MONASH University

Visualising Uncertainty: Conceptual Framework, Software, and Perception

Harriet Mason

B.Comm. (Hons), Monash University

A thesis submitted for the degree of
Doctor of Philosophy
at Monash University in 2026
Department of Econometrics & Business Statistics

Table of contents

Copyright notice

Produced on 28 April 2026.

Abstract

Visualisation is a powerful tool in data analysis, as it allows us to learn from our data, identifying new insights or hypotheses that we would otherwise have missed. These insights are only possible thanks to the tools that allow us to render accurate and reliable representations of our data. Unfortunately, despite the prevalence of random variables in statistical analysis, these powerful tools do not extend to the visualisation of estimates, as we are unable to incorporate uncertainty into our plots. Limiting our visualisations to point estimates restricts our ability to gain insights into uncertain data, which is especially problematic given their prevalence. Working past this limitation has remained difficult, as data visualisation is a broad area, simultaneously touching several fields at once, including philosophy, mathematical statistics, and studies of human perception.

Properly addressing these issues requires a critical assessment of existing uncertainty visualisation approaches under each lens. This is the work done by this thesis, which presents three original contributions. The first contribution is a philosophical argument that untangles the current literature and provides guidance on what it means to make a "good" uncertainty visualisation. The second contribution is both conceptual and practical, as it provides a mathematical formalisation of uncertainty visualisation, which is then used to create the R package `ggdibbler`, which is a flexible uncertainty visualisation software that facilitates exploratory data analysis of random matrices. The final contribution is a human study on the perception of uncertain plots, which evaluates uncertainty visualisations on their ability to align with the conclusions of classical hypothesis tests, and provides some guiding principles on the perception of uncertain plots. These contributions allow us to visualise random variables with as much ease as we do normal data, assured by the knowledge that what we see is actually there.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three research articles, which are at different stages of publication. Chapter 2 has been submitted to Annual Reviews of Statistics and Its Applications as an invited contribution. Chapter 3 is being prepared for submission to the Journal of Computational and Graphical Statistics. Chapter 4 has been submitted to IEEE VIS 2026. The core theme of this thesis is "visual communication of uncertain data". The ideas, development, and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Department of Econometrics & Business Statistics under the supervision of Professor Dianne Cook, Dr Sarah Goodwin, and Dr Susan VanderPlas (University of Nebraska-Lincoln). Chapter 2 was also a collaboration with Dr Emi Tanaka, and Chapter 4 was a collaboration with Dr Rachel Rogers (University of Technology Sydney) and Dr Alison Kleffner (Creighton University). In the case of Chapter 2, Chapter 3, and Chapter 4, my contributions are detailed in the table below.

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

Thesis chapter	Publication title	Status	Nature and % of student contribution	Nature and % of coauthors' contribution	Coauthors are Monash students
2	The Noisy Work of Uncertainty Visualisation	Submitted to Annual Reviews of Statistics and Its Applications	Concept, writing, editing: 80%	Dianne Cook: 10%; Sarah Goodwin: 5%; Emi Tanaka: 2.5%; Susan Vanderplas: 2.5%	No
3	A Mathematical Framework and Software Implementation for Uncertainty Visualisation	Prepared for submission to Journal of Computational and Graphical Statistics	Concept, software design and implementation, writing, editing: 80%	Dianne Cook: 10%; Sarah Goodwin: 5%; Susan Vanderplas: 5%	No
4	Colour Blinded by the Noise	Submitted to IEEE VIS 2026	Concept, experiment design and implementation, analysis, writing, editing: 75%	Rachel Rogers: 10%; Alison Kleffner: 10%; Dianne Cook: 5%	No

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Harriet Mason

Student signature:

Date:

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author, I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Dianne Cook

Main Supervisor signature:

Date:

Reproducibility statement

All materials associated with this thesis are openly available for transparency and following reproducible practice. The thesis is written using Quarto (Allaire & Dervieux 2024) and is available online at harriet-mason.github.io/phd_thesis/. All materials (including the data sets and source files) required to reproduce this document can be found at the Github repository github.com/harriet-mason/phd_thesis.

Acknowledgements

First, I want to thank my supervision team, who put up with my complete inability to stick to deadlines for almost 4 years. You have all forever shaped the way I think about statistics, software, research, and visualisation. To Emi and Ursula, even you were not supervisors for the entire duration of the project, your excellent feedback in the early days of my PhD had a significant influence on my approach to research, for which I am very grateful. To Susan, having someone on the supervision team who could understand my half-formed ideas well enough to help me translate them into something other people could understand was genuinely a lifesaver, and I don't think I would have finished the first paper (let alone the entire PhD) without your help. To Sarah, your exuberant energy and openness to odd ideas are things that I sincerely hope has rubbed off on me as I try to embody that attitude for the rest of my career. To Di, I don't think I have words to express the level of gratitude I have towards you. Thank you for taking me on as a research assistant, then as an honours student, and finally as a PhD student. It is not an overstatement to say I wouldn't even be in research were it not for you. These last 6 years have inspired a level of joy and fulfilment in my work that I didn't think I would ever see in this lifetime, and for that you have my eternal gratitude.

Thank you to Rachel and Alison for your help on the third paper. I had given up on the idea of doing an experiment, but your enthusiasm about the project, delightful company, and, most importantly, ability to do the parts of the research that I couldn't, allowed me to include it in my thesis. The experiment paper was easily the most fun of all the chapters.

Thank you to my fellow PhD students for always being good company in the office, particularly Sherry, Patrick, Heshani, Ze-yu, Floyd, Fillip, Shelly, Kris, Minh, Cash, Tina, Maliny and Vis. An extra special thank you to Cynthia, Mitch, Jayani, Bets, and Janith, who I occasionally sucked into conversations that went on for so long that my housemates would call to check if I had died. Thank you to the NUMBATs research group for being my home while at Monash. In particular, thank you to Michael, Hannah, and Kate for being as easy to talk to as the other PhD students. Thank you to R-ladies Melbourne for being so friendly and inviting, particularly Dionne and Daidai.

To Gael, thank you for letting me into the program (despite my almost failing your class in my

Bachelors), and Catherine, David, and Xibin for taking over the program after Gael retired. To my panel: Catherine, Jess, Michael, and David, thank you for the helpful feedback at my milestones that greatly improved the work. Thank you to everyone at AEMO and the Zema scholarship fund for supporting my research.

To my friends, Monique, Chloe, Sophie, Kat, and Kris, thank you for letting me complain so much in our various group chats and hang outs. I hope this acknowledgement made all my complaining worthwhile. To Eliot, thank you for the eight years we have spent getting coffee and walking laps around Monash University. Conversations with you always brighten my day, and my supervisors can thank you for approximately 50% of my visits to campus.

To my family, Mum, Dad, Grandma, Eloise, Ben, Alastair, Lauren, and Prudence, thank you for offering your homes when I travelled for conferences, participating in my pilot study despite having no idea what it was for, and remaining supportive for this very long degree. To my dog, Bosco, thank you for being a good boy and giving me your company while I worked. Finally, to my partner Tom, thank you for answering my questions on maths definitions (even at three in the morning), helping me bake for the annual NUMBATs group bake off (even at four in the morning), and continuing to give your love and support for the past few months as my entire life was sucked into this thesis.

To all my friends, family, and loved ones, I know these past four years have been almost as hard on you as they have been on me (*almost*), but I wouldn't have gone through them with anyone else.

Chapter 1

Introduction

Uncertainty visualisation is a relatively new field, but it has been poorly defined. Uncertainty itself is a vague term; each paper seems to take the meaning of it as a given, so the visualisation of it is fraught with disagreement and conflict over definitions, applications, and best practices. Incorporating uncertainty into our visualisations is important for transparent graphics and improved decision making, with some authors considering ignoring uncertainty to be tantamount to fraud (Hullman 2020). Despite the importance of uncertainty visualisation, the field has struggled to create an overarching cohesive theory, with current best practices in plot design remaining ad hoc and context specific (MacEachren et al. 2005). These issues would be resolved if uncertainty were formalised using existing structures for statistical graphics, such as the *grammar of graphics* (Wilkinson 2005) and its implementation in `ggplot2` (Wickham 2010). The formalisation of uncertainty in the grammar of graphics has remained difficult due to the confusing nature of uncertainty, a term which has as many definitions as there are discussions on the topic (Spiegelhalter 2017).

This thesis attempts to close these gaps by providing a comprehensive analysis of uncertainty visualisation, including its definition, practical application, and evaluation. First, we deal with the vague definition of uncertainty and interrogate the current meaning of the term "uncertainty visualisation". Building on foundations in statistical inference and graphics, we redefine what it means to visualise uncertainty and establish a statistical foundation for the goals of the field. Next, we leverage this foundation to formalise uncertainty visualisation in the grammar of graphics, implemented in the `ggplot2` extension `ggdibbler`. We show that the formalisation goes beyond mathematical pedantry as it facilitates exploratory data analysis of uncertain data, an application that has remained just out of reach of the field for decades. Finally, we tie these ideas together in an evaluation study on the perception of uncertainty visualisations, and show that the formalisation we designed is not only flexible, but provides a framework for testing and validating these graphics. This validation is

done using a novel experiment design, by drawing parallels in the goals of uncertainty visualisation and standard colour blind tests, repurposing the standard Ishihara colour blind test for uncertainty visualisation. Statistical graphics exist at the intersection of philosophy, statistics, computer science, and psychometrics, and a thorough dive into the problems facing uncertainty visualisation will require us to touch on all four fields.

1.1 Thesis Outline

The thesis is structured as follows.

Chapter 2 provides a comprehensive review of the uncertainty visualisation literature. Existing reviews in the field take the purpose of an uncertainty visualisation to be self-evident, which results in a large amount of conflicting information. The most common approach communicates uncertainty as a probability or a distribution, and focuses on showing the uncertainty as an isolated variable disconnected from its context. We contrast these visualisations with approaches that view uncertainty as noise that should be incorporated to give a holistic view of our data. We coin the term "signal suppression" to describe a visualisation that is designed for preventing false conclusions, as the approach demands that the signal (i.e., the conclusions drawn from the estimates) is suppressed by the noise (i.e., the variance on those estimates). We provide motivation for viewing signal suppression as the most worthwhile goal in uncertainty visualisation, and argue that visualisations that display uncertainty as an isolated variable should not be considered uncertainty visualisations at all. We further discuss difficulties in creating and evaluating the effectiveness of these plots, which further motivates the following chapters of this thesis.

Chapter 3 introduces a mathematical framework for uncertainty visualisation, as well as a new R package, `ggdibbler`. In Chapter 2 we discussed the run-on effects of the ambiguous definition of uncertainty. These problems included difficulty incorporating uncertainty into the grammar of graphics framework, which leads to software that does not always behave in ways that users would expect. Chapter 3 closes this gap by discussing the theoretical framework required to integrate uncertainty into the grammar of graphics, along with the desirable statistical properties that go along with this framework. This framework is demonstrated in `ggdibbler`, a `ggplot2` extension that allows for flexible uncertainty visualisation for exploratory data analysis. The software allows users to replace any vector of variables in a `ggplot2` plot with a vector of random variables, and get an uncertain version of a plot. These plots include uncertainty in such a way that they align with the goals suggested by Chapter 2, and reduce the visibility of statistically invalid signals.

Chapter 4 presents a user study that evaluates human perception of uncertainty visualisations. Chapter 2 highlighted the difficulty in evaluating uncertainty visualisations, as the uncertainty must be evaluated as noise rather than signal. Correctly evaluating uncertainty visualisations, therefore, requires us to design an experiment that measures uncertainty as a latent variable. This section applies the suggested experimental methods and identifies which graphics introduced in Chapter 3 are effective tools for signal suppression. This experiment confirms many of the hypothesis proposed by the literature review in Chapter 2, and suggests that the visualisations made by Chapter 3 are statistically valid.

The material in Chapter 2 was submitted to *Annual Reviews of Statistics and Its Applications* as an invited contribution to Volume 14 of the Journal. The `ggdibbler` software outlined in Chapter 3 is published on *CRAN*. The contribution in Chapter 3 of this thesis was presented at *useR! 2025* in Durham, North Carolina, in August 2025 and *ASC 2026* in Perth, Western Australia, in December 2025. The material in Chapter 3 is prepared for submission to the *Journal of Computational Graphics and Statistics*. The material in Chapter 4 has been submitted to *IEEE VIS 2026*.

Chapter 2

The Noisy Work of Uncertainty Visualisation

2.1 Introduction

What do we mean when we talk about "uncertainty visualisation"? The phrase can feel contradictory to anyone familiar with the term. Among statisticians, "uncertainty" is often discussed as an omnipresent spectre touching every stage of our analysis without ever being fully seen. Authors will often mention that the phrase is vague (Spiegelhalter 2017; Griethe & Schumann 2006), or avoid defining it by describing a list of things uncertainty *could* be (Kinkeldey, MacEachren & Schiewe 2014; Hullman 2016), but rarely do authors attempt to discuss what uncertainty actually *is*. By contrast, visual statistics (information visualisations, data plots) are one of the most powerful tools in the statistician's toolbox, allowing for quick and memorable communication that identifies quirks in our data that we didn't even know to look for. We see this in datasets such as Anscombe's quartet (Anscombe 1973) or the Datasaurus Dozen (Matejka & Fitzmaurice 2017; Locke & D'Agostino McGowan 2018), where visual statistics are able to highlight elements of the data that are invisible to the typical summary statistics. We also see this in recall experiments, where simply sketching a distribution before recalling statistics or making predictions can greatly increase the accuracy of those measures (Hullman et al. 2018; Goldstein & Rothschild 2014). Taken together, uncertainty visualisation implies a need to pull back the curtain and explore the unknowns of our analysis.

As nice as this sentiment is, it turns out to be easier said than done. Reviews on uncertainty visualisation rarely offer tried and tested rules for effective uncertainty visualisation, instead commenting on the difficulties faced when trying to summarise the field. Kinkeldey, MacEachren & Schiewe (2014) found most experimental methods to be ad hoc, with no commonly agreed upon methodology,

formalisations, or a greater goal of describing general principles. Hullman (2016) noticed there is a serious noise issue in the field, with errors from participants misunderstanding visualisations, misinterpreting questions, and incorrectly applying heuristics, overwhelming any information we can glean from studies. MacEachren et al. (2005) identified so many contradictions that they spent an entire page discussing the conflicting evidence for the question "Should I map uncertainty to colour hue?". Spiegelhalter (2017) concluded that different plots are good for different things, arguing against a universal best plot for all people and circumstances. Padilla, Kay & Hullman (2022) summarised several cognitive effects that repeatedly arise in uncertainty visualisation experiments, but these effects were each discussed in isolation as a list of considerations rather than an overarching theory for effective uncertainty visualisation.

> "Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house." - Henri Poincaré (1905)

While these reviews are thorough in scope, none discuss how the existing literature contributes to the broader goal of uncertainty visualisation – that is, despite the wealth of reviews, the field of uncertainty visualisation remains a heap of stones. There is a mountain of work that identifies common heuristics found in uncertainty visualisations, evaluates competing plot designs, or starts a theoretical discussion on a niche aspect of the field. While important, each of these papers offers up its own bespoke motivation and methodology, with little reference to the uncertainty visualisation papers outside their fiefdom. The field is in desperate need of a unifying theory that can tie the conflicting and siloed research together. This review attempts to address this issue by offering a novel perspective on the uncertainty visualisation problem. That is, we will use the wealth of established stone to construct a foundation to build a house.

2.2 The purpose of uncertainty visualisation

Mentions of "uncertainty visualisation" start springing up around 1990, across several different fields (Ibrekk & Morgan 1987; MacEachren 1992), each with its own motivation for the work. In computer science, the area appears to be motivated by issues in the public's perception of random variables, with the hope that visualisations would give laypeople the ability to extract important information from graphical representations (Ibrekk & Morgan 1987). With similar concerns about the public's understanding of randomness, the fields of psychology, statistics, and economics used "uncertainty visualisations" as a communication tool to mitigate the psychological bias associated with the communication of risk, a topic of concern since the early 1980s (Spiegelhalter 2017). In cartography, it was motivated by the inherent uncertainty of geoscience data, the practical use of

visualisation as an exploratory tool, and the constrained visual channels from map representations (MacEachren 1992). These disparate motivations have blended together, and today, uncertainty visualisation is usually motivated by the vague goal of "decision-making". This term has been used to mean the mitigation of psychological bias to ensure economically rational decisions (Padilla et al. 2022, 2021; Kale, Kay & Hullman 2021), the facilitation of trust or confidence (Zhao et al. 2023; Yang et al. 2023), the ability to extract values related to a distribution (Sarma et al. 2023), the prevention of false discovery in plots (Sarma et al. 2024; Koonchanok et al. 2023), or the extraction of some other metric that is vaguely related to "uncertainty" (Chakraborty, Kiefer & Raubal 2024; Ndlovu, Shrestha & Harrison 2023). This gradual scope creep of the field, motivated by the hazy definition of "decision-making", is the most likely culprit for the jumbled literature that makes up the field today.

Given that there is so much subliminal disagreement in uncertainty visualisation, how did these disparate motivations come to be seen as interchangeable? All discussions on uncertainty visualisation seem to have a common thread that connects them: the belief that the ultimate goal of uncertainty visualisation is not trust, rationality, or value extraction, but *transparency*. We see it said directly in reviews of the field (Padilla, Kay & Hullman 2022), or when authors claim that failing to include uncertainty is akin to fraud or lying (Hullman 2020; Manski 2020). We see it when authors assert that uncertainty communicates the legitimacy (or illegitimacy) of the conclusion drawn from visual inference (Kale et al. 2018; Griethe & Schumann 2006). We see it when authors say uncertainty visualisations should communicate a degree of confidence (Correll, Moritz & Heer 2018; Boukhelifa et al. 2012) or validity (Hullman 2020; Griethe & Schumann 2006) in our conclusions. We see it when authors suggest uncertainty visualisation should "guide, qualify, or soften our judgements of uncertain data" (e.g., Wilkinson (2005) in his seminal work on the grammar of graphics). These authors are not wrong about the need for transparency in science communication: a six-month survey of anti-mask groups on Facebook during the COVID-19 pandemic showed that anti-maskers made persuasive arguments by exploiting inherent uncertainty ignored by pro-maskers (Lee et al. 2021).

Uncertainty visualisation is motivated by the need for a sort of "visual hypothesis test", a sentiment expressed by some authors directly (Correll & Gleicher 2014; MacEachren 1992). A successful uncertainty visualisation would act as a "statistical hedge" for any inference we make using the graphic. Since the purpose of a visualisation is to give a quick gist of the information (Spiegelhalter 2017), this hedging should be communicated visually without the need for complicated mental calculations. Therefore, an effective uncertainty visualisation should not just "show" uncertainty; untrustworthy conclusions should *not be visible*. If we refer to the conclusion we draw from a graphic as its "signal", and the uncertainty that should make this signal harder to read as the "noise", we can summarise the above information into three key requirements. A good uncertainty visualisation

should:

1) Reinforce justified signals to encourage confidence in results.
2) Hide spurious signals that are overwhelmed by noise.
3) Perform tasks 1) and 2) in a way that is proportional to the level of confidence in those conclusions.

Usually, visualisations that are unconcerned with uncertainty have no issue showing justified signals, but struggle with the display of unjustified signals. Therefore, we **suggest calling this approach to uncertainty visualisation "signal-suppression"** since it primarily differentiates itself from the normal "noiseless" visualisation approach through criterion (2). This is the main criterion we will use to assess the current literature on uncertainty visualisation.

2.3 Current Approaches

2.3.1 Ignoring uncertainty

The most common way to visualise uncertainty is to simply not. A study conducted by Hullman (2020) found that only a quarter of authors surveyed included uncertainty in 50% or more of their visualisations, in part because authors are not sure how to calculate uncertainty. This is not entirely unreasonable, given that even uncertainty visualisation researchers themselves seem to be in conflict about what exactly uncertainty is. We will start with visualisations that ignore uncertainty with the hope that by looking at where uncertainty isn't, we can better understand where it is.

2.3.1.1 What is uncertainty?

It is surprisingly hard to describe what uncertainty is. Most authors avoid the problem and describe the many characteristics of uncertainty. Often, uncertainty is split by factors such as whether it is due to true randomness or a lack of knowledge (Begg, Welsh & Bratvold 2014; Spiegelhalter 2017; Gustafson & Rice 2019; Padilla, Kay & Hullman 2022; Hullman 2016; Walker et al. 2003); quantifiable or unquantifiable (Spiegelhalter 2017; Walker et al. 2003; Padilla, Kay & Hullman 2022); scientific or human (Benjamin & Budescu 2018; Gustafson & Rice 2019); systematic or random (Sanyal et al. 2009); statistical or bounded (Gschwandtner et al. 2016; Olston & Mackinlay 2002); accuracy or precision (Griethe & Schumann 2006; Benjamin & Budescu 2018; Hullman 2016); etc. There are enough qualitative descriptors of uncertainty to fill a paper, but none of this is particularly helpful in understanding how to integrate it into a visualisation.

Rather than trying to define uncertainty by looking at the myriad ways in which it *does* appear in an analysis, we may find it easier to look at where it *does not*. Descriptive statistics describe our sample as

it is and summarise large data into a usable format, but they are not seen as the primary goal of modern statistics. In 19th-century England, *positivism* was the popular philosophical approach to science (positivists included famous statisticians such as Francis Galton and Karl Pearson). Practitioners of the approach believed statistics ended with descriptive statistics, as science must be based on actual experience and observations (Otsuka 2023). In order to make statements about population statistics, future values, or new observations, we need to perform inference, which requires the assumption of the "uniformity of nature", that is, we need to assume that unobserved phenomena should be similar to observed phenomena (Otsuka 2023). Positivists believed referencing the unobservable was bad science, embracing descriptive statistics due to the inherent certainty associated with them. Since uncertainty is nonexistent in descriptive statistics, it is clear that uncertainty is a by-product of inference: uncertainty is the noise that is both inseparable from our inference and meaningless without it. If we consider uncertainty to be a by-product of statistical inference, then uncertainty visualisations are the plots that depict an estimate, and therefore have an associated uncertainty. The most complete description of these estimates is their distributions. Rather than extracting just one element of the distribution, if you can retain the whole distribution, that not only allows the uncertainty calculation to be reproduced, but also makes it possible to derive other estimates as well. Suggesting distributions as a representation of uncertainty is not new. Kay (2023) originally suggested thinking about uncertainty visualisations as visualisations with distribution inputs, to replace the commonly used mean and standard deviation and remove the assumption of a Gaussian distribution. In practice, uncertainties are sometimes calculated by other people or organisations, and the process used to derive them may not be known. While some researchers believe these abstract notions of uncertainty, such as credibility (Thomson et al. 2005), forecaster confidence (Padilla et al. 2021), or uncertainty about uncertainty (Hadjimichael, Schlumberger & Haasnoot 2024), are too complex to be quantified, this is not necessarily true. Abstract concepts such as human belief or credibility are regularly quantified by Bayesians, and hierarchical approaches are often used to model uncertainty about uncertainty.

2.3.1.2 Example: ignoring uncertainty

If visualising uncertainty is fundamentally visualising a set of random variables, what does "ignoring" uncertainty look like? We investigate this question using Figure 2.1 which shows three different scenarios under which we might want to visualise uncertainty, each with a high or low uncertainty case. Each plot shows the expected value of the input distribution. In plots R1 and R2, the expected value depicts the point estimates from a simple linear regression on a car's miles per gallon (`mpg`) and weight (`wt`) using the `mtcars` data (available in `ggplot2` (Wickham 2010), and originally from Henderson & Velleman (1981)). For the low uncertainty case, the linear regression is calculated on

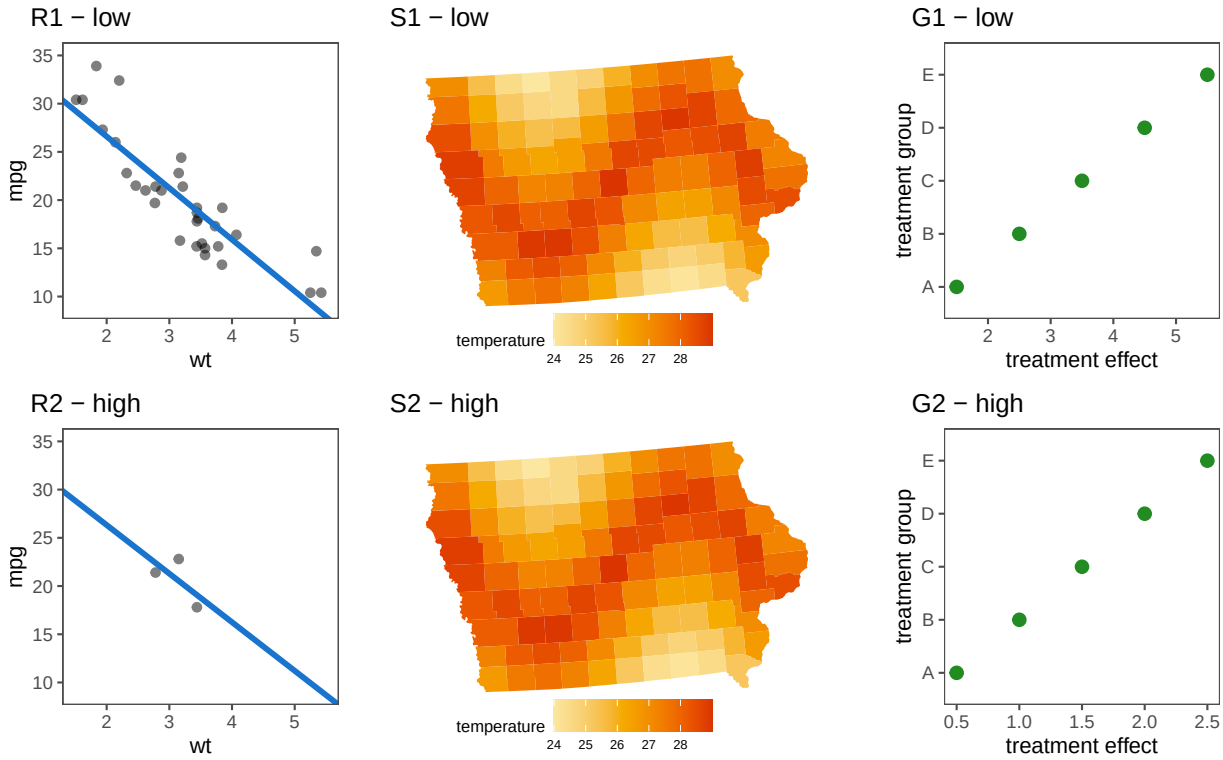


Figure 2.1: *Three example types of data and associated plots that will be used to illustrate various choices of uncertainty representation throughout the paper: scatterplot and regression line (R1, R2), spatial choropleth map (S1, S2), and grouped dotplot (G1, G2). Two levels of uncertainty (low, high) are used with each example. Ignoring the points in R1 and R2, there are no differences between the high and low uncertainty versions. Ignoring uncertainty can lead to misrepresentation of data.*

the full `mtcars` data, but in the high uncertainty case, it is calculated using a subset of three points. The points used to calculate the linear regression are also shown on the plot, however, the fitted line is the only component that should include any uncertainty. Plots S1 and S2 show a choropleth map of Iowa, where counties are coloured according to a simulated temperature measurement, with measurement error being the uncertainty associated with the measuring instrument. Plots G1 and G2 show the central value of five different simulated distributions representing five different treatment groups (A-E). In the linear regression, we see a downward trend, in the map, we see a sine wave, and in the univariate distributions, an incremental increase in treatment effect. The noise in the high uncertainty case is set such that it should overwhelm the trend in all three plots. Can you see a difference between the plots in the top row versus the bottom row? Is the strength of the trend communicated through the visualisation? The answer to both of these questions, for S1, S2, G1, and G2, is no as the high and low variance cases are identical. For R1 and R2, any ability to differentiate the plots would likely relate to the data points, rather than the regression line. The lack of differentiability between the high and low uncertainty case highlights the danger of ignoring uncertainty in our visualisations.

2.3.2 Uncertainty as a statistic

The next most common approach to uncertainty visualisation is to treat uncertainty as another statistic. Examples of this approach appear frequently in the literature. In geospatial visualisation an example can be seen in the exceedance probability map (Kuhnert et al. 2018), where areas are coloured according to the probability the estimate exceeds a specified value. An example from risk communication can be seen in the icon array (or pictograph) (Spiegelhalter 2017) that communicates the relative number of affected individuals for a particular treatment using coloured human icons. An example in statistics can be seen in the summary plot (Potter et al. 2010) which combines the plot of a single set of numeric values with a density estimate, boxplot, and statistical moments. Simply put, these approaches might be described as swapping our original statistic out for an "uncertainty statistic" to give us an "uncertainty visualisation". Some authors take these approaches because they explicitly see uncertainty as a variable of importance in of itself (Blenkinsop et al. 2000), while others straddle the line, asserting uncertainty is acting as signal and noise, and should fulfil both roles (Peña-Araya et al. 2025). This leads to the question, how does visualising uncertainty as a statistic change our view of the data, and the conclusions we draw from our plot?

2.3.2.1 Example: visualising variance

Figure 2.2 depicts the six plots introduced in Figure 2.1 but the central value has been replaced with an uncertainty statistic. Plots R1 and R2 show the residual plot of our linear regression instead of the linear regression itself. Swapping the underlying statistic completely changes the meaning conveyed by the visual structures in the plot. Thee interpretation of the slope of the line in our residual plot has little in common with the slope of our linear regression. Plots S1 and S2 show an exceed probability map, where each county is coloured according to $P(temperature > 27)$. At first glance, we might think this plot is performing signal suppression as the sine wave trend is clearly visible when the error is low, but barely visible with high error. However, just like R1 and R2, the meaning of these plots was changed when we swapped the underlying statistic. The exceedance probability map actually identifies extreme values, and the change in visibility of the signal comes from a uniformity in the shape of the distributions, rather than from signal suppression. Continuing this trend, G1 and G2 visualise the variance of the distributions, and once again show that changing the underlying statistic changes the fundamental meaning of our graphics, but with a slight twist. This plot reveals some interesting information about our data: the variance is constant in G1 but different for each group in G2. Here, we highlight what is *actually* going on when we visualise uncertainty as a statistic. The approach allows us to can glean new information about the variance itself, but this variance is treated as independent of the estimates, and does not recontextualise the signal conveyed by Figure 2.1. This then leads to the question, does visualising uncertainty as a statistic count as

"uncertainty visualisation"?

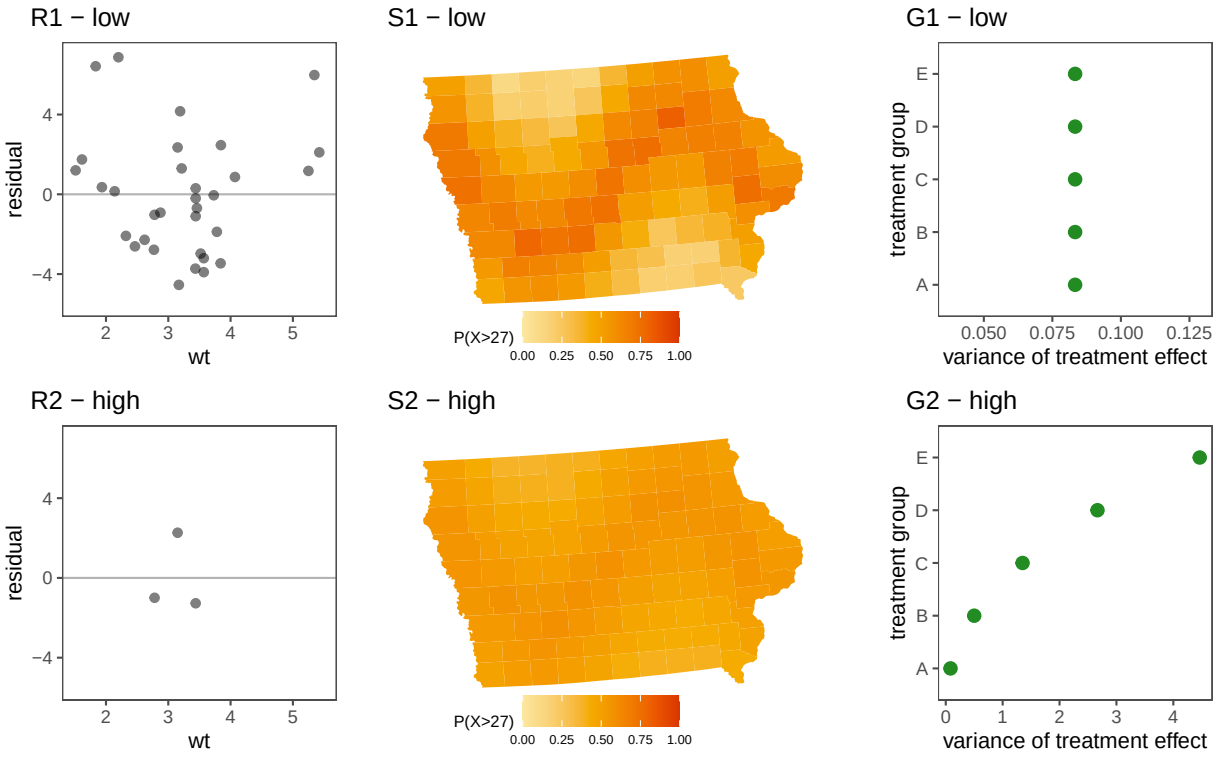


Figure 2.2: *Treating the uncertainty as a statistic, with the same six examples. The regression (R1 and R2) is a scatterplot of residuals vs explanatory variable, separated from the regression, making comparison of uncertainty related to the trend more difficult. For the choropleth (S1, S2): instead of temperature, the probability of exceeding 27°C is shown. This has the effect of highlighting (sine wave) trend in the low error data, and de-emphasising it in the high error data. The plots G1 and G2 have replaced the treatment effect with the variance of our treatment effect. It is a bit nonsensical, but we learn something interesting that was not seen earlier: the variance in G2 is not uniform like that in G1.*

2.3.2.2 What is an uncertainty visualisation, then?

What an uncertainty visualisation is or is not is one of the most pervasive divides in the literature. For example, Wilkinson (2005) mentions that popular graphics, such as pie charts and bar charts, omit uncertainty. Wickham & Hofmann (2011) suggests their product plot framework, for area plots like bar charts and also histograms, needs to be extended to include uncertainty representation. However, pie charts, bar charts, and histograms have all been used in a significant number of experiments as examples of an "uncertainty visualisation" (Ibrekk & Morgan 1987; Olston & Mackinlay 2002; Zhao et al. 2023; Hofmann et al. 2012). What is going on here?

This conflict stems from a subconscious disagreement about the purpose of uncertainty visualisation. If you believe uncertainty visualisation is about communicating risks or probabilities, uncertainty visualisations are just visualisations of "uncertainty statistics". On the other hand, if you believe uncertainty visualisation is about suppressing false visual signals, then you see an uncertainty visualisation as a transformation of an existing graphic that adds the uncertainty in. The former has no

limitation on the visual appearance of an "uncertainty visualisation", allowing pie charts, bar charts, or histograms, so long as the graphic is visualising "uncertainty", while the latter believes uncertainty visualisations only exist in relation to some "normal" visualisation. When we refer to the graphics depicted in Figure 2.2 as "uncertainty visualisations", we are classifying visualisations by the data they display, not their visual features. This is not the standard approach in statistical graphics. A scatter plot that compares means and a scatter plot that compares variances are both scatter plots.

Unlike plots, which are not defined by their underlying statistic, uncertainty can *only* by defined in relation to a particular inferential statistic. This is frequently discussed in the literature as a dependence on the "goals" of our analysis. Meng (2014) commented that what is kept as data and what is tossed away is determined by the motivation of an analysis - what was previously noise can become signal depending on the question. Otsuka (2023) suggested that the process of observing data to calculate statistics is largely dependent on our goals, because the process of boiling real-world entities down into probabilities depends on the relationships we seek to identify within our data. Wallsten et al. (1997) argue that the best method for evaluating or combining subjective probabilities depends on the uncertainty the decision-maker wants to represent, and why it matters. Fischhoff & Davis (2014) suggested we should have methods for communicating uncertainty depending on what the user is supposed to do with it. Spiegelhalter (2017) says we "cannot assess the quality of risk communication unless the objectives are clear". Peña-Araya et al. (2025) asserted that whether or not uncertainty is a source of doubt depends on the context. The sentiment behind this repeated point is clear: the role of uncertainty or signal is not dependent on the "type" of statistic, on the source of the information, or the methods we use; it is determined by the statistic we wish to draw inference on. Therefore, the fundamental problem with the "uncertainty statistic" approach is that the uncertainty in the plot isn't acting as noise; it is acting as signal.

If the uncertainty in a graphic is acting as a signal, there isn't an interesting perceptual challenge associated with the visualisation: the uncertainty can be displayed using standard principles of graphic design. In changing the inferential statistic, we also haven't dealt with the original problem of integrating noise, as these "uncertainty statistics" *also have associated uncertainty in the estimates* (e.g., variance of standard deviation estimate) that is being ignored. There is nothing wrong with explicitly visualising variance, error, bias, or any other statistic. These metrics provide important and useful information for analysis and decisions. The problem with this approach is that it is so broad that it defines everything as an uncertainty visualisation, and if everything is an uncertainty visualisation, nothing is.

2.3.3 Uncertainty as a variable

Another common characterisation of uncertainty is as just another variable to be integrated into the visualisation, which means uncertainty visualisation is, at its core, a high-dimensional visualisation problem (Griethe & Schumann 2006; Peña-Araya et al. 2025). This approach emerges in computer science (Kinkeldey, MacEachren & Schiewe 2014), cartography (MacEachren et al. 2005), and statistical graphics (Wilkinson 2005) alike. Discussion on these visualisations focuses on how "integrated" the uncertainty is with the estimate. Kinkeldey, MacEachren & Schiewe (2014) identified a split between intrinsic plots, where we map uncertainty to the colour or size of the geometric object of our estimate, and extrinsic plots, where uncertainty is mapped to a separate geometric object, such as glyphs or error bars. Similarly, Padilla, Kay & Hullman (2022) classified uncertainty visualisations as graphical annotations (extrinsic), and probability mapped to a visual encoding channel (intrinsic), or a hybrid of the two. It is unclear if these levels of "integration" in a plot design affect its ability to suppress signals.

2.3.3.1 Example: mapping two independent variables

Figure 2.3 shows the six examples introduced in Figure 2.1, with uncertainty mapped to a spare aesthetic in the visualisation. Aesthetics are a component of the grammar of graphics that represent the visual stimuli our variables are mapped to within a plot, such as position, colour, and size. In plots R1 and R2, the standard error of the slope is represented by the width of the line. This is common, but it fails to represent the standard error of the intercept alongside the slope. We might think we can examine the width of the line at `wt=0`, but this would be incorrect. Here, it has been included, less obviously, by mapping the standard error of the intercept to transparency. While these help to give a gestalt of the uncertainty, they are not exacting representations. One would expect that the width above and below the line is one standard error, but why not map the width to two standard deviations, or even three? Interpreting transparency into a numerical quantity is also virtually impossible, so using this aesthetic mapping is effectively useless.

Either way, the trend is still visible in both displays. A bivariate colour palette map is shown in plots S1 and S2. With two dimensions of the plot reserved for spatial position, it is difficult to incorporate the error. The bivariate colour palette maps the estimate to hue and the error to saturation, keeping the signal and noise contained to one visual aesthetic. This has the unfortunate effect of making the signal appear stronger when the error is higher (S2). Colour perception is a wild beast that is hard to tame. The extrinsic approach, shown in plots G1 and G2, has the estimate represented by a point, and uncertainty computed as a 95% confidence interval mapped to line length. It could be argued that where the variance is high, the trend remains a main focus; that is, the display fails to sufficiently suppress the signal. Because all of these graphics visualise the distribution's estimate and uncertainty

as two separate pieces of information, the message of the plot is "here is the trend *and* here is the uncertainty". It is worthwhile to examine why this occurs, to see if we can move towards a version of this plot where we are able to communicate signal and noise simultaneously.

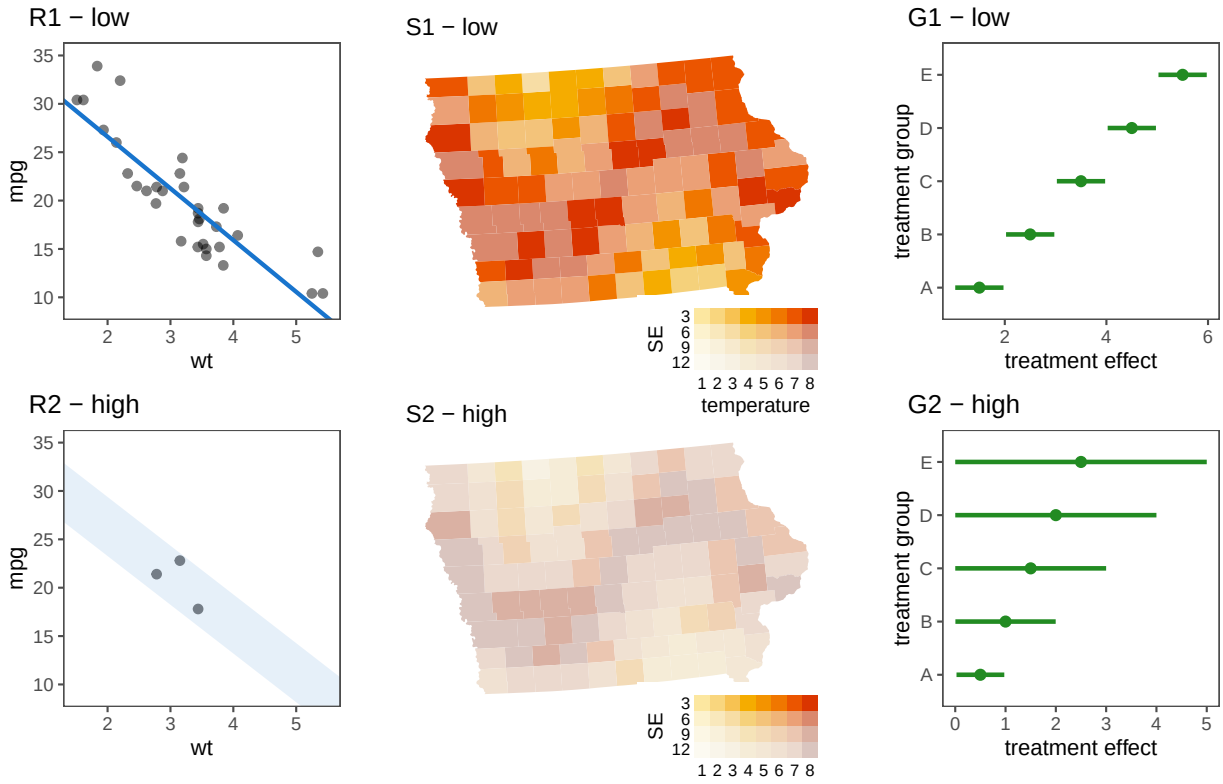


Figure 2.3: *Treating the uncertainty as a variable, with the same six examples. In plots R1 and R2, the standard error of the slope is mapped to the line width. The standard error of the intercept is mapped to the transparency, which is less conventional. In plots S1 and S2, a bivariate colour palette is used with mean mapped to the hue, and error mapped to saturation. Plots G1 and G2 represent the mean of each group as a point, and the variance as an interval. In this example, the uncertainty has been integrated with the signal, but the results are undesirable. This is especially true in the case of the choropleth map, where the signal is easier to see when the error is higher.*

2.3.3.2 Can we visualise a "single integrated uncertain value"?

The reality is, based on our discussion on inferential statistics, uncertainty *isn't* a separate variable: it is a component of the random variable that is indistinguishable from the random variable itself. Similarities between the "as a variable" approach and the "as a statistic" approach become apparent when we read the motivations behind the method. When visualising a point estimate (i.e., Figure 2.1) and variance (i.e., Figure 2.2) side-by-side, one needs to switch focus between two displays, leaving us vulnerable to change blindness (Simons & Levin 1997). The preference for the methods utilised in Figure 2.3 is usually motivated by the difficulties in combining information on two separate graphics (Moritz et al. 2017; Correll, Moritz & Heer 2018), rather than an understanding that the "uncertainty statistic" approach is not philosophically sound. To achieve signal suppression, we need to visualise

noise and signal together as a "single integrated uncertain value" (Kinkeldey, MacEachren & Schiewe 2014) rather than as two separate statistics.

Just because we can still see the signal in Figure 2.3, that does not mean the reading of the estimate is completely independent of the uncertainty. When making any visualisations, we usually want the visual channels to be separable, that is, we don't want the data represented through one visual channel to interfere with the others (Smart & Szafir 2019). Separability may be desirable in standard data visualisation, but in uncertainty visualisation, it allows the estimate and its variance to be read independently, potentially leading to the uncertainty being ignored (Padilla, Kay & Hullman 2022). Therefore, rather than trying to maintain visual separability, the goals of uncertainty visualisation align far better with the pursuit of visual integration. In an ideal system, our estimate and uncertainty would be manipulated separately, but would be so *well-integrated that they are read as a single channel by the human brain*. The problem is that even if we can implement the most extreme versions of integrable, our methods fall short, as illustrated by the bivariate colour palette map in Figure 2.3. Colour hue and brightness are one of the classic examples of integrable variables (Vanderplas, Cook & Hofmann 2020), and decreasing saturation should make the colours harder to distinguish, but the signal is still clearly visible in the high variance case. This is to say nothing of the fact that multi-dimensional colour palettes can make the graphics harder to read and less accessible (VanderPlas & Hofmann 2015).

2.3.3.3 Another example: mapping combined variables

The Value Suppressing Uncertainty Palette (VSUP) (Correll, Moritz & Heer 2018) was designed with the intention of preventing high uncertainty values from being extracted from a map by blending colours together as they become less certain. Figure 2.4 shows the spatial example (S1, S2) using the VSUP approach. Since the palette was designed with the extraction of individual values in mind and it has only been tested on simple value extraction tasks (Correll, Moritz & Heer 2018) or search tasks (Ndlovu, Shrestha & Harrison 2023). We can see that, at least for our example, when the uncertainty is high, the spatial trend has functionally disappeared.

Finally, we have signal suppression! Well, not really, sorry, we tricked you. The two plots depicted in Figure 2.4 actually show the exact same data; they are both the low variance case. If you look closely, you can see that the two plots have a different scale, where plot A has been scaled according to our existing knowledge about this data, while plot B has been scaled using the range of the data passed to the plot. Since the variance and estimate are scaled independently, arbitrary differences in the range of our variance, unrelated to the estimate itself, will have significant impacts on the visual appearance of our plot. The scale issue in VSUP maps was also recognised by Kay (2019), who

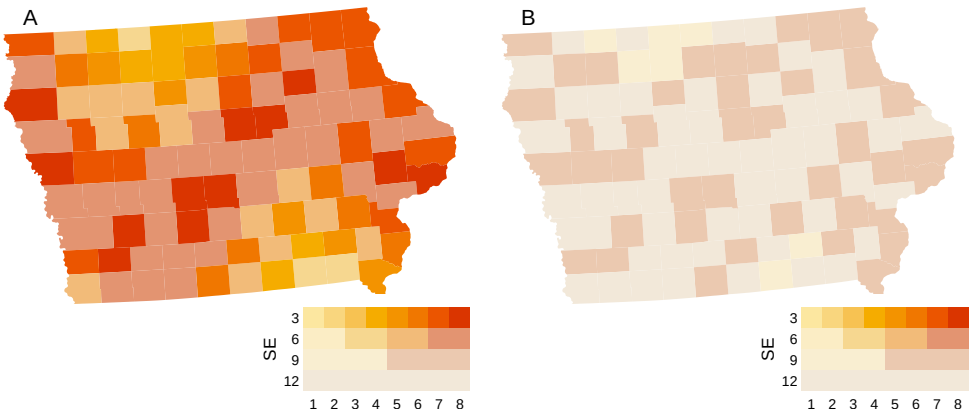


Figure 2.4: *The spatial examples displayed with a choropleth map using a VSUP colour palette, where hue is blended when increased uncertainty. Plot B has successfully produced signal suppression. Look closely at the scales, though: it may have another explanation.*

noted that the suppression of any one hypothesis largely depends on the methods we use to combine the palette, and the variance levels at which the blending occurs. This means that, for us to know that our plot will successfully perform signal suppression, we need to already know what signal we are trying to suppress and set up the VSUP palette accordingly. This means that VSUP maps are not suitable for exploratory data analysis.

2.3.3.4 Uncertainty and exploratory data analysis

The lack of uncertainty in descriptive statistics is due to the lack of inference. Descriptive statistics are actually a small piece of a much larger field, exploratory data analysis (EDA). Tukey et al. (1977) described EDA as the process of searching for interesting hypotheses ("the greatest value of a picture is when it forces us to notice what we never expected to see"), and defined it in relation to confirmatory data analysis (CDA), the process of verifying a hypothesis. There are more subfields of EDA: initial data analysis (Huebner, Vach & le Cessie 2016; Chatfield 1985), which involves checking assumptions and data quality prior to CDA, and model diagnostics (e.g., Belsley, Kuh & Welsch (1980)), including posterior checks of model fit. What binds these pursuits together is their reliance on visual summaries for making assessments and an absence of formal inference.

Hullman & Gelman (2021) argued that the EDA and CDA are not entirely distinct, as it is often difficult to draw a hard line. Our belief, as with many concepts, is that these approaches exist on a continuum, where we have an inherent trade-off between the number of hypotheses we can look for and the certainty of any conclusions reached. It can help to think of the knowledge-generating process of EDA and CDA as the nozzle on a hose with multiple spray options, where EDA is a fine misting spray that touches everything in the room, and CDA is a high-pressure jet capable of obliterating any and all debris from any single spot.

Viewing EDA and CDA as a dichotomy can create some confusion when it comes to understanding the

source of uncertainty in our analysis. This is why we have avoided the topic until now, despite the fact that an uncertainty visualisation system for EDA is one of the most discussed topics in the field (Hadjimichael, Schlumberger & Haasnoot 2024; Peña-Araya et al. 2025; MacEachren et al. 2005; Sarma et al. 2024; Griethe & Schumann 2006). The EDA versus CDA dichotomy can be compared to the dichotomy between induction, for building theories, and deduction, for testing theories, from formal logic. One of the trade-offs in the two methods is that deductive conclusions provide certainty, while conclusions from induction are inherently uncertain. This means that EDA, the inductive counterpart, has uncertainty in any conclusions reached, with the requirement to follow up our newfound hypothesis with CDA if we want true certainty. This adds another layer of confusion to the study of uncertainty visualisation: conflating the uncertainty in our data with the uncertainty that is inherent to an exploratory process (EDA).

Authors often over-compensate for the inherent EDA uncertainty, pre-emptively hedging against every false inference that could possibly be drawn from a graphic. Hullman & Gelman (2021) argues there is no such thing as a "model-free" visualisation. Guo et al. (2025) provides a grammar for visualising statistical model checks. The lineup protocol (Buja et al. 2009) provides the viewer with plots of the data in a field of plots of null data where any patterns seen are due to sampling variability. The Rorschach protocol, from the same paper, shows only null plots to give the reader some intuition for what spurious sampling patterns exist. Savvides et al. (2019) provides a statistical super-test against multiple comparisons driven by probabilistic arguments. There is a CDA quality to these approaches. The sum total of a lot of CDA is not EDA, just as swinging a high-pressure jet around a room is not equivalent to using a misting spray. While EDA and CDA may be along a continuum, we cannot simultaneously perform EDA and CDA as the approaches are, philosophically speaking, perpendicular to one another.

To truly create an uncertainty visualisation approach that is capable of EDA, we need to accept that uncertainty is inherent to the method, and it cannot be pre-emptively removed from the visualisations. If we accept this fact, then the only possible source of uncertainty in an uncertainty visualisation system that performs EDA is from the data itself. That is to say, the only possible way for there to be uncertainty in a visualisation designed for EDA is that the data itself represents inference that has been done earlier in our analysis. We can see this in the original description of Figure 2.1, where the uncertainty in all cases, even measurement error, represents inference that was performed earlier in our analysis.

In the VSUP approach, our ability to arbitrarily decide which values to blend at, or which suppression approach to use, means that the "uncertainty" we are visualising will be informed by the conclusions we

are drawing, and not a product of the data itself – an antithetical approach to EDA. For a visualisation to be suitable for EDA, it should always look the same regardless of what hypothesis we plan to draw from it. As much of the transparency in data visualisation comes from this feature in EDA, it is reasonable to set it as a requirement of our visualisations. Ensuring this property means we cannot treat signal and noise as separate variables, but rather as a single integrated unit.

2.3.4 Uncertainty as a distribution

Rather than trying to reduce random variables to a single value or pair of values, why not visualise the whole distribution? This approach is found in computer science's hypothetical outcome plots (HOPs), which animate a sequence of potential outcomes of a distribution (Hullman, Resnick & Adar 2015), in geoscience's pixel-maps (Lucchesi, Kuhnert & Wikle 2021; Blenkinsop et al. 2000), and in statistics multiple forecasts (Hyndman & Athanasopoulos 2021).

2.3.4.1 Example: visualising samples

Figure 2.3 shows the six examples introduced in Figure 2.1, shown as distributions. The distribution is represented by either a sample of outcomes or a quantile dot plot. Plots R1 and R2 show a linear regression as a sample of possible outcomes from the distribution. The distribution used for both the slope and intercept is normal with the conventional mean and standard error. Plots S1 and S2 show a pixel map, which is a choropleth map where each county is coloured by a sample of outcomes from the distribution. Plots G1 and G2 show each univariate distribution as a quantile dot plot. We can see that the strong downward trend in the linear regression, the sine wave in the choropleth map, and the incremental increase in the univariate distributions are all clearly visible in the low variance case, but disappear in the high variance cases. The graphics have achieved signal suppression. Visualising the estimate as a distribution also gives additional information, such as the previously hidden bimodality of the univariate distributions in G1 and G2.

2.3.4.2 Quantified versus unquantified uncertainty

By showing our distribution as "data", we are able to read the "uncertainty" plots using the same perceptual mechanisms we use to read the "non-uncertainty" plot. This should lead to more effective communication, as people tend to read more complicated visualisations like the bivariate and VSUP plots the same way they read the simple choropleth counterpart (Ndlovu, Shrestha & Harrison 2023). This approach also does not significantly hinder our ability to extract the individual values mapped by the previous plots, as extracting global statistics from a sample can be done with relative ease (Franconeri 2021). We are able to include more information by offloading more computation to visual processing, but what is the limitation on this approach? Which aspects of uncertainty should be processed by our visual system, and which should be processed by our statistical computation? It

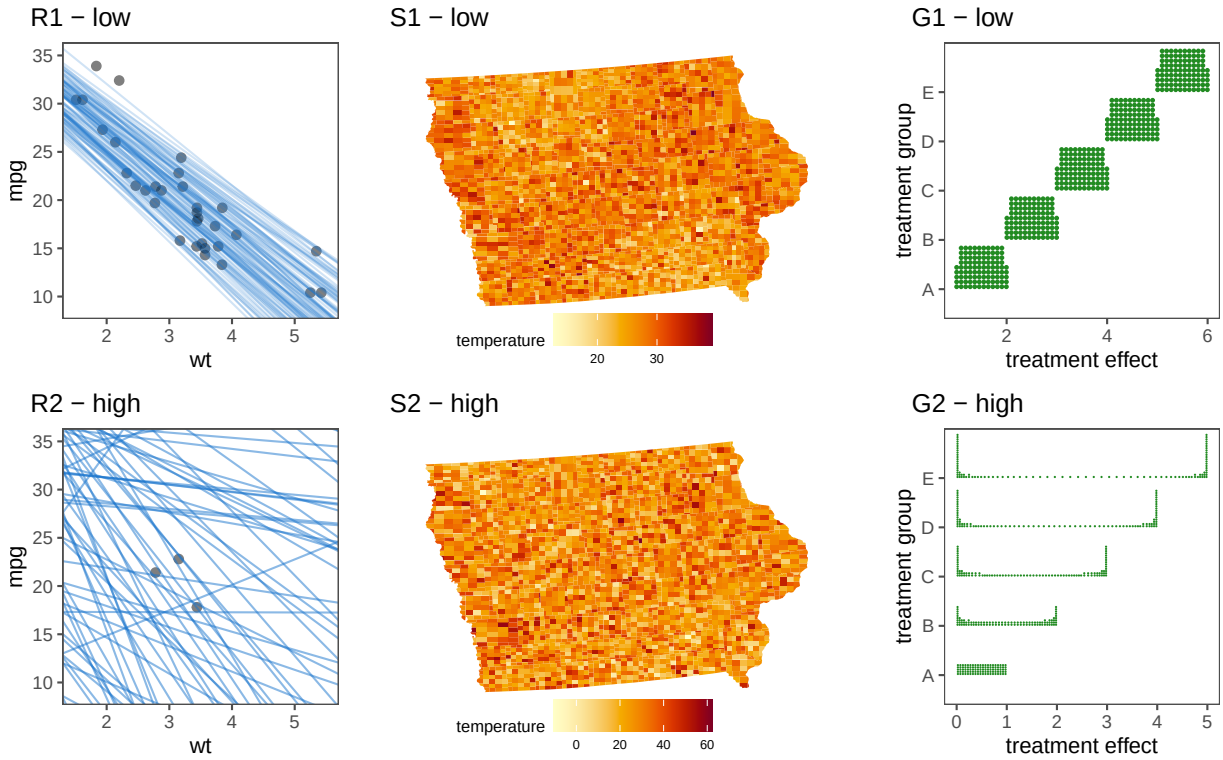


Figure 2.5: *Treating the uncertainty as a distribution, with the same six examples. Plots R1 and R2 show a linear regression as a sample of possible outcomes from the distribution. Plots S1 and S2 show a pixel map. Plots G1 and G2 show the groups as quantile dot plots. In each case, the signal (regression line, sine wave, increasing trend) has disappeared with high uncertainty.*

is not obvious from the question, but this is actually a question about how much of our uncertainty should be quantified.

Quantified uncertainty usually focuses narrowly on concepts such as probability, confidence intervals, variance, error, or precision (Hullman et al. 2018; Maceachren et al. 2012; Thomson et al. 2005), while unquantified uncertainty often includes a broader range of concepts like missing values, reliability, model validity, or source integrity (Griethe & Schumann 2006; Wilkinson 2005; Pang, Wittenbrink & Lodha 1997; Pham, Streit & Brown 2009; Boukhelifa et al. 2017). When discussing uncertainty, we typically include these unquantified uncertainties, not because these things *are* uncertainty, but because they can *create* uncertainty when we perform inference. This is often because these unquantified uncertainties violate our assumptions of the uniformity of nature (Otsuka 2023).

Sometimes we are able to visualise these assumption violations directly. For example, we can check for structure in our missing data using the `naniar` package Tierney & Cook (2023) that allows us to include missing values as a "shadow" alongside our usual visualisations. This approach amounts to just "showing the data", which is a simple but effective option for uncertainty visualisation that is largely overlooked. While this approach is useful for better understanding data, it will not eliminate

trends that have become invalid due to structure in missing data or an invalid model. We can only integrate uncertainty as noise when that uncertainty has been *quantified* as an effect on the estimates we are visualising. This is not to say one method is preferable; visualising both quantified and unquantified uncertainty is necessary for a healthy analysis. Data analysis often works in cycles, where we find assumption violations using EDA, quantify the effect of these violations on inference, and then visualise the output of that inference using uncertainty visualisation.

2.4 Evaluating uncertainty visualisations

Unfortunately, the conflicting results in the field are not limited to plot design and extend to the experimental findings as well (MacEachren et al. 2005; Kinkeldey, MacEachren & Schiewe 2014; Hullman 2016). There are as many explanations for the noisy evaluation studies as there are contradictions in the research itself. Kim et al. (2019) believes there is some interference in results from participants' prior beliefs; Hullman (2016) believes the noise in the literature could come from visual heuristics, subjective probabilities, unknown participant utility functions, or a misunderstanding of statistical concepts (such as confidence intervals). Kinkeldey, MacEachren & Schiewe (2014) suggest the perception of visualisation changes by audience, so we cannot expect the same results between different subpopulations. Brennen & Tuerk (2018) attributes evaluation difficulties to cognitive load from complicated uncertainty visualisations, as well as the participants' prior experience in the topic. While these issues will certainly have some impact on our ability to synthesise, none of them is unique to uncertainty visualisation. Rather, the issue is likely due to a disconnect between the evaluation methods used and the stated goals of each experiment, a common issue in visualisation evaluations (Vanderplas, Cook & Hofmann 2020).

2.4.1 Current evaluation methods

2.4.1.1 Value extraction

Uncertainty visualisations are most commonly evaluated based on how accurately viewers can extract an estimate and its variance (Kinkeldey, MacEachren & Schiewe 2014; Hullman et al. 2019). This is not unusual, as direct observation is the simplest way to verify that information can be accurately read from a graph (Vanderplas, Cook & Hofmann 2020). Unfortunately, this approach doesn't work for uncertainty visualisation. The second we ask a specific question about a statistic, that statistic becomes inferential, even if the plot was not the intent behind the question. By shifting the focus from $\hat{X}$ to $Var(\hat{X})$ or $P(\hat{X} < x)$, we end up evaluating visualisations on their ability to convey uncertainty statistics, rather than their ability to perform signal suppression. Even if the authors do not realise it themselves, there is nothing unique to uncertainty in these studies, so when we boil the findings down

to generalised results, they simply restate existing principles within information visualisation. Some of the findings are obvious: participants were more accurate when reading a probability expressed as text than when they had to extract it from a graphic (Cheong et al. 2016; Savelli & Joslyn 2013). Other studies replicate existing research, such as the finding that a probability mapped to a position is more accurate than one mapped to an area (Ibrekk & Morgan 1987; Gschwandtner et al. 2016), established as part of the hierarchy of perceptual tasks more than 40 years ago (Cleveland & McGill 1984; replicated by Heer & Bostock 2010). This extends beyond simple accuracy evaluations: Sanyal et al. (2009) found that colour was more effective than size when searching for extrema in variances; we have known that pre-attentive aesthetics, such as colour, are more efficient for search tasks since the 1980s (Vanderplas, Cook & Hofmann 2020). By classifying these studies as evaluations of "uncertainty visualisation" while evaluating uncertainty as a signal, we are encouraged to see successful examples of signal suppression as failure. This approach leads authors to advise against particular aesthetic mappings for uncertainty, because they cause participants to have more difficulty extracting values (Blenkinsop et al. 2000). This conclusion is antithetical to the goals of signal suppression and occurs because these methods evaluate uncertainty as a signal, not as noise.

2.4.1.2 Trust, confidence, and risk aversion

If we cannot directly measure uncertainty for fear that it turns into a signal, we might then assume we can measure the secondary benefits of increased transparency. This seems to be the approach of many visualisation authors, as secondary benefits such as trust, confidence, and risk aversion are all frequently used in uncertainty evaluation studies (Hullman et al. 2019). Unfortunately, measuring these secondary effects often leads to confusing conclusions that simultaneously argue for and against the inclusion of uncertainty.

This is most commonly noticed in the use of trust as a measure, as several authors have commented that measuring trust, and not transparency, can lead to a questionable subtext that argues against transparency (Spiegelhalter 2017; O'Neill 2018). We see this directly play out in the visualisation literature, where surveyed visualisation authors explicitly said they didn't include uncertainty due to the fact that they might decrease trust in their conclusions (Hullman 2020). This sentiment is also true for confidence, as Blenkinsop et al. (2000) commented that visually integrable depictions of uncertainty should be avoided, as they decrease the viewer's confidence in their extracted data values.

Another secondary effect that is similar to trust and confidence is risk aversion. Risk aversion is an economic term used to describe an agent who would choose a random variable with a lower expected payout because it also has a lower variance. Risk aversion's role in the uncertainty visualisation is

unclear, as authors will argue uncertainty should elicit more risk aversion in one paper (Hullman et al. 2019), and argue for less risk aversion (by proxy of suggesting rational agents as a benchmark) in the next (Wu et al. 2023). Ultimately, these approaches have similar issues to value extraction studies, except they are slightly more confusing in their goals, leading them to simultaneously argue for and against the inclusion of uncertainty in a visualisation.

2.4.1.3 Alternative approaches

Often, authors understand that the effects of uncertainty are more complicated than simple value extraction. These studies indicate that accurately capturing uncertainty will be more complicated than simply avoiding value extraction or trust as a measure.

One approach is to ask indeterminate questions, such as asking participants for the "best estimate" (Ibrekk & Morgan 1987), or to select which distribution is the "furthest to the right" in a lineup (Hofmann et al. 2012). In both cases, the ground truth is based on the mean of the distribution, which is not as indeterminate as the question. This approach can lead to inconclusive results, as we are left unclear whether it was the phrasing of the question or the plot design that caused the participants to answer incorrectly.

On the other hand, questions that are incredibly specific about the distribution information can confuse the participants and induce noisy results. For example, Hullman, Resnick & Adar (2015) asked participants to compare two normally distributed groups, A and B, and had many participants say that group A was more likely to be bigger, despite group B having a higher mean. Gschwandtner et al. (2016) asked participants the "probability that the interval has already ended at the marked point in time?" and participants replied with the probability that the interval had already started.

The confusion around trying to capture the effects of uncertainty can also (understandably) extend to the authors of the study itself. We can see an example of this in Padilla, Ruginski & Creem-Regehr (2017). In order to answer the question correctly, the first experiment required participants to assume that an oil rig being "more likely to be hit" by a hurricane would *not* translate to the rig sustaining "more damage". The second experiment required participants to assume the opposite.

2.4.1.4 Effective methods

This is not to say all evaluation studies fail to properly evaluate uncertainty as noise. There are several studies that ask participants to identify a particular signal that the noise is trying to obfuscate (Kale et al. 2018; Correll & Gleicher 2014), which seems to be an effective method. The only problem with these studies is that most uncertainty visualisation methods are not integrated into *the grammar of graphics* (Wilkinson 2005; Wickham 2010), so we regularly see comparisons between disparate plots that would never be considered substitutes for one another outside the artificial "uncertainty

visualisation" framework they are placed within. For example, Kale et al. (2018) compared static bar charts with error bars to a bar chart with animated samples, meaning that any difference in participants' ability to read the plot could be due to the statistic (confidence interval versus sample), the geometry (bars versus intervals), or the use of animation (static versus animated plots). This issue was rectified in their second experiment, where they compared overlayed and animated samples, giving us an insight into the types of visualisations that are appropriate to compare in uncertainty visualisation experiments. This means that even when evaluations *are* done correctly, there is no generalisable theory we can take from the results.

The point here is not to accuse the authors of poor academic rigour. The papers are (usually) logically consistent and well-formulated pieces of work. Rather, the point is to illustrate that evaluating uncertainty as noise is surprisingly difficult. Designing tests for signal suppression will require a formalisation of uncertainty within the grammar of graphics, as well as improved evaluation methods.

2.4.2 Implicit Hypothesis Testing

The main problem with current uncertainty visualisation evaluations is that they often require explicit (or convoluted) questions about the variance. Asking direct questions about the statistics or outcomes is not an explicit requirement of visualisation evaluations. In their review of testing statistical graphics, Vanderplas, Cook & Hofmann (2020) drew a distinction between explicit tests, where participants are asked direct questions about specific features of a plot, and implicit testing, where users identify both the purpose and function of the plot. The lineup protocol is the most salient example of the implicit approach. Lineups are a confirmatory visualisation tool where participants are shown a set of M plots, and asked to identify the plot that is the "most different", leaving participants to decide what "most different" means to them, even if it is not what the authors intended (VanderPlas & Hofmann 2017). The implicit test does not limit the versatility of the approach, with the lineup being used to evaluate the effectiveness of different types of plots (Hofmann et al. 2012), colour palettes (Reda & Szafir 2021b), and design decisions (VanderPlas & Hofmann 2017).

Lineup protocols are not only useful for implicit testing: they also have parallels to hypothesis testing that can be leveraged in uncertainty visualisation. The concept of signal suppression is, at its core, an assertion of statistical validity: the visibility of signals should be directly proportional to p-values or some equivalent measure. This comparison is not new in uncertainty visualisation, where parallels have been drawn to frequentist statistics by Correll & Gleicher (2014), who compared results to Cohen's D, and to Bayesian statistics by Kim et al. (2019), who evaluated plots based on their impact on the users' prior beliefs. The comparison to hypothesis testing is far more natural for the lineup protocol, which has a visual test statistic and can be compared to results from Bayesian analysis

(VanderPlas & Hofmann 2017) or standard statistical tests using power curves (Majumder, Hofmann & Cook 2013; Li et al. 2024). The connections between lineups and uncertainty visualisation are numerous and have been previously identified in the development of HOPs (Hullman, Resnick & Adar 2015).

The lineup protocol and uncertainty visualisations are similar: lineups were designed for checking if perceived patterns are real or merely the result of chance (Buja et al. 2009; Wickham et al. 2010). As both approaches are attempting to do the same thing, it is likely that we are unable to leverage the lineup protocol directly to evaluate uncertainty visualisation, but a new evaluation methodology should try to learn from the success of the lineup approach. Designing an implicit testing method for uncertainty visualisation that allows us to draw parallels to standard notions of statistical significance would solve many of the issues with the current evaluation approaches.

2.5 Conclusions and Future Work

This paper examines the literature and provides suggestions for a structural framework to support uncertainty visualisation. Particularly, we propose that uncertainty visualisation should accomplish signal suppression, dampening weak signals, and amplifying strong signals. We have also highlighted several gaps in the existing literature.

Experimental practices on uncertainty visualisation need standards. Some existing evaluation experiments treat uncertainty as a signal, while others treat uncertainty as noise. As a result, it is difficult to combine results from papers to get a meaningful sense of how uncertainty information is understood by a viewer. Researchers need to ensure that when they identify the motivation behind their visualisation technique, their evaluation methods align with the stated goals of the paper.

Experimental methods that evaluate uncertainty as noise need to be developed. Research into separability and integrability of signal and noise is of particular interest to uncertainty visualisation, as it allows assessment of the interference between the two. When designing experiments, authors often choose aesthetics that are visually distinguishable; uncertainty visualisation authors should consider doing the opposite.

Uncertainty needs to be formalised within the grammar of graphics. Some of this formalisation was done by Kay (2023), but it focuses only on the visualisation of univariate distributions. Giving authors the ability to describe uncertainty visualisations in terms of statistics, geometries, and aesthetics will support evaluation experiments that can build towards a cohesive theory of visualising uncertainty.

Software that allows users to easily perform signal suppression is needed. Existing uncertainty visualisation methods view a distribution as its own object, and there are no software options treating "an uncertainty visualisation as a function of an existing visualisation" philosophy.

Signal suppression is an undeveloped area of visualisation research, and developing methods for the practice may require us to challenge our entire notion of what makes a good visualisation.

Reproducibility

The R packages were used for this work were: `tidyverse` (Wickham et al. 2019), `RColorBrewer` (Neuwirth 2022), `scales` (Wickham, Pedersen & Seidel 2025), `sf` (Pebesma & Bivand 2023), `urbnmapr` (Strochak, Ueyama & Williams 2024), `flextable` (Gohel & Skintzos 2024), `colorspace` (Stauffer et al. 2009), `ggdist` (Kay 2023), `ggdibbler` (Mason et al. 2026b), `patchwork` (Pedersen 2025b), `distributional` (O'Hara-Wild et al. 2024), `ggthemes` (Arnold 2024), `broom` (Robinson et al. 2026), and `rgeos` (Bivand & Rundel 2023). The GitHub repository for this paper can be found at https://github.com/harriet-mason/ARSA-UncertaintyLitReview, which contains the files required to reproduce this article in full.

Chapter 3

A Mathematical Framework and Software Implementation for Uncertainty Visualisation

3.1 Introduction

> "Elegant design requires us to think about a theory of graphics, not charts." Wilkinson (2005)

Uncertainty visualisation has suffered from a serious lack of formalisation in recent years, which has resulted in an overwhelming tsunami of named plots, touching every corner of the literature. No visualisation problem is too niche as we see visualisations of 2D-intervals called a cross plot, rectangle plot, segment plot, and dandelion plot (Zhang & Lin 2022). No plot change is too small with simulated outcomes on lines called spaghetti plots (or lasagne plots for gradients) (Swihart et al. 2010), on maps called pixel maps (Lucchesi & Kuhnert 2020), on bar charts called fuzzygrams (Kay 2023), and when the outcomes are animated, we call them hypothetical outcome plots (HOPs) (Hullman, Resnick & Adar 2015). No dead horse is too beaten with probability functions having more names than ways to differentiate them with terms like histogram, density plot, violin plot, ridgeline plot, rain plot, dot plot, and gradient plot, only scratching the surface (Kay 2023). The practice of naming plots turns the field into a race, where authors are incentivised to brand their name on as many graphics as possible, rather than work towards a cohesive theory of visualisation. This problem is as pervasive as it is frustrating.

According to Wilkinson (2005), "the computer woefully focuses the mind", so we might assume that

these hyper-specific naming conventions do not extend to the landscape of uncertainty visualisation software. This assumption would be wrong. Even just within the R ecosystem of `ggplot2` extensions, there is an overwhelming number of choices, leaving users unclear when each package should be used and for what purposes. A density plot can be made using `ggplot2`, `ggbeeswarm`, `ggridges`, `ggrain`, `ggdist`, `ggpointdensity`, and `ggdensity`, all packages with similar descriptions and overlapping designs. This is not to say these packages are pointless: they have thousands of downloads per week and were motivated by a gap in the literature. Rather, this example illustrates the havoc that named plots contribute to our software ecosystem.

With this wealth of choices in plots and software, it would be reasonable to assume that the field is meeting the needs of its users, albeit in a messy, roundabout way, but this assumption would be incorrect. Despite this wealth of choice, each package is brittle and constrained by its own set of assumptions. There is a cry for flexibility among the users of uncertainty visualisation software, dating back more than 30 years (MacEachren 1992). A useful uncertainty visualisation system should allow for exploration with uncertain data from multiple different sources, dimensions, and data types (Hadjimichael, Schlumberger & Haasnoot 2024; Peña-Araya et al. 2025; MacEachren et al. 2005), a flexibility that all current implementations fail to provide.

This implies the real gap in uncertainty visualisation is not new plots or bespoke software, but structure, and the flexibility that comes with it. Structure is not so alien to the visualisation community that this is an unreasonable request. Standard statistical graphics have been formalised within **the grammar of graphics** (Wilkinson 2005; Wickham 2010) for several decades. This structure has largely been ignored by the uncertainty visualisation community, with the exception of Kay (2023) and the `ggdist` R package. Along with this formalisation of visualisations of univariate probability functions, Kay expressed hope for a single coherent uncertainty visualisation framework. This single unified framework, and its implementation in the `ggdibbler` package, is what will be discussed in the rest of this paper.

3.2 A motivating example

We are going to build our uncertainty visualisation system upon a simple assumption based on the existing literature: the input for an uncertainty visualisation is a data set where every cell is a random variable (Kay 2023) which contains all the quantified uncertainty we wish to represent (Mason et al. 2026a). Once we have our distribution inputs, there is only one question left for our uncertainty visualisation system to solve. What exactly should it *do* with these distributions?

Consider the case of a vector, X, that contains the heights of 15 different women, where each $x_i \in \mathbb{R}$ represents one measured height. If we feed this vector into a density plot function, such as `geom_density()` from `ggplot2`, how should it behave? Hopefully, this case is obvious, and the output of this function is shown in the `ggplot2` example in Figure 3.1. A density curve of our data is estimated and displayed using a line geometry. This is a straightforward example, but what happens when our input is a distribution? This can occur if, for example, the tool used to record the heights has an inherent measurement error. Now, we have a vector, $\mathbf{X}$, of 15 distributions, where each $\mathbf{x}_i \sim N(x_i, \sigma_i)$, where $x_i \in X$ is the recorded value, and σ_i is an estimated variance based on the environmental conditions of the measurement. There are two routes we can take when it comes to visualising this estimate. The second plot in Figure 3.1 is made using `ggdist`, and it shows the case where we are interested in the distribution of each individual measurement. The third plot is made using `ggdibbler`, which places the emphasis on the full data density *but* carries forward the variability in the density that comes with having distributional inputs.

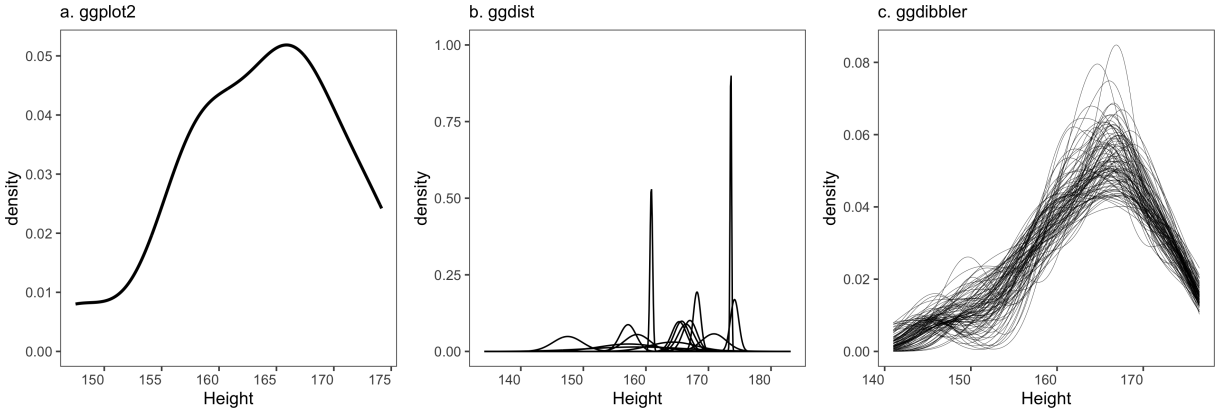


Figure 3.1: *Alternative interpretations of how to render a density plot when the input is a set of distributions describing uncertainty of measurements, according to three plotting packages: (a) `ggplot2` forms the density from the mean values, (b) `ggdist` puts the distribution on each observation, treating uncertainty as signal, (c) `ggdibbler` shows the densities for multiple samples, which puts the focus on how the density might look given the uncertainty. These differences illustrate how uncertainty is interpreted in different ways. Which is correct?*

The distinction between the two approaches presented in Figure 3.1 is the same signal and noise paradigm presented by Mason et al. (2026a). In the `ggdist` plot, we are interested in the shape of the distribution of each observation, so we are visualising the uncertainty as a signal. In the `ggdibbler` plot, we are not interested in the uncertainty in and of itself, but rather, we only included it to see how it would change the conclusions from the `ggplot2` plot, thus, visualising it as noise. Given these two approaches, which plot is the "correct" visualisation depends on the goals of our analysis and what we are looking to infer from making the plot. However, this freedom to choose disappears if we want to design a visualisation system for EDA. If we want to design a visualisation system for EDA, the method must always work, which means any *existing* plot should always have an "uncertain"

counterpart. Therefore, when designing a system for EDA, we are actually interested in a slightly different question: "Which plot is the correct behaviour of `geom_density` with the distributional input, **X**?". The answer to this question is definitively, the `ggdibbler` plot. The reasoning is rooted in strong statistical foundations; we just need to expand the concepts to visual statistics.

3.3 Visual Statistics

3.3.1 The deterministic visual function

When you really think about it, what is a visualisation? One of the most definitive answers to this question can be found in the grammar of graphics, which is a theoretical framework that characterises a visualisation as a series of composite functions (Wilkinson 2005). Under this formalisation, a visualisation is a function that takes a deterministic matrix (Definition 3.1) as input and outputs a statistical graphic. This formalisation is lengthy, so we summarise the key aspects of it in Definition 3.2, and Figure 3.2, which shows each step of the visualisation process with highlighted components indicating the sections we will need to adjust for uncertainty visualisation. By leveraging the graphics pipeline, we make the implicit explicit, which will allow us to compare different uncertainty visualisation systems on a deeper and more objective level (Wickham et al. 2009). Note that these steps summarise the underlying architecture of the `ggplot2` (Wickham 2010) implementation of the grammar, which is an essential foundation for this new software. However, the conceptual framework described here is not dependent on any particular implementation of the grammar and could easily be implemented in another grammar of graphics system, such as Vega-Lite (Satyanarayan et al. 2016). The notation used for visual statistics builds on that developed in Majumder, Hofmann & Cook (2013) where the term was first introduced in the context of conducting statistical inference.

Definition 3.1 (Deterministic matrix)**.** Let **A** be an $m \times n$ matrix on the sample space Ω. A is a deterministic matrix if $a_{i,j} \forall i, j$ are single-value (non-random) entries.

Definition 3.2 (The deterministic visual function)**.** Let X_{n_1,m_1} be a matrix of outcomes with n_1 rows and m_1 columns on the sample space Ω. Let V be a function that maps X from our sample space Ω to the space of all visual statistics, Ψ. We refer to V as a visual function, and its output $V(.)$ as a visual statistic. The visual function, V, can be decomposed into the following composite function:

$$V = A \circ O \circ G \circ S \circ M$$

where:

- $M = [M_1, M_2, ..., M_{m_1}]'$ is a set of m_1 functions that each scale one column of our X_{n_1,m_1} matrix. This function maps the individual cells of X_{n_1,m_1}, $x_{i,j}$ $\forall i = 1, ..., n_1$ and $j = 1, ..., m_1$ from the sample space, Ω, to the space of real numbers, R.
- S is a statistic function that summarises our X_{n_1,m_1} matrix down to an X_{n_2,m_2} matrix. There are no strict requirements for the statistic function. The function is a transformation from R to R.
- G converts two (for a 2D graph) position columns $X_{[.,k]}, X_{[.,l]}$ to values that represent magnitudes in space returning an X_{n_2,m_2} matrix on the bounded plane B^2. We can further decompose G into $G = E \circ P$, where is the geometry function that maps each $x_{i,j} \forall j = k, l$ from R to the bounded plane B^2, and P is a position modifier function that checks for overlapping values in X_{n_2,m_2} and either stacks them on top of each other in the dependent axis, or dodges them next to each other on the independent axis.
- O is a transformation of our coordinate system. The function is a transformation from B^2 to B^2.
- A is an injective function that transforms our X_{n_2,m_2} matrix into physically observable stimuli. The function maps our data from B to the space of visual statistics Ψ.

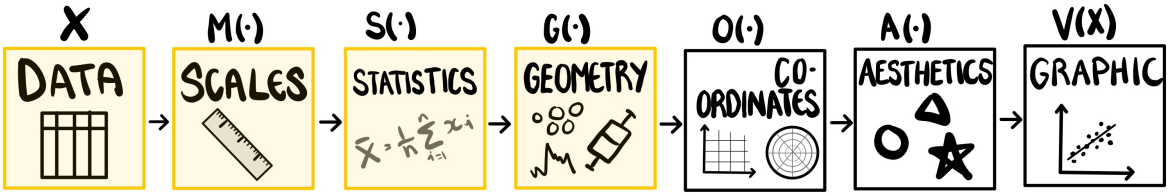


Figure 3.2: *Illustration of the steps to render a plot, as defined by the grammar of graphics. For uncertainty visualisation, changes are needed for the highlighted steps: data, scales, statistics, and the position adjustment within the geometry.*

3.3.2 Random matrices and continuous mapping theorem

Unlike a standard visualisation, an uncertainty visualisation has a random matrix (Definition 3.3) input, where distributions replace the single values of our deterministic matrix (Definition 3.1).

Definition 3.3 (Random matrix)**.** Let **A** be an $m \times n$ matrix-valued random variable on the probability space $(\Omega, \mathscr{F}, Pr)$. This is a matrix where some or all of its entries are random variables drawn from some probability distribution.

Combined, Definition 3.3 and Definition 3.2 boil uncertainty visualisation down to a simple problem. Our visualisation system needs to be designed in such a way that it follows existing mathematical principles of random variables and functions. More specifically, our graphics should uphold the *continuous mapping theorem* (Mann & Wald 1943), Theorem 3.1.

Theorem 3.1 (Continuous mapping theorem)**.** *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be* $n \times m$ *random matrices, and let* Z *be an* $n \times m$ *deterministic matrix. Let* $f : E \longrightarrow E'$ *be a continuous function from one metric space,* E*, to*

another E'. Then:

$$\boldsymbol{X} \longrightarrow \boldsymbol{Y} \Rightarrow f(\boldsymbol{X}) \longrightarrow f(\boldsymbol{Y})$$

and

$$\boldsymbol{X} \longrightarrow Z \Rightarrow f(\boldsymbol{X}) \longrightarrow f(Z)$$

In simple terms, Definition 3.2 and Theorem 3.1 mean that our uncertainty visualisations should have similar convergence properties to their underlying distributions.

If we accept the idea that a visualisation is a continuous function, adhering to Theorem 3.1 is not a nice property or an opinion on how visualisations should behave, but rather a *fundamental requirement* of any visualisation of a random matrix. Or, at least it will be after we establish that the assumptions of Theorem 3.1 are true. This will require us to show that Ω and Ψ are metric spaces, and then use those metric spaces to define convergence for visual functions.

3.3.2.1 Visual metric spaces

The idea that both Ω and Ψ are metric spaces is not particularly strange: one of the core tenets of statistical graphics is that they maintain the link between data and visual aesthetic (Wilkinson 2005). This point was made rather comically by Bartonicek, Urbanek & Murrell (2025), who drew two rectangles stacked on top of each other, and pointed out that it was not a stacked bar chart. The idea that there is some kind of structure, or ordering that we need to maintain, implies that Ω is a metric space. In most cases, $\Omega \subseteq \mathbb{R}^p$ and it immediately follows that the ordered pair (Ω, d) is a metric space, where d is Euclidean distance. In the cases where $\Omega \nsubseteq \mathbb{R}^p$, the first step of the analysis is to scale the data to $\mathbb{R}^p$ using M, so we can say that $(\Omega, d \circ M)$ is a metric space. For our visual space, Ψ, we can actually do exactly the same thing, but in reverse. Since our aesthetic function is defined as $A : B \longrightarrow \Psi$, and $B \subseteq \mathbb{R}^p$, we can set our metric to be the inverse of our aesthetic function, such that the ordered pair $(\Psi, d \circ A^{-1})$ is a metric space.

Ideally, as visualisations are designed to be viewed by humans, we would have set our distance on Ψ to be the human ability to visually differentiate two plots, h. This is similar to the notion of distance that is leveraged by the lineup protocol (Buja et al. 2009). In the lineup protocol, viewers are shown $M-1$ null plots, $V(X_{N_i})$, where X_{N_i} for $i = 1, ..., M-1$ are independent draws generated from some null distribution, and a visual test statistic, $V(X_T)$, where X_T is our actual data. If the test statistic, $V(X_T)$, is significantly different, visually, from $V(X_{N_i}) \forall i$, then viewers will be able to pick it out of the lineup, and we would reject our null hypothesis that X_T was generated from the same distribution as X_{N_i}. This test actually measures Mahalanobis distance on Ψ, as Mahalanobis distance is a multivariate

generalisation of a Z score, and the lineup protocol is the visual equivalent of a hypothesis test. Even though human perception is the natural metric for statistical graphics, (Ψ, h) it is not a metric space. This is because it violates the triangular inequality due to the existence of "just noticeable differences" (JND) (Luce & Edwards 1958). We can show that human perception is not a metric space with a quick proof.

Proposition 3.1. *Let $h : \Psi \longrightarrow \mathbb{R}$ be a piecewise distance function that measures the human perception of statistical graphics, defined as:*

$$h(x,y) = \begin{cases} g(x,y) & g(x,y) \geq \epsilon \\ 0 & g(x,y) < \epsilon \end{cases}$$

where ϵ is the JND of our human observer (with $\epsilon > 0$), and $g(x,y) = d(A^{-1}(x), A^{-1}(y))$ is the Euclidean distance in the rendered graphics; h is not a metric.

Proof. Let $a, b, c \in \Psi$ be three different visualisations, where a and c are plots in the metric space, and b is exactly halfway between them, such that $g(a,c) = \epsilon$, $g(a,b) = \frac{1}{2}\epsilon$, and $g(b,c) = \frac{1}{2}\epsilon$. Therefore, $h(a,c) = \epsilon$, $h(a,b) = 0$, and $h(b,c) = 0$. If we assume h is a metric space, then the triangular inequality will hold, and we can state

$$h(a,c) \leq h(a,b) + h(b,c)$$

which implies

$$0 \leq \epsilon$$

Therefore, by contradiction, h is not a metric. □

As long as our graphics system does not let A map to increments $< \epsilon$, this should not be a problem, but due to differences in human perception, ϵ is not constant for the entire human population, and this system would be impossible to implement. This means that we might not be able to visually distinguish every plot that is different in Ψ, but if we are able to *see* a difference in two plots, it is definitely there.

3.3.2.2 Visual convergence

With the metric space out of the way, we need to define the concept of convergence for visual functions. Thankfully, this is also covered by our definition of a metric space. Two plots have converged if their renderings are identical.

Definition 3.4 (Visual convergence)**.** Let $\mathbf{X}$ and $\mathbf{Y}$, be $n \times m$ random matrices, and let Z be an $n \times m$ deterministic matrix. Let V be a visual function $V\colon \Omega \longrightarrow \Psi$. Let $g\colon \Psi \longrightarrow R$ where $g = d \circ A^{-1}$, and (Ψ, g) is a metric space. We say two random graphics have visually converged, that is, $V(\mathbf{X}) \longrightarrow V(\mathbf{Y})$ when $g(V(\mathbf{X}), V(\mathbf{Y})) = 0$. We say that a random graphic has converged to a deterministic graphic, that is, $V(\mathbf{X}) \longrightarrow V(X)$ when $g(V(\mathbf{X}), V(X)) = 0$.

We will usually approximate this convergence by using visual distinguishability, similar to the approach taken by the lineup protocol.

3.3.3 Returning to the density plot example

In our density example, we defined $\mathbf{X}$ and X such that $\mathbf{x}_i \xrightarrow{p} x_i, \forall i = 1, ..., 15$ as $var(\mathbf{x}_i) \longrightarrow 0$. Therefore, as the variance of our distributions approach zero, $\mathbf{X} \xrightarrow{p} X$ and we should see $V(\mathbf{X}) \xrightarrow{p} V(X)$. That is, as the variance of all the heights in $\mathbf{X}$ approaches zero, our uncertainty visualisation should be visually indistinguishable from the `ggplot2` plot of Figure 3.1. Looking at the plots, we would observe this behaviour in the `ggdibbler` plot, but not the `ggdist` plot. This is why we assert that the `ggdibbler` plot is the random matrix version of the `geom_density` function from `ggplot2`.

3.4 Generalising the visual function

The visual function, as described in the *grammar of graphics*, has one key limitation; it assumes each data point is deterministic. That is, Definition 3.2 in its current form does not allow for random matrix input. This is an issue for uncertainty visualisation, as a random matrix input is one of the core assumptions of the approach. To redefine our visual function such that it accepts random matrix inputs, we will need to adjust the definition of our **scale**, **statistic**, and **geometry** to generalise Definition 3.2. In doing so, we will also ensure we do not make any changes that will result in a violation to Theorem 3.1. When we finish, we should have a generalised visual function that will accept random matrix inputs for any graphic created using the *grammar of graphics*, with the additional property that the plots always obey Theorem 3.1.

3.4.1 The adjustment to `scales`

Scales map our data to a set of real number outcomes; they determine how we perceive the size, shape, and location of our data, and give our data meaning (Wilkinson 2005). This sentiment is echoed in the construction of our visual statistics, as the scale M is used to define the distance metric for our metric space. To use the same scale, M, that was defined for our deterministic matrix X on our random matrix $\mathbf{X}$, we will need to perform a "change of variable". Kay (2023) would refer to this as a "scale aware" requirement for uncertainty visualisation systems. This means we are not changing

the function M itself, but we are just expanding the definition to allow for random matrix input. This formalisation of this scale is summarised in Definition 3.5, and Figure 3.3,

Definition 3.5 (Generalised scale)**.** Let Ω be a sample space and $M\colon\ \Omega \longrightarrow R$ be a scale function that maps our data from Ω to R. Let $\mathbf{X}$ be a random matrix on the probability space $(\Omega, \mathscr{F}, P)$. Then the scale function M, applied to $\mathbf{X}$ will give us $M(\mathbf{X})$ with the induced probability measure $P_{M(X)}(A) = P(M^{-1}(A))$.

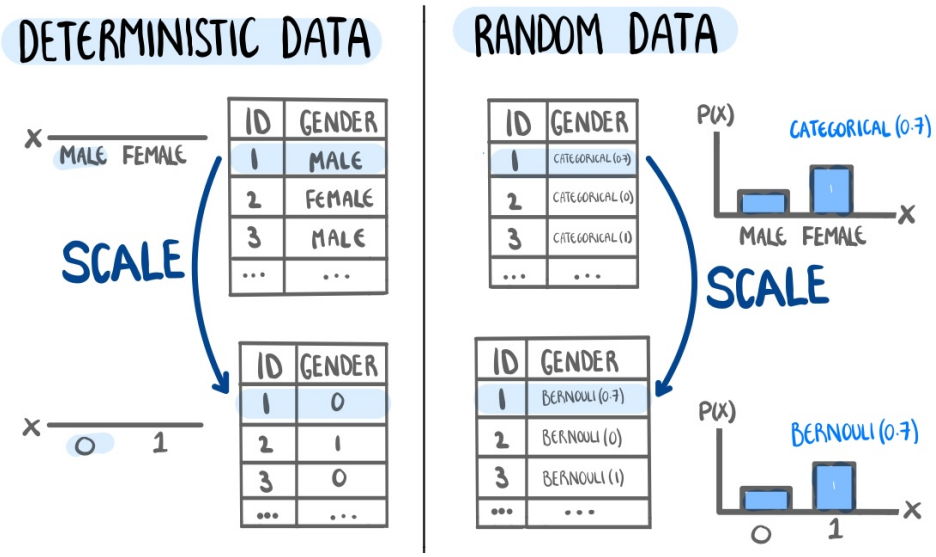


Figure 3.3: *An illustration that highlights the difference between scaling distributions and scaling deterministic variables. Scaling deterministic variables only requires us to map individual values, but scaling a distribution requires us to scale the distribution's domain.*

The illustration in Figure 3.3 shows that our distribution can also changes names (i.e. we scale a binary categorical random variable to {0,1} to create a Bernoulli distribution), but this is simply a function of the scaling, and does not represent any meaningful change in the shape of the distributions.

3.4.2 The adjustment to `statistics`

Statistics provide a summary of our data. Several graphics, such as box plots or bar charts, are inherently linked to a statistic (the five-number summary and summation, respectively). This is the current role of our statistic function, S, which is only well defined for deterministic inputs. Unlike deterministic variables, random variables are more abstract and cannot be expressed as a single concrete value. Therefore, there is a second statistic that needs to be computed, which is the statistic that represents the distribution. The distributional statistic must be calculated at this stage, as the geometry assumes we are working with concrete values that can be mapped to a position in Euclidean space. Therefore, in uncertainty visualisation, there are two statistics that must be defined: the statistic that represents the distribution, and the statistic that summarises the data. The generalised statistic function that accepts random variables is shown in Definition 3.6.

Definition 3.6 (Generalised statistic)**.** Let $\mathbf{X}_{n_1,m_1}$ be a random matrix on the probability space $(R, \mathscr{F}, P)$. Let $S_{sample} : (R, \mathscr{F}, P) \longrightarrow R$ be a function that transforms the random matrix $\mathbf{X}_{n_1,m_1}$ into the deterministic $X_{n_1,m_1,t}$ array, where t is the number of samples drawn from $\mathbf{X}_{n_1,m_1}$. Let $S\colon R \longrightarrow R$ be a function that transforms the deterministic matrix, $X_{n_1,m_1,t}$ into a statistical summary, $X_{n_2,m_2,t}$. Note that S covers all statistics that can be implemented in the deterministic grammar of graphics. We define $S^* : (R, \mathscr{F}, P) \longrightarrow R$ to be the composite function $S = S_{sample} \circ S^*$ that transforms the random matrix $\mathbf{X}_{n_1,m_1}$ into the deterministic matrix, $X_{n_2,m_2,t}$.

The definition presented in Definition 3.6 may leave you with some questions, specifically, why we have limited the distribution representation to a sample of outcomes. This is the question we will spend the rest of this section answering.

3.4.2.1 Representing a distribution

Distributions are abstract concepts, needing some kind of concrete representation to visualise them. This is a common consideration in uncertainty visualisation where we frequently see software that lets us visualise our distribution as a sample of outcomes, a mean and variance, a confidence interval, or even several of these statistics at once (Potter et al. 2010; Kay 2023). Authors often opt for flexibility in the distribution level statistic, with very little consideration as to how this might affect the rest of the plot. It is not unreasonable to assume that the distribution statistic should be interchangeable, as this is how the *rest* of the grammar operates. Even Wilkinson (2005) himself expressed this sentiment in the uncertainty visualisation chapter of The Grammar of Graphics. What this sentiment fails to realise is that the interchangeable nature of the grammar comes from its formalisation, a formalisation that does not properly integrate uncertainty. When we take the time to formalise uncertainty within the visualisation function, we quickly see that the way we represent our distribution does not have a neutral impact on the other components of the grammar.

3.4.2.2 Beyond point estimates

The first problem is obvious: not all distribution representations can show the uncertainty in our random matrix. We briefly illustrate the problem here. Figure 3.4 shows a set of bivariate densities as raster plots visualising the `uncertain_faithfuld` data from `ggdibbler`, which is a random matrix variation of the `faithfuld` data from the `ggplot2` package. Plot (a) visualises only the estimate, while the other three plots represent the data using a sample. As we move from plot (b) to plot (d), the variance in our distribution increases. This variance is independent of our expected value, so the increasing variance has no impact on the expected value of the distribution. Unlike the sample visualisation, the visualisation of our point estimate does not change as the variance increases and is always represented by plot (a), which is equivalent to always conveying a variance of zero. The example shows how `ggdibbler` handles increasing variance in a density plot, as the "graininess" of

the plot increases with the uncertainty. As the variance increases, these grains dominate the plot, making the visualisation harder to read, as it should be.

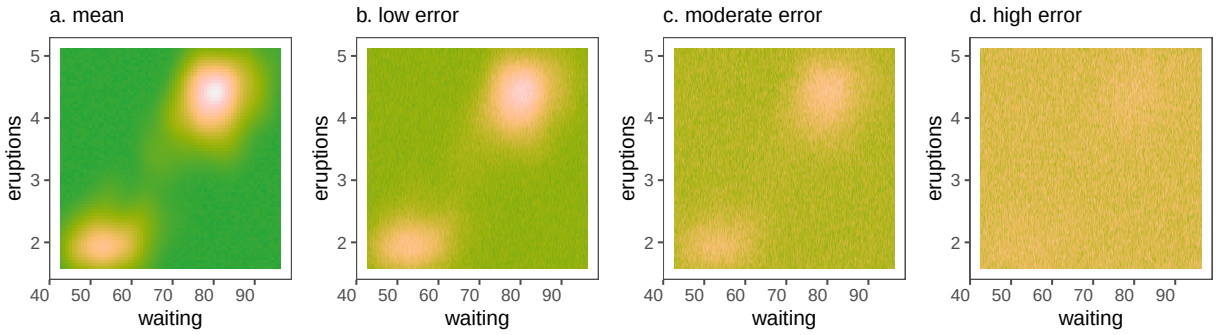


Figure 3.4: *How uncertainty is handled (or not) in raster displays of bivariate density. The axes show the eruption time vs waiting time, and colour indicates density value, with lighter indicating higher density. In plot (a), uncertainty is ignored by showing only the estimate, and plots (b, c, d) show samples reflecting different scales of uncertainty in the density estimate. We can see that as the variance in the estimates increases, the visualisation of the sample becomes harder to read and conveys more uncertainty.*

You may still want to visualise summary statistics alongside our uncertainty representation. This is a common sentiment, and it is the same desire that causes us to visualise a mean and confidence interval on the same plot. Unfortunately, there is a reasonable amount of evidence that visualising summary statistics alongside uncertainty information causes that uncertainty information to be ignored (Padilla, Kay & Hullman 2022). Allowing you to include the point estimates that allow you to ignore the noise, explicitly because the visualisation is too noisy, defeats the entire purpose of the approach. If we are only concerned with convergence to constant values (and not visual convergence to other distributions), then we do not need to include uncertainty at all and Theorem 3.1 would be trivially fulfilled by visualisations of the mean. Therefore, whichever representation we choose, it needs to convey a complete view of this distribution; a point estimate cannot do that.

3.4.2.3 Why not probability functions?

Disallowing point estimates doesn't actually limit our flexibility, as we still have samples, quantiles, and probability functions at our disposal. This is where the "Mr Potato Head" approach to distributional statistics starts to cause problems as we bump up against the orthogonality requirement that is built into the grammar of graphics. Flexibility requires that *every* deterministic graphic should have an uncertain counterpart, even ones that have a pre-defined statistic comprised of point estimates, such as a box plot or bar chart. This is a fundamental requirement, otherwise the system will not be an effective EDA tool. If we allow for any statistic, we will create a mismatch where the values we are trying to feed into our statistic, S, are not on the same domain as expected by the function. To be more explicit, our statistic is expecting values on the domain, $M(\Omega)$. If we define a new statistic, $S^* = S \circ S_{dist}$, where the range of S_{dist} is not $M(\Omega)$, such as $P_{M(\Omega)}(M(\Omega))$, then we have produced

invalid input for the next stage of our visual function, S. For example, if our statistic is expecting heights that range from 150 to 200, we cannot feed in a set of probabilities on [0,1] and expect there to be no issues. The statistics that create a domain mismatch also tend to create an implicit inference problem, as changing the statistic used to represent the random variables can also change the role of uncertainty in our analysis (Mason et al. 2026a). This means that violating this rule will not only result in unusable inputs or nonsensical outputs for S, it will also fundamentally change our visual function, V. This change means our visualisation will not adhere to Theorem 3.1, which is the primary requirement of our system. For these reasons, we can only allow statistical representations that output a range that is equivalent to the input space.

3.4.2.4 Why not quantiles?

This leaves quantiles and samples as our remaining distribution statistics. It makes sense that these methods would be the most flexible, as a sample is just outcomes on $M(\Omega)$, and quantiles are just ordered samples. However, the notion of "ordering", which is required for quantile representations, produces two problems for a flexible visualisation system. The first problem is that quantiles communicate an explicit ordering on Ω. While the data at this stage is technically on the real line $M(\Omega)$, the quantiles will not have meaning if Ω is unordered. Using quantiles to visualise uncertain categorical data will either result in meaningless graphics or an inability to visualise the data at all. This limitation would prevent us from visualising uncertain categorical data, which is a common output of classification models. The second problem with quantiles is that they don't have a natural extension to multivariate data. Quantiles are well-defined for univariate cases, but multivariate spaces require several assumptions on the relative magnitude of our variables, which are unlikely to always be correct. Figure 3.5 visualises the four scenarios that arise from passing a univariate or multivariate random variable, represented as a quantile or a sample, to the slope or intercept of a `geom_abline`. This is not an unreasonable scenario, as a linear regression with a random intercept and slope is a common topic even in introductory statistics courses. We can see that in the univariate case, where the intercept of the line is an $N(0,1)$ distribution, the information conveyed by the quantile (a) and the sample (b) is very similar. This is because quantiles are well defined in the univariate case. However, for the multivariate case, adding a second random variable in the slope (such that we are now visualising a multivariate normal distribution with marginal distributions $N(0,1)$, and a covariance of -0.8) throws our notion of ordering out the window. In this case, quantiles are not well defined, as any quantile, q, will have an infinite number of (intercept, slope) pairs that could produce that probability, and are better conveyed by a function (hence why contour plots are typically used). As we cannot visualise the slope of a line as a function, a sensible alternative might be to use the marginal or equicoordinate quantiles (Bornkamp 2018). We opted to use the marginal quantiles,

which are also used to colour the lines in both representations, but the conclusions are the same if we use the equicoordinate approach instead. This allows us to see the danger of using quantiles as our distribution representation. The first issue is that the notion of ordering we have imposed on the quantiles does not translate in the multivariate case, which can be seen in the haphazard colouring of the lines in the sample, which was not a problem in the univariate case. The implicit pairing of values has also changed the point of intersection of the lines, and the neatness of the quantiles conveys more certainty in our conclusions than is warranted by the actual data. Even if we tried to work around these problems by coming up with some abstract definition of a visual quantile, we wouldn't be able to draw the output with a straight line, which is the only real requirement for `geom_abline`.

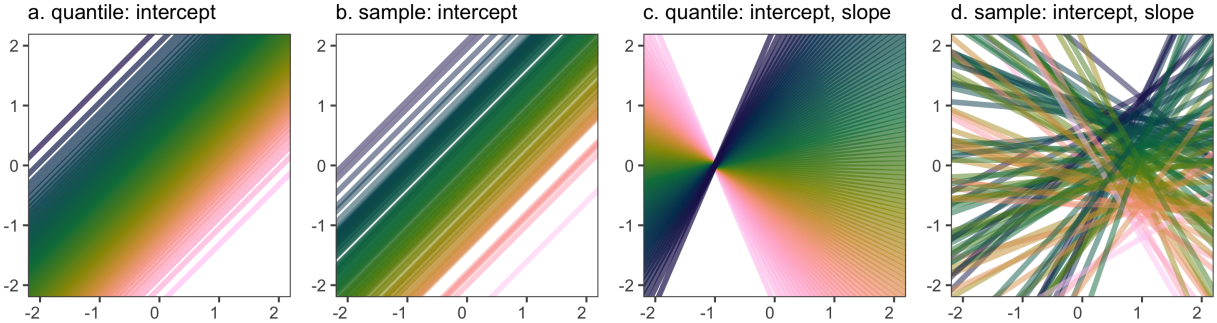


Figure 3.5: *Why quantiles are problematic for representing our distribution variables, using regression coefficients. Intercepts and slopes were simulated using marginal distributions of N(0,1) and a covariance of -0.8. Plots (a) and (b) have only the intercept treated as random, and show the quantiles and samples, respectively. Colour maps to quantile in (a) and to the value of the intercept in (b): both plots convey similar information. It breaks down when both slope and intercept are treated as random, shown as quantiles (c) and samples (d). Colour is mapped to the same notion of distance that is used to construct the quantiles. But distance is not well defined, and we can see the approaches diverge in the erratic colouring of the lines. Ordering beyond one variable makes quantiles inflexible representations of distributions.*

3.4.2.5 Distributions as samples

This means that the only representation of a distribution that is equally as flexible as a point prediction is a sample of outcomes. Of course, we don't want a sample of individual points; we want a sample of geometric objects. To get this, we need to pass the data through the visual function in batches, where each batch represents an outcome of the full random matrix. In practice, this translates to "splitting" on the `drawID` in the Grammar of Graphics (Wilkinson 2005), which involves modifying the `group` variable in `ggplot2` (Wickham 2010) to include the `drawID`. Figure 3.6 demonstrates the effect of grouping in the implementation. Without appropriately handling the grouping in `geom_smooth`, the plot has only one fitted curve with a standard error artificially small due to a larger number of observations, which is wrong. Once the group variable is modified to include the `drawID`, the result is a fitted line for each sample, and also the choice to include a standard error ribbon faintly in the background for each sample.

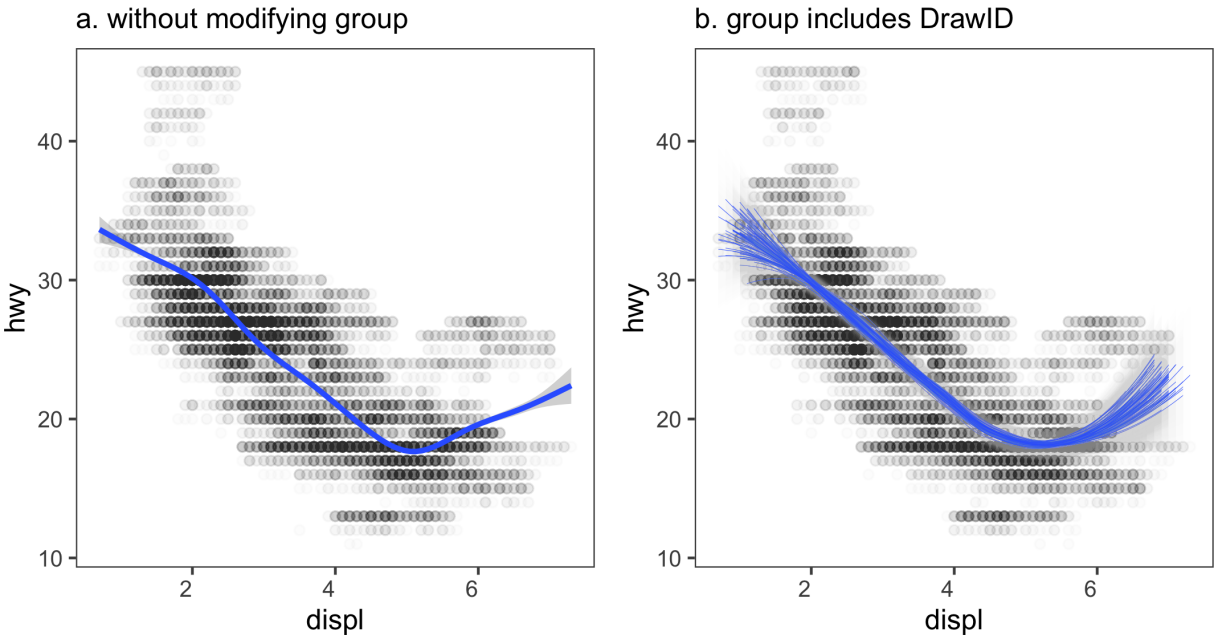


Figure 3.6: *Modifying the group variable is essential for handling samples: (a) not done, giving an incorrect representation of the uncertainty, (b) group variable includes the `drawID`. We can see that we need to pass our samples through the visual function in batches to ensure that the statistics are not artificially changed by the sample size.*

The requirement for samples and *only* samples as our distribution representation is why the formalisation by Kay (2023), despite having the insight to use distributional inputs, did not have the full flexibility required for EDA. By allowing flexible distribution representations, `ggdist` is focused on looking at distribution as values in their own right, rather than integrating uncertainty into existing visualisation systems. This is also how the visual functions of `ggdist` and `ggdibbler` in Figure 3.1 diverge from one another. It is important to understand that neither approach is a subset of the other; they are orthogonal, and most of the plots made in `ggdist` cannot be made using `ggdibbler`. While there are instances where `ggdist` and `ggdibbler` produce similar-looking plots, these plots cannot be made using the same data or the same code. The distinction between the two approaches translates directly from the philosophy of Mason et al. (2026a), who pointed out that the difference between the role of signal and noise is in our inferential statistics. By having the distribution statistic subsume the statistic of the plot, we are changing our inferential statistic and visualising uncertainty as a signal. This is why we repeatedly say that `ggdist` is for looking at uncertainty as signal, and `ggdibbler` is for looking at uncertainty as noise. The visualisations of `ggdist` cannot be made by `ggdibbler` due to the limitations in the distribution statistic, and the visualisations of `ggdibbler` cannot be made by `ggdist` due to the limitations in the `ggplot2` level statistic. Ultimately, both packages are required if we want a complete picture of uncertain data.

3.4.3 The adjustment to `geometry`

The geometry component of the grammar translates our data to a magnitude in space (Wilkinson 2005). By displaying each distribution as a sample, we have distilled uncertainty visualisation down to a simple over-plotting problem; where we previously had a matrix, $X_{n,m}$, we now have an array, $X_{n,m,t}$. Therefore, to pass our data through the following stages of the grammar, we need to flatten our array back into a matrix in such a way that we ensure each outcome from our random matrix is equally weighted. Over-plotting is usually managed by the geometry component of our visual function by using position adjustments such as dodging to prevent overlap or transparency to make the overlapping visible (Cook, Lee & Majumder 2016; Wickham 2010; Wilkinson 2005). Position adjustments are an umbrella term used to describe any small changes to the position of a geometric object, and are usually implemented to ensure different groups are equally visible. Unlike statistics, we can perform position adjustments one after the other, which means position adjustment can be nested within each other. Therefore, we can make the position adjustment required for the overplotting created by the samples *within* any position adjustment that already existed in the plot. This is the final change we will make to our visual function to allow the visualisation of random variables, and it is formalised in Definition 3.7.

Definition 3.7 (Generalised geometry)**.** Let G^*: $R \longrightarrow B$ be a geometry function that transforms an array of data into a matrix of geometric positions. We can further decompose G^* into $G = E \circ P^*$, where $E : R \longrightarrow B$ is the geometry function, and $P^* : B \longrightarrow B$ is the position adjustment. Let $C_{n,m,t}$, $B_{n,m,t}$, $A_{n\times t,m}$ be matrices of geometric positions. We can further decompose P^* into $P^* = P_{within} \circ P_{between}$, where

- P_{within} is a within sample position adjustment that transforms $C_{n,m,t}$ to $B_{n,m,t}$, by identifying overlapping points on $C_{n,m,i}$ $\forall i = 1, ..., t$ and applying the specified sample position function, and.

- $P_{between}$ is a between sample position adjustment that transforms $B_{n,m,t}$ to $A_{n\times t,m}$ by flattening $B_{n,m,t}$ into $B_{n\times t,m}$, identifying overlapping points on $B_{n\times t,m}$ and applying the specified position adjustment.

We will spend the rest of this section detailing the nested position system.

3.4.4 Nested position adjustments

Position adjustments change the location or size of a geometric object to prevent overlap and ensure all geometric objects remain visible. They achieve this by placing objects beside one another (dodging), stacking them on top of each other (stacking), placing them in front of each other and making them see

through (transparency), or showing them one after each other in quick succession (animations). These options make up the four dimensions that we have available for position adjustments: x (dodge), y (stack), z (transparency), and time (animation). Including transparency and time as position adjustments is not the standard approach in the literature, as these plots are typically considered separate axes. This is not unique to uncertainty visualisation, even Wilkinson (2005) did not discuss positions relative to the axis of x-y-z-t, but rather specified position adjustments as being on the measured scale (stack) or in the spare space (dodge). This is an important distinction, but we choose to frame the positions in terms of x-y-z-t to highlight that there may be multiple measured or spare axes in a single plot.

Unlike the statistics, (most) position adjustments do not meaningfully change the inferential statistic of our plot, so we can nest them freely without concern. This is particularly useful because without nested position adjustments, we would need to apply the same position adjustments to the over-plotting caused by both the original grouping and the sampling. Since we cannot use stacking on the measured axis for our samples, this would prohibit us from making uncertain versions of stacked bar charts, which, again, would be an undesirable limitation to our uncertainty visualisation system. We can see this problem in Figure 3.7, which shows the visualisation of a stacked bar chart of the `mpg` data (a), alongside a visualisation of its random counterpart, the `uncertain_mpg` data, visualised using a "stack" (b), "stack_dodge" (c), and "stack_identity" (d), position adjustment. The fact that stacking is not a viable approach should be obvious: the scale has been artificially inflated, and the visualisation provides little to no information about our data. In the case of a bar chart, stacking is aligned with our measurement axis (y), which leads to it being a problematic adjustment, as our between-sample position adjustment, $P_{between}$, can only be implemented on an axis representing spare space. In other words, stacking is only a viable position adjustment when the sum of the stacked groups holds meaning (Wilkinson 2005), which is not true for an arbitrary number of samples. Interestingly, this split of appropriate versus inappropriate position adjustments is opposite to the findings of Bartonicek, Urbanek & Murrell (2025), who found that stacking was the only appropriate axis to use for interactivity, for similar arbitrary scale change issues. This suggests the possibility of an underlying orthogonal relationship between uncertainty and interactivity in statistical graphics that would allow us to implement both interactivity and uncertainty visualisation simultaneously.

Using this framework, we find that a lot of plots with distinct names can actually be described by a single plot with different position adjustments. For example, if we have a map with the fill of each area represented by a random variable, then we could capture this uncertainty using a HOPs (Hullman, Resnick & Adar 2015), a pixel map (Lucchesi & Kuhnert 2020), or a value-suppressing uncertainty palette (Correll, Moritz & Heer 2018). These plots could all be made in `ggdibbler` using

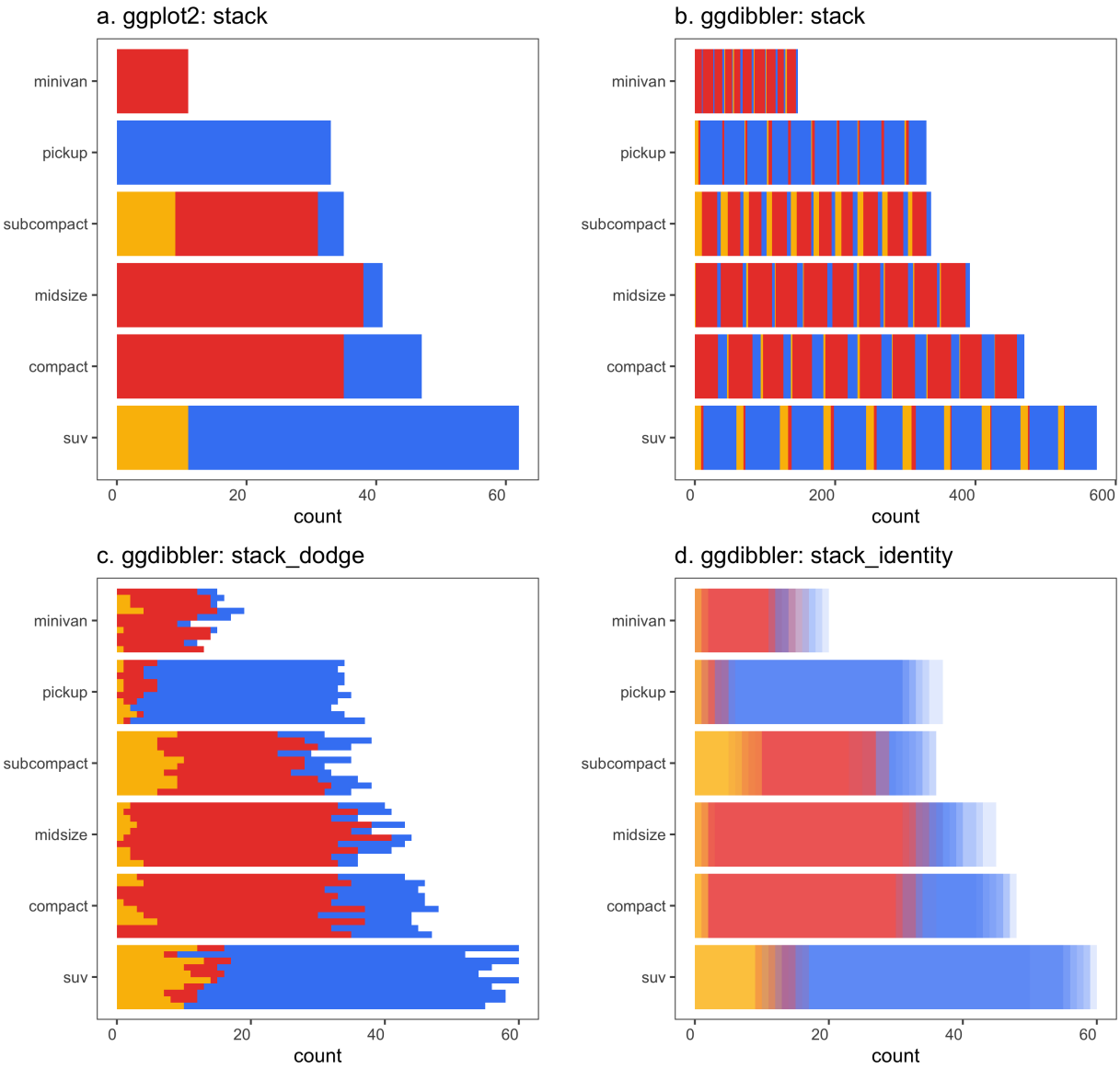


Figure 3.7: *Why we need nested position adjustments illustrated using stacked bar charts made using different position adjustments. Plot (a) shows what a deterministic plot looks like for reference, while plots (b), (c), and (d) use the same visual function, but have a random variable input. We can see that stacking is not viable as plot (b) is unreadable and does not maintain continuity, while dodging (c) and transparency (d) work well. It is clear that we should not use the measurement axis for our samples' position adjustment.*

a `geom_sf_sample` and an animation, subdivide (a simultaneous dodge and stack), or transparency position adjustment, respectively. If we consider a facet to be a "between" plot position adjustment, in contrast to the "within" plot position adjustments we get with dodging and transparency, we can extend this idea further. Under this framework, the uncertainty visualisations that map a null distribution with an alternative on the same plot (Guo et al. 2025; Hullman & Gelman 2021; Savvides et al. 2019; McNutt, Kindlmann & Correll 2020) are just the lineup protocol (Buja et al. 2009) without the "between plot" position adjustment.

The most appropriate position adjustment to use can depend on which aesthetic the random variable is mapped to, and impacts our ability to read the plot. Figure 3.8 illustrates the change in plot appearance when a random variable is mapped to text using a transparency (a) and jitter (b), or mapped to colour using a dodge (c) and transparency (d). We can see that transparency works quite well for text, while position adjustments such as jitter make the overlapping text harder to read, regardless of the uncertainty in the estimate. We can see that colour works well with dodged positions, as it allows us to see the full sample and do the final calculation visually. Managing colour with transparency will still produce a technically correct plot, but it can lead to colours that cannot be matched to the legend, as high variance colours mix and create new colours that do not belong to the palette. Differences in the most appropriate position adjustment can cause conflict when there are multiple sources of uncertainty in a plot. It would be interesting to investigate this further with a perceptual experiment to test the effectiveness of different position adjustments for different aesthetics, but that is well beyond the scope of this paper.

3.4.5 The generalised visual function

If we combine the definitions from Definition 3.2, Definition 3.5, Definition 3.6, and Definition 3.7 we have a generalised visual function that accepts random variable inputs. In this generalised visual function, the deterministic graphics made by Definition 3.2 are simply a special case of Definition 3.8, where every cell is a degenerate distribution.

Definition 3.8 (The generalised visual function)**.** Let $\mathbf{X}$ be a random matrix on the probability space $(\Omega, \mathscr{F}, Pr)$. Let V be a function that maps $\mathbf{X}$ from $(\Omega, \mathscr{F}, Pr)$ to the space of all visual statistics, Ψ. The visual function, V, can be decomposed into the following composite function:

$$V = A \circ O \circ G^* \circ S^* \circ M$$

3.5 Implementation in `ggdibbler`

The visual function given by Definition 3.8 should allow you to make an uncertainty visualisation version of any graphic that can be described with the grammar of graphics. We have implemented this theory in the R package, `ggdibbler`, which is a `ggplot2` extension that allows users to create an uncertain version of any `ggplot2` graphic. Figure 3.9 illustrates this flexibility by showing a collection of plots that were all made using this generalised visual function. We can see that there is flexibility in both plot type, as we display line plots, maps, pie charts, histograms, bubble charts, and network diagrams, and in aesthetics, as position, colour, size, slope, and other aesthetics all have

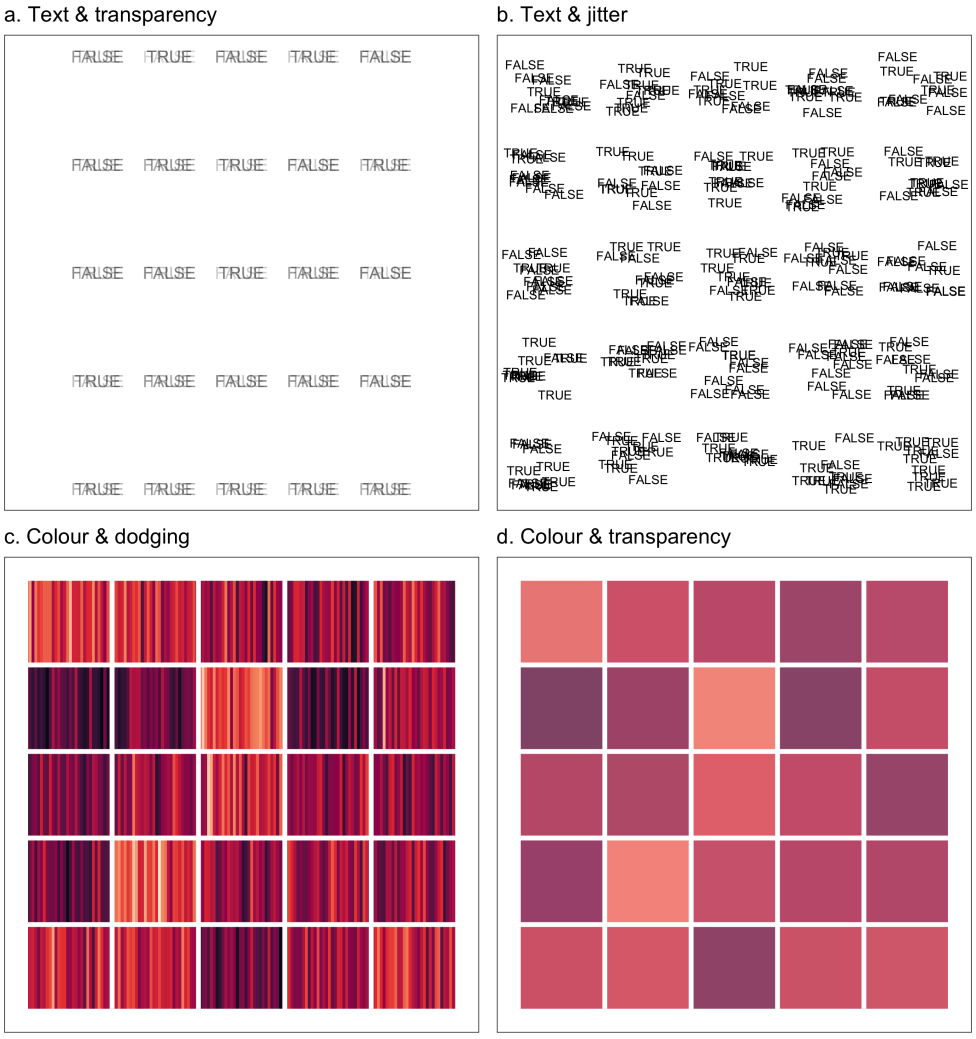


Figure 3.8: *Illustration of the change in plot appearance based on aesthetic mapping and position adjustment. Plots (a, b) map the random variable to text with transparency and jitter, respectively, and plots (c, d) map to tile colour using dodging and transparency. Although this needs experimental evidence, mapping the samples to transparency improves readability for text, but for colour, dodging produces better readability than transparency.*

a random variable mapped to them. A single plot can even have multiple sources of uncertainty simultaneously mapped to different aesthetics. By establishing a set of rules that will almost always work, we save ourselves from having to design bespoke software for every single individual case.

While we have established the conceptual theory that would underpin a flexible uncertainty visualisation system, there are considerations that need to be made when implementing the theory as practical software. Specifically, we should discuss the design of the user interface, the data objects that allow us to work with random matrices, and the computational complexity that comes with uncertain plots. This section will detail these components.

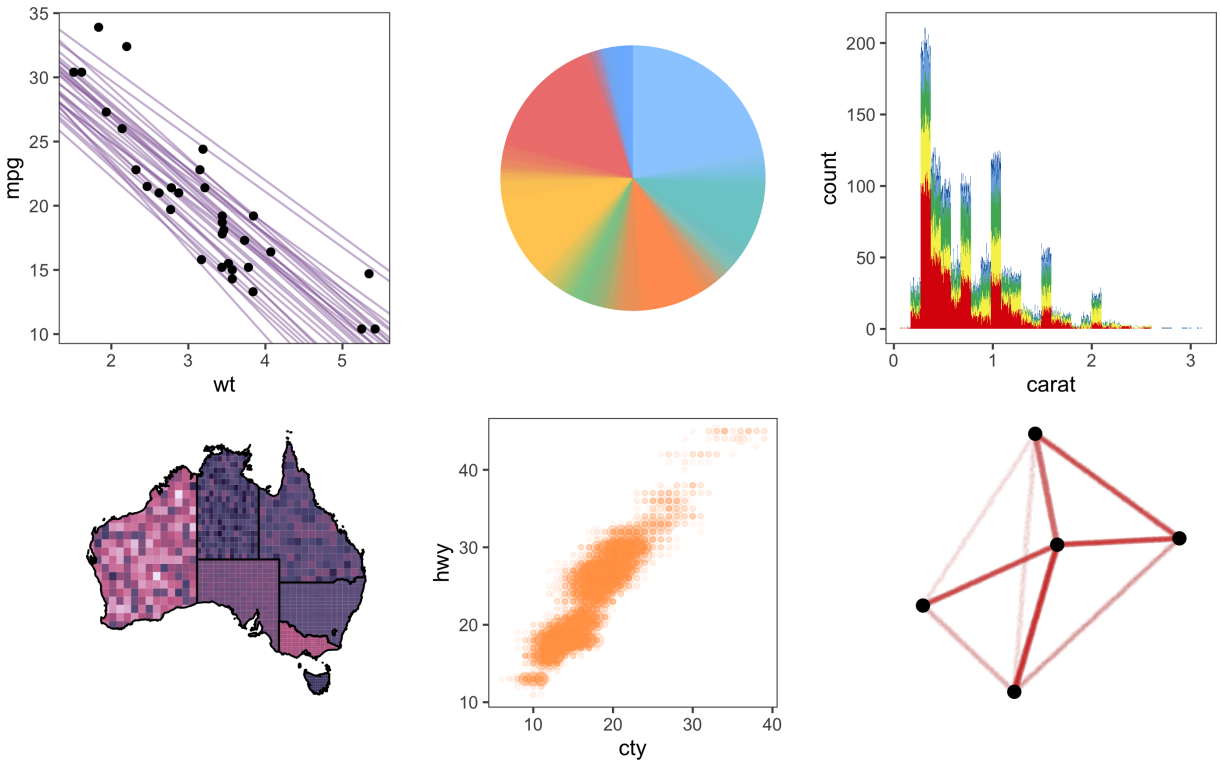


Figure 3.9: *This formalisation of uncertainty visualisation offers extensive flexibility, illustrated by six plots: line, pie chart, histogram, map, bubble chart, and network diagram. These plots are made with almost identical syntax with* `ggdibbler` *as that of the deterministic* `ggplot2` *equivalent. These aesthetics - position, colour, size, slope - are all mapped using random variables.*

3.5.1 The software design

The way that a user interacts with a piece of software can help communicate the mechanisms or theory that underpins it. The statistical theory behind random matrices, visual convergence, and the continuous mapping theorem that underpins the `ggdibbler` package might be too complicated for someone trying to make a simple scatter plot, but a basic understanding of these ideas is required to use the package correctly. Therefore, when designing the functions, we opted to subtly communicate these ideas through coding paradigms and function names.

Readers familiar with programming paradigms might look at Definition 3.2 and Theorem 3.1 and immediately think of object-oriented programming (OOP). These readers would be right, `ggplot2` can be considered to be an object-oriented system with a functional-feeling interface. Actually, the bulk of functionality underlying R structures can be considered to be object-oriented.

In OOP systems, data is stored as objects and methods that operate on these objects. The user interacts with these objects only through the methods, not by directly inspecting the elements. Objects can inherit, so special objects have features that will work in a variety of settings, and also some additional special features. Different objects can respond to the same method call in different ways

(polymorphisms).

Today's R contains several choices in data management: S3, S4, R6, and the latest, S7. S3 forms the original framework, and an example of the polymorphism is the `print()` function, where what is printed will change depending on the object provided. For example, a `data.frame` will be printed differently from an `lm` (linear model object). It lacks the full characteristics of OOP, though, because there are no formal class definitions, and it is easy to misuse. S4 is stricter and underlies all of the Bioconductor (Gentleman, Carey, Bates, et al. 2004), but more cumbersome for the user. S7 is designed to have the ease-of-use of S3 but the strict checks of S4. Of course, this is not an exhaustive list, and there are several other choices floating around for specific use-cases, including ggproto and R6.

The main reason to use OOP is polymorphism, which allows us to apply the same function to different classes of input (Wickham 2019). For example, this paradigm would allow us to create a `draw` function that accepts both `polygons` and `lines`, drawing either one without any special input from the user. The `sf` package's `geom_sf` function allows users to indiscriminately pass points, lines, and polygons to the function, which responds exactly as the user expects, plotting geoms appropriate to the spatial class passed in. OOP is the logical approach for uncertainty visualisation, as we want to convey that V is the same function regardless of the input type. From the perspective of the user, the function should be the same whether we input $\mathbf{X}$ or X.

Not only is OOP theoretically consistent with our goals, but it is also the most practical approach. Using OOP allows us to hide unnecessary complexity, which is particularly helpful for uncertainty visualisation, as the abstract nature of uncertainty causes it to be quite error-prone. Visualising uncertainty without the guardrails put up by `ggdibbler` can lead to cases where uncertainty is visualised as a separate variable, and the graphic does not have the statistical properties guaranteed by the package. This approach will also make it easier to perform EDA, as the most flexible programming systems make no assumptions about our data; instead, they react to the object input by the user to determine the initial mapping (Wilkinson 2005). This principle is already carried through in `ggplot2`, where plots automatically adapt to categorical, continuous, discrete, or date-time inputs by having an OOP scale, M. While it would be nice to implement `ggdibbler` as a scale, an uncertainty visualisation system requires more intensive changes to V.

The implicit relationship between `ggdibbler` and `ggplot2` is communicated through the syntax of the code. Ideally, if all of `ggplot2` were built on an OOP system, we could just create an "uncertainty" version of all the `ggplot2` functions, allowing users to pass distributions without even noticing the change in the underlying package. That is, both the `ggplot2` and `ggdibbler` plots from

Figure 3.1 would use the syntax, `ggplot(data = density_data) + geom_density(aes(x = x))` where the visualisation software runs the `ggdibler` code if `x` is a `distributional` object. As `ggplot2` is not built on an OOP system, this is not possible, so instead, `ggdibbler` adds a `*_sample` suffix in the function name. This allows us to maintain similar naming conventions to the related `ggplot2` function, while also being explicit about what the function does. For example, the code that makes the `ggplot2` density is `ggplot(data = density_data) + geom_density(aes(x = xmean))`, while the code that makes the `ggdibbler` density is `ggplot(data = density_data) + geom_density_sample(aes(x=xdist))`. This syntax still conveys the idea that the visual function is identical; it is only the input that has changed.

This strong theoretical foundation not only gives us an intuitive function design, but it also allows us to have a lot of versatility built on a shockingly simple code base. Despite the package covering the full range of geoms in `ggplot2`, the bulk of `ggdibbler` is a single function that does the sample and group adjustment, and a second function that nests the positions (if needed). Every `geom_*` has a `geom_*_sample` counterpart that calls these core functions, and then passes the data through the standard `geom_*` pipeline, with little to no bespoke adjustments for individual geoms. A significant amount of the simplicity of the design comes from the packages dependencies: `distributional` (O'Hara-Wild et al. 2024), `ggplot2` (Wickham 2010), `dplyr` (Wickham et al. 2023), `rlang` (Henry & Wickham 2026b), `lifecycle` (Henry & Wickham 2026a), `scales` (Wickham, Pedersen & Seidel 2025), `tidyr` (Wickham, Vaughan & Girlich 2025), `tibble` (Müller & Wickham 2025), `cli` (Csárdi 2025), and `sf` (Pebesma 2018).

3.5.2 Representing uncertainty using `distributional`

The theoretical framework we have presented assumes you already have a random matrix input, and thus far, we have somewhat ignored how you would go about making one. This is because quantifying uncertainty and representing it as a data object is fundamentally part of the data manipulation stage of our analysis, so it should be kept as separate as possible from the visualisation stage (Wickham 2010). This gives users full transparency in *what* precisely they are visualising (Wickham 2010). In `ggdibbler`, this separation is created by leveraging the `distributional` package (O'Hara-Wild et al. 2024), which allows users to store distributions inside data frames as individual cells. `ggdibbler` works from the assumption that users have already quantified the uncertainty as distributions before attempting to visualise it, building a system that allows for `distributional` inputs.

The `ggdibbler` package is designed to accept any `distributional` input. There are two ways you can define a distribution in `distribtional`:

1. A theoretical distribution defined by the distribution and its parameters, or

2. An empirical distribution defined by a set of samples

Most types of uncertainty can be represented as one of the two cases, so the software is surprisingly flexible. The purpose of the distribution is to accommodate classical statistics, or even Bayesian thinking, where the distribution is known or specified. The reason for being able to specify the distribution through samples is to accommodate the common situation today that the distribution may not be known, but we can describe it with samples. Such a situation might arise when we have bootstrap samples describing the variability.

The `ggdibbler` software implements a full uncertain replication of `ggplot2`, including the examples. This requires `ggdibbler` to have uncertain versions of all the data sets used in the`ggplot2` examples, made in distributional, including (but not limited to) `uncertain_mpg`, `uncertain_faithful`, and `uncertain_economics`. These examples make use of both the theoretical and empirical distribution types. This allows users to immediately start using `ggdibbler` with familiar data and plots, making the system less daunting to integrate into an existing data analysis pipeline.

3.5.3 Additional computational complexity

Replacing a single plot with a large sample of plots can significantly increase the computational complexity of the visualisation. The more random variables we feed into the plot, the bigger the sample size needs to be, but the bigger the sample size gets, the more computationally expensive the plots become to make and render. These are simply manifestations of the classic statistics trade-off between computational cost and accuracy, and the curse of dimensionality.

In `ggdibbler`, this sample size is controlled by the `times` argument, which is set to `10` by default. Sometimes this is appropriate, sometimes it is not; it depends on the variance of the distributions, the particular plot type, the number of random variables fed in, etc. We cannot set a reliable and sensible `times` argument that works for every plot in every situation, so instead we advise you to pick a sample size that allows your plot to visually converge. This is a different type of convergence from the convergence to a deterministic `ggplot2` discussed earlier. Technically, a `ggdibbler` visualisation is a random variable, so every time you print one, it will draw a new random sample and look slightly different from the previous renderings. If your sample size is big enough, the variability between each visualisation should be small enough that your conclusions do not change between renderings. We could take this a little further and suggest that your sample size should be large enough that different renderings are visually indistinguishable from one another. In other words, they have visually converged by Definition 3.4. To be more specific, let $V(\mathbf{X})_i$ be the i^{th} rendering of a `ggdibbler` plot, then we would say that the plot has converged (and our sample size is large enough) when $d(V(\mathbf{X})_i, V(\mathbf{X})_j) = 0 \forall i, j$. This process of visual convergence is shown in Figure 3.10. We can see that

the full shape of the distribution becomes more visible as the `times` argument increases (the input is a scatter plot of uniform distributions).

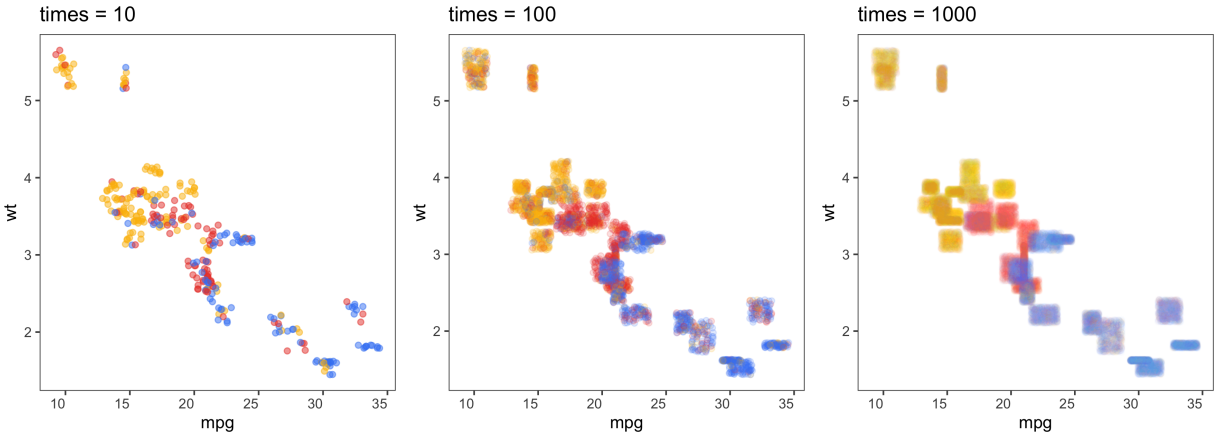


Figure 3.10: *A scatter plot of a random matrix version of the `mtcars` data from the 'datasets' R package, with the aesthetic mapping `x=mpg`, `y=wt` and `colour=cyl`. All three variables are random variables. This plot shows the impact of an appropriately chosen times argument. We can see that as the sample size increases, the distributions form cohesive units and stop looking like a collection of separate points with little connection. This is not always achievable due to computational cost, but we should, at the very least, select a sample size that means our conclusions are not changing between renderings of the plot.*

Given this additional complexity, we might opt to skip the entire sampling procedure and instead just map the "uncertainty" to some kind of aesthetic. After all, despite only using samples to depict uncertainty, the `ggdibbler` documentation (along with this paper) is littered with plots that appear to be using the explicit mappings that are commonly associated with uncertainty. Blur, fuzziness, colour lightness, and shape/size frequently appear despite never being controlled through the aesthetics. Therefore, it might be more computationally effective to just control these elements directly.

While this idea is not unreasonable and might be possible to implement in the future, we do not take this approach in `ggdibbler`, as directly mapping uncertainty seems to be fraught with danger. The source of the uncertainty and its visual expression are tightly linked in a `ggdibbler` plot. Take, for example, the appearance of blurring and fuzziness shows in Figure 3.11, which depicts `ggdibbler` fuzziness in a bubble chart, `ggdibbler` blur in a dotplot, and `ggdist` blur in a dotplot. Both blur and fuzziness use a transparent position adjustment, but blur is created through uncertainty in position, while fuzziness is created through uncertainty in size. This allows us to convey multiple types of uncertainty at once, as blur makes it more difficult to read the position but does not affect our ability to read the size, and vice versa for fuzziness. The inability to convey multiple types of uncertainty is the most common difficulty faced by uncertainty visualisation approaches (Peña-Araya et al. 2025; Hadjimichael, Schlumberger & Haasnoot 2024; MacEachren et al. 2005). The blur and fuzziness in the `ggdibbler` plots is more of an "emergent" aesthetic, which appears through the combination

of transparency and randomness; this is in direct contrast to the `ggdist` blur, which is controlled top-down. This type of plot is rather unusual for the `ggdist` package, as the uncertainty is controlled by a standard error, rather than a distributional input, and that error is mapped to a "blur" aesthetic. Looking closely at the plot, we can see a blurred cliff effect, where the blurring of some dots extends over the dots beneath them. Since the dots in a dotplot must be stacked on top of each other, this type of blurring would not be possible. This makes it unclear as to how the blur should be interpreted, and it indicates some kind of breakdown in the relationship between the data and its visual representation.

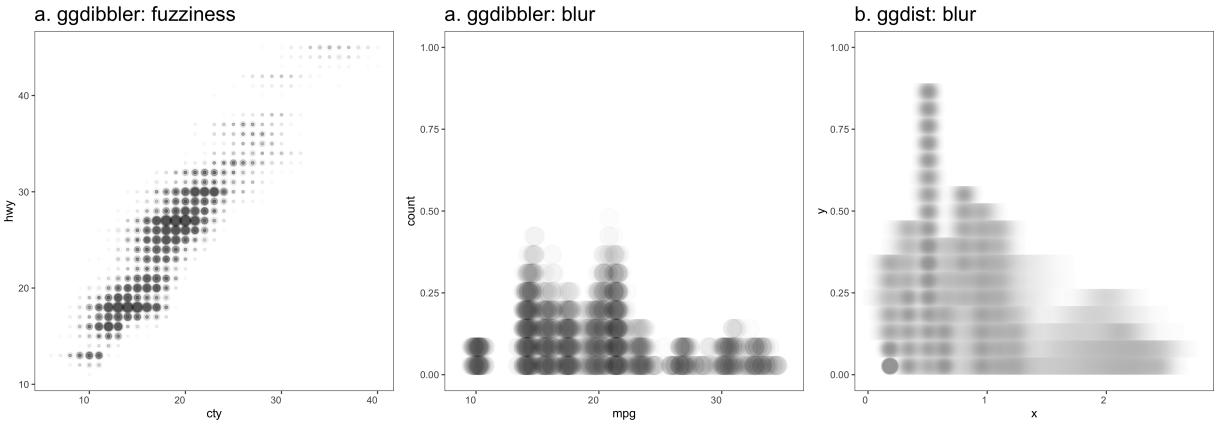


Figure 3.11: *Two examples from the `ggdibbler` documentation, and one example from the `ggdist` documentation to illustrate the difference in the top-down versus emergent aesthetic approach. The blur and fuzziness emerge from the `ggdibbler` plots due to the sampling procedures, while the blur in `ggdist` is added manually as a top-down aesthetic. We can see that the 'cliff' effect in the `ggdist` plot is not visible in the blurred `ggdibbler` plot, because it would be impossible to generate that appearance from the underlying data.*

Breaking the connection between the data and its visual representation would result in us losing the desirable statistical properties that are guaranteed by `ggdibbler`. This breakdown appears to be quite common when we try to manually map uncertainty to an aesthetic (Mason et al. 2026a). Additionally, a flexible EDA system should allow *any* combination of *any* uncertain aesthetics, and working out how these combinations of random variables should appear would be incredibly laborious, if not impossible. Additionally, this approach would almost defeat the purpose of the system, as the whole point of visualising the data to begin with is that we *don't already know* what it should look like. For these reasons, trying to directly map uncertainty to an aesthetic might be better computationally, but would likely result in more headaches than it would be worth, if it were possible at all.

3.6 Conclusions and future research

This paper set out to formalise uncertainty visualisation in order to design a flexible uncertainty visualisation system that maintains nice statistical properties. The value of this formalisation is not

only in the power of a truly flexible system, but also in the mathematical foundation itself that ensures the connection between data and visual aesthetic is always maintained.

By defining the visual function mathematically, we leveraged the concept of continuity to evaluate the behaviour of visual statistics. This approach uncovers a wealth of other statistical concepts we could translate to statistical graphics. Building on this work, we should investigate other concepts like bias/variance trade-off, statistical sufficiency, and other convergence properties for visual statistics. In this vein, we should also identify the minimum level of variance in a plot that is required for the human perception of visual convergence, related to the concept of "just noticeable differences" in psychophysics. Similarly, it would be useful to know if there are principles for determining the sample size required for visual convergence, as it would give us a convenient ceiling on the computational complexity of the plots.

On the topic of convergent graphics, it would be interesting to find out if all plots with the same limiting visual statistic have some underlying computation in common. This curiosity is quite similar to the question posed by Wickham (2010), where they asked if plots that are visually identical, that are made using different grammar adjustments, have some underlying principle in common. When looking into interactive visualisations Bartonicek, Urbanek & Murrell (2025) also found underlying links between the mathematical functions we plot and the appropriate visual representations, so we wonder if there is a more "fundamental" version of the grammar of graphics that can make this link more explicit. In mathematics, all functions can essentially be broken down into a basic increment/step function (something akin to stacking in graphics), so we do wonder if a similar principle can be applied to plots.

The motivation for a more fundamental building block of statistical graphics can also be found in our brief discussion on emergent aesthetics. The emergent aesthetics blur the line between the statistic and geometric stages of plot building, begging the question "Should we explicitly map a statistic, or allow it to emerge through the visualisation process?" To solve this problem, we would need to start by untangling which aesthetics are primary aesthetics and which are emergent aesthetics. We would also need to understand the conditions under which each of these aesthetics arises and the position adjustments that lead to them.

Position adjustments have historically been an afterthought when building graphics, but this work motivates a more thorough investigation into how they affect the readability of a plot. We discussed how text/shape works well with alpha, and dodging/subdividing is good for colour, but this is just an anecdotal observation. A more formal theory on how position adjustments change our ability to read a plot is much needed.

Finally, the `ggdibbler` software has room for improvement. Currently, the software only accepts individual distributions and, if multiple distributions are passed, they are assumed to be independent. The `distributional` software allows for joint distributions, which are functionally a random matrix collapsed into a vector. Expanding `ggdibbler` to accept these joint distributions is a natural extension for the software. Additionally, `ggdibbler` does not work with `ggplot2` extensions due to the way the software is implemented. We plan to extend the package to make this possible. A full list of the planned changes is available, along with the package source code in the `ggdibbler` GitHub.

3.7 Acknowledgements

The first author of this paper is supported in part by a scholarship from the the Australian Energy Market Operator. This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship [DOI: https://doi.org/10.82133/C42F-K220]. The first author would also like to thank Mitchell O'Hara-Wild and Cynthia Huang for their comments and feedback which substantially improved the work, as well as Ze-Yu Zhong for several interesting examples that ended up being used in this paper. The R packages used for this work were: `tidyverse` (Wickham et al. 2019), `distributional` (O'Hara-Wild et al. 2024), `ggdist` (Kay 2023), `ggdibbler` (Mason et al. 2026b), `patchwork` (Pedersen 2025b), `khroma` (Frerebeau 2025), `tidygraph` (Pedersen 2024), `colourspace` (Stauffer et al. 2009), `ggraph` (Pedersen 2025a), `ozmaps` (Sumner 2021), `sf` (Pebesma 2018), and `ggthemes` (Arnold 2024). The GitHub repository for this paper can be found at https://github.com/harriet-mason/paper-ggdibbler, which contains the files required to reproduce this article in full.

Chapter 4

Colour Blinded by the Noise

4.1 Introduction

Uncertainty is routinely present and often ignored in data visualisation. Anything other than "raw" data (and it is debated whether or not "raw" data exists at all (Bokulich & Parker 2021)) will involve some sort of modelling, and therefore, uncertainty (Otsuka 2023). Because this uncertainty can change the conclusions we draw from our data, it is important that it is incorporated into our visualisations in a way that is intuitive and easy to understand.

There are many practical reasons to visualise uncertainty. Visualisations that do not include uncertainty can be misleading. Authors argue that effective uncertainty visualisations should soften unjustified conclusions (Wilkinson 2005), prevent the identification of false discoveries (Sarma et al. 2024; Koonchanok et al. 2023), or improve decision making (Padilla, Kay & Hullman 2022). Uncertainty visualizations may also be more transparent. Hullman (2020) likens failing to visualise uncertainty to fraud or lying, while Zhao et al. (2023) found that including uncertainty can improve trust in our models or analysis. These notions of trust, clarity, and transparency imply that uncertainty visualisation should act as a sort of visual hypothesis test, where statistically valid signals are visible, while statistically spurious signals are not. Under this framework, the best uncertainty visualisations will simultaneously minimise the chance of seeing something that isn't there (type I error) as well as the chance of missing something that is there (type II error) (MacEachren 1992), in a way that does not rely on any statistical expertise from the viewer (Correll & Gleicher 2014). For this to be true, it is not enough to simply include uncertainty in our visualisation; it needs to be incorporated in such a way that false signals become invisible without damaging the visibility of true signals. This property, where statistical validity translates to visibility of signal in a plot is called "signal suppression" (Mason et al. 2026a). This is what it means to visualise uncertainty in a way that it can be **seen**.

Interestingly, despite the wealth of evaluation studies in uncertainty visualisation (Hullman et al. 2019) and the fact that the connection between signal visibility and statistical validity originates with the field itself (MacEachren 1992), this simple hypothesis has not been tested. Uncertainty visualisation studies have recorded accuracy, decision-making quality, confidence, trust, risk aversion, cognitive load, and intuitiveness of encoding, to name a few (Hullman et al. 2019). However, there is little evidence of studies comparing statistical validity to signal visibility. While there are a handful of studies that compare participant responses to the results from statistical tests (Correll & Gleicher 2014; Kale et al. 2018), they often frame questions in a way that is entangled with human judgment (e.g., "Do you think there is a trend in this line?" or "Who will win this election?"), which introduces noise into the results (Hullman 2016). However, this is not the only source of noise in the field. Kinkeldey, MacEachren & Schiewe (2014) attributed a lot of the conflict in the research to an engineering approach to visualisation, which approaches the issue from a usability perspective, rather than asking why some representations do or do not work. This noise makes it difficult to synthesise the results of evaluation studies into a cohesive framework or set of recommendations (Kinkeldey, MacEachren & Schiewe 2014; Bostrom, Anselin & Farris 2008). Synthesising results is so difficult, in fact, that MacEachren et al. (2005) wondered whether we should be visualising uncertainty at all, or if it would be better to just suppress it. Addressing this problem will require boiling the evaluation of uncertainty visualisation down to a simple question: "Are statistically insignificant signals visible in this visualisation, and if not, why?" Answering the "Why?" component of this question requires a systematic comparison of graphics.

Systematically changing plots becomes far easier when our scope is narrowed to a specific scenario or plot type. Maps have been one of the key focal points for uncertainty visualisation, in part because they offer a particularly challenging case study, as many familiar statistical visualisation tools become unavailable when data is referenced on a map (Waller 2024). Cartography also has a tradition of attention to data quality and a strong desire for visualisations that ensure the accuracy and reliability of their conclusions (MacEachren 1992). Geoscience also offers a classic case study in the importance of conveying uncertainty through the communication of climate change or extreme weather events. We see these scenarios frequently in evaluations of uncertainty visualisations as participants are asked to make decisions about sea level projections (Benjamin & Budescu 2018), flood uncertainties (Lim, Brandt & Seipel 2016), wildfire hazards (Cheong et al. 2016), and hurricane forecasts (Padilla, Ruginski & Creem-Regehr 2017). While narrowing our scope down to mapping makes sense given the context of the field, it is still not simplified enough for our purposes. A single map may have multiple measurements, with multiple sources of uncertainty (MacEachren et al. 2005; Kinkeldey, MacEachren & Schiewe 2014), and trying to test all these sources of uncertainty at once will overcomplicate our

experiment. Effective evaluations require us to isolate and test marginal units. Trying to test too many sources of uncertainty at once will be as effective as testing none at all. For this reason, we choose to focus on a single commonly-used aesthetic within a map: colour.

Choropleth maps are visualisations that use colour to represent a statistic aggregated over a region (such as a county, state, or country), with their first uses dating back over 200 years (Dupin 1826). This aggregation is one of the key sources of uncertainty in maps (Xiao 2021). We can see an example of this in the American Community Survey (ACS), an annual collection of socio-economic data that is then aggregated over spatial regions, resulting in a large margin of error, which is not always considered in associated visualisations or analyses (Jung, Thill & Issel 2019). These aggregation problems go beyond cartography alone. On-Farm Precision Experiments (OFPE) will often aggregate measurements on yield, nitrogen rate, or seed levels (Bullock et al. 2019; Kyveryga 2019) to reduce clutter, removing the uncertainty inherent to the calculations.

While our study will focus on choropleth maps, our experimental hypotheses are built upon foundations in perception and colour theory. We present an experimental approach for evaluating uncertainty as noise, in the hopes that it can provide a foundation for evaluation studies that will go beyond our modest choropleth map.

4.2 Background

4.2.1 Lineups and uncertainty visualisation

Visualisations that seek to align the visibility of a pattern with classical hypothesis tests are nothing new, and have a well-established foundation in graphical inference with the lineup protocol (Buja et al. 2009; Wickham et al. 2010). The connection between the two approaches was originally identified over a decade ago by Hullman, Resnick & Adar (2015). Like uncertainty visualisation, the lineup protocol was developed in pursuit of the goal of ensuring the inferences we draw from our visualisations are statistically valid, and it is seen as a visual parallel to hypothesis testing. In the lineup protocol, as described in Buja et al. (2009), viewers are shown M plots: $M-1$ generated from some null hypothesis, and 1 generated from the real data, called the target plot. Viewers are then asked to identify which plot is the "most different", where the definition of "most different" is left to the subjective discretion of the viewer. If viewers can consistently identify the target plot from the lineup, then the null hypothesis can be rejected, meaning that the data used in the target plot is different from the data used to generate the other $M-1$ plots.

While both uncertainty visualisations and the lineup protocol provide a mechanism to perform visual

hypothesis testing, their sources of uncertainty are different. In a lineup, the source of uncertainty is shown through the null distribution (or future inference), whereas the uncertainty in an uncertainty visualisation is shown as a feature of the data itself (typically as distributional inputs (Kay 2023; Mason et al. 2026b)). This distinction is the main difference between the two approaches.

The different sources of uncertainty between lineups and uncertainty visualisation have a run-on effect in how conclusions are drawn from each approach to inference. One of these run-on effects is in how the methods show uncertainty, as lineups show outcomes across multiple plots, while uncertainty visualisations show the outcomes within a single plot. While some visualisations disobey this distinction and display the distribution of a null plot alongside the data (Guo et al. 2025; Savvides et al. 2019), where the true data is coloured differently from the null hypothesis, this approach is somewhat antithetical to the goals of both uncertainty visualisation and the lineup protocol. Uncertainty visualisation and lineup protocols are built on the idea that statistical significance should imply visual distinction. If a target plot is significantly different, it should be visually distinguishable from the null distribution; if a pattern is statistically significant in an uncertainty visualisation, it should be visible to viewers. By colouring the target data differently, the visual differentiability is built into the plot design and therefore cannot be tested, as would be done in a lineup scenario. We can think of uncertainty visualisation as the visual parallel to a confidence interval, while the lineup protocol is the visual parallel to hypothesis testing. Both can be connected to a statistical hypothesis.

4.2.2 Implicit testing

One of the key benefits of a lineup protocol is its ability to measure a pattern's visibility without participants explicitly needing to interpret, understand, or make judgments about the pattern. This is because lineups leverage implicit testing (Vanderplas, Cook & Hofmann 2020). Implicit tests, where participants must infer the question from the provided stimuli, exist in direct contrast to explicit tests, which ask users to extract a specific statistic (Vanderplas, Cook & Hofmann 2020). Existing uncertainty visualisation research takes the explicit approach to evaluation: authors ask participants specific questions or set goals that the participants must use the visualisation to fulfil. These studies conflate a participant's ability to see a statistical signal with outside influences such as misunderstanding of statistical concepts, error from cognitive overload, and interference from participants' prior beliefs or utility functions (Bostrom, Anselin & Farris 2008; Hullman 2016; Kim et al. 2019). Shifting our evaluation from interpretation to simple signal visibility with pre-attentive processing will also reduce the cognitive load required to read the plot (Vanderplas, Cook & Hofmann 2020), which is a common issue with uncertainty visualisations due to their additional complexity (Brennen & Tuerk 2018).

While explicit testing is perfectly fine for plots that have been constructed to showcase a *specific*

structure in the data (Vanderplas, Cook & Hofmann 2020), it severely handicaps our ability to evaluate a visualisation for exploratory data analysis (EDA). Data visualisations, just like other stimuli, suffer from inattentional blindness, so explicitly extracting a statistic can disconnect us from the exploratory process (Boger, Most & Franconeri 2021). This is particularly a problem for uncertainty visualisation in geoscience as several authors have expressed a desire for methods that are suitable for EDA (Hadjimichael, Schlumberger & Haasnoot 2024; MacEachren et al. 2005). To evaluate a plot's usefulness as a tool for EDA, we must evaluate our ability to see a signal, before we even know what that signal might be. Evaluating plots for these purposes will require a more implicit approach relative to existing experiments.

Ideally, we would be able to directly translate the lineup protocol to evaluate different types of uncertainty visualisations. However, because uncertainty visualisations and lineup protocols are two different methods of visual inference, this will not be possible. The lineup protocol can still have a visible pattern in the data when noise itself generates an interesting pattern (see the LDA lineup in Roy Chowdhury et al. (2015)). This means that a rejection, or failure to reject, in the lineup protocol does not align with signal visibility, making it an unsuitable approach for evaluating the specific goals of uncertainty visualisation.

It is unlikely that we will be able to design a fully implicit test for uncertainty visualisation, as the assumptions of the approach require the viewer to bring a null hypothesis with them in the form of an explicit question. However, we can still use some of the lineup protocol's design principles as guidance when designing our implicit tests. One of the unusual design elements of a lineup protocol is that all the context is removed, as the point is to "see" the patterns in the plot, unhindered by prior beliefs (Cook, Reid & Tanaka 2021). Therefore, if we can boil our question down to such an intuitive level of psychophysical stimuli that we don't even need scales to interpret, we can leverage some of the benefits that come with the lineup protocol. As a matter of fact, if we do not need the context of a statistical graphic, we can look outside standard visualisation evaluation approaches to come up with an effective method.

4.2.3 The Ishihara colour blind test

At its core, the goal of this uncertainty visualisation experiment is to measure the effect of noise on the visibility of signal. To put it another way, we are trying to measure the effect of a latent variable by measuring its impact on a primary signal variable. When we restrict this to the case of a choropleth map, we are trying to measure the conditions under which the latent variable collapses one colour channel (signal) down into another colour channel (noise). When the latent variable we are trying to measure is a colour vision deficiency, this is known as a colour blind test.

The connection between statistical maps and colour blind tests is not particularly unusual. According to a 1930 review of colour blind tests, methods such as sorting or matching coloured objects, naming coloured lights, and distinguishing objects presented in complementary colours have been used as tests for red-green distinction (Haupt 1930), tasks that would not be out of place in a visual evaluation study. In particular, we are interested in pseudo-isochromatic colour blind tests, the most popular of which involve identifying patterns on coloured cards that are visible to those with standard colour vision, but invisible to those with colour vision deficiencies (Haupt 1930). The Ishihara test is a specific version of a pseudo-isochromatic test that is commonly used today (Gobira et al. 2025; Plutino et al. 2023), where the coloured dots form numbers or paths on a white background (Tamura et al. 2017; Plutino et al. 2023). Notably, these tests use familiar shapes (in the form of numbers), which provide readily identifiable spatial clusters without the need for complex definitions. Additionally, it allows us to perform one of the most under-researched uncertainty visualisation tasks, aggregations of uncertain information over an area (Kinkeldey, MacEachren & Schiewe 2014). The widespread use of this test also lends an air of familiarity to test-takers, and the simple numerical response reduces the amount of time needed for each individual trial, allowing for a single individual to view many plots.

There are some considerations we need to make when translating colour blind tests to statistical graphics, many of which are common to the evaluation of colour-dependent visualisations. Perceptions of physical Ishihara test panels are sensitive to the colour of lighting/wavelengths used (Tamura et al. 2017), and viewing colours electronically is not always consistent, and may depend on gamma values (Ihaka 2003). However, evidence seems to show that colour blind tests maintain accuracy when implemented electronically (Gobira et al. 2025; Khizer et al. 2022) and have displayed consistency across Android and Apple devices (Khizer et al. 2022). Therefore, even though there are differences in the perception of colour in a digital colour blind test, they do not seem to have a significant impact on the results of these studies and may be unlikely to impact a crowd-sourced experiment.

Combined, this work suggests that a “noise” blind test, which uses variance rather than colour vision deficiency, to make a signal invisible, would be an effective way to evaluate our choropleth maps.

4.2.4 Choropleth maps and the grammar of graphics

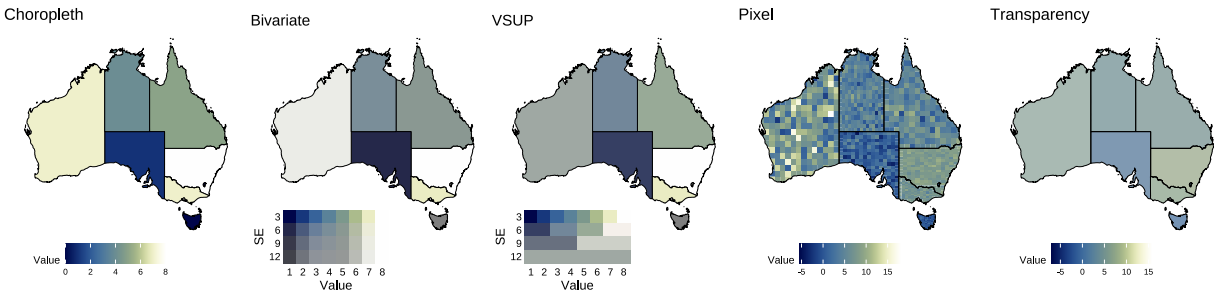
Evaluating our graphics using a systematic approach will require a theory of the “difference” in plots. The seminal work of Wilkinson (2005), *the grammar of graphics*, provides a framework for us to understand the information conveyed by our graphics. The grammar’s influence can be seen in its implementation, as the theory is the foundational structure for widely used visualisation software such as `ggplot2` (Wickham 2010) and Vega Lite (Satyanarayan et al. 2016). The grammar describes

the transformation of visualisations as they move through several layers, which are, in order: data, scales, statistics, geometry, co-ordinates, and aesthetic mappings. Using the grammar, elements of a visualisation can be incrementally changed, keeping the underlying data constant, and the effects of these changes can be isolated to particular components of the grammar (Vanderplas, Cook & Hofmann 2020). This approach is usually utilised in conjunction with the lineup protocol, and the combined pair has been used to compare polar and cartesian coordinates (Hofmann et al. 2012), geometric distribution representations (Hofmann et al. 2012), and colour palettes (Reda & Szafir 2021a). Until recently, it would have been difficult (if not impossible) to implement this approach for uncertainty visualisation, as uncertainty visualisations were not properly described by the grammar of graphics (Wilkinson 2005). While we would have no issues representing the choropleth map in the grammar, the uncertain variations would be impossible to describe. However, thanks to the recent work by Kay (2023) and Mason et al. (2026b), the uncertainty variations used in this study can be fully explained within the grammar.

There are a few restrictions on the type of choropleth map that will be evaluated in this experiment. Our first restriction is that we will not look at maps that primarily change the underlying statistic we are visualising, such as exceedance probability maps (Smemoe 2004) or Bayesian surprise maps (Correll & Heer 2016). We set this restriction because changing the underlying statistic changes the fundamental meaning of our plot, rather than integrating uncertainty into an existing visualisation (Mason et al. 2026a). Additionally, largely due to computational costs, our approach is restricted to static plots, ruling out visualisations such as HOPs plots (Hullman, Resnick & Adar 2015). Finally, to ensure we are comparing variations of the same choropleth map, we will only compare maps that are identical at their deterministic limit (Mason et al. 2026b). That is, all the maps we compare should create an identical choropleth map when the variance is zero.

Given these restrictions, this study will consider five map types: the choropleth, bivariate, Value-Suppressing Uncertainty Palettes (VSUP), pixel, and transparency maps, shown in Figure 4.1, alongside the mapping of their (relevant) components in the grammar of graphics. All visualisations will have a distribution input for each area, and the components that are not included in the table are constant for all visualisations. These map types were chosen not only for their ubiquity, but also due to what their success or failure can teach us about uncertainty visualisation. Our goal is to identify the design choices that lead to effective signal suppression, rather than pick out a "best" plot.

The first map in the graphic is the classic choropleth map, which doesn't include any uncertainty at all, and will serve as a baseline in the experiment to compare other methods against. Both Kay (2023) and Mason et al. (2026b) establish that a statistic that represents the distribution is required



	Name	Distributional Statistic	Geometric Position	Coordinate (colour)	Aesthetic (colour)
	Choropleth	Mean	Identity	Cartesian	Hue
	Bivariate	Mean & Variance	Identity	Cartesian	Hue & Saturation
	VSUP	Mean & Variance	Identity	Warped Cartesian	Hue & Saturation
	Pixel	Sample	Subdivide	Cartesian	Hue
	Transparency	Sample	Transparency	Cartesian	Hue

Figure 4.1: *The five map designs we will be evaluating, illustrated using the state boundaries of Australia. The distributions visualised in the maps are the same data, but randomly generated. Along with an example with each map, we also have the breakdown of how each map is created, a grammar of graphics breakdown of its important components. We can use this table to understand how each map diverges from one another (as well as the standard choropleth map) to understand how our findings translate to generalisable findings about uncertainty visualisation.*

for visualising uncertainty, and for cases that ignore uncertainty, such as the standard choropleth map, we typically represent a distribution by its mean. The choropleth map's grammar of graphics mapping is made using the `ggplot2` pseudocode `ggplot(map_data) + geom_sf(aes(fill = mean(value), geometry = geometry))`.

Bivariate colour maps were first used by the US Census Bureau in the 1970s (Meyer, Broome & Jr. 1975), and they are one of the older variations of the choropleth map we will investigate. This approach diverges from the choropleth map by extracting the variance of the distribution alongside the mean, and mapping that variance to colour saturation, creating a 2D colour palette. Mason et al. (2026a) hypothesised that the bivariate map would be insufficient for signal suppression, as mapping an estimate and its uncertainty on the parallel channels of hue and saturation, even if the channels are visually integrable and perceived as a single unit (Vanderplas, Cook & Hofmann 2020), will not introduce enough interference to hide statistically invalid signals.

An alternative approach to the bivariate map is the Value Suppressing Uncertainty Palette (VSUP) proposed by Correll, Moritz & Heer (2018). Fundamentally, this visualisation is a bivariate map

with a warping coordinate transformation applied to the colour space, as discussed by Wilkinson (2005). In the VSUP coordinate space warping, nearby colours are blended together to reduce their discriminability as the variance in the estimate increases. Unfortunately, this approach has significant dependence on the scaling of the variables and the method taken to blend the colours (Kay 2019). While this method is almost certain to result in *some* interference, there is no reason to believe that separately extracting a mean and variance, only to recombine them using a coordinate transformation, would result in visual interference that is statistically sound. Uncertainty visualisation is fundamentally a question about statistical validity, which suggests the interference should be managed at the statistic level of the grammar.

Rather than change the palette, another alternative to the choropleth map is the pixel map (Blenkinsop et al. 2000; Lucchesi & Wikle 2017). This map is identical to the choropleth map in palette, but represents the distribution as a sample of n draws, instead of as a mean and variance. This approach implements an implicit visualisation of uncertainty (Correll & Gieicher 2015) and relies on human perception to extract estimates such as mean or variance. To manage overplotting and ensure each draw from the sample is evenly weighted, the area is subdivided into a grid of outcomes, creating the pixelated appearance that gives the plot its name. If this plot is more effective than the VSUP or bivariate, it would suggest that effective signal suppression requires the visualisation of a full distribution. This would provide a perceptual argument for the formalisation of uncertainty as a distribution, as suggested by Kay (2023). While we will be using a sample, theoretically, any statistic that gives a full picture of a distribution (such as a set of quantiles, the PMF, or the PDF) should have similar results, but investigating these alternatives is outside the scope of this paper.

Finally, we will test the transparency map, which is identical to the pixel map, but does not perform the subdivision. Instead, to maintain equal weighting of every draw, the transparency is set to $\frac{1}{n}$, meaning we represent each area using a single colour, but that colour will not necessarily be able to be matched to a palette. This is not a problem for the artificial environment of our colour blind tests, but it would be for standard statistical graphics. As we draw meaning from our visual signals using the legend of our plot, having values that are impossible to identify on the scale will leave us unable to interpret our visual signal. Unlike the other maps, this plot is not a standard alternative to the choropleth map, but we have included it in our evaluation as it provides some insight into the design requirements for uncertainty visualisation. Both the pixel and transparency maps are made using the `ggdibbler` (Mason et al. 2026b) R package, with the only distinction between them being the position adjustment. The pixel map diverges from the choropleth map in two ways: through the statistic it visualises, as well as the number of values it shows. If the pixel map results in successful signal suppression, the transparency map will allow us to isolate which of those design changes

caused it.

All the maps only utilise colour to convey information with no additional glyphs or aesthetics, so there is minimal chance for interference from other visual channels. However, in designing the experiment, we found that the palette choice was not neutral in its influence on the readability of the plot. For this reason, we considered using a rainbow colour palette, as Reda & Szafir (2021a) found that palettes possessing more uniquely identifiable colours (such as rainbow palettes) provided higher discernment between visual models than more monotone palettes. However, rainbow palettes have a disadvantage in that those with colour vision deficiency may struggle to distinguish between shades, and the colours used are not uniformly perceived (Crameri, Shephard & Heron 2020; Vanderplas, Cook & Hofmann 2020). High chroma colours are also disadvantageous in their production of after-image effects (Ihaka 2003), which would be a particular issue in our fast moving colour blind test. Ultimately, we settled on using the continuous `davos` palette by Crameri (2018) as it is part of a collection of perceptually uniform, ordered, and colour blind friendly palettes designed for use by the scientific community. This means that it likely reflects a good example of an actual palette used by a visualisation author.

4.3 Experimental Design

4.3.1 Hypothesis

Our experiment has three main hypotheses, listed below.

H1: Mapping standard deviation to a second channel in our visualisation is insufficient to create signal suppression. If this is true, the probability of reading the signal in the bivariate and choropleth maps will be similar and independent of the standard deviation in our estimates.

H2: For interference to be statistically valid, it depends on the visualisation of the correct statistical information, rather than post-hoc adjustments in the later stages of the grammar. If this is true, the interference in the VSUP map will not be proportional to statistical significance.

H3: The visualisation of a statistic that conveys the full size of the distribution, even if it is visualised as a single value, is the primary requirement for signal suppression. If this is true, the probability of reading the signal in the pixel and transparency maps will be similar.

4.3.2 Stimuli

We generated our pseudo-isochromatic plates using functions from the `ishihara` package (Tierney 2020). Each "plate" of the test was made up of about 1000 circles with approximately 20% belonging to the number group, N, and the remaining 80% belonging to the standard background colour, B. The colour of each circle, P_i, is represented by a continuous value, C_i. The standard assumption in

uncertainty visualisation is that every variable is represented by a distribution, with its own central value and standard deviation. We replicated this scenario using a Bayesian hierarchical model with $C_i \sim N(\bar{n}_i, V^2) \; \forall \; P_i \in N$, or $C_i \sim N(\bar{b}_i, V^2) \; \forall \; P_i \in B$, where $\bar{n}_i$ is an outcome from $\bar{N} \sim N(0.5 * D, 1)$ and $\bar{b}_i$ is an outcome from $\bar{B} \sim N(-0.5 * D, 1)$. This allows us to replicate the appearance of a standard Ishihara test, even without variance in the individual estimates, as the colours are mottled by the distribution of the means. To keep the experiment simple and reduce the confounding signal from the variance, we kept the standard deviation constant for all C_i within a plate. The relative distance between these two groups, D, as well as the standard deviation within each observation, V, were factors of interest when generating the data, as both affect the visibility of the numbers. This resulted in three factors, each with five levels:

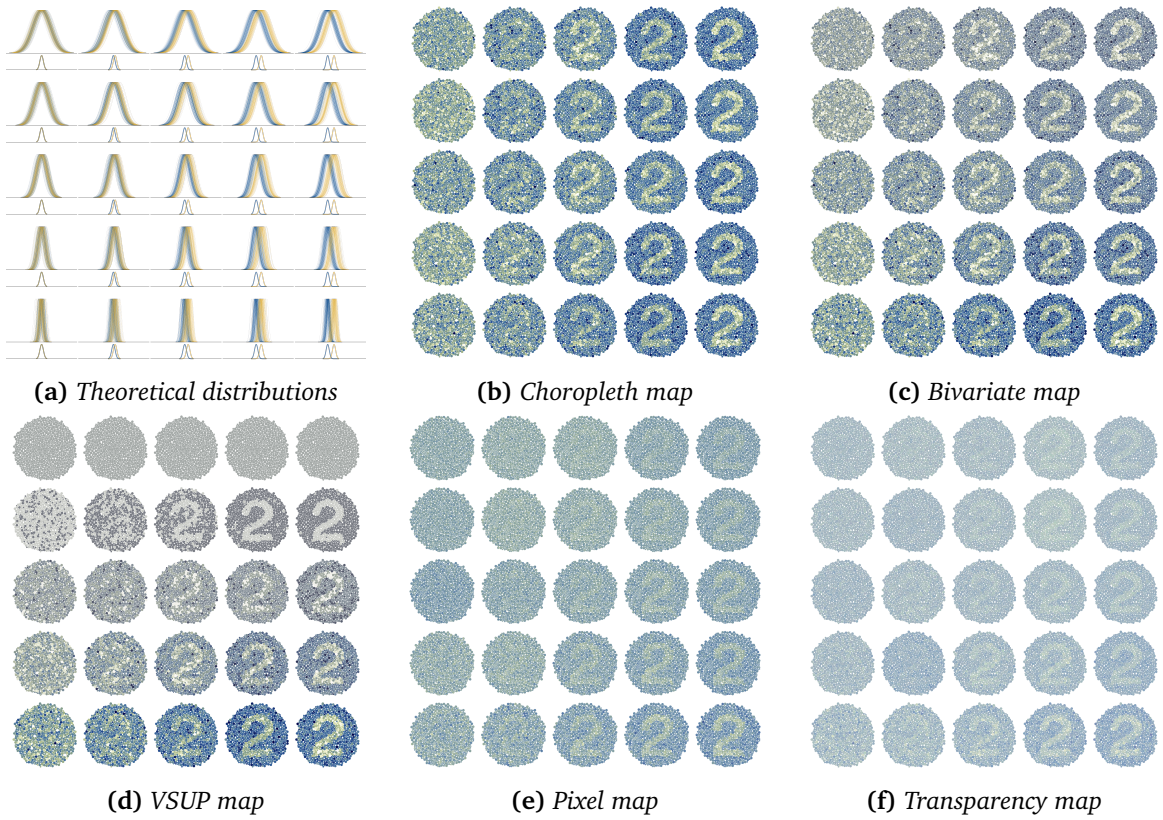
- The five plot types of interest: choropleth, bivariate, VSUP, pixel, and transparency.
- The group difference, D, with values of 0, 1, 2, 3, and 4 were chosen, where 0 results in no difference between the two groups.
- The standard deviation within each observation, V, was set at 1, 2, 3, 4, and 5.

This resulted in a 5x5x5 factorial design, for a total of 125 experimental plots. Additionally, as the type of number displayed may affect the readability of the plot, the number displayed in each plot was randomly assigned for each participant. The order of plots was completely randomised for each participant.

4.3.3 Task and procedure

Participants were asked to specify their country of residence, age, pronouns, education level (based on the International Standard Classification of Education (UNESCO 2026)), and whether or not they had colour vision deficiency. Participants were then instructed that they would be asked to identify numbers in a plot, and that, although the test may resemble a colour blind test, the test was not meant to measure colour vision deficiency. They were asked to set their screen brightness to at least 75% and to turn off colour filters to ensure consistency in results (Gobira et al. 2025). Additionally, plot sizes were standardised. When viewing the plots, participants were able to select a digit ranging from 0 to 9, or indicate that no number was visible within the plot using the interface shown in Figure 4.3. They also had the option to revisit the previous plot, allowing them to correct mistaken selections. After every 25 plots, they were shown a plot with a clearly visible number as an attention check. At the end of the study, participants were able to leave comments. The study was estimated to take approximately 10 to 15 minutes to complete, and responses were collected via Prolific, an online survey website.

General demographic information is meant to provide an overview of our study sample, while



(a) *Theoretical distributions* **(b)** *Choropleth map* **(c)** *Bivariate map*

(d) *VSUP map* **(e)** *Pixel map* **(f)** *Transparency map*

Figure 4.2: *An illustration of the 5x5x5 factorial design used in the experiment, along with the underlying distributional data that makes the plots. The theoretical distribution is a visualisation of 100 of the 1000 individual distributions that make up each dot in the colour blind test, where each individual distribution either belongs to the number (the yellow group) or the background (the blue group). The data is visualised for all levels of D (0,1,2,3,4) on the x-axis, and V (1,2,3,4,5) on the y-axis, across all five plot types (choropleth, bivariate, VSUP, pixel, transparency). We can see that the bottom right corner, when $D = 5$ and $V = 1$, is where all plots are the most visible. An effective uncertainty visualisation should become harder to read as* either *D decreases or V increases, and have a visible triangle stemming from the bottom right corner.*

information regarding colour vision deficiency, when combined with the results, allows us to evaluate unintentional issues in colour selection and plot design. The number of correct selections will be compared across plot types and at the selected treatment distances (D) and standard deviations (V). Ultimately, these values will be compared to generated significance tests in order to compare user output to accepted measures of difference between groups.

4.3.4 Statistical methods

4.3.4.1 Deciding on a ground truth

The extensive literature on the lineup means that there is a wealth of information we can utilise to design our uncertainty visualisation experiment. Designing an experiment where we have an equivalent classical test that can be easily calculated is actually the worst-case scenario for the lineup (Majumder, Hofmann & Cook 2013), and by extension, uncertainty visualisation. When we compare

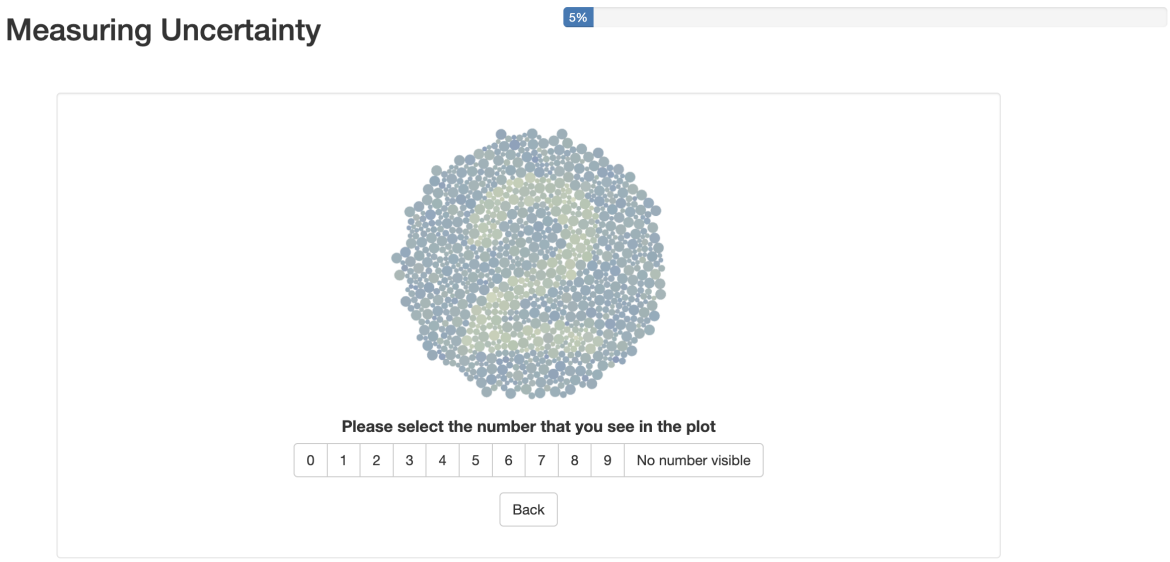


Figure 4.3: *A screenshot of the app as it appeared to participants while answering the stimulus question.*

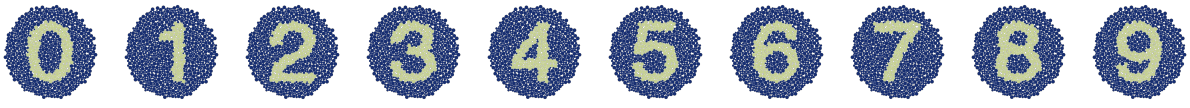


Figure 4.4: *The ten attention checks used in the experiment, included here to illustrate the appearance of all 10 number plates, as well as the clarity of the attention checks.*

our uncertainty visualisations to classical statistical tests, the idea is that, if the visualisations behave well when we *do* know the equivalent hypothesis test, they should also behave well when we don't (Majumder, Hofmann & Cook 2013). Or even better, they will work when we don't even know what we want to test, which is the case for exploratory data analysis.

The correct null hypothesis for any one uncertainty visualisation is not set in stone, and can depend on the rationale participants use to identify the numbers. Strictly looking at the data-generating process, we can see that the data will fulfil the assumptions of a t-test, the classic test for a difference between two groups. This would suggest a t-test is the appropriate ground truth for our experiment. However, in the context of a map, participants might use clustering of lighter or darker spots to make a guess at a shape, even if the full shape cannot be identified. In these cases, tests of spatial autocorrelation, such as Moran's I and Geary's C (Cliff & Ord 1981), that measure clustering of similar data values in areal regions, would be the more appropriate choice. One test is more appropriate given the data-generating context, but the other is more appropriate given the spatial context in which the participants read the plot. For the sake of completeness, we will evaluate participants' responses against both models.

Using accuracy to compare the plots and hypothesis tests might feel like a natural approach to

comparison, but it would be naive as it would end up penalising visualisations that do exactly what we have designed them to do, hide statistically spurious signal. This approach would always suggest our choropleth map to be the best approach, for the very reason we want to avoid using it in the first place. We are not looking for across-the-board "accuracy", but rather, a visualisation that is accurate when a statistical test would be accurate, and inaccurate when a statistical test would be inaccurate. Therefore, we opt to compare visualisations using power curves, which is the same approach taken by the lineup literature (Majumder, Hofmann & Cook 2013). Power curves show the probability of a statistical test detecting an effect, given that the effect actually exists, across a range of effect sizes. Power curves allow us to compare the efficiency of different tests, so, for a given significance level, α, a higher power curve (a higher probability of correctly rejecting a false hypothesis) makes a better test. Ultimately, our goal is for our test to minimise error (type I or type II), and power curves allow us to compare hypothesis tests on this metric. Often, hypothesis tests will cross, so there is no "uniformly most powerful" test. An additional rule of thumb for comparison is "the steeper the curve, the better", as that indicates the test has high sensitivity. Usually, this power curve analysis is performed using effect size, but effect size can be difficult to calculate for many statistical tests, and some tests, such as Moran's I, have no effect size equivalent at all. Therefore, we used D and V as proxies for effect size; however, comparing signal suppression methods using effect size might create smoother results, and is a potential future area of research.

4.3.4.2 Experimental power curve

Participant ability to select the correct number in the plot was modelled using a generalised linear mixed model with a binomial response. Participant correctness is treated as the response variable, and D, V, and plot type are treated as explanatory variables, with up to the three-way interaction between these variables included in the model. D and V are both continuous, while the plot type is discrete. It is standard practice to account for individual ability to read plots using a random block effect (Majumder, Hofmann & Cook 2013), so our model also includes participants as a random effect. The number in the plate is likely to have a similar impact, so that is also included as a random effect.

4.3.4.3 Theoretical power curves

Hypothesis tests are usually specified in terms of a test statistic, which is a function of a sample (Casella & Berger 2024). This is a bit of a problem, because uncertainty visualisations, by their very nature, are designed for situations where we *don't* have a set sample, and instead are working with distributions. There are t-test equivalents, such as the pooled t-test, that allow us to compare two distributions, but this does not exist for our spatial tests, and certainly not with the Bayesian hierarchical model we have used to generate the data. That is, these common spatial autocorrelation metrics fail to take uncertainty into account in their calculations by assuming the variance of the

areal estimates is equal, which results in biased estimates of the spatial structure (Koo, Wong & Chun 2019; Waldhör 1996; Jung, Thill & Issel 2019). For example, Koo, Wong & Chun (2019) showed that measures of spatial autocorrelation are more extreme when uncertainty is not included. Jung, Thill & Issel (2019) warns that spatial autocorrelation measures are problematic without considering uncertainty information, especially in neighbourhood-based studies of public health data, which they displayed using data from the American Community Survey. Few alternatives have been suggested to incorporate uncertainty within spatial autocorrelation metrics. Waldhör (1996) suggests a method to incorporate population size with Moran's I through changes to the covariance matrix, as it is common for population size to vary within areal units, which leads to different variances. Another alternative proposed by Koo, Wong & Chun (2019) is the Spatial Chattacharyya Coefficient, which incorporates the distribution underneath the estimates. However, integrating uncertainty using these theoretical approaches will often only work for one test or the other and cannot be implemented with both the t-test and Moran's I test simultaneously.

Given that our data-generating process was known, we were able to use it to perform a Monte Carlo simulation, where we generated 100 samples of each of the 250 data sets used in the experiment. For each simulated data set, Moran's I and its corresponding p-value were calculated for the alternative hypothesis of positive spatial autocorrelation (or spatial clustering). A permutation test was used for the Moran's I calculation to establish the null distribution of the I statistic, and neighbouring regions were defined as only needing one shared boundary, as the areal regions are circles. The t-test and its corresponding p-value were also calculated, with the alternative hypothesis being a difference between the two means. These p-values are then used to calculate the theoretical power curves.

Estimating power curves for the theoretical hypothesis tests is not straightforward, so we will adapt the methods used by Li et al. (2024). When comparing hypothesis tests, it is standard to restrict the comparison to tests that have the same Type I error probability (Casella & Berger 2024). Using the null plots, that is, when $D = 0$, we can estimate the probability of rejecting H_0 when H_0 is true, which we denote as $\hat{\alpha}_k$, for each plot type, k. Using this $\hat{\alpha}_k$ value, we can set the α for our theoretical tests, turning our p-values into accept/reject outcomes. These data points are comparable to the accept/reject outcomes we were able to observe in the participants' responses. We then use these data points to model the theoretical power of the test using a generalised linear mixed model of the response variable, where D and V are treated as explanatory variables, with a two-way interaction. The correct number is also included as a random effect.

4.4 Results

4.4.1 Participant Information

137 individuals completed the study and passed the attention check. The median study duration was 8.18 minutes with an interquartile range of 4.72 minutes. Because individuals were provided a 'back' button in the case of accidental submissions, only the last observation entered by an individual for each plot is considered. A Pearson's chi-squared test for independence indicates a non-significant relationship between plot type and number present in the data, indicating that random assignment of numbers was successful (p-value of 0.92).

Demographically, participants tended to be younger, use he/him pronouns, and have a tertiary education. 37.96% of individuals identified as being between 18 and 25, while 38.69% identified as 26 to 35; the remaining individuals identified as older than 35. 64.96% of individuals use he/him pronouns, 32.85% use she/her, and the remaining individuals use they/them, or prefer not to answer. 86.86% of individuals indicated that their highest education level was tertiary, while the remaining individuals identified their education level as below tertiary or post-secondary non-tertiary education. A large number of participants were from Africa and Europe (41.61% and 30.66%, respectively), and 5.11% of participants indicated that they were colour blind or unsure.

4.4.2 Results Overview

Figure 4.5 shows the percent of participants who were able to make a correct selection for each type of plot, based on the distance (D) between the number and non-number groups, and the standard deviation (V) of the subsamples. The grid for the choropleth and bivariate maps appears to be quite similar, with no visible effect from the change in V, only the change in D. Looking at the VSUP map, we do get some interference, as expected, but the interference appears to be limited to the cases when the $D = 1$ or $V = 5$. Outside of those two scenarios for the VSUP map, there does not appear to be a clear relationship with V. The pixel and transparency map appear to have the lower triangles of visibility. However, the theoretical hypothesis data does not appear to have the same sensitivity as the transparency or pixel map.

4.4.3 Power analysis

4.4.3.1 Estimating significance levels

To set the α for the theoretical test, we need to calculate the $\hat{\alpha}$ value for each plot type, k, where $k \in \{choropleth, bivariate, vsup, pixel, transparency\}$. While the lineup protocol has a theoretical calculation that can be used to estimate α due to the fact that any plot being picked by chance is $\frac{1}{M}$ for a lineup with M plots (assuming there is no dependence structure between the plots) (Majumder,

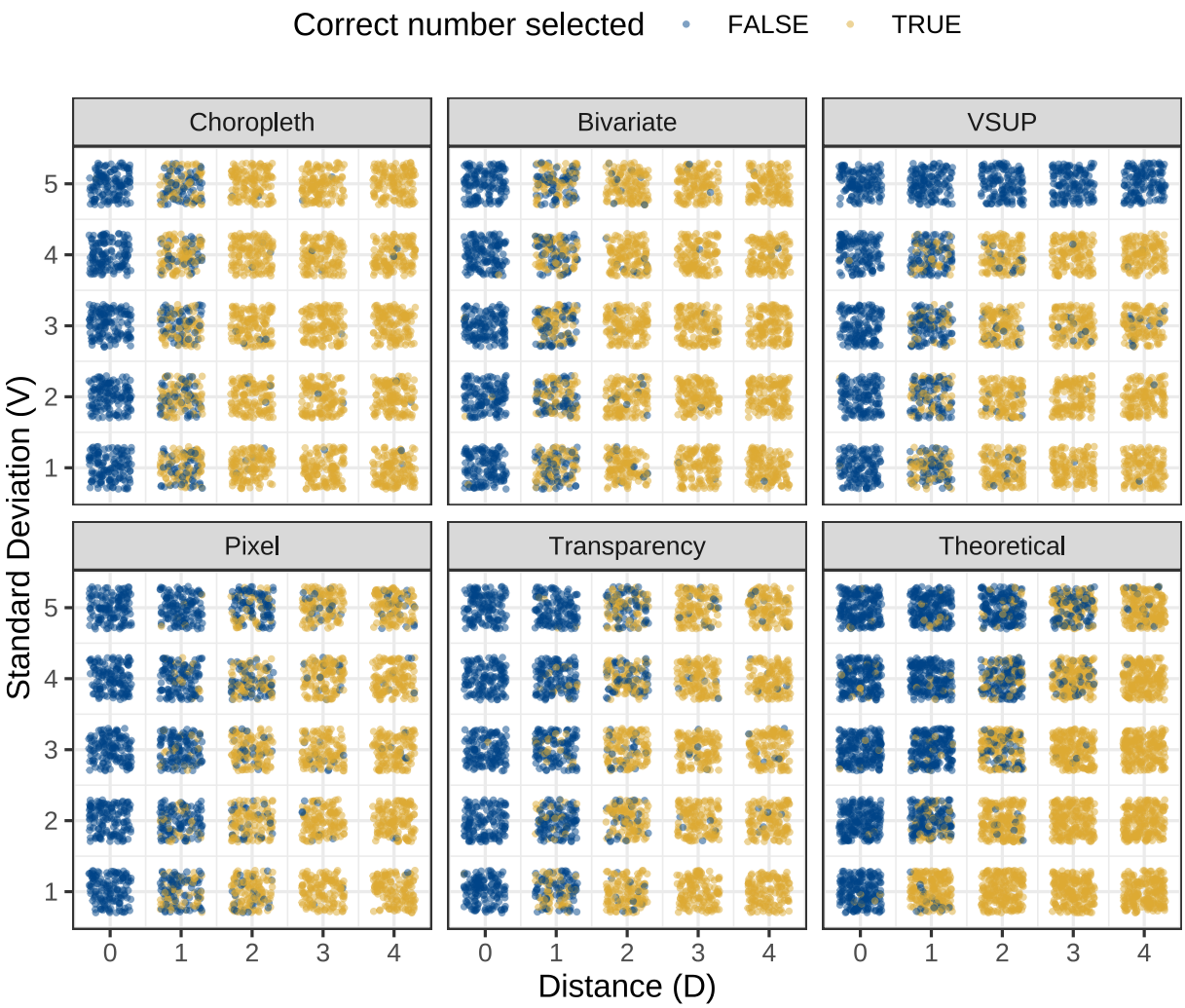


Figure 4.5: *The full set of participant responses, coloured by whether or not they were able to identify the correct number in the plot, grouped according to the distance (D) between the distributions, and the standard deviation (V) in each individual estimate. The theoretical simulation from the t-test and Moran's I calculations is also included as a point of reference. We can see that the choropleth and bivariate map have an I-shape, the VSUP map makes an L-shape, and the pixel and transparency maps make an upper triangular shape. The I-shape indicates the variance has no impact on the visibility, the L-shape indicates the variance only has an impact on visibility at its maximum, and the upper triangular shape shows a consistent impact on visibility from the variance. None of the plot types perfectly match the pattern in the theoretical visualisation.*

Hofmann & Cook 2013; VanderPlas et al. 2021), this calculation does not translate to uncertainty visualisation. As graphics do not come with a significance threshold (VanderPlas et al. 2021), and a significance threshold is required to compare our plots to classic hypothesis tests, we estimate $\hat{\alpha}_k$ using the proportion of participants who identified a number in a plot that was just random noise.

Figure 4.6 shows the bootstrapped distribution for the significance level of each plot, $\hat{\alpha}_k$. We can see that there is quite a large range in the distribution, which is likely caused by individual differences in risk aversion. Reading the comments in the study, we found that some participants thought they

were always supposed to see a number, and would make a guess if they had the slightest inkling towards a specific number, while other participants would only make a guess if they could fully make out the number's shape. These tactics align with Moran's I test and t-test, respectively. It may be more appropriate to also treat our value of $\hat{\alpha}_k$ as a random effects model, but this will theoretically and computationally overcomplicate the model. Therefore, we will use the average number of false positives within each plot type to estimate $\hat{\alpha}_k$.

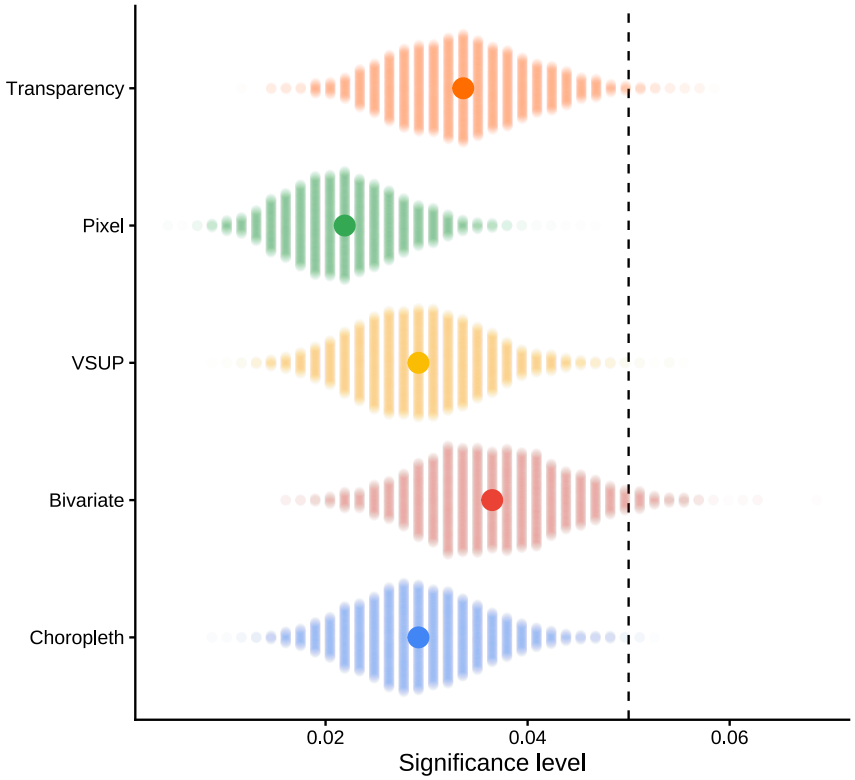


Figure 4.6: *Bootstrapped significance levels for each plot type, shown as jittered dotplots, computed using the null plots, $D = 0$. The large points are the full sample values that are used in the power analysis. The vertical line indicates the conventional significance level, 0.05. We can see that all the significance levels seem to hover around 0.03, with pixel maps sitting slightly lower than the other plot types.*

4.4.3.2 Random effects model

Figure 4.7 shows the random effects model and fixed effects components for each plot type, D and V, with a black dashed line indicating the theoretical power curve of our hypothesis tests. The two different dashed lines show the difference between the simulated t-test and Moran's I, and are slightly different shapes as the power curve is set based on that plot type's significance level, $\hat{\alpha}_k$. The random effects model allows us to see the variation in individual participants, while the fixed effects average allows us to easily compare the different plot types. In the random effects plot, the variability in the black dashed lines comes from the random effect of the numbers shown in the plate, while the variability in the experimental results comes from both the random effects of the numbers in the plate

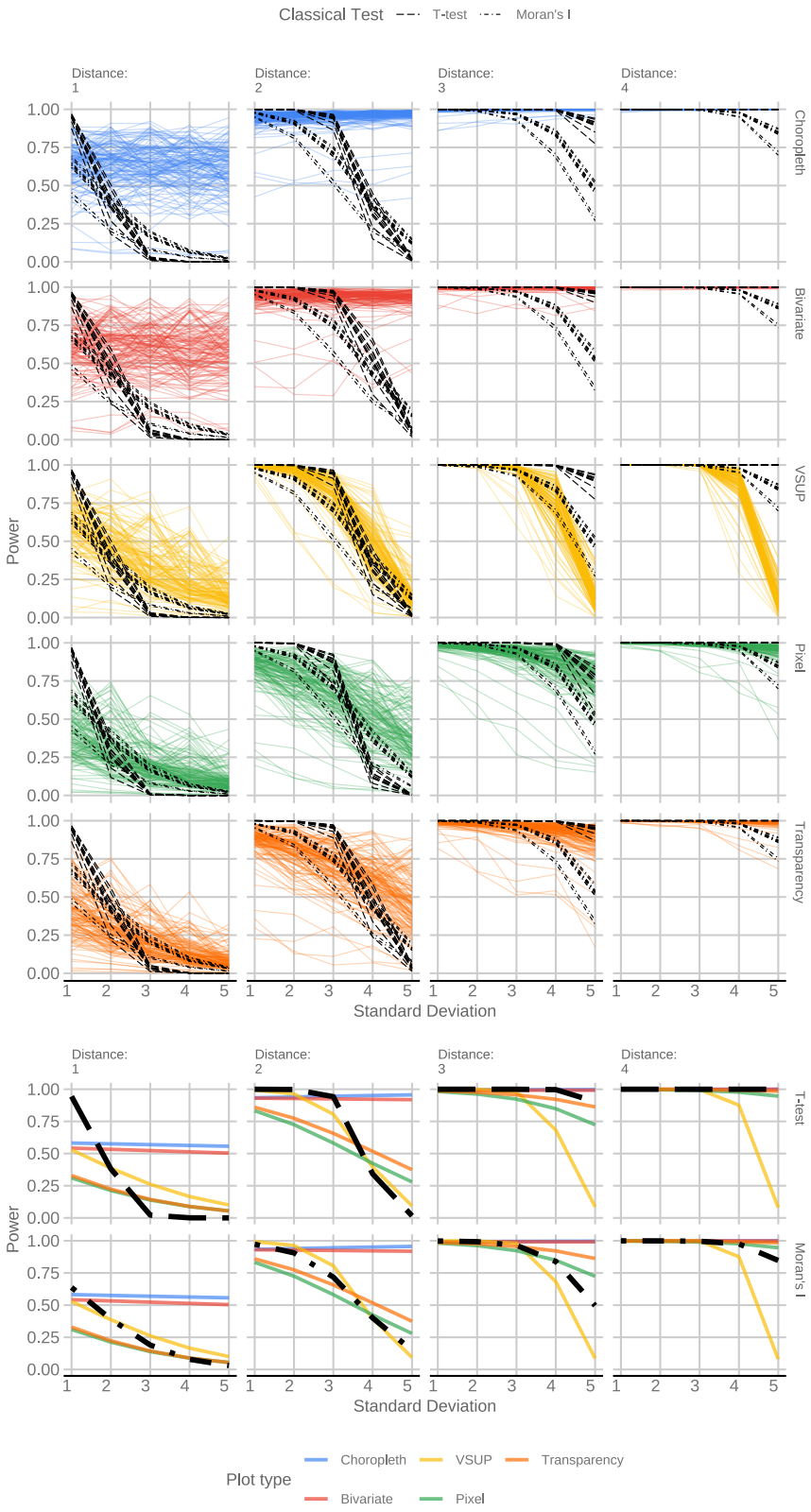


Figure 4.7: *A visualisation of the experimental power curves, estimated using a generalised linear random effects model (top) and generalised linear fixed effects model (bottom) for all 125 experimental factors: V (standard deviation), D (Distance), and K (plot type). The results of the theoretical hypothesis tests, the t-test, and Moran's I test, are shown using two different black dashed lines. We can see that the variance has little impact on the signal visibility in the choropleth and bivariate map, the visibility of the signal in the VSUP map shrinks to zero when V=5, regardless of the effect size, and the pixel and transparency map broadly follow the shape of the theoretical tests.*

and the participants' ability to read the plot. As the variance increases, we would expect the signal in the plate to become harder to see, as its statistical significance drops off. Ideally, this drop off would occur at a rate that is similar to the hypothesis tests, represented by the black dashed lines. If the coloured line for a plot sits above the hypothesis test, then the plot is showing a pattern that the classical hypothesis tests would consider to be insignificant. While this can be a desirable property when comparing standard hypothesis tests, this is not necessarily a desirable property for uncertainty visualisation. It is important to remember that the key issue with choropleth maps, and the reason we need uncertainty visualisation at all, is that they *always* show statistical signal, even if that signal is spurious. We have designed this experiment such that the visibility *does* link to a signal, but that will not always be the case. Therefore, the power curve of the most appropriate plot will not necessarily sit above the black dashed lines, but rather, should have a shape that is similar to the hypothesis test curves across the different values of D.

These charts seem to largely agree with our initial findings. The choropleth and bivariate maps are constant, and the visibility of the plot appears to be largely independent of V, only changing with an increase in the average effect. The VSUP appears to follow the theoretical line for $D = 1, 2$, but diverges for $D = 3, 4$ as the statistically significant difference at $V = 5$ is hidden by the monochrome plate. This divergence suggests that the alignment when $D = 1, 2$ is a coincidence born from the fact that when $D = 1, 2$, the power of the hypothesis test should be zero, and $V = 5$ will also produce a power of zero, independent of the effect size in the plot. Finally, we see that the pixel and transparency maps appear to do a reasonably good job of following the power curve of the theoretical hypothesis test, but the power is falling short at both ends of the spectrum. The signal is not visible enough when variance is low, but it is too visible when variance is high. This indicates that the test is not quite sensitive enough to changing values of V. This issue in the pixel and transparency maps becomes less prevalent as D increases, and at higher values of D, the visualisation actually sits between the results of the t-test and Moran's I test.

A closer inspection of the random effects models reveals some patterns in the participants' ability to read the plot. There is quite a bit of variation among participants, especially when $D = 1$. As the D increases, the variance in the participants' responses decreases, a pattern that is less pronounced in the pixel and transparency map. This indicates that, as the plots become easier to read, either due to decreasing V or increasing D, the variance among viewers will also decrease. Some participants perform rather badly across the board, but we can see others manage to track the theoretical power curve rather well. This indicates that, unlike in traditional hypothesis tests, there is a degree of skill involved in identifying a signal in uncertainty visualisations. The importance of this skill becomes more pronounced as the effect size gets smaller. This indicates that training could potentially improve

Table 4.1: *Effect of Standard Deviation (V) by Plot Type, Averaged Over Distance (D)*

Plot Type	V Effect	SE	Z Ratio	P-Value
Choropleth	0.179	0.144	1.238	0.216
Bivariate	-0.044	0.121	-0.362	0.718
VSUP	-2.486	0.106	-23.359	0.000
Pixel	-0.700	0.058	-12.015	0.000
Transparency	-0.604	0.068	-8.843	0.000

Table 4.2: *Selected Comparison of Plot Types for Standard Deviation (V), Averaged Over Distance*

Contrast	Estimate	SE	Z Ratio	P-Value
Choropleth - Bivariate	0.223	0.188	1.181	0.475
Transparency - Pixel	0.096	0.090	1.075	0.565

the power of plots as statistical tests.

4.4.3.3 Hypothesis tests

Relating to H1: Changes to the standard deviation in the distributions will result in no meaningful difference in our ability to read the number in the choropleth and bivariate map.

Table 4.1 shows the estimated marginal effect of V for each plot type, averaging over the effect of D. If the marginal effect is significantly different from zero, then V has some effect on the visibility in the signal in the plot, but if it isn't significantly different from zero, then the variance does not have any impact on signal visibility at all. We can see that the effect associated with V for each plot type is insignificant in the case of the bivariate map and choropleth map, but significant for all other map types. These conclusions remain true if we perform significance tests at set $D = 1, 2, 3, 4$, instead of looking at the average across D. These distance-based results are available in Chapter 6. This indicates that V has no significant effect on the visibility of the signal in the bivariate and choropleth maps, given the generalised linear mixed effects model.

Relating to H1 and H3: The probability of correctly reading the transparency and pixel map, as well as the probability of correctly reading the choropleth and bivariate map, will be similar.

Comparisons of the standard deviation effect between plot types of interest are shown in Table 4.2. The p-values were adjusted for multiple comparisons using the Sidak adjustment. If the statistical information conveyed by two plots is equivalent, the signal visibility between the two plots will be equivalent. If this is true, the probability of correctly reading the transparency and pixel maps will be similar. For the pairwise comparisons of the transparency/pixel map and the choropleth/bivariate map, significance tests at $D = 1, 2, 3, 4$ result in the same conclusions in terms of significance. Distance-based results as well as all pairwise comparisons are available in the Chapter 6. The non-significant difference between the transparency and pixel map, as well as between the choropleth and bivariate

map, indicates that standard deviation effects are similar between the plot types, as expected.

4.5 Discussion

All hypotheses posed at the beginning of the experiment were either fully or partially confirmed, creating a foundation for a theory of signal suppression in uncertainty visualisation.

Unsurprisingly, the choropleth map had an insignificant relationship with standard deviation, as the standard deviation of each distribution was not included at all in the choropleth map. What *is* of note is that the bivariate map *also* had a coefficient of zero, and its coefficients on the mixed effects model were not significantly different from the choropleth map. This finding is of particular note, as bivariate maps are frequently suggested as an alternative to choropleth maps, specifically for the purposes of suppressing statistically invalid signals. However, these results indicate that, if your goal is to make statistically invalid signals invisible, *a bivariate map is functionally identical to a choropleth map*. Any benefit in using a bivariate map will need to come from explicit calculation, and a visualisation that requires a mental calculation to be understood defeats the purpose of making a visualisation in the first place.

The results partially supported our second hypothesis. The VSUP map, unlike the bivariate map, did achieve visual interference and was not functionally identical to the choropleth map; however, this interference did not align with the theoretical hypothesis tests. The fact that every plate at $V = 5$ was completely monochrome disproportionately pulled down the slope of the random effects model. This caused better alignment with the theoretical tests for $D = 2$, but incredibly poor alignment when $D = 4$ and $D = 5$. The VSUP and bivariate maps were designed with the implicit knowledge of the maximum and minimum values of the data-generating process, which was only possible due to the artificial scenario of the experiment. This requirement of knowledge could be considered a form of data snooping that disproportionately benefits the evaluation of the VSUP and bivariate maps, but the maps are impossible to calculate without it, and it serves as one of the main limitations to the approach. Usually, the scaling of our axis or colours is done automatically, but locally scaled bivariate and VSUP maps, with no attention to the relationship between standard deviation and estimate value, are completely meaningless. One could argue that a workaround for this issue is leveraging the monotonic shrinkage algorithms for VSUP palettes suggested by Kay (2019). However, as Wickham (2010) pointed out, the distinction between a coordinate transformation and a statistical transformation is not always so clear-cut. Therefore, adjusting the coordinate to get a specific statistical output is indistinguishable from adjusting the statistic itself. Not only would this violate one of our original limitations for plot selection, but it would also significantly increase the number of plots

we would need to evaluate. Therefore, we opted to test the classic VSUP map and did not extend to the methods suggested by Kay (2019). The VSUP approach also has several other implicit limitations that we did not cover in this experiment. For example, it cannot handle discontinuous distributions that oscillate between two distant outcomes, rather than maintaining a smooth transition between values. The VSUP approach also only works for ordered variables, and does not make intuitive sense if we are uncertain about a categorical fill. These limitations are absent in the pixel and transparency map.

Finally, the results for the pixel and transparency maps have some interesting implications for the statistical properties of plots. Given that the two plots were aligned, we can isolate the benefit in the approach to the display of a full distribution, rather than the representation of a distribution as multiple values. This means we can still effectively perform signal suppression in a situation where the pixel map is inappropriate. These plots showed that we *can* achieve a signal that aligns with hypothesis testing (or at least a signal that has a consistently similar shape). There are likely several design changes that could improve this alignment, such as an optimised colour palette or sample size.

Our results also have implications for the broader concepts of design in uncertainty visualisations. The confirmation of **H1** indicates that treating uncertainty as a second variable, independent of the estimate, will not facilitate the suppression of false signals. Plotting the estimate and uncertainty to separate aesthetic channels will prevent the uncertainty from visually interfering with the estimate, which means it has no mechanism to make more uncertain estimates harder to see, and will not prevent the visualisation from displaying spurious signals. The confirmation of **H2** indicates that the correct representation of an estimate should be one that completely describes the distribution. These findings provide empirical evidence for the formalisation suggested by Kay (2023). The confirmation of **H3** suggests that, if we visualise our distribution as a sample, and ensure all the draws are equally weighted, they will provide an identically valid statistical signal. This is particularly useful for the `ggdibbler` (Mason et al. 2026b) R package, which leverages this perpendicularity in the grammar of graphics to make a range of flexible uncertainty visualisations for EDA. If this hypothesis had proven to be false, these position adjustments would not be independent of the statistical information, and changing the position adjustment would change the signal suppression in the plot, making the suggested methods ineffective. In the case where multiple random variables are being fed into the system at once, transparency is the only viable position adjustment to allow for even weighting of all sample outcomes. Our results show that the system maintains statistical validity, even when we are unable to translate the results back to a numerical scale using the legend.

4.6 Contributions, limitations, and future research

Several contributions were made in this paper. We designed a visualisation experiment that is able to measure uncertainty as a latent variable, capturing its effect as noise, rather than signal. We also translated the statistical approaches from the lineup protocol to uncertainty visualisation, allowing us to compare pattern visibility to the outcomes of a t-test and Moran's I test. By constructing the plots with respect to the grammar of graphics, we were able to understand *why* some approaches may, or may not, work and contribute towards a general theory of uncertainty visualisation, rather than limit ourselves to sweeping statements about plot types.

While these contributions are quite substantial, the work is not without its limitations. The first limitation is that we were strict in our data-generating process. We decided to keep the data generation very simple, to avoid confounding data issues with other untested aspects of the design, but this did result in quite strict assumptions. Our distribution shapes were limited to normal distributions, every observation had an identical standard deviation with no interfering pattern, and the only variance between observations came from the estimate's central values. While this data-generating process is limited, deviations from this strict scenario are more likely to help, rather than hinder, the uncertainty visualisation in comparison to the theoretical hypothesis tests. There is ample evidence that classical hypothesis tests will outperform visual inference when the assumptions of the test are true, but the visual tests will do better when the assumptions are false (Majumder, Hofmann & Cook 2013; Li et al. 2024; Hofmann et al. 2012). This indicates that more unusual data-generating processes that are more likely to violate traditional test assumptions may improve the performance of the visualisations.

There were quite a few factors in the plot design that likely impact the readability of the figures, but were outside the scope of this experiment. Colour palette choice can affect how patterns are perceived in statistical graphics (Reda & Szafir 2021a). Anecdotally, when designing the experiment, we found that some palettes had a significant difference in visibility relative to others. A different choice in colour palette may produce different results, although we hypothesise that the standard deviation effects would remain similar. Therefore, values resulting from visibility calculations should not be treated as definitive values for the data used to generate the plots.

There were also several extensions to this work that might improve the sensitivity issues in the pixel and transparency maps. One is the effect of sample size in the visualisations. We set the number of samples within each plot to be 50, which reduces the variance within the visualisation, but also compresses the colour scale, making individual colours harder to discern. There may be a sample size trade-off in designing these visualisations that is having an effect on their sensitivity. Theoretically, the pixel map approach could be generalised to any visualisation that can be made in `ggplot2`

(Mason et al. 2026b), so the signal suppression approach could be investigated for other aesthetics in other types of plots. Additionally, we could extend our question to the cases where we have multiple uncertain variables at once, rather than only allowing variability in colour. The size of the plot also seemed to have an effect on the visibility of the pattern, with smaller plots making the pattern easier to see.

Given the fact that the transparency map was effective, it would become a viable visualisation approach if an interpretable legend could be presented alongside the map. In its current state, as can be seen in Figure 4.1, the colours do not exactly map to the legend, and as the standard error increases, that issue only gets worse as sample outcomes jump wildly around the colour scale and mix the colours shown in the plot.

There are also several potential extensions or issues we encountered when implementing the colour blind test that should be of note to anyone who wants to implement a similar method. We noticed an afterimage occasionally when doing the test, so including an image as a mask between plots may help alleviate this issue. We did not conduct testing of font types, so the tendency for participants to get confused between numbers (specifically 6, 3 and 8) might be mitigated with a better font choice. There are also colour blind tests that ask participants to trace a shape, rather than identify a number, which are less susceptible to memorisation. Using software that allows participants to draw on data visualisations, such as the `r2d3` (Robinson, Howard & VanderPlas 2023) R package, would allow researchers to implement that variant of the noise blind tests. Alternatively, including a variety of shapes in the test may reduce the probability of random guessing by the participants.

Ethics declaration

Ethics approval for the online survey was granted by Monash University Human Research Ethics Committee (Project ID 51214). All applicants provided informed consent prior to participating in this research.

4.7 Acknowledgements

The first author of this paper is supported in part by a scholarship from the Australian Energy Market Operator. This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship [DOI: https://doi.org/10.82133/C42F-K220]. We thank Susan VanderPlas, Sarah Goodwin, and Emily Robinson for their insightful comments and feedback, which substantially improved the work. The R packages used for this work were: `tidyverse`

(Wickham et al. 2019), `distributional` (O'Hara-Wild et al. 2024), `ggdist` (Kay 2023), `ggdibbler` (Mason et al. 2026b), `patchwork` (Pedersen 2025b), `khroma` (Frerebeau 2025), `colourspace` (Stauffer et al. 2009), `ozmaps` (Sumner 2021), `sf` (Pebesma 2018), `ggthemes` (Arnold 2024), `MASS` (Venables & Ripley 2002), `shadowtext` (Yu 2025), `flextable` (Gohel & Skintzos 2024), `emmeans` (Lenth 2025), `kableExtra` (Zhu 2024), `broom` (Robinson et al. 2026), `lme4` (Bates et al. 2015), `car` (Fox & Weisberg 2019), `janitor` (Firke 2024), `packcircles` (Bedward, Eppstein & Menzel 2024), `gglogo` (Hofmann, Hare & GGobi Foundation 2026), `scales` (Wickham, Pedersen & Seidel 2025), `glue` (Hester & Bryan 2024), `digest` (Eddelbuettel 2025), `ggbeeswarm` (Clarke, Sherrill-Mix & Dawson 2025), `conflicted` (Wickham 2023), `sp` (Bivand, Pebesma & Gomez-Rubio 2013), and `spdep` (Pebesma & Bivand 2023). The GitHub repository for this paper can be found at https://github.com/harriet-mason/uncertainty-experiment, which contains the files required to reproduce this article in full.

Chapter 5

Conclusion

The three papers presented in this thesis share a common theme of communicating uncertainty in a way that it can be *seen*. These papers present a cohesive vision for the future of uncertainty visualisation, one that is defined within the infrastructure set by statistical graphics. We established the foundations of this vision by approaching the uncertainty visualisation problem from three separate angles: the philosophical goals of the field, the practical limitations of the mathematical objects we are working with, and the human perception of uncertainty in visualisation.

5.1 Contributions

This thesis makes several important contributions to the uncertainty visualisation literature.

First, Chapter 2 offers a philosophical foundation for the future of uncertainty visualisation that resolves the ongoing conflicts that currently plague the field. By connecting our evaluation of uncertainty visualisation back to its primary purpose, we are able to map out what it actually means to "see" uncertainty in a visualisation. Using this context, we illustrate that much of the conflict in the literature comes from a mismatch between the stated goals and evaluation methods utilised by the field. These insights allow us to present a cohesive vision for the future of uncertainty visualisation, built upon the goal of visualising uncertainty as "noise".

Second, Chapter 3 presents two key contributions: a mathematical framework for uncertainty visualisation, and the introduction of the `ggdibbler` R package. The mathematical framework presents the argument that the grammar of graphics is a continuous function, and therefore, statistical graphics should adhere to the continuous mapping theorem. The key insights of this formalisation are translated into the flexible uncertainty visualisation software, `ggdibbler`, the second contribution

from this chapter. The `ggdibbler` software allows us to substitute any vector of values with a vector of random variables, and make an uncertain version of any graphic that can be made in `ggplot2`. This software received the 2026 John Chambers Award for Statistical Software from the ASA Sections on Statistical Computing and Statistical Graphics, it has 21 stars on GitHub, and has been downloaded over 1800 times on CRAN since it was uploaded in July 2025. This indicates the value the statistics community has placed on flexible uncertainty visualisation software that can be used for EDA.

Finally, Chapter 4 uses the mathematical formalisation of Chapter 3 in a perceptual evaluation of choropleth maps to confirm the hypotheses discussed in Chapter 2. Evaluating uncertainty as noise required the development of a new experimental methodology that is able to measure the effect of a latent variable (noise) on an observable variable (signal). The results of this paper show that visualising a distribution as a set of samples is not only the most flexible approach (as established by Chapter 3), but it is also the approach that most accurately conveys the results of standard statistical tests.

5.2 Future work

There are several avenues one could take to improve upon this work; we will list the most important ones here.

5.2.1 `ggdibbler` software

The full list of improvements and fixes that should be made to `ggdibbler` can be seen on the package's GitHub issues. In this section, we will touch on some of the high-level improvements that could be made to the software.

One of the core assumptions of the software is that it assumes each individual cell of our table is a distribution, and these distributions are completely independent. This assumption can be quite strict, and the most natural improvement to the software would be to allow users to pass joint distributions to allow for dependency between cells in the random matrix. The theoretical framework already allows for this, as does the underlying `distributional` package; the only limitation currently imposed is by the `ggdibbler` software. However, implementing this change is not straightforward, as it would require us to reach inside the joint distribution to map variables, and could possibly require bespoke syntax to get the extension working.

Another improvement to the software is the flexibility to allow any object type. Similarly to the joint distributions, this is already allowed by the theoretical foundations as well as `distributional`. Practically, this would require us to set up a nested distribution scale for all scale types. That is,

currently the software replaces `scale_x_continuous` with `scale_x_continuous_distribution`, which works for continuous data, but `ggplot2` also has scales such as `scale_x_datetime`, which does not have a `scale_x_datetime_distribution` counterpart. While it would be straightforward to implement this with the existing `ggplot2` supported scales, much of the benefit of `ggplot2` comes from its network of extensions that allow for many different object types. Implementing `ggdibbler` with `ggplot2` extensions is an issue that goes beyond scales, and similar limitations exist with stats and positions. Implementing some kind of "function factory" or wrapper solution might give us the flexibility needed to allow `ggdibbler` to work with all input types, these developments are still a work in progress.

5.2.2 Latent variable testing for all aesthetics

While a variation of the Ishihara colour blind tests allowed us to measure uncertainty as a latent variable, this approach is only viable when we have mapped our random variable to colour. The theoretical framework developed for uncertainty visualisation should allow us to map uncertainty to any aesthetic in any combination. Designing similar experiments for the most commonly used aesthetics (including position, shape, length, etc.) would serve as a value baseline for the evaluation of uncertainty in statistical graphics.

5.2.3 A fundamental theory of visualisation

Many elements of this thesis point towards a more fundamental theory of visualisation that can be tied to the grammar of graphics. This alternative connection would allow us to better formalise our graphics as mathematical objects, which, as illustrated by Chapter 3 and Chapter 4, would facilitate more flexible plot design and more informative perceptual experiments. This fundamental theory of visualisation would allow us to treat visualisations as statistics with desirable statistical properties, such as statistical sufficiency, bias, and variance. This line of thinking is similar to the argument made by Wickham & Hofmann (2011) in their discussion of product plots. The authors draw a link between statistical graphics and their underlying distributions using the fact that both geometry and probability are born from measure theory. Strengthening this connection is likely a good avenue for future research.

Bibliography

Allaire, J & C Dervieux (2024). *quarto: R interface to quarto markdown publishing system*. R package version 1.4.4. https://CRAN.R-project.org/package=quarto.

Anscombe, FJ (1973). Graphs in statistical analysis. *The American Statistician* **27**(1), 17–21.

Arnold, JB (2024). *ggthemes: Extra themes, scales and geoms for 'ggplot2'*. R package version 5.1.0. https://CRAN.R-project.org/package=ggthemes.

Bartonicek, A, S Urbanek & P Murrell (2025). No more, no less than sum of its parts: Groups, monoids, and the algebra of graphics, statistics, and interaction. en. *Journal of Computational and Graphical Statistics* **34**(3), 1063–1074.

Bates, D, M Mächler, B Bolker & S Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**(1), 1–48.

Bedward, M, D Eppstein & P Menzel (2024). *packcircles: circle packing*. R package version 0.3.7. https://CRAN.R-project.org/package=packcircles.

Begg, SH, MB Welsh & RB Bratvold (2014). Uncertainty vs. variability: what's the difference and why is it important? In: *SPE Hydrocarbon Economics and Evaluation Symposium*. Vol. SPE Hydrocarbon Economics and Evaluation Symposium. https://doi.org/10.2118/169850-MS.

Belsley, D, E Kuh & R Welsch (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley Series in Probability and Statistics. New York: Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/0471725153.

Benjamin, DM & DV Budescu (2018). The role of type and source of uncertainty on the processing of climate models projections. *Frontiers in Psychology* **9**(MAR), 1–17.

Bivand, R, E Pebesma & V Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. https://asdar-book.org/.

Bivand, R & C Rundel (2023). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. R package version 0.6-3. https://CRAN.R-project.org/package=rgeos.

Blenkinsop, S, P Fisher, L Bastin & J Wood (2000). Evaluating the perception of uncertainty in alternative visualization strategies. *Cartographica* **37**(1), 1–13.

Boger, T, SB Most & SL Franconeri (2021). Jurassic mark: Inattentional blindness for a datasaurus reveals that visualizations are explored, not seen. In: *2021 IEEE visualization conference (VIS)*. IEEE, pp.71–75.

Bokulich, A & W Parker (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science* **11**(1), 31.

Bornkamp, B (2018). Calculating quantiles of noisy distribution functions using local linear regressions. *Computational Statistics* **33**(1), 487–501.

Bostrom, A, L Anselin & J Farris (2008). Visualizing seismic risk and uncertainty: A review of related research. In: *Annals of the New York Academy of Sciences*. Vol. 1128. Blackwell Publishing Inc., pp.29–40.

Boukhelifa, N, A Bezerianos, T Isenberg & JD Fekete (2012). Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics* **18**(12), 2769–2778.

Boukhelifa, N, ME Perrin, S Huron & J Eagan (2017). How data workers cope with uncertainty: A task characterisation study. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, pp.3645–3656. https://doi.org/10.1145/3025453.3025738.

Brennen, A & S Tuerk (2018). An instrument for evaluating uncertainty visualization techniques. *Conference on Human Factors in Computing Systems - Proceedings* **2018-April**. ISBN: 9781450356206, 1–6.

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne & H Wickham (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4361–4383.

Bullock, DS, M Boerngen, H Tao, B Maxwell, JD Luck, L Shiratsuchi, L Puntel & NF Martin (2019). The data-intensive farm management project: Changing agronomic research through on-farm precision experimentation. *Agronomy Journal* **111**(6), 2736–2746. (Visited on 03/25/2026).

Casella, G & R Berger (2024). *Statistical inference*. Chapman and Hall/CRC.

Chakraborty, S, P Kiefer & M Raubal (2024). The influence of uncertainty visualization on cognitive load in a safety- and time-critical decision-making task. en. *International Journal of Geographical Information Science* **38**(8), 1583–1610. (Visited on 11/11/2025).

Chatfield, C (1985). The initial examination of data. *Journal of the Royal Statistical Society. Series A (General)* **148**(3), 214–253. (Visited on 03/19/2026).

Cheong, L, S Bleisch, A Kealy, K Tolhurst, T Wilkening & M Duckham (2016). Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty. *International Journal of Geographical Information Science* **30**(7), 1377–1404.

Clarke, E, S Sherrill-Mix & C Dawson (2025). *ggbeeswarm: Categorical scatter (violin point) plots*. R package version 0.7.3. https://CRAN.R-project.org/package=ggbeeswarm.

Cleveland, WS & R McGill (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* **79**(387), 531–554.

Cliff, AD & J Ord (1981). *Spatial processes: models & applications*. London: Pion.

Cook, D, EK Lee & M Majumder (2016). Data visualization and statistical graphics in big data analysis. en. *Annual Review of Statistics and Its Application* **3**(1), 133–159. (Visited on 10/07/2025).

Cook, D, N Reid & E Tanaka (2021). The foundation is available for thinking about data visualization inferentially. en. *Harvard Data Science Review*. (Visited on 11/04/2025).

Correll, M & M Gieicher (2015). Implicit uncertainty visualization: Aligning perception and statistics. In: *Workshop on Visualization for Decision Making under Uncertainty. https://api. semanticscholar. org/CorpusID*. Vol. 16691049.

Correll, M & M Gleicher (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics* **20**(12), 2142–2151.

Correll, M & J Heer (2016). Surprise! Bayesian weighting for de-biasing thematic maps. *IEEE transactions on visualization and computer graphics* **23**(1), 651–660.

Correll, M, D Moritz & J Heer (2018). Value-suppressing uncertainty palettes. *Conference on Human Factors in Computing Systems - Proceedings* **2018-April**, 1–11.

Crameri, F (2018). Geodynamic diagnostics, scientific visualisation and StagLab 3.0. English. *Geoscientific Model Development* **11**(6). Publisher: Copernicus GmbH, 2541–2562. (Visited on 03/11/2026).

Crameri, F, GE Shephard & PJ Heron (2020). The misuse of colour in science communication. en. *Nature Communications* **11**(1). Publisher: Nature Publishing Group, 5444. (Visited on 03/12/2026).

Csárdi, G (2025). *cli: Helpers for Developing Command Line Interfaces*. R package version 3.6.5. https://CRAN.R-project.org/package=cli.

Dupin, C (1826). *Carte figurative de l'instruction populaire de la France*. ark : /12148 / btv1b530830640.

Eddelbuettel, D (2025). *digest: Create compact hash digests of R objects*. R package version 0.6.38.

Firke, S (2024). *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.2.1. https://CRAN.R-project.org/package=janitor.

Fischhoff, B & AL Davis (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13664–13671.

Fox, J & S Weisberg (2019). *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Franconeri, SL (2021). Three perceptual tools for seeing and understanding visualized data. *Current Directions in Psychological Science* **30**(5), 367–375.

Frerebeau, N (2025). *khroma: Colour schemes for scientific data visualization*. R package version 1.17.0. Université Bordeaux Montaigne. Pessac, France. https://packages.tesselle.org/khroma/.

Gentleman, RC, VJ Carey, DM Bates, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**(10), R80.

Gobira, M, V Freire, C Tinoco, GS Avelino, P Carricondo, A Dias & MA Negreiros (2025). Assessing the accuracy of a digital color vision test. *Archivos de la Sociedad Española de Oftalmología (English Edition)* **100**(12), 781–787. (Visited on 03/04/2026).

Gohel, D & P Skintzos (2024). *flextable: Functions for tabular reporting*. R package version 0.9.6. https://CRAN.R-project.org/package=flextable.

Goldstein, DG & D Rothschild (2014). Lay understanding of probability distributions. *Judgment and Decision Making* **9**(1), 1–14.

Griethe, H & H Schumann (2006). The visualization of uncertain data: Methods and problems. In: *SimVis*. Vol. 6, pp.143–156.

Gschwandtner, T, M Bögl, P Federico & S Miksch (2016). Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics* **22**(1), 539–548.

Guo, Z, A Kale, M Kay & J Hullman (2025). VMC: A grammar for visualizing statistical model checks. *IEEE Transactions on Visualization and Computer Graphics* **31**(1), 798–808.

Gustafson, A & RE Rice (2019). The effects of uncertainty frames in three science communication topics. *Science Communication* **41**(6), 679–706.

Hadjimichael, A, J Schlumberger & M Haasnoot (2024). Data visualisation for decision making under deep uncertainty: current challenges and opportunities. en. *Environmental Research Letters* **19**(11), 111011. (Visited on 11/11/2025).

Haupt, IA (1930). Tests for color-blindness: A survey of the literature with bibliography to 1928. *The Journal of General Psychology* **3**(2). Publisher: Routledge, 222–267. (Visited on 03/04/2026).

Heer, J & M Bostock (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: Association for Computing Machinery, pp.203–212. https://doi.org/10.1145/1753326.1753357.

Henderson & Velleman (1981). Building multiple regression models interactively. *Biometrics* **37**, 391–411.

Henry, L & H Wickham (2026a). *lifecycle: Manage the life cycle of your package functions*. R package version 1.0.5. https://CRAN.R-project.org/package=lifecycle.

Henry, L & H Wickham (2026b). *rlang: Functions for base types and core R and 'Tidyverse' features*. R package version 1.2.0. https://CRAN.R-project.org/package=rlang.

Hester, J & J Bryan (2024). *glue: Interpreted string literals*. R package version 1.8.0. https://CRAN.R-project.org/package=glue.

Hofmann, H, L Follett, M Majumder & D Cook (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics* **18**(12), 2441–2448. (Visited on 12/22/2022).

Hofmann, H, E Hare & GGobi Foundation (2026). *gglogo: Geom for logo sequence plots*. R package version 0.1.5, commit 2dc11eae5b50190684dd7e655969e7978c73ac81. https://github.com/heike/gglogo.

Huebner, M, W Vach & S le Cessie (2016). A systematic approach to initial data analysis is good research practice. *The Journal of Thoracic and Cardiovascular Surgery* **151**(1), 25–27.

Hullman, J (2016). Why evaluating uncertainty visualization is error prone. *ACM International Conference Proceeding Series* **24-October**, 143–151.

Hullman, J (2020). Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics* **26**(1), 130–139. eprint: 1908.01697.

Hullman, J & A Gelman (2021). Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 1–70.

Hullman, J, M Kay, YS Kim & S Shrestha (2018). Imagining replications: Graphical prediction discrete visualizations improve recall estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics* **24**(1), 446–456.

Hullman, J, X Qiao, M Correll, A Kale & M Kay (2019). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics* **25**(1), 903–913.

Hullman, J, P Resnick & E Adar (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS ONE* **10**(11).

Hyndman, R & G Athanasopoulos (2021). *Forecasting: principles and practice, 3rd edition*. Melbourne, Australia: OTexts. https://otexts.com/fpp3/.

Ibrekk, H & MG Morgan (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis* **7**(4), 519–529.

Ihaka, R (2003). Colour for presentation graphics. en. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Jung, PH, JC Thill & M Issel (2019). Spatial autocorrelation and data uncertainty in the American Community Survey: A critique. *International Journal of Geographical Information Science* **33**(6), 1155–1175. (Visited on 03/16/2026).

Kale, A, M Kay & J Hullman (2021). Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics* **27**(2), 272–282. eprint: 2007.14516.

Kale, A, F Nguyen, M Kay & J Hullman (2018). Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics* **25**(1), 892–902.

Kay, M (2019). *How much value should an uncertainty palette suppress if an uncertainty palette should suppress value? Statistical and perceptual perspectives*. https://doi.org/10.31219/osf.io/6xcnw.

Kay, M (2023). ggdist: Visualizations of distributions and uncertainty in the grammar of graphics. *IEEE Transactions on Visualization and Computer Graphics* **30**(1), 414–424.

Khizer, MA, U Ijaz, TA Khan, S Khan, T Liaqat, A Jamal, I Zahid, HG Shah & MA Zahid (2022). Smartphone color vision testing as an alternative to the conventional Ishihara booklet. *Cureus* **14**(10), e30747. (Visited on 03/04/2026).

Kim, YS, LA Walls, P Krafft & J Hullman (2019). A Bayesian cognition approach to improve data visualization. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. arXiv: 1901.02949.

Kinkeldey, C, AM MacEachren & J Schiewe (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *Cartographic Journal* **51**(4), 372–386.

Koo, H, DW Wong & Y Chun (2019). Measuring global spatial autocorrelation with data reliability information. *The Professional Geographer : the Journal of the Association of American Geographers* **71**(3), 551–565. (Visited on 03/16/2026).

Koonchanok, R, GY Tawde, GR Narayanasamy, S Walimbe & K Reda (2023). Visual belief elicitation reduces the incidence of false discovery. In: *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp.1–17.

Kuhnert, PM, DE Pagendam, R Bartley, DW Gladish, SE Lewis & ZT Bainbridge (2018). Making management decisions in the face of uncertainty: A case study using the Burdekin catchment in the Great Barrier Reef. *Marine and Freshwater Research* **69**(8), 1187–1200.

Kyveryga, PM (2019). On-farm research: Experimental approaches, analytical frameworks, case studies, and impact. *Agronomy Journal* **111**(6), 2633–2635. (Visited on 03/25/2026).

Lee, C, T Yang, GD Inchoco, GM Jones & A Satyanarayan (2021). Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox science online. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp.1–18.

Lenth, RV (2025). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.11.1. https://CRAN.R-project.org/package=emmeans.

Li, W, D Cook, E Tanaka & S VanderPlas (2024). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *Journal of Computational and Graphical Statistics*, 1–19. (Visited on 09/24/2024).

Lim, NJ, SA Brandt & S Seipel (2016). Visualisation and evaluation of flood uncertainties based on ensemble modelling. *International Journal of Geographical Information Science* **30**(2), 240–262.

Locke, S & L D'Agostino McGowan (2018). *datasauRus: Datasets from the Datasaurus Dozen*. R package version 0.1.4. https://CRAN.R-project.org/package=datasauRus.

Lucchesi, L & P Kuhnert (2020). *Vizumap: Visualizing uncertainty in spatial data*. https://lydialucchesi.github.io/Vizumap/.

Lucchesi, L, P Kuhnert & C Wikle (2021). Vizumap: an R package for visualising uncertainty in spatial data. *Journal of Open Source Software* **6**(59), 2409.

Lucchesi, LR & CK Wikle (2017). Visualizing uncertainty in areal data with bivariate choropleth maps, map pixelation and glyph rotation. *Stat* **6**(1), 292–302.

Luce, RD & W Edwards (1958). The derivation of subjective scales from just noticeable differences. *Psychological review* **65**(4), 222.

MacEachren, AM (1992). Visualizing uncertain information. *Cartographic Perspectives* (13), 10–19.

MacEachren, AM, A Robinson, S Hopper, S Gardner, R Murray, M Gahegan & E Hetzler (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* **32**(3). ISBN: 1523040054738, 139–160.

Maceachren, AM, RE Roth, J O'Brien, B Li, D Swingley & M Gahegan (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics* **18**(12), 2496–2505.

Majumder, M, H Hofmann & D Cook (2013). Validation of visual statistical inference, applied to linear models. en. *Journal of the American Statistical Association* **108**(503), 942–956. (Visited on 11/11/2025).

Mann, HB & A Wald (1943). On Stochastic Limit and Order Relationships. *The Annals of Mathematical Statistics* **14** (3), 217–226.

Manski, CF (2020). The lure of incredible certitude. *Economics and Philosophy* **36**(2), 216–245.

Mason, H, D Cook, S Goodwin, E Tanaka & S VanderPlas (2026a). *The noisy work of uncertainty visualisation research*. arXiv: 2411.10482 [cs.HC]. https://arxiv.org/abs/2411.10482.

Mason, H, D Cook, S Goodwin & S VanderPlas (2026b). *ggdibbler: Add uncertainty to data visualisations*. https://github.com/harriet-mason/ggdibbler.

Matejka, J & G Fitzmaurice (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, pp.1290–1294. https://doi.org/10.1145/3025453.3025912.

McNutt, A, G Kindlmann & M Correll (2020). Surfacing visualization mirages. en. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, pp.1–16. https://dl.acm.org/doi/10.1145/3313831.3376420 (visited on 09/05/2025).

Meng, XL (2014). A trio of inference problems that could win you a nobel prize in statistics (if you help fund it). *Past, Present, and Future of Statistical Science*, 537–562.

Meyer, MA, FR Broome & RHS Jr. (1975). Color statistical mapping by the U.S. Bureau of the Census. *The American Cartographer* **2**(2), 101–117. eprint: https://doi.org/10.1559/152304075784313250.

Moritz, D, D Fisher, B Ding & C Wang (2017). Trust, but verify: Optimistic visualizations of approximate queries for exploring big data. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp.2904–2915.

Müller, K & H Wickham (2025). *tibble: Simple data frames*. R package version 3.3.0. https://CRAN.R-project.org/package=tibble.

Ndlovu, A, H Shrestha & LT Harrison (2023). Taken by surprise? Evaluating how Bayesian surprise & suppression influences peoples' takeaways in map visualizations. In: *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, pp.136–140.

Neuwirth, E (2022). *RColorBrewer: ColorBrewer palettes*. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer.

O'Hara-Wild, M, M Kay, A Hayes & R Hyndman (2024). *distributional: Vectorised probability distributions*. R package version 0.5.0. https://CRAN.R-project.org/package=distributional.

O'Neill, O (2018). Linking trust to trustworthiness. *International Journal of Philosophical Studies* **26**(2), 293–300.

Olston, C & JD Mackinlay (2002). Visualizing data with bounded uncertainty. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS* **2002-Janua**, 37–40.

Otsuka, J (2023). *Thinking about atatistics: The philosophical foundations*. 1st. New York: Routledge, p. 204.

Padilla, L, H Hosseinpour, R Fygenson, J Howell, R Chunara & E Bertini (2022). Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Scientific Reports 2022 12:1* **12**(1), 1–14. (Visited on 04/15/2022).

Padilla, L, M Kay & J Hullman (2022). "Uncertainty visualization". In: *Computational Statistics in Data Science*. Ed. by WW Piegorsch, RA Levine, HH Zhang & TCM Lee. Hoboken, NJ: John Wiley & Sons. Chap. 22, pp.405–426.

Padilla, L, M Powell, M Kay & J Hullman (2021). Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology* **11**.

Padilla, L, I Ruginski & S Creem-Regehr (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications* **2**(1).

Pang, AT, CM Wittenbrink & SK Lodha (1997). Approaches to uncertainty visualization. *Visual Computer* **13**(8), 370–390.

Pebesma, E (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**(1), 439–446.

Pebesma, E & R Bivand (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://r-spatial.org/book/.

Pedersen, TL (2024). *tidygraph: A Tidy API for Graph Manipulation*. R package version 1.3.1. https://CRAN.R-project.org/package=tidygraph.

Pedersen, TL (2025a). *ggraph: An implementation of grammar of graphics for graphs and networks*. R package version 2.2.2. https://CRAN.R-project.org/package=ggraph.

Pedersen, TL (2025b). *patchwork: The composer of plots*. R package version 1.3.2. https://CRAN.R-project.org/package=patchwork.

Peña-Araya, V, CM Fontaine, X Wei, G Delpech & A Bezerianos (2025). Uncertainty in science is malleable. Advocating for user-agency in defining uncertainty in visualizations: a case study in geology. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp.1–18.

Pham, B, A Streit & R Brown (2009). "Visualization of information uncertainty: Progress and challenges". In: *Advanced Information and Knowledge Processing*. Vol. 36. Springer-Verlag London Ltd, pp.19–48.

Plutino, A, L Armellin, A Mazzoni, R Marcucci & A Rizzi (2023). Aging variations in Ishihara test plates. en. *Color Research & Application* **48**(6), 721–734. (Visited on 03/04/2026).

Potter, K, J Kniss, R Riesenfeld & CR Johnson (2010). Visualizing summary statistics and uncertainty. *Computer Graphics Forum* **29**(3), 823–832.

Reda, K & DA Szafir (2021a). Rainbows Revisited: Modeling Effective Colormap Design for Graphical Inference. eng. *IEEE transactions on visualization and computer graphics* **27**(2), 1032–1042.

Reda, K & DA Szafir (2021b). Rainbows revisited: Modeling effective colormap design for graphical inference. en. *IEEE Transactions on Visualization and Computer Graphics* **27**(2), 1032–1042. (Visited on 02/23/2026).

Robinson, D, A Hayes, S Couch & E Hvitfeldt (2026). *broom: Convert statistical objects into tidy tibbles*. R package version 1.0.12. https://CRAN.R-project.org/package=broom.

Robinson, EA, R Howard & S VanderPlas (2023). 'You Draw It': Implementation of visually fitted trends with r2d3. *Journal of Data Science* **21**(2), 281–294.

Roy Chowdhury, N, D Cook, H Hofmann, M Majumder, EK Lee & AL Toth (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Computational Statistics* **30**(2), 293–316. (Visited on 12/28/2022).

Sanyal, J, S Zhang, G Bhattacharya, P Amburn & RJ Moorhead (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics* **15**(6), 1209–1218.

Sarma, A, S Guo, J Hoffswell, R Rossi, F Du, E Koh & M Kay (2023). Evaluating the use of uncertainty visualisations for imputations of data missing at random in scatterplots. *IEEE Transactions on Visualization and Computer Graphics* **29**(1), 602–612.

Sarma, A, X Pu, Y Cui, M Correll, ET Brown & M Kay (2024). Odds and insights: Decision quality in exploratory data analysis under uncertainty. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3641995.

Satyanarayan, A, D Moritz, K Wongsuphasawat & J Heer (2016). Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics* **23**(1), 341–350.

Savelli, S & S Joslyn (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology* **27**(4), 527–541.

Savvides, R, A Henelius, E Oikarinen & K Puolamäki (2019). Significance of patterns in data visualisations. en. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: Association for Computing Machinery, pp.1509–1517. https://doi.org/10.1145/3292500.3330994.

Simons, DJ & DT Levin (1997). Change blindness. *Trends in Cognitive Sciences* **1** (7), 261–267.

Smart, S & DA Szafir (2019). Measuring the separability of shape, size, and color in scatterplots. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14.

Smemoe, CM (2004). *Floodplain risk analysis using flood probability and annual exceedance probability maps*. Brigham Young University.

Spiegelhalter, D (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application* **4**, 31–60.

Stauffer, R, GJ Mayr, M Dabernig & A Zeileis (2009). Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society* **96**(2), 203–216.

Strochak, S, K Ueyama & A Williams (2024). *urbnmapr: State and county shapefiles in sf and tibble format*. R package version 0.0.0.9002. https://github.com/UrbanInstitute/urbnmapr.

Sumner, M (2021). *ozmaps: Australia Maps*. R package version 0.4.5. https://CRAN.R-project.org/package=ozmaps.

Swihart, BJ, B Caffo, BD James, M Strand, BS Schwartz & NM Punjabi (2010). Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology* **21**(5), 621–625.

Tamura, S, Y Okamoto, S Nakagawa, T Sakamoto, M Ando & Y Shigeri (2017). Light wavelengths of LEDs to improve the color discrimination in Ishihara test and Farnsworth Panel D-15 test for deutans. en. *Color Research & Application* **42**(4), 424–430. (Visited on 03/04/2026).

Thomson, J, E Hetzler, A MacEachren, M Gahegan & M Pavel (2005). A typology for visualizing uncertainty. *Visualization and Data Analysis 2005* **5669**(March 2005), 146.

Tierney, N (2020). *ishihara*. https://github.com/njtierney/ishihara.

Tierney, N & D Cook (2023). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software* **105**(7), 1–31. eprint: 1809.02264.

Tukey, JW et al. (1977). *Exploratory data analysis*. Vol. 2. Springer.

UNESCO (2026). *International Standard Classification of Education - ISCED | Institute for Statistics (UIS)*. en. https://www.uis.unesco.org/en/methods-and-tools/isced (visited on 03/30/2026).

Vanderplas, S, D Cook & H Hofmann (2020). Testing statistical charts: What makes a good graph? *Annual Review of Statistics and Its Application* **7**(1), 61–88.

VanderPlas, S & H Hofmann (2015). Signs of the sine illusion—why we need to care. *Journal of Computational and Graphical Statistics* **24**(4), 1170–1190.

VanderPlas, S & H Hofmann (2017). Clusters beat trend!? Testing feature hierarchy in statistical graphics. *Journal of Computational and Graphical Statistics* **26**(2), 231–242.

VanderPlas, S, C Röttger, D Cook & H Hofmann (2021). Statistical significance calculations for scenarios in visual inference. *Stat* **10**(1), e337.

Venables, WN & BD Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Waldhör, T (1996). The spatial autocorrelation coefficient Moran's I under heteroscedasticity. *Stat Med* **15**(7-9), 887–892. (Visited on 03/18/2026).

Walker, WE, P Harremoes, J Rotmans, JP Van Der Sluijs, MBA Van Asselt, P Janssen & MP Krayer Von Krauss (2003). Defining uncertainty. *Integrated Assessment* **4**(1), 5–17.

Waller, LA (2024). Maps: A Statistical View. *Annual Review of Statistics and its Application* **11**, 75–96.

Wallsten, TS, DV Budescu, I Erev & A Diederich (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making* **10**(3), 243–268.

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

Wickham, H (2019). *Advanced R*. Chapman and Hall/CRC. https://adv-r.hadley.nz.

Wickham, H (2023). *conflicted: An alternative conflict resolution strategy*. R package version 1.2.0. https://CRAN.R-project.org/package=conflicted.

Wickham, H, M Averick, J Bryan, W Chang, L D'Agostino McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo & H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, D Cook, H Hofmann & A Buja (2010). Graphical inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics* **16**, 973–979.

Wickham, H, R François, L Henry, K Müller & D Vaughan (2023). *dplyr: A grammar of data manipulation*. R package version 1.1.4. https://CRAN.R-project.org/package=dplyr.

Wickham, H & H Hofmann (2011). Product plots. *IEEE Transactions on Visualization and Computer Graphics* **17**(12), 2223–2230.

Wickham, H, M Lawrence, D Cook, A Buja, H Hofmann & DF Swayne (2009). The plumbing of interactive graphics. *Computational Statistics* **24**(2), 207–215.

Wickham, H, TL Pedersen & D Seidel (2025). *scales: Scale functions for visualization*. R package version 1.4.0. https://CRAN.R-project.org/package=scales.

Wickham, H, D Vaughan & M Girlich (2025). *tidyr: Tidy messy data*. R package version 1.3.2. https://CRAN.R-project.org/package=tidyr.

Wilkinson, L (2005). *The grammar of graphics*. Berlin, Heidelberg: Springer-Verlag.

Wu, Y, Z Guo, M Mamakos, J Hartline & J Hullman (2023). *The rational agent benchmark for data visualization*. arXiv: 2304.03432 [cs.HC].

Xiao, J (2021). Spatial aggregation entropy: A heterogeneity and uncertainty metric of spatial aggregation. eng. *Annals of the American Association of Geographers* **111**(4). Publisher: Taylor & Francis Ltd, 1236–1252. (Visited on 02/20/2026).

Yang, F, M Cai, C Mortenson, H Fakhari, AD Lokmanoglu, J Hullman, S Franconeri, N Diakopoulos, EC Nisbet & M Kay (2023). Swaying the public? Impacts of election forecast visualizations on emotion, trust, and intention in the 2022 US midterms. *IEEE Transactions on Visualization and Computer Graphics* **30**(1), 23–33.

Yu, G (2025). *shadowtext: Shadow text grob and layer*. R package version 0.1.6. https://CRAN.R-project.org/package=shadowtext.

Zhang, M & DKJ Lin (2022). Visualization for interval data. en. *Journal of Computational and Graphical Statistics* **31**(4), 960–975. (Visited on 05/07/2025).

Zhao, J, Y Wang, MV Mancenido, EK Chiou & R Maciejewski (2023). Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics* **30**(7), 4093–4107.

Zhu, H (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax*. R package version 1.4.0. https://CRAN.R-project.org/package=kableExtra.

Chapter 6

Appendix A — Supplementary Material for "Colour Blinded by the Noise"

6.1 Full app screenshots

Screenshots from the app, in the order that the participants would experience them.

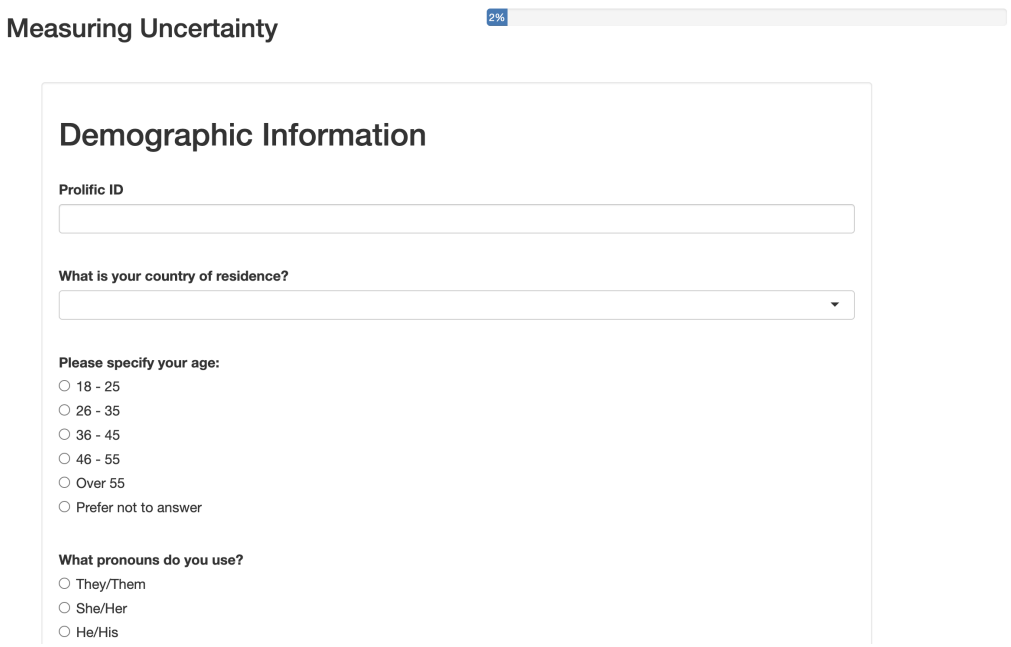


Figure 6.1: *Demographics 1*

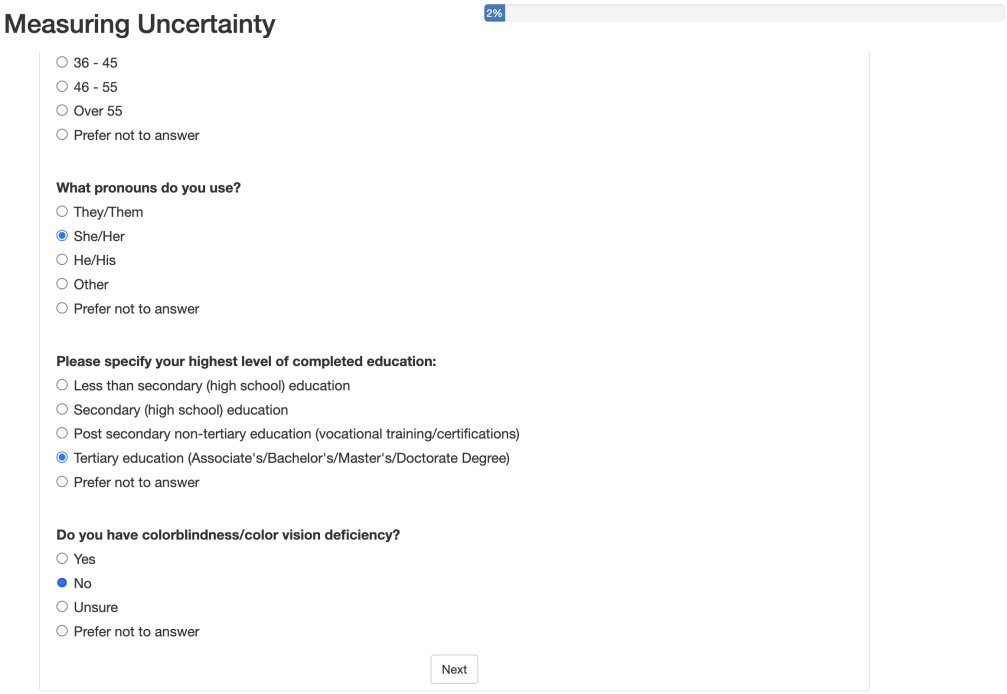


Figure 6.2: *Demographics 2*

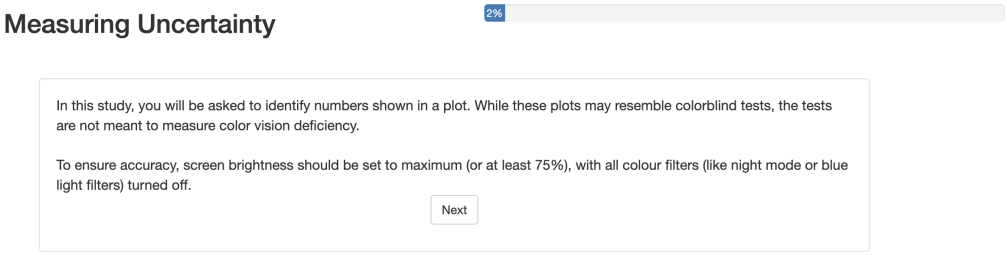


Figure 6.3: *Additional instructions*

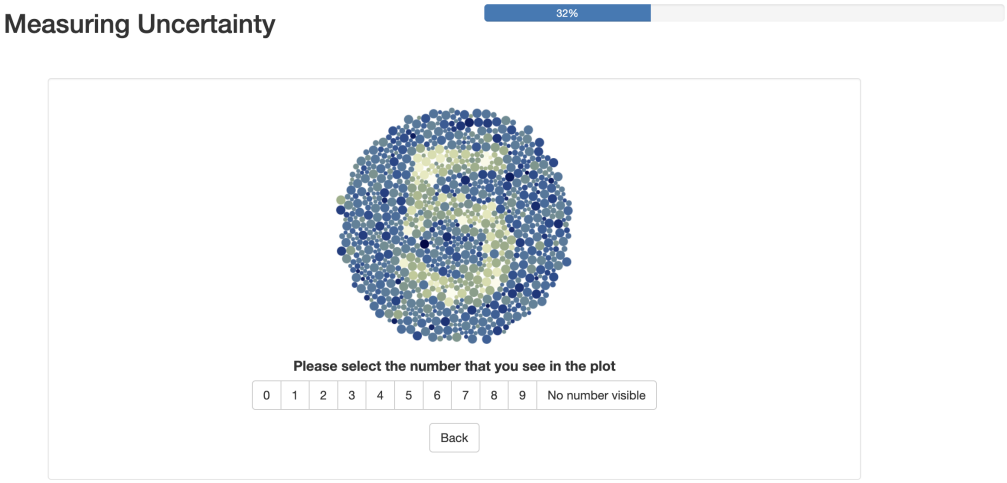


Figure 6.4: *Stimuli example A*

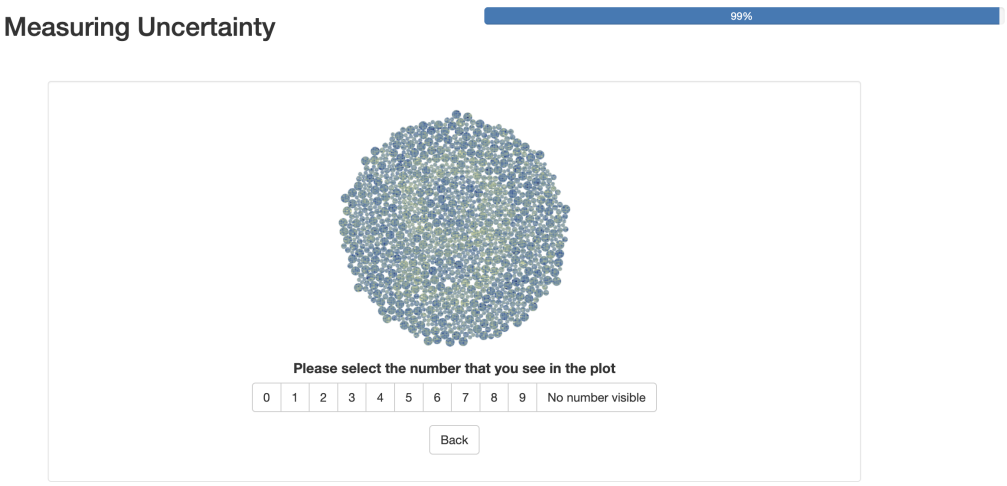


Figure 6.5: *Stimuli example B*

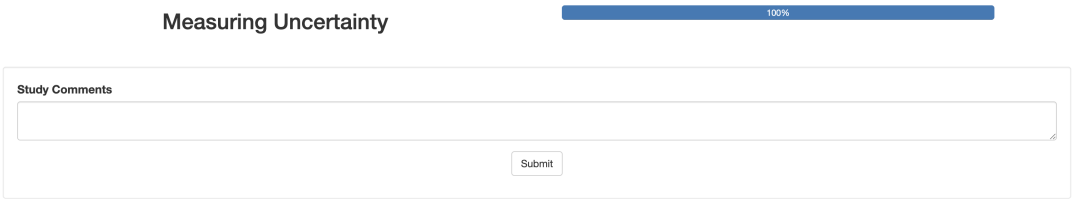


Figure 6.6: *End of study comments*

6.2 Confusion matrix of numbers

To understand the influence of the number displayed on the participant response, we can look at a confusion table of the number responses (Table 6.1).

For the cases where there actually was a number visible, we can see that participants typically got the number right, or selected no number visible, rather than making an incorrect guess. When there was no number, participants seemed to guess 3 more often than the other numbers.

There are also a few numbers that participants seemed to get confused more often than others. If we focus on the cases where 50 or more incorrect guesses were made, we can see that 3, 6 and 9 were frequently reported as an 8, and 5 was frequently reported as a 6. This makes sense as we could consider the dots that make up 3, 6, and 9 to be a subset of those covered by 8, with a similar relationship existing between 5 and 6. Interestingly, the converse is not true. That is, 8 was not mistaken for a 3, 6, or 9, and 6 was not mistaken for 5. This seems to suggest that, when participants could not make out the number with confidence, they seemed to have a tendency to add in structure that wasn't there, rather than miss structure that was there.

Table 6.1: *A confusion matrix of the number participants gave, alongside the true number in the plot. The off-diagonal elements are incorrect responses, with any concentration of numbers indicating values that are frequently mistaken for one another: 8 for 3, 6 and 9, and 6 for 5 are the most common.*

Correct Number	Number Selected										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N**
0	1,091	6	0	7	1	4	32	2	17	15	207
1	6	991	2	11	31	2	2	4	1	0	324
2	1	5	1,178	7	4	0	2	3	10	1	165
3	2	0	1	953	1	8	3	2	104	18	228
4	0	0	3	3	1,073	2	2	2	0	0	282
5	1	3	0	21	2	1,000	82	0	24	1	258
6	3	0	2	9	1	13	1,019	2	79	5	220
7	0	6	8	2	2	2	5	1,038	1	2	306
8	4	1	2	25	0	16	24	3	1,025	5	251
9	12	0	6	25	1	10	5	3	61	1,056	227
N	5	6	9	27	13	9	7	11	11	5	3,321

Table 6.2: *Numbers that were most commonly identified as 'no number visible', ordered from most to least frequent. The order roughly coincides, inversely, with the number of 'dots' that make up the number, suggesting that numbers constructed with fewer dots are harder to identify.*

Selected	Correct	Total	Dots
No number visible	1	324	107
No number visible	7	306	150
No number visible	4	282	181
No number visible	5	258	203
No number visible	8	251	236
No number visible	3	228	219
No number visible	9	227	222
No number visible	6	220	238
No number visible	0	207	211
No number visible	2	165	221

Additionally, the number 1 (and possibly 7) were more frequently reported as no number visible

relative to the other numbers (Table 6.2). This might be due to those numbers having less circles in the "number" group relative to the "background" group, as we can see the top 3 numbers reported as "no number" also had the lowest number of "number" dots relative to those in the background. However, this trend seems to drop off after 1, 7, and 4.

6.3 Duration Analysis

The trend in the amount of time participants spent on each question seems to align with the probability of getting the question correct. Figure 6.7 shows the median amount of seconds spent on each D, V, and plot type. Unsurprisingly, the most amount of time across all plots was on the $D = 1$ case, when the signal was not particularly strong. The pixel and transparency maps have a lower triangle of easy to see numbers, that become harder to extract as both D decreases and V increases. It is also clear that participants rarely spent more than a few seconds on each plot. This also highlights that, by making uncertainty something that should be visibly seen, a well designed uncertainty visualisation can be correctly within a few seconds.

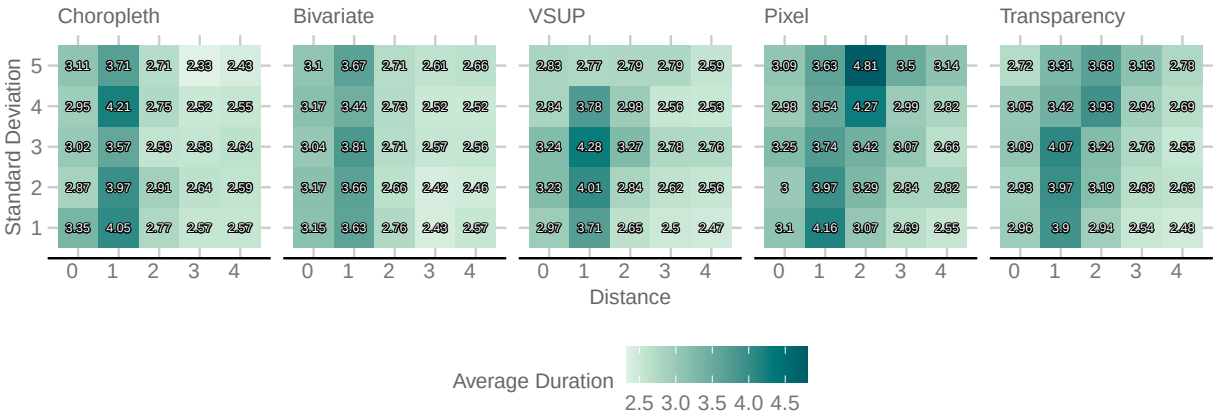


Figure 6.7: *A heatmap showing the median duration accross each D, V, and plot type.*

6.4 Demographic Analysis

The demographic analysis indicates no relationship between the demographic details and the proportion of correct responses.

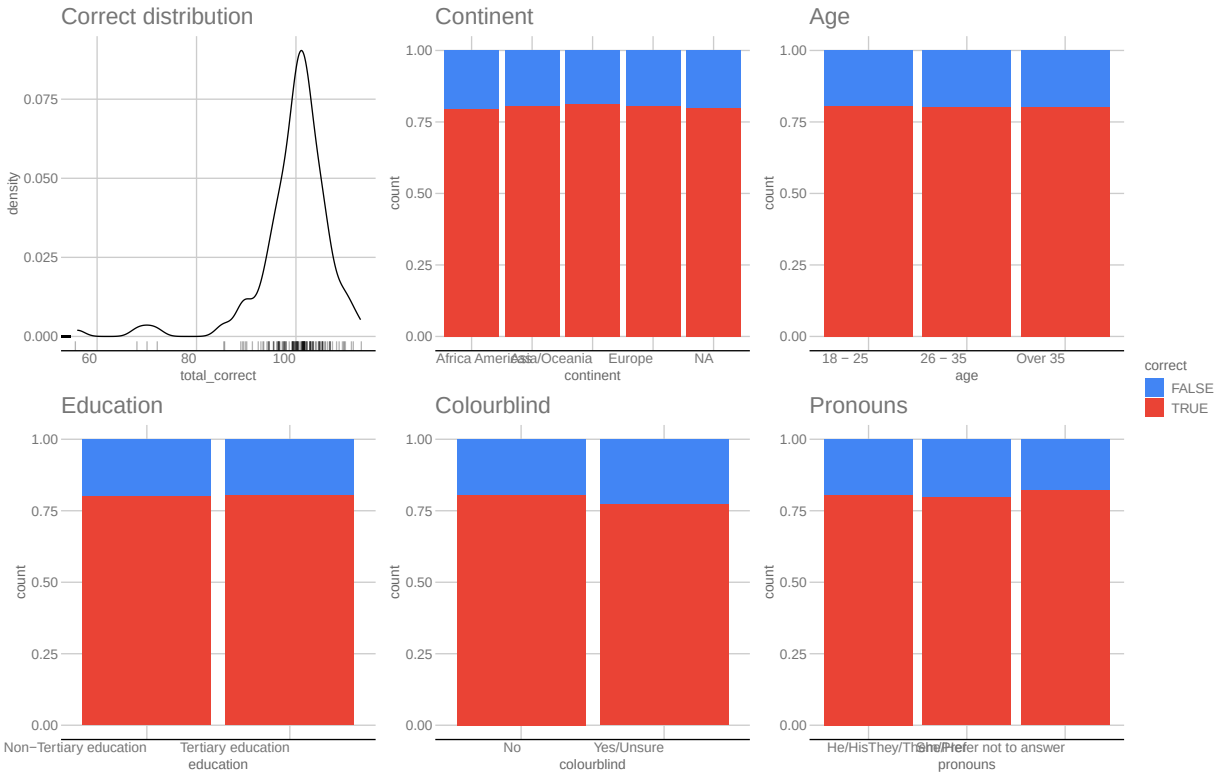


Figure 6.8: *Plots showing the overall distribution of correct answers from the participants, as well as the relationship betweeen the five demographic responses and the probability of correctly identifying the number in the stimuli.*

6.5 Additional model comparison results

The distance-based results as well as all pairwise comparisons, as mentioned in the main text.

Table 6.3: *Results for Standard Deviation Trend by Plot Type at Distance = 1*

plot_type	V.trend	SE	z.ratio	p.value
Choropleth	-0.026	0.056	-0.460	0.646
Bivariate	-0.039	0.055	-0.701	0.484
VSUP	-0.577	0.059	-9.712	0.000
Pixel	-0.515	0.063	-8.234	0.000
Transparency	-0.538	0.066	-8.216	0.000

Table 6.4: *Results for Standard Deviation Trend by Plot Type at Distance = 2*

plot_type	V.trend	SE	z.ratio	p.value
Choropleth	0.110	0.097	1.132	0.258
Bivariate	-0.042	0.082	-0.511	0.609
VSUP	-1.849	0.077	-23.942	0.000
Pixel	-0.639	0.047	-13.731	0.000
Transparency	-0.581	0.051	-11.454	0.000

Table 6.5: *Results for Standard Deviation Trend by Plot Type at Distance = 3*

plot_type	V.trend	SE	z.ratio	p.value
Choropleth	0.245	0.190	1.288	0.198
Bivariate	-0.045	0.160	-0.278	0.781
VSUP	-3.121	0.139	-22.458	0.000
Pixel	-0.762	0.077	-9.870	0.000
Transparency	-0.624	0.092	-6.753	0.000

Table 6.6: *Results for Standard Deviation Trend by Plot Type at Distance = 4*

plot_type	V.trend	SE	z.ratio	p.value
Choropleth	0.381	0.290	1.316	0.188
Bivariate	-0.047	0.245	-0.193	0.847
VSUP	-4.394	0.209	-21.040	0.000
Pixel	-0.886	0.124	-7.154	0.000
Transparency	-0.667	0.149	-4.475	0.000

Table 6.7: *Results for Standard Deviation Trend by Plot Type at Distance = 1*

contrast	estimate	SE	z.ratio	p.value
Choropleth - Bivariate	0.013	0.078	0.166	1.000
Choropleth - VSUP	0.551	0.081	6.763	0.000
Choropleth - Pixel	0.490	0.084	5.847	0.000
Choropleth - Transparency	0.513	0.086	5.965	0.000
Bivariate - VSUP	0.538	0.081	6.639	0.000
Bivariate - Pixel	0.476	0.083	5.715	0.000
Bivariate - Transparency	0.500	0.086	5.835	0.000
VSUP - Pixel	-0.062	0.086	-0.716	0.953
VSUP - Transparency	-0.038	0.088	-0.435	0.993
Pixel - Transparency	0.023	0.090	0.257	0.999

Table 6.8: *Results for Standard Deviation Trend by Plot Type at Distance = 2*

contrast	estimate	SE	z.ratio	p.value
Choropleth - Bivariate	0.153	0.128	1.190	0.757
Choropleth - VSUP	1.960	0.125	15.674	0.000
Choropleth - Pixel	0.749	0.109	6.897	0.000
Choropleth - Transparency	0.693	0.111	6.265	0.000
Bivariate - VSUP	1.807	0.113	15.974	0.000
Bivariate - Pixel	0.596	0.095	6.289	0.000
Bivariate - Transparency	0.540	0.097	5.563	0.000
VSUP - Pixel	-1.211	0.090	-13.515	0.000
VSUP - Transparency	-1.267	0.092	-13.766	0.000
Pixel - Transparency	-0.056	0.069	-0.821	0.924

Table 6.9: *Results for Standard Deviation Trend by Plot Type at Distance = 3*

contrast	estimate	SE	z.ratio	p.value
Choropleth - Bivariate	0.292	0.253	1.157	0.776
Choropleth - VSUP	3.369	0.238	14.130	0.000
Choropleth - Pixel	1.009	0.208	4.843	0.000
Choropleth - Transparency	0.873	0.215	4.068	0.000
Bivariate - VSUP	3.076	0.214	14.361	0.000
Bivariate - Pixel	0.716	0.180	3.972	0.001
Bivariate - Transparency	0.580	0.188	3.094	0.017
VSUP - Pixel	-2.360	0.159	-14.864	0.000
VSUP - Transparency	-2.496	0.167	-14.939	0.000
Pixel - Transparency	-0.136	0.121	-1.124	0.794

Table 6.10: *Results for Standard Deviation Trend by Plot Type at Distance = 4*

contrast	estimate	SE	z.ratio	p.value
Choropleth - Bivariate	0.432	0.385	1.122	0.795
Choropleth - VSUP	4.778	0.361	13.238	0.000
Choropleth - Pixel	1.268	0.319	3.975	0.001
Choropleth - Transparency	1.052	0.330	3.191	0.012
Bivariate - VSUP	4.346	0.325	13.357	0.000
Bivariate - Pixel	0.836	0.278	3.003	0.022
Bivariate - Transparency	0.620	0.291	2.133	0.206
VSUP - Pixel	-3.509	0.243	-14.453	0.000
VSUP - Transparency	-3.725	0.257	-14.486	0.000
Pixel - Transparency	-0.216	0.195	-1.107	0.803