



HAL
open science

**Support pédagogique - SIG niveau avancé -
Modélisation des dynamiques spatio-temporelles des
abondances des vecteurs du paludisme au Burkina Faso
avec le langage de programmation R**

Paul Taconet

► **To cite this version:**

Paul Taconet. Support pédagogique - SIG niveau avancé - Modélisation des dynamiques spatio-temporelles des abondances des vecteurs du paludisme au Burkina Faso avec le langage de programmation R. Master. Montpellier, France. 2023. hal-04616224

HAL Id: hal-04616224

<https://hal.science/hal-04616224>

Submitted on 18 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

TD SIG niveau avancé - Modélisation des dynamiques spatio-temporelles des abondances des vecteurs du paludisme au Burkina Faso avec le langage de programmation R

Paul Taconet, post-doctorant, IRD

Janvier 2023

Table des matières

1	Introduction	2
1.1	Préambule	2
1.2	Présentation du contexte et des données	2
2	En apéritif : Explorer et visualiser les données	5
3	Plat principal : Modéliser l'effet du paysage et des conditions météorologiques sur les abondances des anophèles	5
3.1	Charger et visualiser les données dans R	6
3.2	Modéliser l'effet du paysage	9
3.3	Modéliser l'effet des conditions météorologiques	10
4	Pour aller plus loin	13

1 Introduction

1.1 Préambule

Ce document est un exercice à destination de personnes souhaitant approfondir leurs connaissances en manipulation de données spatiales et spatio-temporelles sur R. Nous abordons des notions avancées de SIG (manipulation de données d'occupation du sol, extraction de séries spatio-temporelles au format NetCDF, modélisation spatiale) à travers un cas d'étude lié à la santé publique : la modélisation des dynamiques spatio-temporelles des abondances des vecteurs du paludisme en fonction des conditions environnementales. Notre zone d'étude est le district sanitaire de Diébougou, situé au Burkina Faso.

Cet exercice requiert les logiciels QGIS et R. Après une rapide présentation du cas d'étude théorique, nous présentons les données utilisées dans l'exercice, puis nous déroulons l'exercice. Afin de contrôler le bon avancement du tutoriel, des questions sont régulièrement posées tout au long de l'exercice. Les réponses sont disponibles en fin de document.

Prérequis pour aborder sereinement le document : Connaissances en SIG, utilisation de QGIS (manipulations basiques de données vectorielles et rasterisées), connaissances de base en R

Notions et concepts techniques abordés : extraction de données spatio-temporelles, séries spatio-temporelles issues d'images d'observation de la Terre, NetCDF, coefficient de corrélation, modélisation statistique

1.2 Présentation du contexte et des données

L'objectif principal de l'exercice est de modéliser l'abondance spatio-temporelle de vecteurs du paludisme en fonction des conditions environnementales paysagères et météorologiques - autrement dit, d'identifier les déterminants environnementaux de la présence et de l'abondance des vecteurs. En effet, les conditions météorologiques (températures, précipitations) et paysagères (utilisation, occupation du sol), impactent de nombreux traits de vie des vecteurs : émergence, croissance, survie, dispersion, activité, etc. Par exemple, la température affecte le moustique, à chaque étape de son cycle de vie (croissance larvaire, survie des adultes, etc.). De leur côté, les précipitations remplissent ou créent les gîtes larvaires et expliquent ainsi, en partie, la saisonnalité de l'abondance de certaines espèces d'anophèles. La fréquence des précipitation, leur abondance, leur durée, sont donc des paramètres essentiels pour expliquer les densités des vecteurs.

Comprendre de quelle manière l'environnement impacte la distribution et la densité des anophèles, et être en mesure de prédire ces densités à fine échelle spatio-temporelle, peut *in fine* aider à concevoir et déployer des interventions de lutte anti-vectorielle.

1.2.1 Contexte

L'exercice proposé utilise des données de terrain collectées dans le cadre du projet *REACT : Gestion de la résistance aux insecticides au Burkina Faso et en Côte d'Ivoire : recherche sur les stratégies de lutte anti-vectorielle*, mené en partenariat entre l'Institut de Recherche pour le Développement (IRD, France), l'Institut de Recherche en Sciences de la Santé (IRSS, Burkina Faso) et l'Institut Pierre Richet (IPR, Côte d'Ivoire). Ce projet était financé par L'Initiative 5%. L'objectif principal de ce projet, dont la phase de terrain s'est déroulée entre les années 2016 et 2018, était d'évaluer l'impact de l'utilisation de mesures de lutte anti-vectorielles complémentaires à la moustiquaire imprégnée d'insecticide, sur la transmission et l'épidémiologie du paludisme à travers un essai randomisé contrôlé (ERC). A cette fin, deux zones d'études ont été sélectionnées dans deux pays d'Afrique de l'ouest : le Burkina Faso (BF) et la Côte d'Ivoire (CI). Ces deux pays sont situés en zone endémiques du paludisme à *P. falciparum*.

Chaque zone d'étude du projet REACT couvre environ la surface d'un district sanitaire rural ouest-africain (~2500 km²). Il s'agit de zones principalement rurales. Pour le projet REACT, un total de 55 villages (27 au Burkina Faso, 28 en Côte d'Ivoire) a été sélectionné au sein de ces zones pour mener l'ERC selon les critères suivants : accessibilité pendant la saison des pluies, 200 à 500 habitants par village, et distance entre les villages supérieure à 2 km. La figure @ref(fig:study-areas) présente la localisation géographique des zones et des villages sélectionnés ; ainsi que le chronogramme de collectes de données effectuées dans le cadre du projet REACT.

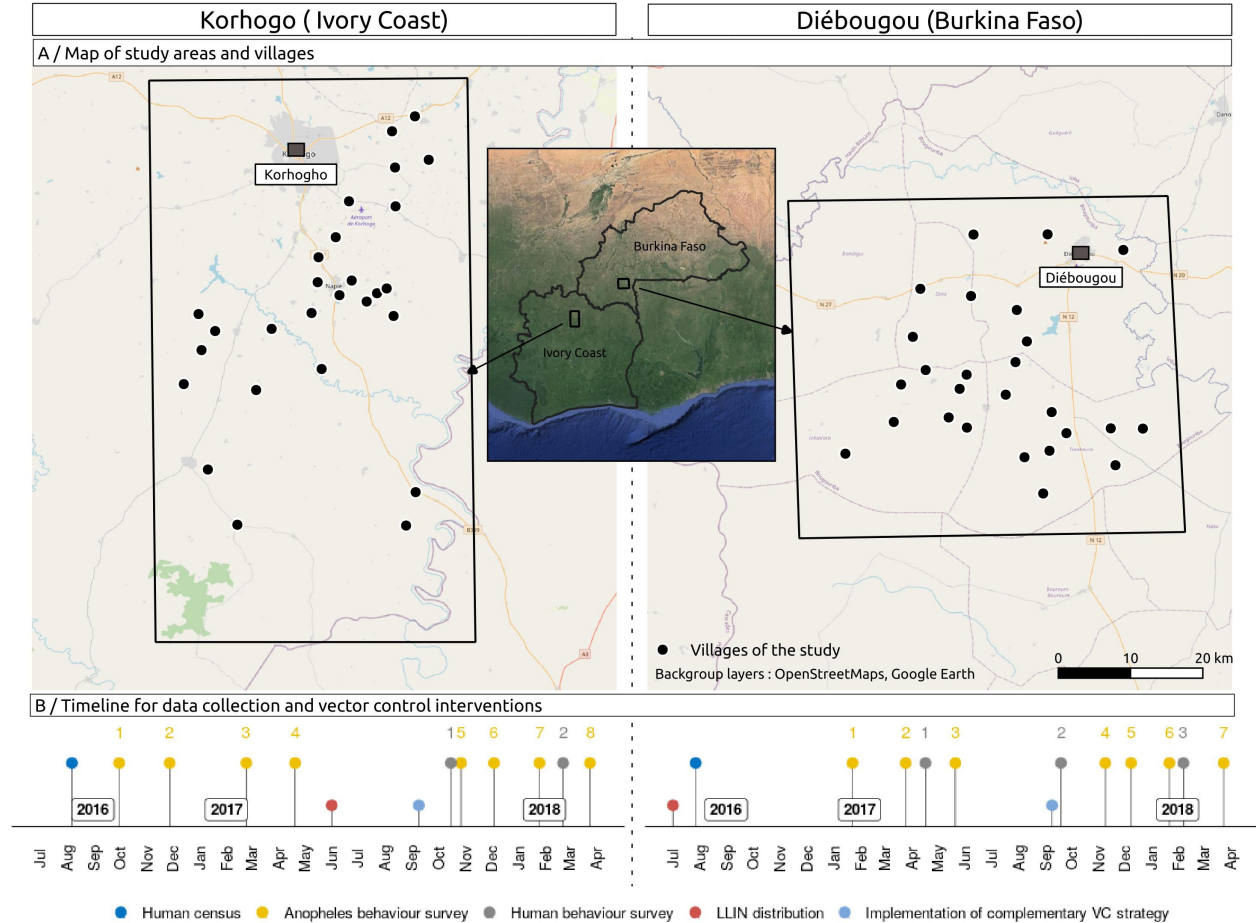


Figure 1: Projet REACT : zones d'étude, villages et dates de collectes des données

Dans cet exercice, nous allons nous focaliser sur la zone Burkinabé du projet REACT. Cette zone couvre la région de Diébougou, au sud-ouest du pays, en région bioclimatique soudanienne. Le climat y est caractérisé par une saison sèche d'octobre à avril (incluant une période 'froide' de décembre à février et une période 'chaude' de mars à avril) et une saison pluvieuse de mai à septembre. La végétation naturelle est dominée par la savane arborée parsemée de forêts ripicoles.

1.2.2 Données

Dans le cadre du projet REACT, plusieurs enquêtes entomologiques ont été effectuées dans chaque village au cours des 2 années du projet. Les données entomologiques utilisées dans cet exercice ont été recueillies dans le cadre du projet REACT. Sept enquêtes entomologiques ont été effectuées dans chaque village au cours des 2 années du projet. Les moustiques ont été collectés en utilisant la technique de la capture sur sujet humain, de 17h00 à 09h00, à raison de quatre sites (habitations) par village. Les anophèles ont ensuite été identifiés à l'espèce. Trois espèces d'anophèles principales ont été identifiées : *An. coluzzii*, *An. gambiae s.s.*, *An. funestus*. Dans cet exercice, afin de faciliter les traitements, nous allons conserver et nous focaliser sur une seule espèce : *An. gambiae s.s.*

Afin de modéliser l'effet du paysage et de la météorologie sur les abondances des moustiques, ces données de terrain ont été complétées par des données environnementales issues d'images satellitaires d'observation de la Terre.

Une carte d'occupation du sol à Très Haute Résolution Spatiale (THRS) couvrant la zone d'intérêt a été produite par classification supervisée orientée objet de produits satellitaires multisources (SPOT 6/7, Sentinel 2, MNT SRTM). Une autre carte d'occupation du sol à Haute Résolution Spatiale (HRS), a été extraite du prototype de classification de l'occupation du sol à l'échelle du continent africain produit par l'Agence Spatiale Européenne en 2016 (<https://2016africalandcover20m.esrin.esa.int/viewer.php>).

Par ailleurs, des données météorologiques sur nos zones d'étude ont été recueillies :

- données de température de surface (LST) hebdomadaires recueillies par l'instrument MODIS embarqué à bord du satellite Terra de la NASA (collection MOD11A2.061). Ces données ont été extraites en utilisant l'outil [AppEARS](#).
- données de Normalized Difference Vegetation Index (NDVI) bi-mensuelles recueillies par l'instrument MODIS embarqué à bord du satellite Terra de la NASA (collection MOD13Q1.061). Ces données ont été extraites en utilisant l'outil [AppEARS](#).
- données de cumul des précipitations par période de 10 jours issues de la base de données publique [TAMSAT](#). Ces données ont été extraites en utilisant [l'outil d'extraction des données](#) du site Tamsat.


Ces données sont disponibles dans le dossier data qui est à votre disposition :

- fichier **entomo_data.gpkg** : base de données au format Geopackage contenant les données entomologiques (table *capture_anopheles*) ainsi que certaines données géographiques contextuelles (villages de l'étude (table *villages*), points de capture des anopheles (table *metadonnees_capture*), frontières des pays d'Afrique de l'Ouest (table *contexte_paysafriqueouest*)).
- fichier **ROI.kml** : contour de la zone d'intérêt.
- fichiers **landcover_1-5m.tif**, **landcover_1-5m_pixval.csv** et **landcover_1-5m_styleQGIS.qml** : carte d'occupation du sol THRS de la zone d'intérêt au format rasterisé + table attributaire du raster (i.e. classes d'occupation du sol associées au pixels) + fichier de style pour visualiser la carte sous QGIS.
- fichiers **landcover_20m.tif**, **landcover_20m_pixval.csv** et **landcover_20m_styleQGIS.qml** : carte d'occupation du sol HRS de la zone d'intérêt au format rasterisé + table attributaire du raster (i.e. classes d'occupation du sol associées au pixels) + fichier de style pour visualiser la carte sous QGIS.
- fichier **modis_lst.nc** : fichier contenant la série spatio-temporelle MODIS LST au format NetCDF.
- fichier **modis_ndvi.nc** : fichier contenant la série spatio-temporelle MODIS NDVI au format NetCDF.
- fichier **tamsat_rainfall.nc** : dossier contenant la série spatio-temporelle TAMSAT au format NetCDF.


2 En apéritif : Explorer et visualiser les données


Avant de se lancer tous azimuts dans les analyses, nous allons tenter de comprendre le contenu des données en les cartographiant.

Cartographier le contexte dans QGIS


 Utilisez les données de contexte et les données d'occupation du sol pour réaliser une carte de la zone d'étude sous QGIS : zone d'intérêt, villages de l'étude, occupation du sol, etc.


Cartographier les données entomologiques dans QGIS

 Utilisez les données entomologiques pour réaliser une carte des abondances des vecteurs dans chacun des villages lors de chaque mission entomologique.

 : il y a plusieurs manières de faire... Vous pouvez par exemple utiliser l'outil "Agrégation" de la boîte à outils de traitements afin de moyenniser les données de captures d'anophèles par {mission, village, espèces}, suivi de l'outil "Séparer une couche vecteur" afin de séparer la couche résultante selon les missions (enquêtes) entomologiques..

Visualiser les séries spatio-temporelles dans QGIS

 Importez dans QGIS les données MODIS de températures diurnes hebdomadaires. Cartographiez, dans un dégradé (palette) de couleur rouge, la température moyenne sur la zone d'étude la semaine du 1er Novembre 2017.

 Importez dans QGIS les données TAMSAT de précipitations. Cartographiez, dans un dégradé (palette) de bleu, le cumul des précipitations au cours de cette même période de temps.

Les travaux sous QGIS sont maintenant terminés... place à R !

3 Plat principal : Modéliser l'effet du paysage et des conditions météorologiques sur les abondances des anophèles

Dans cette partie, nous allons expliciter les relations entre les densités anophéliennes et l'environnement. Pour cela, nous allons dans un premier temps charger les données entomologiques et environnementales dans R. Ensuite, nous allons constituer un ensemble de variables paysagères et météorologiques à partir des données d'occupation du sol et des séries spatio-temporelles. Ces variables vont être calculées en extrayant les valeurs des pixels des couches géographiques dans des zones tampons autour des points de capture. Enfin, nous allons visualiser la relation statistique existante entre l'abondance des vecteurs (variables à expliquer) et l'environnement (variables explicatives). A partir de l'étude de ces relations, nous allons tenter d'émettre des hypothèses sur les déterminants environnementaux des abondances des vecteurs, en lien avec leur bio-écologie.

3.1 Charger et visualiser les données dans R

3.1.1 Charger les bibliothèques R

Avant toute chose, chargeons les bibliothèques R qui vont nous servir pour la suite de l'exercice. Si elles ne sont pas installées, vous pouvez le faire en utilisant la fonction `install.packages()`.

- Bibliothèque `tidyverse`: pour la préparation des données (“data wrangling”)
- Bibliothèque `raster`: pour manipuler des données géographiques rasterisées
- Bibliothèque `stars`: bibliothèque puissante pour manipuler les fichiers rasterisés très volumineux ou les séries spatio-temporelles au format NetCDF
- Bibliothèque `sf`: pour manipuler des données géographiques vectorielles
- Bibliothèque `landscapemetrics`: pour calculer des métriques paysagères afin de constituer les variables explicatives issues de la couche d'occupation du sol

```
# Charger les bibliothèques dans R
library(tidyverse)
library(raster)
library(stars)
library(sf)
library(landscapemetrics)
```

3.1.2 Charger les données entomologiques et contextuelles

Nous allons à présent charger dans R les données entomologiques et les données contextuelles. Ces données étant vectorielles, nous allons utiliser la bibliothèque R `sf`. Cette bibliothèque fait référence pour la manipulation de données spatialisées vectorielles.

```
# lien vers la base de données geopackage comprenant les données entomologiques :
path_to_db <- "data/BF/entomo_data.gpkg"

## Importer les données
# données de captures :
captures_ano <- sf::st_read(path_to_db,"captures_anopheles")
# données de localisation et dates des captures :
metadonnees_captures <- sf::st_read(path_to_db,"metadonnees_captures")
# données de localisation des villages :
villages <- sf::st_read(path_to_db,"villages")
```

Les données sont à présent chargées dans R. Nous pouvons les ouvrir et les visualiser.

```
## Visualiser des données
# Pour visualiser les 10 premières lignes du fichier des captures, tapez simplement :
captures_ano

# Pour cartographier les données, tapez simplement :
plot(metadonnees_captures)
```

🔧 Décrivez, en une ou deux phrases simples, la table de capture des anopheles (granulométrie des données, type d'informations qui y sont disponibles, etc.) .

🔧 Faites de même pour la table des données de localisation et dates des captures.

🔧 Créez un histogramme des abondances d'anophèles capturés (indice : utilisez la fonction `hist()` sur la bonne colonne de la table `capture_ano`).

3.1.3 Charger les données paysagères

Chargeons à présent dans R les données d'occupation du sol. Pour cela, nous allons utiliser la librairie `raster`. Si `sf` fait référence pour la manipulation de données vectorisées, `raster` fait référence pour la manipulation des données rastérisées.

```
# importer les données d'occupation du sol
raster_lulc <- raster::raster("data/BF/landcover_1-5m.tif")

# importer également la table attributaire des pixels du raster
pixval_lulc <- read.csv("data/BF/landcover_1-5m_pixval.csv")
```

Comme précédemment, ouvrons les données et visualisons-les :

```
# pour voir les informations relatives au raster, tapez simplement :
raster_lulc

# pour plotter la carte d'occupation du sol :
plot(raster_lulc)
```

🔧 Identifiez les informations suivantes concernant le raster d'occupation des sols :

- résolution spatiale précise
- système de projection et de coordonnées
- nombre total de pixels
- nom de la classe d'occupation du sol à laquelle correspond les pixels dont la valeur est : 2

🔧 Maintenant que vous savez ouvrir des données rastérisées sur R, ouvrez le fichier d'occupation du sol à 20 m de résolution spatiale (`landcover_20m.tif`) et répondez aux mêmes questions que précédemment.

🔧 Plottez le raster.

3.1.4 Charger les données météorologiques

Les données météorologiques disponibles sont des séries spatio-temporelles. Le format NetCDF est particulièrement adapté à ce type de données. En plus d'être en mesure de gérer les aspects à la fois spatiaux et temporels de ces données, NetCDF est un format « auto-documenté », c'est-à-dire qu'il existe un en-tête qui décrit la disposition des données dans le fichier.

Il existe de nombreuses bibliothèques permettant de lire et manipuler les données au format NetCDF dans R : `raster`, `ncdf4`, `stars`, etc.. Nous allons ici en tester deux : `raster` et `stars`.

Commençons par la bibliothèque `stars`.

```
# Ouvrir la série MODIS LST avec la bibliothèque stars
modis_lst_stars <- read_stars("data/BF/modis_lst.nc")

# pour voir les informations relatives au netcdf, tapez simplement :
modis_lst_stars

# pour plotter la série temporelle des températures diurnes :
plot(modis_lst_stars)
```

🔧 Identifiez les informations suivantes concernant la série temporelle MODIS LST :

- système de projection et de coordonnées
- nombre et noms des attributs
- nombre total de “couches” temporelles
- dates minimum et maximum de la série temporelle
- unités des températures

🔧 Sur le plot de l'ensemble des couches temporelles, à quoi correspondent les carrés blancs (sans couleur) ?

À présent, tentons d'ouvrir les mêmes données avec la bibliothèque `raster`. Pour ce faire, nous allons utiliser la fonction `brick` du package `raster`. Cette fonction permet de charger un raster contenant plusieurs bandes (à l'image des fichiers NetCDF). Nous devons également spécifier le nom de l'attribut que nous voulons charger.

```
# Ouvrir la série MODIS LST avec la bibliothèque raster
modis_lst_raster = raster::brick("data/BF/modis_lst.nc", varname = "LST_Day_1km")
# Noter l'argument "varname", qui permet de sélectionner l'attribut désiré

# La bibliothèque n'identifie pas le système de projection. Nous allons donc le définir
# grâce à la fonction `crs`
crs(modis_lst_raster) <- "+proj=longlat +datum=WGS84 +no_defs +type=crs"

# plotter la série temporelle
plot(modis_lst_raster)

# Plotter uniquement la 4ème date :
plot(modis_lst_raster[[4]])
```

🔧 Maintenant que vous savez ouvrir des données NetCDF sur R, ouvrez le fichier de précipitations sur la zone d'étude (`tamsat_rainfall.nc`) et répondez aux mêmes questions que précédemment.

🔧 Plottez la série temporelle des précipitations (au choix avec `raster` ou `stars`)

Nous avons maintenant chargé et visualisé les différentes données. Nous sommes prêts à modéliser l'effet du paysage dans un premier temps, puis de la météorologie dans un second temps, sur les abondances des vecteurs.

3.2 Modéliser l'effet du paysage


Dans cette première partie de l'analyse, nous allons modéliser l'effet du paysage sur l'abondance des anophèles. Nous allons dans un premier temps constituer un ensemble de variables explicatives paysagères à partir des données d'occupation du sol. Ensuite, nous allons calculer une relation statistique entre ces variables et la variable à expliquer : l'abondance des anophèles.

3.2.1 Constitution des variables explicatives

Pour constituer nos variables explicatives paysagères, nous allons extraire un certain nombre de métriques paysagères (en anglais : *landscape metrics*) dans des zones tampons autour des village de l'étude. Il existe dans l'absolu un grand nombre de métriques paysagères. Dans cet exercice, nous allons calculer le pourcentage de surface utilisé par chaque classe d'occupation du sol dans une zone tampon d'un rayon de 2 km autour de chaque village.

Afin de calculer ces métriques, nous allons utiliser la librairie `landscapemetrics`.

```
# Tout d'abord, convertissons le système de coordonnées de la couche géographique des villages,  
# afin qu'il corresponde à celui du raster d'occupation du sol  
villages <- sf::st_transform(villages, crs(raster_lulc))  
  
# Puis, calculons les métriques paysagères à l'aide de la fonction `sample_lsm` du package  
# `landscapemetrics`.  
  
# La fonction qui suit permet de calculer le % de surface utilisé par chaque classe  
# d'occupation du sol dans une zone tampon d'un rayon de 2 km autour de chaque village.  
# Pour davantage de détail sur la fonction, tapez : help(sample_lsm)  
  
df_lsm <- landscapemetrics::sample_lsm(landscape = raster_lulc,  
                                     y = villages,  
                                     plot_id = villages$codevillage,  
                                     what = "lsm_c_pland",  
                                     shape = "circle",  
                                     size = 2000,  
                                     progress = T)  
  
# Afin d'obtenir le nom des classes d'occupation du sol (et pas seulement le numéro des pixels),  
# joignons la table attributaire du raster  
df_lsm <- dplyr::left_join(df_lsm, pixval_lulc, by = c("class"="pixval"))  
  
# ouvrons le fichier résultant  
df_lsm
```

 Exprimez en langage naturel (i.e. en français) l'information que fournit la première ligne du tableau.

3.2.2 Visualisation de la relation entre l'abondance des anophèles et les conditions paysagères

Afin de visualiser la relation statistique entre la variable à expliquer (l'abondance des anophèles) et les variables explicatives (% d'occupation du sol utilisé par chaque classe autour des points de capture), nous devons maintenant joindre ces deux tableaux. Nous allons les joindre en utilisant la colonne commune dans les deux tableaux : `code_village`.


```
# Tout d'abord, nous passons le tableau des métriques paysagères du format "long" au format "large"
df_lsm <- df_lsm %>%
  dplyr::select(value,plot_id,pixlabel) %>%
  pivot_wider(names_from = pixlabel, values_from = value, values_fill = 0)


# puis nous joignons les tableaux de capture des anophèles à celui des métriques paysagères
captures_ano_paysage <- dplyr::left_join(captures_ano, df_lsm, by = c("codevillage"="plot_id"))

# Visualisons à présent la relation entre le nombre d'anophèles capturés et le %
#d'occupation du sol occupé par les prairies autour des points de capture.
ggplot(captures_ano_paysage, aes(x=prairies_seches, y=n_capt)) +
  geom_point() +
  geom_smooth(method = "lm")

# Faisons de même avec la classe "marecage"
ggplot(captures_ano_paysage, aes(x=marecage, y=n_capt)) +
  geom_point() +
  geom_smooth(method = "lm")
```

 A votre avis, que représente la courbe bleue sur les graphiques précédents ?

 Selon les données, quelles classes d'occupation du sol sont positivement corrélées avec l'abondance des anophèles ? A l'inverse, quelles classes sont négativement corrélées ?

 Pouvez-vous émettre des hypothèses scientifiques pour expliquer ces relations ?

3.3 Modéliser l'effet des conditions météorologiques

Nous allons à présent étudier la relation entre les abondances des vecteurs et les conditions météorologiques, en utilisant les séries spatio-temporelles MODIS et TAMSAT.

3.3.1 Constitution des variables explicatives

Pour constituer nos variables explicatives météorologiques, nous allons extraire la moyenne des données spatio-temporelles dans des zones tampons autour des villages de l'étude, et à différents intervalles de temps précédant les captures. Nous allons ici faire le travail pour les données MODIS LST. Si vous les avez perdues, chargez de nouveau les données entomologiques et météorologiques dans R (voir section "Charger et visualiser les données dans R")

```
# La fonction suivante extrait les valeurs moyennes de chaque bande de la série MODIS LST
#dans une zone tampon de rayon 2000 m autour des villages.
df_temperature <- raster::extract(x = modis_lst_raster,
```

```

        y = villages,
        fun = mean,
        buffer = 2000,
        na.rm = T,
        sp = T)

# Observons les 10 première lignes du fichier que l'on vient de créer
head(df_temperature)

```

 Expliquez la structure d'ensemble du jeu de données `df_temperature`.

Nous allons travailler un peu le jeu de données afin de le croiser, dans la suite, avec les données d'abondance des anophèles.

```

df_temperature <- as.data.frame(df_temperature)
df_temperature <- df_temperature %>%
  dplyr::select(-c(nomvillage, population,X,Y,coords.x1, coords.x2))

# Nous passons d'une structure de données "large" à une structure "longue" :
df_temperature <- pivot_longer(df_temperature,-codevillage)

# nous convertissons la colonne des dates au format Date
df_temperature$date <- gsub("X","",df_temperature$name)
df_temperature$date <- gsub("\\\\.", "-",df_temperature$date)
df_temperature$date <- as.Date(df_temperature$date)
df_temperature$name <- NULL

# ici nous soustrayons 273.15 à la valeur de température, afin d'obtenir des températures en °C :
df_temperature$value <- df_temperature$value - 273.15

# Plottons l'évolution de la température dans les différents villages en fonction du temps :
ggplot(df_temperature, aes(x=date, y = value, group = codevillage, color = codevillage)) + geom_line()

```

A présent, il nous faut rattacher ces données de température aux données entomologiques. Pour chaque point de capture et mission entomologique, nous allons extraire les données de température pour chacune des quatre semaines précédant la capture.

```

# Nous n'avons pas besoin du caractère spatial des données de métadonnées de capture.
#Nous le supprimons donc.
metadonnees_captures <- st_drop_geometry(metadonnees_captures)
# Nous convertissons la date de capture en format Date
metadonnees_captures$date_capture <- as.Date(metadonnees_captures$date_capture)

# La fonction suivante calcule, pour chaque point de capture et enquête entomologique,
#la température moyenne pour chacune des 4 semaines précédant la date de capture
fun_get_meteo_weeks_before <- function(df_meteo, metadonnees_captures, meteo_variable){

  meteo_weeks_before <- NULL
  for(i in 1:nrow(metadonnees_captures)){

```

```

th_date_capture <- metadonnees_captures[i,c("idpointdecapture", "codevillage",
                                             "nummission", "date_capture")]
df_meteo$diff_time <- as.numeric(th_date_capture$date_capture - df_meteo$date)
th_dates <- df_meteo[which(df_meteo$diff_time>=0 &
                          df_meteo$diff_time<=40 &
                          df_meteo$codevillage==th_date_capture$codevillage),]
th_dates <- th_dates[1:4,]
th_dates$week_before_capture <- c(4,3,2,1)
th_dates <- th_dates[,c("value", "week_before_capture")]
th_dates <- pivot_wider(th_dates,
                        names_from = week_before_capture,
                        names_prefix = paste0(meteo_variable, "_week_before_"),
                        values_from = value)
th_dates <- rev(th_dates)
th_date_capture <- cbind(th_date_capture, th_dates)

meteo_weeks_before <- rbind(meteo_weeks_before, th_date_capture)
}

meteo_weeks_before$nummission <- as.numeric(meteo_weeks_before$nummission)
return(meteo_weeks_before)
}

# Executons la fonction que nous venons de créer sur le tableau 'df_temperature'
df_temperature <- fun_get_meteo_weeks_before(df_temperature, metadonnees_captures, "temp")

```

3.3.2 Visualisation de la relation entre l'abondance des anophèles et les températures

Notre jeu de données est fin prêt pour visualiser la relation entre l'abondance des anophèles et la température, à l'image de ce que nous avons fait pour les données paysagères.

```

# Joignons le tableau des températures aux données de captures d'anophèles
captures_ano$date_capture <- as.Date(captures_ano$date_capture)
captures_ano_temp <- dplyr::left_join(captures_ano, df_temperature)


# Visualisons la relation entre le nombre d'anophèles capturés et les températures moyennes
# la semaine précédant les captures
ggplot(captures_ano_temp, aes(x=temp_week_before_1, y=n_capt)) +
  geom_point() +
  geom_smooth(method = "lm")


# Maintenant, visualisons la relation entre le nombre d'anophèles capturés et les températures moyennes
# 3 semaines précédant les captures
ggplot(captures_ano_temp, aes(x=temp_week_before_3, y=n_capt)) +
  geom_point() +
  geom_smooth(method = "lm")

```

 A votre avis, que représente la courbe bleue sur les graphiques précédents ?

 Selon les données, quel est le sens de la relation entre les températures et l'abondance des anophèles ?

 Pouvez-vous émettre des hypothèses scientifiques pour expliquer ces relations ?

 Tentez de reproduire l'ensemble de l'exercice avec (i) les données précipitations puis (ii) les données de NDVI.

4 Pour aller plus loin

A l'issue de ce travail, nous n'en sommes qu'au début de l'aventure !

Nous pourrions, pour commencer, chiffrer précisément l'effet et la significativité des relations statistiques entre les variables environnementales et les abondances des vecteurs (autrement dit, construire des modèles statistiques explicatifs). A cette fin, nous pourrions construire des modèles linéaires généralisés, à effets mixtes afin de tenir compte de l'autocorrélation spatiale liée à la proximité des différents points de captures dans les villages).

Pour continuer, nous pourrions ensuite construire des modèles prédictifs de la distribution spatio-temporelle de l'abondance des vecteurs. Pour ce faire, nous pourrions construire des modèles d'apprentissage automatique multivariés (type forêts aléatoires), qui utiliseraient à la fois les variables paysagères et climatiques. Ces modèles prédictifs pourraient finalement être utilisés pour cartographier les abondances des vecteurs dans tous les villages de la zone d'étude.

Si vous souhaitez en lire davantage sur ce genre d'applications réalisées avec ces données, n'hésitez pas à consulter l'article scientifique *Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso* disponible ici : <https://doi.org/10.1186/s13071-021-04851-x>

Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale 4.0 International.