BIOS0052: Human And Ecosystem Health In A Changing World

Week 3 practical session

Measuring the impacts of climate on *Aedes albopictus* abundance

In today's practical we will use mosquito surveillance data to understand how climatic factors influence the population dynamics of *Aedes albopictus*, the Asian tiger mosquito, in Northern and Central Italy. We will build on last week's mapping and simple threshold-based models of heat hazard, to more formally develop statistical models that relate mosquito abundance to environmental factors. We will be using spatial data tools (e.g. the *terra* and *sf* packages), but will also use *mgcv*, a powerful and flexible R package for fitting **generalised additive models** (more on these later).

The tiger mosquito (*Ae. albopictus*) has gradually colonised Europe over the last half-century, creating a new series of escalating risks for mosquito-borne disease. This species is a competent vector for several viruses that pose major threats to human health, in particular dengue and chikungunya. First found in Italy in the 1990s, it has now established populations across most of the country, and during the last decade several Italian cities have seen significant, sustained outbreaks of dengue and chikungunya. In response Italian health authorities have been conducting year-on-year mosquito surveillance (trapping) to understand this species' populations and the factors that create outbreak risks. A large set of surveillance data from several Italian provinces was recently published - you can read the paper here - and this dataset forms the basis for today's practical.

This workbook is structured as a series of code snippets with short exercises interspersed. The solutions for the short exercises are available in the Rmarkdown script in the GitHub folder, but *please avoid looking at these* until you have tried to solve them!

At the end, there are some **longer extension exercises** to allow you to apply your knowledge and prepare you for the assessments. The solutions for these will be uploaded to Moodle next week.

All the data you'll need for the workshop are in the GitHub, in the "Week3-Measuring-Env-Effects" folder. Please download the whole folder using download-directory.github.io. Set this folder as your working directory, then all the materials you will need are contained within the "data" subfolder.

```
# package dependencies
# use the "install.packages()" command if not already installed
library(terra); library(dplyr); library(magrittr); library(ggplot2); library(sf)
library(rstudioapi); library(tidyr); library(stringr); library(mgcv); library(gratia)

# automatically set working directory
# (or if this doesn't work,
# manually set your working directory to the folder "Week3-Measuring-Env-Effects")
PATH = dirname(rstudioapi::getSourceEditorContext()$path)
setwd(PATH)
```

Our research questions

The biological processes and population dynamics of invertebrates - including mosquitoes - are highly sensitive to local climate, and this in turn can influnce risks of mosquito-borne disease emergence and spread. So our analyses will focus on the question: **what is the effect of local temperature on the abundance of *Aedes albopictus*?** We will develop models to ask this question, consider possible confounders, and explore what the shape of this relationship might be.

Understanding the dataset

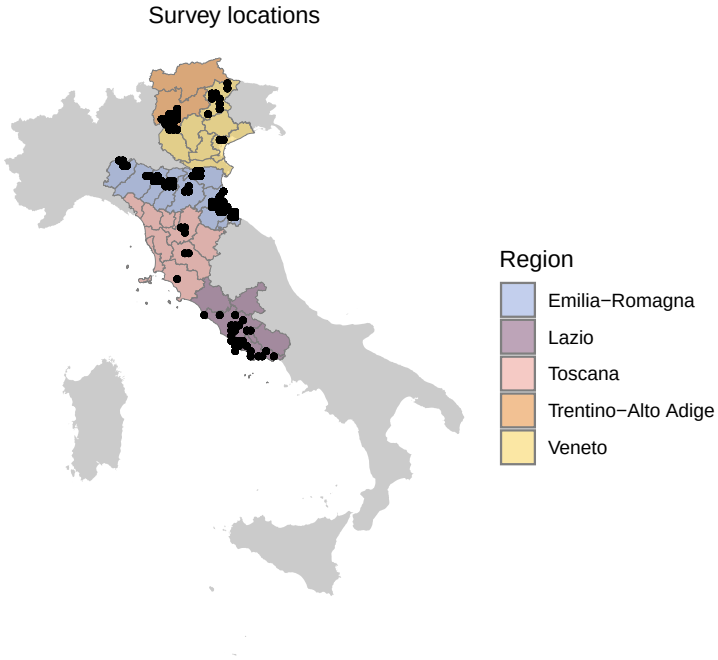
To answer this question, we will use a subset of a larger mosquito surveillance database that was recently published in *Scientific Data* (Da Re et al, 2024). Our data consist of monthly counts of *Ae. albopictus* eggs, collected using ovitraps, for several months each year between 2010 and 2022. Ovitraps provide a water container in which female mosquitoes lay their eggs, which are subsequently collected and counted. Repeat surveys were conducted monthly at almost 100 unique locations - specified as the column "*ID*" in the dataset - within five Italian regions (Trento, Emilia-Romagna, Lazio, Tuscany and Veneto). Here, we can use **egg counts as a proxy for mosquito abundance**, and begin to ask what influences abundance.

```
# the mosquito surveillance data are stored as a csv
# mosquito counts are stored in the "count" column
dd = read.csv("./data/vectabundance/italy_vectabundance_summercounts_monthly.csv")

# an accompanying shapefile of italian municipalities
shp = sf::st_read("./data/shapefiles/italy_vectabundance_regions.shp")

# shapefile of italian border (for mapping)
shp_ita = sf::st_read("./data/shapefiles/gadm41_ITA_0.shp")
```

- **Q1**: Explore the surveillance dataset. What months of the year do we have data for? What other information does the data frame contain, in addition to the monthly mosquito counts?

At the beginning of an analysis project, we need to understand what exactly the data are, their spatial and temporal distribution, and any additional information that has been collected. These data contain spatial survey locations (in the x and y column) so a useful starting point would be to map them.

```
map = ggplot() +
  geom_sf(data=shp_ita, color=NA, fill="grey60", alpha=0.5) +
  geom_sf(data=shp, color="grey50", aes(fill=Region), alpha=0.5) +
  theme_void() +
  geom_point(data = dd, aes(x, y), color='black', size=1) +
  MetBrewer::scale_fill_met_d(name="Archambault") +
  ggtitle("Survey locations") +
  theme(legend.position="right",
        plot.title=element_text(size=11, hjust=0.5))
map
```



- **Q2**: Plot a histogram or bar plot of the number of observations across years, and across months. Is the sampling relatively even across years and months, or does it vary?

Before we start analysing the data it's also critical to visualise and understand **variation in the response variable**. This is stored in the column "count". Looking at the dataset (*hint: you can use R's "head()" function to examine it*) it's clear that we have a form of nested, hierarchically-structured time series data. At each survey location (**denoted by the column "ID"**) we have monthly counts during the summer months (June to September), over multiple years.

- **Q3**: How many survey locations are in our dataset? (*Hint: the unique() or n_distinct() functions might be helpful*)

We can visualise these time series for some example survey locations, to demonstrate the structure of the data. It's clear that there is a lot of variation in abundance between different locations, and also between years.

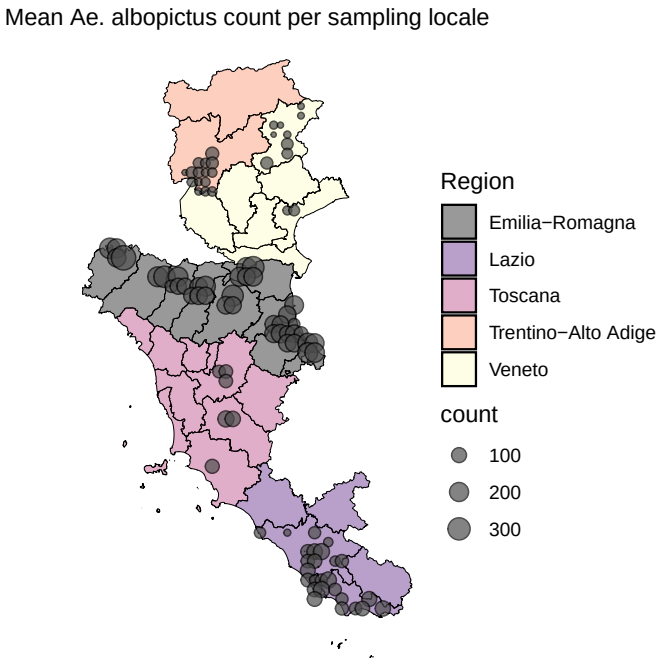
```
# subset to a random selection of locations
random_locs = dd %>%
```

```
  dplyr::filter(ID %in% c(6777, 8189, 7490, 14199))

# create a "date" column for easier plotting
random_locs$date = as.Date(paste(random_locs$year, random_locs$month, "01", sep="-"))

rl_plot = random_locs %>%
  ggplot() +
  geom_line(aes(date, count, group=year)) +
  geom_point(aes(date, count, group=year)) +
  theme_classic() +
  facet_wrap(~ID, ncol=1) +
  xlab("Month") + ylab("Ae. albopictus abundance") +
  ggtitle("Aedes albopictus count time-series for example locations") +
  theme(strip.background = element_blank(),
        plot.title=element_text(size=11, hjust=0.5))
rl_plot
```

- **Q4**: Visualise and compare the distributions of counts across each survey year, for example using boxplots. Are mosquito abundance levels quite similar during most years, or is there interannual variability? (*Hint: geom_boxplot() or geom_violin() might both be helpful here*)

It is often useful to map our response variable, to assess whether there are any obvious spatial trends. Since we have multiple counts for each location, we can't effectively show everything on a map at once. One option is to instead visualise the mean count at each location. When you plot this map, what do you notice about how *Ae. albopictus* is distributed geographically? What do you think might be causing this?

```
# calculate mean count per location across all surveys
mean_counts = dd %>%
  dplyr::group_by(ID) %>%
  dplyr::summarise(count = mean(count, na.rm=TRUE),
                   x = head(x, 1),
                   y = head(y, 1))

# plot the map of average differences between sampling locations
mean_counts %>%
  ggplot() +
  geom_sf(data=shp, color="black", aes(fill=Region), alpha=0.4) +
  theme_void() +
  geom_point(aes(x, y, size=count), fill="grey30", alpha=0.7, pch=21) +
  scale_color_viridis_c(name="Mean\nsummer\ncount") +
  scale_fill_viridis_d(option="magma") +
  scale_size(range=c(0.8, 5)) +
  ggtitle("Mean Ae. albopictus count per sampling locale") +
  theme(plot.title=element_text(size=11, hjust=0.5))
```



Investigating the role of temperature in shaping monthly *Aedes albopictus* abundance

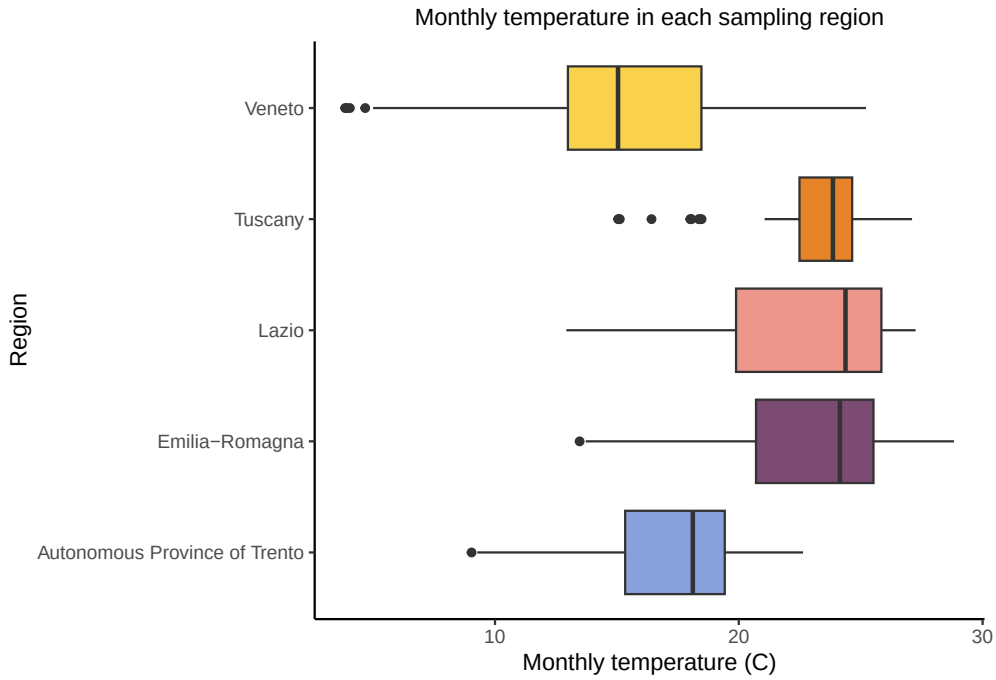
Our maps indicate that *Ae. albopictus* appears to be particularly abundant in the Emilia-Romagna region, and less so further to the north and south. Italy is a very mountainous country so it's possible that local climatic factors are playing a role. The dataset contains a column - *"temperature_c"* - which contains the monthly average temperature during the month preceding surveys, accesssed from ERA5-Land.

First, let's examine the correlation between monthly temperature and *Ae. albopictus* egg count.

- **Q5**: Plot a scatter plot of the relationship between temperature and egg counts. Is there visual evidence of a relationship?

We can also use boxplots or scatter plots to examine differences in temperature between the different regions where our data were collected. What do you notice about the regional differences in temperature? Do you think these could be playing a role?

```
# scatter plot of temp vs counts
ggplot(dd) +
  geom_boxplot(aes(Region, temperature_c, fill=Region)) +
  theme_classic() +
  xlab("Region") + ylab("Monthly temperature (C)") +
  MetBrewer::scale_fill_met_d(name = "Archambault") +
  ggtitle("Monthly temperature in each sampling region") +
  theme(plot.title=element_text(size=11, hjust=0.5), legend.position="none") +
  coord_flip()
```



Fitting generalised additive models to estimate the effect of temperature

Now we will investigate our research question using statistical models, to assess the support for the relationship between temperature and *Ae. albopictus* abundance. Recall that, because our response variable is *count* data, it is more appropriate to use an error distribution suited for count data. Here, we will start with a *Poisson mixed-effects regression* with a *log link function*, where we estimate the effect of covariates on the **log *Ae. albopictus* abundance**.

The model would be formulated as:

$Y_i \sim Pois(\lambda_i)$

$log(\lambda_i) = \beta_0 + X\beta$

where Y_i is observed *Ae. albopictus* egg count, and λ_i is the mean expected value of a Poisson distribution, which we are estimating as a logarithmic function of covariates. β is a vector of slope parameters, and X is a matrix of covariates.

You may remember Poisson models from last term or other previous statistics courses. In today's workshop, we will be implementing these models in the R package **mgcv**, which provides a powerful, fast and flexible framework for fitting a wide variety of models. mgcv is specifically designed for fitting *generalised additive models (GAMs)* which contain mixtures of both linear and nonlinear terms, but it can also fit mixed-effects models (i.e. with random effects). For a great beginner's introduction to the principles of GAMs in mgcv, see: https://noamross.github.io/gams-in-r-course/chapter1.

In the following code block we will prepare the data then fit a Poisson model in mgcv. This model includes a random intercept for survey location (which we specify using the *s(ID, bs="re")* term) since we have multiple sampling events at each location. We will also include a categorical fixed effect for Region (to account for unmeasured regional differences), and a linear fixed effect for temperature. (*Note that we centre and scale the temperature covariate beforehand, i.e subtract the mean and divide by the standard deviation. This helps to stabilise the inference, and ensures that slope parameters always describe the change in Y for 1 standard deviation change in X, regardless of what units X was measured in.*)

```r
# ensure all categorical variables are coded as factors
# so model doesn't interpret them as numeric!
dd$ID = as.factor(dd$ID)
dd$year = as.factor(dd$year)
dd$month = as.factor(dd$month)

# specify a GAM with count as response variable
# scaled linear effect of temperature
# random intercept of survey location
# categorical fixed effect for region
m2 = mgcv::gam(count ~ scale(temperature_c) + s(ID, bs="re") + Region, # formula
               data=dd,
               family="poisson",
               method="REML")

# model diagnostics: examine pattern of residuals

# define a function to extract and plot residuals vs fitted values
residFittedPlot = function(model){

  r = data.frame(
    fitted = fitted(model),
    resid = resid(model, type="pearson")
  )
  ggplot(r) +
    geom_point(aes(fitted, resid), color="blue", pch=21, fill=NA) +
    theme_classic() +
    geom_hline(yintercept=0, lty=2) +
    xlab("Fitted") + ylab("Residual (Pearson)")

}

# apply function to plot residuals - how does this look?
# (these should be a nice even cloud of points)
residFittedPlot(m2)

# finally view summary of fitted model to see fixed effects estimates
summary(m2)
```

Have we accounted for all potential unmeasured sources of variation in this model? What about seasonal and interannual differences that are not related to temperature?

- **Q6**: Modify the above code to fit another model called "m3", adding fixed effects for year and month. Examine the model summary and residuals. Has accounting for unexplained seasonal and interannual variation changed your estimate of the effect of temperature?

When we visualise these residuals, we see the fit isn't too bad, but still shows some very high residual values. The Poisson distribution makes strict assumptions about the error variance around the fitted mean (which must be equal to the mean), and our model is probably violating these. We can relax this assumption by **changing the model's likelihood to a *negative binomial* distribution** - this is a more general case of the Poisson distribution that allows for much higher variance in error, independent of the fitted mean.

```r
# fit negative binomial model
m4 = mgcv::gam(count ~ scale(temperature_c) + s(ID, bs="re") + Region + year + month,
               data=dd,
               family="nb", # change family to "NB"
               method="REML")

# plot residuals
# looks numerically better, except a few outliers at low fitted values
residFittedPlot(m4)
```

Recall that one way to compare the fit betwen models with the same error family is to use **Akaike Information Criterion**, an estimate of goodness-of-fit penalised by model complexity. A lower AIC value denotes a better fitted model, so we can use this to assess which of our models is the most appropriate fit to the data.

- **Q7**: Using the AIC function, compare AIC values between your 3 fitted models. Which fits the data best? Examine the findings of this best-fitted model. What does it suggest about the effect of temperature on mosquito abundance?

Modifying our model to test for a nonlinear effect of temperature

Given what we have recently learned about the thermal sensitivity of mosquito biology, it seems surprising to find a negligible marginal effect of temperature on *Ae. albopictus* abundance, after adjusting for the temporal and spatial structure of the dataset. Can you think of a possible reason for this, either biological or statistical?

One plausible reason is that the effect of temperature might be nonlinear, potentially with a thermal optimum beyond which abundance starts to decline, due to detrimental impacts on survival and reproduction. In such a case, it's possible that a linear effect would not detect a more complex relationship like this. One of the most powerful aspects of *mgcv* is its ability to fit complex splines and other nonlinear functions to data, as the sum of a set of basis functions (for a beginner-friendly technical introduction, see Noam Ross' tutorial).

Let's modify our model to instead fit our temperature effect as a penalised thin-plate regression spline. The *"k"* argument specifies the number of basis functions ('knots') and helps determine how wiggly our spline could be. Compare this model to the model with a linear temperature effect using AIC - which is more strongly supported?

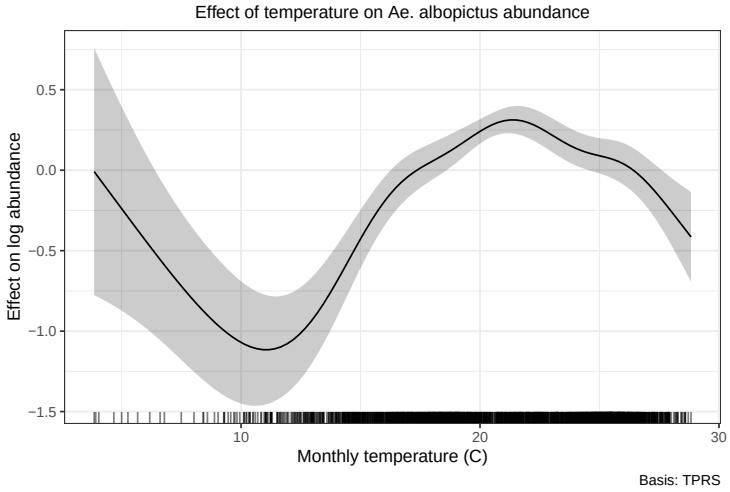
```r
# adding a nonlinear temperature effect
m5 = mgcv::gam(count ~ s(temperature_c, k=9, bs="tp") + s(ID, bs="re") + Region + year + month,
               data=dd,
               family="nb",
               method="REML")
```

After fitting the model, use *gratia*'s draw function to plot the spline from the model.

What does this show about the effect of temperature? What has including a nonlinear term helped us to understand?

```r
# gratia::draw() is a useful function to plot all fitted functions from an mgcv model
# here we customise to just plot the temperature effect and improve aesthetics
gratia::draw(m5, select=1) +
  xlab("Monthly temperature (C)") +
  ylab("Effect on log abundance") +
```

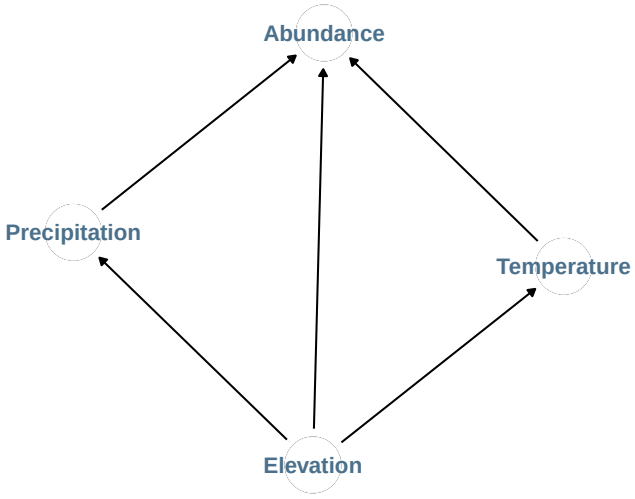
```
  theme_bw() +
  ggtitle("Effect of temperature on Ae. albopictus abundance") +
  theme(plot.title=element_text(size=11, hjust=0.5))
```



Addressing potential confounders of the temperature-abundance relationship

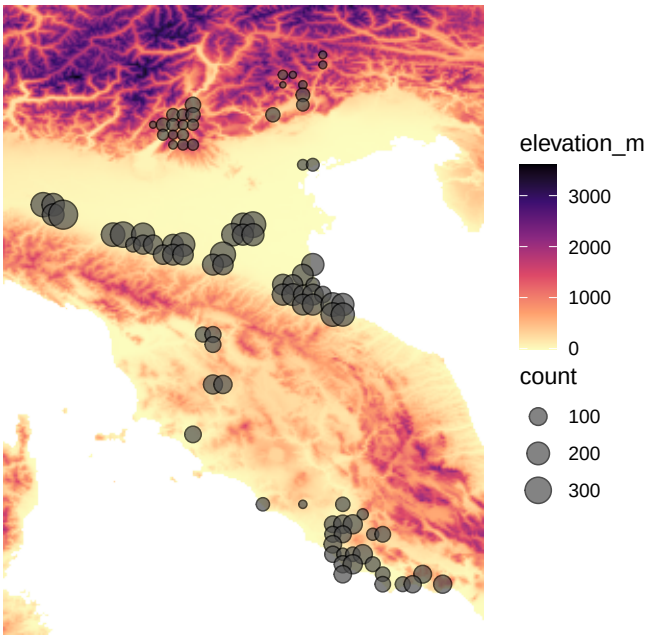
Recall this morning we discussed the concept of a **confounder** - a variable that influences both the exposure variable (temperature) *and* the outcome variable (mosquito abundance). If unaccounted for, confounders can bias our results and provide a misleading conclusion. **What could be important confounder(s) in the temperature-abundance relationship?**

One obvious candidate might be altitude, in such a topographically complex region - altitude influences both temperature and rainfall, and can also influence mosquito abundance directly, for example via distance and isolation effects, or constraining long-term niche suitability. Let's draw a simple causal diagram to explore this possibility. **Does this suggest that altitude could be a confounder? Why?**



We can address this formally by examining whether our results change when we include altitude in the model. The *data* folder includes a raster of elevation for Italy, so let's read it in and examine it.

```r
# read in and plot
alt = terra::rast("./data/elevation/altitude_metres_italy.tif")

# visualise altitude in relation to our mean counts at each location
# definitely looks like it could be playing a role
alt %>%
  as.data.frame(xy=TRUE) %>%
  ggplot() +
  geom_raster(aes(x, y, fill=elevation_m)) +
  theme_void() +
  geom_point(data=mean_counts, aes(x, y, size=count), fill="grey30", alpha=0.7, pch=21) +
  scale_color_viridis_c(name="Mean\nsummer\ncount") +
  scale_fill_viridis_c(option="magma", direction=-1) +
  coord_fixed()
```



As with last week's practical, next we extract the elevation raster values at each sampling location, then add them to our data frame.

```r
# create an sf object of survey locations
locs = dd %>%
  dplyr::select(ID, x, y) %>%
  dplyr::distinct() %>% # keep 1 record per locaton
  sf::st_as_sf(coords = c("x", "y")) %>% # set coordinates for sf
  sf::st_set_crs(crs(alt)) # harmonise CRS with altitude raster

# check
plot(locs$geometry)

# extract from the altitude raster at each location
alt_extr = terra::extract(alt, locs, ID=FALSE, na.rm=TRUE) %>%
```

```
  dplyr::mutate(ID = locs$ID)

# use left join to add this to our dataframe
dd = dplyr::left_join(dd, alt_extr)
```

Now we can use these altitude data to examine their relationship to temperature and *Aedes albopictus* abundance.

- **Q8**: Use scatter plots to investigate the relationships between temperature, elevation and mosquito abundance. Do you think elevation could be confounding this relationship?
- **Q9**: Modify your model with the nonlinear effect of temperature (*"m5"*) to also include a scaled linear effect of elevation. Fit the model and examine the estimated effects for elevation and temperature. What does this model suggest about the role of elevation in mosquito abundance? Does this model suggest that elevation is a significant confounder of the temperature-abundance relationship?

Extension exercises

The following extension exercises provide an opportunity further expand your analyses of this dataset and practice exploring these modelling methods. **I strongly recommend working through them**, either during the practical session if there is time, or during your own time. The solutions will be uploaded to Moodle next week, with an opportunity to discuss them in class.

The effect of precipitation on *Ae. albopictus* abundance

Precipitation is another critical climatic driver of mosquito population dynamics, since rainfall creates the bodies of standing water that most species need to lay their eggs ("breeding sites"). So far we have investigated the role of temperature, but the dataset also includes a column "precip_mm" that contains the total precipitation over the preceding month before sampling. In these extension exercises, we will investigate the role of precipitation.

- **Q10**: Use scatterplots and boxplots to examine the distribution of precipitation across regions and sampling locations, and to check whether there is a visual relationship between precipitation and mosquito egg counts. What do these exploratory analyses suggest?
- **Q11**: Modify your code above to fit a negative binomial model with fixed effects for region, year and month, a random intercept for survey location, and a scaled linear effect for precipitation. What does this model suggest about the effect of precipitation on abundance?
- **Q12**: Similarly to temperature, it is plausible that the effect of precipitation is nonlinear. Modify your code from Q11 to fit a model with a nonlinear effect (spline) for precipitation. Use AIC to compare the fit of this model to the model with a linear precipitation effect. Which is more strongly supported?
- **Q13**: It is possible that temperature is a confounder of the relationship between precipitation and abundance. Temperature and precipitation are negatively correlated, so hotter temperatures tend to be associated with drier conditions (*hint: you can plot a scatter plot of temperature and precipitation to see this*). Investigate whether including temperature in your model changes your estimate of the effect of precipitation. What do the findings of this model suggest?