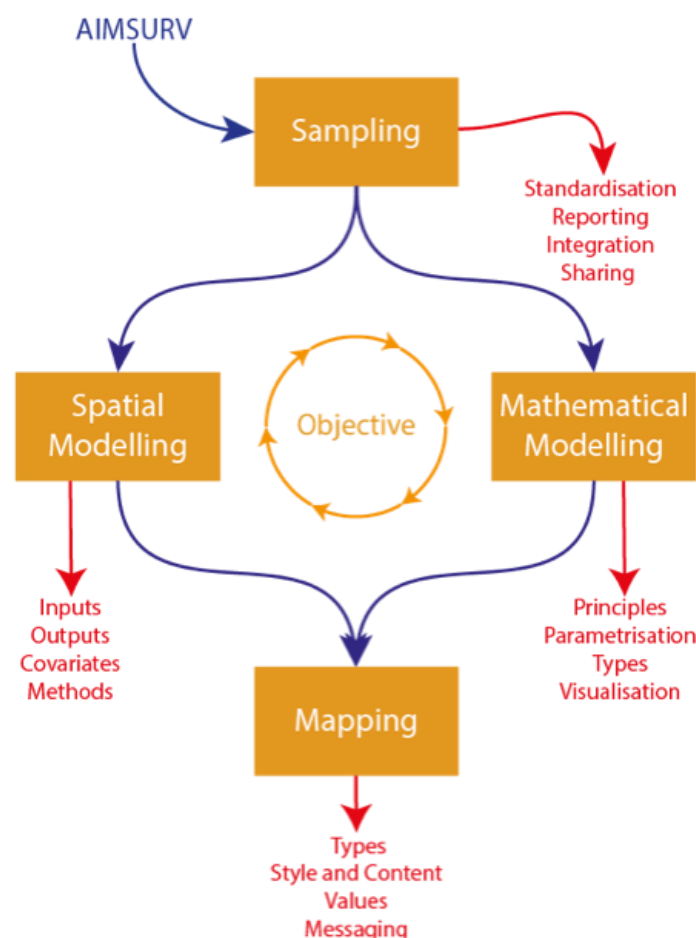
As sampling, by definition, leaves gaps between the samples in our maps, we also have to find ways to fill the gaps. Providing there is enough sample data to calibrate them, this can be done by using modelling techniques and by calculating habitat suitability based on environmental limits. By filling the gaps, the models give detailed distributions, and so are useful to identify likely areas to which the vectors can spread and establish, as well as when and for how long they are active during the year.

This document is intended as a general high-level overview of the entire chain from mosquito sampling through to modelling and mapping, and of the visualisation of the outputs. The overview, summarised in Figure 1, is not therefore an exhaustive guide on how to conduct each step, but rather an extended aide memoire containing recommendations for best practice aimed at a wide range of non-specialist stakeholders and professionals. It does not cover specific control methods nor address training needs. Throughout it is assumed that data are collected and processed to be part of an integrated chain of components – *e.g.* field sampling results will be, by default, mapped and modelled, rather than simply collected in isolation.

2 Sampling, data exploration and preparation

- **Sampling should be in line with the analysis and objectives**
- **Distinction between cross-sectional and longitudinal sampling**
- **Cross-sectional sampling is more appropriate for spatial modelling**
- **Longitudinal sampling is more appropriate for mathematical modelling**
- **All collected data will need to be curated and cleaned in line with the analysis and objectives**
- **AIM-COST provided a standardised sampling protocol via the AIMSURV initiative**

Any modelling and mapping work should start by precisely defining the objectives in the light of the spatial and temporal detail required, and what data already exist or can be collected from the literature or from fieldwork.

Before we can start any field work, we still need to determine a sampling strategy. Two elements may matter: the practicalities of sampling (for planning, it may be important to consider accessibility, whether dictated by road access or private land restricting access); and what you will consider affects representativity of your sample: how many locations should we sample and where should they be placed? You should make sure that the sampling provides the right data for the type of analysis you are intended to conduct. The more locations you sample, the more confident you can be that the findings from the samples truly reflect the situation for the whole region under investigation. However, the larger your sample size is, the more labour intensive, costly and time consuming your project will be. Finding a good balance between collecting enough samples to draw robust conclusions and optimising time, labour and cost effectiveness is sometimes hard to achieve.

There are several different sampling methods available, but the two main ones are longitudinal sampling (sampling the same points several times) and cross-sectional sampling (always sample different points). A more detailed description of sampling strategies can be found in the Appendix Section 7.1.

A key constraint is to ensure that any sampling is repeatable, and is sufficiently standardised to produce results that are comparable between different locations and times. Details of some of the factors involved in identifying sample locations are provided in Appendix Section 7.1.2. One of the important initiatives within AIM-COST has been to harmonize the sampling activities and protocols used by European entomologists. This was developed as the **AIMSURV initiative** (Miranda et al., 2022) which implemented a sampling programme involving standardised sampling activities across Europe, and also, subsequent data sharing and publication processes.

To account for differences in resources available to field teams, a set of minimum and a set of more resource intensive recommended sampling protocols were developed so that teams with fewer resources could still

contribute to a standardised sampling across Europe (see details in Appendix Section 7.1.3). More details on the different field surveillance methodologies for sampling different *Aedes* life stages, frequency and minimum length of sampling period, and data reporting can be found in the publication.

2.1 Basic data cleaning and curation

> • **Exploring data will help with interpretations later on.**
>
> • **The type of data preparation is determined by the type of modelling**
>
> • **Data cleaning can be a significant time sink**

Exploratory data analysis (EDA), provides a basic understanding of the data available. How good (that is, informative for answering the question we are looking at) or bad are the data? What needs to be cleaned? What data are available and what are missing? Are the data categorical, numeric, or something else? Are there any data points that are outside the main body of the dataset ("outliers") or are clearly anomalous in relation to the data nearby?. Much of this information can be obtained by simple examination of the data tables, or by visualising the data as maps, and is needed before the datasets can be cleaned to remove errors. Once cleaned, further preparation for the analysis and modelling steps can be made.

Data pre-processing can be very time consuming and there is no universal way to go about this. If you are combining different data sources you will need to do some data normalisation and standardisation to make sure the different sources are compatible and, for example, are correctly georeferenced, cover the same area, use the same units and are collected using compatible sampling methods. The more the data sampling and collection is standardised, the less cleaning is likely to be needed.

One thing you will almost always need to consider is data imbalances. If for example you have 1000 samples in one class and only 10 in another, it will create an imbalance in your dataset and it can bias your maps and skew your model. In cases like this, you may want to consider reducing the number of data points in the most common classes to produce a less biased sample. These techniques can also be very useful if your data are clustered in overlapping groups, which can also bias your analyses (Hendrickx et al., 2021).

2.2 Occurrence data

Depending on the sampling methods, the two main types of occurrence data produced by field surveillance or extracted from published literature are presence/absence and abundance data (*e.g.* counts per trap). Presence/absence data tend to be more robust and more widely modelled than abundance data as the accurate counts needed to provide abundance data for modelling require clearly defined and standardised sampling effort (*e.g.* trap type, trap set-up etc*.*) to ensure counts are comparable between different times and locations.

2.2.1 Absence data

As its name implies, presence and absence modelling requires both presence and absence records as training data. So, it is essential that absence data are recorded in the field. However, absence records are often not routinely recorded by field sampling, or if they are, they are frequently not reported in survey reports and publications. It is therefore sometimes necessary to create absence records without using field sampling. This can be achieved using different methods (Barbet-Massin et al., 2012).

The easiest method is to use simple geographical distance from known presences, whereby "pseudo-absences" are added in areas between a minimum and maximum distance from any presence location. This method assumes that there are no presences beyond the sampled range and so can be misleading, and the results should be used with care. A better method is to use some kind of logic-based definition of habitat suitability which identifies unsuitable areas based on known environmental limits (Figure 3) or identifies areas environmentally dissimilar to those where presences are recorded (using *e.g.* the

Mahalanobis distance) (Iturbide et al., 2015). Absences can then be assigned to the 'unsuitable' areas with some confidence.

Figure 2: Example of data used to generate Absences
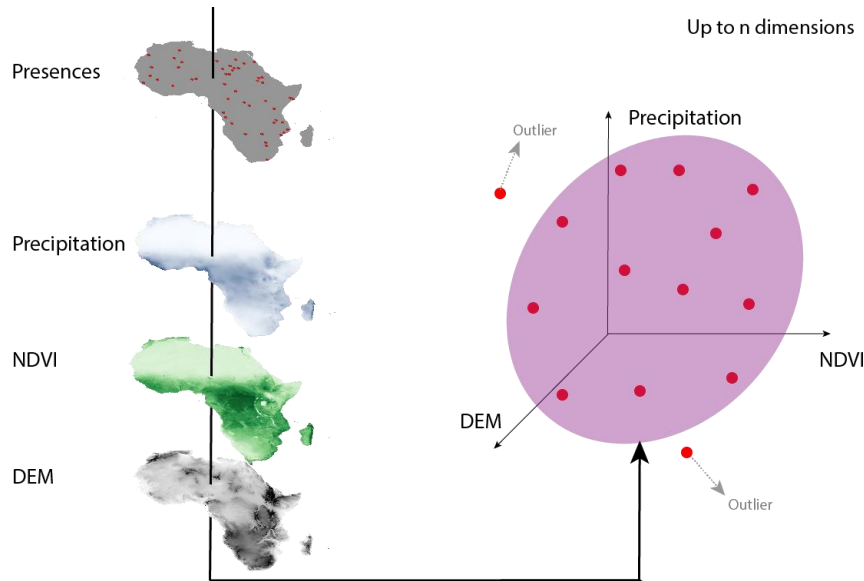
2.3   Passive surveillance

Sampling methods introduced above belong to the active surveillance category, which involves sampling by trapping. Complementary to trapping are passive approaches, which mobilize citizens, enthusiasts, students - any volunteer - to report a relevant sighting. Public engagement in surveillance and control comes naturally with passive surveillance through the establishment of communication channels between experts, decision makers, and the public. Citizen reports can be used to inform active sampling especially for invasive species, which rapidly spread and appear in unexpected distances. The approach has proven successful in surveying vector fauna, determining the extent of invasion by alien species, and assisting in the early recognition of changes that may have an impact on biting habits and the risk of pathogen transmission (Bartumeus et al., 2018; Kampen et al., 2015).

Passive surveillance has the advantage of generating data continuously. However, reliable data collection rests heavily on citizen engagement, which declines in time and correlates with geographic and socio-economic factors. Citizen activity must be maintained through communication and by providing feedback. Nevertheless, sampling bias is inevitable and must be addressed before attempting spatial or mathematical modelling. Palmer *et al.* (2017) presents a successful approach involving modelling user behaviour, and thus, removing the sampling bias (Palmer et al., 2017). Scalability of this approach, best practices of designing citizen science surveillance programmes, and standardizing the entire process are hot topics of study.

Citizen science data collection is included in the AIMSURV initiative and is recommended for integration into routine surveillance programmes for early detection and determination of spread for *Aedes* invasive mosquitoes (Južnič-Zonta et al, 2023).

2.4   Data reporting

As mentioned before, AIM-COST has aimed to provide a standardised framework so data (and outputs) can be shared and compared at a local, regional, and international level. In order to make sample reports

comparable from study to study, recorded numbers need to be standardised according to sampling effort, and the samples need to be adequately geo-referenced and time stamped. Importantly, it is increasingly accepted that zeros and absence records are just as important as presences or sample numbers

Many studies that attempt to assess species distributions – either presence or abundance - do so by extracting data from published reports and papers. This extracted information needs to be adequately georeferenced and standardised if it is to be comparable between studies and sample programmes. This is particularly true if the extracted information is abundance which requires a number per standard measure of sampling effort (such as number per square metre or number per trap per day) to be applicable across many studies and locations.

It is therefore important to ensure that the data are reported in a way that provides all the information that readers might need for downstream analysis. The minimum requirements for reporting can be summarised as follows: trap type/sample methods, number of traps/sample events, start and end collection dates, trap/sample geographical coordinates and location descriptors, number of specimens caught by species (including zeros), method of species identification. Details are provided in the Appendix Section 7.2.

3 Modelling

- A number of factors need to be considered before modelling is done:
    - Scale and resolution
    - Covariates

- **Both have significant influence on the results**

- **Two major types of modelling are: Spatial modelling for distributions; and Mathematical modelling for dynamics**

- **Both input data available and aim will determine what type of modelling is possible and appropriate**

- **Mathematical modelling is more appropriate for dynamics**

- **Spatial modelling is more appropriate for distribution**

A number of issues should be considered before any model is run – in particular the extent of the area to be modelled, the scale of the data, what level of detail ("resolution") is required for the outputs. Note that these are issues of relating to the data inputs available and what needs to be modelled. Scale and resolution also affect the visualisation of the data, and are also therefore discussed in those contexts in Section 4, below.

3.1 Scale and Resolution

We all have an intuitive understanding of scale but it is worth formalising our understanding to distinguish between data scale discussed here, and map scale discussed later in this document. In the broadest sense, geographers define scale as including both extent (how much of the Earth's surface is covered) and resolution (in how much detail, what is the smallest element represent on the map) (Figure 3). Though we focus here on the spatial scale, the same principles can also be applied to temporal scale.

Figure 3: Different scales and resolutions
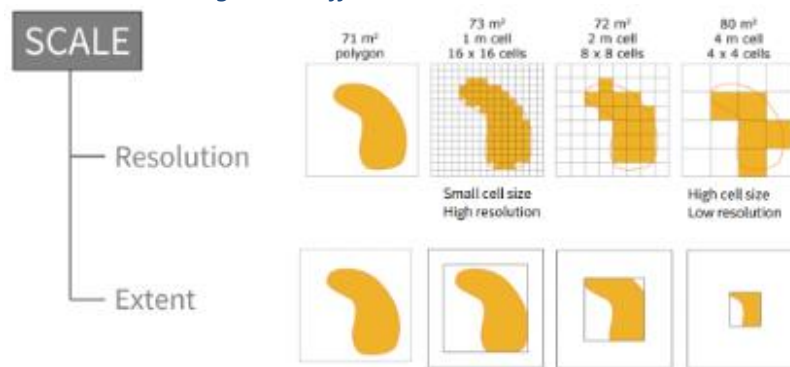
3.1.1 Data scale

It is worth establishing that data themselves have a scale, and that how these differ between vector maps (point or polygons) and raster images built from pixels. The well-known map in Figure 4 covers the full extent of Europe and its neighbouring countries, and with an intermediate resolution based on administrative units. Though convention would usually require a scale, this map has no need of one because it is self-explanatory (everyone can "size" the European continent and its provinces, and this map is not going to be used to assess distance. It is worth noting that using administrative units may however conceal how spatially complete the observations are: *e.g.* mapping mosquito presence in the whole of Belgium could be based on trapping in Antwerp only.

Figure 4: VectorNet AIM map

Source: https://www.ecdc.europa.eu/en/disease-vectors/surveillance-
and- disease-data/mosquito-maps
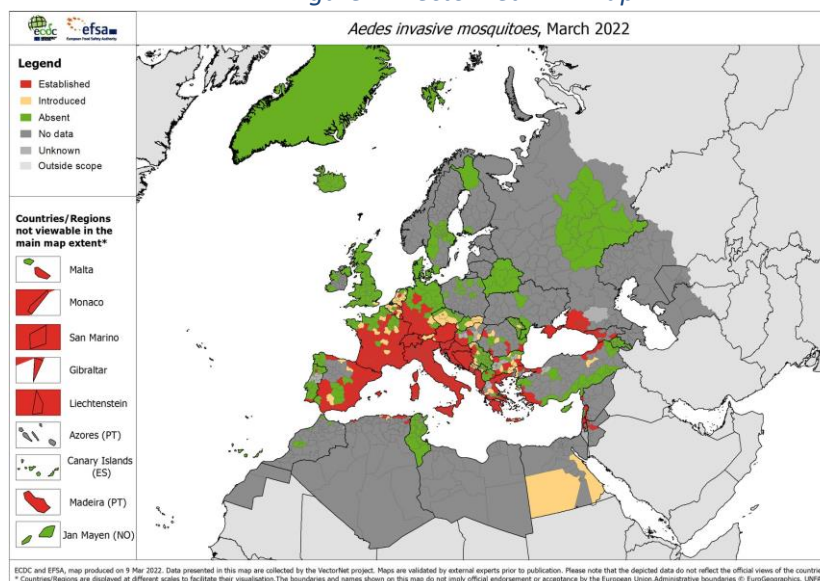
In the case of pixel-based data, resolution is given by the pixel size. The spatial resolution of the covariate data, usually used as raster/pixel-based data, will determine the resolution of the model output. Care should be taken to ensure that the resolution of the covariate data is appropriate for the input training data (Chauvier et al., 2022). Other aspects of resolutions are discussed in the data visualisation section.

3.1.1 Data aggregation

Sometimes the data available for mapping or modelling are very unevenly distributed, perhaps because they are collected from a wide range of different surveillance programmes or from samples taken at different times. This can lead to data clusters where several data values are very close together or even

overlapping, whilst in other places, they are widely dispersed. In these cases, it is often useful to aggregate the data as samples repeated over time may provide different values because of the difference in trapping conditions like the weather. An example of this process would be to take all sample records within a 10 or 20 sq. km. grid cell and take the maximum recorded number during the year. Further aspects of spatial aggregation for mapping are discussed in Section 4.4.
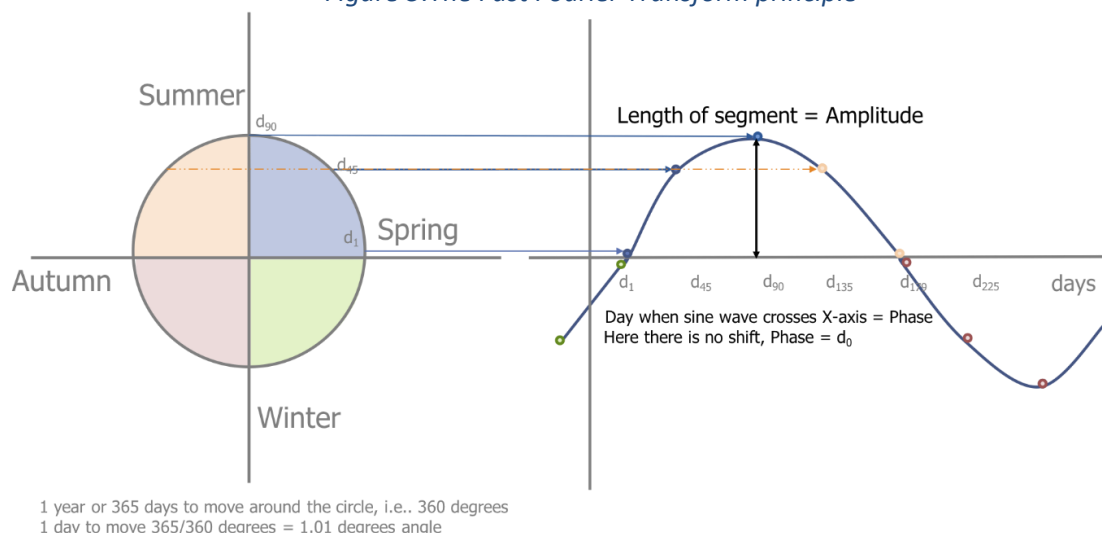
3.2 Covariates

- **Covariates are the explanatory values that are used to fill in the gaps**
- **Covariate selection has an influence on the modelling result**

3.2.1 Covariate selection and processing

Species distributions are determined by abiotic and biotic factors and species distribution models frequently use the abiotic environmental factors as predictor variables (Elith & Leathwick, 2009). Commonly used covariates for mosquito distribution modelling include things such as NDVI (normalized difference vegetation index), temperature, rainfall etc. A list of useful covariates and links to the sources is provided in the Appendix Section 7.2.6.

In recent years, many environmental datasets have become very extensive and complex. For example, temperature datasets may be available for every 500m of the globe for every day for many decades. This makes it difficult to choose which particular dataset should be used as a predictor, and so it is often necessary to use data reduction techniques to provide summary statistics. Whilst means, minima or maxima can be useful, there are other techniques that can also provide indices of seasonality. One such are bioclimatic summary variables based on temperature and rainfall data – such as temperature of the warmest quarter which are readily available online. A potential problem with such data is that the individual parameters are often correlated – vegetation is usually linked to temperature and rainfall – which interferes with the statistical modelling processes.

Figure 5:The Fast Fourier Transform principle

1 year or 365 days to move around the circle, i.e.. 360 degrees
1 day to move 365/360 degrees = 1.01 degrees angle

An alternative is to use temporal Fourier analysis (Figure 5) which provides biologically relevant indicators and removes the inherent correlations between different environmental parameters. These are however complex and time intensive calculations (Scharlemann et al., 2008) and it may be preferable to find out whether GIS colleagues have these datasets prepared rather than trying to produce them yourself.
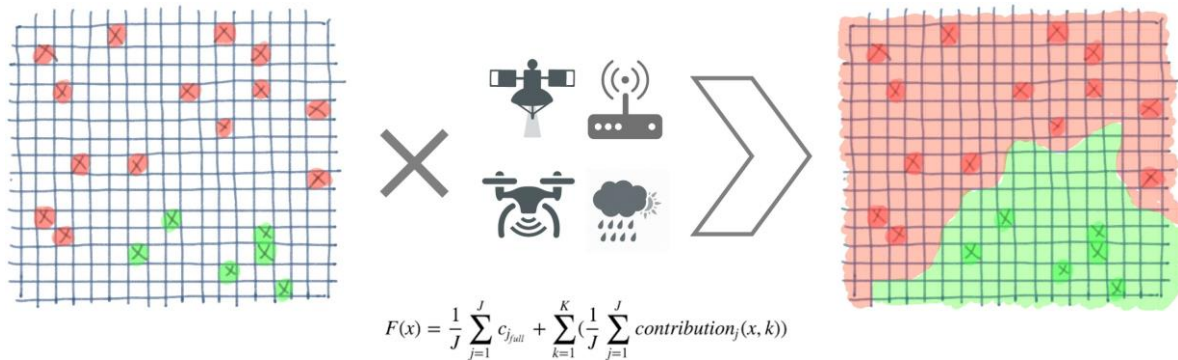
3.3   Modelling Methods

Modelling in this context can be conducted in either of two ways: Spatial modelling which is based on the statistical relationship between the target vector and a series of covariate drivers; and mathematical modelling which is based on measured relationships between the target vector and defined processes like growth, birth or mortality rates. Spatial modelling usually produces maps, whilst mathematical modelling typically provides graphs of population changes over time.

3.3.1 Spatial modelling

Spatial distribution modelling consists of two components: The species data, often drawn from sampling or extracted from the literature, and the covariates used as predictors. The modelling process first establishes a relationship between species data and these predictors. The species data often consist of the presence or abundance of the target variable for a series of sample locations. The modelling then applies those relationships to all areas for which data are not known (Figure 6) (Araújo & Guisan, 2006).

The output maps of these species distribution models are usually static maps and do not reflect temporal or seasonal changes. Proxies for these dynamic factors can, however, be derived from proxies like length, peak or start of season (Petrić et al., 2021).

Figure 6:Species distribution modelling concept

$$F(x) = \frac{1}{J} \sum_{j=1}^{J} c_{j_{full}} + \sum_{k=1}^{K} (\frac{1}{J} \sum_{j=1}^{J} contribution_j(x, k))$$

3.3.2 Modelling method selection

There are a plethora of models to choose from (Figure 7). All these different models have advantages and disadvantages. Depending on your data, area of interest and covariates, some models will perform better than others. It is often advisable to test different models to see what works best in your particular case (Uusitalo et al., 2021). Statistical and machine learning models are most commonly used (Beery et al., 2021).

Statistical models focus on inference, creating a mathematical representation of the data generation process to understand the behaviour of a system or test a hypothesis and to distinguish between a genuine effect and noise. To specify the model, identified parameters of special interest are used. Assumptions can be verified when enough data are available and if the model is refined enough. Statistical models take uncertainty into account.

In contrast, machine learning models concentrate on finding an unobserved outcome or future behaviours, by using algorithms to find patterns in the data. This method is useful when dealing with 'wide data' where the number of input variables exceeds the number of subjects. Machine learning is more empirical and data-driven which means more complex relationships can be included in the modelling. At the same time, machine learning does not attempt to isolate the effect of a single variable (Merow et al., 2014).

Figure 7:A graphical representation of different data model families
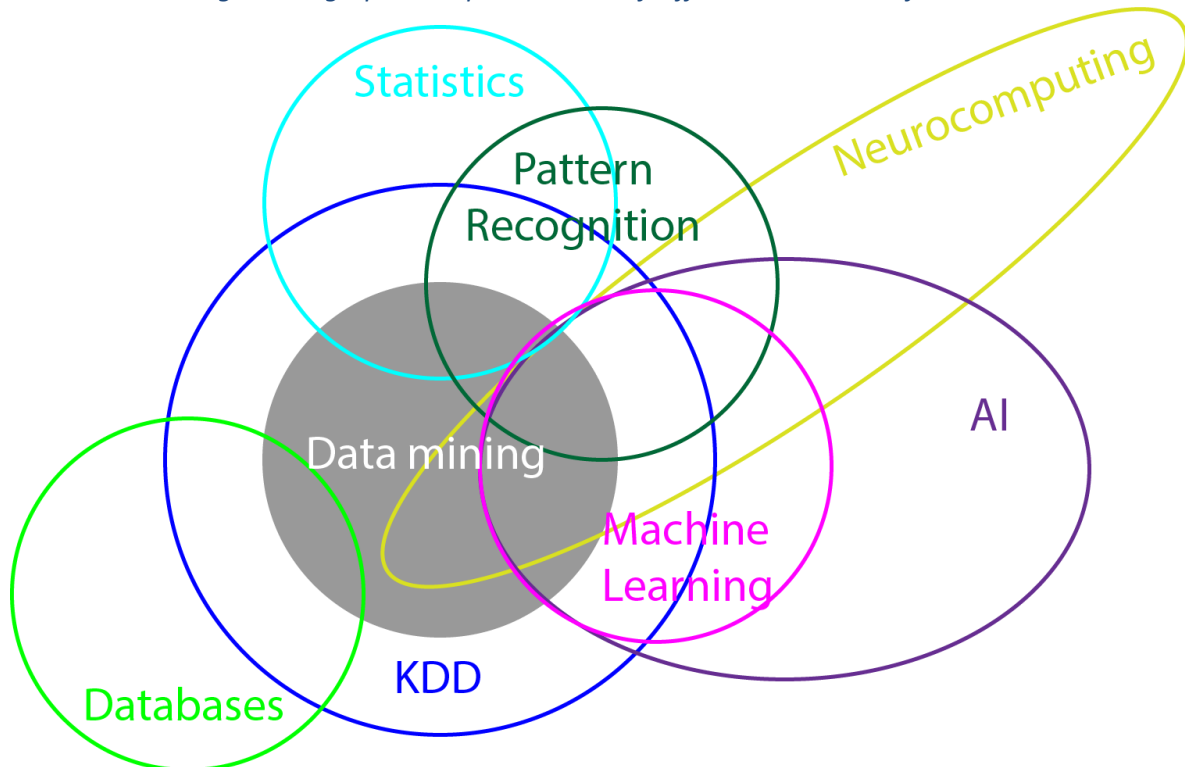
(AI=artificial intelligences, KDD=Knowledge Discovery in Databases)

Statistical models are the better choice if the signal/noise ratio is small, the model needs to be interpretable and if the sample size is smaller. These are a good choice when the effects of a variable or a small number of variables need to be isolated. Machine learning is also ideal when you are not interested in (directly) estimating uncertainty. A few examples of the most common used models include: Non-Linear Discriminant Analysis and Generalised Linear Model (GLMM) as examples of statistical models, while Random Forest and Boosted Regression Trees (BRT) are examples of machine learning models (Merow et al., 2014). Further details are given in the Appendix Section 7.3.

3.3.2.1 Bootstrapping

Spatial modelling is a statistical process and so it is not enough to run a model just once. Uncertainty analysis can be conducted by running many repeats of each model to assess which areas are reliably predicted and which are not and would benefit from additional data. Bootstrap sampling is one of several methods normally used for testing the accuracy of predictive models. After the data are cleaned, you typically split your data in 2 to 3 subsets. The largest subset will be the training sets (70–80%). This set is used to train your model. A second subset, the validation set (10–15%), can be optional and is used to validate and tune the model outputs. The last subset is the test set (10–15%) and is used to evaluate model performance. The reason for splitting the data is to have independent sets at each stage. If you would use the same sets of data to train and test the model you will get a self-fulfilling prophecy.

A bootstrap sample is the sub-sample of a set of training data that is used to make one prediction of a species' distribution. Multiple bootstrap samples are taken. A prediction is then made for each and finally, the entire set of predictions is combined to produce a single, average prediction. Bootstrap samples are taken from the training set with replacement because we assume that the training set itself is a sample of reality, and the occurrence of any one observation within it is essentially random. A different training set could contain that observation once, more than once, or not at all. One advantage of bootstrap sampling with replacement is that within any one model the samples can be arranged to have equal numbers of presence and absence observations. Recent work suggests that this situation produces model outputs with the greatest accuracy.

3.3.2.2 Ensemble modelling

Bootstrapping occurs within the same model but you can leverage this and combine different model methods into an ensemble model that usually performs better than the individual models (although this is not always the case). As an example, BRT tends to over fit but GLMM tends to smooth out the infrequent, but likely, observations. So, a combination of both is likely to avoid these flaws. Ensemble modelling creates a more robust model result, but at the trade-off of increased complexity and calculating time (Hao et al., 2020).

3.3.1 Mathematical models

> • **Mathematical models represent biological and ecological systems as a set of equations.**
>
> • **Model calibration is performed using longitudinal observations collected regularly at the same location over an extended time period.**
>
> • **Combining complementary datasets, such as meteorological data and experimentally derived physiology, is informative, when compatibility and applicability are ensured.**
>
> • **Multiple data sources, including experimental and field data, are needed to develop mathematical models. Often, such data do not exist or are not publicly available.**
>
> • **There is a need of a standard language of model definition to improve accessibility and reproducibility.**
>
> • **Demand for computational resources increase as models become more complex and data accumulate.**

Mathematical models are idealised representations of biological and ecological systems formulated into a set of equations. In contrast to spatial modelling, mathematical modelling is process based, or mechanistic. While spatial modelling relies on statistical relationships of predictors with species' presence or abundance, mathematical models rely on measured or experimentally determined relationships with developmental or population processes as are so much more complex to construct.

For mathematical models, the essential components of a system are specified in a formal framework and interactions among these components and links to internal or external drivers are defined without ambiguity. For example, temperature-dependent adult *Anopheles gambiae* mortality can be modelled with a polynomial equation, $aT^2 + bT + c$ (Lunde, Bayoh, et al., 2013). The symbols *a*, *b*, and *c* represent parameters, values of which are justified by laboratory experiments or field observations. The modeller then employs analytical and numerical methods, rooted in well-defined mathematical theories, to investigate the system *in silico*.

Good models do not always need extensive detail, but merely to offer useful insights into the system. The level of detail and the extent of interactions formulated in a model depend heavily on its purpose. Mathematical models of vector populations and vector-borne disease transmission can be developed to address, among others, population response to changing environmental conditions, the environmental and climatic drivers of a species, and the impact of control interventions.

For instance, the malaria transmission model of Ross and Macdonald, with only two components (the prevalence of malaria and the rate of humans acquiring infections), shows that a reasonable reduction of mosquito population below a threshold is sufficient to interrupt transmission cycle (D. L. Smith et al., 2012).

3.3.1.1 Model calibration and identifiability

Although a mathematical model needs to be well-defined, the natural biological or ecological process it represents can be complex and largely unknown. A model is inevitably a coarse approximation compared to the endless complexity of the actual system. The process of model construction, re-structuring knowledge into a set of well-defined abstract formulas, helps to critically evaluate observations and identify gaps in knowledge. In addition to the limitations of *in silico* representation, several factors may contribute to uncertainties in model output and deviations from observations. These include measurement error, chance (expected randomness of a natural process), and errors in defining relationships (parameter identifiability).

Measurement error is quite common in life and environmental sciences, and is minimised through training and practice. Random or chance variation in natural processes is addressed though stochastic models (see Appendix Section 7.4). These account for the expected variation in, for instance, development times and survival, and generate a slightly different number of adults emerging at a slightly different time in each model simulation. Overall, stochastic models better represent the range of outcomes of laboratory experiments or field observations.

Limited parameter identifiability is the hardest to deal with, as it may require extensive data collection and expert validation. The parameters of a model, e.g., a, b, and c in the equation of adult *Anopheles gambiae* mortality, may be derived from laboratory experiments in controlled conditions, though the results are difficult to translate to field conditions which are often more spatially and temporally variable.

Most state-of-the-art gridded meteorological datasets used to apply the models are available at resolutions measured in kilometres so are often difficult to reconcile with micro environments the vector actually inhabits. Data loggers may be used to obtain measurements at collection sites but mobility of adult mosquitoes may enable them use significantly different micro-environments than the ones around the measurement sites.

Certain model parameters may be unavailable or difficult to identify by laboratory experiments. Examples include (i) the average volume of breeding pools, (ii) extent of population reduction following the administering of vector control, and (iii) mosquito biting rate. Such cases pose an inverse problem where a subset of parameters, or the entire parameter set, is adjusted to make the model output match observations (Erguler & Stumpf, 2011; González et al., 2016; Koons et al., 2017). Any observation that can be compared with a particular model output can be used for calibration, such as (i) number of eggs in an ovitrap, (ii) number of adults caught in an adult trap, and (iii) number of bites reported by a person.

Longitudinal data are the preferred means of calibrating a mathematical model. Such data comprise field observations performed regularly at the same location for, ideally, more than one season with the same protocol. By doing so, potential impact of an environmental variable, current or historical, on the dynamics of a population can be captured. The frequency of collections and the number of collection sites are planned to ensure capturing transient changes and resolving the differences in neighbourhoods and land types. Collection of longitudinal data is expensive as it requires designing sampling strategies to sufficiently sample from a single population. It is advisable when performing field observations to (i) combine different life stages, (ii) collect from various environmental backgrounds, (iii) extend data collection beyond the anticipated time frame of an active season, and (iv) aggregate observations in multiple, ideally subsequent, years (Erguler et al., 2019).

3.3.1.2 Risk mapping

Mathematical models may represent spatial and temporal dynamics of vector populations including adult activity and abundance, duration of the peak season, and first emergence of adults. They may be used to anticipate disease transmission by explicitly representing vector-host interactions and mobility. One can perform risk assessment by exploiting predicted abundance, biting activity, force of infection, and the expected number of disease cases. Such indicators can be summarised into future projections, for the dynamics, and maps, for the geospatial distribution, of risk. Climate- and environment-driven mathematical models have been developed to represent local vector populations and perform risk assessment mainly at the scale of cities (Annelise Tran et al., 2019; Guzzetta et al., 2016; Tran et al., 2013). Extending applicability towards larger scales requires incorporating variable environmental factors and breeding site dynamics into

population dynamics models. In addition, compatibility of laboratory-derived physiological parameters in the context of large-scale land and climate conditions also needs to be ensured. Recently, Erguler et al. proposed re-calibrating all parameters (including laboratory-derived dependencies obtained at fixed micro-climate conditions) with respect to field observations as a means to develop large-scale models (Erguler et al., 2016, 2017). Nevertheless, there is a large room for improvement concerning accuracy and applicability as more data accumulate and better methods are developed.

3.3.2 Modelling tools

Several tools and software are available to conduct the types of work and analysis discussed in this document. These tools range from scripting to full-fledged software packages.

VECMAP is a full-fledged software package that offer the full chain of sampling to species distribution modelling with a Graphical User Interface (GUI) for those that do not want to deal with scripting. If you don't mind scripting then R is a good option. The R scripting package offers several libraries for species distribution modelling, such as the caret and biomod2 packages, and mathematical modelling, such as the pop, stagePop, albopictus, and dynamAedes packages. It is convenient as you can also do all your data preparations and covariate creating in R as well. There are also other software packages that are capable of doing part of the processing like image manipulation (IDRISI) or general GIS aspects and spatial data visualisation (QGIS).

In contrast to the situation with spatial models, many of the published mathematical models do not describe how to run them and very few provide the code used, or have been developed into software tools (Ryan et al., 2022). Code availability is essential to ensure reproducibility of the entire model development process and should be accompanied by the exact model framework, parameters used, and any tricks the authors used while transferring the model into code (such details may not be included in model description). To ensure this, a common unified all-encompassing formal language of model definition is needed. The Systems Biology Markup Language (SBML (Hucka et al., 2019)) commonly used for biological modelling could be given as a template, and could be accompanied by a comprehensive curated model repository similar to the BioModels Database (Li et al., 2010) to promote visibility, facilitate adoption, and encourage validation by fellow experts.

4 Data visualization and mapping

- **Maps are communication devices and need to be elaborated considering what the purpose is: what is the message or information we are trying to bring across?**

- **Common sense is very useful for elaborating maps, but cartography rules and practices also help avoiding confusing the message or misleading – even inadvertently – the reader**

- **How the map will be diffused (on paper or online) offers various options and what works in one context may not be ideal in another**

- **All maps are a simplification of reality, that may also include a spatial bias. Generalisation is necessary for conveying a clear message, but sometimes comes at a price.**

There are many reasons to produce a map as they are very efficient ways of rapidly displaying complex information about a single parameter. Many analysts use them to validate data entry and to locate outliers or anomalous values which may represent errors. They are most used however as ways to

convey information to a wide range of technical and lay audiences. As for text or graphs, how maps are presented will affect how understandable and useful it will turn out to be. This section covers some basic principles that are useful for using maps to communicate, plan and make decisions, and then describes some situations typical for entomologists. Everything presented here is software-independent and the information presented is meant to be of use whether you produce maps yourself or not.

Like any document stemming from scientific information, maps are generally meant to communicate a message, but as a summary to facilitate discussion and decisions or disseminate information. While a map can be as basic as visualising data for organising a field collection campaign or to explore data, the purpose will always be to try to allow interpretation: the data (such as mosquito numbers at a set of locations) are the measures you can compile in a table, and the interpretation such as risk assessment or a control strategy will be the conclusions drawn.

More specifically, we could map presence/absence in November, or date of peak abundance, or any other element. This nuance matters because when making the maps we must be aware of the information, the message we are trying to convey. They all potentially lead to different information being conveyed by the map. What we need to keep in mind: there is no single way of mapping: the cartographer bears the responsibility of presenting something that allows the reader to draw relevant conclusions.

The following sections consider the main topics for which tried and trusted conventions exist that help present the data, convey information and help the reader to grasp the message that the mapper intended. We only cover broad topics here, so some additional resources are list in the Appendix.

4.1   Map projection, scale and generalisation

The Earth is (roughly) spherical, and paper or digital maps are flat. Any flat map therefore distorts the actual distance in one or both directions. To try to minimise such distortions maps can be drawn using different 'projections', where each projection represents the distance on the ground differently on a map. This is a very complex field, and there are thousands of projections, each defined to minimise distortions in particular situations (countries, small or large areas etc). Each looks different on the page (see Appendix Figure 22) and so maps can only be combined if they have the same projection. It is therefore essential to know which projection a map uses. Further details are given in Appendix Section 7.6.1.

Figure 8: Effect of scale (Source: Eurostat country maps)
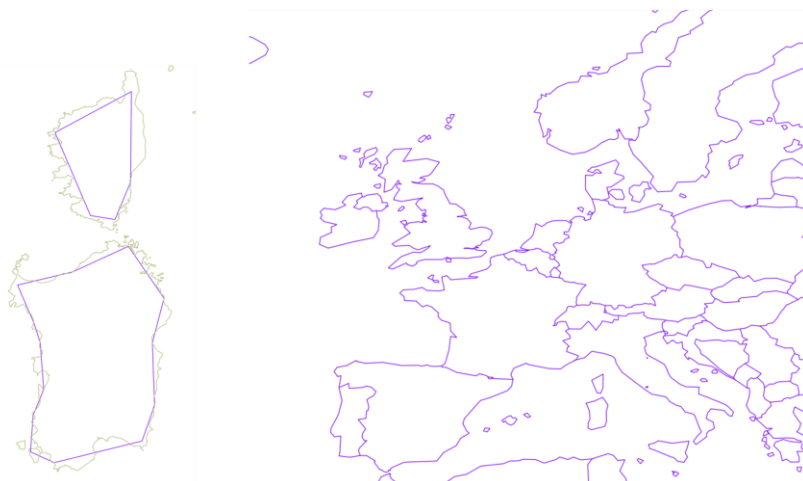
Data scale was covered previously in this document. The map scale results from the combination of the scales of the data presented and the mapping choices. Sometimes scales are shown as a fraction, which remains a useful notion because it relates coarsely to levels of detail. However, because we now view such documents most often on a digital screen, the fraction is either no longer useful and replaced by a linear scale, or needs to change as we scroll and zoom in and out, as we can see on platforms such as Google maps. Figure 8 provides two examples of scale effects: Corsica and Sardinia look very different in these two maps where one is at a scale of 1/1000000 (green) and the other at 1/60000000 (purple).

Both are the Eurostat world country map and when zoomed out to the whole continent the latter scale is sufficient.

In relation to scale, we must recognize that maps always are and should be a simplified representation of reality. What data should then be shown to convey the intended message, and what data can be omitted? Distracting, irrelevant, or unreadable elements have no place on a map, so if we map Europe, for example, we do not need to include in the map the thousands of minuscule, uninhabited islands off the coast of Sweden. Remember it is important to be consistent and to harmonise by using the same amount of detail across a map.

4.2   Spatial bias and representativeness

One element that is not explicitly covered in scale but that may matter very much is whether the information presented is equally representative for an entire map: for example our dataset may be focused on residential areas and therefore ignores less built-up areas. This is what sampling strategies discussed in Section 2 aim to cover. Also, how were residences and neighbourhood selected? The sampling strategy description should make such details explicit. In the case of citizen science contributions (*e.g.* Figure 9) which are less controlled than formal sampling programmes, the potential for spatial bias is much more difficult to grasp and difficult to account for. People's concern for the matter or interest for science may strongly affect the pattern of observation, so it is uncertain whether Figure 9 provides and exhaustive view of *Ae. albopictus* distribution in Europe.

Figure 9: All reported observations of Aedes albopictus *on iNaturalist.*

Source:https://www.inaturalist.org/observations?place_id=any&subview=map&
taxon_id=62984

4.3   Variable types

Data types that can be mapped include:

- Qualitative: nominal or categorical (*e.g.* urban, pasture, field, forest); ordinal (*e.g.* low density, medium density, high density)

- Quantitative: stock or absolute (you can sum values); ratio (the sums do not make sense).

The spatial data (*e.g.* trapping locations) have associated numerical or categorical attributes which may include their geographical coordinates, dates of trapping, trap type, person trapping, number of mosquitoes collected. Any parameter (like model outputs) can be an attribute as long as they are linked to the spatial entity by identifiers or geographical coordinates.

Many published maps include a default background proposed by GIS software – such as topography or roads and settlements. Very often this information is either too detailed and obscures the main data displays or

represents irrelevant information that has nothing to do with the data in the map. Use these default layers with care.

4.4   Aggregation

In the preceding sections we have discussed various reasons to aggregate data such as improving statistical clarity, to improving visibility, or to maintain confidentiality (see below). Two sorts of aggregation are widely used – spatial and attribute aggregation.

For spatial aggregation the idea is to summarize the spatial displays into coarser units - such as a series of point records of population number into population totals for a municipality or province. Administrative units are a common unit of aggregation which have the advantage of being widely recognised on a map as well as summarising mapping results in terms of operational structures for decision-making. Here the main caveats pertain to the potential variation that various aggregation options will produce on a same dataset as is shown in Figure 10 below. There are risks involved in such manipulations – an example being "gerrymandering" in which electoral district boundaries are adjusted in relation to likely voter intentions to bias and election result.

Figure 10:Spatial aggregation
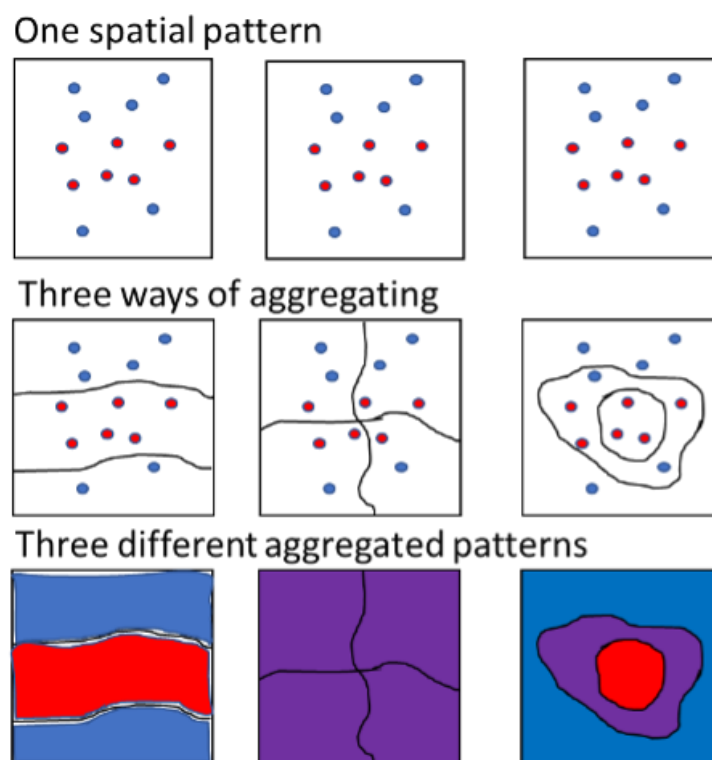
Adapted from Lambert & Zanin, 2016

Attributes themselves can also be aggregated by for example grouping categories. The nested legend of Europe's CORINE land cover map is such an example: the separate coniferous, broad-leaved and mixed forest categories can be aggregated to an overall category of "Forests and semi-natural areas". Continuous variables can also be 'binned' in various ways using category threshold intervals such as quartiles, equal intervals or manually defined boundaries each of which will result in different class memberships and so different outcome maps.

4.5   Map types and content

Depending on the attribute (variable) mapped and the objective of the map, various tools can be used. In broad terms, variables representing stocks (absolute numbers, *e.g.* total population, number of schools, number of disease cases…) are mapped using symbols of varying size. For ratios (proportions, e.g. population densities, schools per inhabitant, disease incidence rate…) choropleth maps are more

appropriate. These are maps in which an entire area, typically an administrative unit, is given a colour. Stocks mapped as choropleths is possibly the most widespread mistake made in cartography (see more details and an example in the Appendix Section 7.5).

Once the map type has been selected, many options exist for dressing the map: the use of colour and texture can help (or hinder!) readability of the message. A summary of the options for graphic semiology (as cartographers call their palette) is found in Appendix section 7.5. A clear title, legend, and graphic scale will all contribute to convey the message clearly and convincingly.

4.5.1 Mapping for typical situations

A few typical scenarios in which a map may be deemed helpful include: planning fieldwork for surveillance, planning monitoring for interventions or management of decision making. To take the first example – there are a number of questions that need to be answered before the field work can go ahead – these include what variable to measure, and what spatial and temporal scale to adopt in the sampling – these have been discussed earlier in the document. Once these have been addressed, a number of scenarios may present themselves to be mapped, such as: Where has been surveyed? and How do the results compare with others?.

4.5.1.1 Where has been surveyed

In these cases, simply mapping sampling locations may be enough. The spatial scale needs to display the area or interest: city, province, country. In the latter case, when the scale is broader, aggregating some of the points to more generalised locations (*e.g.* displaying sampled cities rather than individual sampling points within cities) may help clarity. It is important to decide what maps to use as a background to this information- some alternatives are road network, administrative boundaries, or land use type – as the background should be relevant to the sampling strategy but also not be too cluttered.

If you intend to cover neighbourhoods of various socio-economic levels, mapping your sampling over *e.g.* census based mean income data could be useful. If you are more concerned about building transects of increasing distance from the potential source points (*e.g.* a harbour or logistics centre), then showing those elements on the map will be the primary objective.

Figure 11: Sample location maps

Figure 11 provides a detailed example: Map A (left): Three zones are delineated to prioritise the surveillance of container-breeding mosquitoes: blue = very important (inner circles, up to 500m), orange = important (centre circles, up to 1000m), orange red = less important (outside circles, up to 1500m). Green areas = forests; cyan dots = all urban sites/units within the target municipality that could be sampled; yellow triangles = rural sites/units. Map B (right): Some urban sites were randomly selected (dark and light purple dots for municipality X, orange dots for municipality Y); most of the sites were sampled (light purple and orange dots); traps were placed in forested sites (green dots) and along rivers (blue dots).

4.5.1.2 How do our results compare with other results?

Multiple indicators are always challenging. The map should not be overcrowded and the data from each set of results should be comparable (was the same collection method used?). If the data are comparable then showing the difference or the trend could be useful, depending on what the key message is.

This scenario raises an important question, valid elsewhere as well: is a map the answer? A map is only necessary or useful if we consider that spatial heterogeneity matters. If one trap or model, or the aggregated (or maximum value across places) observation or prediction provides the information needed, then a map is unnecessary and a graph following the progress of values may be more useful. In this context, comparisons often involve the use of confidence intervals to validate whether differences are real. Confidence intervals are difficult to represent on the same map as the source values, and so are rarely mapped but efforts are increasingly being made to find ways to integrate them when they matter. It may involve adding two supplementary maps that show the low and high thresholds for the chosen confidence interval, or greying out areas that fall outside the confidence interval.

4.5.2 Communication

Whatever the audience, you must first identify the message you want your map to convey, and then how the map can help you get that message across. The map needs to do that by itself, *i.e.* be clear and easy to read. You will need to remember most of your readers are unfamiliar with the data and in some instances, may be unfamiliar with the issue.

Figure 12 shows an example of communication by the French Ministry of Health to the general public. It shows which "Départements" are known to have established *Aedes albopictus* populations.

Figure 12: Tiger mosquito map, French Ministry of Health
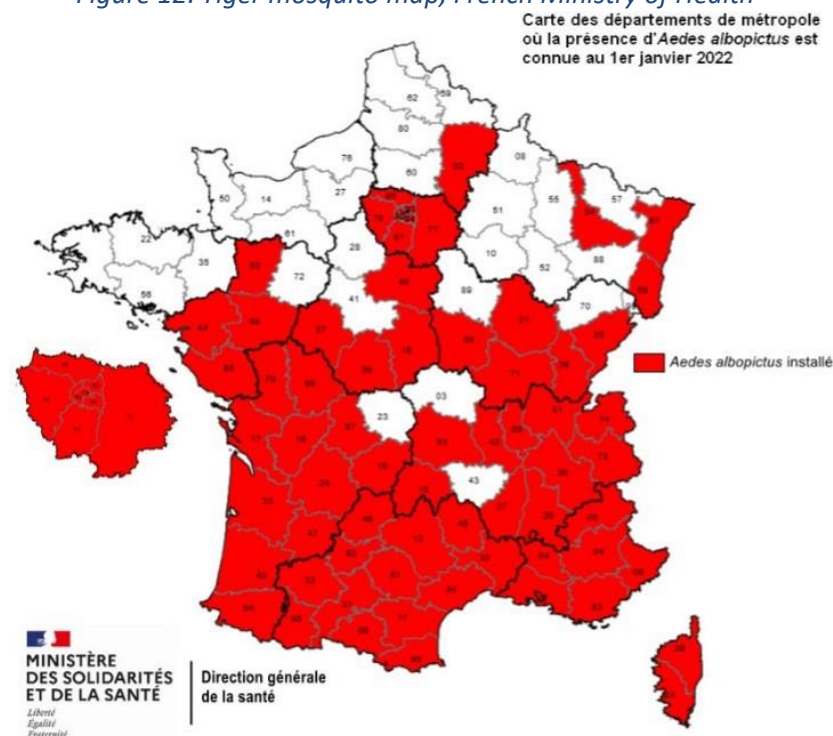
There are a number of comments we can make about this map:

- It focuses on the "metropole" (so does not include any Overseas Territories).
- The legend has one item: *Aedes albopictus* is established in a given department.
- It has a map inset on the left, giving more detail on the region of Ile-de-France (Paris and suburbs)
- It has no detail inside the Départements.

- It has no distance bar or north arrow so it is assumed that any reader is familiar enough with metropolitan France to do without.

This map therefore fails to provide some details that are formally necessary, but largely succeeds in conveying a simple message: the Tiger mosquito is now found in numerous areas (67 out of 96 Départements). It leaves out the detail of precise location where it was observed, thus leaving aside issues concerning sampling effort, population density, recent introductions and spread. It also does not map the other departments as "absence" thus implying that the situation can evolve and/or that the mosquito may spread.

4.5.1 A word on confidentiality

When data are mapped, the risk of a breach in confidentiality increases substantially. This is an obvious concern for health-related data, but there may be other reasons to "blur" the exact position of a mapped element, so as not to reveal the exact location of sightings for very rare species, or to protect equipment from theft or vandalism. The use of a big symbol is insufficient as the location to which it is fixed can be easily deduced. To overcome this, some software packages are able to "jitter" positions. Removing identifying elements may help. Aggregating data (*e.g.* mapping admin units rather than points) is also a good option.

5 Integration and implementation

Whilst the previous sections provide an aide memoire of activities needed for effective surveillance, mapping and modelling of invasive mosquitoes, they do not suggest when to implement these activities in a real-life context. To identify these implementation steps, breakout discussions were held at the final AIMCOST network meeting in Rome, February 2023, to define what activities are needed when an invasive mosquito is first detected in a region. The recent detection of *Aedes aegypti* in Europe is used to provide an example. The list of activities described below have been distilled from the group discussions about the **integration and implementation of surveillance field activities and use of modelling and outputs such as maps** to improve sampling design and communication.

A number of component activities were defined and are expanded in subsequent paragraphs:

- Establish situation: Confirm true presence/absence of the vector in the region of first detection, and define sampling strategy needed to provide that information, including defining where to put traps using existing vector models or habitat suitability if available

- Use modelling to predict likely direction and extent of spread, using the predictions to define wider scale surveillance to monitor spread; to validate model in field

- Communication: To provide maps of sampling results and distribution models to stakeholders

a) Establish true presence/absence, and define sampling needed to acquire that information;

After the initial detection of *Ae. aegypti* in a particular area, the first recommended step is to confirm the presence and spread of *Ae. aegypti* around the point of detection, *i.e.* to establish whether the record represents a true presence, and over what area the vector is present. This is more likely to be achieved by sampling of the immediate area rather than modelling (Lana *et. al.* 2018), unless local distribution models are already available (se also below).

Sampling effort needed to determine a true "presence" or "absence**"** (*i.e.* The number and locations of either oviposition or adult traps, needed to determine a true "presence" or "absence") requires specialist entomological knowledge to define and would depend on the likely mosquito abundance, conditions during the period of sampling, and the specificity and sensitivity of the sampling methods,.

    I.   **Abundance and extent of introduced population:** This factor is usually unknown, as depending on the season of introduction, breeding can take place rapidly, leading to an increasing population from the moment of entry. Single detection at PoE (i.e., airports) are

possible by using different techniques described above, such as direct observations/ capture and relative sampling methods (traps) (Ibáñez-Justicia *et al*., 2017). In addition, some assumptions about the area of spread can be made, for example according to ECDC guidelines (2012), it could be assumed that "the population remains restricted to a small area with an indicative maximum area of 25 km²"

II. **Weather conditions during the sampling period:** The impact of abiotic factors like climate (*e.g.* temperature) on vector distributions can be used to refine appropriate surveillance strategies that focus on environmentally suitable areas. Early detection of mosquito activity then occurs when environmental conditions favour adult activity (flight, feeding, oviposition). Mathematical models using environmental profiles in each area may also be used to help to predict periods and levels of adult activity which may then be validated against existing data (Lana et al., 2014) and used to dictate surveillance strategies (Pulecio-Montoya *et al*., 2021)

III. **Sensitivity and specificity of the sampling methods** The sampling methods must be tailored for the target species.   For example, those widely used for for Aedes (*e.g.* ovitraps and adult traps, use of Lactic Acid rather than CO_2 in targeted attractants) are rather specific and are generally ineffective for Culicine mosquitoes.. In addition, if several AIM species are thought to be in the same area, the sampling used methods should be less specific. Direct sampling of breeding sites is also recommended, as it provides a high sensitivity and specificity.

As well as these, other factors, such as feasibility and economic viability must be taken into account . The ECDC guidelines (2012) provide indications about the sampling effort in different scenarios, including the surveillance of introductions at Points of Entry. In this case, a density between $1/2500m^2$ and $1/5000m^2$ is recommended. As a part of the implementation and integration, the sampling effort needed and the seasonal activity can be estimated with the aid of mathematical models than can identify likely periods and areas of peak activity (Lana et al., 2018).

The biology of the targeted species in relation to environmental constraints can be used to develop habitat suitability models, which will identify areas most likely to support the vector, and so dictate optimal locations for sampling. Maps presenting modelling outputs of habitat suitability can be usefully supplemented by maps showing access, transports and other factors that might affect the feasibility of setting traps. An example from the south of Belgium to sample the extent of a *Aedes japonicus* establishment and to coordinate its control efforts is shown below.

Figure 13: Worldcover maps of the area around a positive sample for Ae. japonicus
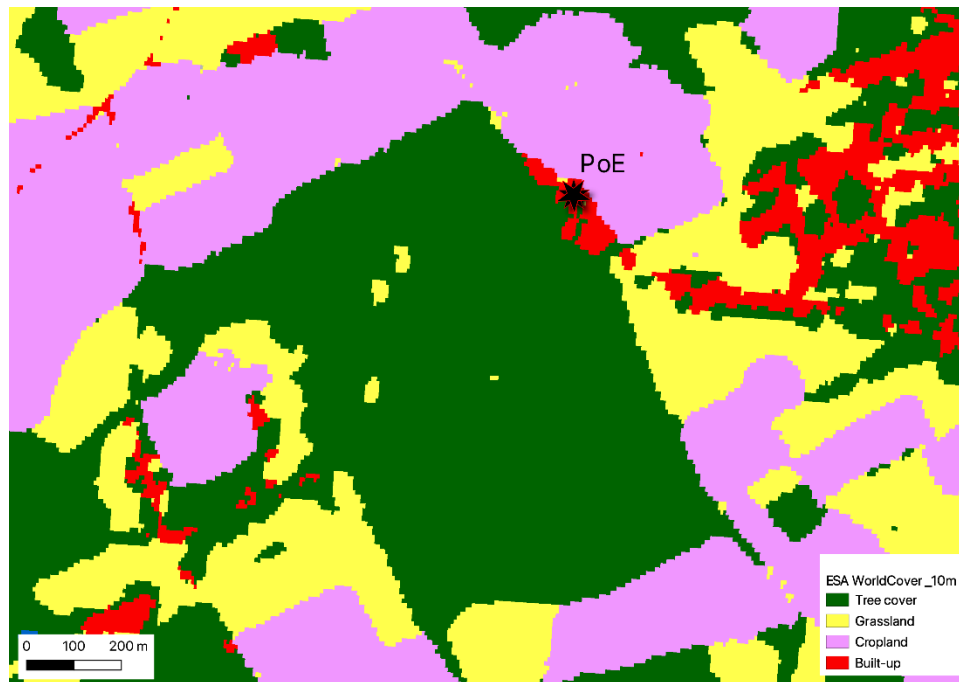
An example from the south of Belgium to sample the extent of a *Aedes japonicus* establishment and to coordinate its control efforts is shown above (Figure 13). From previous PoE monitoring, a positive PoE was found. As *Ae. japonicus* is a tree hole breeder, a land cover analysis of the surrounding area was conducted to identify suitable areas for breeding, and a forested area behind the PoE was identified and selected for surveillance**Error! Reference source not found.**. Both larval site sampling and ovitraps were used as they w ere deemed the most appropriate based on the location and biology of the species. This method proved to be effective in tracking vector invasion within the forest.

A second example from invasive mosquito monitoring in Belgium shows all positive and negative traps, their nature , provincial boundaries and major roads and shows clearly where the mosquitoes have been detected and where they were not (Figure 14). Note that the negative points are essential to provide a complete overview of the likely distribution

Figure 14: Mapped output of initial invasive mosquito surveillance, Belgium
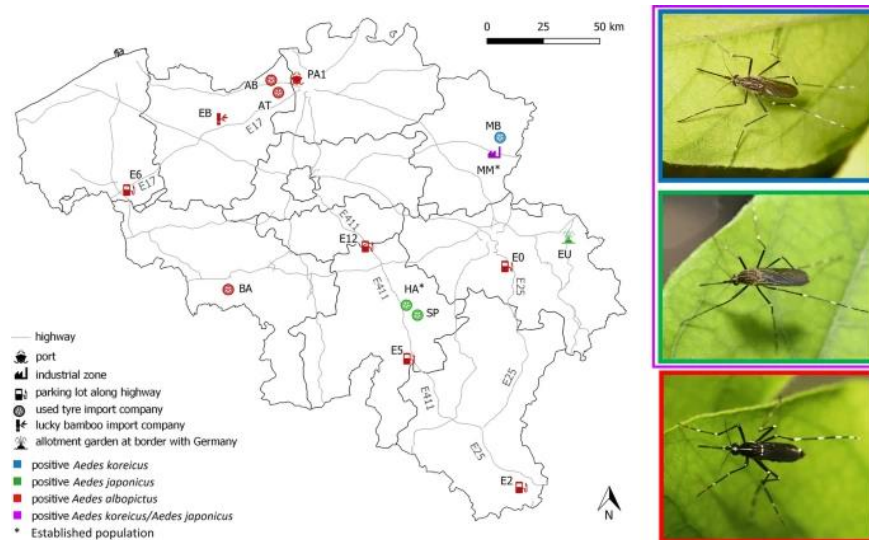
The results of these surveillance efforts need to be mapped to provide stakeholders with the outputs as soon as possible. They will help grasp the general situation, give a first assessment of potential control challenges, and help generate hypotheses for the source of introduction. Such maps will be of primary use for entomologists so they can summarise the situation for decision makers. The scale and extents of sampling should initially be local to the reported site of introduction, and the production of more extensive models can then be considered in order to predict likely spread and establishment.

b) Modelling to predict direction and distance of spread and to target surveillance

Though modelling plays a relatively minor role in the initial assessments of true presence or absence following a recorded introduction, they are an effective way of predicting spread from the point of introduction. Such models tend to cover regions well beyond the initial introduction site (Takahashi et al., 2005) and should ideally be available 'on the shelf' as part of any preparedness plan. It is likely that available models may not be specifically designed for the newly invaded region, but they are still likely to be able to provide initial predictions of suitability, activity or abundance at the time of the year, and potential spread to nearby locations. These may need some regional customisation, or be revised to incorporate newly acquired distribution data to improve their relevance.

If not available, then their production should be initiated as soon as possible after the introduction is recorded. These models can be based on habitat suitability or climate envelopes, or on known time series of vector distributions in other areas (e.g. Kraemer et al 2019). In addition, models including potential movement of species due to human transportation (i.e., inside cars or by leapfrog spread) (Eritja et al., 2017*;* Tavecchia et al., 2017) may be useful complements. In the case of *Ae.aegypti* in Europe, from which the species is largely absent, historical distributions can be used to infer likely suitability (Wint et al., 2022).

Basic distribution models, whether off the shelf or newly produced, may need further modifying when disseminated to a wider audience in order to show uncertainties and best or worst case spread extents that can be better interpreted by Public Health planners and strategists. Maps of modelled potential distribution are likely to be of most use to entomologists who can then use them to define surveillance and monitoring sample sites in areas with the highest predicted suitability, and to provide early warning of potential spread to areas far from the initial introduction.

c) Channels of communication

After the detection of an invasive mosquito species, such as *Aedes aegypti*, it is of great importance to establish the adequate communications channels with all the actors involved. The message intended for each sector should be clearly defined to match the sector priorities and perspectives.

The roles of each actor from national to local level should be defined according to the situation. For example, actions for professionals at each geographic level are defined in the ECDC guidelines for IMS (2012). In the practical case presented here, considering the first introduction of *Ae. aegypti* in a given area, definition and implementation of surveillance methods, data reporting and analysis, and means for communication (*i.e.* maps or tables) are key aspects to consider, as well as the involvement of each actor.

Of all mapping contexts, this is the most challenging, but should not be overlooked as it is also a very efficient way of communicating information. While many stakeholders are now familiar with maps there is no guarantee that they can interpret them properly, and so clarifying the message to be conveyed before making the map is of paramount importance. The example in Figure 15 provides a clear picture of where the Tiger mosquito has been recorded and where it is thought to be absent, but is still partially ambiguous in that the grey areas which represent no data are often mistakenly interpreted as not present. Ideally therefore such material should be accompanied by clear statements of the intended messages.

Figure 15: ECDC VectorNet map of Aedes albopictus in Europe
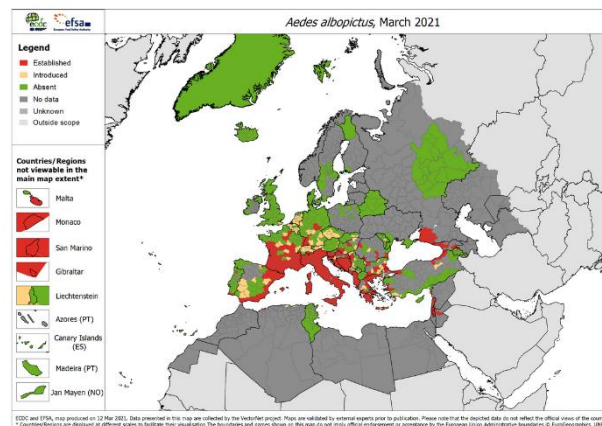
At the national and local level, information can be more spatially precise or specific. At this point considerations of resolution and extent will take into account who is the intended audience and the objective of the map. Informing the population surrounding a PoE to garner their help for surveillance (e.g. **Figure 16**) may focus on the local scale and local landscape features indicating proximity to residential or recreational areas, for example, but generating political attention and support to sustained funding for surveillance may need use national level data.

Figure 16: Mosquito records in Belgium. Left: local level; Right: national level

https://mosquito.sciensano.be/map
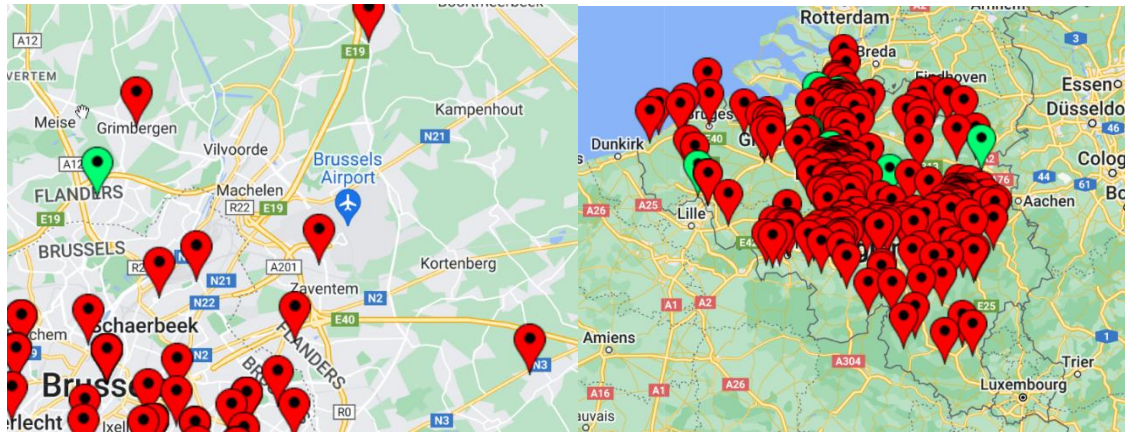
 Special attention should be given to what platform will be used for disseminating the maps: paper/fixed vs. online/dynamic. Simplicity and clarity need to be prioritised. Online mapping can be attractive when the map reader wants to be able to interact with the content but the absence of a clear message (as can be the case with data presentation) may interfere with your ability to advocate with audiences not interested with processing data and information themselves or interact with content in any way.

In conclusion: Sampling, modelling and mapping are tools in a set of tools that complement each other in various context. None of them covers the full set of activity, and the blend of each will depend on the need. This should encourage also to combine expertise through collaborations. Local context matters a great deal when faced with an introduction, so that broad scale models or maps may take a secondary role when confirmation and local extent are what is at stake. When considering the situation at the national or continental level, models and maps may become key in getting a useful overview. But many situations exist between these two extremes that need to combine modelling, spatial visualisation and thorough understanding of sampling features.

6 References

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338.

Bartumeus, F., Oltra, A., & Palmer, J. R. B. (2018). Citizen Science: A Gateway for Innovation in Disease-Carrying Mosquito Management? *Trends in Parasitology*, *34*(9), 727–729. https://doi.org/10.1016/j.pt.2018.04.010

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species distribution modeling for machine learning practitioners: a review. *ACM SIGCAS Conference on Computing and Sustainable Societies*, 329–348.

Buffoni, G., & Pasquali, S. (2007). Structured population dynamics: continuous size and discontinuous stage structures. *Journal of Mathematical Biology*, *54*(4), 555–595. https://doi.org/10.1007/s00285-006-0058-2

Chauvier, Y., Descombes, P., Guéguen, M., Boulangeat, L., Thuiller, W., & Zimmermann, N. E. (2022). Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity. *Ecography*, *2022*(10), e05973.

ECDC (2012). ECDC guidelines for the surveillance of invasive mosquitoes in Europe. *Eurosurveillance. European Centre for Disease Prevention and Control (ECDC).*

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, *40*(1), 677–697.

Erguler, K. (2020). sPop: Age-structured discrete-time population dynamics model in C, Python, and R. *F1000Research*, *7*, 1220. https://doi.org/10.12688/f1000research.15824.3

Erguler, K., Chandra, N. L., Proestos, Y., Lelieveld, J., Christophides, G. K., & Parham, P. E. (2017). A large-scale stochastic spatiotemporal model for Aedes albopictus-borne chikungunya epidemiology. *PLOS ONE*, *12*(3), e0174293. https://doi.org/10.1371/journal.pone.0174293

Erguler, K., Mendel, J., Petrić, D. V., Petrić, M., Kavran, M., Demirok, M. C., Gunay, F., Georgiades, P., Alten, B., & Lelieveld, J. (2022). A dynamically structured matrix population model for insect life histories observed under variable environmental conditions. *Scientific Reports*, *12*(1), 11587. https://doi.org/10.1038/s41598-022-15806-2

Erguler, K., Pontiki, I., Zittis, G., Proestos, Y., Christodoulou, V., Tsirigotakis, N., Antoniou, M., Kasap, O. E., Alten, B., & Lelieveld, J. (2019). A climate-driven and field data-assimilated population dynamics model of sand flies. *Scientific Reports*, *9*(1), 2469. https://doi.org/10.1038/s41598-019-38994-w

Erguler, K., Smith-Unna, S. E., Waldock, J., Proestos, Y., Christophides, G. K., Lelieveld, J., & Parham, P. E. (2016). Large-Scale Modelling of the Environmentally-Driven Population Dynamics of Temperate Aedes albopictus (Skuse). *PLOS ONE*, *11*(2), e0149282. https://doi.org/10.1371/journal.pone.0149282

Erguler, K., & Stumpf, M. P. H. (2011). Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Molecular BioSystems*, *7*(5), 1593–1602. http://www.ncbi.nlm.nih.gov/pubmed/21380410

González, E. J., Martorell, C., & Bolker, B. M. (2016). Inverse estimation of integral projection model parameters using time series of population-level data. *Methods in Ecology and Evolution*, *7*(2), 147–156. https://doi.org/10.1111/2041-210X.12519

Gurney, W. S. C., Nisbet, R. M., & Lawton, J. H. (1983). The Systematic Formulation of Tractable Single-Species Population Models Incorporating Age Structure. *Journal of Animal Ecology*, *52*(2), 479–495.

Guzzetta, G., Montarsi, F., & Baldacchino..., F. (2016). Potential risk of dengue and chikungunya outbreaks in northern italy based on a population model of aedes albopictus (diptera: Culicidae). *PLoS Negl Trop {…}.* http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004762

Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, *43*(4), 549–558.

Hendrickx, A., Marsboom, C., Rinaldi, L., Vineer, H. R., Morgoglione, M. E., Sotiraki, S., Cringoli, G., Claerebout, E., & Hendrickx, G. (2021). Constraints of using historical data for modelling the spatial distribution of helminth parasites in ruminants. *Parasite*, *28*.

Hucka, M., Bergmann, F. T., Chaouiya, C., Dräger, A., Hoops, S., Keating, S. M., König, M., Novère, N. le, Myers, C. J., Olivier, B. G., Sahle, S., Schaff, J. C., Sheriff, R., Smith, L. P., Waltemath, D., Wilkinson, D. J., & Zhang, F. (2019). The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2. *Journal of Integrative Bioinformatics*, *16*(2). https://doi.org/10.1515/jib-2019-0021

Ibañez-Justicia, A., Gloria-Soria, A., Den Hartog, W., Dik, M., Jacobs, F. &Stroo, A. (2017) The first detected airline introductions of yellow fever mosquitoes (Aedes aegypti) to Europe, at Schiphol International Airport, the Netherlands. Parasites & Vectors, 10, 603.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., & Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, *312*, 166–174.

Južnič-Zonta Ž, Sanpera-Calbet I, Eritja R, Palmer JRB, Escobar A, Garriga J, et al. Mosquito alert: leveraging citizen science to create a GBIF mosquito occurrence dataset. Gigabyte. 2022 May 30;2022:1–11.

Kampen, H., Medlock, J. M., Vaux, A. G. C., Koenraadt, C. J. M., van Vliet, A. J. H., Bartumeus, F., Oltra, A., Sousa, C. A., Chouin, S., & Werner, D. (2015). Approaches to passive mosquito surveillance in the EU. *Parasit Vectors*, *8*, 9. https://doi.org/10.1186/s13071-014-0604-5

Koons, D. N., Arnold, T. W., & Schaub, M. (2017). Understanding the demographic drivers of realized population growth rates. *Ecological Applications*, *27*(7), 2102–2115. https://doi.org/10.1002/eap.1594

Kraemer MUG, Reiner RC Jr, Brady OJ, Messina JP, Gilbert M, Pigott DM, Yi D, Johnson K, Earl L, Marczak LB, Shirude S, Davis Weaver N, Bisanzio D, Perkins TA, Lai S, Lu X, Jones P, Coelho GE, Carvalho RG, Van Bortel W, Marsboom C, Hendrickx G, Schaffner F, Moore CG, Nax HH, Bengtsson L, Wetter E, Tatem AJ, Brownstein JS, Smith DL, Lambrechts L, Cauchemez S, Linard C, Faria NR, Pybus OG, Scott TW, Liu Q, Yu H, Wint GRW, Hay SI, Golding N. Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. Nat Microbiol. 2019 May;4(5):854-863. doi: 10.1038/s41564-019-0376-y. Epub 2019 Mar 4. Erratum in: Nat Microbiol. 2019 Mar 21;: Erratum in: Nat Microbiol. 2019 Apr 8;: PMID: 30833735; PMCID: PMC6522366.

Lefkovitch, L. P. (1965). The Study of Population Growth in Organisms Grouped by Stages. *Biometrics*, *21*(1), 1–18.

Leslie, P. H. (1945). On the Use of Matrices in Certain Population Mathematics. *Biometrika*, *33*(3), 183–212. https://doi.org/10.1093/nq/s3-XI.286.498-b

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J. L., Hucka, M., Novère, N. le, & Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol*, *4*, 92. https://doi.org/10.1186/1752-0509-4-92

Lunde, T. M., Bayoh, M. N., & Lindtjørn, B. (2013). How malaria models relate temperature to malaria transmission. *Parasit Vectors*, *6*, 20. https://doi.org/10.1186/1756-3305-6-20

Lunde, T. M., Korecha, D., Loha, E., Sorteberg, A., & Lindtjørn, B. (2013). A dynamic model of some malaria-transmitting anopheline mosquitoes of the Afrotropical region. I. Model description and sensitivity analysis. *Malaria Journal*, *12*(1), 28. https://doi.org/10.1186/1475-2875-12-28

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*(12), 1267–1281.

Miranda, M. Á., Barceló, C., Arnoldi, D., Augsten, X., Bakran-Lebl, K., Balatsos, G., Bengoa, M., Bindler, P., Boršová, K., Bourquia, M., & others. (2022). AIMSurv: First pan-European harmonized surveillance of Aedes invasive mosquito species of relevance for human vector-borne diseases. *Gigabyte*, *2022*, 1–11.

Nisbet, R. M., & Gurney, W. S. C. (1983). The systematic formulation of population models for insects with dynamically varying instar duration. *Theoretical Population Biology*, *23*(1), 114–135. https://doi.org/10.1016/0040-5809(83)90008-4

Palmer, J. R. B., Oltra, A., Collantes, F., Delgado, J. A., Lucientes, J., Delacour, S., Bengoa, M., Eritja, R., & Bartumeus, F. (2017). Citizen science provides a reliable and scalable tool to track disease-carrying mosquitoes. *Nature Communications*, *8*(1), 916. https://doi.org/10.1038/s41467-017-00914-9

Pasquali, S., Soresina, C., & Gilioli, G. (2019). The effects of fecundity, mortality and distribution of the initial condition in phenological models. *Ecological Modelling*, *402*(September 2018), 45–58. https://doi.org/10.1016/j.ecolmodel.2019.03.019

Petrić, M., Ducheyne, E., Gossner, C. M., Marsboom, C., Venail, R., Hendrickx, G., Schaffner, F., & others. (2021). Seasonality and timing of peak abundance of Aedes albopictus in Europe: Implications to public and animal health. *Geospatial Health*, *16*(1).

Pulecio-Montoya, A. M., López-Montenegro, L. E., & Medina-García, J. Y. (2021). Description and analysis of a mathematical model of population growth of Aedes aegypti. *Journal of Applied Mathematics and Computing, 65(1–2), 335–349. https://doi.org/10.1007/s12190-020-01394-9*

Ross, R. (1908). *Report on the prevention of malaria in Mauritius*. Waterlow and Sons Limited.

Ryan, S. J., Lippi, C. A., Lowe, R., Johnson, S., Diaz, A., Dunbar, W., & others. (2022). *Landscape mapping of software tools for climate-sensitive infectious disease modelling*.

Scharlemann, J. P. W., Benz, D., Hay, S. I., Purse, B. v, Tatem, A. J., Wint, G. R. W., & Rogers, D. J. (2008). Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PloS One*, *3*(1), e1408.

Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W., & McKenzie, F. E. (2012). Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. *PLoS Pathogens*, *8*(4), e1002588. https://doi.org/10.1371/journal.ppat.1002588

Smith, N. R., Trauer, J. M., Gambhir, M., Richards, J. S., Maude, R. J., Keith, J. M., & Flegg, J. A. (2018). Agent-based models of malaria transmission: a systematic review. *Malaria Journal*, *17*(1), 299. https://doi.org/10.1186/s12936-018-2442-y

Takahashi, L.T., Maidana, N.A., Ferreira, W.C. *et al. Mathematical models for the Aedes aegypti dispersal dynamics: Travelling waves by wing and wind. Bull. Math. Biol. 67, 509–528 (2005).* https://doi.org/10.1016/j.bulm.2004.08.005

Tavecchia, G., Miranda, M. A., Borrás, D., Bengoa, M., Barceló, C., Paredes-Esquivel, C., & Schwarz, C. (2017). Modelling the range expansion of the Tiger mosquito in a Mediterranean Island accounting for imperfect detection. *Frontiers in Zoology, 14(1).* https://doi.org/10.1186/s12983-017-0217-x

Tran, A., L'ambert, G., Lacour, G., Beno\^\it, R., Demarchi, M., Cros, M., Cailly, P., Aubry-Kientz, M., Balenghien, T., & Ezanno, P. (2013). A rainfall- and temperature-driven abundance model for Aedes albopictus populations. *IJERPH*, *10*(5), 1698–1719. https://doi.org/10.3390/ijerph10051698

Tran, A., Fall, A. G., Biteye, B., Ciss, M., Gimonneau, G., Castets, M., Seck, M. T., & Chevalier, V. (2019). Spatial Modeling of Mosquito Vectors for Rift Valley Fever Virus in Northern Senegal: Integrating Satellite-Derived Meteorological Estimates in Population Dynamics Models. *Remote Sensing*, *11*(1024).

Uusitalo, R., Siljander, M., Culverwell, C. L., Hendrickx, G., Lindén, A., Dub, T., Aalto, J., Sane, J., Marsboom, C., Suvanto, M. T., & others. (2021). Predicting spatial patterns of sindbis virus (Sinv) infection risk in finland using vector, host and environmental data. *International Journal of Environmental Research and Public Health*, *18*(13), 7064.

Wint, W., Jones, P., Kraemer, M., Alexander, N., & Schaffner, F. (2022). Past, present and future distribution of the yellow fever mosquito Aedes aegypti: The European paradox. *Science of the Total Environment, 847.* *https://doi.org/10.1016/j.scitotenv.2022.157566* .

7 Appendix

7.1   Sampling

Defining a sampling strategy is a complex process. It involves deciding how often to sample, defining the sampling locations, and defining the protocols which set out what parameters to record during your sampling.

7.1.1 Sampling strategies

7.1.1.1   Cross-sectional sampling

A cross-sectional sampling is a sampling where you sample the location once and then move on to the next location, this method therefore creates a dataset with a specific sampling location at a single time point. This way of sampling allows to cover a wider area with the same number of traps. The results of this sampling method can be used for the presence and absence modelling of species. Especially in areas with an unknown distribution this a good method. If the aim of the sampling campaign is to detect the presence of the species, factors such as seasonality need to be taken into account.

7.1.1.2   Longitudinal sampling

Longitudinal sampling differs from cross-sectional sampling in that the same location is sampled at multiple time points. This method allows to monitor population dynamics over time which makes this data better suited for mathematical modelling. The time- interval between location visits has an important influence on the data. The aim of the monitoring will also influence the time-interval. The time interval also doesn't have to be consistent throughout the season, *e.g.* higher frequency during the peak of the season or daily sampling during one week per month.

The third option constitutes a combination of both methods where you do a repeated cross-sectional sampling where the same location is not constantly monitored but is revisited several times throughout the season. This method still might provide some information on dynamics and effectiveness of the cross-sectional sampling.

7.1.2 Choosing sampling locations

If you have no idea about what determines a species' location, it is advisable to carry out sampling throughout your area of interest and to set the sample size so as not to miss habitats that are less well represented, but which could be important for your species. Sampling locations can be assigned as points, along lines (transects) or inside quadrants. The location of the sampling locations can be assigned in a number of ways: completely random, randomly within defined zones (stratified random) or systematically (Figure 17).

If you already have an idea of the environmental preferences of your species, you can expect certain variables to influence the absence or presence of your species. In these cases, sampling locations are assigned to different geographical zones or strata related to the biology of the vector like vegetation category. Sampling locations can be assigned to ensure that each zone is sampled at the same intensity, or alternatively, so that the more favourable areas are sampled more frequently.

Figure 17:ways to place sample locations
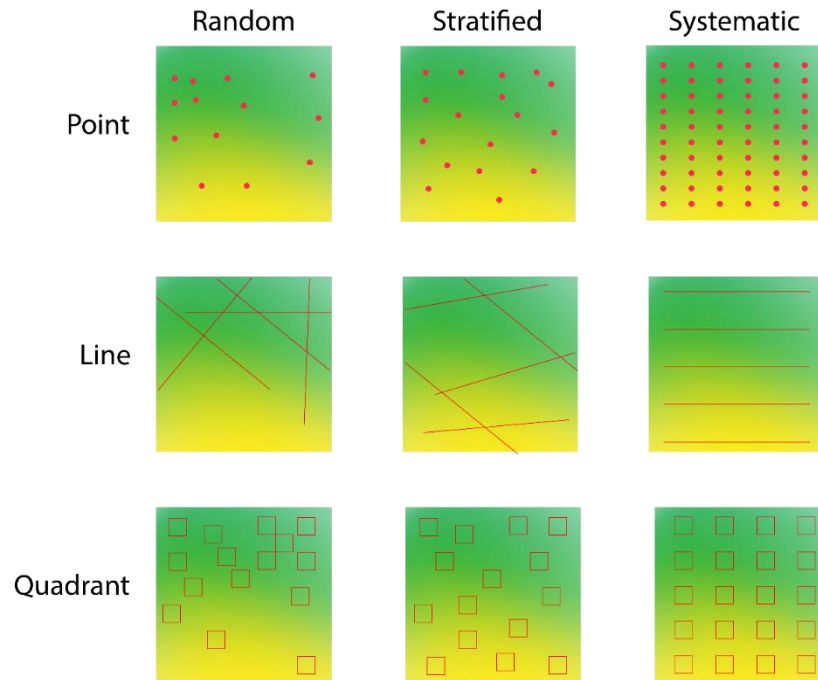
7.1.3 AIMSURV sampling protocols

The AIMSURV Initiative was launched in 2020. Its primary objectives were to conduct standardised surveillance for invasive mosquito species at a continental level, while recognising that field teams had access to different level of support for these activities. As a result, two protocols were developed: a set of minimum requirements designed to capture the peak annual population levels; and a more detailed sampling strategy designed to sample vector population throughout the year

7.1.3.1 Minimum requirements

- **Traps**
    - Density of 5 ovitraps per location with 15 to 100 m of distance between each of the traps, on a minimum of 3 locations, distant by 10 Km or (better) more; The sampling locations should share similar environment, *e.g.* garden of single family houses in residential urban/peri-urban areas, public parks near residential areas, recreational areas.
    - As a substrate for AIM spp. female oviposition, a wood tongue depressor (1.7*15 cm) should be used.
- **Period of sampling**
    - A minimum of 3 months sampling is required, making sure that the population peak of the targeted species is included within that period (*e.g.* Spain: from September to November).
    - This applies to ovitraps and Mosquito Alert.
- **Frequency of sampling**
    - Conduct sampling every two weeks during the three minimum months of sampling.
- **Parameters to record**
    - Geolocation: latitude and longitude of the position of each trap; Use the decimal system (*i.e.* 46.759463 N, 3.568237 E) and not the degree, minutes, seconds system.
    - Name of municipality/county/district (according to the country) and locality (see format in VECMAP® guidelines).
    - Start and End date refer to the trapping event for which the data are reported (*e.g.* a period of 14 days / 2 weeks for ovitraps), in order to get, for the final analysis, numbers per trap/night.
    - Land use category (see VECMAP® guidelines page 13 for possible options of the land use field).
    - Trap status: report technical issues that could have influenced the trapping result (negative or not), *e.g.* in case of trap missing or broken, oviposition support missing, battery out, etc.
    - When/where no AIM eggs are sampled or adults are caught, 0 (absences) have to be reported.

7.1.3.2   Recommended AIMSURV sampling requirement

More locations and dates samples to assess seasonal occurrence.

- **Locations**
  - Same density of ovitraps (5 per site) but conducted at more than 3 locations, distant as much as possible from each other to cover a wide area, at locations either sharing similar environments or showing different conditions (e.g. urban areas, rural areas, high altitude areas...).
  - Adult traps can be also used, BG-Sentinel™ trap baited with BG-Lure™ and C02 is recommended as the standard; One trap-night per site per week is recommended.
- **Period of sampling**
  - Whole mosquito season; The period of sampling is suggested to be increased in order to ensure to record the start, the peak and the end of the population activity in each site of sampling (*e.g.* May – November in Central Europe).
  - This applies to ovitraps, adult traps, and Mosquito Alert.
- **Frequency of sampling**
  - Weekly sampling during the three minimum months of sampling.
  - Weekly sampling during the whole season (start-peak-end) for adults and/or eggs.
- **Parameters to record**
  - Same as for minimum requirement, plus:
  - Daily or weekly record of meteorological parameter (maximum, minimum, average temperature) per site, collected by using data loggers or local weather stations.
  - A map showing the sampling locations (numbered) and countries' administrative units can be provided.

7.2   Reporting requirement details

Published results are often incomplete, incompatible or inconsistent. This makes assembling databases of vector distribution extracted from the literature very challenging, and the process would yield much more useful outputs if all published vector distribution reports contained a minimum set of standard components, as detailed below.

7.2.1 Sampling Location

Sample site locations should, if possible, be provided as geographical xy point coordinates in latitude and longitude, with some indication of precision. If point georeferencing is not possible, then settlement and admin unit names with xy coordinates of their centres should be given. It should be noted that data locations with a precision of less than NUTS2 level or its equivalent are only suitable for large scale mapping, and cannot be used for detailed analysis. Many journals now insist on the data used for publications being made available with the publication, and so it is also recommended that the sample locations are provided as an ESRI compatible vector format file. Additional sample site details, such as whether indoors or outdoors should also be provide where relevant.

7.2.2 Sampling details

The sampling method used affects not only the species that could theoretically be collected but also the number and type (*e.g.* life stage) of specimens caught. Reported results should therefore include sample method. For each trap the sampling dates, duration (*e.g.* days), and sample (*e.g.* trap) numbers. The start and end date of each sample event should be given.

7.2.3 Sampled species and numbers

The most basic of results consist of simple presence records for each species. Sampled numbers are preferable, if available, but it is emphasised that these are only useful if the sampling details provided allow the sample numbers to be standardised to number per trap per day (or hour or week). It is often a temptation to provided results only for the academically or epidemiologically important species, and to leave out the

information about less high-profile species in the interest of space. As a result, a lot of distribution information is discarded, which could be supplied as Supplementary Information even if there is not enough room in the main text. Such non-priority samples could also be provided as just presence rather than numbers.

7.2.4 Absences

It is just as important to know where a species in NOT found as to record its presence or abundance, especially at the beginning or end of a season, or on the edge of its range. Absence or zero values can be inferred for all species that are recorded at a particular location during a sampling programme: if, for example it is recorded once, then zero or absent can be attributed to every other sample in that series.

7.2.5 Vector species

The method of specimen identification for each species, species group or species complex should be specified. Specimens that are not reliably identified to the level of species, species group or species complex should not be reported. If the identification method used is not accepted as definitive, the specimen should not be reported. If the number of specimens caught is not reported, the "reported" status is a mandatory minimum data requirement.

7.2.6 Useful Covariates

Basic Covariates suggested for modelling vectors

The most widely used are climatic (temperature and precipitation), vegetation (Normalised Difference Vegetation Index, Enhanced Vegetation Index), Land Cover, Elevation, and human population density. All these except elevation are available for time series.   Some pointers for obtaining these datasets are given below.

Daily, monthly and annual Temperature data are available from:

- Satellite imagery (for example MODIS,), referred to as Land Surface temperature. Both day and night temperature are provided at daily, 8 day, monthly or annual intervals, and at 500m, 1, 5 or 8 kilometer resolutions, from 2000 onwards. Two common satellite sources are available: MODIS from 2000 onwards but unreliable since mid 2022, and VIIRS from 2012 onwards. These can be downloaded from https://search.earthdata.nasa.gov/search, usually as hdf format files. EU Sentinel satellite data is also regularly produced, but is not yet provided in a user friendly format
- Another common source of temperature data is the ECWMF weather station based data. This also provides global daily, monthly and annual temperature time series data, but it is air temperature rather than land surface temperature. Unlike the satellite data, these are also available as minimum, mean and maximum values. The data come as netcdf or GRIB format files and can be downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form. They are usually available at about 25km resolution, but scripting tools can be used to generate and download higher resolutions (https://cds.climate.copernicus.eu/toolbox-editor/9616)

Daily Monthly and Annual Precipitation data are also available from the Era5 Climate Data Store:

Long term averages such as for 1961-2000 and ***projected estimates*** (e.g. 2050 and 2080 for different climate change scenarios) for temperature and precipitation are available at https://worldclim.org. This site also provides a set of long term "bioclimatic indicators" such as temperature of warmest quarter, which are popular choices for model covariates.

***Vegetation indices and other indicators are available from*:**

- Satellite imagery via MODIS (since 2001) and VIIRS (since 2012) from Earthdata (https://search.earthdata.nasa.gov/search). Like temperature they are available at different resolutions and time periods (daily, 16day, monthly and annual). Two indices are available: NDVI and EVI, - with the latter being used more widely in recent -years.   The European Sentinel vegetation greenness data are coming online through the Copernicus and Earthdata sites, but have only been available since 2018, so are not yet suitable for longer time series.
- Land cover – *i.e.* estimates of shrubland or forest are also useful – especially to produce masks. Good sources are Corine (for EU only) as well as http://www.earthenv.org/ and ESA CCI (previously Globcover) at http://maps.elie.ucl.ac.be/CCI/viewer/download.php.

GMTED (https://topotools.cr.usgs.gov/gmted_viewer/gmted2010_global_grids.php) is a good source of **elevation dat**a produced by USGS. This produced minimum, mean and maximum values for each pixel location.

Estimates of human population at up to 100m resolutions, by age and gender, and from 2001 is available from https://topotools.cr.usgs.gov/gmted_viewer/gmted2010_global_grids.php.

7.3   Spatial modelling methods

Non-Linear Discriminant Analysis (NLDA) are best suited to presence/absence (0, 1) data or discrete (classified) data. It is possible to model continuous data if these are first binned into discrete classes using the threshold clustering tool. Better results may be obtained by using a model that is specifically geared to continuous data. NLDA is often considered superior to logistic regression as it models a complete multivariate normal distribution rather than a monotonic logit curve. A normal curve is more ecologically interpretable. Models can be bootstrapped. Clustering of data handles spatial heterogeneity of habitat niches and zones.

Random Forest can be run in regression mode (to predict continuous data) or classification mode (to predict presence/absence data or categories). Often quoted to handle complex interactions and correlated covariates very well. Useful for datasets with a high number of covariates compared to data points as it can remove redundant variables. Can be used to split   the study area into several ecozones with a pre-defined zone layer. Variable importance metrics are generated giving a clear indication of the strongest covariates. Can elegantly handle non-linear effects in covariates as trees are grown via a binomial split. Relatively fast, compared to a bootstrapped GLM or NLDA.

Generalised Linear Model (GLM) methodology Is more suited for continuous data, but can handle presence/absence data as well (with Model family= binomial). Can be used to split the study area into several eco-zones with a pre-defined zone layer to deal with spatial variability. Many options are available so the model can be tailored to your data. For example, count data with negative binomial or Poisson families, proportion data with the binomial family, zero-weighted continuous data with the Tweedie family, for continuous data, a least-squares regression equivalent with the Gaussian family and nonlinear distributed responses with the exponential family. Models can be bootstrapped. Can handle non-linear effects in covariates (by opting to pair each covariate with its square). GGWR option allows investigation of spatial variability in regression coefficients to see if spatial heterogeneity is an important issue in your training set. It has functionality to remove temporal trends and to account for spatial autocorrelation by using an autoregressive term or mixture model.

Boosted Regression Trees (BRT) is an evolved form of Random Forest that can elegantly handle non-linear effects in covariates as trees are grown via a binomial split. Can be seen as an improvement on random forests as the algorithm learns during an iterative process, rather than just outputting the average of a set of independent trees. Most of the tools for BRT also provide visualisation and quantification of variable interactions.

7.4   Mathematical modelling frameworks

Numerous mathematical frameworks exist, which can be used to represent climate- and environment-driven population dynamics. Compartmental models represent components of a system as groups of entities (individuals) in distinct states, such as groups of "mosquito larva" or "susceptible human". One of the most common mathematical frameworks used to develop compartmental models of vector populations and disease transmission is the ordinary differential equation (ODE), which dates back to the early twentieth century (Ross, 1908; D. L. Smith et al., 2012). ODE models have been developed to represent climate-driven population dynamics of many vector species (Lunde, Korecha, et al., 2013; Tran et al., 2013). Alternatively, using the delayed differential equations (DDEs) framework, Nisbet, Gurney, and Lawton developed a stage-structured DDE model to represent insect population dynamics (Gurney et al., 1983; Nisbet & Gurney, 1983), which links present dynamics with past environmental conditions using time lags. Partial differential equations (PDEs) incorporating life processes as part of the drift and diffusion components have also been utilised for the same purpose (Buffoni & Pasquali, 2007; Pasquali et al., 2019).

Matrix population models (MPMs) are discrete-time structured population models, making use of carefully designed projection matrices and matrix algebra to project population structure from one census date to the next. MPMs are commonly used to represent population heterogeneity by grouping individuals into distinct age (Leslie, 1945) and/or development stage classes (Lefkovitch, 1965). Recently, they have been applied to representing life processes driven by age as well as heat accumulation (Erguler, 2020; Erguler et al., 2022).

Alternatively, agent-based models (ABMs) explicitly represent each individual (even be it a single mosquito) and associate respective state variables, response functions, and links to internal and external drivers for each (N. R. Smith et al., 2018). ABMs are inherently stochastic and they readily represent spatial heterogeneity (*e.g.,* vector movement or heterogeneous control applications) and intrinsic stochasticity (*e.g.,* different survival and development rates). However, such flexibility requires programmers experienced in model construction and extensive computational resources to run simulations, which are common restrictions in the adoption of such methods.

7.5   Map types and content

Many options exist for mapping but not all can be used for all variables. We only mention here the most widespread.

Figure 18: Representations offered by the free online mapping tool Magrit.
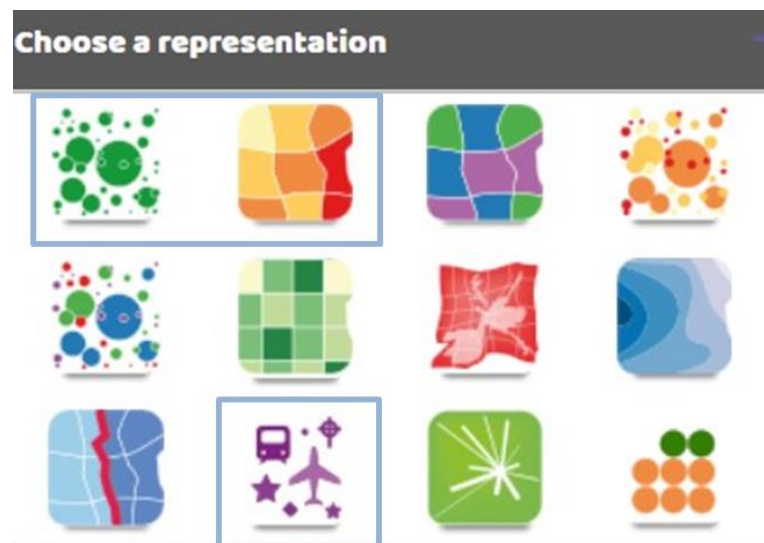
In the blue box at the top left of Figure 18 you can see two maps used to map continuous variables (so with different shades of a single colour): a proportional symbol representation on the left, and a choropleth map on the right. The third map in the top line is another choropleth with categorical data (so with different colours), and the first map on the second line is a dual variable symbol map with a categorical variable. The box on the last line highlights a simple symbol map, useful for mapping objects such as landmarks or trapping sites. More detail here https://magrit.cnrs.fr/docs/carto_fr.html.

This demonstrates that different types of data should be mapped using different types of maps. The two common map types are those using proportional symbols so that the symbol size relates directly to the data value (Figure 19); and "choropleth" maps, in which the colour used to fill a polygons such as an administrative unit is determined by the data value (Figure 20).

Figure 19: Proportional symbol map, dengue cases, ECDC)
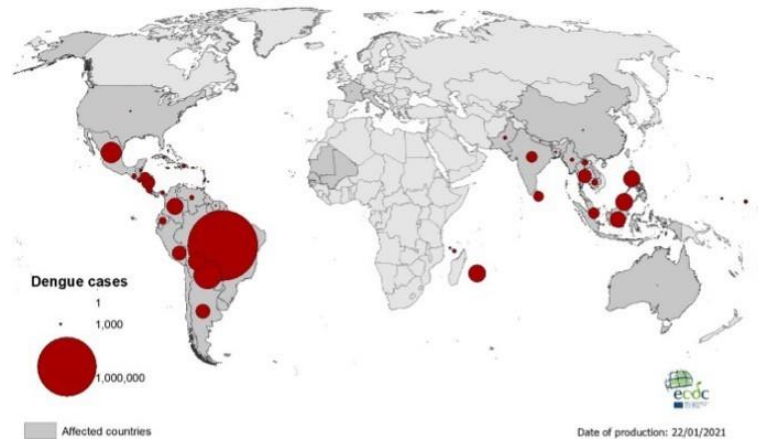
Proportional symbol maps are most appropriate for absolute numbers, whilst ratio variables can be mapped as choropleth, as illustrated in in Figure 19 and Figure 20, where the absolute number of cases for comparing South East Asia and South Asia can be very different whilst the incidence rates are more comparable. Choropleths are unfortunately often used to map simple numbers for variable which can produce very misleading results and colour differences may be too limited to represent large value ranges.

*Figure 20: Choropleth map, ratio variable: dengue incidence, ECDC*
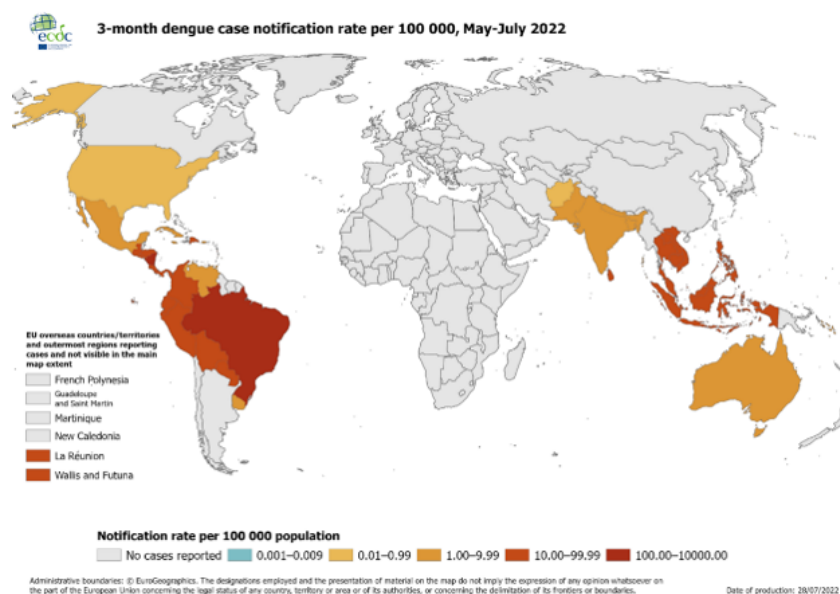
7.5.1 *The language of maps: graphic semiology*

The language of maps pertains to the colours, textures and symbols that we can use to map various phenomena. As pictured in Figure 21, various options exist to show differences between spatial objects, that include colour (and its variation), size, shape and texture. Coherence and readability are key here. Coherence with common place understanding of colour (*e.g.* green/blue usually perceived as "cold"/low/ positive and red as "hot"/high/negative or dangerous) and habits in a field. Readability as helping to fulfil the map's objective to convey a specific message.

A summary of the foundational text by French cartographer Jacques Bertin can be found here: https://innovis.cpsc.ucalgary.ca/innovis/uploads/Courses/InformationVisualizationDetails/09Bertin.pdf . It provides a good introduction to these important choices made in cartography – everything you could read about this will be based on these principles, even though further evolution is to be expected in the context of the growing prominence of on-screen viewing.

Figure 21: Example legends
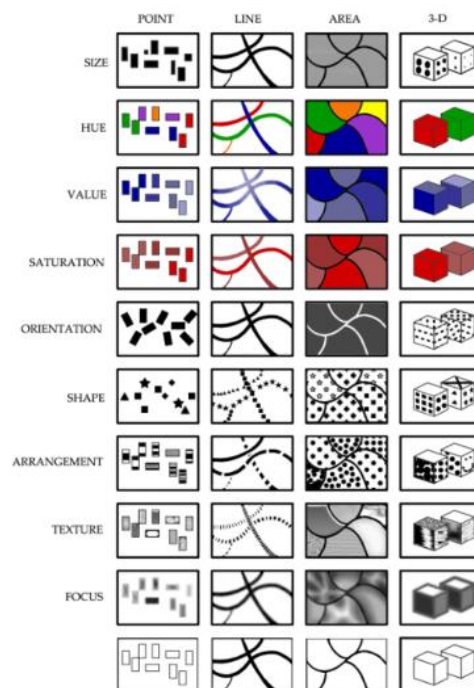
Source: National Information Security, Geospatial Technologies Consortium (NISGTC)

Source QGIS documentation

7.6   Other mapping issues

7.6.1 *Projection systems*

Spatial analysis data is inherently linked with projection systems. All spatial data is projected using a spatial or coordinate reference system (SRS or CRS. During analysis it is important that all data and co- variates are projected and projected with the same CRS. Therefore, it is important to always check the CRS of the different data and convert when necessary.

Three common projection families are cylindrical, conical and planar (Figure 22). All projections distort (stretch or compress) reality when drawn on paper. The rectangles on each map in Figure 22 are an equal number of degrees wide and deep, and the projection sets how long and wide each degree is on paper in different ways. As a result, they look very different on a map depending on the projection used, and so drawing a map in one projection on top of one in a different projection can cause spectacular errors.

Figure 22: The three major map projections systems
a) cylindrical, b) conical c) planar.
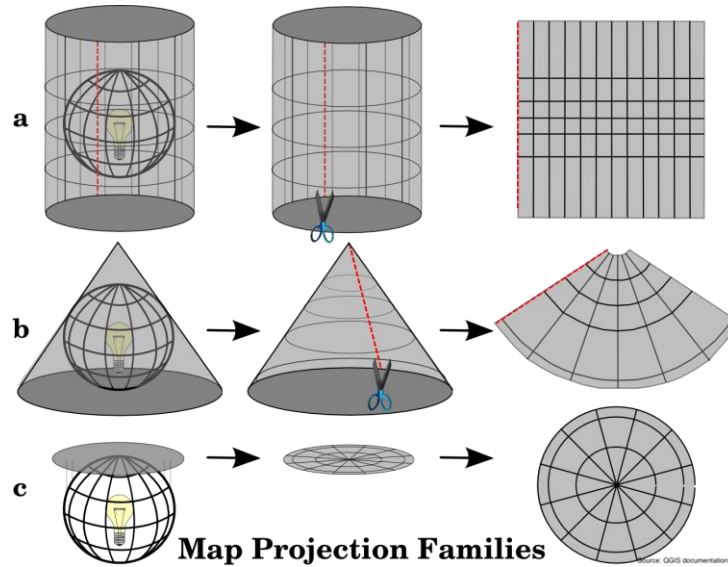
Each projection is designed to be useful in different situations. Cylindrical projections, for example, show correct shapes and directions, but the larger the area the less accurate are the distances measured from the map. Conical projections are good for mapping regions aligned east to west, and planar ones are better for mapping circular regions rather than rectangular ones.

There are thousands of projections. A good and easy system to distinguish and identify a CRS is by its EPSG code (https://epsg.io/). The Most common is EPSG:4326, which is a global CRS. A common one for Europe is EPSG:3035 which project European countries in a more realistic way.

7.7   Resources for mapping

7.7.1   Online

https://www.bloomberg.com/news/articles/2015-06-25/how-to-avoid-being-fooled-by-bad-maps (last checked August 9th, 2022)

Map projection (6min video) https://www.youtube.com/watch?v=kIID5FDi2JQ&t=225s Course on cartography https://storymaps.arcgis.com/collections/bc79ea24ec354f77bfa7616b247ac986

7.7.2   Books

Monmonier M, 2018, How to lie with maps. Third edition. Chicago University Press

Kraak M.J., Ormeling F., 2020, Cartography. Visualization of Geospatial data. Fourth edition. CRC Press

Buffoni, G., & Pasquali, S. (2007, 3). Structured population dynamics: continuous size and discontinuous stage structures. Journal of Mathematical Biology , 54 , 555-595.

Retrieved from http://link.springer.com/10.1007/s00285-006-0058-2   doi:10.1007/s00285-006-0058-2

Erguler, K., Pontiki, I., Zittis, G., Proestos, Y., Christodoulou, V., Tsirigotakis, N., . . . Lelieveld, J. (2019, 12). A climate-driven and field data-assimilated population dynamics model of sand flies. Scientific Reports, 9 , 2469. Retrieved from https:// doi.org/10.1038/s41598-019-38994-w doi: 10.1038/s41598-019-38994-w

Erguler, K., & Stumpf, M. P. H. (2011). Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability   in systems biology models. Molecular bioSystems, 7 , 1593-602. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21380410

González, E. J., Martorell, C., & Bolker, B. M. (2016, 2). Inverse estimation of integral projection model parameters using time series of population-level data. Methods in Ecology and Evolution, 7 , 147-156. Retrieved from https://onlinelibrary.wiley..com/doi/abs/10.1111/2041-210X.12519 doi: 10.1111/2041-210X.12519

Gurney, W. S. C., Nisbet, R. M., & Lawton, J. H. (1983). The systematic formulation of tractable single-species population models incorporating age structure. Journal of Animal Ecology , 52 , 479-495.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., My- ers, C. R., & Sethna, J. P. (2007, 1). Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol , 3 , e189. Retrieved from http://compbiol.plosjournals.org/perlserv/?request=get document&doi=10.1371/journal.pcbi.0030189.eor

Hucka, M., Bergmann, F. T., Chaouiya, C., Dräger, A., Hoops, S., Keating, S. M., . . .Zhang, F. (2019, 6). The systems biology markup language (sbml): Language spec- ification for level 3 version 2 core release 2. Journal of Integrative Bioinformatics, 16 . doi: 10.1515/jib-2019-0021

Koons, D. N., Arnold, T. W., & Schaub, M. (2017, 10). Understanding the demographic

drivers of realized population growth rates. Ecological Applications, 27 , 2102-2115. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/eap.1594 doi: 10.1002/eap.1594

Lefkovitch, L. P. (1965). The study of population growth in organisms grouped by stages.

Biometrics, 21 , 1-18. Leslie, P. H. (1945). On the use of matrices in certain population mathematics.

Biometrika, 33 , 183-212. doi: 10.1093/nq/s3-XI.286.498-b

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., . . . Laibe,

C. (2010, 1). Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol , 4 , 92. doi: 10.1186/ 1752-0509-4-92

Lunde, T. M., Bayoh, M. N., & Lindtjørn, B. (2013, 1). How malaria models relate temperature to malaria transmission. Parasit Vectors, 6 , 20. Retrieved from http:// www.parasitesandvectors.com/content/6/1/20 doi: 10.1186/1756-3305-6-20

Lunde, T. M., Korecha, D., Loha, E., Sorteberg, A., & Lindtjørn, B. (2013, 1). A dynamic model of some malaria-transmitting anopheline mosquitoes of the afrotropical region. i. model description and sensitivity analysis. Malaria Journal , 12 , 28. Re- trieved from https://malariajournal.biomedcentral.com/articles/10.1186/ 1475-2875-12-28 doi: 10.1186/1475-2875-12-28

Nisbet, R. M., & Gurney, W. S. (1983). The systematic formulation of population models for insects with dynamically varying instar duration. Theoretical Population Biology , 23 , 114-135. doi: 10.1016/0040-5809(83)90008-4

Pasquali, S., Soresina, C., & Gilioli, G. (2019). The effects of fecundity, mortality and distribution of the initial condition in phenological models. Ecological Modelling , 402 , 45-58. Retrieved from https://doi.org/10.1016/j.ecolmodel.2019 .03.019 doi: 10.1016/j.ecolmodel.2019.03.019

Ross, R. (1908). Report on the prevention of malaria in mauritius. Waterlow and Sons Limited.

Ryan, S. J., Lippi, C. A., Lowe, R., Johnson, S., Diaz, A., Dunbar, W., et al. (2022). Landscape mapping of software tools for climate-sensitive infectious disease mod- elling.

Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W., & McKenzie, F. E. (2012, 4). Ross, macdonald, and a theory for the dynamics and control of mosquito- transmitted pathogens. PLoS Pathogens, 8 , e1002588. Retrieved from https:// doi.org/10.1371/journal.ppat.1002588 doi: 10.1371/journal.ppat.1002588

Smith, N. R., Trauer, J. M., Gambhir, M., Richards, J. S., Maude, R. J., Keith, J. M., & Flegg, J. A. (2018, 12). Agent-based models of malaria transmission: a systematic review. Malaria Journal , 17 , 299. Retrieved from https://malariajournal.biomedcentral.com/articles/10.1186/s12936-018-2442-y doi:10.1186/ s12936-018-2442-y

Tran, A., L'ambert, G., Lacour, G., Benoˆ, R., Demarchi, M., Cros, M., . . . Ezanno, P. (2013, 5). A rainfall- and temperature-driven abundance model for aedes albopictus populations. IJERPH , 10 , 1698-1719. Retrieved from http://www.mdpi.com/ 1660-4601/10/5/1698 doi: 10.3390/ijerph10051698