Vignette: commonweights package

Biostat Global Consulting

Package installation

The easiest way to use the `commonweights` package is to install it directly from GitHub:

```
if (!requireNamespace("pak")){install.packages("pak")}
pak::pkg_install("BiostatGlobalConsulting/commonweights")
```

Once the package is installed, load it:

```
library(commonweights)
```

Harmonia dataset

We'll use an example dataset with three survey samples from the imaginary country of Harmonia.

```
data("fauxdata")
```

The dataset includes samples from 2005, 2010, and 2015. In addition to the year variable, the dataset has stratum and cluster IDs, a variable indicating the child's age in years, the `psweight` variable which shows each child's survey weight (the number of children in the population represented by this respondent), and three dichotomous variables: possession of a vaccination card, vaccination with DTP3, and vaccination with MCV1.

Here are the first few rows of the faux dataset:

geo	year	stratumid	clusterid	age	psweight	has_card	dtp3_vx	mcv_vx
District 01	2005	1	1	1	3411.1	1	1	1
District 01	2005	1	1	1	3411.1	1	0	1
District 01	2005	1	1	1	3411.1	1	1	0
District 01	2005	1	1	1	3411.1	0	0	1
District 01	2005	1	1	1	3411.1	0	0	1
District 01	2005	1	1	1	3411.1	1	1	1

Using the cwt_output function

The core function in the commonweights package is called `cwt_output`. It has the following arguments:

data Dataset to analyze. Should have one row per respondent and contain survey data from at least two years.

outcomevars Names of dichotomous outcome variable(s) coded `0/1/NA`. Provide as a single variable name e.g. `"dtp3_vx"` or as multiple variables in a concatenated list, e.g. `c("dtp3_vx", "mcv1_vx")`

outcomenames Names of the outcomes in `outcomevars` to use in tables and plots; must be the same length as `outcomevars`. Optional, defaults to NULL. E.g. `c("DTP3", "MCV1")`

geovar Name of geography variable in `data`. Must have the same levels across all years in `years`.

yearvar Variable in dataset indicating survey year

weightvar Variable in dataset containing survey weights for each respondent

years Data years to analyze. Defaults to `"all"`, which will analyse all years in `yearvar`. To analyze a subset of years, provide a vector of years to include, e.g. `c(2000, 2005, 2010)`

weightopt Which year(s) to use as common weights. `"earliest"` will calculate outcomes post-stratified to the earliest year in `years`. `"previous"` will calculate outcomes for each survey year using the weights from the previous survey year. `"custom"` allows the user to select a year to use weights from.

weightyear Year to use for post-stratifying if `weightopt = "custom"`

age Age to filter by. Defaults to NULL. Provide as a single value of `agevar` to keep or provide as a pair of values for minimum and maximum ages, e.g. `age = c(12, 23)` will keep `agevar >= 12` and `<= 23`.

agevar Age variable in dataset; used if `age` is not NULL. Defaults to NA.

ci a logical value indicating whether confidence intervals should be calculated for outcomes. Defaults to FALSE

cilevel Defaults to 0.95

clustvar Variable in the dataset identifying which cluster each respondent belongs to. Used in confidence interval calculations to take survey design into account. Provide if `ci = TRUE`. Defaults to NULL.

stratvar Variable in the dataset identifying which stratum each respondent belongs to. Used in confidence interval calculations to take survey design into account. Provide if `ci = TRUE`. Defaults to NULL.

countryname Character, name of the country (or other top-level grouping) that `data` is from

geolabels Character vector of labels corresponding to the levels of `geovar`. Defaults to NULL

palette Color palette for the levels of `geovar` in output plots. Defaults to NULL, in which case a default ggplot2 palette will be used.

If you don't want to use the default ggplot2 colors in the subnational weight plot, define a color palette with at least as many levels as your geography variable. The Harmonia dataset has ten districts, so the palette we define should have at least ten colors.

```
pal <- c("#f7abc1", "#a2d3fa", "#69c9c8", "#f1cfa1", "#a33c51",
         "#ed8e1c", "#298667", "#0d6eba", "#b8f5f2", "#081116")
```

```
harmonia <- cwt_output(
  data = fauxdata,
  outcomevars = c("has_card", "dtp3_vx", "mcv_vx"),
  outcomenames = c("Vaccination Card", "DTP3", "MCV1"),
  geovar = "geo",
  yearvar = "year",
  weightvar = "psweight",
  years = "all",
```

```
  weightopt = "previous",
  age = 1,
  agevar = "age",
  ci = TRUE,
  cilevel = 0.95,
  clustvar = "clusterid",
  stratvar = "stratumid",
  countryname = "Harmonia",
  geolabels = NULL,
  palette = pal
)
```

The `cwt_output` function returns eight kinds of objects:

```
#> weightdata
#> outcomes
#> subnational_outcomes
#> tables
#> weight_plot
#> subnational_outcome_plots
#> national_outcome_plots
#> combined_plots
```

1. `weightdata` is a data frame showing the relative weight for each level of the `geo` variable for each year in the dataset. This table is the basis for the visualization of weights over time in the `weight_plot` output.

Area	Ratio	Year
District 01	0.171	2005
District 02	0.050	2005
District 03	0.102	2005
District 04	0.085	2005
District 05	0.088	2005
District 06	0.033	2005
District 07	0.147	2005
District 08	0.190	2005
District 09	0.092	2005
District 10	0.041	2005
District 01	0.194	2010
District 02	0.063	2010
District 03	0.138	2010
District 04	0.105	2010
District 05	0.090	2010
District 06	0.026	2010
District 07	0.115	2010
District 08	0.161	2010
District 09	0.076	2010
District 10	0.032	2010
District 01	0.179	2015
District 02	0.066	2015
District 03	0.097	2015
District 04	0.087	2015
District 05	0.102	2015
District 06	0.027	2015
District 07	0.147	2015
District 08	0.174	2015
District 09	0.092	2015
District 10	0.028	2015

2. `outcomes` is a data frame that contains, for each outcome and each year specified in the function call, the point estimate using original weights (`Outcome_Original`), a survey-adjusted Wilson confidence interval for that point estimate (`CI_LB_Original` and `CI_UB_Original`), and the point estimate when weights are post-stratified.

Note that confidence intervals may be too narrow if the weights you are post-stratifying to are *estimated* totals rather than population values. See Jill Dever & Richard Valliant (2010), "A comparison of variance estimators for poststratification to estimated control totals," Survey Methodology, 36(1), 45-56.

Country	OutcomeVar	OutcomeName	Outcome_Original	CI_LB_Original	CI_UB_Original	Outcome_PS	DataYear	PSWeightYear
Harmonia	has_card	Vaccination Card	75.5	74.0	77.0	75.5	2005	2005
Harmonia	dtp3_vx	DTP3	59.7	57.0	62.4	59.7	2005	2005
Harmonia	mcv_vx	MCV1	78.5	76.1	80.7	78.5	2005	2005
Harmonia	has_card	Vaccination Card	76.9	75.3	78.4	76.7	2010	2005
Harmonia	dtp3_vx	DTP3	58.0	56.0	59.9	60.2	2010	2005
Harmonia	mcv_vx	MCV1	76.1	74.3	77.7	78.1	2010	2005
Harmonia	has_card	Vaccination Card	77.1	75.3	78.9	77.4	2015	2010
Harmonia	dtp3_vx	DTP3	59.4	57.4	61.3	58.1	2015	2010
Harmonia	mcv_vx	MCV1	78.1	76.4	79.8	76.7	2015	2010

3. `subnational_outcomes` contains subnational outcome estimates for each level of `geovar`. Recall that post-stratifying does not affect outcome estimates at this lower level – it only impacts how the subnational estimates are aggregated to produce a national estimate. So, there are no "original" and "post-stratified" outcome estimates in this table.

Country	Area	Year	has_card	dtp3_vx	mcv_vx
Harmonia	District 01	2005	0.8	0.4	0.6
Harmonia	District 02	2005	0.8	0.5	0.7
Harmonia	District 03	2005	0.8	0.5	0.7
Harmonia	District 04	2005	0.8	0.4	0.6
Harmonia	District 05	2005	0.8	0.6	0.8
Harmonia	District 06	2005	0.8	0.7	0.9
Harmonia	District 07	2005	0.8	0.7	0.8
Harmonia	District 08	2005	0.7	0.8	0.9
Harmonia	District 09	2005	0.8	0.8	1.0
Harmonia	District 10	2005	0.7	0.7	0.9
Harmonia	District 01	2010	0.8	0.4	0.6
Harmonia	District 02	2010	0.8	0.5	0.8
Harmonia	District 03	2010	0.8	0.5	0.7
Harmonia	District 04	2010	0.8	0.4	0.6
Harmonia	District 05	2010	0.7	0.6	0.8
Harmonia	District 06	2010	0.8	0.7	0.8
Harmonia	District 07	2010	0.8	0.7	0.8
Harmonia	District 08	2010	0.7	0.8	0.9
Harmonia	District 09	2010	0.8	0.8	1.0
Harmonia	District 10	2010	0.7	0.6	0.9
Harmonia	District 01	2015	0.8	0.4	0.6
Harmonia	District 02	2015	0.8	0.5	0.7
Harmonia	District 03	2015	0.8	0.6	0.7
Harmonia	District 04	2015	0.8	0.4	0.6
Harmonia	District 05	2015	0.8	0.6	0.8
Harmonia	District 06	2015	0.7	0.7	0.9
Harmonia	District 07	2015	0.8	0.6	0.8
Harmonia	District 08	2015	0.7	0.8	0.9
Harmonia	District 09	2015	0.8	0.8	1.0
Harmonia	District 10	2015	0.7	0.7	0.9

4. `tables` is a list containing a data frame for each outcome specified. Each row represents a survey pair, and contains original and post-stratified outcome estimates for both years in the survey

pair, as well as calculations of relative difference due to state weights (RDSW) and difference-in-differences. `tables` also contains an openxlsx Worksheet object called `excel`, which is ready to export as a spreadsheet; more on that below.

```
#> Harmonia_has_card_table
#> Harmonia_dtp3_vx_table
#> Harmonia_mcv_vx_table
#> excel
```

You can view the individual data frames with syntax like:

```
harmonia$tables$Harmonia_dtp3_vx_table
```

Country	OutcomeVar	OutcomeName	l_year	k_year	pl	pk
Harmonia	dtp3_vx	DTP3	2010	2005	58.0	59.7
Harmonia	dtp3_vx	DTP3	2015	2010	59.4	58.0

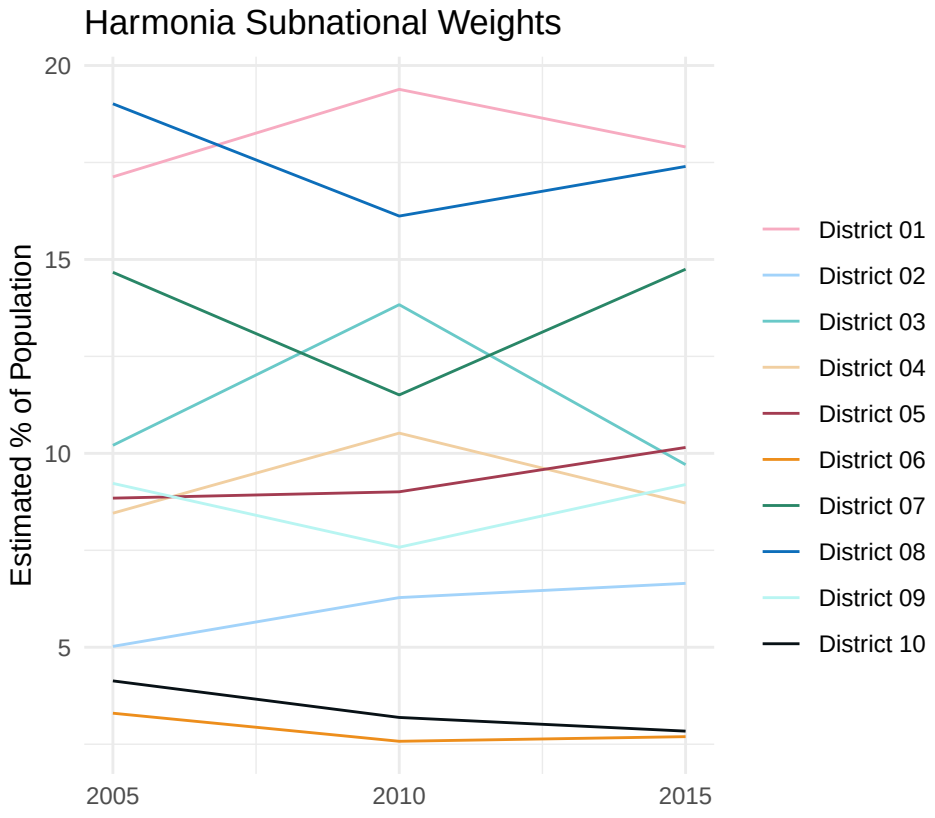
plps	pkps	poststratyear	pl_minus_pk	plps_minus_pkps	rdsw	diff_in_diffs
60.2	59.7	2005	-1.8	0.4	124.8	2.2
58.1	58.0	2010	1.4	0.2	87.5	-1.2

You can save the Excel object - basically, formatted versions of the individual outcome tables - with syntax like the following:

```
openxlsx::saveWorkbook(harmonia$tables$excel, file = "HarmoniaTables.xlsx")
```

The Excel spreadsheet will have a tab for each outcome.

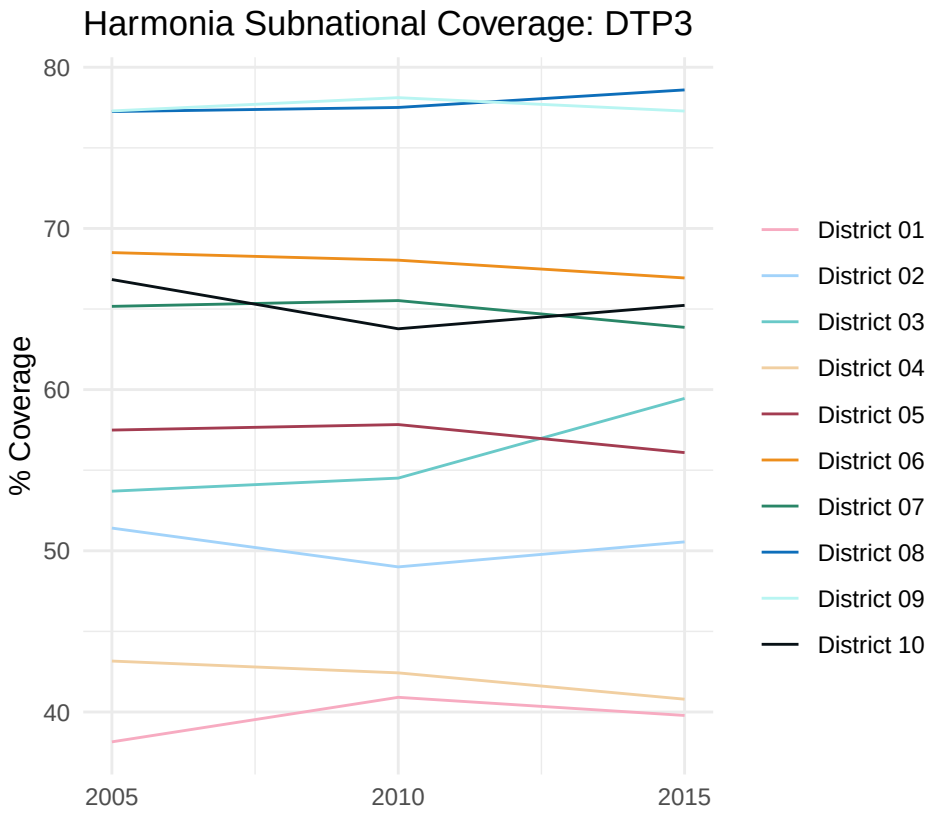
5. `weight_plot` is a plot showing how subnational weights change over time.



6. `subnational_outcome_plots` is a list of plots, each showing outcomes in each geographic area over time.

For example, to look at the plot for DTP3 coverage by district:

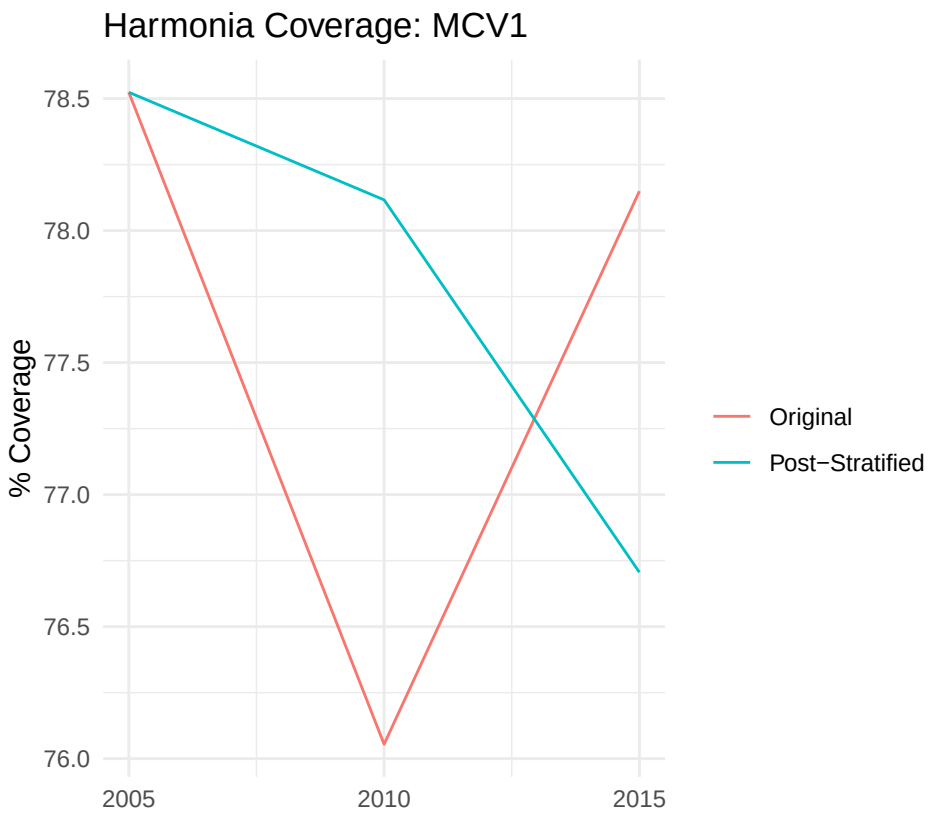
```
harmonia$subnational_outcome_plots$subnational_plot_dtp3_vx
```



7. `national_outcome_plots` is a list of plots, each showing original and post-stratified outcomes over time.

For example, to look at the plot for MCV1 coverage:

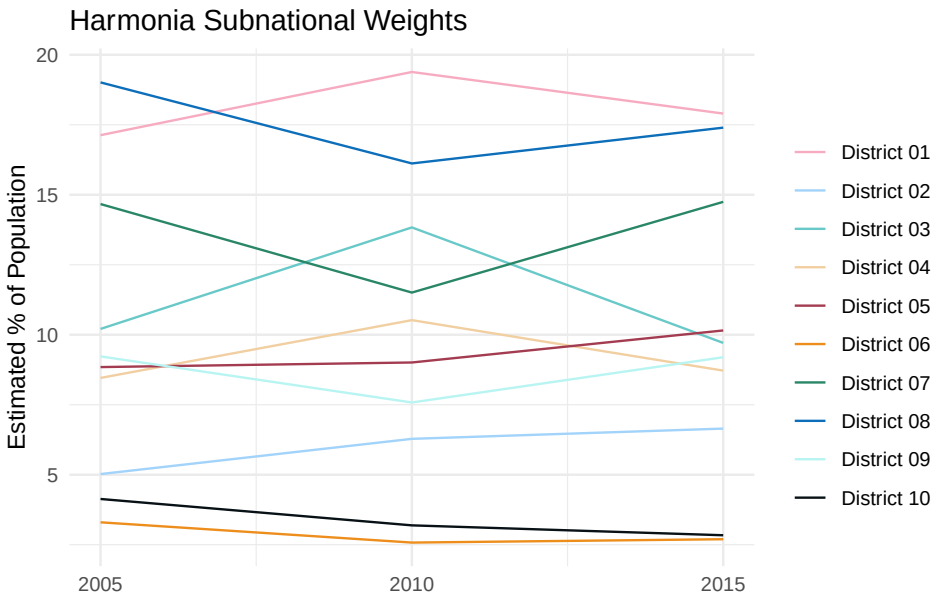
```
harmonia$national_outcome_plots$national_plot_mcv_vx
```

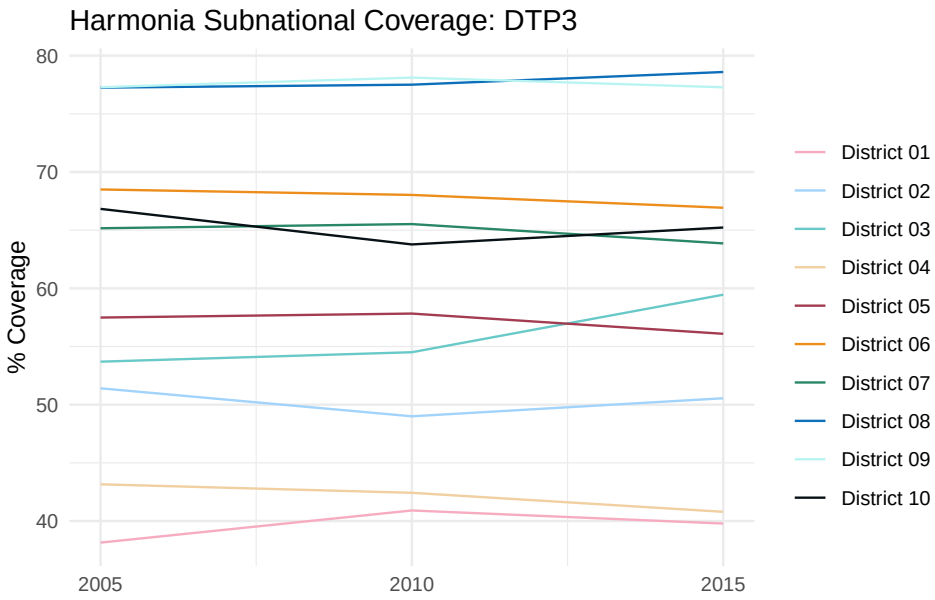


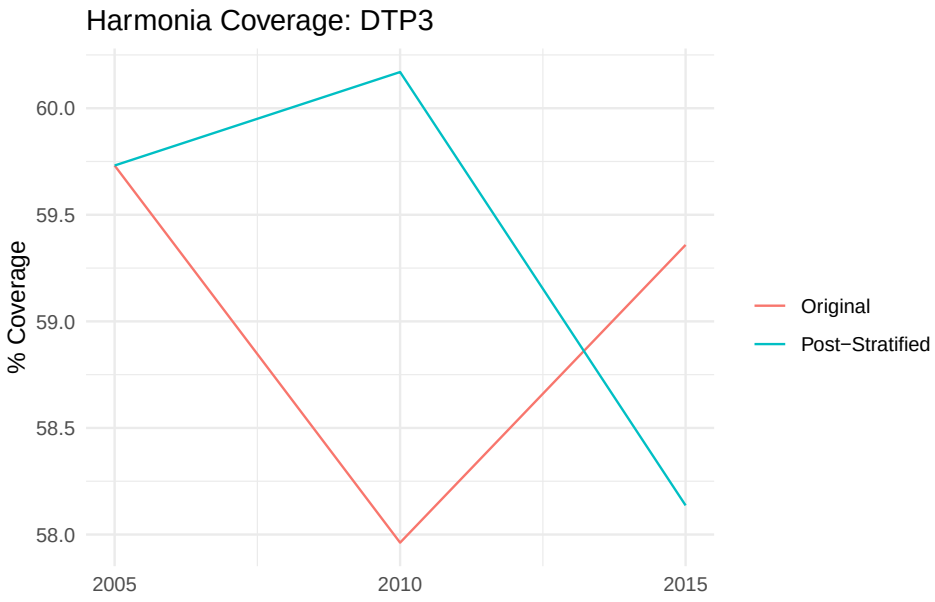
8. `combined_plots` aggregates `weight_plot` and the subnational and national plots for each outcome. The weight and subnational outcome plots help put the plot of original and post-stratified outcomes in context. There is one combined plot for each outcome variable.

You can display these plots with `grid::grid.draw` to view in an R session:

```
grid::grid.draw(harmonia$combined_plots$combined_plot_dtp3_vx)
```


Harmonia Subnational Weights
Estimated % of Population
20
15
10
5
2005
2010
2015
District 01
District 02
District 03
District 04
District 05
District 06
District 07
District 08
District 09
District 10


Harmonia Subnational Coverage: DTP3
% Coverage
80
70
60
50
40
2005
2010
2015
District 01
District 02
District 03
District 04
District 05
District 06
District 07
District 08
District 09
District 10


Harmonia Coverage: DTP3
% Coverage
60.0
59.5
59.0
58.5
58.0
2005
2010
2015
Original
Post−Stratified

Or you can save the combined plots with ggsave:

```
ggplot2::ggsave("CombinedPlotDTP3.png"
       harmonia$combined_plots$combined_plot_dtp3_vx,
       height = 12, width = 6, units = "in")
```