Foundational Soft Skills

MACEPA Data Fellowship - Training Materials

Table of contents

Foundational Soft Skills Module **3**
What You Will Learn 3

1 Module Content **5**

2 Tutorial - How to turn a bad slide into a good slide **6**
2.1 Libraries and Data Preparation 6
2.1.1 Data Preparation 7
2.2 Initial Visualization 9
2.2.1 The Raw Plot 9
2.2.2 Aesthetic Improvements 11
2.2.3 Scientific story telling issues 13
2.3 Technical Enhancements 14
2.3.1 Population-Adjusted Plot 14
2.4 Exercises 18

3 True Colours Exercise **19**

Foundational Soft Skills Module

Welcome to the Foundational Skills module of the MACEPA Data Fellowship Program. This module is designed to equip fellows with the critical soft skills necessary to succeed in a fast-paced, collaborative, and data-driven environment.

What You Will Learn

Throughout this module, we will explore key competencies required for effective work in MACEPA and beyond, including:

- **Organizational Structure and Collaboration:** Understanding how to navigate MACEPA's work environment and effectively collaborate with your team and external partners.
- **Time Management and Prioritization:** Developing strategies to manage competing priorities and deadlines while maintaining productivity.
- **Feedback Culture:** Mastering the art of giving and receiving feedback to foster professional growth and strengthen team dynamics.
- **Addressing Common Challenges:** Identifying and overcoming challenges in data and analytics work, from data accessibility to engaging stakeholders.
- **Effective Communication:** Learning how to convey complex information clearly and impactfully through writing, presentations, and visualizations.

Success in the field of malaria data analytics requires more than technical expertise. It calls for strong interpersonal skills, strategic planning, and effective communication. This module provides a foundation to help you navigate the complexities of your role and make meaningful contributions to the fight against malaria.

1 Module Content

Find below the slide deck for this module - there are several sub topics contained within this work with links to interactive videos and courses included so make sure you take your time and work through all of this.

2 Tutorial - How to turn a bad slide into a good slide

This tutorial follows on from the How to turn a bad slide into a good slide section of the powerpoint presentation. It will walk through the code used to generate the plots and includes some follow on exercises for you to complete with this same dataset.

2.1 Libraries and Data Preparation

First lets load in the necessary R packages and the dataset for this tutorial - which you can download from the box folder.

> **Tip**
>
> Make sure you save the data in the same folder as your associated R Project for this module.
> If you haven't already installed the necessary R packages for this tutorial you will need to call the install.packages before loading the library.
> For help installing `PATHtoolsZambia` see the package web-page for more details.

```
# Load libraries
library(tidyverse)
library(PATHtoolsZambia)
library(scales)
library(sf)
library(ggpubr)

# Load the data
```

```r
dat <- read.csv("monthly-cases.csv")
```

2.1.1 Data Preparation

Our motivating question here is: **What is the malaria trend in Northern Province Zambia since 2018?**

Lets take a quick look at this dataset and see what kind of data we are working with. We have 5 columns `period` with monthly values from 2018 - June 2024, `reported_district` with names of all the districts in Northern Province, `data_type` shows our case data is in long format with values of clinical, confirmed and Confirmed_Passive_CHW, `age_group` again is in long format with categories of Under 5 and Over 5 and finally the `total` column that provides the number of cases reported.

```r
# see column names
head(dat)
```

```
      period reported_district data_type age_group total
1 2018-01-01           Chilubi  Clinical   Under 5    14
2 2018-01-01           Chilubi  Clinical    Over 5    30
3 2018-01-01           Chilubi  Clinical   Under 5     6
4 2018-01-01           Chilubi Confirmed   Under 5  1015
5 2018-01-01           Chilubi Confirmed    Over 5  1657
6 2018-01-01           Chilubi Confirmed   Under 5   372
```

```r
# Display each column's unique values to explore data options
dat %>% select(-total) %>% map(~ table(.))
```

```
$period
.
2018-01-01 2018-02-01 2018-03-01 2018-04-01 2018-05-01 2018-06-01 2018-07-01
        59         60         60         52         58         54         49
2018-08-01 2018-09-01 2018-10-01 2018-11-01 2018-12-01 2019-01-01 2019-02-01
        55         54         50         55         56         55         51
2019-03-01 2019-04-01 2019-05-01 2019-06-01 2019-07-01 2019-08-01 2019-09-01
        57         62         60         64         62         63         61
```

```
2019-10-01 2019-11-01 2019-12-01 2020-01-01 2020-02-01 2020-03-01 2020-04-01
        56         59         75         75         76         75         77
2020-05-01 2020-06-01 2020-07-01 2020-08-01 2020-09-01 2020-10-01 2020-11-01
        76         76         72         68         68         70         72
2020-12-01 2021-01-01 2021-02-01 2021-03-01 2021-04-01 2021-05-01 2021-06-01
        72         74         74         72         75         73         68
2021-07-01 2021-08-01 2021-09-01 2021-10-01 2021-11-01 2021-12-01 2022-01-01
        66         69         70         71         80         81         80
2022-02-01 2022-03-01 2022-04-01 2022-05-01 2022-06-01 2022-07-01 2022-08-01
        80         80         76         77         81         78         75
2022-09-01 2022-10-01 2022-11-01 2022-12-01 2023-01-01 2023-02-01 2023-03-01
        74         81         73         74         79         80         84
2023-04-01 2023-05-01 2023-06-01 2023-07-01 2023-08-01 2023-09-01 2023-10-01
        87         86         83         81         79         88         83
2023-11-01 2023-12-01 2024-01-01 2024-02-01 2024-03-01 2024-04-01 2024-05-01
        82         82         89         89         96         90         91
2024-06-01
        91

$reported_district
.
  Chilubi    Kaputa    Kasama     Lunte Lupososhi   Luwingu     Mbala Mporokoso
      581       368       510       431       426       379       606       323
 Mpulungu    Mungwi     Nsama     Senga
      464       423       487       608

$data_type
.
             Clinical             Confirmed Confirmed_Passive_CHW
                 1803                  2810                   993

$age_group
.
 Over 5 Under 5
   1946    3327
```

```r
# correct date data from character string to date variables:
dat$period <- as.Date(dat$period)
```

The next step is to aggregate this data up across all of the Districts in Northern Province as we are interested in the Province

as a whole, we will also aggregate across `age_groups` so we have the total population level totals for each `data_type`.

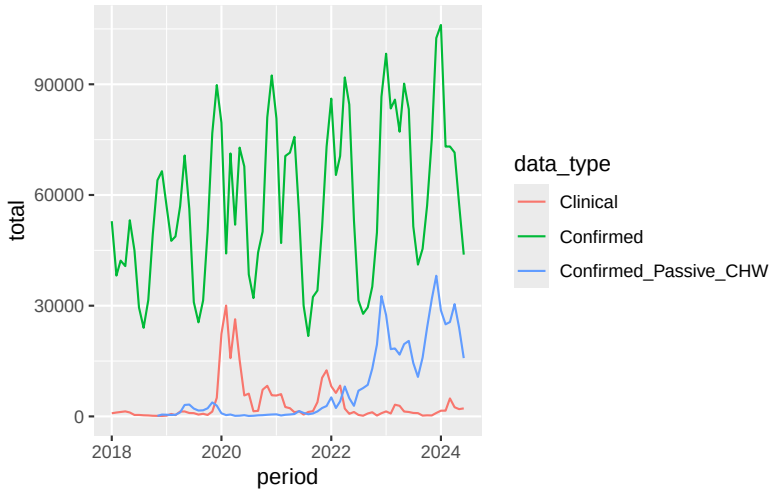
```
# Summarize data at the province-month level across all age groups
dat_sum <-
  dat %>%
  group_by(period, data_type) %>%
  summarise(total = sum(total, na.rm = TRUE)) %>%
  ungroup() %>%
  filter(data_type %in% c("Clinical", "Confirmed", "Confirmed_Passive_CHW"))
```

2.2 Initial Visualization

2.2.1 The Raw Plot

Lets create a simple `ggplot()` of this dataset. What do we notice? Malaria appears to be increasing in Northern Province over the period from 2018 - 2024.

```
# Create an initial raw plot
ggplot(dat_sum,
       aes(x = period, y = total, color = data_type, group=data_type)
       ) +
  geom_line()
```



This plot can show us that Malaria appears to be increasing in Northern Province over the period from 2018 - 2024. However, there are several problems with this plot:

- No plot title
- Thin lines that are hard to see
- Hard to see how all the data types add together to get a total malaria trend
- Boring use of color (and potential issues for red-green color blind people)
- Untidy legend (no proper title, underscores between words)
- Figure labels too small
- Timeseries are squashed
- Axes not informatively labeled
- Numbers don't have commas (i.e 60,000)

2.2.2 Aesthetic Improvements

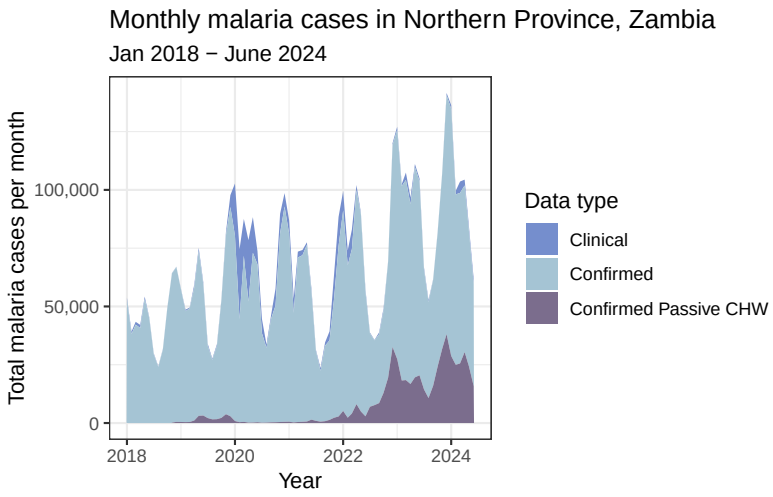
The best approach to making improvements to plots is to iteratively make small changes to enahance each aspect of the plot. Such as the following steps:

Lets first remove the "_" from one of our `data_type` variables for better readability.

```
# where data_type value is "Confirmed_Passive_CHW" replace with "Confirmed Passive CHW" othe
dat_sum <-
  dat_sum %>%
  mutate(nice_names =
           ifelse(data_type == "Confirmed_Passive_CHW", "Confirmed Passive CHW", data_type)
         )
```

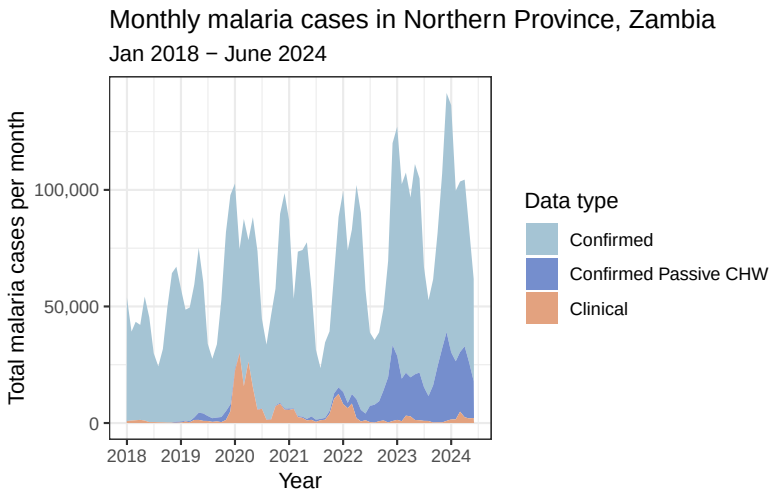
Then we can make some additions to the `ggplot()` code to further enhance the plot - each is described in the code below:

```
# Create an improved plot
ggplot(dat_sum,
       aes(x = period, y = total, fill = nice_names, group=nice_names) # use nice_names for
       ) +
  geom_area() + # replace lines with a shaded area plot
  scale_y_continuous(labels = comma) + # add comma separators to the numbers on the yaxis
  labs(x = "Year", y = "Total malaria cases per month") + # include informative axis labels
  ggtitle("Monthly malaria cases in Northern Province, Zambia", # include a plot title and s
          subtitle = "Jan 2018 - June 2024"
          ) +
  scale_fill_manual("Data type", values = c("#758ECD", "#A5C4D4", "#7B6D8D")) + # change the
  theme_bw() # use an inbuilt ggplot theme
```



This is a good step in the right direction but it is still difficult to see the number of clinical cases when stacked at the top and the colours we have chosen are also slighty hard to distinguish so lets address these two aspects:

```
# use factoring to alter the order of data_type in the plot
dat_sum <-
  dat_sum %>%
  mutate(nice_names = factor(nice_names, levels = c("Confirmed",
                                                    "Confirmed Passive CHW",
                                                    "Clinical")
                             )
         )

# Plot this data with new colurs
ggplot(dat_sum, aes(x = period, y = total, fill = nice_names, group=nice_names)) +
  geom_area() +
  scale_y_continuous(labels = comma) +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  labs(x = "Year", y = "Total malaria cases per month") +
  ggtitle("Monthly malaria cases in Northern Province, Zambia",
          subtitle = "Jan 2018 - June 2024") +
  scale_fill_manual("Data type", values = c("#A5C4D4","#758ECD", "#E3A27F")) + #changes colo
  theme_bw()
```



2.2.3 Scientific story telling issues

As noted in the slides, while this plot is a visual improvement on the first we need to also address how we interpret this plot and the potential missing pieces to answer our analysis question.

- Our initial description of the results: "Malaria appears to be increasing in Northern Province over the period from 2018 - 2024" isn't very informative and is a limited description of the results.
- Plot shows total malaria cases – doesn't account for population growth
- Cant really tell the extent to which cases are increasing over time – hard to aggregate visually over 12 months of data to assess annual trends

Is more context needed to starting thinking of solutions? Are these increases occurring over all districts in the province? Are increases greater in under 5s or over 5s? Has coverage of interventions decreased over time?

2.3 Technical Enhancements

To start addressing some of these issues we can add in some additional context to this plot - through using population data we can add in an understanding of how malaria incidence as well as raw case counts are changing over time.

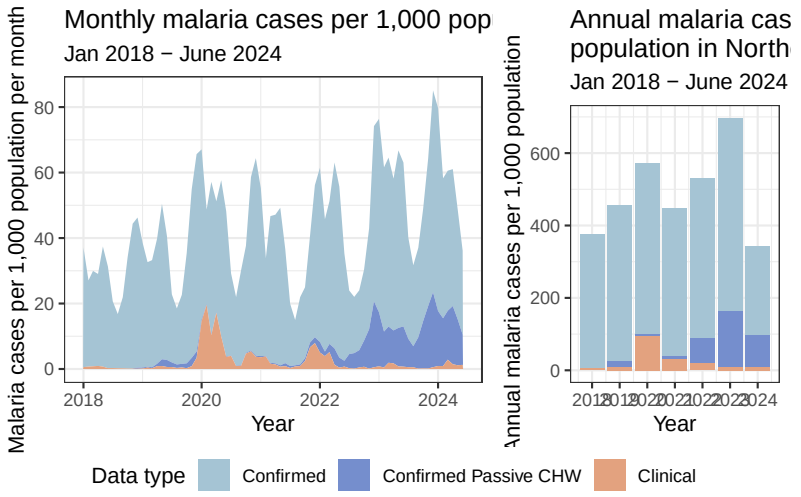
```
# Retrieve population data from the PATHToolsZambia package - this is the population totals
pop_northern_2022 <-
  retrieve("province-shp") %>%
  st_drop_geometry() %>%
  filter(geo_province == "Northern") %>%
  pull(census_pop_22)

# Calculate incidence
dat_sum_inc <-
  dat_sum %>%
  # data is from 2022 so rename column for ease of use
  mutate(pop_22 = pop_northern_2022) %>%
  # extract year data from the period column
  mutate(year = year(period)) %>%
  # scale the population total data for the years that we are missing this data assuming a p
  mutate(pop = scale_pop_growth_annual(initial_pop = pop_22, new_year = year, initial_year =
  # calculate incidence per 1000 population
  mutate(monthly_inc_per_1000 = total / pop * 1000)
```

2.3.1 Population-Adjusted Plot

We can view this incidence data both as a timeseries and summarised at the annual level.

```
# Create area plot with population-adjusted data
p1 <-
  ggplot(dat_sum_inc, aes(x = period, y = monthly_inc_per_1000, fill = nice_names)) +
  geom_area() +
  scale_y_continuous(labels = comma) +
  labs(x = "Year", y = "Malaria cases per 1,000 population per month") +
  ggtitle("Monthly malaria cases per 1,000 population in Northern Province, Zambia",
          subtitle = "Jan 2018 - June 2024") +
```

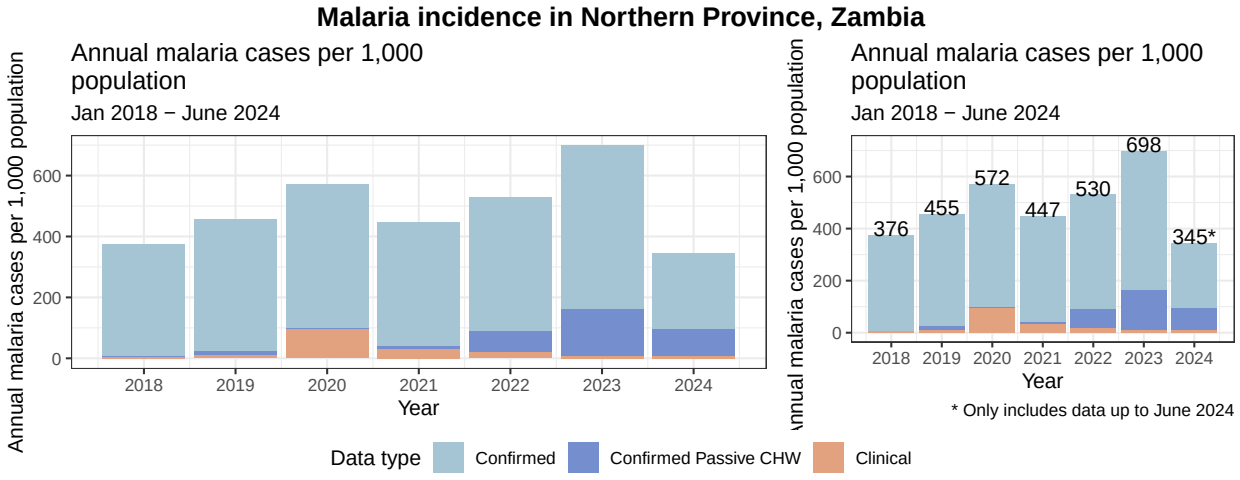
```r
  scale_fill_manual("Data type", values = c("#A5C4D4", "#758ECD", "#E3A27F")) +
  theme_bw()

# summing data to the yearly level for each data type
dat_sum_inc_annual <-
  dat_sum_inc %>%
  group_by(nice_names, year) %>%
  summarise(total_inc = sum(monthly_inc_per_1000)) %>%
  ungroup()

# create a bar chart of annual data
p2 <-
  ggplot(dat_sum_inc_annual, aes(x = year, y = total_inc, fill = nice_names)) +
  geom_col() +
  scale_x_continuous(breaks = 2018:2024) +
  labs(x = "Year", y = "Annual malaria cases per 1,000 population") +
  ggtitle("Annual malaria cases per 1,000 \npopulation in Northern Province, Zambia",
          subtitle = "Jan 2018 - June 2024") +
  scale_fill_manual("Data type", values = c("#A5C4D4","#758ECD", "#E3A27F")) +
  theme_bw()

# combine the plots into a single image
p_comb <- ggpubr::ggarrange(p1, p2, common.legend = TRUE, legend = "bottom",
                  widths = c(2,1.2))

p_comb
```



The addition of the bar graph makes it easier to see the trends, but would be even easier if we could read off the total incidence on each bar, and it is not necessarily clear to the reader that 2024 only includes 6 months of data. In addition we have cut the title off the plots short with the sizing so lets fix all of that.

```
# calculating bar totals to add to plot - this is the combined total of each of the data_typ
dat_sum_inc_annual_tot <-
  dat_sum_inc_annual %>%
  group_by(year) %>%
  summarise(total = sum(total_inc)) %>% #summing total values
  ungroup() %>%
  mutate(bar_label = round(total, 0)) %>% #rounding totals to remove decimal places
  mutate(bar_label = ifelse(year ==  2024, paste0(bar_label, "*"), bar_label)) #including *

# Plot these changes
p3 <-
  ggplot(dat_sum_inc_annual, aes(x = year, y = total_inc, fill = nice_names)) +
  geom_col() +
  scale_x_continuous(breaks = 2018:2024) +
  labs(x = "Year", y = "Annual malaria cases per 1,000 population") +
  ggtitle("Annual malaria cases per 1,000 \npopulation",
          subtitle = "Jan 2018 - June 2024") +
```

```
  scale_fill_manual("Data type", values = c("#A5C4D4","#758ECD", "#E3A27F")) +
  theme_bw()


p4 <-
  ggplot() +
  geom_col(data = dat_sum_inc_annual, aes(x = year, y = total_inc, fill = nice_names)) +
  geom_text(data = dat_sum_inc_annual_tot, aes(x = year, y = total + 25, label = bar_label))
  scale_x_continuous(breaks = 2018:2024) +
  labs(x = "Year", y = "Annual malaria cases per 1,000 population",
       caption = "* Only includes data up to June 2024" #include caption note about 2024 mis
       ) +
  ggtitle("Annual malaria cases per 1,000 \npopulation",
          subtitle = "Jan 2018 - June 2024") +
  scale_fill_manual("Data type", values = c("#A5C4D4","#758ECD", "#E3A27F")) +
  theme_bw()

p_comb2 <- ggpubr::ggarrange(p3, p4, common.legend = TRUE, legend = "bottom",
                             widths = c(2,1.2))

# add a combined figure title
annotate_figure(p_comb2, top = text_grob("Malaria incidence in Northern Province, Zambia",
                                         face = "bold", size = 14))
```



With this updated figure we can provide clearer key messages

to our audience:

- There was an increase in malaria cases in 2023, to 698 cases per 1,000 population a **32% increase** on 2022
- Since 2022, there has been an increasing proportion of malaria cases **detected in the community**, potentially contributing to increased case reports

2.4 Exercises

Can you use a similar approach to provide a slide to answer the following question: **Is this increase consistent across all districts, or is it focused in a few places?**

> Tip
>
> `ggplot` includes an excellent faceting feature (`facet_wrap` and `facet_grid`) that you might find useful to answer this question.
> `retrieve("district-shp")` will be useful when retrieving the population data at the district level.

3 True Colours Exercise

We covered the True Colours Excercise during our Welcome Week sessions - the slides can be found here for your reference.