Thesis title

A thesis submitted for the degree of

Master of Business Analytics

by

Dewi Lestari Amaliah

31251587

Department of Econometrics and Business Statistics

Monash University

Australia

October 2021

Contents

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Acknowledgements

I would like to thank my pet goldfish for …

Chapter 1

Introduction

Start with visualisation And then survey And MRP

Chapter 2

Systematic Literature Review

This study is performed using a systematic review method. This method collects empirical evidence explicitly and systematically using pre-specified eligibility criteria to answer a specific research question (Green et al., 2008). Systematic literature reviews also enable the process of finding the gap in a field of science, such as understanding what has been done and what needs to be done (Linnenluecke, Marrone, and Singh, 2020). Hence, in this case, systematic literature review could assist us to understand the common practice in MRP visualisations so that we can explore how to improve.

According to Brown University Library (2021), the key criteria of the systematic literature review are: *"a clearly defined question with inclusion & exclusion criteria; rigorous & systematic search of the literature; critical appraisal of included studies; data extraction and management; analysis & interpretation of results; and report for publication."* Hence, to conform with these criteria, this study incorporates the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)'s checklist and flow diagram. The following subsections discuss the steps conducted following these criteria.

2.1 Literature Identification

MRP is applied in various scientific fields, ranging from social and political science to public health. Therefore, to identify relevant literature, this study refers to research databases instead of field-specific journals. Those databases are JSTOR, EBSCO, and

PubMed. The first two databases are chosen due to their broad range of field coverage, while the latter is chosen since MRP is sometimes also applied in the health and medical fields. These databases were also chosen to represent the heterogeneity of the field, which is one of the important factors in a systematic literature review (Schweizer and Nair, 2017).

From these databases we identify relevant articles using the combination of several search terms. Generally the search terms include the term "multilevel regression", "post-stratification", "poststratification", and "multilevel model". Our target literature is articles that are written in English. We exclude all of the publications before 1997 since this was the first proposal date for MRP. Initially we included only the title/abstract when searching these databases. However, using this method limits the set of potential articles to only include those with the search term in the abstract/title. To rectify this, we also include a search with "all field" in the search criteria. Note that for EBSCO, we directly apply the search for all fields. The detailed literature identification is shown in Table 2.1.

The total number of articles from this search criteria are 327. Next, we utilize the literature manager, EndNote X9, to manage these articles and to find duplicate articles. After removing those duplicate articles, we have 212 articles to be screened in the next stage.

Table 2.1: *Detail of literature identification*

Database	Search Terms	Search Field	Inclusion	Exclusion	Number Returned
JSTOR	(multilevel regression and poststratification) OR (“post-stratification”)	Abstract	Article, content I can access, English	anything before 1997	44
JSTOR	(("multilevel regression" AND ("post-stratification" OR Poststratification)) OR ("multilevel model" AND ("post-stratification" OR Poststratification)))	All field	Article, English	anything before 1997	142
EBSCO	"multilevel regression with post-stratification" OR "multilevel regression with poststratification" OR "multilevel regression and Poststratification" OR "multilevel regression and Post-stratification"	All field	Academic (Peer-Reviewed) Journals, English	anything before 1997	42
EBSCO	(multilevel regression AND post-stratification) OR (multilevel model AND post-stratification) OR (multilevel regression AND poststratification) OR (multilevel model AND poststratification)	All field	Academic (Peer-Reviewed) Journals, English	anything before 1997	45
PubMed	"multilevel regression with post-stratification" OR "multilevel regression with poststratification" OR "multilevel regression and Poststratification" OR "multilevel regression and Post-stratification"	Title/Abstract	Article, English	anything before 1997	26
PubMed	(multilevel regression AND post-stratification) OR (multilevel model AND post-stratification) OR (multilevel regression AND poststratification) OR (multilevel model AND poststratification)	All field	Article, English	anything before 1997	28

2.2 Screening and Eligibility Criteria

We screen all of the articles based on predetermined criteria. We find that 3 articles are apparently not research papers. This results in 209 abstracts to be screened. To screen efficiently, we use two stages. The first stage is a review of abstracts, the second a full manuscript review.

2.2.1 Stage 1: Review of abstracts

In the first stage DA and LK independently review all article abstracts with the following eligibility criteria:

1. The abstract should mention analysis of data or creation of simulation data.
2. The abstract should mention the use of MRP or multilevel models to make population estimates or the use of other regression models (BART, spatial, stacking, trees) to make population estimates.

During the screening, DA and LK agreed that 61 articles meet the eligibility criteria listed above, while 104 articles do not meet the criteria. The two reviewers disagreed on 44 articles. Accordingly, DA and LK skim the full manuscript to decide whether the paper could be included in the next stage or not. As the result, an additional 22 more articles are moved to stage 2, making a total of 83.

2.2.2 Stage 2: Full manuscript review

DA reviews the full manuscript on 83 articles based on a second set of criteria. The aim of this stage is to get the list of the final articles that would be included in the study. We set the criteria of inclusion as follow:

1. It should apply MRP as its method.
2. It should contain at least one plot relate to MRP findings.

During this stage, we exclude 4 articles as they do not meet the first criteria. Further, 7 articles are excluded as they do not meet the second criteria. Also, an article is not included

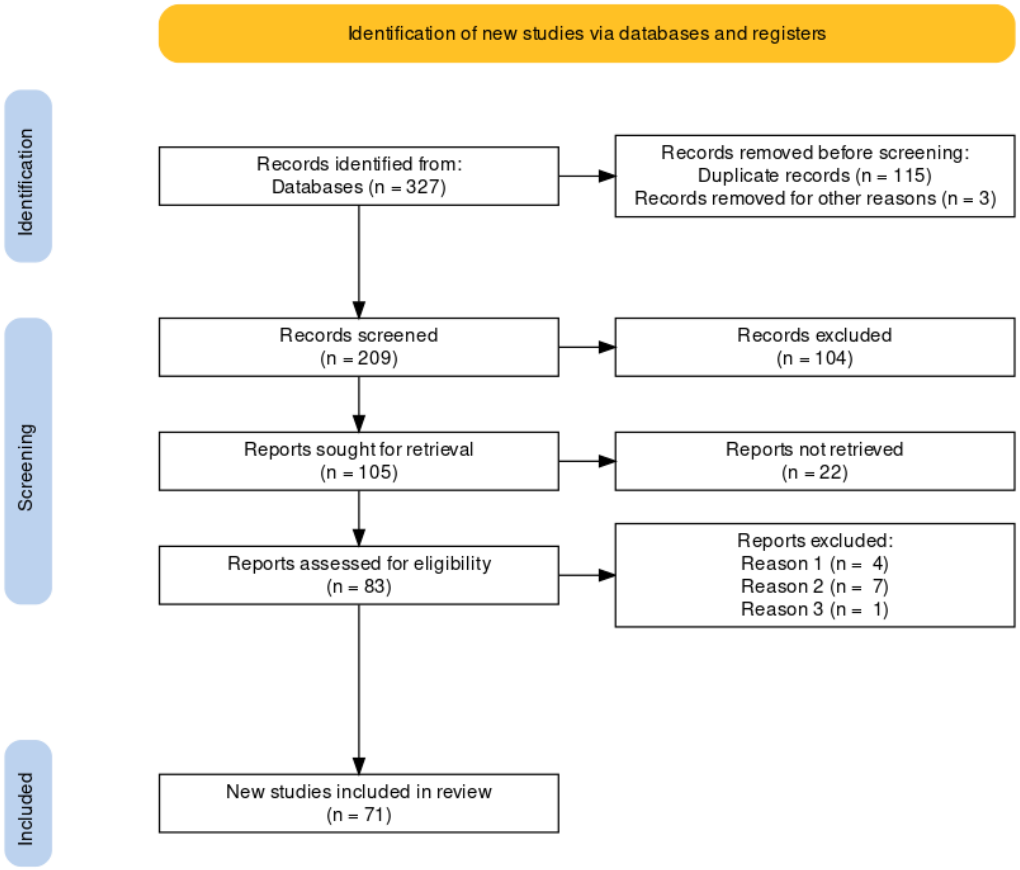

Figure 2.1: *PRISMA flow chart of this systematic literature review.*

because it is a duplicate that was not detected automatically by Endnote X9. Finally, we have 71 articles to be reviewed in the next stage. Figure 2.1 displays the PRISMA flow chart of this study. This figure is generated using `PRISMA2020` (Haddaway, Pritchard, and McGuinness, 2021).

2.3 Data Extraction and Analysis

We focus the data extraction on the MRP-related plot. We manually create a metadata for each plot (included in the supplementary material). We will use this metadata to analyse the current reporting practices with MRP. This metadata will also ensure the reproducibility of the analysis and to maintain the transparency of the systematic literature review process.

We code the plots according to their type, i.e., communication (coded to 0) and diagnostic plot (coded to 1). For diagnostic plots, we examine whether the plots compare MRP with other estimates, which are:

1. Raw (direct estimates or direct disaggregation);
2. Ground truth;
3. Weighted estimates;
4. Estimates from other MRP models, for example, a paper build several MRP models from various simulation scenarios or using different covariates;
5. Estimates from another study/survey;
6. Estimates from another method, for example comparing MRP with Bayesian Additive Tress with Post-Stratification(BARP).

Plots that show a comparison of MRP with the above list would be coded to 1, otherwise coded to 0. Diagnostic plots also categorised based on how they compare the performance of MRP. The five observed criteria are:

1. Bias;
2. Mean Absolute Error (MAE);
3. Mean Square Error (MSE)/ Relative Mean Square Error (RMSE);
4. Standard Error (SE);
5. Correlation.

Each plot is assessed based on the use of the performance metric. For each metric is scored based on whether it is used (coded 1) or not (coded 0).

We also review other features of the plot using the grammar in `ggplot2` (Wickham, 2016) as a framework. The common grammar used in practice allows us to understand to what extend MRP models are effectively visualised. It is worth noting that there is no specific convention or well-documented recommendation on how data should be visualised as building a graph more often involves choice or preference (Midway, 2020). For example, there is no specific convention on which variable should be put on the x and y-axis in a

scatter plot, even though it has been common knowledge to put the response variable on the y-axis and the explanatory variable on the x-axis. Hence, grammar assists us in evaluating well-formed graphics (Wickham, 2010). In addition, Vanderplas, Cook, and Hofmann (2020) mention that classifying and comparing graphs according to their grammar is more robust and more elegant.

Accordingly, we examine the facet, geom, axis, color, and shape. For reproducibility, the metadata also contains the article's author/s, publication year, title, and corresponding figure number as it appeared in the article. After the extraction, we analyze the data using graphical visualization with `ggplot2` (Wickham, 2016). The result will be discussed in the following subsection.

2.4 Common practices in MRP visualisations

In this study, graphics are classified into two types, i.e., communication and diagnostic plots. A plot is classified as a communication plot if the plot's goal is solely to convey the MRP result. At the same time, a diagnostic plot displays the MRP estimation by showing the performance metrics or compares it with other estimation methods. From 71 articles, we extract the data of 243 plots. 47.33 % of these plots are diagnostics plots, while the remaining are communication plots.

2.4.1 Performance metrics used in MRP

According to Botchkarev (2019), performance metrics is *"a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted"* RMSE and MAE are among the most common methods used in many studies (Botchkarev, 2019). However, Willmott and Matsuura (2005) states that RMSE should not be reported in any studies since it could be multiinterpreted because it does not describe average error alone and MAE is more appropriate metrics. This argument is denied by Chai and Draxler (2014) who argue that RMSE is not ambiguous and better than MAE if the distribution of model's error is normal. Accordingly, there is no single metrics that fits for all (Chai and Draxler, 2014).

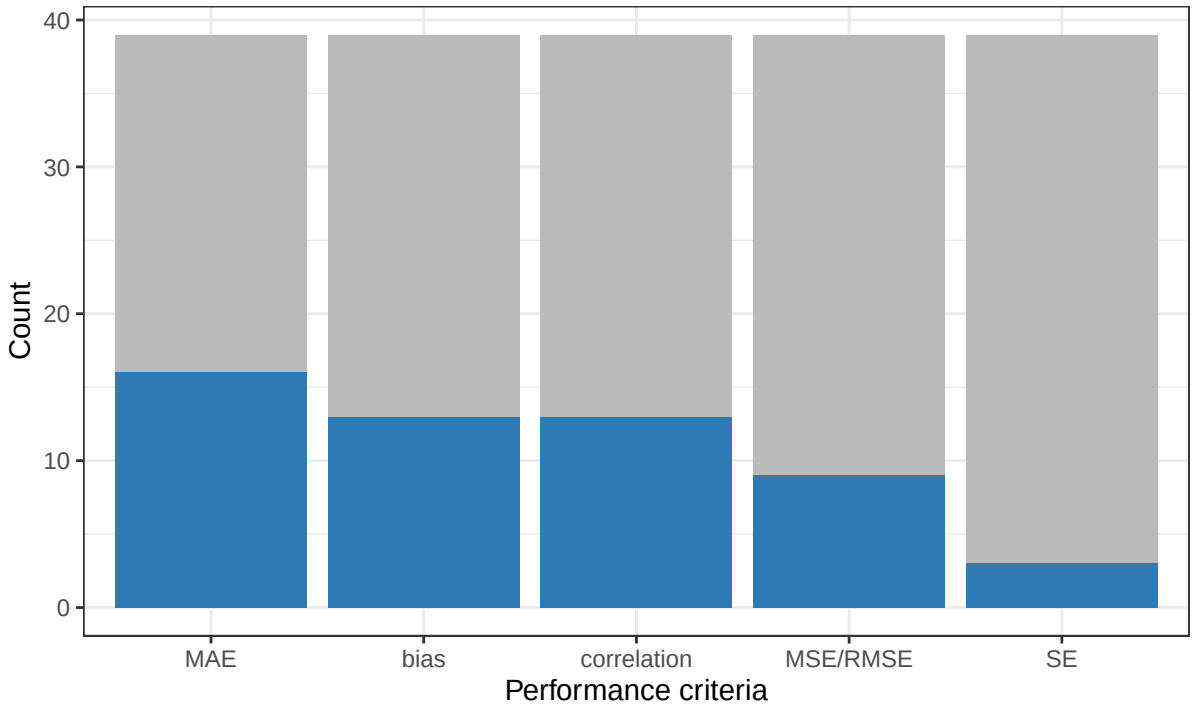

Figure 2.2: *There are five performance metrics used in study: Mean Absolute Error (MAE), bias, correlation, Mean Square Error/Root Mean Square Error (MSE/RMSE), and Standard Error (SE). The blue shade represents the number of articles that show performace metrics in plot, while the grey shade represents the number of articles that show performance of MRP but did not use the corresponding metrics.*

In this study, we find that there are 39 plots out of 115 diagnostic plots (about 34%) display performance measures. As seen in Figure 2.2, we find that MAE is the most widely demonstrated performance metric by MRP visualisations. Bias, which is interpreted similarly to MAE, is also widely used. Meanwhile, the square error measures, which are MSE/RMSE and standard error, are only shown by a few plots. It is interesting that correlation, which is not a common metric use as a performance indicator, is more widely used than square error metrics.

It is worth noting that most of these metrics only refer to point estimates, i.e., the distance between the predicted value and the actual values. Also, these metrics mainly measure bias. However, MRP is a model in which bias-variance is applied. Therefore, other measures are also needed that reflect the degree of uncertainty and variations in the predicted value. Measures such as length of confidence or credibility interval can be used, in which the narrower the value, the more precise the estimates.

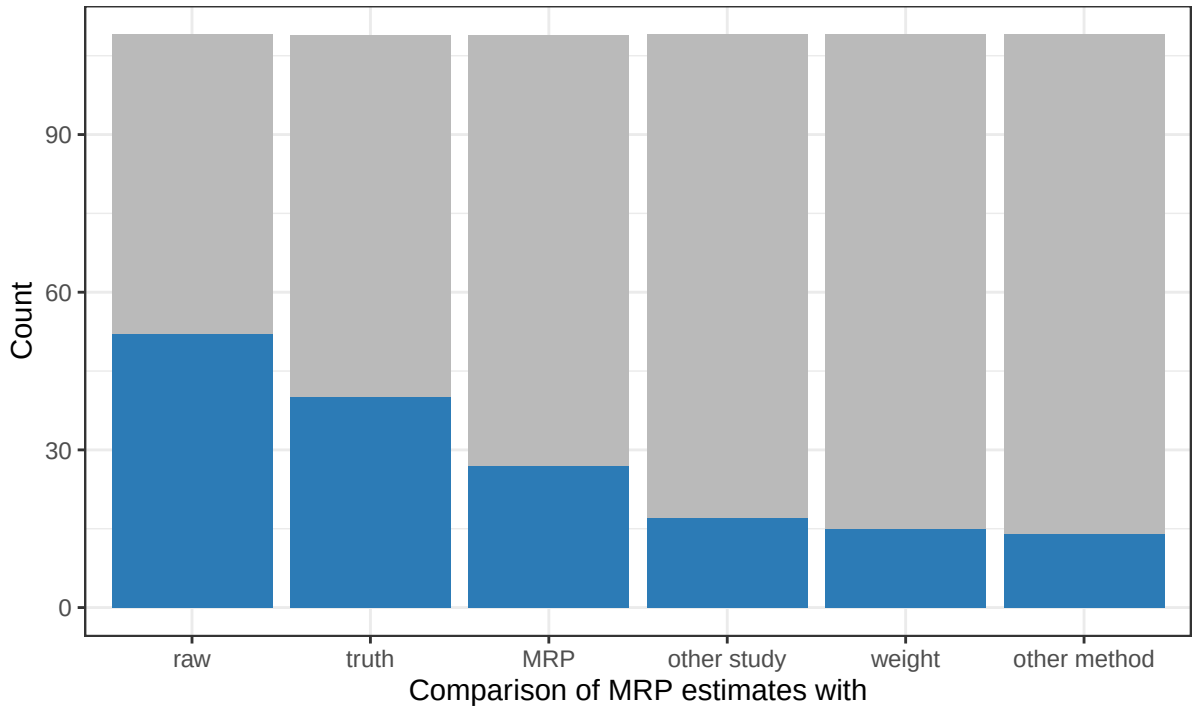

Figure 2.3: *Estimates to compare with MRP. The blue shade represents the number of articles compare MRP estimates with the result of other estimation methods, while the grey shade represents the number of articles that also showed comparison of MRP but did not use the corresponding estimates.*

2.4.2 Common estimates to compare with MRP estimates

The goals of MRP are to estimate the population and usually to adjust the estimation from an unrepresentative survey. Thus, in practice, MRP is usually compared with the actual value, which, in this case, is called truth. Regarding the objective of improving estimates from unrepresentative surveys, MRP estimates are usually compared with direct estimates (raw). In addition, MRP is often compared with weighted estimates.

This study finds that from 115 diagnostic plots, 109 (about 95%) of which compares MRP estimates with estimates from other methods. Figure 2.3 shows the distribution of the estimation methods to compare with MRP estimates. It can be seen that MRP estimates are mostly compared to direct estimates and ground truth. Some studies also compare estimates from several MRP models (usually with different model specifications). We also find that there are not many plots showing the comparison between MRP estimates and weighted estimates.

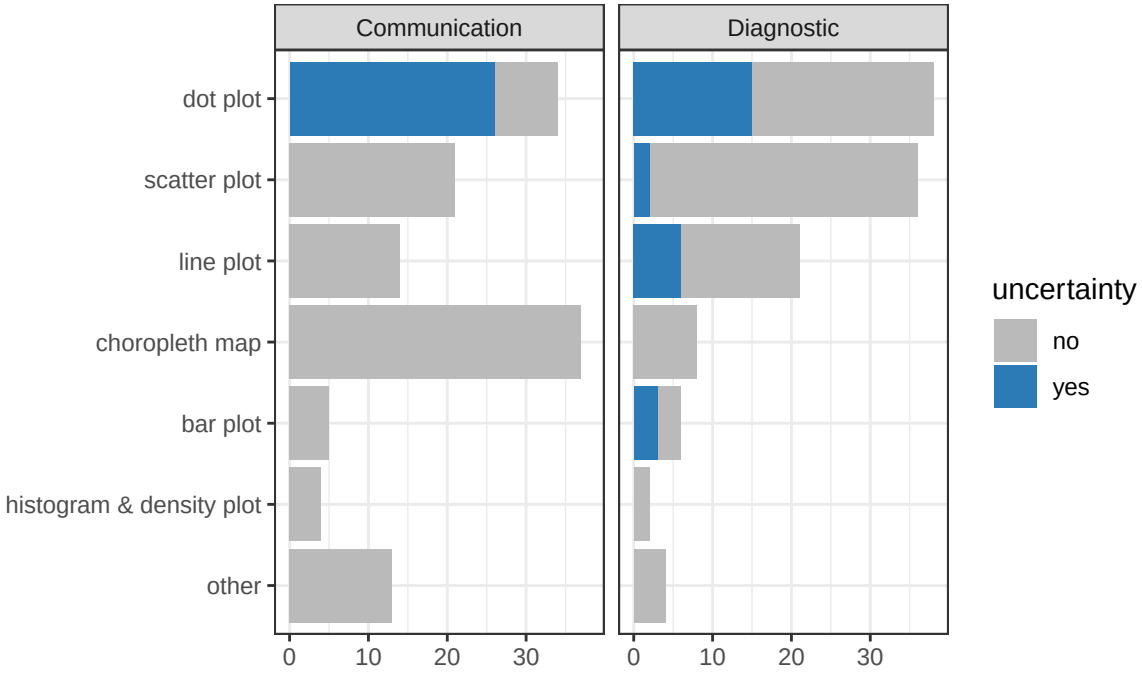

Figure 2.4: *Common plot types used in MRP visualisations. The blue shade display the number of plots that showed uncertainty, while the grey shade display the number of plots that did not show uncertainty. Both communication and diagnostics plots rarely displayed uncertainty.*

2.4.3 Common grammar in MRP visualisations

Plot type

Plot type, referred to as `geom` in the grammar of graphics, represents the shape and features displayed in the graph. Figure 2.4 suggests that communication and diagnostic plots have a different pattern in the plot type's distribution. Communicating MRP estimates are mostly done using a choropleth map as MRP is often deployed for small area estimation. For diagnostics purposes, dot plots are mostly used to compare more than two estimation methods or to show some performance metrics.

Notice that Figure 2.4 also displays the usage of uncertainty in MRP model visualisations. According to Midway (2020), displaying uncertainty in the statistical graphs is essential as the absence of this measure would produce a misleading interpretation and hinder parts of statistical messages. However, he further states that uncertainty is often neglected in

data visualisation. This is what we find in this study in which the uncertainty is not often seen in the plots.

Values put in x and y-axis

Main things to explain: - There are no strict rules on what to put in a and y-axis. However there are some conventions, for example fixed value is put in the x-axis, while random variable is put in the y-axis. - Time variable is always put in x axis.

Facet

Other features used

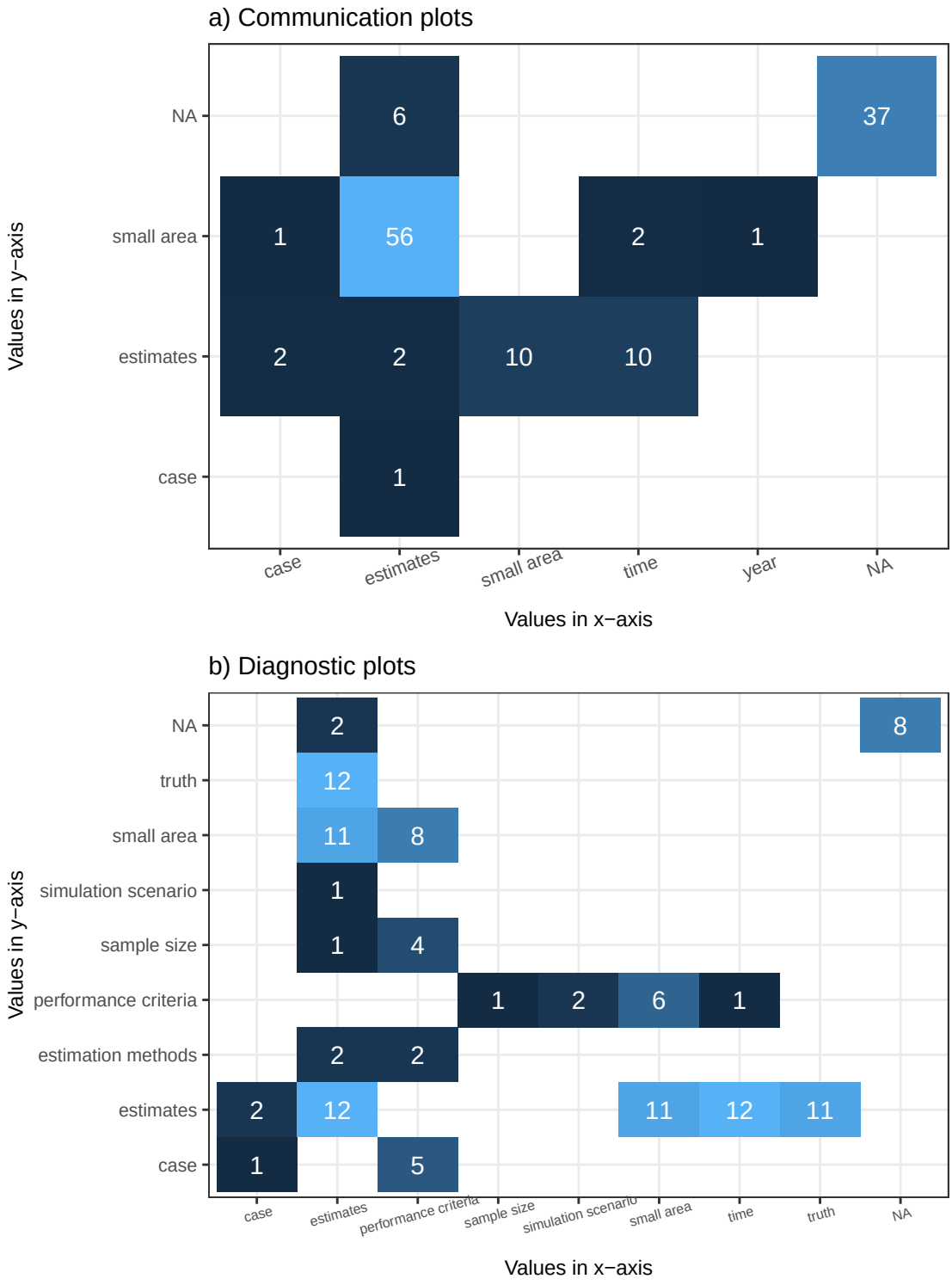

Figure 2.5: *Common values put in plots' axis. Axis in diagnostic plots more varied compared to communication plot.*

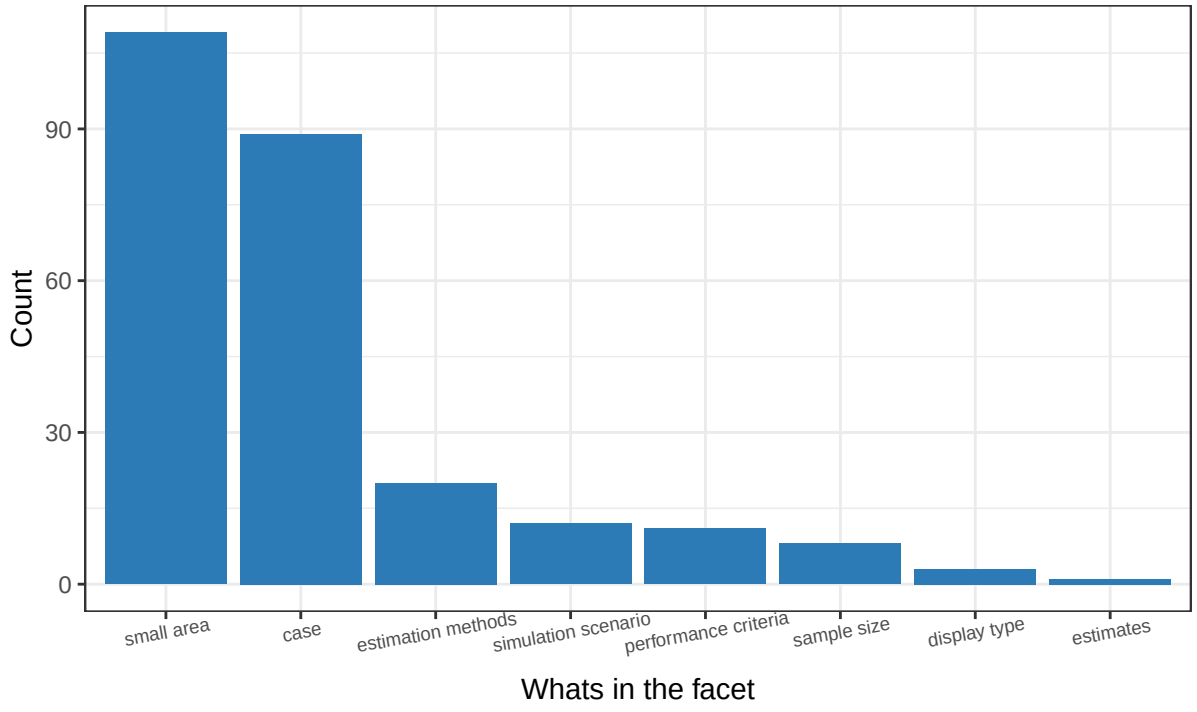

Figure 2.6: *Measures faceted in MRP visualisations*

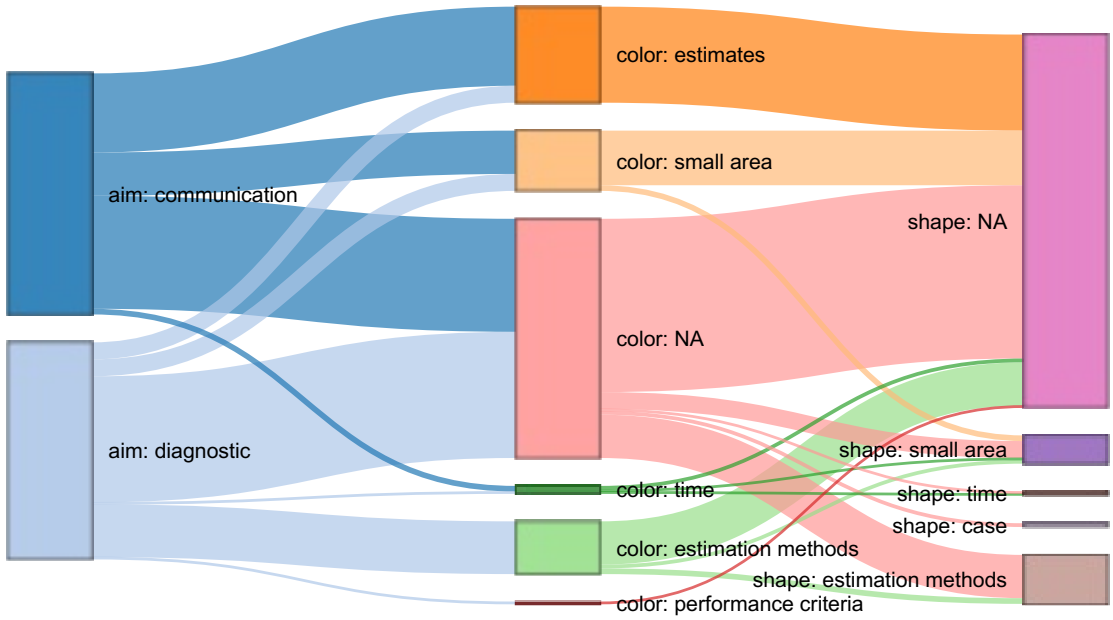

Figure 2.7: *Color and shape commonly used in MRP visualisations. Both communication and diagnostic plots rarely use color and shape features.*

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Botchkarev, A (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. eng. *Interdisciplinary journal of information, knowledge, and management* **14**, 45–76.

Brown University Library (2021). *Scientific Literature Review Resources and Services*. https://libguides.brown.edu/Reviews/types.

Chai, T and RR Draxler (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. eng. *Geoscientific model development* **7**(3), 1247–1250.

Green, S, JP Higgins, P Alderson, M Clarke, CD Mulrow, and AD Oxman (2008). "Introduction". In: *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd. Chap. 1, pp. 1–9. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470712184.ch1. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470712184.ch1.

Haddaway, NR, CC Pritchard, and LA McGuinness (2021). *PRISMA2020: R package and ShinyApp for producing PRISMA 2020 compliant flow diagrams (Version 0.0.2)*.

Linnenluecke, MK, M Marrone, and AK Singh (2020). Conducting systematic literature reviews and bibliometric analyses. eng. *Australian journal of management* **45**(2), 175–194.

Midway, SR (2020). Principles of Effective Data Visualization. *Patterns* **1**(9), 100141.

Schweizer, ML and R Nair (2017). A practical guide to systematic literature reviews and meta-analyses in infection prevention: Planning, challenges, and execution. eng. *American journal of infection control* **45**(11), 1292–1294.

Vanderplas, S, D Cook, and H Hofmann (2020). Testing Statistical Charts: What Makes a Good Graph? *Annual Review of Statistics and Its Application* **7**(1), 61–88.

Wickham, H (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics* **19**(1), 3–28.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Willmott, C and K Matsuura (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. eng. *Climate research* **30**(1), 79–82.