Exploring and visualizing meta-analytical datasets using R

Damien Beillouin

2024-09-26

Contents

Chapter 1

About

This training module is designed to provide an introduction to data exploration and visualization using R, with a particular focus on meta-analysis. The course is built around hands-on exercises aimed at developing essential skills for manipulating, summarizing, and presenting data in a clear, visually appealing format. By leveraging R packages such as ggplot2, cowplot, and rnaturalearth, participants will learn how to generate meaningful visual representations, such as histograms, maps, treemaps, heatmaps, and bubble plots.

The data used in this course comes from various scientific studies that analyze the impact of human interventions on various outcomes. The exercises cover a wide range of techniques, from simple bar charts to more complex visualizations like interactive maps and dynamic tables. The primary goal is to equip participants with practical tools to explore large datasets, summarize findings, and effectively communicate results in scientific contexts.

Whether you're new to R or looking to expand your data visualization skills, this module will provide you with a strong foundation in the fundamentals of data analysis and graphical presentation, emphasizing clarity, accuracy, and aesthetic quality in your work.

Chapter 2

Before starting the analyses

In the context of evidence maps and meta-analyses, data files typically contain structured information derived from primary studies. A well-organized dataset is essential for ensuring transparency, reproducibility, and clarity in statistical analyses. The structure of these files plays a crucial role in data management and visualization, particularly when handling large datasets that summarize study characteristics, interventions, and outcomes. Below are some best practices and examples to follow when preparing and using such files.

2.1 Toward transparency and reproducibility

The **FAIR principles** (Findable, Accessible, Interoperable, and Reusable) provide a framework for improving the management and sharing of digital research assets. These principles ensure that data is discoverable through search engines, accessible with appropriate authorization, interoperable with other datasets, and reusable for various purposes. A key aspect is **machine-actionability**, enabling computers to process and understand data without significant human intervention.

France has made significant strides in promoting **open science**. The **French National Plan for Open Science (2021-2024)** mandates open access to both scientific publications and research data generated with public funding.

The overarching objective of this plan is to promote transparency, accessibility, and the preservation of scientific knowledge. By mandating open access, France ensures that research funded by public resources benefits the wider global community, fostering international collaboration and cross-disciplinary advancements.

A cornerstone of this effort is the adoption of the FAIR principles (Findable, Accessible, Interoperable, Reusable), which are integral in addressing common

challenges in data management. By adhering to these principles, research data becomes more reliably reusable, supporting better documentation, accessibility, and data compatibility across different systems and disciplines.

Breakdown of FAIR Principles:

- **Findable**: Data must be easy to locate for both humans and machines. This involves assigning globally unique identifiers (such as DOIs) to datasets and ensuring that metadata is searchable and indexed in databases.
- **Accessible**: Data must be accessible under clear and transparent terms. This means storing datasets in repositories that guarantee long-term access, either through open-access platforms or specialized data journals that maintain the integrity of the data over time.
- **Interoperable**: Data should be compatible with other datasets and tools. Standardized formats (e.g., CSV, JSON) and recognized metadata structures like Dublin Core help ensure that datasets can be integrated and compared across different systems.
- **Reusable**: Data must be well-documented, with detailed metadata providing sufficient context to allow future researchers to reuse it effectively. This includes information on the dataset's provenance, context, and usage conditions, ensuring that it can be reliably understood and repurposed in new research contexts.

2.1.0.1 Reporting Standards in systematic reviews: PRISMA, ROSES, and Beyond

In systematic reviews and meta-analyses, following standardized reporting guidelines is essential for transparency and reproducibility. The most widely used framework is the **PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)** guideline, which outlines the minimum information that should be included in a systematic review, covering everything from search strategies to result synthesis. PRISMA encourages the use of flow diagrams to illustrate the study selection process, making the review process clear and replicable.

For environmental and social sciences, the **ROSES (Reporting Standards for Systematic Evidence Syntheses)** framework offers a tailored alternative. It includes checklists and flow diagrams similar to PRISMA but adapted for the specific challenges of conducting systematic reviews in complex, interdisciplinary fields like ecology, conservation, and agriculture.

Using these frameworks ensures:

1. **Transparency in Study Selection and Data Extraction**: Flow diagrams such as the PRISMA diagram clearly document how many studies were identified, screened, and ultimately included in the synthesis. This transparency helps prevent biases in study selection and allows future researchers to see the logic behind inclusion and exclusion criteria.

2. **Comprehensive Reporting of Methods and Results**: Both PRISMA and ROSES encourage detailed reporting of the data extraction process, statistical methods used in meta-analyses, and sensitivity analyses, which are crucial for assessing the robustness of results.

3. **Enhanced Reproducibility**: These guidelines ensure that other researchers can reproduce the review process, validate findings, and use the extracted data for new meta-analyses, secondary syntheses, or policy assessments.

2.2 Publishing Fully Reproducible Protocols

While pre-registration of protocols has become standard practice in fields like medicine—facilitated by platforms such as PROSPERO—it is still in the early stages of adoption within agronomy and ecology. In evidence synthesis and meta-analysis, publishing detailed and reproducible research protocols is increasingly recognized as essential for enhancing transparency and minimizing bias. This approach is well-established in medical research, where systematic reviews and meta-analyses typically adhere to stringent pre-registration guidelines. However, it has yet to gain similar traction in agronomy and ecology, highlighting an important area for growth and improvement in these disciplines. Encouraging the use of pre-published protocols in these fields would improve methodological rigor, comparability of results, and overall transparency in environmental and agricultural research. Protocols describe the step-by-step methodologies researchers intend to follow before conducting a study. They ensure transparency, reproducibility, and consistency in systematic reviews, meta-analyses, and other research designs by pre-registering the research questions, criteria for study inclusion, and planned analytical methods. This practice minimizes bias, prevents selective reporting, and enhances the credibility of findings.

2.2.0.1 Key Components of a Research Protocol

1. **Research Objectives and Questions**: Clearly defines the goals of the study and the specific research questions to be addressed.

2. **Eligibility Criteria**: Specifies which studies will be included or excluded based on predefined parameters (e.g., study design, population characteristics, intervention type).

3. **Search Strategy**: Describes the databases, search terms, and timeframe for literature searches.

4. **Data Extraction and Coding**: Outlines the methods for extracting, coding, and managing data, including variable definitions and metadata structures.

5. **Risk of Bias and Quality Assessment**: Details the criteria and tools used to assess the quality and potential biases of included studies.

6. **Analytical Plan**: Pre-specifies statistical methods, models, and subgroup analyses to be used, ensuring that analytical choices are not influenced by observed results.

2.2.0.2 Importance in Meta-Analyses and Evidence Synthesis

Publishing a detailed protocol before initiating a meta-analysis or systematic review is crucial for avoiding bias and maintaining scientific rigor. Protocols act as a roadmap, guiding researchers through the review process and serving as a reference point against which deviations can be assessed. This is particularly important for high-stakes reviews, such as those informing policy decisions or large-scale evidence syntheses in public health and environmental sciences.

Well-developed protocols also enhance collaboration and standardization within research communities by enabling other researchers to replicate or build upon the same methodology. In ecological and agronomic meta-analyses, where diverse study designs and heterogeneous data sources are common, robust protocols are indispensable for harmonizing evidence and ensuring comparability across studies.

2.2.0.3 Standards and Guidelines

Several frameworks provide comprehensive guidance for developing and publishing reproducible protocols:

- **Cochrane Handbook for Systematic Reviews**: The Cochrane Collaboration sets the gold standard for systematic reviews in health and medical research. Its protocols follow a highly structured format that emphasizes transparency, replicability, and methodological rigor.

- **ROSES (RepOrting standards for Systematic Evidence Syntheses)**: Tailored for ecological and environmental sciences, the ROSES framework outlines specific guidelines for planning and reporting systematic reviews and maps in these fields.

- **PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols)**: PRISMA-P is designed to standardize the reporting of protocols for systematic reviews and meta-analyses, ensuring all critical elements are included.

2.2.0.4 Journals Specializing in Protocols

Several specialized journals focus on publishing research protocols, providing a platform for researchers to share detailed methodological plans and facilitate reproducibility:

- **BMC Systematic Reviews**: Publishes protocols and reviews in health, social, and environmental sciences. BMC Systematic Reviews requires that all protocols adhere to PRISMA-P or similar reporting standards.
- **Protocols.io**: An open-access platform that allows researchers to publish detailed experimental protocols, workflows, and analysis pipelines. It is widely used across disciplines to promote transparent research.
- **BMJ Open**: Features protocols for any research area, including environmental, health, and social sciences. The journal emphasizes open science and reproducibility.
- **Nature Protocols**: Focuses on detailed experimental protocols in life sciences. Although primarily designed for laboratory research, it offers high visibility for methodological papers.
- **PROSPERO**: An international database for pre-registering protocols of systematic reviews focused on health and social care.

2.2.0.5 Example of a Protocol Publication

Rousset, C., Segura, C., Gilgen, A., Alfaro, M., Mendes, L.A., Dodd, M., Dashpurev, B., Bastidas, M., Rivera, J., Merbold, L. and Vázquez, E., 2024. What evidence exists relating the impact of different grassland management practices to soil carbon in livestock systems? A systematic map protocol. *Environmental Evidence*, *13*(1), p.22.

This protocol describes a systematic review and meta-analysis aimed at adapting health systems in crisis settings. The document pre-specifies all methodological details, including eligibility criteria, data extraction strategies, and planned analyses, ensuring reproducibility and transparency throughout the study.

2.2.0.6 Useful Links for Protocol Standards and Templates

- **Cochrane Handbook for Systematic Reviews of Interventions**: Cochrane Handbook
- **PRISMA-P Reporting Guidelines**: PRISMA-P Checklist
- **ROSES Guidelines for Environmental Sciences**: ROSES Reporting Standards
- **Equator Network**: A comprehensive resource for research reporting guidelines and protocol standards: Equator Network

```
useful links: https://environmentalevidencejournal.biomedcentral.com/submission-guideli
```

2.3 Publish a DataPaper

A Data Paper is a publication dedicated to describing the structure, collection, and value of a dataset. Unlike traditional research papers, which focus on findings and interpretations, Data Papers emphasize the metadata, methodology, and potential uses of the dataset itself. They offer detailed insights into how the data was gathered, processed, and structured, which is essential for reproducibility in scientific studies. Key Components of a Data Paper:

- Dataset Overview: Provides a summary of the dataset, including its purpose and potential applications.
- Metadata: Describes each variable, including units of measurement, data types, and any transformations applied.
- Collection Methods: Details the experimental or observational methods used to gather the data.
- Limitations and Uncertainties: Discloses any potential biases, gaps, or limitations in the dataset.
- Data Access: Specifies how the data can be accessed and reused, often with a permanent DOI link.

In evidence mapping and meta-analyses, the publication of Data Papers ensures that large datasets, which could be difficult to interpret otherwise, are accompanied by clear, accessible documentation. This reduces barriers to data reuse and promotes collaboration across research communities.

Journals Publishing Data Papers

Several specialized journals focus on publishing Data Papers, promoting high-quality data curation and sharing. Scientific Data(by Nature Research) and Data in Brief (by Elsevier) are prominent examples, offering platforms for data-specific publications. These journals often require the dataset to be archived in an open-access repository, accompanied by rich metadata, and adhere to rigorous peer review processes. For example, Biodiversity Data Journal also publish data papers focused on biodiversity and ecological datasets. This ensures that the data shared is of high quality, reusable, and follows best practices for transparency and openness.

Example of a Data Paper:

- Beillouin, Damien, Marc Corbeels, Julien Demenois, David Berre, Annie Boyer, Abigail Fallot, Frédéric Feder, and Rémi Cardinael. "A global meta-analysis of soil organic carbon in the Anthropocene." *Nature Communications* 14, no. 1 (2023): 3700.

- Byun, E., Müller, C., Parisse, B., Napoli, R., Zhang, J.B., Rezanezhad, F., Van Cappellen, P., Moser, G., Jansen-Willems, A.B., Yang, W.H. and Urakawa, R., 2024. A global dataset of gross nitrogen transformation rates across terrestrial ecosystems. *Scientific Data*, *11*(1), p.1022.

```
useful links:
-   CIRAD publier un Datapaper https://coop-ist.cirad.fr/gerer-des-donnees/publier-un-data-paper/

-   CINES, 2017. Les formats de fichier. https://www.cines.fr/archivage/des-expertises/les-format

-   CNRS, 2023 (version 2.0) . Guide de bonnes pratiques sur la gestion des données de recherche.

-   DoRANum, 2018. La minute Publier un Data paper. https://doi.org/10.13143/4mhn-mq42
```

2.3.1 Publishing in Open Access Journals

When submitting research for publication, consider choosing open access journals, particularly those that operate on a non-profit basis. This approach ensures that publicly funded research is readily accessible to the public, promoting transparency and facilitating broader dissemination of knowledge. Open access publishing removes paywalls, allowing researchers, practitioners, and policymakers to engage with your work without financial barriers, thereby enhancing the impact and reach of your findings. Prioritizing non-profit journals also supports sustainable publishing practices that align with the principles of open science.

2.4 Best practices for structuring Meta-analysis DataFiles

2.4.1 Generalities

1. **Consistent Naming Conventions**: Ensure that file names are clear, consistent, and meaningful. For example, naming columns such as `Study_ID`, `Outcome`, `Intervention`, and `Effect_Size` helps in avoiding confusion during data manipulation. Avoid special characters in column names, and use underscores or camel case for readability (e.g., `StudyName` or `study_name`).

2. **Comprehensive Metadata**: Metadata should accompany the main data file, providing explanations of each column and the coding used (e.g., what constitutes "intervention type" or "effect size unit"). A "Data Dictionary" should always be part of your dataset, explaining variables such as:

   - `Outcome`: The primary outcome measured in the study.
   - `Intervention`: Types of interventions, such as "land-use change" or "management."
   - `Effect_Size`: Numeric or categorical data on effect size (e.g., Hedge's g or Cohen's d).

3. **Wide vs. Long Format**: Choose the format that best suits your analysis:

- *Wide Format*: Used when each row represents a study, with multiple columns for each outcome (e.g., separate columns for effect sizes).

Field	Soil pH	Nitrogen Content (%)	Crop Yield (kg/ha)
1	6.5	45	3000
2	6.8	50	3200
3	6.2	40	2800

- *Long Format*: More suitable for meta-analysis and visualization in R. Each row contains a single observation or a study's outcome, which allows for easier aggregation, filtering, and plotting.

Field	Variable	Value
1	Soil pH	6.5
1	Nitrogen Content	45
1	Crop Yield	3000
2	Soil pH	6.8

Field	Variable	Value
2	Nitrogen Content	50
2	Crop Yield	3200
3	Soil pH	6.2
3	Nitrogen Content	40
3	Crop Yield	2800

4. **Handling Missing Data**: It's common to encounter missing data in meta-analyses. Best practices include:

    - Using a consistent code for missing values, such as `NA`.
    - Avoiding empty cells, which can cause issues when importing data into R.
    - Documenting missing data in the metadata.

5. **Version Control**: Ensure version control for your datasets. Tools like Git or a simple versioning system (e.g., `dataset_v1.csv`, `dataset_v2.csv`) can help track changes and maintain the integrity of your data over time.

6. **Data Cleanliness**: Ensure all numeric data are formatted correctly (e.g., avoid mixing numbers and text in the same column). Double-check for typographical errors, duplicates, and inconsistencies in categorical data. Tools like `dplyr::mutate()` and `tidyr::pivot_longer()` can aid in cleaning and restructuring data for analysis.

2.4.2 harmonisable classifications of practices and outcome

Meta-analysis and evidence synthesis necessitate consistent and harmonized classifications of interventions, practices, and outcomes to ensure the comparability of findings across studies and geographic contexts. In agricultural and ecological research, the diversity of practices, variations in terminology, and the complex relationships between interventions and their impacts on multiple outcomes complicate this classification task. This chapter highlights the importance of employing ontologies as a foundational step in developing harmonizable classifications. Investing the time to establish clear definitions and boundaries between classes for practices, outcomes, and site descriptions is crucial. A well-defined research question can further refine the scope, facilitating the classification process. By systematically categorizing agricultural practices and outcomes, researchers can enhance the rigor and relevance of meta-analytical studies, ultimately contributing to more robust evidence synthesis

2.5 Example: Meta-analysis datasets general datasets

To explore and utilize meta-analysis datasets, you can refer to the metadat package in R, which provides a comprehensive collection of datasets tailored for teaching, illustrating meta-analytic methods, and validating published analyses. You can install the package from CRAN using:

Once installed, you can browse available datasets by using:

```
# install metadat package
#install.packages("metadat")

# load metadat package
library(metadat)

#List of dataset included
help(package = metadat)
```

Each dataset is well-documented with metadata, including concept terms such as research field, outcome measures, and analytic models. These metadata provide insight into the structure and purpose of each dataset. Additionally, the datsearch() function allows you to search for datasets based on specific concept terms or perform a full-text search through their documentation.

The datasets in metadat follow structured formats, typically containing variables related to effect sizes, moderators, and sample information. To contribute or explore more in-depth examples, visit the package's online documentation at metadat GitHub, where you can also view the output of example analyses for each dataset

```
# load curtis databse
dat <- dat.curtis1998

# Explore curtis data
#install.packages("skimr")
library(skimr)
head(dat)
```

```
##   id paper   genus     species fungrp co2.ambi co2.elev units time pot method stock
## 1 21    44   ALNUS       RUBRA  N2FIX      350      650  ul/l   47 0.5     GC  SEED
## 2 22    44   ALNUS       RUBRA  N2FIX      350      650  ul/l   47 0.5     GC  SEED
## 3 27   121    ACER      RUBRUM  ANGIO      350      700   ppm   59 2.6     GH  SEED
## 4 32   121 QUERCUS      PRINUS  ANGIO      350      700   ppm   70 2.6     GH  SEED
## 5 35   121   MALUS   DOMESTICA  ANGIO      350      700   ppm   64 2.6     GH  SEED
```

```
## 6 38   121    ACER SACCHARINUM  ANGIO      350      700   ppm   50 2.6     GH  SEED NONE
##       m1i      sd1i n1i    m2i      sd2i n2i
## 1  6.8169 1.7699820   3 3.9450 1.1157970   5
## 2  2.5961 0.6674662   5 2.2512 0.3275839   5
## 3  2.9900 0.8560000   5 1.9300 0.5520000   5
## 4  5.9100 1.7420000   5 6.6200 1.6310000   5
## 5  4.6100 1.4070000   4 4.1000 1.2570000   4
## 6 10.7800 1.1630000   5 6.4200 2.0260000   3
```

2.5.1 Dataset for Our Exercises

We will be using the dataset titled A Global Database of Diversified Farming Effects on Biodiversity and Yield. This comprehensive dataset includes 4,076 comparisons of biodiversity outcomes and 1,214 comparisons of yield in diversified farming systems, contrasting these outcomes with two reference systems.

The dataset encompasses evidence from 48 countries and evaluates the effects of diversified farming systems on species across 33 taxonomic orders, including insects, plants, birds, mammals, eukaryotes, annelids, fungi, and bacteria. It specifically addresses systems that produce both annual and perennial crops across 12 commodity groups.

This dataset serves as a valuable resource for researchers and practitioners, facilitating access to critical information regarding the positive contributions of diversified farming systems to both biodiversity and food production outcomes.

2.5.2 Steps to Access the Dataset

1. **Load the File from Harvard Dataverse**
   Visit the Harvard Dataverse website.
   Download all files and save them in your current working directory.

By following these steps, you will be well-equipped to utilize the dataset for the upcoming exercises in this module.

```r
#installer le package pour lire des données depuis Excel
#install.packages("readxl")
#Charger le package
library(readxl)
# charge le fichier
Meta_Data <- read_excel("data/Dataset 1_sources.xlsx", sheet = "Literature_screened")
head(Meta_Data)
```

```
## # A tibble: 6 x 19
##      ID Article_source      Inclusion_yes_no Exclusion_reasion_pico Exclusion_reason
##   <dbl> <chr>               <chr>            <chr>                  <chr>
## 1    13 Stakeholder recomm~ Yes              NA                     NA
## 2    15 Scopus or WoS       No               Unsuitable outcomes    Effect on yield
## 3    18 Stakeholder recomm~ Yes              NA                     NA
## 4    25 Scopus or WoS       No               Unsuitable population  Irrelevant
## 5    30 Stakeholder recomm~ Yes              NA                     NA
## 6    34 Stakeholder recomm~ Yes              NA                     NA
## # i 11 more variables: Source.title <chr>, Volume <chr>, Issue <chr>, Art_No <chr>,
## #   Page_end <chr>, Page_count <chr>, DOI <chr>, Link <chr>, ISSN <chr>, ISBN <chr>
```

```
# For detailled summary use:
#skim(Outcome)

Outcome <- read_excel("data/Dataset_2_outcomes.xlsx", sheet = "Data")
head(Outcome)
```

```
## # A tibble: 6 x 105
##      ID Experiment_stage Comparison_ID_C Comparison_class_C Crop_C Crop_FAO_C
##   <dbl> <chr>            <chr>           <chr>              <chr>  <chr>
## 1  1033 1                C1              Natural            Forest NA
## 2  1033 1                C2              Simplified         Coffee 12 - STIMULANT CR
## 3  1033 2                C1              Natural            Forest NA
## 4  1033 2                C2              Simplified         Coffee 12 - STIMULANT CR
## 5  1033 3                C2              Simplified         Coffee 12 - STIMULANT CR
## 6  1033 4                C1              Natural            Forest NA
## # i 98 more variables: Crop_woodiness_C <chr>, crops_all_common_C <chr>, crops_all_s
## #   crops_all_scientific_level_C <chr>, System_raw_C <chr>, System_details_C <chr>,
## #   Fertiliser_C <chr>, Fertiliser_chem_C <chr>, Pesticide_C <chr>, Pesticide_quanti
## #   Soil_management_C <chr>, Time_state_C <chr>, Study_length_C <chr>, Sampling_unit
## #   B_error_measure_C <chr>, B_error_value_C <dbl>, B_error_range_l_C <chr>, B_error
## #   B_value_C <dbl>, B_SD_C <dbl>, B_N_C <dbl>, Yield_value_C <chr>, Yield_SD_C <chr
## #   Yield_error_measure_C <chr>, Yield_error_value_C <chr>, Yield_error_range_l_C <c
```

```
# For detailled summary use:
#skim(Outcome)
```

Chapter 3

Temporal Analysis for evidence maps

Temporal analysis is a critical aspect of evidence maps and meta-analyses, allowing researchers to understand how knowledge and research trends evolve over time. This analysis can reveal patterns in the accumulation of evidence, shifts in research focus, and emerging areas of interest. By systematically examining temporal trends, researchers can identify gaps in the literature, inform future research directions, and contextualize findings within broader historical or socio-political frameworks.

However, it is essential to approach temporal analysis with caution. As the volume of published literature continues to grow, there is a risk of drawing hasty conclusions based solely on trends in publication counts. Researchers must critically evaluate the context behind the data, considering factors such as changes in research funding, emerging technologies, or shifts in societal concerns that may influence publication rates.

3.1 Recommended Graph Types

When visualizing temporal data, the choice of graph type is crucial for accurately conveying trends:

- **Cumulative Sum Graphs**: These graphs represent the total number of publications over time, allowing researchers to visualize the overall growth of knowledge in a field. They can illustrate how research attention has increased, revealing long-term trends and shifts in focus.

- **Count Graphs**: These graphs show the number of publications per time period (e.g., year), enabling the identification of specific periods of increased research activity. They can highlight trends that may warrant further investigation or indicate reactions to external events.

In some cases, a combination of both graph types can provide a comprehensive view of temporal trends, illustrating both cumulative growth and specific spikes in research output.

3.2 Example in R

To illustrate temporal analysis in R, we will use the `ggplot2` and `tidyverse` packages to create both cumulative sum and count graphs. Below is an example of how to plot these two types of graphs using fictional publication data.

3.2.0.1 Initial Data Manipulation

```
# Install necessary packages (if not already installed)
# install.packages("ggplot2")
# install.packages("tidyverse")

# Load necessary libraries
library(ggplot2)
library(tidyverse, warn.conflicts = FALSE)

# Merge the outcome dataset with metadata
# TAB should be your combined dataset, with relevant variables
TAB <- left_join(Outcome, Meta_Data)
```

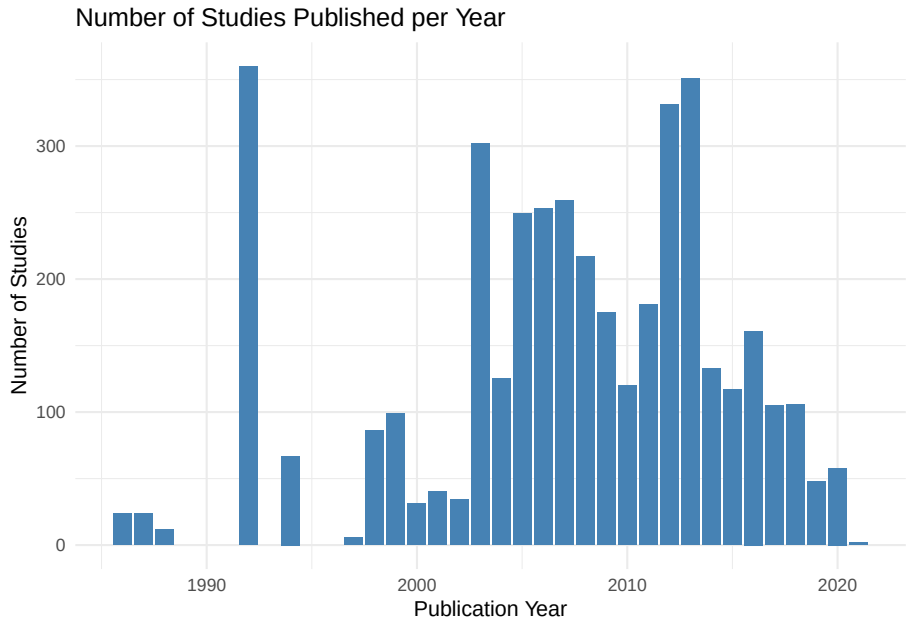
```
## Joining with `by = join_by(ID, Authors, Title, Year)`
```

3.2.1 Count Graph

This graph visualizes the count of different studies published each year. It uses a bar graph to display the number of studies per publication year, making it easy to compare across years

```
# Count Graph
count_data <- TAB %>%
  group_by(Year) %>%  # Group data by publication year
```

```
  summarise(Count = n(), .groups = "drop")  # Count studies per year

# Plot the count graph
ggplot(count_data, aes(x = Year, y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +  # Create a bar graph
  labs(title = "Number of Studies Published per Year",
       x = "Publication Year",
       y = "Number of Studies") +
  theme_minimal()  # Use a minimal theme for clarity
```

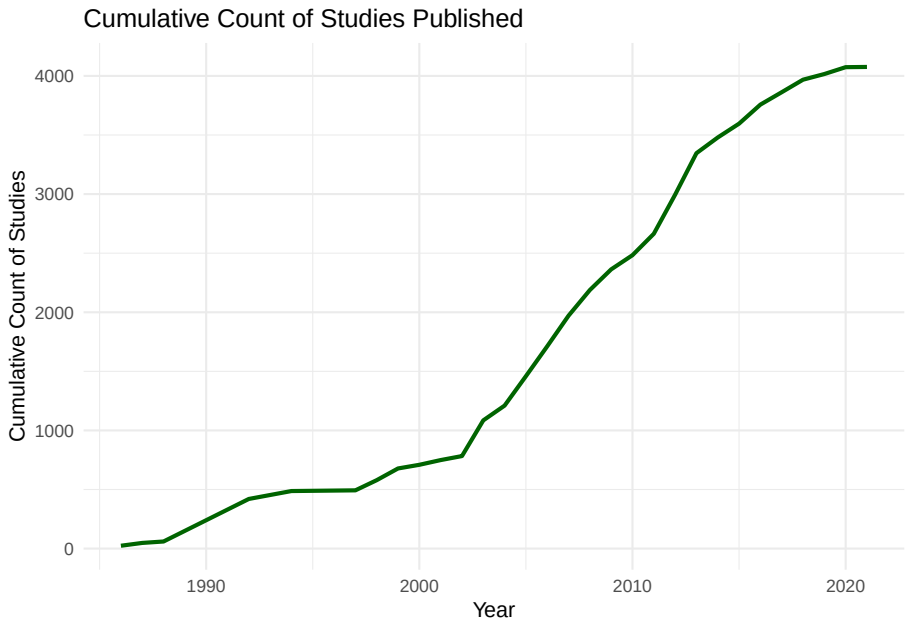


3.2.2 Cumulative Sum Graph

This graph displays the cumulative count of studies published over the years. It helps visualize how the total number of studies increases as time progresses.

```
# Cumulative Sum Graph
cumulative_data <- count_data %>%
  arrange(Year) %>%  # Ensure data is sorted by year
  mutate(Cumulative = cumsum(Count))  # Calculate cumulative count

# Plot the cumulative sum graph
ggplot(cumulative_data, aes(x = Year, y = Cumulative)) +
  geom_line(size = 1, color = "darkgreen") +  # Create a line graph
```

```
  labs(title = "Cumulative Count of Studies Published",
       x = "Year",
       y = "Cumulative Count of Studies") +
  theme_minimal()  # Use a minimal theme for clarity
```

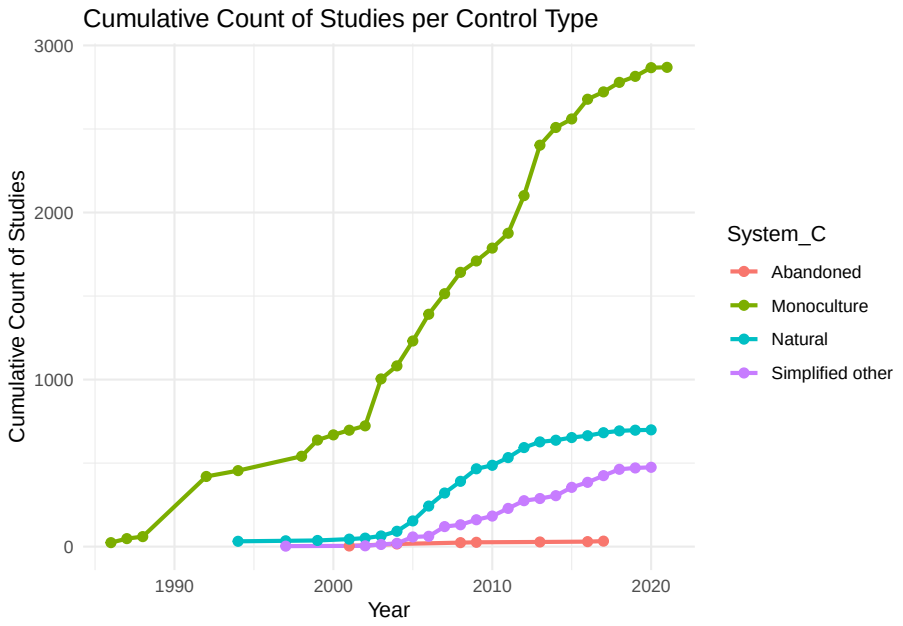


3.2.3 Grouped Graph by categories

This graph shows the cumulative count of studies based on different crop types. It allows for comparison of study trends across various crop systems.

```
# Grouped Graph by Crop
grouped_data <- TAB %>%
 group_by(Year, System_C) %>% summarise(Count = n(),.groups = "drop") %>%
  arrange(Year) %>%
  group_by(System_C) %>%
mutate(Cumulative = cumsum(Count)) %>%
  ungroup()

# Plot the grouped cumulative count graph

ggplot(grouped_data, aes(x = Year, y = Cumulative, color = System_C, group = System_C))
  geom_line(size = 1) +
  geom_point(size = 2) +
```

```
  labs(title = "Cumulative Count of Studies per Control Type",
       x = "Year",
       y = "Cumulative Count of Studies") +
  theme_minimal()
```



3.3 Brief Note on Cumulative Meta-Analyses

Cumulative meta-analyses are another important aspect of temporal analysis. They allow researchers to assess how the effect sizes of interventions or phenomena change as new studies are added over time. This approach can provide valuable insights into the robustness of findings and help track the evolution of evidence on specific topics. Cumulative meta-analyses can further enrich evidence maps by providing a more nuanced understanding of how knowledge accumulates and shifts within a research domain.

By integrating temporal analysis into evidence maps and meta-analyses, researchers can enhance the depth and relevance of their findings, ultimately contributing to a more robust and informed understanding of research trends and their implications.

Chapter 4

Spatial Analyses and Exploration for Evidence Maps

Spatial analysis is an invaluable tool for meta-analyses and evidence maps, enabling researchers to explore how intervention outcomes and environmental impacts vary across different geographic contexts. In meta-analytic research, spatial dimensions often add a critical layer of understanding that cannot be captured through temporal or non-spatial methods alone. For example, the geographic distribution of evidence can reveal clusters of research in certain regions or highlight underrepresented areas that might need targeted studies or interventions. This spatial context becomes even more relevant when interventions are applied in fields such as agriculture or ecology, where environmental heterogeneity significantly influences intervention outcomes.

In meta-analyses, spatial data is often available in the form of GPS coordinates (point data), though precision can vary. This variability poses challenges but also offers opportunities for spatial exploration at different scales. Depending on the granularity and completeness of spatial data, researchers can choose from a range of spatial units for analysis, including administrative boundaries (e.g., country, region), environmental zones (e.g., biomes), or climate classifications. By considering these spatial frameworks, it becomes possible to detect location-specific patterns, assess the transferability of interventions across ecological zones, and draw more nuanced conclusions that account for geographic variability.

Additionally, experimental data can be enriched by integrating geolocated climate or soil data from global or regional databases, such as WorldClim, TerraClimate, or the Harmonized World Soil Database. These external datasets

allow researchers to map experimental sites onto broader environmental gradients, thereby capturing key contextual factors like precipitation, temperature, or soil texture that might moderate intervention effects. This comprehensive spatial integration provides a richer understanding of the environmental conditions underlying intervention success and helps refine location-specific recommendations

4.0.1 Recommended Graph Types for Spatial Analysis

1. **Choropleth Maps**: Choropleth maps use different color intensities to represent the value of a variable across predefined spatial units (e.g., countries, regions, or districts). These maps are effective for visualizing spatial distributions of interventions or outcomes and can highlight regions of high or low research density.

2. **Whitaker Plots:** Whitaker plots are particularly useful for visualizing the distribution of study sites or intervention outcomes across ecological or climatic gradients, such as biomes or climate zones. This visualization technique is valuable in meta-analyses focusing on agroecological or environmental interventions, as it emphasizes the interaction between climatic conditions and intervention effectiveness. By illustrating how study outcomes vary across different ecological contexts, Whitaker plots enable researchers to identify key environmental factors that may influence the success of interventions, facilitating more tailored and effective recommendations.

4.0.2 Examples of Spatial Analysis in R

For practical implementation, we recommend utilizing a suite of R packages that facilitate efficient spatial data handling and visualization, including `sf`, `sp`, `terra`, `raster`, and `ggplot2`.

The `sf` package (Pebesma, 2022a) is designed for representing and working with spatial vector data, such as points, polygons, and lines, along with their associated attributes. It employs **sf** objects, which extend data frames to contain collections of simple features or spatial objects with potentially linked data.

The `terra` package (Hijmans, 2022) provides robust functions for creating, reading, manipulating, and writing both raster and vector data. Raster data is particularly valuable for representing spatially continuous phenomena, as it divides the study area into a grid of equally sized cells or pixels, each assigned a value corresponding to the variable of interest. Notably, `terra` is the latest and most powerful tool for raster analyses, offering enhanced functionality and improved performance for working with spatial data.

Furthermore, the `naturalearth` package streamlines access to country boundaries directly from the internet, thereby eliminating the need for manual downloads and ensuring users have the most current spatial data available.

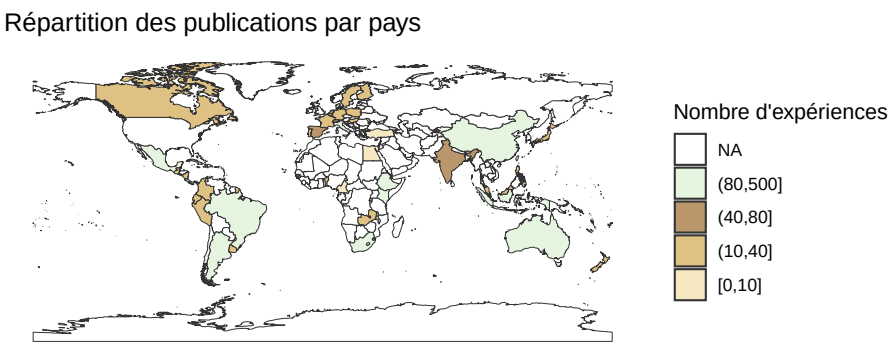
When performing spatial analyses, it is crucial to ensure that all datasets are in the same coordinate projection before extracting values based on latitude and longitude. Inconsistent projections can lead to inaccuracies in analysis, so careful attention to this detail is essential for reliable results. Below, we present examples of spatial analyses using sample data on intervention outcomes across different regions.

4.0.2.1 Example 1: Creating a Choropleth Map

This example demonstrates how to create a basic choropleth map using `ggplot2` and the `sf` package. The dataset represents intervention effectiveness scores across different European regions.

```
# Install necessary packages (uncomment to install if not already done)}
# install.packages("rnaturalearth")
# install.packages("sf")
# install.packages("dplyr")
# install.packages("ggplot2")

# Load required libraries for map creation and data manipulation
library(rnaturalearth)  # For world map data
library(sf)             # For spatial data handling
library(dplyr)          # For data manipulation
library(ggplot2)        # For visualization

# Load the dataset from the chosen CSV file
DATA <- Outcome %>%
  mutate(Country = factor(tolower(Country))) %>%
  rename(geounit = Country) %>%
  group_by(geounit) %>%
  count()

# Load the world map with medium scale
world <- rnaturalearth::ne_countries(scale = "medium", returnclass = "sf") %>%
  mutate(geounit = tolower(geounit)) %>%
  # Rename certain countries for consistency
  mutate(geounit = case_when(
    geounit == "united kingdom" ~ "uk",
    geounit == "united states of america" ~ "usa",
    TRUE ~ geounit
  ))
```

```r
# Merge the world map with the study count data
world2 <- merge(world, DATA, by = "geounit", all = TRUE) %>%
  # Create a discrete variable for legend representation
  mutate(cut_n = cut(n, breaks = c(0, 10, 40, 80, 500), include.lowest = TRUE))

# Create the map visualizing the number of publications
map_plot <- ggplot(world2) +
  geom_sf(aes(fill = cut_n), size = 0.2, color = "gray20") +
  guides(fill = guide_legend(reverse = TRUE)) +
  labs(
    fill = 'Nombre d\'expériences',
    title = 'Répartition des publications par pays',
    x = NULL,
    y = NULL
  ) +
  theme_classic() +
  scale_fill_manual(values = c('#f6e8c3', "#dfc283", '#ba966c', '#e5f5e0', '#a1d99b',

# Display the map
map_plot
```

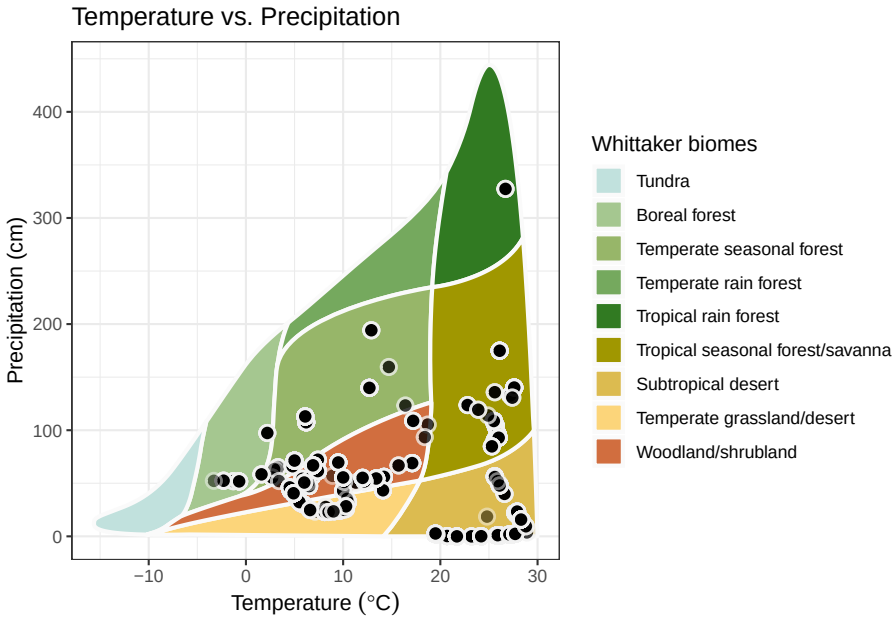


4.0.3 Example 2: Creating a Whitaker Plot

This example illustrates how to create a Whitaker plot using `ggplot2`.

```
# Install the plotbiomes package from GitHub (uncomment to install if not done yet)
# install.packages("devtools")
# devtools::install_github("valentinitnelav/plotbiomes")

# Load necessary libraries
library(plotbiomes)
library(sp)
library(raster)
library(ggplot2)

# Load temperature and precipitation raster data
path <- system.file("extdata", "temp_pp.tif", package = "plotbiomes")
temp_pp <- raster::stack(path)
names(temp_pp) <- c("temperature", "precipitation")

# Prepare spatial coordinates from Outcome dataset
coordinates <- cbind(as.numeric(Outcome$Lat_C), as.numeric(Outcome$Lat_T))
coordinates[is.na(coordinates)] <- 1  # Handle NA values
spatial_points <- SpatialPoints(coordinates)

# Extract temperature and precipitation values from the raster datasets
extractions <- raster::extract(temp_pp, spatial_points, df = TRUE)

# Adjust temperature values (WorldClim temperature data has a scale factor of 10)
extractions$temperature <- extractions$temperature / 10

# Convert precipitation from mm to cm
extractions$precipitation <- extractions$precipitation / 10

# Create a Whittaker base plot and add the temperature-precipitation data points
whittaker_base_plot() +
  geom_point(data = extractions,
             aes(x = temperature,
                 y = precipitation),
             size = 3,
             shape = 21,
             color = "gray95",
             fill = "black",
             stroke = 1,
             alpha = 0.5) +
  theme_bw() +
  labs(
```

```
    title = "Temperature vs. Precipitation",
    x = "Temperature (°C)",
    y = "Precipitation (cm)"
  )
```



Chapter 5

2D Analyses for Meta-analyses and Evidence Maps

Two-dimensional (2-D) analyses are essential tools in meta-analyses and evidence maps, allowing researchers to explore relationships between variables, detect patterns, and effectively summarize complex data. These techniques visualize interactions and distributions across two variables, providing a structured way to present findings. From contingency tables that organize categorical data to heatmaps that reveal patterns of concentration, 2-D visualizations are highly versatile and can accommodate both qualitative and quantitative information.

This chapter presents key 2-D analytical techniques—including contingency tables, heatmaps, and interactive tables—that are particularly useful in meta-analytic contexts. We will discuss when to use each approach, provide practical examples using R, and highlight their value for synthesizing and interpreting research evidence.

5.1 Key 2-D Analysis Techniques

- Contingency Tables: Contingency tables, also known as cross-tabulations, summarize the distribution of two categorical variables in a matrix format. They display the frequency or proportion of observations that fall into each category combination, helping to identify relationships or dependencies between variables. Use Case: Useful for summarizing the types of interventions applied across different farm types or regions, or for examining the association between two categorical outcomes (e.g., intervention success vs. failure across different management practices).

- Heatmaps: Heatmaps use color gradients to represent the values of a variable within a 2-D space, making it easy to spot high and low concentrations. In meta-analysis, heatmaps can visually summarize study characteristics, intervention effects, or evidence distribution across multiple dimensions. Use Case: Ideal for visualizing the intensity of evidence coverage (e.g., number of studies by region and intervention type) or effect sizes across multiple subcategories.

- Interactive Tables : Interactive tables enable researchers to explore data dynamically by sorting, filtering, and aggregating information directly within the table. Tools like DT in R or pivot tables in Excel provide flexibility for engaging with complex datasets, facilitating deeper exploration and comparison of study characteristics. Use Case: Effective for summarizing large evidence databases where users need to explore specific subsets (e.g., interventions, regions, outcomes) without losing track of the broader dataset.

5.2 Practical Examples in R

To implement these 2-D analysis techniques in R, we recommend using packages such as tidyverse, tableone, gplots, and DT for creating contingency tables, heatmaps, and interactive tables. Below, we illustrate how to construct and interpret each of these visualizations using sample data.

5.2.1 Example 1: Creating a Simple Contingency Table

A contingency table helps summarize the frequency of two categorical variables. In this example, we cross-tabulate intervention type and region to identify patterns in research distribution.

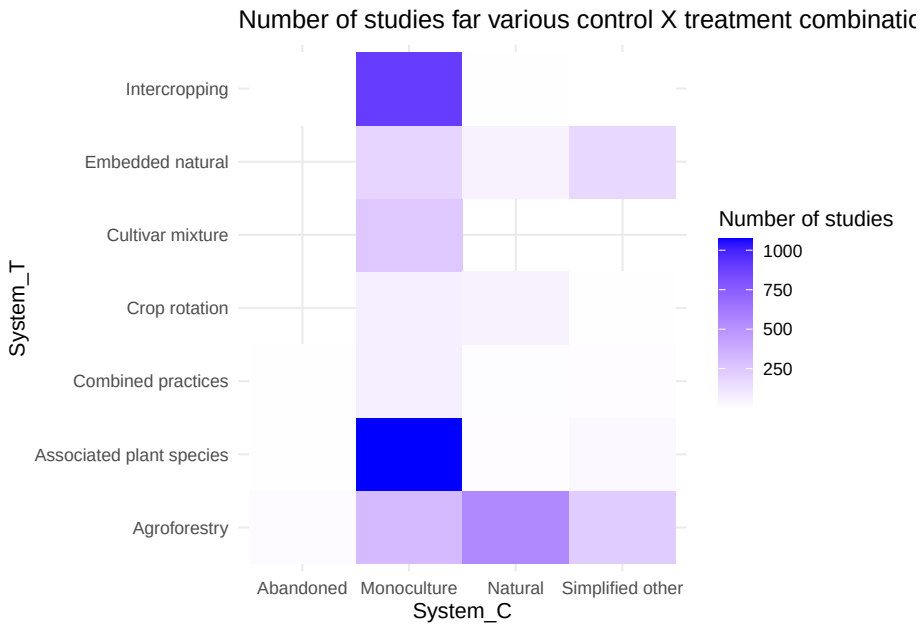
```
# Load required libraries
library(tidyverse)

contingency <- Outcome %>%
  group_by(System_C,System_T) %>%
  count()

# Create contingency table
contingency_table <- table(Outcome$Crop_FAO_C, Outcome$System_C)

# Display as a data frame
as.data.frame.matrix(contingency_table)
```

```
##                                                  Abandoned Monoculture Natural Simplified c
## 1 - CEREALS AND CEREAL PRODUCTS                          0         841       0
## 10 - SPICES                                              0           8       0
## 11 - FODDER CROPS AND PRODUCTS                           0          29      26
## 12 - STIMULANT CROPS AND DERIVED PRODUCTS                0         194       0
## 2 - ROOTS AND TUBERS AND DERIVED PRODUCTS                0          28       0
## 3 - SUGAR CROPS AND SWEETENERS AND DERIVED PRODUCTS      0           2       0
## 4 - PULSES AND DERIVED PRODUCTS                          0           4       0
## 5 - NUTS AND DERIVED PRODUCTS                            0          18       0
## 6 - OIL-BEARING CROPS AND DERIVED PRODUCTS               0         222       0
## 7 - VEGETABLES AND DERIVED PRODUCTS                      0         560       0
## 8 - FRUITS AND DERIVED PRODUCTS                          0         588       0
## 9 - FIBRES OF VEGETAL AND ANIMAL ORIGIN                  0         305       0
## NA                                                      25           0     673
## Other or nd                                              8          70       0
```

5.2.2 Example 2: Generating a Heatmap

Heatmaps are excellent for visualizing interactions or intensities. This example uses a heatmap to display intervention effect sizes across different farm types.

```r
# Load required libraries
library(ggplot2)

# Create sample data
heatmap_data <- Outcome %>%
  group_by(System_C,System_T) %>%
  count()

# Create heatmap
ggplot(heatmap_data, aes(x = System_C, y = System_T, fill = n)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Number of studies far various control X treatment combination",
       fill = "Number of studies")
```

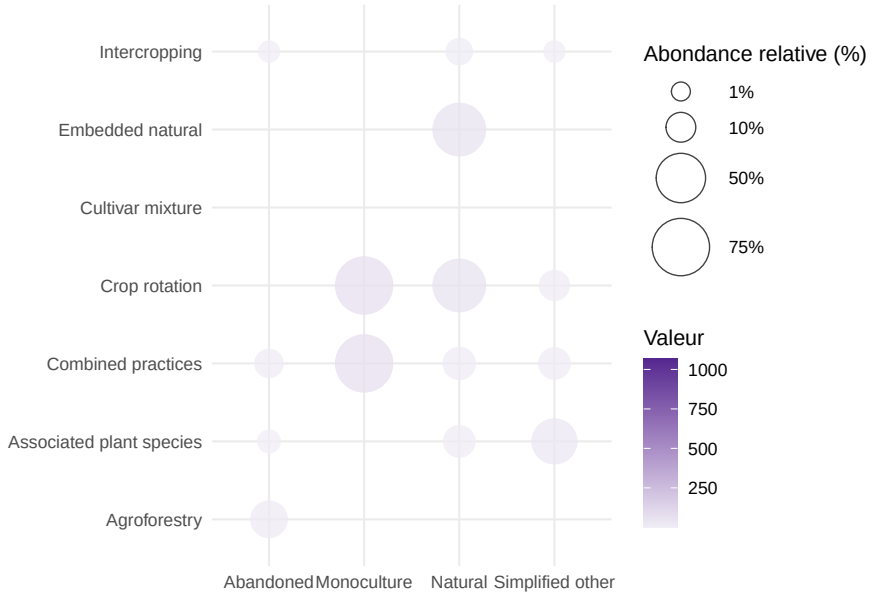


5.2.3 Example 3: Bubble plots

Bubble plots are useful for showing interactions between different categories and the size of intervention effects. They encode information through the position, size, and color of the bubbles, making it easy to see patterns and differences. For example, a bubble plot can show intervention effects across various farm types, with larger bubbles indicating stronger effects. This helps highlight which interventions work best in different contexts.

```
# Load required libraries
library(ggplot2)

# Create sample data
bubble_data <- Outcome %>%
  group_by(System_C,System_T) %>%
  count()

# Créez un graphique à bulles
bubble_plot <- ggplot(bubble_data, aes(x = System_C, y = System_T)) +
  geom_point(aes(size = n, fill = n, color = n), alpha = 0.75, shape = 21) +
  # Échelle personnalisée pour la taille des bulles
  scale_size_continuous(
    limits = c(0.000001, 100),  # Limites personnalisées
    range = c(3, 15),           # Plage de tailles personnalisées
```

```
    breaks = c(1, 10, 50, 75),  # Points de rupture pour les étiquettes de taille
    labels = c("1%", "10%", "50%", "75%")  # Étiquettes pour les tailles
  ) +
  # Personnalisez les légendes
  labs(
    x = "",
    y = "",
    size = "Abondance relative (%)",
    fill = "Valeur",
    color = "Valeur"
  ) +
  # Personnalisez les thèmes pour une meilleure lisibilité
  theme_minimal() +
  # Personnalisez la palette de couleurs
  scale_fill_gradient(low = "#F1EEF6", high = "#54278F") +
  scale_color_gradient(low = "#F1EEF6", high = "#54278F")

bubble_plot
```



5.2.4 Example 3: Building an Interactive Table

Interactive tables are useful for large datasets that require filtering or detailed inspection. The `DT` package in R makes it easy to create tables that users can

sort, search, and explore.

```
library(reactable)
library(dplyr)
library(webshot)

# Create a summary of grouped data
GROUP <- Outcome %>%
  group_by(System_C) %>%
  summarize(Number = n(), .groups = "drop")


react<-reactable(
  GROUP,
  details = function(index) {
    # Filter the details for the selected group
    details_data <- filter(Outcome, System_C == GROUP$System_C[index])

    # Create a reactable for the detailed view
    tbl <- reactable(details_data,
          columns = list(
  System_C = colDef(
    style = function(value) {
      if (value > 0) {
        color <- "#008000"
      } else if (value < 0) {
        color <- "#e00000"
      } else {
        color <- "#777"
      }
      list(color = color, fontWeight = "bold")
    }
  )
))
    htmltools::div(style = list(margin = "12px 45px"), tbl)
  },
  onClick = "expand",  # Set click behavior to expand row details
  rowStyle = list(cursor = "pointer")  # Change cursor style to pointer for rows
)

#react
```

Bibliography

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL http://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2024. URL https://github.com/rstudio/bookdown. R package version 0.40.