Improving the robustness and reliability of population-based global biodiversity indicators

Shawn Arthur Dove

A thesis submitted for the degree of:

Doctor of Philosophy

University College London

September 2022

Primary supervisor:
Dr. David Murrell

Secondary supervisor:
Dr. Robin Freeman

Tertiary supervisor:
Dr. Monika Böhm

Institution:
Centre for Biodiversity and Environment Research,
Department of Genetics, Evolution and Environment,
University College London,
London, WC1E 6BT

I, Shawn Dove, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The current global biodiversity crisis is complicated by a data crisis. Reliable tools are needed to guide scientific research and conservation policy decisions, but the data underlying those tools is incomplete and biased. For example, the Living Planet Index (LPI) tracks the changing status of global vertebrate biodiversity, but gaps, biases and quality issues plague the aggregated data used to calculate trends. Unfortunately, we have little understanding of how reliable biodiversity indicators are. In this thesis I develop a suite of tools to assess and improve the reliability of trends in the LPI and similar indicators. First, I explore distance measures as a flexible toolset for comparing time series and trends. I test distance measures for properties related to time series comparisons and rate their relative sensitivities, then expand the results into a framework for choosing an appropriate distance measure for any time series comparison task in ecology. I use the framework to select an appropriate metric for determining trend accuracy. Second, I construct a model of trend reliability from accuracy measurements of sampled trend replicates calculated from artificially generated time series datasets. I apply the model to the LPI to reveal that the majority of trends need more data to be considered reliable, particularly across the global south, and for reptiles and amphibians everywhere. Finally, I develop a method to account for sampling error and serial correlation in confidence intervals of indicators that use aggregated abundance data from different sources. I show that the new method results in more robust and accurate confidence intervals across a wide range of dataset parameters, without reducing trend accuracy. I also apply the method to the LPI to reveal that the current method used by the LPI results in inaccurate and overly wide confidence intervals.

Impact Statement

Biodiversity is the key to healthy ecosystems that provide essential services to sustain human life. Medicines, clean water, breathable air, a functioning climate, food, and many other life-sustaining services depend on biodiversity. But biodiversity is increasingly under threat from anthropogenic forces. To prevent its continued decline, there is a growing need for reliable data on the changing status of global biodiversity. Governments, NGOs, and scientists rely upon biodiversity indicators to provide this information to make decisions about policies, conservation, and research, but we have little understanding of their reliability. The work presented in this thesis will greatly help to improve our understanding of the state of knowledge on biodiversity trends, as well as pinpoint data deficient taxa and regions on which to focus research efforts. More importantly, it provides flexible methodological tools to aid and enable further investigations and discoveries.

The selection method presented in Chapter 2 is much more flexible and broader than previous methods, making use of both existing and new research to aid scientists in choosing an appropriate distance measure for any time series comparison task. The work is presented in an ecological context and will expand the scope for using distance measures to answer ecological questions; however, it is also broadly applicable and may be useful or inspirational to scientists from many other disciplines. The modelling approach presented in Chapter 3 will move forward the field of indicator assessment by providing a flexible modelling framework that can be used to predict indicator accuracy instead of just testing responses to modelled scenarios. I demonstrate its potential by using it to assess the level of data deficiency underlying all regional taxonomic group trends in the Living Planet Index and quantifying the number of populations needed to overcome data deficiency for each trend. This work can be used to direct data-aggregation efforts for the Living Planet Index, and to focus future data-collection efforts where they are most needed. It will also provide policy makers with information on which trends are reliable before making consequential decisions. Chapter 4 introduces a new approach to calculating confidence intervals for biodiversity indicators that use aggregated data from multiple sources. Current methods are unable to produce reliable confidence intervals as they fail to take sampling error and serial correlation into account. The new approach solves these issues, which I demonstrate using

the Living Planet Index. Both the modelling framework and the method of calculating confidence intervals can be easily adapted to other biodiversity indicators.

All the code I produced for my projects is available online, open source. I have already adapted parts of the code from Chapter 3 for use in a real-time online-accessible version of the Species Information Index. I also used my understanding of biodiversity trends and indicators to contribute to a manuscript (currently on bioRxiv but intended to be submitted for publication) that used counterfactual analysis to reveal the impact of conservation actions on populations in the Living Planet Database.

Acknowledgements

The past four years have been at times interminably long, and at times far too brief. But through it all some very special people have supported me and my work. First, my supervisors, David Murrell, Monika Böhm, and Robin Freeman. Dave, you've always been supportive, available, and kept my best interests in mind. Monni, you not only kept things light with your wonderful sense of humour, but you managed to do all the supervisory things too. Despite moving countries partway through, you have remained a valuable friend and supportive supervisor. Rob, you never cease to amaze me with how quickly you can put my work into context and come up with valuable and insightful suggestions on how to make it better.

A huge thanks to my colleagues and very good friends, Sean Jellesmark, Gonzalo Albaladejo Robles, and Bouwe Reijenga. You guys were there all the way, listening to my problems and rants, providing helpful scientific suggestions, endless jokes to keep me smiling, and incredibly valuable emotional support all the way through.

Thank you to all my colleagues at Inspire4Nature. I will always be sad that covid put a stop to our regular meetings, which were tremendously valuable and a lot of fun. You are all brilliant and wonderful people. Special thanks to Ana Rodriguez for making it all happen and being an inspiring mentor, Pauline Roger for so brilliantly taking care of all planning and admin, and Tim Blackburn for dealing patiently with all the admin woes that Sean, Gonzalo, and myself unwittingly visited upon you.

To all my colleagues at CBER, you folks are amazing, fun, kind, and lovely people. It's been a fun ride.

Thank you to my parents, Ray and Dawn, who despite living on a different continent, were always supportive, great listeners, and amazing people. Last but most importantly, my wife, Andrea, and my children, Tilly and Sammy. You have kept my life interesting, fun, and full of love and purpose, and have always been there for me. I love you.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766417.

Thesis outline of contents and collaborators

Chapter 1

Introduction.

In this chapter, I introduce the concepts and focus of this thesis, and highlight important gaps in the field of biodiversity indicators that this thesis aims to fill.

Chapter 2

Selecting appropriate distance measures to compare ecological time series

In this chapter, I present an objective method to select the most appropriate distance measures for any ecological research that involves comparing time series. The work was conducted in collaboration with Monika Böhm, Robin Freeman, Sean Jellesmark and David J. Murrell. I conceived the study with input from DJM, RF, and MB. I produced all the code, synthetic data, and figures, and designed and conducted all analyses. Wading bird indices, along with percentage improvement and t-test results, were produced by SJ. Time series used for uncontrolled testing were obtained freely from the UCR Time-Series Classification Archive (Dau et al., 2019). I wrote the paper, with critical feedback from all authors. This work is available on bioRxiv under the title '*A user-friendly guide to using distance measures to compare time series in ecology*' and is under review at *Methods in Ecology and Evolution*. Original code is available online at https://github.com/shawndove/Trend_compare. Wading bird indices will be archived at Zenodo upon acceptance of the manuscript for publication. Datasets from the UCR Time Series Classification Archive are available at https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Chapter 3

Reliability and data deficiency in global vertebrate biodiversity trends

In this chapter, I simulate datasets of population abundance time series and use them to build a model of accuracy for sampled trends. I then use the model to analyse the data underlying each regional taxonomic group trend in the Living Planet Index, assign reliability ratings to each trend, and determine the number of additional population time series

needed to reach a threshold of reliability. This work was carried out in collaboration with MB, RF, Louise McRae, and DJM. I conceived the study with input from DJM, RF, and MB. I produced all the code, synthetic data, and figures, and designed and conducted all analyses. I wrote the paper, with critical feedback from all authors. Original code is available at https://github.com/shawndove/DD_LPI. Some of the time series from the Living Planet Database are not available for public use, but a limited public version of the database is available at https://www.livingplanetindex.org/data_portal. Simulated datasets used for analysis comprise more than 250 GB and are thus too large to archive online but can be approximately reproduced using the available code and parameter settings.

Chapter 4

Accounting for sampling and measurement error in aggregated abundance-based biodiversity indicators

In this chapter, I introduce a new method of calculating confidence intervals for aggregated abundance-based biodiversity indicators that accounts for sampling error and serial correlation. The work was carried out in collaboration with MB, RF, and DJM. RF, DJM, and I jointly conceived the study. RF provided the GAM resampling code. I wrote the rest of the code, produced the synthetic data and figures, and designed and conducted all analyses. I wrote the paper, with critical feedback from all authors. Original code is available at https://github.com/shawndove/LPI_Sampling_Error. Some of the time series from the Living Planet Database are not available for public use, but a limited public version of the database is available at https://www.livingplanetindex.org/data_portal. Simulated datasets used for analysis comprise more than 200 GB and are thus too large to archive online but can be approximately reproduced using the available code and parameter settings.

Chapter 5

Discussion and synthesis

In this chapter, I evaluate the key findings and methodological contributions of this thesis and place them into context. I also discuss some of the limitations of the work, and explore challenges and future directions of biodiversity indicator research.

Appendices

Appendix 1

Supplementary materials for Chapter 2

Appendix 2

Supplementary materials for Chapter 3

Appendix 3

Supplementary materials for Chapter 4

Table of Contents

List of main text figures 14
List of main text tables 15
Chapter 1: Introduction 16
1.1. Measuring Biodiversity 17
1.2. Types of Biodiversity Indicators 19
1.3. Data Deficiency in Biodiversity Indicators 21
1.4. Assessment of Biodiversity Indicators 22
1.5. Comparing Trends of Biodiversity Indicators 24
1.6. Reliability in Biodiversity Indicators 25
1.7. Uncertainty in Biodiversity Indicators 26
1.8. Conclusion 28
1.9. Bibliography 28
Chapter 2: Selecting Appropriate Distance Measures to Compare Ecological Time Series 39
2.1. Abstract 39
2.2. Introduction 40
2.3. Methods 43
2.3.1. Metric properties (adapted from McCune & Grace, 2002) 43
2.3.2. Value-based properties 44
2.3.3. Time-based properties 46
2.3.4. Other properties 48
2.3.5. Metric properties tests 49
2.3.6. Time-based and value-based properties tests 50
2.3.7. Controlled testing 51
2.3.8. Correlation between distance measures 53
2.3.9. Uncontrolled testing 54
2.3.10. Selection process 55
2.3.11. Real-world example dataset 56
2.4. Results 56
2.4.1. Metric test results 56
2.4.2. Sensitivity test results 59
2.4.3. Time-based invariances and other test results 61
2.4.4. Selection process 64

2.5. Discussion 72
2.6. Conclusion 76
2.7. Bibliography 78
Chapter 3: How much data do we need? Reliability and data deficiency in global vertebrate biodiversity trends 82
3.1. Abstract 82
3.2. Introduction 82
3.3. Material and Methods 85
3.3.1. Synthetic data generation 87
3.3.2. Observation error 89
3.3.3. Data degradation 90
3.3.4. Sampling 90
3.3.5. Calculation of sampled trends 91
3.3.6. Calculation of the 'true' trend 92
3.3.7. Comparison of trends 92
3.3.8. Generation of datasets 93
3.3.9. Multiple regression model 93
3.3.10. Model validation 93
3.3.11. Maximum trend deviation value 94
3.3.12. Minimum sample size for regional taxonomic groups 96
3.3.13. Assigning reliability ratings to regional taxonomic groups 97
3.3.14. Correlations between reliability rating and LPI relative weighting 97
3.3.15. Modelling potential solutions 97
3.3.16. Coding and data 97
3.4. Results 98
3.4.1. Regression model 98
3.4.2. Maximum trend deviation value 98
3.4.3. Minimum sample size 99
3.4.4. Trend reliability 100
3.4.5. Modelling potential solutions 104
3.5. Discussion 105
3.6. Conclusion 110
3.7. Bibliography 112

Chapter 4: Accounting for sampling and measurement error in aggregated abundance-based biodiversity indicators....116
4.1. Abstract....116
4.2. Abbreviations....116
4.3. Introduction....117
4.4. Material and Methods....120
4.4.1. Synthetic data generation, observation error, data degradation, and sampling 120
4.4.2. Calculation of sampled trends for GRRE method....121
4.4.3. Calculation of sampled trends using the LPI (GC) method....122
4.4.4. Calculation of sampled trends using a modified LPI (GO) method....122
4.4.5. Calculation of the 'true' trend & trend comparison....122
4.4.6. Confidence intervals for sampled trends using the GRRE method....123
4.4.7. Confidence intervals for sampled trends by bootstrapping the species rates of change (GC and GO methods)....123
4.4.8. Percentage of 'true' trend captured within confidence interval of sampled trends (capture percentage)....124
4.4.9. Mean normalized width of confidence intervals of sampled trends....124
4.4.10. Comparison of confidence interval methods....124
4.5. Results....125
4.6. Discussion....131
4.7. Conclusion....135
4.8. Bibliography....136
Chapter 5: Discussion and Synthesis....140
5.1. Distance measures....141
5.2. Modelling approach....142
5.3. Populations vs species....145
5.4. Challenges, remaining questions, and the future....147
5.5. Conclusions....151
5.6. Bibliography....152
Appendices....159
Appendix 1: Supplementary materials for Chapter 2....160
S2.1. Descriptions and formulas for selected distance measures....160
S2.2. Distance measure properties....170
S2.3. Uncontrolled testing....172

S2.4. Metric test results 174
S2.5. Controlled test results 175
S2.5.1. Sensitivity tests 175
S2.5.2. Time-based invariances and other tests 176
S2.5.3. Pairwise correlations between distance measures 177
S2.6. Uncontrolled test results 178
S2.7. Example dataset results 184
S2.8. Speeding up DTW 185
S2.9. Plots of controlled test results for all distance measures 186
S2.10. Tables of controlled test results for all distance measures 228
S2.11. Plots of wading bird rankings for all distance measures 271
S2.12. Bibliography for Appendix 1 313
Appendix 2: Supplementary materials for Chapter 3 315
Appendix 3: Supplementary materials for Chapter 4 321

List of main text figures

Figure 2.1. Illustration of time series distortions. 48
Figure 2.2. Illustration of antiparallelism bias. 49
Figure 2.3. Examples of time series used for uncontrolled testing. 55
Figure 2.4. Metric test results for 42 distance measures. 58
Figure 2.5. Sensitivity test results for 42 distance measures. 60
Figure 2.6. Test results for antiparallelism bias, non-positive value handling, and time-related invariances for 42 distance measures. 62
Figure 2.7. Reserve and counterfactual trends for five wading bird species that breed on RSPB lowland wet grassland reserves in the UK. 63
Figure 2.8. Comparative rankings of conservation impact on five wading bird species. 67
Figure 2.9. Comparative rankings of conservation impact on unsmoothed trends of five wading bird species. 68
Figure 2.10. Comparative rankings of conservation impact on smoothed trends for five wading bird species. 69
Figure 2.11. Decision tree to aid in choosing a distance measure category. 70
Figure 2.12. Decision tree to aid in choosing a sub-category of shape-based distance measures. 71
Figure 3.1. Modelling trend accuracy in the LPI: an overview. 86
Figure 3.2. Sqrt-log model of trend deviation value (TDV) vs sample size with optimal cut point. 96
Figure 3.3. Proportion of the total amount of time series data needed to achieve the trend reliability threshold that each regional taxonomic group in the LPD currently contains. 102
Figure 3.4. Reliability of regional taxonomic group trends in the LPI, grouped by system, realm, and taxon. 103
Figure 3.5. Trend reliability of regional taxonomic groups in the LPD vs the relative weighting applied to each group when calculating aggregated LPI trends. 104
Figure 3.6. The effect on trend accuracy of potential solutions to data deficiency in LP regional taxonomic groups. 105
Figure 4.1. Mean trend deviation value of sampled trends. 127
Figure 4.2. Percentage of the ‘true’ trend captured within sample confidence intervals. 128

Figure 4.3. Normalized width of sampled confidence intervals. 129
Figure 4.4. LPI trends for four regional taxonomic groups. 130

List of main text tables

Table 2.1. Solutions to potential issues in the data. 72
Table 3.1. Parameters with value ranges for simulated datasets, degraded samples, and the LPD. 93
Table 3.2. Multiple regression model of ln(TDV). 98
Table 3.3. Estimated TDV and number of populations needed to meet the threshold for all regional taxonomic groups in the LPD. 99
Table 4.1. Parameters with values for simulated datasets and degraded samples. 125

Chapter 1: Introduction

The reign of modern humans is devastating global biodiversity (Primack, 2018). Coral reefs are dying as the ocean warms, old-growth forests are being cut down or burned to make way for agriculture, and overexploitation is pushing many species to the brink of extinction. In the geological eye-blink since we evolved, we have accelerated the pace of species extinctions to a thousand times the normal background rate (de Vos et al., 2015). If we do not act soon to prevent it, we could cause a 6th mass extinction event (Barnosky et al., 2011).

The goal of conservation science, and by extension, conservation scientists, is to protect biodiversity (Primack, 2018). But to protect the world's biodiversity, we need to understand it. What is out there, where is it, and how is it doing? More importantly, we need to know how the situation is changing, and be able to relate changes to drivers so that we can influence them. The problem is that our understanding of global biodiversity and its trends is limited by lack of data (Hortal et al., 2015; Turak et al., 2017). So far, we have described an estimated 14% (Mora et al., 2011) of extant species, and assessed less than 2.7% of described species for extinction risk (Barnosky et al., 2011), and we have detailed demographic information for only 1.3% of tetrapod species (Conde et al., 2019). In the context of a global biodiversity crisis, this constitutes an urgent data crisis.

Decisions about which species, ecosystems, or geographical areas to protect hinge on knowing which ones need protecting, why, and how successful any previous efforts have been. Attempts to comprehensively assess global biodiversity (e.g., the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, IPBES), and to set policies and goals that will halt or reverse its loss (e.g., the Convention on Biological Diversity, CBD, and Sustainable Development Goals, SDGs), need reliable and up-to-date scientific information (Jetz et al., 2019). Governments, conservation organizations, and researchers contribute, but most studies and tracking programs are either species- or region-focused, temporally limited and inherently biased, leaving large geographic and taxonomic knowledge gaps (Hortal et al., 2015; Jetz et al., 2019; Meyer et al., 2015; Proença et al., 2017; Turak et al., 2017). Advances in technologies such as camera tracking, satellite sensors, digital image recognition, network speed and capacity, data access, and mobile

devices are improving our ability to track and count populations of birds and mammals (Lausch et al., 2016; Nichols et al., 2011; Rose et al., 2015), but even for these species our datasets are far from complete. The situation is worse for amphibians, reptiles, insects, and other groups, for which many species have yet to even be described (Mora et al., 2011).

We need tools to improve our understanding of global biodiversity within the limitations imposed by our biased and incomplete datasets. Mace & Baillie (2007) suggested a solution: develop indicators based on existing data, understand data biases, and develop methods to reduce the bias. Biodiversity indicators summarize complex scientific information in a simple way, often serving as a bridge between science and policy (Secretariat of the Convention on Biological Diversity, 2006). Biodiversity indicators are used for monitoring biodiversity trends (Jones et al., 2011), progress towards conservation goals (Buckland et al., 2012), and impacts of biodiversity policies (Nicholson et al., 2012), and for understanding the impacts humans have on the environment (Watermeyer et al., 2021). They also aid in management decisions and serve political purposes, spurring community dialogue and influencing resource allocation toward urgent conservation issues (Robertson & Hull, 2001). But given the limitations imposed by the state of biodiversity knowledge, indicators can only summarize a fraction of the biodiversity they purport to measure. Therefore, to what extent can we rely on biodiversity indicators to present a true picture of the changing state of global biodiversity? That is the fundamental question I address in this thesis.

1.1. Measuring Biodiversity

The concept of biodiversity lies at the core of conservation science, yet there is no scientifically agreed upon definition of what biodiversity is (Heink & Kowarik, 2010b; Newman et al., 2017). The United Nations' Convention on Biological Diversity (CBD; United Nations, 1992, p. 3) defined biological diversity as "the variability among living organisms from all sources […] includ[ing] diversity within species, between species and of ecosystems." The U.S. Congress (1987, p. 3) defined it to encompass "different ecosystems, species, genes, and their relative abundance." Noss (1990) divided biodiversity into structural, functional, and compositional aspects, and three main levels of organization: ecosystem, species, and gene, with greater diversity at any of these levels or in any of these aspects making a sample more diverse. More recently, the concept of Essential Biodiversity

Variables (EBVs) was developed by the Group on Earth Observations Biodiversity Observation Network (GEO BON) to reduce the variety of possible biodiversity measurements to a manageable common framework of key measurements (Pereira et al., 2013). There are six classes of EBVs, including genetic composition, species populations, species traits, community composition, ecosystem structure, and ecosystem function; however, debate continues over which of the many potential EBVs to codify within these classes (Schmeller et al., 2018). While there are many important biodiversity measures, this thesis will focus on species populations.

Biodiversity is most commonly measured as species richness (Hillebrand et al., 2018; Redford & Sanderson, 1992), meaning simply the number of species. However, species are not equally common; in an environment or taxonomic group, typically a few species will be very abundant, many will be rare, and some will be moderately abundant (Magurran, 2004). Diversity indices such as the Simpson's Diversity Index (Simpson, 1949) and the Shannon Diversity Index (Shannon, 1948) take abundance into account, describing some combination of species richness and species evenness (the similarity in abundances between species).

With an awareness of ongoing human impacts to the environment, such as anthropogenic climate change, deforestation, and agricultural expansion and intensification, there has been an increase in focus on measuring changes in biodiversity over time, referred to as biodiversity trends. Species richness and species evenness are inadequate for measuring temporal changes in biodiversity. If abundance declines evenly across species, there will be no change in evenness. Neither will richness change unless and until extinctions occur (and are recorded). Furthermore, invasive species introductions can contribute to an increase in richness while simultaneously causing declines in abundance (Hillebrand et al., 2018).

Comparing biodiversity across sites or time periods also requires consideration of scale, particularly if using a measure that involves species richness, as richness is an absolute, and therefore scale-dependent, value (Chase & Knight, 2013). Furthermore, richness estimates depend on total and relative abundance, spatial aggregation, and density, as e.g., less abundant and less aggregated species are less likely to be discovered or counted during surveys (Hillebrand et al., 2018). Finally, richness may not reflect changes in species composition, as immigrations due to human-mediated dispersal or human-caused

environmental change may meet or exceed extinctions, and often occur far in advance of extinctions (Elahi et al., 2015; Hillebrand et al., 2018; Sax et al., 2002). With immigration and range expansions occurring on a wide scale, richness can show markedly different changes at different scales, with regional increases belying global declines while local richness remains stable (Dornelas et al., 2019; Thomas, 2013; Vellend et al., 2013).

Biodiversity is widely believed to be declining globally, but the scale of this decline is a topic of controversy and may depend on how biodiversity is defined and measured (Dornelas et al., 2014, 2019; Gonzalez et al., 2016; McGill et al., 2015; Vellend et al., 2013, 2017). There are many ways to quantify or measure biodiversity change. McGill et al. (2015) defined fifteen different types of biodiversity trends. However, biodiversity cannot be measured comprehensively even on a small scale; therefore, indicators are used (Duelli & Obrist, 2003). In general, indicators are surrogate measures intended to be simpler or easier to measure than the parameter they indicate (Gregory & van Strien, 2010). However, indicators may be simple or complex, one or multi-dimensional, measure directly or indirectly, be descriptive or normative, and may or may not be a component of the indicandum, *i.e.*, the parameter being indicated (Heink & Kowarik, 2010a). Biodiversity indicators may have to contend not only with the question of how biodiversity is to be defined, often further complicated by spatial considerations, but also with the question of relative importance (Duelli & Obrist, 2003). For example, are some species more valuable (in conservation terms) than others due to rarity, life history traits, extinction risk, evolutionary uniqueness, etc.?

1.2. Types of Biodiversity Indicators

State indicators measure biodiversity change. There are also indicators measuring pressures and drivers that influence biodiversity change, indicators measuring conservation measures taken in response to biodiversity change, and indicators measuring the socioeconomic impacts of biodiversity change (Biggs et al., 2007). However, this thesis is concerned only with measuring biodiversity change. Therefore, henceforth the term 'biodiversity indicator(s)' will refer only to state indicators.

State indicators can be further divided. Area-based indicators measure biodiversity change according to the fraction of occupied area in comparison to a reference year, e.g., of forests (FAO, 2020), coral reefs (Wilkinson, 2000), or mangroves (Wilkie & Fortuna, 2003). Fragmentation indicators measure the level of fragmentation of habitats, e.g., by mean fragment size or by using the density of roads as a proxy (Biggs et al., 2007). Extinction-risk indicators measure species' risk of extinction. The Red List Index (Butchart et al., 2004) does this using empirical data on population sizes and trends, range extent and occupancy, population fragmentation, and threats, along with expert knowledge and/or judgment. Population-based indicators measure changes in abundance of populations. Some, such as the UK Farmland Bird Indicator (Gregory et al., 2004) or the European Grassland Butterfly Index (van Swaay et al., 2019), do this using a small number of species with high-quality data compiled at national or regional scales, while others, such as the Living Planet Index (McRae et al., 2017), utilize as much data as possible from various sources and quality levels, using a more complex methodology in an attempt to overcome biases and data quality issues at a global scale.

Population-based indicators, particularly the Living Planet Index, are the focus of this thesis. The Living Planet Index is a global biodiversity indicator tracking the changing state of the world's vertebrate biodiversity over time via population time series, beginning at a base year of 1970. While 1970 is modern enough that much of biodiversity decline likely happened prior to that date, population abundance data before 1970 is scarce due to a lack of monitoring programmes (Collins et al., 2020). The LPI is one of the oldest and best-known biodiversity indicators, with a 25-year development history (Ledger et al., 2022). It remains under constant development, with new time series added regularly, as well as methodological updates to improve indicator accuracy and address biases and criticism. The Living Planet Database underlying the LPI currently has more than 38,000 populations representing more than 6,000 vertebrate species (Ledger et al., 2022). Recently, the LPI has come under fire, with papers criticizing it for being biased towards decline due to random population fluctuations (Buschke et al., 2021) and oversensitivity to outliers (Leung et al., 2020). The claims were themselves controversial, with Leung et al. (2020) receiving four published responses (Loreau et al., 2022; Mehrabi & Naidoo, 2020; Murali et al., 2022; Puurtinen et al., 2022). Nonetheless, the fallout resulted in the LPI being removed from the

Convention on Biological Diversity (CBD), where it was previously a headline indicator. Its long history, status and up-to-dateness, and the recent criticism, make the LPI an ideal candidate for examining the reliability and robustness of population-based global biodiversity indicators.

1.3. Data Deficiency in Biodiversity Indicators

We lack the data to fully understand how biodiversity is changing globally. Biodiversity indicators can aid in this by calculating proxy measurements from samples of the world's biodiversity. But to provide accurate proxy measurements, samples must be sufficiently large and sufficiently representative. The quality and abundance of data varies widely between taxa and geographical regions (Boakes et al., 2010; Collen et al., 2008; Conde et al., 2019; Hortal et al., 2015; McRae et al., 2017; Oliveira et al., 2016; Oliver et al., 2021; Scheele et al., 2019; Yesson et al., 2007). This is because resources are limited and data availability is generally a product of interest or convenience, rather than importance (Cardoso et al., 2011; Oliveira et al., 2016). Birds are the best-known organisms, in no small part due to public interest, which leads to increased funding and the availability of citizen science data (Oliver et al., 2021). Birds are charismatic, visible and audible almost everywhere, even in big cities, and relatively easy to identify. By contrast, apart from a few charismatic or highly visible species, invertebrates remain largely unnoticed and poorly known, with millions likely still undescribed (Cardoso et al., 2011), and only 1.7% of described invertebrates assessed for extinction risk (Hochkirch et al., 2021). Even more extreme are microbes, which are generally invisible to the unaided eye, difficult to identify without genetic sequencing, and of very limited interest to the public; they are very poorly known, despite being the most abundant and diverse organisms on Earth (Shoemaker et al., 2017). The tropics are much more diverse than colder northern regions but are poorly studied and therefore much of their biodiversity remains undocumented or poorly understood (Collen et al., 2008). Marine and freshwater biodiversity are less understood than terrestrial biodiversity (Bouchet, 2006; Darwall et al., 2011; Miqueleiz et al., 2020), likely because underwater areas are less accessible and thus less visible.

On a small scale, data deficiency can be evaluated and even compensated. Expert assessment makes clear which species on the IUCN Red List are too data poor to reliably determine extinction risk, and these species are listed as data deficient. In many cases trait data and/or geographic distributions are available for data deficient species (Bland, Orme, et al., 2015), and this information may be used to indirectly assess them. Machine learning algorithms have shown promise for relieving data deficiency in the Red List Index (Bland et al., 2015; Bland & Böhm, 2016; Caetano et al., 2022), as have trait-based models (Luiz et al., 2016; Walls & Dulvy, 2019; Welch & Beaulieu, 2018). However, these methods are only useful to predict individual species. Determining change in extinction risk requires repeated assessments of entire groups; therefore, a trend can only be calculated from the first year of assessment for that group. An advantage of a time-series based indicator like the LPI is that trends can be calculated using existing data. The Living Planet Database contains data from 1950 and resulting LPI trends are assessed either from 1970 or from the earliest date for which time series exist in the database for a given group. On the level of individual species or populations, data deficiency is present in the form of old or short time series. Unlike the Red List Index, data deficient time series can still contribute to LPI trends, although in a reduced capacity. This is because the trends are calculated through a system of hierarchical aggregation of interannual changes, thus allowing individual time series to begin and end at any year in the index. There have to date been no successful attempts to use predictive methods to lengthen or update short or old time series, although efforts are underway (Ledger et al., 2022). But data deficiency can also be considered in relation to biodiversity trends; if a taxonomic group contains too little data to calculate an accurate trend, then that group can be considered as data deficient.

1.4. Assessment of Biodiversity Indicators

Given their use in management and policy decisions, as well as conservation, ecology, and environment research, it is important that biodiversity indicators be assessed to make sure they do what they claim. Biodiversity indicators have been assessed on several criteria, including feasibility, efficiency, sensitivity, specificity, measurability, predictability, complementarity, uncertainty, responsiveness, timeliness, relevance, design, effectiveness, and how fit for purpose they (Halouani et al., 2019; Jones et al., 2011; Link et al., 2009;

Mace & Baillie, 2007; Rowland et al., 2020; Watermeyer et al., 2021). In addition, it has been pointed out that there are spatial and taxonomic biases, and information gaps in underlying data (Hortal et al., 2015; Jetz et al., 2019; Meyer et al., 2015; Proença et al., 2017; Turak et al., 2017). But what has not been assessed, and is rarely discussed, is how accurately biodiversity indicators measure the indicandum. Consider a thermometer that has a scale of degrees and a liquid that expands and contracts when the temperature changes, but the internal diameter of the glass tube is inconsistent and the ticks on the scale are too far apart. This thermometer does what it is supposed to do; it measures temperature. But it is not a thermometer anyone would want to rely on because it will report inaccurate temperatures and respond unpredictably to temperature changes. However, this would only be clear when testing the thermometer against one known to be accurate. Likewise, a biodiversity indicator may do what it claims to but mislead by measuring biodiversity changes inaccurately. The only way to determine this is by testing it against a reference.

The problem is that there is no reference. There is no direct way to measure the accuracy of biodiversity indicator trends because there is no basis for comparison. Biodiversity indicators are unique; what we know about the indicandum comes from the indicator. If the 'true' situation was known the indicator would not be needed. However, there are indirect ways to approach the problem. Indicators are often tested using simulated data models that mimic real-world systems (Fulton et al., 2005; Halouani et al., 2019; Hill et al., 2016; Mccarthy et al., 2014; Rowland et al., 2020). This approach has the advantage that the parameters of the modeled data can be fully known, which is not possible in the real world due to insufficient data (Rowland et al., 2018). An alternative that may be feasible when real-world data is comprehensive is to use a sampled approach, as Baillie et al. (2008) did when testing the minimum sample size needed to achieve reliable trends when developing the sampled approach to the Red List Index (sRLI). This approach involves taking sub-samples of existing real-world data and comparing the resulting trends to the trend of the full sample (Baillie et al., 2008).

1.5. Comparing Trends of Biodiversity Indicators

Baillie et al. (2008) used a simple one-tailed test to compare the direction of linear trends in the sRLI, as the groups they tested were declining and their interest was in avoiding falsely positive trends. Henriques et al. (2020) updated this to include falsely neutral trends. They also added slope comparison, with stronger negative slopes in the samples considered correct and equal or less negative slopes considered false. Non-linear trends, such as those presented in the Living Planet Index, can be compared in various ways. One way is to treat them as linear and only compare the endpoints. It would then be possible to borrow the simplistic sRLI method with adaptations to account for groups with positive trends. However, this ignores important information contained in the more complex trends of the LPI, such as slope and directional changes. A trend could fall steeply, then change directions to become positive, yet still be below its starting index value. Treating such a trend as a linear negative trend would ignore the fact that biodiversity change had been strongly reversed, potentially leading to egregiously wrong results. As an extreme example, consider that a trend with the opposite trajectory (rising at the start, then changing direction and falling steeply) but the same end value would be considered accurate, although representing a negative scenario rather than a positive one. Information is valuable and should not be ignored without good reason. Other options include a statistical test, such as a t-test, or dividing each sampled trend into linear segments (e.g., one segment per year) and calculating the number of times it deviates in direction and/or slope from the reference trend. While these methods capture more information than an end point comparison by allowing for non-linearity, they still decouple the temporal aspect by ignoring the shape of the trend and treating it as a cloud of unconnected points. The temporal information contained in the order of points can be retained by comparing each segment or point on the sampled trend to its corresponding segment or point on the reference trend. One way to do this is by using a distance measure.

The concept of a distance measure is as straightforward as it sounds, a measure of the distance between points. However, the way distance is calculated varies from the simple and familiar Euclidean distance, which uses the Pythagorean theorem, to distances based on complex algorithms that can match multiple points to a single point to account for time

series distortions (e.g., the Dynamic Time Warping distance – Berndt & Clifford, 1994; Mori et al., 2016), and distances based on particular features of time series, such as their estimated partial autocorrelation coefficients (the Partial Autocorrelation-based dissimilarity – Galeano & Peña, 2000; Montero & Vilar, 2014). There are many distance measures that either have been, or could be, used to compare time series. Distance measures are widely used by ecologists in time series comparison tasks, including classification, clustering, prediction, and anomaly detection (e.g., Capinha, 2019; Capinha et al., 2020; Marques et al., 2018; Potamitis et al., 2015; Priyadarshani et al., 2020). However, there is little discussion in the literature of how to select an appropriate distance measure for a given time series comparison task. While a few studies have analysed the classification accuracy of distance measures across different datasets (Bagnall et al., 2017; Paparrizos et al., 2020; Pree et al., 2014; Wang et al., 2013), they only discussed overall accuracy, ignoring dataset- and task-related differences. In my second chapter I fill this hole in the literature by developing an objective method to determine appropriate distance measures to use for any ecological time series comparison task. I evaluate 42 distance measures for 16 properties related to comparing time series, then show how to use the test results, along with a decision tree derived from existing literature, to choose a distance measure fit for the user's purpose. I demonstrate the selection method on a set of UK bird population trends from a study of the effectiveness of conservation measures (Jellesmark et al., 2021). This distance measure selection method not only provides an objective basis to choose a distance measure for comparing biodiversity indicator trends, which I utilize in Chapters 3 and 4, but also provides a generalized method to aid ecologists and scientists from other disciplines who need to compare time series in selecting appropriate distance measures for their own projects.

1.6. Reliability in Biodiversity Indicators

The Living Planet Index is an aggregate index. It calculates a global biodiversity trend from population time series through a series of hierarchical averaging steps, and as such can also be disaggregated into three system trends (Terrestrial, Freshwater, Marine), sixteen realm trends (geographical areas within a given system), and fifty-seven regional taxonomic trends (taxa within realms) (McRae et al., 2017). The accuracy of each trend is a product of the

quality and quantity of the underlying data, as well as the design of the indicator (Collen & Nicholson, 2014). The LPI calculates trends using the geometric mean approach, which has been shown to be the most appropriate for abundance trends (Buckland et al., 2005; van Strien et al., 2012). However, the underlying data in the Living Planet Database (LPD) is known to be biased both geographically and taxonomically (McRae et al., 2017). The LPI has implemented a representative weighting system to account for this, which weights each disaggregated trend according to the number of species in that realm and taxon relative to others (McRae et al., 2017). However, this ignores that trends based on poorly represented regional taxonomic groups might be less accurate due to data paucity. I expect that the quantity of time series (sample size) used to calculate each trend, as well as underlying properties of the time series data, such as the length and the variance in growth rates, determine accuracy. I examine this idea in my third chapter, using a modeling approach to derive a formula for accuracy based on underlying properties of the population time series data the LPI is based on. I create simulated time series datasets using a generalized time series model with varied parameters. I then randomly sample from the datasets and compare the sampled trends to the trends of the full datasets using a distance measure chosen via the method developed in Chapter 2. I then apply the formula to the regional taxonomic groups of the LPI to determine reliability ratings for the disaggregated regional taxonomic trends, as well as the number of populations that would be required to achieve a reasonable threshold of reliability. While others have tested for appropriate responses of biodiversity indices in response to specific modeled scenarios (Fulton et al., 2005; Halouani et al., 2019; Hill et al., 2016; Mccarthy et al., 2014; Rowland et al., 2020), this is to my knowledge the first time anyone has created a generalized model of accuracy that can be used to assess the reliability of real-world trends.

1.7. Uncertainty in Biodiversity Indicators

All biodiversity indicators have some level of uncertainty associated. First, monitoring surveys introduce observational error (measurement error) through e.g., species misidentification and non-detection, errors in counting, and inaccurate plot area measurements (Elphick, 2008; Holdaway et al., 2014). Often, estimates of observational error are not reported (Morrison, 2016), although they may be substantial (Alldredge et al.,

2008; Strickfaden et al., 2020). Second, as biodiversity cannot be comprehensively assessed, sampling must be involved, usually on multiple levels. Any form of sampling introduces sampling error. Population sizes of some large animals can be directly counted, but often it is inaccurate and too expensive, so various direct or indirect estimation methods may be used instead (Fryxell et al., 2014). For example, density can be estimated by averaging counts from a sample of plots (Fryxell et al., 2014), or mark-recapture can be used to indirectly estimate population sizes from samples by capturing a certain number of animals, marking them, releasing them, then capturing more animals and determining how many are marked. Standard errors can be calculated for population size or density estimates but are not always reported. Further, incorporating them into biodiversity indices can be more complicated. Extinction risk in the Red List Index is determined according to objective criteria. Those criteria include quantitative estimates, and may factor in standard errors, but extinction risk categories are qualitative and therefore do not have associated uncertainty (Akçakaya et al., 2000). Index values in the RLI are quantitative and therefore trends do have associated uncertainty in the form of confidence intervals. However, the confidence intervals only incorporate sampling variability (Baillie et al., 2008). To whatever extent, if any, uncertainty in quantitative estimates is accounted for in assessments of extinction risk, that uncertainty does not make it into the index. The Living Planet Index is purely quantitative. However, the Living Planet Database contains data from a wide variety of sources, including grey literature (McRae et al., 2017), which do not always provide estimates of observational or sampling error. Further sources of uncertainty in the LPI are introduced during calculation of the index, including sampling error when species trends are estimated from a sample of the populations within that species, and sampling error when taxonomic or regional trends are estimated from a sample of the species within that taxon or region. The LPI does present confidence intervals, but they only incorporate interannual variation in the species indices (Soldaat et al., 2017). An alternative Monte Carlo estimation method suggested by Soldaat et al. (2017) takes sampling error into account but is not compatible with the LPI because the indicator lacks site-based data (data collected systematically through repeated assessments at the same sites over many years). I address this issue in my fourth chapter by presenting a new method of generating confidence intervals for the LPI that accounts for multiple sources of sampling error without requiring site-based data. I apply a model-based resampling approach to population time series,

modeling each time series with a Generalized Additive Model and using the variance inherent in the model to generate variants of each time series to account for observation error. This method addresses criticism by accounting for sampling error to improve the relevance and accuracy of confidence intervals for LPI trends, and thereby improves the robustness of the LPI.

1.8. Conclusion

Biodiversity is complex and difficult to define, but there is little doubt that it is under threat. As we simultaneously face a climate crisis, biodiversity crisis, and data crisis, biodiversity indicators are becoming increasingly important to understanding the state of biodiversity across the planet so that we can respond appropriately with policies and conservation actions. The research I present in this thesis not only represents an important contribution to the field of biodiversity indicator research but will improve and enlarge the methodological toolbox for other ecologists and conservation scientists to utilize. In my final chapter, I place my research into context and explain its importance to the Living Planet Index, indicator research, and ecology in general.

1.9. Bibliography

Akçakaya, H. R., Ferson, S., Burgman, M. A., Keith, D. A., Mace, G. M., & Todd, C. R. (2000). Making Consistent IUCN Classifications under Uncertainty. *Conservation Biology*, *14*(4), 1001–1013. https://doi.org/10.1046/j.1523-1739.2000.99125.x

Alldredge, M. W., Pacifici, K., Simons, T. R., & Pollock, K. H. (2008). A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities. *Journal of Applied Ecology*, *45*(5), 1349–1356. https://doi.org/10.1111/j.1365-2664.2008.01517.x

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Baillie, J. E. M., Collen, B., Amin, R., Akcakaya, H. R., Butchart, S. H. M., Brummitt, N., Meagher, T. R., Ram, M., Hilton-Taylor, C., & Mace, G. M. (2008). Toward monitoring global biodiversity. *Conservation Letters*, *1*(1), 18–26. https://doi.org/10.1111/j.1755-263x.2008.00009.x

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? In *Nature* (Vol. 471, Issue 7336, pp. 51–57). https://doi.org/10.1038/nature09678

Berndt, D. J., & Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. In *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop* (pp. 359–370). AAAI Press. https://www.aaai.org/Library/Workshops/ws94-03.php

Biggs, R., Scholes, R. J., ten Brink, B. J. E., & Vačkář, D. (2007). Biodiversity indicators. In T. Hák, B. Moldan, & A. L. Dahl (Eds.), *Sustainability indicators: A scientific assessment* (pp. 249–270). Island Press.

Bland, L. M., & Böhm, M. (2016). Overcoming data deficiency in reptiles. *Biological Conservation*, *204*, 16–22. https://doi.org/10.1016/j.biocon.2016.05.018

Bland, L. M., Collen, B., Orme, C. D. L., & Bielby, J. (2015). Predicting the conservation status of data-deficient species. *Conservation Biology*, *29*(1), 250–259. https://doi.org/10.1111/cobi.12372

Bland, L. M., Orme, C. D. L., Bielby, J., Collen, B., Nicholson, E., & Mccarthy, M. A. (2015). Cost-effective assessment of extinction risk with limited information. *Journal of Applied Ecology*, *52*(4), 861–870. https://doi.org/10.1111/1365-2664.12459

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-Qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6). https://doi.org/10.1371/journal.pbio.1000385

Bouchet, P. (2006). The magnitude of marine biodiversity. In C. M. Duarte (Ed.), *The Exploration of Marine Biodiversity: Scientific and Technological Challenges* (pp. 31–62). Fundación BBVA.

Buckland, S. T., Baillie, S. R., Dick, J. M. P., Elston, D. A., Magurran, A. E., Scott, E. M., Smith, R. I., Somerfield, P. J., Studeny, A. C., & Watt, A. (2012). How should regional biodiversity be monitored? *Environmental and Ecological Statistics*, *19*(4), 601–626. https://doi.org/10.1007/s10651-012-0202-7

Buckland, S. T., Magurran, A. E., Green, R. E., & Fewster, R. M. (2005). Monitoring change in biodiversity through composite indices. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 243–254. https://doi.org/10.1098/rstb.2004.1589

Buschke, F. T., Hagan, J. G., Santini, L., & Coetzee, B. W. T. (2021). Random population fluctuations bias the Living Planet Index. *Nature Ecology and Evolution*, *5*(8), 1145–1152. https://doi.org/10.1038/s41559-021-01494-0

Butchart, S. H. M., Stattersfield, A. J., Bennun, L. A., Shutes, S. M., Akçakaya, H. R., Baillie, J. E. M., Stuart, S. N., Hilton-Taylor, C., & Mace, G. M. (2004). Measuring global trends in the status of biodiversity: Red list indices for birds. *PLoS Biology*, *2*(12). https://doi.org/10.1371/journal.pbio.0020383

Caetano, G. H. de O., Chapple, D. G., Grenyer, R., Raz, T., Rosenblatt, J., Tingley, R., Böhm, M., Meiri, S., & Roll, U. (2022). Automated assessment reveals that the extinction risk of reptiles is widely underestimated across space and phylogeny. *PLOS Biology*, *20*(5), e3001544. https://doi.org/10.1371/journal.pbio.3001544

Capinha, C. (2019). Predicting the timing of ecological phenomena using dates of species occurrence records: a methodological approach and test case with mushrooms. *International Journal of Biometeorology*, *63*(8), 1015–1024. https://doi.org/10.1007/s00484-019-01714-0

Capinha, C., Ceia-Hasse, A., Kramer, A. M., & Meijer, C. (2020). Deep learning classification of temporal data in ecology. *BioRxiv*. https://doi.org/10.1101/2020.09.14.296251

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, *144*(11), 2647–2655. https://doi.org/10.1016/j.biocon.2011.07.024

Chase, J. M., & Knight, T. M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity: Why standardised sampling is not enough. *Ecology Letters*, *16*(SUPPL.1), 17–26. https://doi.org/10.1111/ele.12112

Collen, B., & Nicholson, E. (2014). Taking the measure of change. *Science*, *346*(6206), 166–167. https://doi.org/10.1126/science.1255772

Collen, B., Ram, M., Zamin, T., & Mcrae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88.

Collins, A. C., Böhm, M., & Collen, B. (2020). Choice of baseline affects historical population trends in hunted mammals of North America. *Biological Conservation*, *242*. https://doi.org/10.1016/j.biocon.2020.108421

Conde, D. A., Staerk, J., Colchero, F., da Silva, R., Schöley, J., Maria Baden, H., Jouvet, L., Fa, J. E., Syed, H., Jongejans, E., Meiri, S., Gaillard, J.-M., Chamberlain, S., Wilcken, J., Jones, O. R., Dahlgren, J. P., Steiner, U. K., Bland, L. M., Gomez-Mestre, I., … Analyzed, J. W. (2019). *Data gaps and opportunities for comparative and conservation biology*. https://doi.org/10.5061/dryad

Darwall, W. R. T., Holland, R. A., Smith, K. G., Allen, D., Brooks, E. G. E., Katarya, V., Pollock, C. M., Shi, Y., Clausnitzer, V., Cumberlidge, N., Cuttelod, A., Dijkstra, K. D. B., Diop, M. D., García, N., Seddon, M. B., Skelton, P. H., Snoeks, J., Tweddle, D., & Vié, J. C. (2011). Implications of bias in conservation research and investment for freshwater species. *Conservation Letters*, *4*(6), 474–482. https://doi.org/10.1111/j.1755-263X.2011.00202.x

de Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R., & Pimm, S. L. (2015). Estimating the normal background rate of species extinction. *Conservation Biology*, *29*(2), 452–462. https://doi.org/10.1111/cobi.12380

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., & Magurran, A. E. (2014). Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. *Science*, *344*(6181), 296–299. https://doi.org/10.1126/science.1248484

Dornelas, M., Gotelli, N. J., Shimadzu, H., Moyes, F., Magurran, A. E., & McGill, B. J. (2019). A balance of winners and losers in the Anthropocene. *Ecology Letters*, *22*(5), 847–854. https://doi.org/10.1111/ele.13242

Duelli, P., & Obrist, M. K. (2003). Biodiversity indicators: The choice of values and measures. *Agriculture, Ecosystems and Environment*, *98*(1–3), 87–98. https://doi.org/10.1016/S0167-8809(03)00072-0

Elahi, R., O'Connor, M. I., Byrnes, J. E. K., Dunic, J., Eriksson, B. K., Hensel, M. J. S., & Kearns, P. J. (2015). Recent Trends in Local-Scale Marine Biodiversity Reflect Community Structure and Human Impacts. *Current Biology*, *25*(14), 1938–1943. https://doi.org/10.1016/j.cub.2015.05.030

Elphick, C. S. (2008). How you count counts: The importance of methods research in applied ecology. *Journal of Applied Ecology*, *45*(5), 1313–1320. https://doi.org/10.1111/j.1365-2664.2008.01545.x

FAO. (2020). Global Forest Resources Assessment 2020: Main Report. *Food and Agricultural Organization of the United Nations*. https://doi.org/10.4060/ca9825en

Fryxell, J. M., Sinclair, A. R. E., & Caughley, G. (2014). *Wildlife Ecology, Conservation, and Management* (3rd ed.). Wiley Blackwell.

Fulton, E. A., Smith, A. D. M., & Punt, A. E. (2005). Which ecological indicators can robustly detect effects of fishing? *ICES Journal of Marine Science*, *62*(3), 540–551. https://doi.org/10.1016/j.icesjms.2004.12.012

Galeano, P., & Peña, D. (2000). *Multivariate Analysis in Vector Time Series*.

Gonzalez, A., Cardinale, B. J., Allington, G. R. H., Byrnes, J., Arthur, K., Brown, D. G., Hooper, D. U., Isbell, F., O'connor, M. I., & Loreau, M. (2016). Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology*, *8*, 1949–1960.

Gregory, R. D., Noble, D. G., & Custance, J. (2004). The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. *Ibis*, *146*(Suppl. 2), 1–13.

Gregory, R. D., & van Strien, A. (2010). Wild bird indicators: using composite population trends of birds as measures of environmental health. *Ornithological Science*, *9*, 3–22.

Halouani, G., le Loc'h, F., Shin, Y. J., Velez, L., Hattab, T., Romdhane, M. S., & ben Rais Lasram, F. (2019). An end-to-end model to evaluate the sensitivity of ecosystem indicators to track fishing impacts. *Ecological Indicators*, *98*, 121–130. https://doi.org/10.1016/j.ecolind.2018.10.061

Heink, U., & Kowarik, I. (2010a). What are indicators? On the definition of indicators in ecology and environmental planning. *Ecological Indicators*, *10*(3), 584–593. https://doi.org/10.1016/j.ecolind.2009.09.009

Heink, U., & Kowarik, I. (2010b). What criteria should be used to select biodiversity indicators? *Biodiversity and Conservation*, *19*(13), 3769–3797. https://doi.org/10.1007/s10531-010-9926-6

Henriques, S., Böhm, M., Collen, B., Luedtke, J., Hoffmann, M., Hilton-Taylor, C., Cardoso, P., Butchart, S. H. M., & Freeman, R. (2020). Accelerating the monitoring of global biodiversity: Revisiting the sampled approach to generating Red List Indices. *Conservation Letters*, *13*(3). https://doi.org/10.1111/conl.12703

Hill, S. L. L., Harfoot, M., Purvis, A., Purves, D. W., Collen, B., Newbold, T., Burgess, N. D., & Mace, G. M. (2016). Reconciling Biodiversity Indicators to Guide Understanding and Action. *Conservation Letters*, *9*(6), 405–412. https://doi.org/10.1111/conl.12291

Hillebrand, H., Blasius, B., Borer, E. T., Chase, J. M., Downing, J. A., Eriksson, B. K., Filstrup, C. T., Harpole, W. S., Hodapp, D., Larsen, S., Lewandowska, A. M., Seabloom, E. W., van de Waal, D. B., & Ryabov, A. B. (2018). Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *Journal of Applied Ecology*, *55*(1), 169–184. https://doi.org/10.1111/1365-2664.12959

Hochkirch, A., Samways, M. J., Gerlach, J., Böhm, M., Williams, P., Cardoso, P., Cumberlidge, N., Stephenson, P. J., Seddon, M. B., Clausnitzer, V., Borges, P. A. V., Mueller, G. M., Pearce-Kelly, P., Raimondo, D. C., Danielczak, A., & Dijkstra, K. D. B. (2021). A strategy for the next decade to address data deficiency in neglected biodiversity. *Conservation Biology*, *35*(2), 502–509. https://doi.org/10.1111/cobi.13589

Holdaway, R. J., McNeill, S. J., Mason, N. W. H., & Carswell, F. E. (2014). Propagating Uncertainty in Plot-based Estimates of Forest Carbon Stock and Carbon Stock Change. *Ecosystems*, *17*(4), 627–640. https://doi.org/10.1007/s10021-014-9749-5

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Jellesmark, S., Ausden, M., Blackburn, T. M., Gregory, R. D., Hoffmann, M., Massimino, D., McRae, L., & Visconti, P. (2021). A counterfactual approach to measure the impact of wet grassland conservation on U.K. breeding bird populations. *Conservation Biology*, *35*(5), 1575–1585. https://doi.org/10.1111/cobi.13692

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology and Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Jones, J. P. G., Collen, B., Atkinson, G., Baxter, P. W. J., Bubb, P., Illian, J. B., Katzner, T. E., Keane, A., Loh, J., Mcdonald-Madden, E., Nicholson, E., Pereira, H. M., Possingham, H. P., Pullin, A. S., Rodrigues, A. S. L., Ruiz-Gutierrez, V., Sommerville, M., & Milner-Gulland, E. J. (2011). The Why, What, and How of Global Biodiversity Indicators Beyond the 2010 Target. *Conservation Biology*, *25*(3), 450–457. https://doi.org/10.1111/j.1523-1739.2010.01605.x

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J. M., Heurich, M., Jung, A., Klenke, R., Neumann, C., Pause, M., Rocchini, D., Schaepman, M. E., Schmidtlein, S., Schulz, K., Selsam, P., Settele, J., Skidmore, A. K., & Cord, A. F. (2016). Linking Earth Observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecological Indicators*, *70*, 317–339. https://doi.org/10.1016/j.ecolind.2016.06.022

Ledger, S. E. H., McRae, L., Loh, J., Almond, R., Böhm, M., Currie, J., Deinet, S., Galewski, T., Grooten, M., Jenkins, M., Marconi, V., Painter, B., Scott-Gatty, K., Young, L., & Hoffmann, M. (2022). Past, present, and future of the Living Planet Index. *BioRxiv*. https://doi.org/10.1101/2022.06.20.496803

Link, J. S., Yemane, D., Shannon, L. J., Coll, M., Shin, Y.-J., Hill, L., de Fatima Borges Link, M., & Link, J. S. (2009). Relating marine ecosystem indicators to fishing and environmental drivers: an elucidation of contrasting responses. *ICES Journal of Marine Science*, *67*, 787–795. http://www.primer-e.com

Loreau, M., Cardinale, B. J., Isbell, F., Newbold, T., O'Connor, M. I., & de Mazancourt, C. (2022). Do not downplay biodiversity loss. *Nature*, *601*, E27–E28. https://doi.org/10.1038/s41586-021-04179-7

Luiz, O. J., Woods, R. M., Madin, E. M. P., & Madin, J. S. (2016). Predicting IUCN Extinction Risk Categories for the World's Data Deficient Groupers (Teleostei: Epinephelidae). *Conservation Letters*, *9*(5), 342–350. https://doi.org/10.1111/conl.12230

Mace, G. M., & Baillie, J. E. M. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, *21*(6), 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

Magurran, A. E. (2004). *Measuring biological diversity*. Blackwell Publishing.

Marques, A. R., Forde, H., & Revie, C. W. (2018). Time-series clustering of cage-level sea lice data. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0204319

Mccarthy, M. A., Moore, A. L., Krauss, J., Morgan, J. W., & Clements, C. F. (2014). Linking Indices for Biodiversity Monitoring to Extinction Risk Theory. *Conservation Biology*, *28*(6), 1575–1583. https://doi.org/10.1111/cobi.12308

McGill, B. J., Dornelas, M., Gotelli, N. J., & Magurran, A. E. (2015). Fifteen forms of biodiversity trend in the anthropocene. *Trends in Ecology and Evolution*, *30*(2), 104–113. https://doi.org/10.1016/j.tree.2014.11.006

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0169156

Mehrabi, Z., & Naidoo, R. (2020). Shifting baselines and biodiversity success stories. *Nature*, *601*, E17–E18. https://doi.org/10.1038/s41586-021-03750-6

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*(8221). https://doi.org/10.1038/ncomms9221

Miqueleiz, I., Bohm, M., Ariño, A. H., & Miranda, R. (2020). Assessment gaps and biases in knowledge of conservation status of fishes. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *30*(2), 225–236. https://doi.org/10.1002/aqc.3282

Montero, P., & Vilar, J. A. (2014). TSclust: An R Package for Time Series Clustering. *JSS Journal of Statistical Software*, *62*. http://www.jstatsoft.org/

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*(8). https://doi.org/10.1371/journal.pbio.1001127

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). *Distance Measures for Time Series in R: The TSdist Package*.

Morrison, L. W. (2016). Observer error in vegetation surveys: A review. *Journal of Plant Ecology*, *9*(4), 367–379. https://doi.org/10.1093/jpe/rtv077

Murali, G., de Oliveira Caetano, G. H., Barki, G., Meiri, S., & Roll, U. (2022). Emphasizing declining populations in the Living Planet Report. *Nature*, *601*, E20–E22. https://doi.org/10.1038/s41586-021-04165-z

Newman, J. A., Varner, G., & Linquist, S. (2017). *Defending biodiversity: environmental science and ethics*. Cambridge University Press.

Nichols, J. D., O'Connell, A. F., & Karanth, K. U. (2011). Camera traps in animal ecology and conservation: What's next? In *Camera Traps in Animal Ecology: Methods and Analyses* (pp. 253–263). Springer Japan. https://doi.org/10.1007/978-4-431-99495-4_14

Nicholson, E., Collen, B., Barausse, A., Blanchard, J. L., Costelloe, B. T., Sullivan, K. M. E., Underwood, F. M., Burn, R. W., Fritz, S., Jones, J. P. G., McRae, L., Possingham, H. P., & Milner-Gulland, E. J. (2012). Making robust policy decisions using global biodiversity indicators. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041128

Noss, R. F. (1990). Indicators for Monitoring Biodiversity: A Hierarchical Approach. *Conservation Biology*, *4*(4), 355–364.

Oliveira, U., Paglia, A. P., Brescovit, A. D., de Carvalho, C. J. B., Silva, D. P., Rezende, D. T., Leite, F. S. F., Batista, J. A. N., Barbosa, J. P. P. P., Stehmann, J. R., Ascher, J. S., de Vasconcelos, M. F., de Marco, P., Löwenberg-Neto, P., Dias, P. G., Ferro, V. G., &

Santos, A. J. (2016). The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions*, *22*(12), 1232–1244. https://doi.org/10.1111/ddi.12489

Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K., & Jetz, W. (2021). Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biology*, *19*(8). https://doi.org/10.1371/journal.pbio.3001336

Paparrizos, J., Liu, C., Elmore, A. J., & Franklin, M. J. (2020). Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1887–1905. https://doi.org/10.1145/3318464.3389760

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*(6117), 277–278. https://doi.org/10.1126/science.1229931

Potamitis, I., Rigakis, I., & Fysarakis, K. (2015). Insect biometrics: Optoacoustic signal processing and its applications to remote monitoring of McPhail type traps. *PLoS ONE*, *10*(11). https://doi.org/10.1371/journal.pone.0140474

Pree, H., Herwig, B., Gruber, T., Sick, B., David, K., & Lukowicz, P. (2014). On general purpose time series similarity measures and their use as kernel functions in support vector machines. *Information Sciences*, *281*, 478–495. https://doi.org/10.1016/j.ins.2014.05.025

Primack, R. B. (2018). *Essentials of Conservation Biology* (6th ed.). Sinauer Associates.

Priyadarshani, N., Marsland, S., Juodakis, J., Castro, I., & Listanti, V. (2020). Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods in Ecology and Evolution*, *11*(3), 403–417. https://doi.org/10.1111/2041-210X.13357

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R. D., Honrado, J. P., Jürgens, N., Opige, M., Schmeller, D. S., Tiago, P., & van Swaay, C. A. M. (2017). Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. *Biological Conservation*, *213*, 256–263. https://doi.org/10.1016/j.biocon.2016.07.014

Puurtinen, M., Elo, M., & Kotiaho, J. S. (2022). The Living Planet Index does not measure abundance. *Nature*, *601*, E14–E15. https://doi.org/10.1038/s41586-021-03708-8

Redford, K. H., & Sanderson, S. E. (1992). The Brief, Barren Marriage of Biodiversity and Sustainability? *Bulletin of the Ecological Society of America*, *73*(1), 36–39.

Robertson, D. P., & Hull, R. B. (2001). Essays Beyond Biology: toward a More Public Ecology for Conservation. *Conservation Biology*, *15*(4), 970–979.

Rose, R. A., Byler, D., Eastman, J. R., Fleishman, E., Geller, G., Goetz, S., Guild, L., Hamilton, H., Hansen, M., Headley, R., Hewson, J., Horning, N., Kaplin, B. A., Laporte, N., Leidner, A., Leimgruber, P., Morisette, J., Musinsky, J., Pintea, L., … Wilson, C. (2015). Ten ways remote sensing can contribute to conservation. *Conservation Biology*, *29*(2), 350–359. https://doi.org/10.1111/cobi.12397

Rowland, J. A., Lee, C. K. F., Bland, L. M., & Nicholson, E. (2020). Testing the performance of ecosystem indices for biodiversity monitoring. *Ecological Indicators*, *116*. https://doi.org/10.1016/j.ecolind.2020.106453

Rowland, J. A., Nicholson, E., Murray, N. J., Keith, D. A., Lester, R. E., & Bland, L. M. (2018). Selecting and applying indicators of ecosystem collapse for risk assessments. *Conservation Biology*, *32*(6), 1233–1245. https://doi.org/10.1111/cobi.13107

Sax, D. F., Gaines, S. D., & Brown, J. H. (2002). Species invasions exceed extinctions on islands worldwide: a comparative study of plants and birds. *The American Naturalist*, *160*(6).

Scheele, B. C., Legge, S., Blanchard, W., Garnett, S., Geyle, H., Gillespie, G., Harrison, P., Lindenmayer, D., Lintermans, M., Robinson, N., & Woinarski, J. (2019). Continental-scale assessment reveals inadequate monitoring for threatened vertebrates in a megadiverse country. *Biological Conservation*, *235*, 273–278. https://doi.org/10.1016/j.biocon.2019.04.023

Schmeller, D. S., Weatherdon, L. v, Loyau, A., Bondeau, A., Brotons, L., Brummitt, N., Geijzendorffer, I. R., Haase, P., Kuemmerlen, M., Martin, C. S., Mihoub, J.-B., Rocchini, D., Saarenmaa, H., Stoll, S., Regan, E. C., Fondazione,), & Mach, E. (2018). A suite of essential biodiversity variables for detecting critical biodiversity change. *Biological Reviews*, *93*, 55–71. https://doi.org/10.1111/brv.12332

Secretariat of the Convention on Biological Diversity. (2006). *Global Biodiversity Outlook 2*. https://www.cbd.int/gbo2/

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shoemaker, W. R., Locey, K. J., & Lennon, J. T. (2017). A macroecological theory of microbial biodiversity. *Nature Ecology and Evolution*, *1*(5). https://doi.org/10.1038/s41559-017-0107

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*, 688–688. https://doi.org/10.1038/163688a0

Soldaat, L. L., Pannekoek, J., Verweij, R. J. T., van Turnhout, C. A. M., & van Strien, A. J. (2017). A Monte Carlo method to account for sampling error in multi-species indicators. *Ecological Indicators*, *81*, 340–347. https://doi.org/10.1016/j.ecolind.2017.05.033

Strickfaden, K. M., Fagre, D. A., Golding, J. D., Harrington, A. H., Reintsma, K. M., Tack, J. D., & Dreitz, V. J. (2020). Dependent double-observer method reduces false-positive errors in auditory avian survey data. *Ecological Applications*, *30*(2). https://doi.org/10.1002/eap.2026

Thomas, C. D. (2013). Local diversity stays about the same, regional diversity increases, and global diversity declines. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19187–19188. https://doi.org/10.1073/pnas.1319304110

Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., Finlayson, C. M., Ferrier, S., Freyhof, J., Hermoso, V., Juffe-Bignoli, D., Linke, S., Nel, J., Patricio, H. C., Pittock, J., Raghavan, R., Revenga, C., Simaika, J. P., & de Wever, A. (2017). Essential Biodiversity Variables for measuring change in global freshwater biodiversity. *Biological Conservation*, *213*, 272–279. https://doi.org/10.1016/j.biocon.2016.09.005

United Nations. (1992). *Convention on biological diversity*. https://www.cbd.int/doc/legal/cbd-en.pdf/

U.S. Congress. (1987). *Technologies to maintain biological diversity*. Office of Technology Assessment. https://www.biodiversitylibrary.org/bibliography/4178

van Strien, A. J., Soldaat, L. L., & Gregory, R. D. (2012). Desirable mathematical properties of indicators for biodiversity change. *Ecological Indicators*, *14*(1), 202–208. https://doi.org/10.1016/j.ecolind.2011.07.007

van Swaay, C. A. M., Dennis, E. B., Schmucki, R., Sevilleja, C., Balalalaikins, M., Botham, M., Bourn, N., Brereton, T., Cancela, J. P., Carlisle, B., Chambers, P., Collins, S., Dopagne, C., Escobés, R., Feldmann, R., Fernández-García, J. M., Fontaine, B., Gracianteparaluceta, A., Harrower, C., … Roy, D. B. (2019). *The EU Butterfly Indicator for Grassland species: 1990-2017*. www.butterfly-monitoring.net

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., de Frenne, P., Verheyen, K., & Wipf, S. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19456–19459. https://doi.org/10.1073/pnas.1312779110

Vellend, M., Dornelas, M., Baeten, L., Beauséjour, R., Brown, C. D., de Frenne, P., Elmendorf, S. C., Gotelli, N. J., Moyes, F., Myers-Smith, I. H., Magurran, A. E., McGill, B. J., Shimadzu, H., & Sievers, C. (2017). Estimates of local biodiversity change over time stand up to scrutiny. *Ecology*, *98*(2), 583–590. https://doi.org/10.1002/ecy.1660

Walls, R. H. L., & Dulvy, N. K. (2019). Predicting the conservation status of Europe's Data Deficient sharks and rays. *BioRxiv*. https://doi.org/10.1101/614776

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5

Watermeyer, K. E., Guillera-Arroita, G., Bal, P., Burgass, M. J., Bland, L. M., Collen, B., Hallam, C., Kelly, L. T., McCarthy, M. A., Regan, T. J., Stevenson, S., Wintle, B. A., & Nicholson, E. (2021). Using decision science to evaluate global biodiversity indices. *Conservation Biology*, *35*(2), 492–501. https://doi.org/10.1111/cobi.13574

Welch, J. N., & Beaulieu, J. M. (2018). Predicting extinction risk for data deficient bats. *Diversity*, *10*(3). https://doi.org/10.3390/d10030063

Wilkie, M. L., & Fortuna, S. (2003). *Status and trends in mangrove area extent worldwide*. Forest Resources Assessment Programme. Working Paper (FAO).

Wilkinson, C. R. (2000). *Status of coral reefs of the world: 2000*. Australian Institute of Marine Science.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the global biodiversity information facility? *PLoS ONE*, *2*(11). https://doi.org/10.1371/journal.pone.0001124

Chapter 2: Selecting Appropriate Distance Measures to Compare Ecological Time Series

2.1. Abstract

1. Time series are a critical component of ecological analysis, used to track changes in biotic and abiotic variables. Information can be extracted from the properties of time series for tasks such as classification, clustering, prediction, and anomaly detection. These common tasks in ecological research rely on the notion of (dis-) similarity which can be determined by using distance measures. A plethora of distance measures have been described, predominantly in the computer and information sciences, but many of them have not been introduced to ecologists. Furthermore, little is known about how to select appropriate distance measures and the properties they focus on for time-series related tasks.

2. Here I describe 16 potentially desirable properties of distance measures, test 42 distance measures for each property, and present an objective method to select appropriate distance measures for any task and ecological dataset. I then demonstrate my selection method by applying it to a set of real-world data on breeding bird populations in the UK. I also discuss ways to overcome some of the difficulties involved in using distance measures to compare time series.

3. The real-world population trends exhibit a common challenge for time series comparison: a high level of stochasticity. I demonstrate two different ways of overcoming this challenge, first by selecting distance measures with properties that make them well-suited to comparing noisy time series, and second by applying a smoothing algorithm before selecting appropriate distance measures. In both cases, the distance measures chosen through my selection method are not only fit-for-purpose but are consistent in their rankings of the population trends.

4. The results of my study should lead to an improved understanding of, and greater scope for, the use of distance measures for comparing ecological time series, and allow for the answering of new ecological questions.

2.2. Introduction

Time series are a critical component of ecological analysis: ecologists use time series to track changes in biotic variables, such as population sizes and mean growth rates of individuals, as well as abiotic variables, such as temperature and atmospheric carbon dioxide. Time series provide insight into food web and ecosystem function and the causes and effects of environmental change, and are vital to any scientific approach to environmental management (Boero et al., 2015). Time series datasets may contain thousands or even millions of time series (e.g., The Living Planet Index – WWF, 2020; BioTIME - Dornelas et al., 2018; the North American Breeding Bird Survey - Pardieck et al., 2020; the British Trust for Ornithology Breeding Bird Survey - Harris et al., 2020; and the Continuous Plankton Recorder Survey - Edwards et al., 2016). Ecologists make inferences through time series comparisons. For example, one might look for similarities or differences in climate change response between populations within or across geographic or taxonomic groups. However, examining and analysing each time series by hand is unwieldy.

Data mining of time series is the process of extracting information from the properties of time series for tasks such as classification, clustering, prediction, and anomaly detection (Esling & Agon, 2012). These tasks are common in ecology, e.g., clustering time series of parasite counts to identify infection patterns (Marques et al., 2018); predicting the emergence of fruiting bodies by classifying time series of environmental drivers (Capinha, 2019); identifying insect species by classifying wingbeat frequency signals (Potamitis et al., 2015); surveying bird population sizes by classifying recorded calls (Priyadarshani et al., 2020); and predicting species distributions based on time series of environmental variables (Capinha et al., 2020). These tasks all rely on the notion of (dis-) similarity. Clustering involves grouping similar time series together by maximizing the similarity within groups and minimizing the similarity between groups (Aghabozorgi et al., 2015; Esling & Agon, 2012; Warren Liao, 2005). Classification is like clustering, except labels are predefined and new time series are assigned to existing clusters to which they are most similar (Keogh & Kasetti, 2003). Prediction may rely on similarity to determine accuracy by comparing predicted time series against the originals (Capinha, 2019; Esling & Agon, 2012). Finally,

anomaly detection involves comparing time series against an anomaly-free model to determine if they fall outside of a similarity threshold (Esling & Agon, 2012; Teng, 2010).

Similarity between time series can be determined by using distance measures to measure its inverse: dissimilarity. Dissimilarity is more intuitive as a measurement because a value of zero occurs when two time series are identical (while similarity is at a scale-dependent maximum value). Distance measures can be broadly categorized into four different types: shape-based, feature-based, model-based, and compression-based. Shape-based distances compare the shapes of time series by measuring differences in the raw data values (Aghabozorgi et al., 2015; Esling & Agon, 2012) and can be further divided into lock-step measures and elastic measures. Lock-step measures compare each time point of one time series to the corresponding time point of another time series, while elastic measures allow a single point to be matched with multiple points or no points (Wang et al., 2013). Elastic measures fall into two groups. The first, Dynamic Time Warping (DTW), computes an optimal match between two time series by allowing single points to be matched with multiple points, thus allowing local distortion or 'warping' of the time dimension (Esling & Agon, 2012). The second comprises edit distances, which compare the minimum number of 'edits,' or changes, required to transform one time series into another (Esling & Agon, 2012). They are based on the concept of transforming one string into another by changing one letter at a time, with each 'edit' being an insertion, deletion, or substitution. Feature-based distances compute some feature of time series, such as Discrete Fourier Transforms or autocorrelation coefficients, and use either a specialized or common distance function (e.g., the Euclidean distance) to determine the distance between the computed features (Mori et al., 2016a). Model-based distances compare the parameters of models fitted to the time series, such as autoregressive moving average (ARMA) models, with the advantage that they can incorporate knowledge about the process used to generate the time series data (Esling & Agon, 2012). Finally, compression-based distances assess the similarity of two digital objects according to how well they can be 'compressed' when connected (Cilibrasi & Vitanyi, 2018; Esling & Agon, 2012); the more similar the objects, the better they compress when joined in series (Esling & Agon, 2012). Although there are comparatively few model-based and compression-based distance measures, there are many shape-based and feature-based measures available.

The choice of distance measure for any task should depend on the properties of the data to be analysed and the nature of the task (Esling & Agon, 2012). In practice, choosing a distance measure often becomes a matter of convenience. For example, the well-known and easy to use Euclidean distance is among the most widely used distance measures, although there are often better choices (Paparrizos et al., 2020; Wang et al., 2013). When investigating the performance of five distance measures for comparing animal movement trajectories, Cleasby et al. (2019) found that the most used measure was the least appropriate choice. One problem is that many distance measures originate within computer science, information science, systems science, and mathematics, and few are in common use within ecology. Another problem is that information on the strengths, weaknesses, and appropriate uses of distance measures is limited and often difficult to find. Some reviews of distance measures have been published (Esling & Agon, 2012; Lhermitte et al., 2011; Montero & Vilar, 2014; Mori et al., 2016a; Liao, 2005), but are not generally aimed at ecologists (but see Lhermitte et al., 2011); analysis of the properties of distance measures is limited, and guidance on how to choose an appropriate distance measure is either missing or very general. Other studies have analysed the classification accuracy of multiple distance measures across a variety of datasets (Bagnall et al., 2017; Paparrizos et al., 2020; Pree et al., 2014; Wang et al., 2013), but pooled the results to give overall performance scores. This ignores the fact that different distance measures perform better on different datasets and for different tasks. Kocher & Savoy (2017) tested 24 distance measures for six properties, then compared their effectiveness in classification on 13 real-world datasets. However, the study focused on a single task (author profiling, i.e., determining demographic information about the author of a document based on the document itself) and did not present a general method for selecting distance measures for other tasks. Furthermore, the distance measures that demonstrated all proposed properties did not perform best on real-world datasets. Mori et al. (2016b) developed an automated process for selecting distance measures based on nine quantifiable properties of datasets. However, their classifier is limited to clustering tasks, and only includes five common distance measures. I am not aware of any more generalized method of distance measure selection.

In this study, I present a generalized, objective, user-driven method of choosing fit-for-purpose distance measures for time-series comparison. I evaluate 42 distance measures for

16 properties related to time series comparison. I then demonstrate my selection method by applying it to a set of real-world UK bird population trends from a study of the effectiveness of conservation measures (Jellesmark *et al*., 2021). Finally, I discuss how to select appropriate distance measure(s) for any dataset and task.

2.3. Methods

I selected 42 distance measures from the literature (see Table S2.1 in Appendix 1 for a detailed list). I chose measures that had already been implemented in publicly accessible R packages, and that represented each of the categories we defined in the introduction, as well as a variety of potential use cases. Eighteen of the distance measures I selected are implemented in the R package 'TSclust' (version 1.3.1) and have been studied for use in clustering time series (Montero & Vilar, 2014). The other twenty-four are implemented in the R package 'philentropy' (version 0.5.0; Drost, 2018).

I defined a set of 16 properties of distance measures that may be of interest in time series comparison: four metric properties, six value-based properties, five time-based properties, and one uncategorized property. Metric properties define whether dissimilarity is measured in metric space (a space that has physical meaning). Distance measures that do not demonstrate all the metric properties (semi-metrics and non-metrics; McCune & Grace, 2002) are useful, but less intuitive (e.g., negative distances, or distances between identical objects may be non-zero). Value-based properties focus on dissimilarities on the y-axis (differences in values; Figs 2.1-2.2), while time-based properties focus on dissimilarities on the x-axis (differences in time; Fig. 2.1).

2.3.1. Metric properties (adapted from McCune & Grace, 2002)

M1. Zero distance. d(X, X) = 0. The dissimilarity value between a time series and itself should be zero.

M2. Symmetry. d(X, Y) = d(Y, X). The dissimilarity value should be the same regardless of the order in which time series are compared, X to Y or Y to X. A distance measure without symmetry might, for example, cluster a collection of time series differently depending on how the time series are ordered. In the real world, distances within city road networks are often non-symmetric due to one-way streets. Animal

migration times might be non-symmetric if they are moving uphill in one direction and downhill in the other.

M3. Triangle inequality. $d(X, Y) \leq d(X, Z) + d(Y, Z)$. Given three time series, the distance between any pair of them should never be larger than the sum of the distances between the other two pairs of time series. This property is related to Euclidean geometry (one side of a triangle cannot be longer than the other two combined). A non-metric or semi-metric that does not satisfy the triangle inequality can cause errors for many clustering algorithms (Jacobs et al., 2000). On the other hand, some time series classification problems require a distance measure that does *not* satisfy the triangle inequality, e.g., when it is important to ignore outliers or whole subsets of observations (Weinshall et al., 1998). Matching many points to a single point, which allows for warping invariance (T3 below) would not be possible with a metric distance. Therefore, comparing animal calls or movement patterns or other time series that may have a similar pattern but with one time series stretched relative to the other may require a semi-metric (e.g., DTW) or non-metric for accurate classification.

M4. Non-negativity. $d(X, Y) \geq 0$. The dissimilarity value should never be less than zero. Mathematically, this must be true if properties M1, M2, and M3 are true. However, some distance measures that do not satisfy the triangle inequality can return negative dissimilarity values.

2.3.2. Value-based properties

V1. Translation invariance (also called amplitude shifting invariance or offset invariance; Fig. 2.1a). $d(X + q, Y) = d(X, Y)$, where q is any real number (Batyrshin et al., 2016). Increasing the value of all observations of one time series by the same amount q should not change the dissimilarity value. Translation *sensitivity* can be defined where the dissimilarity between X and Y increases relative to the value of q, and translation *insensitivity* where the dissimilarity between X and Y increases by an amount that is independent of q. Translation sensitivity can be measured in relative terms, allowing comparison between distance measures. Invariance to translation can be useful when time series have different starting values, e.g., time series of radiation spikes with different background levels. However, it is often easier to apply

a normalization or scaling adjustment so that all time series start at the same value. See Section S2.2 in Appendix 1 for a more detailed discussion of this property.

V2. Amplitude sensitivity (Fig. 2.1b). Translation sensitivity can be defined on a local scale (sensitivity to translation of a section of a time series) and in that case will be referred to as amplitude sensitivity. If a vertical shift transformation, $f(t) = t + q$, is applied to one or more observations t of time series X to form time series Y, $d(X, Y) > d(X, X)$ and $d(X, Y)$ increases with q (sensitivity). This could be important, for example, in determining deviations in the strength of seasonal temperature patterns. See Section S2.2 in Appendix 1 for a more detailed discussion of this property.

V3. White noise invariance (invariance against random noise; Fig. 2.1c). $d(X + f(X), Y) \approx d(X, Y)$, where f(X) is a function that adds a small pseudo-random value from a normal distribution with a mean of zero and standard deviation q to each observation of time series X (adapted from Lhermitte et al., 2011). Adding a random noise term to one time series from a pair should have an inconsequential effect on the dissimilarity value between them. A distance measure sensitive to white noise will show an increase in dissimilarity values relative to q, allowing us to obtain a relative measure of robustness against white noise. Robustness against white noise might be desirable, e.g., when comparing trends of stochastic processes, such as population growth.

V4. Biased noise invariance (invariance against non-random noise, i.e., noise in a single direction; Fig. 2.1d). $d(X + f(X), Y) \approx d(X, Y)$, where f(X) is a function that adds a small non-random value q to half of the observations (randomly chosen) of time series X (adapted from Lhermitte et al., 2011). Biased noise is different from random noise in that it is in a single direction and therefore more likely to be systematic. An invariance or low sensitivity to biased noise might be important, e.g., if comparing time series of vegetation density calculated from satellite images biased by differential cloud cover.

V5. Outlier invariance (Fig. 2.1e). $d(X + f(X), Y) \approx d(X, Y)$, where f(X) is a function that adds a large pseudo-random value q to a single randomly chosen observation of time series X. Outlier sensitivity is thus defined as the dissimilarity value increasing with q, and is a specific case of amplitude sensitivity limited to a single time point. Sensitivity

to outliers is useful for detecting anomalies or disruptive events, but robustness may be preferred where outliers represent measurement errors or irrelevant anomalies.

V6. Antiparallelism bias (see Fig. 2.2). Antiparallelism refers to line segments or trends which have slopes with the same value but opposite signs, while parallelism refers to those which have identical slopes in both value and sign. A distance measure with positive antiparallelism bias ignores the sign of the slope and treats antiparallel and parallel trend curves the same. A distance measure with negative antiparallelism bias treats trend curves with opposite signs as more dissimilar than those with identical signs. Distance measures with no antiparallelism bias (neutral) measure absolute differences on the y-axis, without respect to slope or direction. Mathematically, if Y = f(X), where f(X) is a function that reflects X across the axis of t_0 (for all t in X, $Y_{t_i} = 2X_{t_0} - X_{t_i}$), and Z = g(X), where g(X) is a function that applies a scale transformation to X relative to t_0 such that the absolute difference in summed values between Z and X is the same as that between Y and X (for all t in X, $Z_{t_i} = 3X_{t_i} - 2X_{t_0}$), then d(X,Z) > d(X,Y) for positively biased distance measures; d(X,Z) < d(X,Y) for negatively biased measures; and d(X,Z) = d(X,Y) for neutral measures. Whether and which kind of antiparallelism bias is desirable depends on the application. For example, it might be important to differentiate between positive and negative fluctuations from a baseline value of energy flow, which would require a distance measure with a positive or negative antiparallelism bias; conversely, if the only concern were the magnitude of fluctuation, a neutral distance measure might be preferred.

2.3.3. Time-based properties

T1. Phase invariance (Fig. 2.1f). $d(X_{i+p}, Y_i) = d(X_i, Y_i)$ (adapted from Lhermitte et al., 2011). Phase invariance is the x-axis equivalent of translation invariance. If all observations of X are shifted temporally by the same value p, it should not affect the dissimilarity value. Phase invariance may be a desirable property to detect similarities that occur separated in time. For example, when matching audio recordings of bird songs, it is likely that similar songs occur at different time points in different recordings. Conversely, when comparing population trends of different species within a community or geographical area to see which ones responded similarly to a

disruptive event occurring at time t, phase invariance is not a desirable property as responses should match in time.

T2. Time scaling invariance (Fig. 2.1g). $d(X_{pi}, Y_i) = d(X_i, Y_i)$ (adapted from Esling & Agon, 2012). If one time series is expanded or compressed along its time axis, the dissimilarity value should not change. This property is useful for certain applications, such as comparing animal behaviour patterns occurring at different speeds.

T3. Warping invariance (Fig. 2.1h). Time scaling invariance can be defined locally, i.e., involving the expansion or compression of one or more sections of a time series, rather than the entire series (Batista et al., 2011). If a function $f(S_i) = S_{pi}$ is applied to expand or compress S, where S is any subset of X, $S \subseteq X$, to form time series Y, then d(X,X) = d(X,Y). Warping invariance is particularly useful when matching similar time series which have plateaus or valleys of uneven lengths. For example, recordings of bird calls may have pauses of different lengths.

T4. Frequency sensitivity (Fig. 2.1i). If time series Y is obtained by applying the same transformation f(t) to one or more observations t of time series X, such that d(X, Y) > d(X, X), then the dissimilarity value will depend on the number of observations to which the transformation f(t) is applied. In other words, if a distance measure is sensitive to frequency, increasing the number of differences between two time series should increase the dissimilarity value. This could be important, for example, to rank a set of environmental time series according to the number of aberrations.

T5. Duration sensitivity (Fig. 2.1j). If time series Y is obtained by applying the same transformation f(t) to one or more consecutive observations of time series X, such that d(X, Y) > d(X, X), then the dissimilarity value will depend on the number of consecutive observations to which the transformation f(t) is applied. This property is a special case of frequency sensitivity. Distance measures which are sensitive to duration must be sensitive to frequency, but the converse is not true. Continuing the example from T4, ranking a set of environmental time series according to the number of aberrations without respect to the lengths of those aberrations would require a distance measure sensitive to frequency but not duration. See Section S2.2 in Appendix 1 for a more detailed discussion of this property.

2.3.4. Other properties

N1. Non-positive value handling. Some distance measures will not return results if the data contains negative values or zeros. This has implications e.g., for tasks such as classification, where it is common to first perform min-max normalization to rescale time series values to [-1,1].

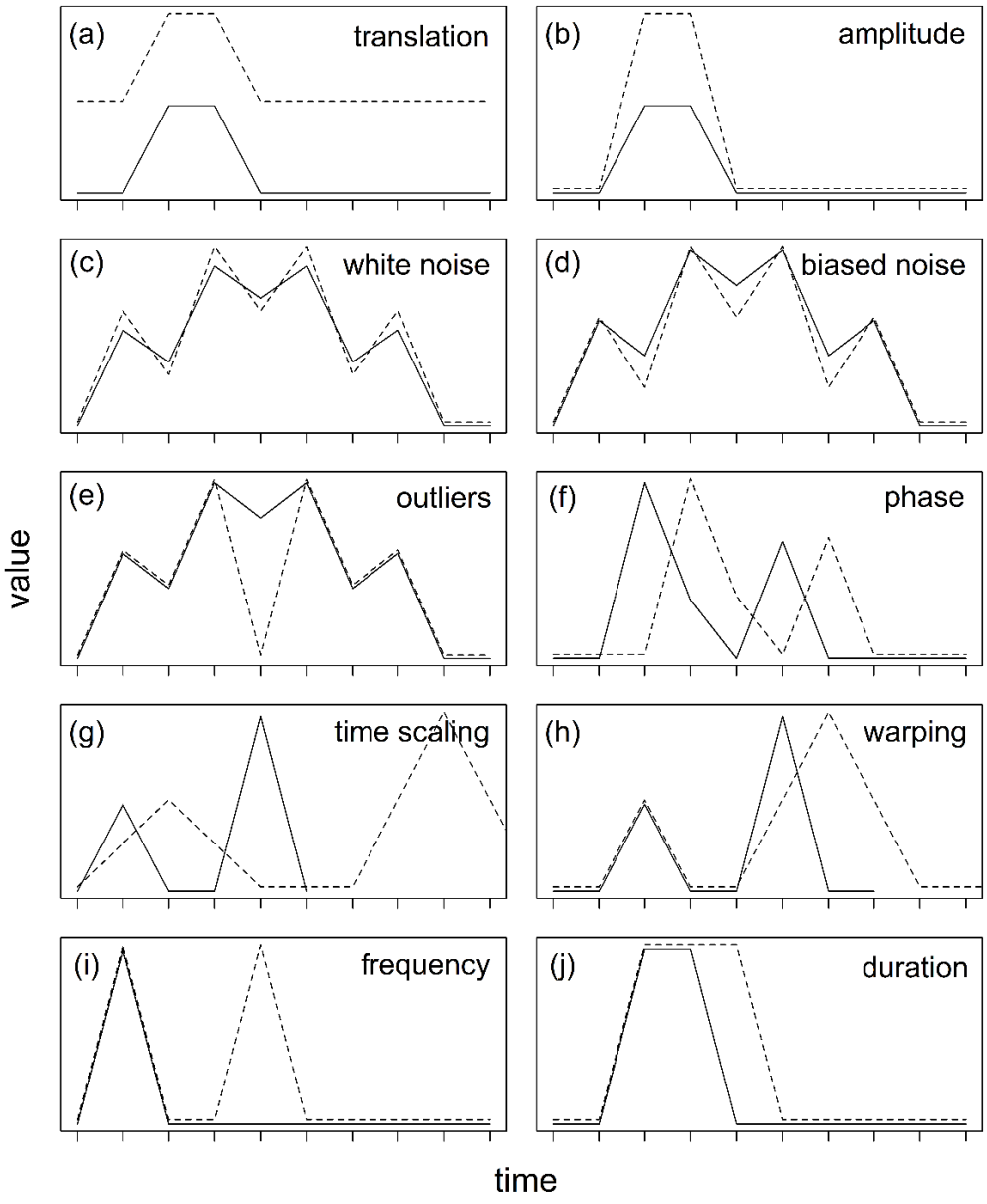


Figure 2.1. Illustration of time series distortions. They are used to demonstrate sensitivities or invariances of distance measures to a) translation; b) amplitude; c) white noise; d) biased noise; e) outliers; f) phase; g) time scaling; h) warping; i) frequency; and j) duration. A dissimilarity value of zero (or equivalent, for any distance measure not demonstrating uniqueness) between any of the illustrated pairs of time series would indicate an invariance to that type of distortion.

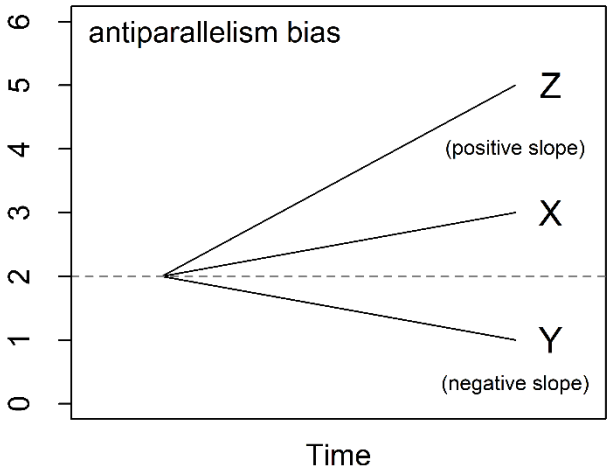


Figure 2.2. Illustration of antiparallelism bias. Time series X and Y are antiparallel (Y has the same slope as X, but in the opposite direction), while Z has a different slope than X, but in the same direction. The total difference in values between X and Z is the same as that between X and Y. Distance measures with positive antiparallelism bias rate time series X as more dissimilar to time series Z than to time series Y, while the opposite is true for those with negative antiparallelism bias. Distance measures with neutral antiparallelism bias rate the time series pairs as equally dissimilar.

2.3.5. Metric properties tests

The metric properties of some distance measures are specified in the literature, but for others it is unclear. Therefore, I devised a set of tests for metric properties. I confirmed the robustness of my tests by comparing my results to the literature for distance measures with known metric properties.

The test for uniqueness was conducted by comparing a time series first to itself, and then to a similar time series with a value difference at a single point. For distance measures with threshold settings (e.g., EDR), I set the threshold to zero to ensure they would recognize the difference. Any distance measure that returned a value of zero when comparing the time series against itself, and any non-zero value when comparing it against a time series with a value difference at a single point, was considered to demonstrate uniqueness.

Symmetry was tested by comparing a pair of different time series, X and Y, in both forward order, d(X, Y), and reverse order, d(Y, X). If the two values returned were identical, the distance measure was considered to demonstrate symmetry. I ensured that the time series

were different enough that no distance measure returned zero for both forward and reverse order.

The triangle inequality and nonnegativity properties were tested by comparing thousands of short, randomized time series generated by a stochastic exponential model. Shorter time series were better at detecting violations, so I set the length to five. I generated 300,000 time series and divided them into 100,000 sets of three. Within each set of three, I considered each time series to represent one corner of a triangle and compared them pairwise, with the resulting distances representing the sides of the triangle. I then subtracted the two shorter sides from the longest side. If the difference was greater than zero for any of the 100,000 sets, then the distance measure was considered to violate the triangle inequality. Additionally, if any of the 300,000 time series comparisons produced a negative value, the distance measure was considered to violate nonnegativity. I set the time series generator such that zeros and negative values were included in some time series, as some distance measures satisfy the triangle inequality and/or non-negativity only when all input values are positive or non-negative.

Distance measures were classified as 'Full' for full metric if they passed all metric tests, 'Semi' for semi-metric if they passed all tests except the triangle inequality, or 'Non' for non-metric if they failed one or more of the other tests.

Settings for adaptive distance measures (distance measures with settings that can be changed to alter their behaviour) were set at defaults given in examples from the documentation of the TSdist R package (Mori *et al*., 2016). For triangle inequality and nonnegativity tests, I kept the same settings for initial testing. If they passed the tests at those settings, I tested them over a range of settings. If they failed at default settings, there was no need for further testing.

2.3.6. Time-based and value-based properties tests

I performed two types of testing for non-metric properties in this study. Controlled testing was performed on sets of short, simple time series to clearly demonstrate specific properties. However, the demonstrated properties may not translate as clearly onto real-world datasets, and the behaviour of distance measures may vary depending on the types of

time series involved (see Lhermitte et al., 2011). Therefore, I employed uncontrolled testing by applying functions to real-world time series from the UCR Time-Series Classification Archive (Dau et al., 2019) to induce differences, then comparing the altered time series to their unaltered counterparts. I applied the functions over a range of parameters, then plotted the resulting curves to show how responses of distance measures vary with magnitude.

2.3.7. Controlled testing

I used the Manhattan distance as a basis for devising controlled sensitivity tests for translation, amplitude, duration, frequency, white noise, biased noise, and outliers. The Manhattan distance is the summed absolute difference between each pair of points in a time series. It is a simple-to-calculate metric and demonstrates all the sensitivities I tested for. Furthermore, it responds to sensitivity tests in a linear manner. These properties make the Manhattan distance an ideal basis for comparison of other distance measures.

For each sensitivity test, I constructed a series of five time series with linearly increasing differences, T_1, T_2, ..., T_5, such that the differences in absolute value between point pairs of any consecutive pair of time series summed to one. Thus, the Manhattan distance between any pair of consecutive time series, T_i and T_{i+1}, was one, and between any non-consecutive pair of time series, T_i and T_{i+j}, is j. For example, the Manhattan distance between T_1 and T_2 would be one, between T_2 and T_3 would be one, and between T_1 and T_5 would be five.

Sensitivity tests were conducted for each distance measure by comparing each time series T_i in the set T_1, T_2, ..., T_n, to T_1. Any distance measure returning a dissimilarity value of zero for *every* pair of time series for a given sensitivity test would be considered as invariant for that property, while a distance measure returning the same *non-zero* value for *every* time series pair would be considered as insensitive (note that invariance implies insensitivity, but insensitivity is not the same as invariance. Distance measures that demonstrate insensitivity to a property register differences as binary—different or not different—while those demonstrating invariance do not register differences at all).

Sensitivity is calculated as the mean of all distances between consecutive time series,

$$s = \frac{\sum_{i=1}^{n-1} d(T_i, T_{i+1})}{n} \ , \tag{2.1}$$

where s is sensitivity, $d(T_i, T_{i+1})$ is the distance between a pair of consecutive time series T_i and T_{i+1}, and n is the total number of time series being compared.

Given that s is an absolute sensitivity value, its interpretation is dependent on the scale of the distance measure. A scale-independent relative sensitivity is obtained by

$$rs_x = \frac{s_x}{s_\mu} \ , \tag{2.2}$$

where rs_x is the relative sensitivity to property x, s_x is the absolute sensitivity to that property, and s_μ is the mean of absolute sensitivities to all tested properties.

The sensitivity values for all distance measures are separated into five bins and designated as 'Very Low,' 'Low,' 'Medium,' 'High,' or 'Very High.' The sensitivity value for the Manhattan distance is one for every property and serves as the median value for the bins, which are: less than 0.2, 0.2 to 0.75, 0.75 to 1.25, 1.25 to 2.5, and greater than 2.5, respectively. Note, however, that the equation for sensitivity is derived from the linear slope equation, but the sensitivity for many distance measures is non-linear. The calculated sensitivity is a linear approximation along the tested range.

Phase invariance testing was conducted in a similar way to sensitivity testing, with T_1, T_2, ..., T_5 representing a set of time series, with the difference in phase increasing with i in T_i. However, the Manhattan distance could not be used as a basis for comparison. This is because lock-step distance measures (those that match every time point one-to-one), including the Manhattan distance, do not respond to time translation in a way that can be interpreted by a function. Distance measures were designated as 'Inv' (meaning they demonstrated phase invariance) when the dissimilarity between *every* pair of time series was 0, 'Ins' (insensitive) when *every* pair of time series returned the same *non-zero* dissimilarity value, 'Sens' (sensitive) when the dissimilarity value was dependent on i, or 'Unp' (unpredictable) when dissimilarity values differed but did not depend on i. For those distances with window size settings (e.g., some distance measures that act stepwise along time series have a setting to control how many time points are considered in each step), I

set the window large enough to cover the maximum difference in phase (that between T_1 and T_n).

Time scaling invariance was tested using a set of time series in which T_{i+1} was stretched compared to T_i. This involved lengthening the time series T_i, keeping the first and last time points the same while altering the values at each time point in between to fit the shape change. Warping was tested by stretching only one horizontal section of a time series, such that a set was formed, with T_{i+1} longer than T_i. As with phase invariance, results for time scaling invariance and warping invariance were not compared against the Manhattan distance. The Manhattan distance and other lock-step distance measures are unable to handle time series with different lengths and therefore by default are not invariant to uniform time scaling or warping. Thus, elastic distance measures were tested and designated as either 'Inv' (invariant) if *all* returned dissimilarities were zero, 'Ins' (insensitive) if *all* returned dissimilarities were *identical* and *non-zero*, 'Sens' (sensitive) if the returned value depended on the degree of time scaling or warping, or 'Unp' (unpredictable) if returned values differed but did not depend on the degree of time scaling or warping. All lock-step distance measures were designated as 'n/a.'

Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. Distance measures were designated as having 'Positive' bias if they gave a greater dissimilarity value to pairs of time series differing in the same direction than to pairs differing in the same direction, 'Negative' bias if they gave a greater dissimilarity value to those differing in opposite directions, or 'Neutral' if they assigned each pair of time series the same dissimilarity value.

2.3.8. Correlation between distance measures

I used the relative sensitivity values (before binning) for translation invariance, amplitude sensitivity, white noise invariance, biased noise invariance, outlier invariance, frequency sensitivity, and duration sensitivity to test for correlations between distance measures, to determine how similarly related and unrelated distance measures responded to my properties tests. First, I calculated the Pearson correlation between each pair of distance measures. I then separated the results into pairwise correlations of distance measures

within the same families and pairwise correlations of unrelated distance measures, and performed a Welch two sample t-test to determine if distance measures within the same family or group are more closely correlated than unrelated distance measures.

2.3.9. Uncontrolled testing

Uncontrolled tests were performed on two real-world time series (Fig. 2.3) from the UCR Time-Series Classification Archive (Dau et al., 2019), an archive of 128 time-series datasets intended for testing of classification algorithms. One was randomly selected from the Yoga dataset and represents body movement during pose transitions. Captured images of actors were converted to one-dimensional time series by calculating the distance between the outline and its centre. The other time series was randomly selected from the Synthetic Control dataset and is a synthetically generated pattern designed to be quantifiably similar or dissimilar to other time series in the dataset. Neither of these are ecological time series, but it does not matter for the purpose of generalized testing.

I created a function for each property to be tested, which applies a transformation to one or more time points of a real-world time series. Each function accepts a value q, the purpose of which varies depending on the function (see section S2.3 in Appendix 1 for details). For example, the translation function adds a real number q to every observation value of a time series. The transformed time series is returned as output and compared against its unaltered counterpart. I applied the functions to a range of q in increments, then graphed the results as response curves (Figs S2.2-S2.5 in section S2.6 of Appendix 1). I did not compare them against a reference or assign sensitivity ratings as they were intended only as a confirmatory check against the results of controlled testing.

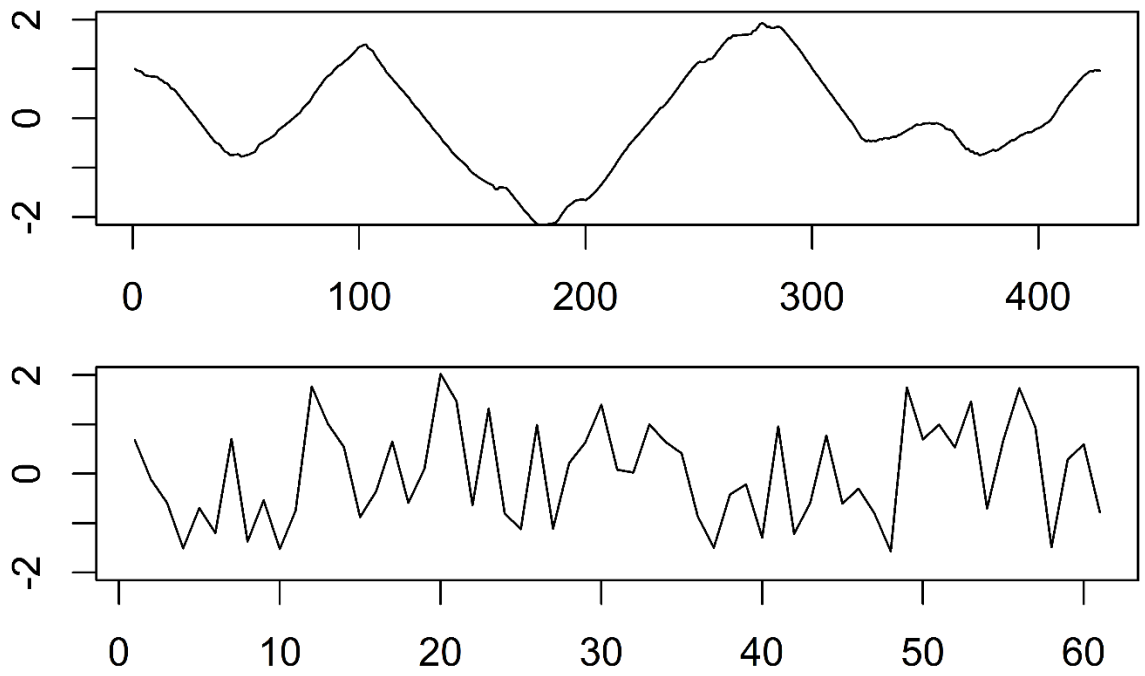


Figure 2.3. Examples of time series used for uncontrolled testing. One time series from each of the Yoga (top) and Synthetic Control (bottom) datasets of the UCR Time-Series Archive (Dau et al., 2019). Time series in the archive are z-normalized. Therefore, I applied a translation shift before testing to ensure compatibility with distance measures that are unable to handle zeros or negative values.

2.3.10. Selection process

I devised a three-step selection process to guide researchers through determining the most appropriate distance measure(s) for their intended application. The selection process utilizes a set of purpose-built tools that I created by combining the results of my properties tests with existing knowledge from the literature (especially Esling & Agon, 2012). The first step is to use a decision tree (Figs 2.11-2.12) to select a general category of distance measures. Step two is to use Table 2.1 to determine which pre-processing steps might be necessary to prepare the dataset and/or to further narrow the choice of distance measures. The final step is to determine which properties will be most important to achieve the desired outcome and use Figs 2.4-2.6 to narrow the selection to the distance measure(s) that exhibit these properties.

2.3.11. Real-world example dataset

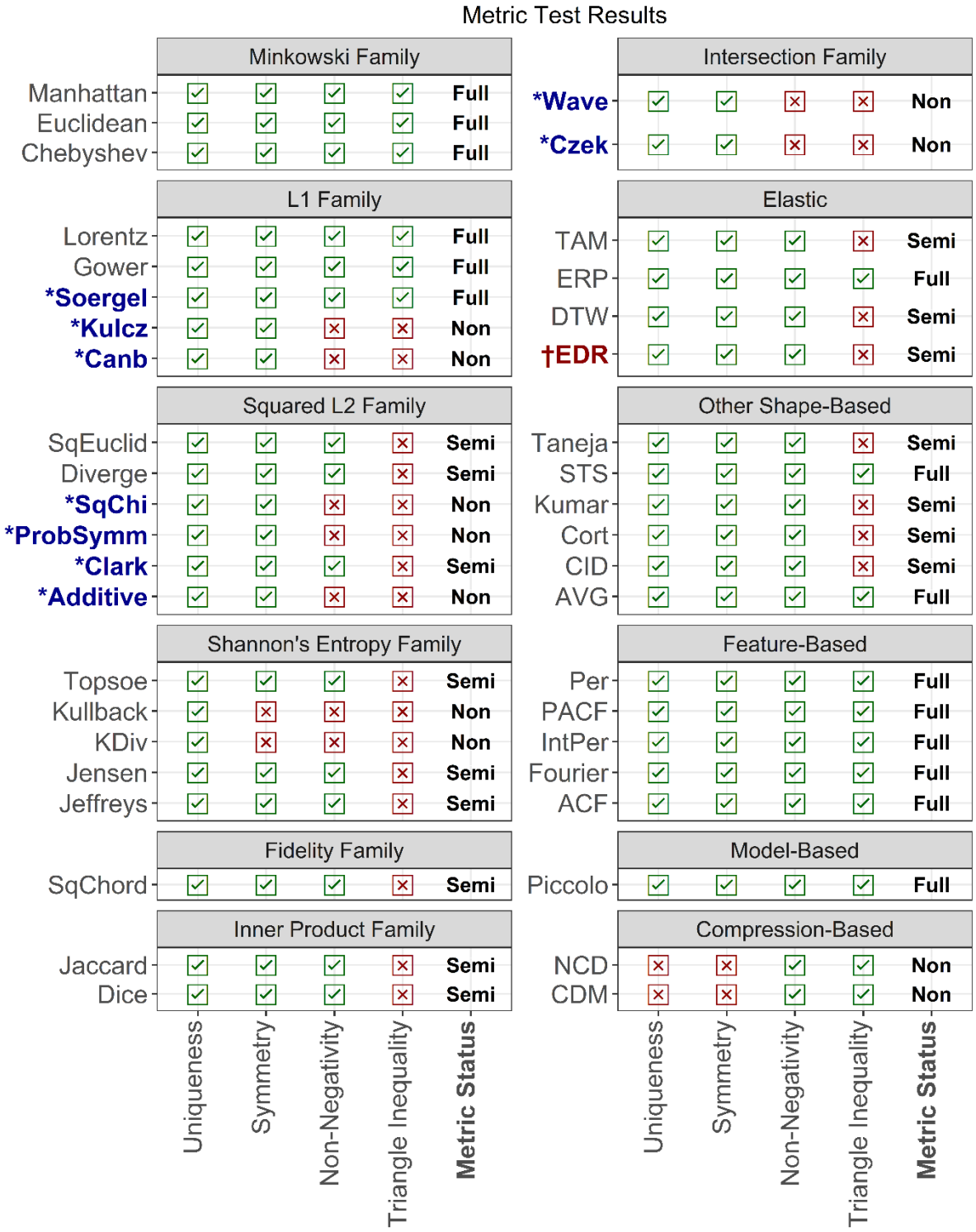
To demonstrate the selection process and add real-world context, I used a dataset from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021). The dataset consists of 25 years of breeding pair count data for five wading bird species, from within and outside of reserves. The within-reserves data came from 47 RSPB lowland wet grassland reserves, while the counterfactual (outside of reserves) data was selected from the UK Breeding Bird Survey data. Data were matched to select sites that represent how reserve land would look in the absence of conservation measures. The reserve and counterfactual count data were aggregated into species trends, then converted to indices by dividing each annual species count total by the first-year species count total. Thus, each of the five bird species was represented with a reserve trend index and a matched counterfactual trend index. Jellesmark et al. (2021) compared each pair of indices to determine the effects of conservation efforts on each bird species, by calculating the percentage improvement of reserve indices over counterfactual indices and performing t-tests to determine significance and effect size of the difference. I ranked the results of Jellesmark et al. (2021) according to both percentage improvement and effect size. I then applied my selection method to select appropriate distance measures, ranked the dissimilarity results returned by each selected distance measure, and examined the rankings with respect to Jellesmark et al. (2021). I also ranked the results returned by unselected distance measures for comparison.

2.4. Results

2.4.1. Metric test results

Fourteen out of 42 distance measures were identified as full metrics, meaning they passed the metric tests for uniqueness, symmetry, non-negativity, and the triangle inequality (see Fig. 2.4). Sixteen distance measures were identified as semi-metrics (failed the triangle inequality test but passed the other three tests) and 12 were identified as non-metrics (failed at least one of the tests for uniqueness, symmetry, or non-negativity; Fig. 2.4). However, in some cases results depended on settings or input values (some distance measures passed the triangle inequality and/or non-negativity tests only when inputs were

constrained to non-negative real numbers). All tested feature-based and model-based distances were full metrics, while all tested compression-based distances were non-metrics. Shape-based measures showed mixed results, even within families and groups. See section S2.4 in Appendix 1 for additional results.



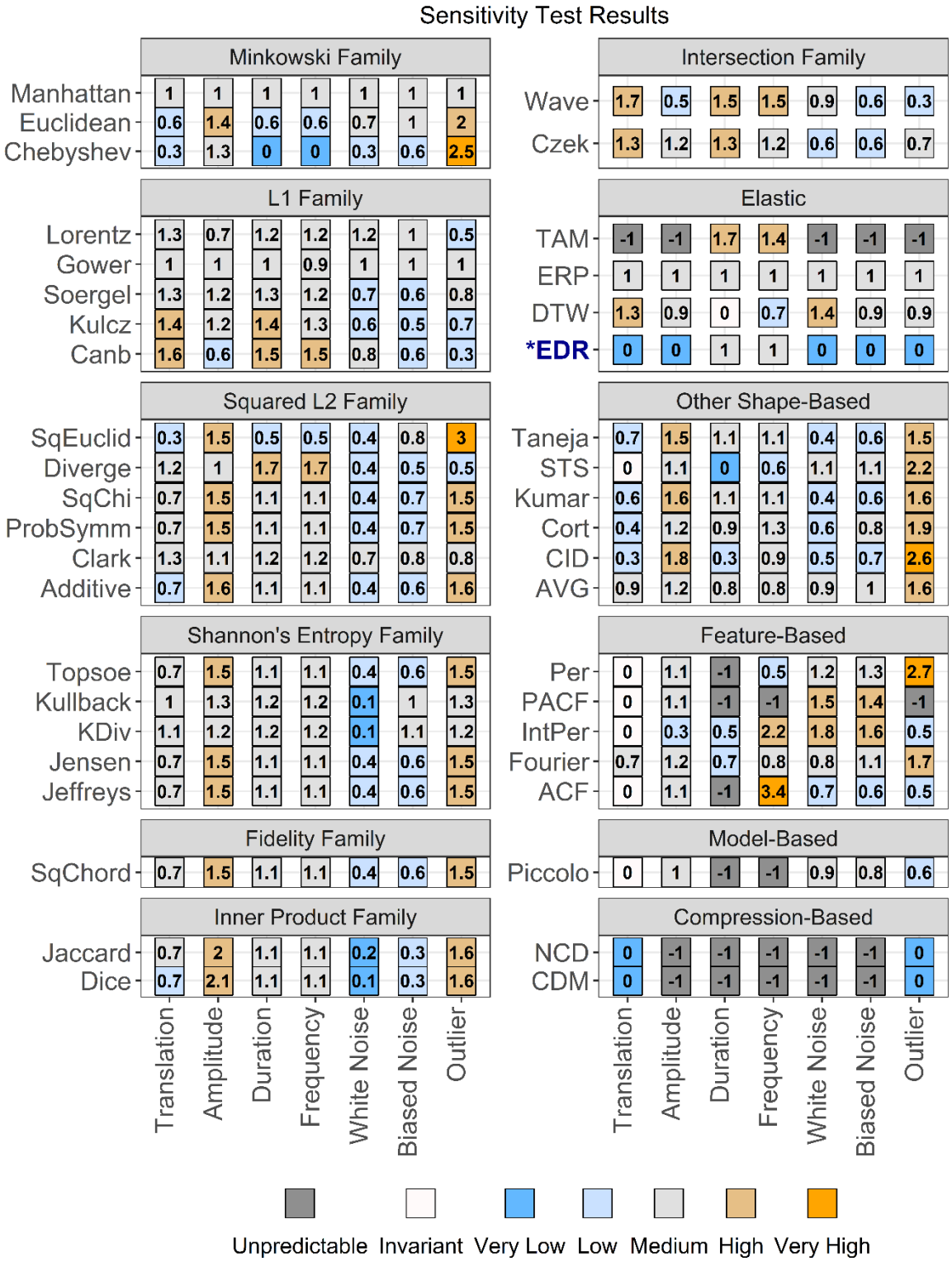
*These distances respond differently when inputs are constrained to non-negative real numbers. As we included negative values in our tests, our results for these measures may differ from others (e.g. Kocher and Savoy, 2017).

†This distance is a full metric when the threshold value (epsilon) is set at 0.

Figure 2.4. Metric test results for 42 distance measures. Results are arranged by family (for lock-step shape-based measures) or type.

2.4.2. Sensitivity test results

Lock-step shaped-based measures varied in the strength of responses to the sensitivity tests, but none tested as unpredictable and only two (the Chebyshev distance and the Short Time Series, or STS, distance) showed any invariances or insensitivities (Fig. 2.5; see Figs 2.1-2.2 for illustrations of the time-series distortions I used to test for sensitivities and invariances). The Welch two sample t-test shows that correlations between distance measures within families or groups (mean Pearson correlation = 0.48) are significantly stronger than between unrelated distance measures (mean Pearson correlation = 0.15): $t = 5.5$, $df = 82.3$, $p < 0.001$. However, not all related distance measures were closely correlated (see Fig. S2.1 in section S2.5.3 of Appendix 1), nor were there clear differences between families of distance measures. Elastic, feature-based and model-based distances showed greater variation in responses, with insensitivities, invariances, and unpredictability being common. The two compression-based distances I tested responded unpredictably to all controlled tests except translation and outliers; they responded unpredictably to *all* uncontrolled tests without exception. See Section S2.5.1 in Appendix 1 for additional results.



Sensitivity Ranges: Very Low: < 0.2, Low: 0.2 - 0.7, Medium: 0.7 - 1.3, High: 1.3 - 2.5, Very High: > 2.5.

*The results for EDR strongly depended on the threshold setting, epsilon. Here, it was set to 0.1.

Figure 2.5. Sensitivity test results for 42 distance measures. Results are arranged by family (shape-based measures) or type, and colour-coded according to sensitivity value.

2.4.3. Time-based invariances and other test results

All distance measures except the Time Alignment Measurement (TAM) distance responded unpredictably to phase invariance testing (Fig. 2.6; see Figs 2.1-2.2 for illustrations of the time-series distortions I used to test for sensitivities and invariances). TAM was sensitive to phase changes, however the response curve in uncontrolled testing was not smooth, suggesting some level of unpredictability. The Edit Distance with Real Penalty (ERP) distance was sensitive to uniform time scaling, while all other distances either responded unpredictably or were unable to be tested due to an inability to handle unequal-length time series. Warping sensitivity was more common, occurring in three elastic distance measures. DTW tested as invariant to warping and was thus the only distance measure I tested with any time-based invariances. Elastic measures were the only group of distance measures that showed any predictable time-based sensitivities or time-based invariances.

Two distance measures in the Shannon's entropy family were unable to deal with zeros, while the entire family was unable to deal with negative values. Three other lock-step shape-based measures also showed an inability to deal with negative values. Antiparallelism bias showed no obvious group-based patterns, but negative antiparallelism bias was most common and positive bias was least common. See Section S2.5.2 in Appendix 1 for additional results.

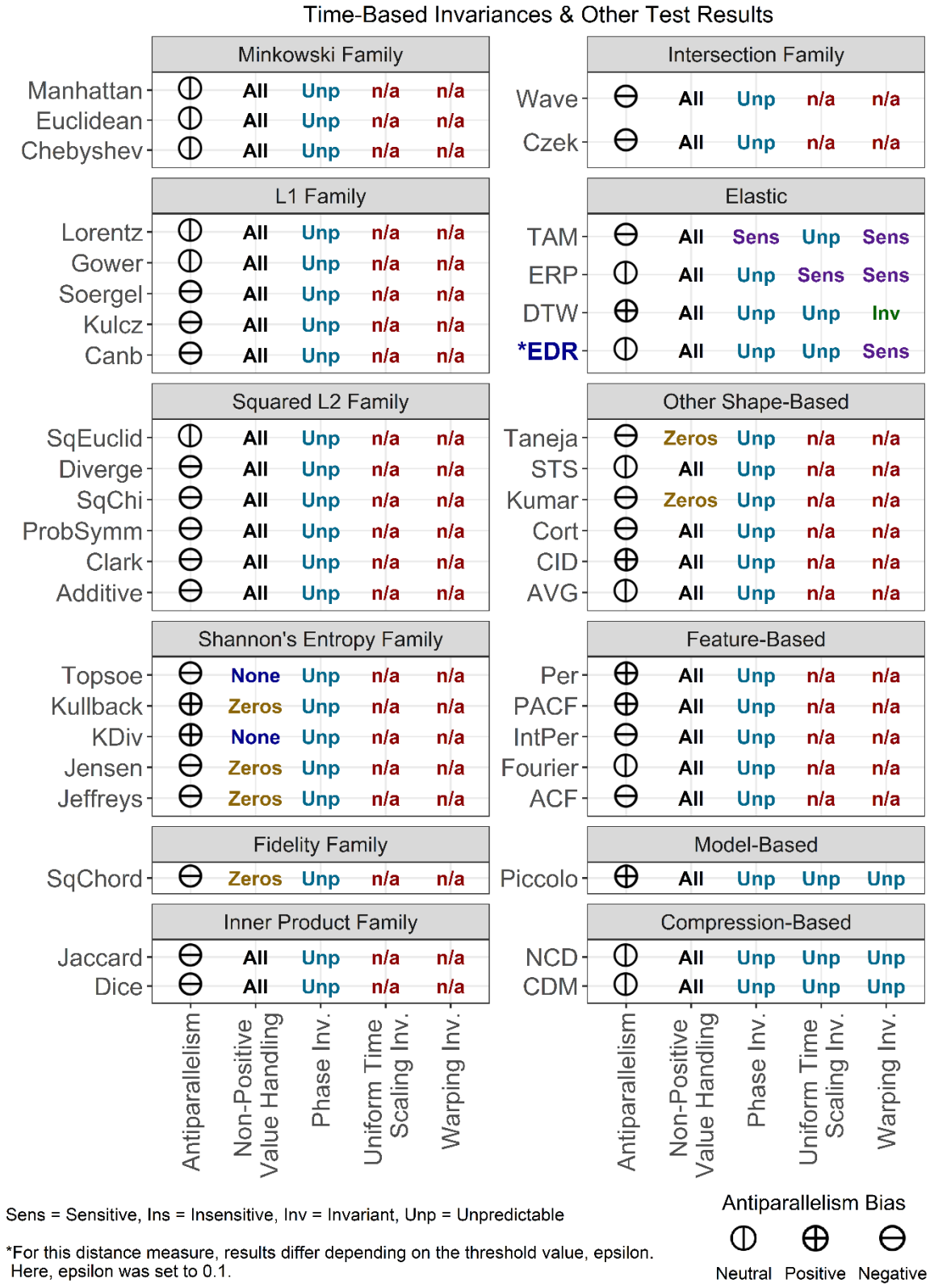


Figure 2.6. Test results for antiparallelism bias, non-positive value handling, and time-related invariances for 42 distance measures. Results of 'n/a' for uniform time scaling invariance and warping invariance mean that the distance measure in question is unable to handle unequal length time series and therefore could not be tested for those properties.

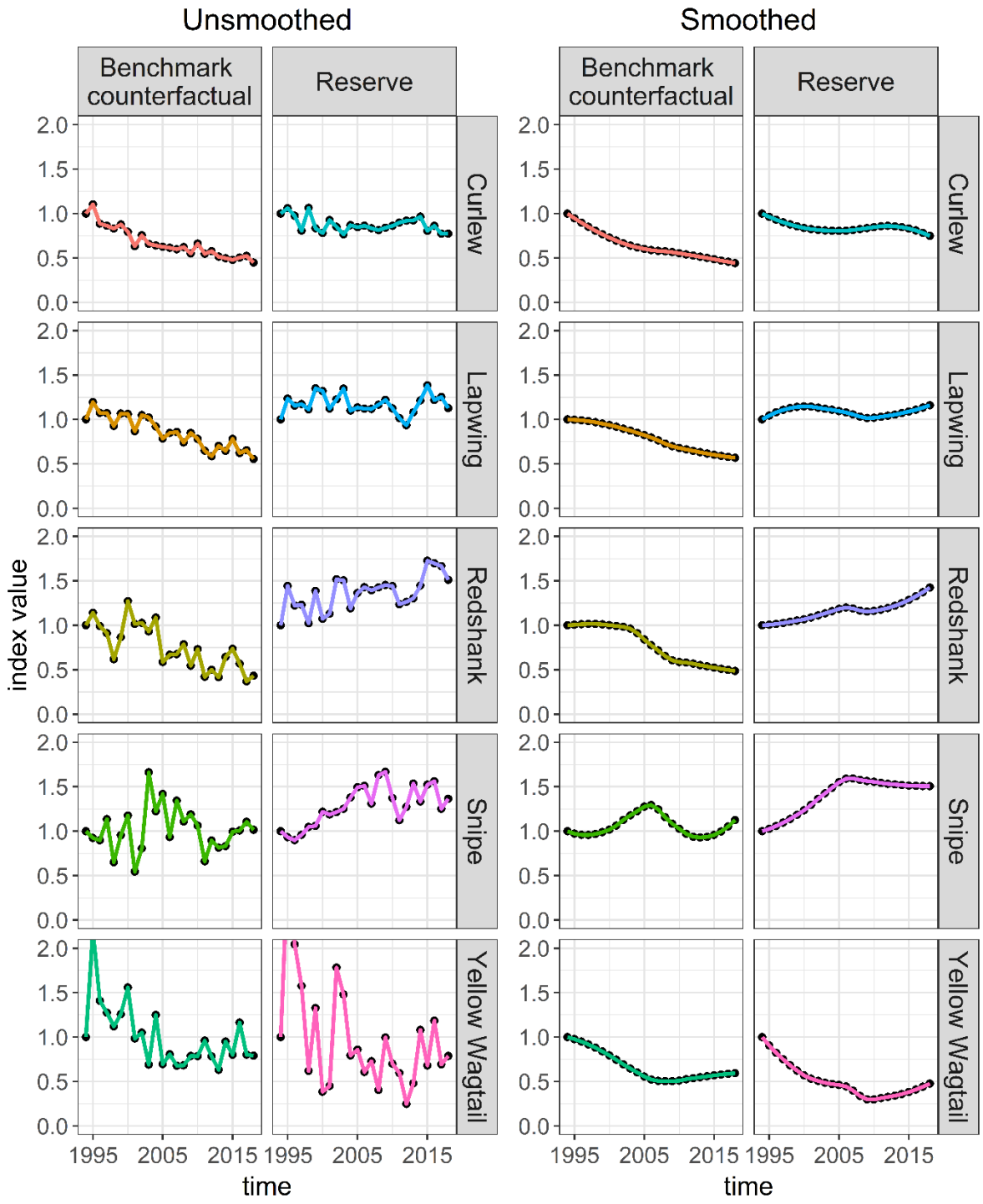


Figure 2.7. Reserve and counterfactual trends for five wading bird species that breed on RSPB lowland wet grassland reserves in the UK. Left: Unsmoothed trends based on original data presented in Jellesmark et al. (2021). Right: LOESS smoothed trends with a span setting of 0.75.

2.4.4. Selection process

The distance measure selection process I describe and demonstrate here was developed using the results from this study in combination with existing literature, and is intended to be useful for any dataset and task the user might have in mind. The first step in the selection process should be to determine the task to be performed. Applications for time series comparisons typically fall into the four main categories described in the introduction: clustering, classification, prediction, and anomaly detection. However, there are other less well-known applications, such as content queries, hypothesis testing, and accuracy assessment. Distance measures can also be used for pattern matching against databases to identify animal species or biological or ecological events from recorded or streaming data sources, such as video, audio, photographs, motion capture, temperature monitors, or other types of sensors. In addition, there are many other types of time series that one might wish to compare, such as activity patterns, biomass, nutrient uptake, growth rates, entropy, etc.

Both the dataset and the intended task are important in selecting an appropriate distance measure. For example, in classification, generally the entire shape of the time series is important, while anomaly detection might work best with distance measures that are especially sensitive to outliers. Classifying bird species according to their songs may require flexibility on the time axis (e.g., warping invariance), while clustering fish populations according to changes in biomass over a set time period does not.

I demonstrated the process of selecting an appropriate distance measure using a real-life example dataset from a study that used trends from wading birds inside and outside of reserves to determine the conservation impact of reserves (see detailed description in Section 2.3.10; also Jellesmark et al., 2021). A greater difference between the trend within reserves and the corresponding counterfactual trend outside of reserves means greater conservation impact on a given wading bird species. I chose this example because it is a type of application that many readers will be unfamiliar with in the context of distance measures, and because the results can be compared with other methods.

I began by examining the wading bird dataset in context of the decision trees in Figs 2.11-2.12. The dataset consisted exclusively of short (25 data points), non-stationary time series.

Following Fig. 2.11, I focused on shape-based distance measures, which compare raw data values. As the time series were of equal-length, in phase, using the same time scale, and without any missing data points, both lock-step and elastic measures would be appropriate (Fig. 2.12).

Next, I worked through Table 2.1. As the wading bird trends were indexed to a starting value of one (Fig. 2.7), they had the same starting value and the same value scale. There were no negative values because the trends were indexed and based on wetland bird counts; nor were there any zeroes. However, I did notice that some of my time series were noisy (Fig. 2.7), which could obscure the trends. Noise is a common characteristic of population data, largely due to the stochasticity of population dynamics and the environmental variables they depend on (Vasseur & Yodzis, 2004). While this noise is often white (random, uncorrelated), biased 'red' noise (positively autocorrelated, tending toward a single direction) is also common, e.g., when environmental conditions are above or below average for an extended period (van de Pol et al., 2011; Vasseur & Yodzis, 2004). Biased noise is therefore more likely to represent a legitimate difference in trends. There are multiple ways to deal with noisy time series (Table 2.1). I first tried the properties-based solution (Table 2.1; see below for the pre-processing solution). Using Fig. 2.5, I filtered out all shape-based distance measures with a white noise sensitivity category of medium or higher (a sensitivity value of 0.7 or more). Next, I required biased noise to be at least two categories higher in sensitivity than white noise (Fig. 2.5; e.g., if white noise sensitivity was very low, biased noise sensitivity must be at least medium). My choices here were based on practicality; sensitivity categories are arbitrary (I categorized them for convenience), so I wanted to avoid being too specific while ensuring that any chosen distance measure exhibited a non-trivial difference in sensitivity between white noise and biased noise.

Finally, I considered the remaining properties in the context of my intended task and desired outcome. I deemed amplitude sensitivity to be important, as I was interested in the overall divergence between population indices within and outside reserves. Duration sensitivity was also important, as I considered population indices which diverged more steeply or for a longer period to be more different, i.e., that conservation measures had a stronger effect on these species. Therefore, both amplitude and duration sensitivity had to be at least low (a

sensitivity value of 0.2 or higher; Fig. 2.5). Again, I could have chosen a different (higher) category, but I was more concerned with making sure the distance measures exhibited *some* sensitivity to these properties than the exact degree of sensitivity. I did not filter for antiparallelism bias, as the high stochasticity in some of the time series (Fig. 2.7) would dilute the signal too much for it to matter.

This selection process left me with two distance measures: the K-Divergence (KDiv) and the Kullback-Leibler distance (Kullback), both of which returned the same rankings that Jellesmark et al. (2021) obtained using percent improvement (Figs 2.8-2.9). Only two of the 40 unselected distance measures, the Edit Distance for Real Sequences (EDR) and TAM, returned the same rankings (Fig. 2.9).

Another way of dealing with noisy time series is by applying a smoothing algorithm (Table 2.1). I applied a LOESS smoothing algorithm (span = 0.75) to all time series in the dataset to remove the noise and reveal the trends (Fig. 2.7). I then re-ran the selection process using the same settings, except that I did not filter for noise sensitivity, and I added a filter for antiparallelism bias. Antiparallelism bias is not very important when dealing with highly stochastic time series because the signals for slope and direction are muddied by noise; however, smoothing introduces strong positive autocorrelation, making the slope and direction signals clear. I selected neutral for antiparallelism bias (Fig. 2.6) because I was more interested in relative differences in the population indices than the direction of change.

I was left with seven distance measures: ERP, the Euclidean distance, the Manhattan distance, the Gower distance, the Lorentzian distance (Lorentz), the Average distance (AVG), and the Squared Euclidean distance (Sq. Euclid). All seven selected distance measures agreed on the following order: Redshank, Snipe, Lapwing, Curlew, Yellow Wagtail (Figs 2.8 & 2.10). Four of the 35 unselected distance measures returned the same results (Fig. 2.10).

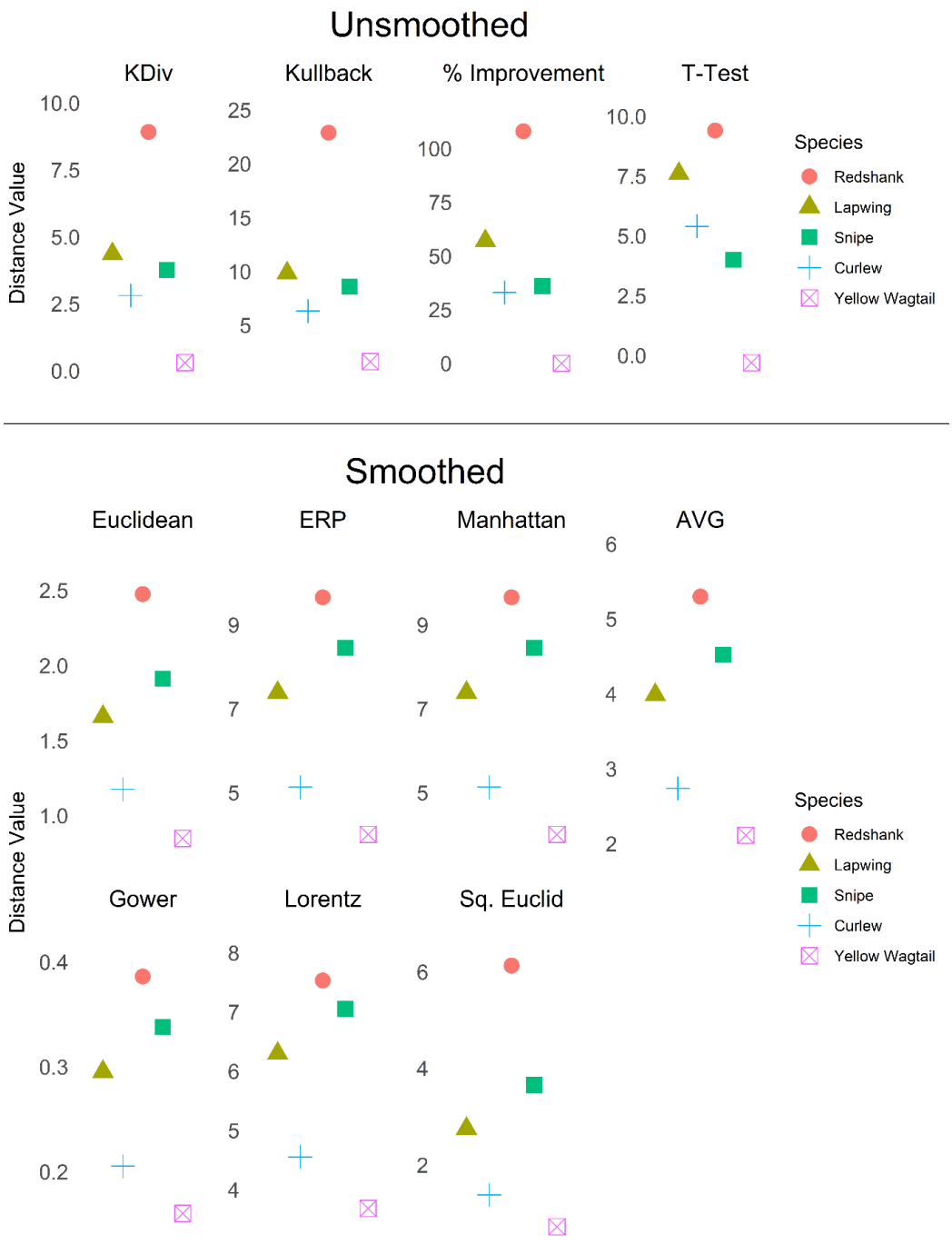


Figure 2.8. Comparative rankings of conservation impact on five wading bird species. Values on the y-axis represent the distance between unsmoothed (top) or LOESS smoothed (bottom) reserve and counterfactual trends for each species. Results are from the distance measures chosen by my selection process, as well as the percent improvement and t-test methods (top) used by Jellesmark et al. (2021). Percent improvement is the difference (multiplied by 100) between the final year index values of the two trends (within and outside of reserves) for a given bird species, while the t-test represents the results of a Welch 2-sample t test between the two trends.

Figure 2.9. Comparative rankings of conservation impact on unsmoothed trends of five wading bird species. Species are ranked according to percent improvement, t-test, and distance measures. Species ranked first had the greatest difference between trends. *Starred distance measures were chosen by my selection process. †Daggered distance measures were not chosen but returned the same results as the chosen measures.

Figure 2.10. Comparative rankings of conservation impact on smoothed trends for five wading bird species. Species are ranked according to percent improvement, t-test, and distance measures. Species ranked first had the greatest difference between trends. *Starred distance measures were chosen by my selection process. †Daggered distance measures were not chosen but returned the same results as the chosen measures.

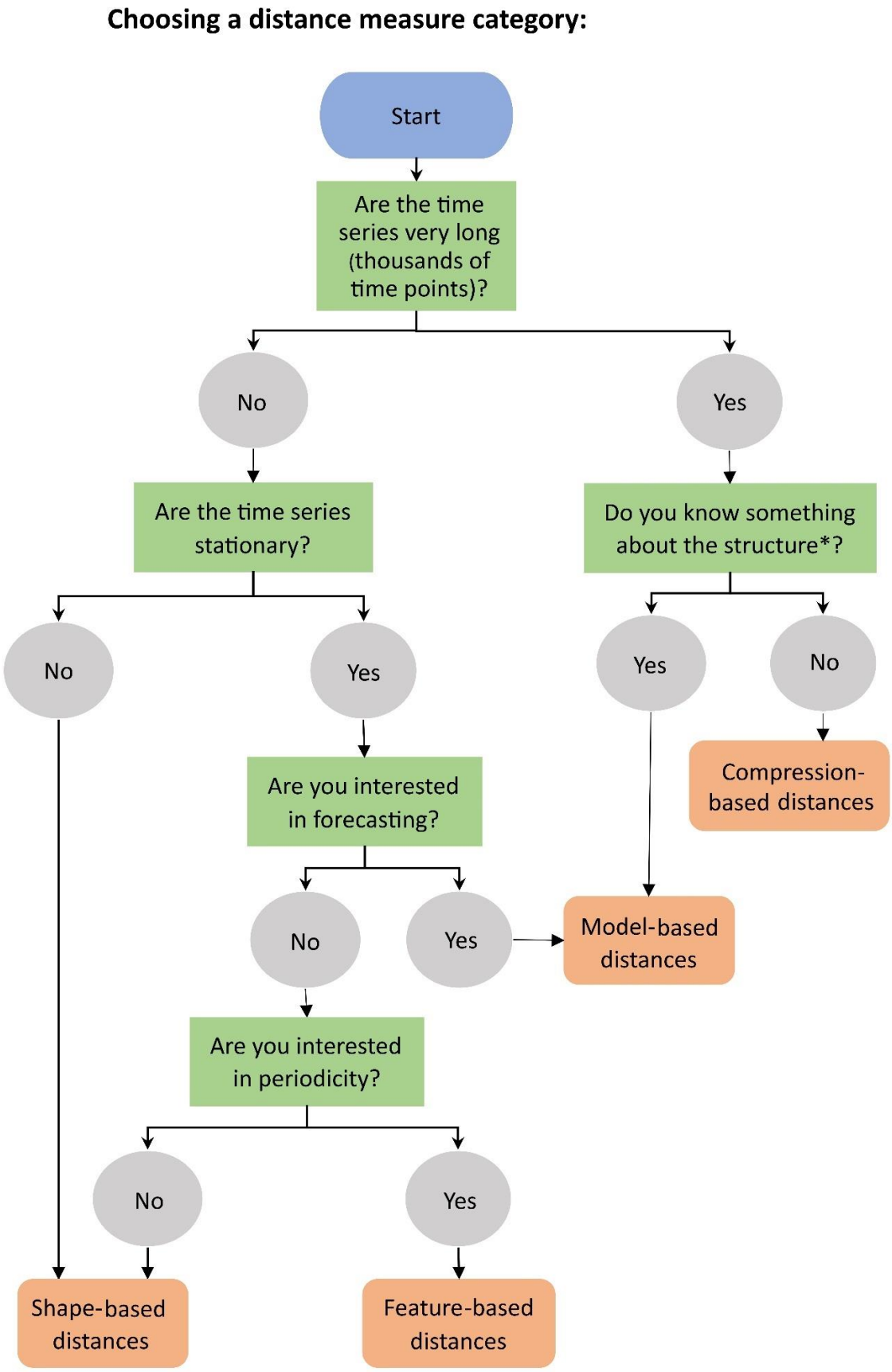


Figure 2.11. Decision tree to aid in choosing a distance measure category. *Structure refers to trends, repeated patterns, spikes, etc.

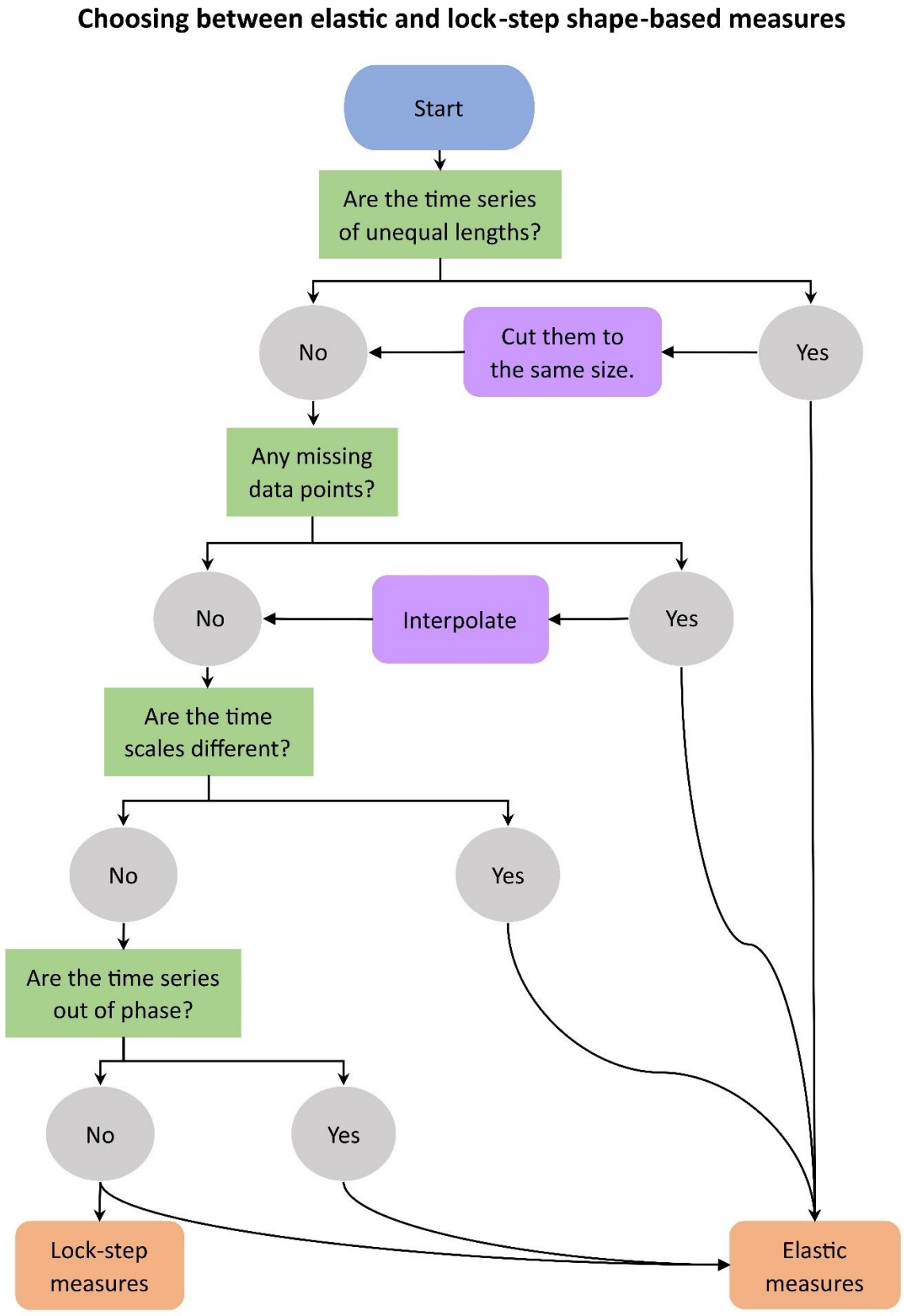


Figure 2.12. Decision tree to aid in choosing a sub-category of shape-based distance measures.

Table 2.1. Solutions to potential issues in the data. Note that choice of invariance or sensitivity as a solution should depend on whether the difference in question is important.

Problem	Pre-processing solution	Properties-based solution
Missing data points	Interpolate missing values.	Choose an elastic distance. They handle gaps through one-to-none or one-to-many point matching.
Different starting values but similar value scales	Apply a translation shift.	Choose a distance measure invariant (or sensitive) to translation.
Different value scales	Normalize or standardize data.	
Zeroes or negative values	Transform data to obtain positive values.	Choose a distance with non-positive value handling.
Noise	Apply a smoothing algorithm.	Choose a distance measure robust (or sensitive) to the type of noise that is of concern.
Out of phase		Choose a phase invariant (or phase sensitive) distance measure.
Unequal lengths	Cut all time series to the same length.	Choose an elastic, model-based or compression-based distance measure.
Different time scales		Choose a distance measure invariant (or sensitive) to uniform time scaling.
Nonuniform sampling intervals	Interpolate intermediate values.	Choose a distance measure that incorporates temporal information, such as the STS distance.

2.5. Discussion

The aim of this study was to provide enough information to make informed, objective decisions about which distance measures to use. I tested 42 distance measures for 16 properties and presented an objective method of selecting distance measures for any task based on those properties. I demonstrated the viability of the method on a real-world dataset by selecting distance measures to rank differences between pairs of wading bird population trends (within and outside of reserves) and showing that the distance measures I selected were fit-for-purpose and consistent in their rankings. The method is user-directed;

therefore, success depends on an understanding of the dataset, the task to be performed, and the hoped-for outcome.

Time series length and stationarity inform what category of distance measures the user should focus on (Fig. 2.11). Shape-based distances are best for short time series with differences that are easy to visualize, while longer, stationary time series may be better suited to feature-based, model-based, or compression-based distance measures (Esling & Agon, 2012).

The majority of distance measures I tested are lock-step measures. While I have categorized many of them by family, it is not evident from my testing that there is enough similarity between distance measures within families for this categorization to be of much use. While there are clear differences in sensitivities (Fig. 5) between lock-step measures, they share a rigidity in their treatment of time, comparing all point pairs 1-to-1, and most lack invariances. This makes them best-suited to applications where sampling is repetitive (e.g., yearly) and standardized in time, such as long-term population trends. Elastic measures, such as DTW, have tremendous flexibility due to their ability to match multiple time points to a single time point, and are therefore best used when time series have different time structures, such as recordings of animal calls or movements.

The broadest difference in use-cases occurs between shape based and non-shape based distance measures. Feature-based and model-based measures are typically used to compare stationary time series, which are time series characterized by repeating patterns rather than stochasticity. Model-based and feature-based measures identify particular aspects of these repeating patterns, thus their uses tend to be more specific than shape-based measures. They are especially useful for prediction, as repeating patterns can be forecast into the future (prediction is generally not applied to non-stationary time series, as stochasticity is by definition unpredictable). For example, they might be used to classify or predict time series of environmental parameters (temperature, pollution, etc), or events or changes that fluctuate or reoccur seasonally or diurnally, and are therefore likely to be stationary. Compression-based measures are designed to be extremely general and can theoretically be applied to any kind of time series. However, in practice I did not find them to be of any use on the time series I used for testing. They were unpredictable and did not demonstrate their

purported metric properties. They are better suited to much longer time series (many thousands or even millions of time points), but there is little I can say about them here.

The results of my properties tests showed a variation in strength of sensitivity to different properties in different distance measures (Fig. 2.5), although most distance measures were highly sensitive to outliers (Fig. 2.5). Invariances were uncommon among the distance measures I tested (Figs 2.5 & 2.6), although several distance measures did demonstrate invariance to translation (Fig. 2.5). Some distance measures, such as EDR and ERP, have tuning parameters that may affect their behaviour. In the case of ERP, these parameters can determine whether and how sensitive it is to missing values, while in the case of EDR, the threshold setting determines how far apart values must be to be considered different, and therefore serves to toggle responses to multiple properties between invariance and sensitivity.

When dealing with time series of unequal length or missing data points, distance measures that allow unequal matching (e.g., matching multiple points to one point), such as DTW, or that allow gaps, such as ERP, may be the solution. Alternatively, pre-processing of data may remove such concerns. For example, missing data points can be filled in by interpolation, or longer time series can be cut to the same length as shorter ones (only attempt such solutions if they make sense for the data).

Elastic measures, such as DTW, EDR, and ERP, are the most versatile distance measures, able to handle many common complications of datasets with little or no pre-processing. For general tasks, they are often a good option (see Figs 2.11-2.12). However, for tasks involving large datasets containing thousands of time series, some elastic measures may be impractical due to processing speed. Much of the research into speeding up time series comparisons for large datasets has focused on a select few distance measures, especially the Euclidean Distance and DTW. While the Euclidean Distance is faster, better known, and still widely used in some fields, an extensive body of research has shown DTW to be more accurate (Dau et al., 2019; Paparrizos et al., 2020; Zhu et al., 2012) and it is considered the *de facto* standard for accuracy in classification (note that it is still important to consider the properties of DTW in relation to the data, as it does not perform well in every case). Despite this, it is rarely used in ecology (Hegg & Kennedy, 2021). Note, however, that DTW is

computationally expensive and therefore can be slow for large datasets (for discussion on ways to speed up DTW, see Section S2.8 in Appendix 1).

For many analyses involving distance measures, researchers may first want to normalize or standardize their data or translate it along the y-axis. This may be an important step if the time series use different scales or have different starting values. For example, when performing classification or clustering tasks, it is common to apply z-normalization to rescale time series to a mean of zero and standard deviation of one (Rakthanmanon et al., 2013). Min-max normalization to a scale of [0,1] or [-1,1] is also common for datasets that are not normally distributed. Be aware, however, that these transformations may affect the subsequent choice of distance measures, as some cannot handle zeros or negative values and some metrics are non-metric when there are negative values present (see Fig. 2.4).

Although I ignored the metric properties of distance measures for my real-world example, they are very important for some tasks. For example, many algorithms for classification and clustering are designed to work only in metric space and may return unexpected results for non-metric distances, while some classification and clustering problems require a semi- or non-metric to get meaningful results (Weinshall et al., 1998). Another thing to be aware of is that output values (distances) returned by distance measures can be on dramatically different scales. Some, such as the Jaccard distance, are confined to [0,1], while others go to positive infinity [0,∞) (e.g., the Euclidean distance), or even include negative values (any distance that does not satisfy non-negativity, e.g., the Canberra distance). Depending on the intended application, the output scale could affect analysis, so may be worth considering.

Noise is a common aspect of ecological time series, as environmental and population dynamics are stochastic. There are several potential ways to deal with noisy time series. Some distance measures, such as EDR, have threshold settings; any difference between time series that falls below the threshold will be ignored. If the noise is relatively uniform in amplitude, this may be a simple solution if the distance measure in question meets all other requirements. Other distance measures, such as KDiv, are relatively robust against white noise although lacking a sensitivity setting, and may be more appropriate if the noise is less uniform. A more drastic solution is to apply a smoothing algorithm as a pre-processing step, though this should be approached with caution. Smoothing will remove noise and outliers

but may distort the time series and increase bias in the process. Therefore, it is important to avoid over-smoothing. Smoothing time series that have sudden and/or drastic value changes may also be problematic, particularly if these changes are an important aspect of differentiation between time series.

My demonstration using wading bird trends from Jellesmark et al. (2021) served to illustrate both the potential benefits and complications introduced by smoothing. When I filtered by noise sensitivity, I was left with two distance measures; both returned the same results as the percentage difference calculations by Jellesmark et al. (2021). When I ran the method after applying a smoothing algorithm, I was left with a larger choice of seven distance measures. Although the ordering differed slightly from Jellesmark et al. (2021), all seven distance measures agreed. The slight difference in ordering (Snipe vs Lapwing, ambiguous from visual inspection of the trends; Fig. 2.7) is unsurprising given that the smoothing algorithm removed all noise from the trends, while the distance measures we selected using noise filtering, although demonstrating very low sensitivity to white noise, were not invariant to it. Smoothing in this case gave me more distance measures to choose from, but with the added complication of not knowing whether I had improved or distorted my results.

While in both cases (smoothed and unsmoothed trends) there were distance measures that gave the same rankings as Jellesmark et al. (2021) despite not matching my selection criteria (Figs 2.9-2.10), the distance measures I selected were all in agreement. Had I been less specific when choosing important properties, I would have risked including measures that were not fit-for-purpose. A single suitable distance measure is better than any number of ill-suited measures.

2.6. Conclusion

Distance measures are widely used in ecology, but the selection of distance measures described in the ecological literature is limited and their use is often poorly understood, leading to misuse. In the wider literature, there are hundreds of distance measures, with new ones frequently described. This study introduces a selection of 42 distance measures for the purpose of ecological time series analysis and describes an objective method for

choosing an appropriate distance measure for any task involving time series. I have used a suite of criteria to uncover their properties and my tests can be applied to distance measures not included in this study. I have provided the first general selection method for choosing a distance measure based on their properties, and I believe this will be useful for a large range of ecological problems that require comparisons of time series. My work should lead to an improved understanding of, and greater scope for, the use of distance measures for comparing time series within the field of ecology. Nonetheless, it is up to the user to think their way through the process. There are hundreds of potential cases for using distance measures to compare time series in ecology, and as many potential issues that may arise in the process. Most of them are beyond the scope of this study. However, my framework can easily be adapted to incorporate other properties to select a distance measure that is appropriate for the task. There is not always a right choice of distance measure, but there are wrong ones, and my main goal is to help avoid those.

2.7. Bibliography

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Batista, G. E. A. P. A., Wang, X., & Keogh, E. J. (2011). A Complexity-Invariant Distance Measure for Time Series. *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)*, 699–710. https://doi.org/10.1137/1.9781611972818.60

Batyrshin, I., Solovyev, V., & Ivanov, V. (2016). Time series shape association measures and local trend association patterns. *Neurocomputing*, *175*, 924–934. https://doi.org/10.1016/j.neucom.2015.05.127

Boero, F., Kraberg, A. C., Krause, G., & Wiltshire, K. H. (2015). Time is an affliction: Why ecology cannot be as predictive as physics and why it needs time series. *Journal of Sea Research*, *101*, 12–18. https://doi.org/10.1016/j.seares.2014.07.008

Capinha, C. (2019). Predicting the timing of ecological phenomena using dates of species occurrence records: a methodological approach and test case with mushrooms. *International Journal of Biometeorology*, *63*(8), 1015–1024. https://doi.org/10.1007/s00484-019-01714-0

Capinha, C., Ceia-Hasse, A., Kramer, A. M., & Meijer, C. (2020). Deep learning classification of temporal data in ecology. *BioRxiv*. https://doi.org/10.1101/2020.09.14.296251

Cilibrasi, R., & Vitanyi, P. (2018). Clustering by Compression. *IEEE Transactions on Information Theory*, *51*(4), 1523–1545. https://doi.org/10.1109/TIT.2005.844059

Cleasby, I. R., Wakefield, E. D., Morrissey, B. J., Bodey, T. W., Votier, S. C., Bearhop, S., & Hamer, K. C. (2019). Using time-series similarity measures to compare animal movement trajectories in ecology. *Behavioral Ecology and Sociobiology*, *73*(11). https://doi.org/10.1007/s00265-019-2761-1

Dau, H. A., Keogh, E., Kamgar, K., Yeh, C. C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Chen, Y., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., & Hexagon, M. L. (2019). *The UCR Time Series Classification Archive*. https://doi.org/10.1109/JAS.2019.1911747

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time

series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Drost, H. G. (2018). Philentropy: Information Theory and Distance Quantification with R. *Journal of Open Source Software*. https://doi.org/10.21105/joss.00765

Edwards, M., Halaouet, P., Johns, D. G., Batten, S., Beaugrand, G., Chiba, S., Hall, J., Head, E., Hosie, G., Kitchener, J., Koubbi, P., Kreiner, A., Melrose, C., Pinkerton, M., Richardson, A. J., Robinson, K., Takahashi, K., Verheye, H. M., Ward, P., & Wootton, M. (2016). Global Marine Ecological Status Report: Results from the global CPR survey 2014/2015. *SAHFOS Technical Report*, *10*, 1–37.

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1). https://doi.org/10.1145/2379776.2379788

Harris, S. J., Massimino, D., Balmer, D. E., Eaton, M. A., Noble, D. G., Pearce-Higgins, J. W., Woodcock, P., & Gillings, S. (2020). The Breeding Bird Survey 2019. *BTO Research Report*, *726*.

Hegg, J. C., & Kennedy, B. P. (2021). Let's do the time warp again: non-linear time series matching as a tool for sequentially structured data in ecology. *Ecosphere*, *12*(9). https://doi.org/10.1002/ecs2.3742

Jacobs, D. W., Weinshall, D., & Gdalyahu, Y. (2000). Classification with nonmetric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(6), 583–600. https://doi.org/10.1109/34.862197

Jellesmark, S., Ausden, M., Blackburn, T. M., Gregory, R. D., Hoffmann, M., Massimino, D., McRae, L., & Visconti, P. (2021). A counterfactual approach to measure the impact of wet grassland conservation on U.K. breeding bird populations. *Conservation Biology*, *35*(5), 1575–1585. https://doi.org/10.1111/cobi.13692

Keogh, E., & Kasetti, S. (2003). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371.

Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. *Information Processing and Management*, *53*(5), 1103–1119. https://doi.org/10.1016/j.ipm.2017.04.004

Lhermitte, S., Verbesselt, J., Verstraeten, W. W., & Coppin, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, *115*(12), 3129–3152. https://doi.org/10.1016/j.rse.2011.06.020

Liao, Warren T. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Marques, A. R., Forde, H., & Revie, C. W. (2018). Time-series clustering of cage-level sea lice data. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0204319

McCune, B., & Grace, J. B. (2002). *Analysis of ecological communities*. MjM Software Design.

Montero, P., & Vilar, J. A. (2014). TSclust: An R Package for Time Series Clustering. In *JSS Journal of Statistical Software* (Vol. 62). http://www.jstatsoft.org/

Mori, U., Mendiburu, A., & Lozano, J. A. (2016a). *Distance Measures for Time Series in R: The TSdist Package*.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016b). Similarity Measure Selection for Clustering Time Series Databases. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 181–195. https://doi.org/10.1109/TKDE.2015.2462369

Paparrizos, J., Liu, C., Elmore, A. J., & Franklin, M. J. (2020). Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1887–1905. https://doi.org/10.1145/3318464.3389760

Pardieck, K. L., Ziolkowski Jr., D. J., Lutmerding, M., Aponte, V. I., & Hudson, M.-A. R. (2020). *North American Breeding Bird Survey Dataset 1966-2019: U.S. Geological Survey data release*. https://doi.org/10.5066/P9J6QUF6

Potamitis, I., Rigakis, I., & Fysarakis, K. (2015). Insect biometrics: Optoacoustic signal processing and its applications to remote monitoring of McPhail type traps. *PLoS ONE*, *10*(11). https://doi.org/10.1371/journal.pone.0140474

Pree, H., Herwig, B., Gruber, T., Sick, B., David, K., & Lukowicz, P. (2014). On general purpose time series similarity measures and their use as kernel functions in support vector machines. *Information Sciences*, *281*, 478–495. https://doi.org/10.1016/j.ins.2014.05.025

Priyadarshani, N., Marsland, S., Juodakis, J., Castro, I., & Listanti, V. (2020). Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods in Ecology and Evolution*, *11*(3), 403–417. https://doi.org/10.1111/2041-210X.13357

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., & Keogh, E. (2013). Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data*, *7*(3). https://doi.org/10.1145/2500489

Teng, M. (2010). Anomaly detection on time series. *Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing*, *1*, 603–608. https://doi.org/10.1109/PIC.2010.5687485

van de Pol, M., Vindenes, Y., Sæther, B. E., Engen, S., Ens, B. J., Oosterbeek, K., & Tinbergen, J. M. (2011). Poor environmental tracking can make extinction risk insensitive to the colour of environmental noise. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1725), 3713–3722. https://doi.org/10.1098/rspb.2011.0487

Vasseur, D. A., & Yodzis, P. (2004). The color of environmental noise. In *Ecology* (Vol. 85, Issue 4). https://doi.org/10.1890/02-3122

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5

Weinshall, D., Jacobs, D. W., & Gdalyahu, Y. (1998). Classification in Non-Metric Spaces. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (pp. 838–846). NIPS 1998.

WWF. (2020). *Living Planet Report 2020 - Bending the curve of biodiversity loss.* (R. E. A. Almond, M. Grooten, & T. Petersen, Eds.). WWF.

Zhu, Q., Batista, G., Rakthanmanon, T., & Keogh, E. (2012). A Novel Approximation to Dynamic Time Warping allows Anytime Clustering of Massive Time Series Datasets. In J. Ghosh, H. Liu, I. Davidson, C. Domeniconi, & C. Kamath (Eds.), *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM)* (pp. 999–1010).

Chapter 3: How much data do we need? Reliability and data deficiency in global vertebrate biodiversity trends

3.1. Abstract

Global biodiversity is facing a crisis, which must be solved through effective policies and on-the-ground conservation. But governments, NGOs, and scientists need reliable indicators to guide research, conservation actions, and policy decisions. Developing reliable indicators is challenging because the data underlying those tools is incomplete and biased. For example, the Living Planet Index tracks the changing status of global vertebrate biodiversity, but gaps, biases and quality issues plague the aggregated data used to calculate trends. But without a basis for real-world comparison, there is no way to directly assess an indicator's accuracy or reliability. Instead, a modelling approach can be used.

I developed a model of trend reliability, using simulated datasets as stand-ins for the real world, degraded samples as stand-ins for datasets in the Living Planet Database, and a distance measure to quantify reliability by comparing sampled to unsampled trends. The model revealed that the proportion of species represented in the database is not always indicative of trend reliability. Important factors are the number and length of time series, as well as their mean growth rates and variance in their growth rates, both within and between time series. I found that many trends in the Living Planet Index are too data-poor to be considered reliable, particularly trends across the global south. In general, bird trends are the most reliable, while reptile and amphibian trends are most in need of additional data. I simulated three different solutions for reducing data deficiency and found that adding existing data (to the extent that it is available) is the most efficient way to improve trend reliability, and that revisiting previously studied populations is a quick and efficient way to improve trend reliability until new long-term studies can be completed and made available.

3.2. Introduction

An urgent data crisis complicates the global biodiversity crisis (Turak et al., 2017). Attempts to assess global biodiversity (e.g., the Intergovernmental Science-Policy Platform on

Biodiversity and Ecosystem Services, IPBES), and to set policies and goals that will halt or reverse its loss (e.g., the Convention on Biological Diversity, CBD, and Sustainable Development Goals, SDGs), need reliable and up-to-date scientific information (Jetz et al., 2019). Yet most studies and tracking programs are either species- or region-focused, temporally limited and inherently biased, leaving large geographic and taxonomic knowledge gaps (Hortal et al., 2015; Jetz et al., 2019; Meyer et al., 2015; Proença et al., 2017; Turak et al., 2017). Advances in technologies such as camera tracking, satellite sensors, digital image recognition, network speed and capacity, data access, and mobile devices are improving our ability to track and count populations of birds and mammals (Lausch et al., 2016; Nichols et al., 2011; Rose et al., 2015), but our datasets are far from complete. The situation is worse for amphibians, reptiles, insects, and other groups, for which many species have yet to even be described (Mora et al., 2011).

We need tools to improve our understanding of global biodiversity within the limitations imposed by biased and incomplete datasets. Mace & Baillie (2007) suggested a solution: develop indicators based on existing data, understand data biases, and develop methods to reduce the bias. Biodiversity indicators summarize complex scientific information in a simple way, often serving as a bridge between science and policy (Secretariat of the Convention on Biological Diversity, 2006). But what can we expect from indicators that summarize a fraction of the biodiversity they purport to measure? To what extent can we rely on them to present a true picture of the state of global biodiversity?

Two of the best-known biodiversity indicators are the Living Planet Index (LPI), which tracks vertebrate population trends (McRae et al., 2017), and the Red List Index (RLI), which tracks extinction risk trends (Butchart et al., 2005). The RLI is based on extinction risk classifications at the species-level, created by expert assessment using an objective set of criteria (IUCN Species Survival Commission, 2012). By contrast, the LPI uses continuous population data collected by scientific surveys. But, as intensive global long-term studies do not exist for most species, the LPI calculates trends from data compiled from a variety of sources, including grey literature (McRae et al., 2017). This means a lack of standardization in study design (individual population time series are standardized, but there is no standardization between populations), monitoring strategy, frequency of assessment, monitoring intensity and effort, even data type (densities, counts of individuals or breeding pairs or even nests,

and population size estimates are mixed). The LPI has taxonomic and geographical imbalances (Collen et al., 2009; McRae et al., 2017), a problem found also in other global biodiversity datasets (Boakes et al., 2010; Collen et al., 2008; Yesson et al., 2007). Further, many included time series are short (McRae et al., 2016; Proença et al., 2017; Saha et al., 2018), and shorter trends tend to be less accurate than longer ones (Arkilanian et al., 2020; Wauchope et al., 2019). Recognizing these weaknesses, the LPI employs statistical techniques to increase the accuracy and precision of trends. Generalized Additive Models or log-linear interpolation are used (depending on the length of a given time series) to fill in missing values in time series, bootstrapping is used to generate confidence intervals (Collen et al., 2009), and a hierarchical weighting system is applied to account for geographical and taxonomic bias (Collen et al., 2009; McRae et al., 2017).

Without a basis for real-world comparison, there is no way to directly assess an indicator's accuracy or reliability. However, there are ways to address this question indirectly. Baillie et al. (2008) employed one solution when they developed the sampled approach to the Red List Index (sRLI). To determine the minimum representative sample size that would provide accurate trends, they chose two comprehensively assessed taxonomic groups (Birds and Mammals) in the Red List Index and compared trends generated from thousands of subsamples of different sizes to trends generated from the complete dataset (Baillie et al., 2008). When the probability of falsely showing a positive trend for a given assessment period was less than 5% (trends for those groups in the complete dataset were negative), the sample was considered large enough (Baillie et al., 2008). The sRLI method was later updated to include more groups as well as a minimum sample size for detecting change in slope instead of just slope direction (Henriques et al., 2020).

Two challenges presented by the LPI prompted me to take a different approach than the sRLI. First, LPI trends are based on population time series that are often short and/or infrequently measured, and there are no regional or taxonomic groups within the LPI where the data is comprehensive enough to be certain of the real-world trend. Therefore, comparing sampled trends to LPI trends would tell little about how the sampled trends might compare to reality. Second, the LPI uses non-linear trends that change slope and direction over time, so trends should be compared in a way that reflects this. To overcome these challenges, I used a modeling approach. I generated thousands of datasets of

synthetic population time series with variations in the underlying properties of the data to represent regional taxonomic groups in the real world, then took samples from those datasets, degraded the samples by randomly removing observations and adding observation error to resemble regional taxonomic groups in the Living Planet Database (LPD, the database underlying the LPI), then compared the trends calculated from the samples with those from the complete datasets using a distance measure. I constructed a multiple regression model of the distance value to understand how accuracy responds to variations in properties of the data. By selecting a threshold value for accuracy and applying the model to the LPI, I was able to quantify the reliability of disaggregated LPI trends and determine the number of additional time series needed to meet the threshold. Finally, I modelled and compared three different solutions for reducing data deficiency: a) tracking unstudied populations for a decade to generate new time series for the LPD, b) resampling previously studied populations to update old time series in the LPD, and c) gathering more time series from existing studies to add to the LPD. The results from this study can be used to focus data-gathering and data-collation efforts on the regions, taxa, and populations that would be of greatest benefit to improving our understanding of the state of global vertebrate biodiversity.

3.3. Material and Methods

Fig. 3.1 shows an overview of my methods, with each numbered step corresponding to a numbered subheading in the text.

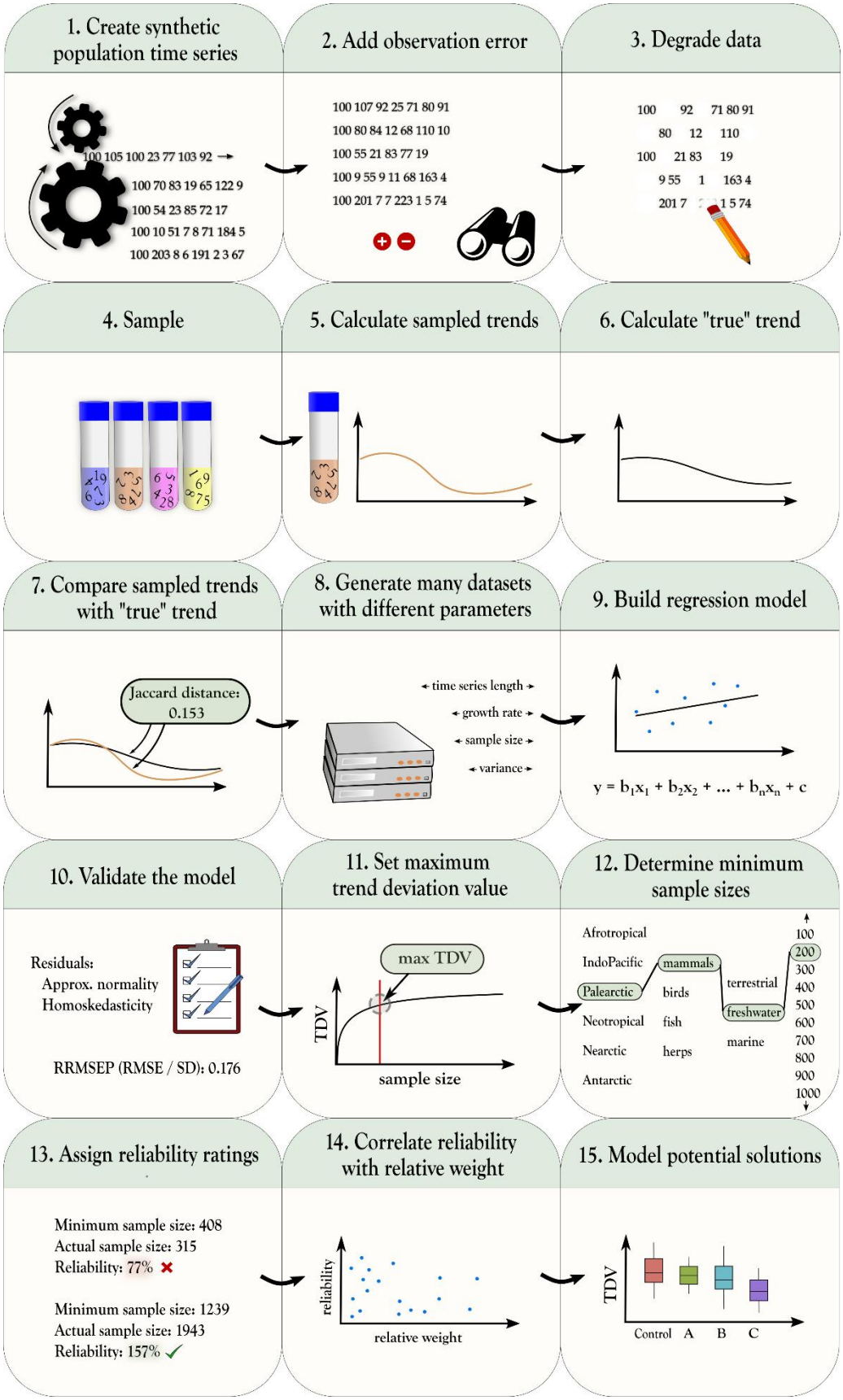


Figure 3.1. Modelling trend accuracy in the LPI: an overview.

3.3.1. Synthetic data generation

I first created simulated datasets to represent real-world regional vertebrate groups for which the LPI calculates biodiversity trends. The LPI is often represented as a single global index trend but can also be disaggregated into hierarchical groups: first into systems (terrestrial, marine, freshwater), then geographical realms within each system, and finally taxonomic groups within each realm. It is this lowest level of the hierarchy, the regional taxonomic groups, which I simulate. From here on each simulated regional taxonomic group will be referred to as a dataset. The base units of the LPI, and of my synthetic datasets, are population time series, which I will refer to simply as populations. These populations are grouped into species, and species are grouped into regional taxonomic groups, or datasets.

My procedure that simulates a dataset requires six parameters: 1) the total number of populations to simulate (set to 10,000), 2) the mean number of populations assigned to each species (set to 10), 3) the number of years (length of trend) to simulate (set to 50), 4) the mean of the population mean growth rates (μ_{ds}), 5) the standard deviation of the population mean growth rates (variation among populations, σ_{ds}), and 6) the mean of the population standard deviations of the growth rate (process error, μ_{η}). The first three parameters were fixed. The first, total populations, affects trend accuracy only when greater than half of all populations in a dataset are sampled (see Fig. S3.1 in Appendix 2), a situation that is unlikely for regional taxonomic groups in the LPD, as it is rare even at the species level (see taxonomic representativeness in McRae et al., 2017). The second parameter, the mean number of populations per species, has no effect on trend accuracy within the wide range of values I tested (see Fig. S3.2 in Appendix 2). The third, trend length, is constant across regional taxonomic groups in the LPD. However, it does affect trend accuracy (see Fig. S3.3 in Appendix 2) and would therefore need to be set appropriately if adapting the model for a different indicator. Parameters four through six are variable in the LPD and affect trend accuracy, and are therefore set to vary in the simulations.

Growth rates can be expressed as discrete annual growth rates,

$$\lambda = (N_{t+1}/N_t) \quad (1)$$

where N_t is population size at year t, or as instantaneous growth rates,

$$r = \ln(N_{t+1}/N_t) \quad (2)$$

Therefore,

$$r = \ln(\lambda) \quad (3)$$

I will discuss both λ and r, but to avoid confusion, I will refer to discrete annual growth rates as lambdas. When I refer simply to the growth rate, I mean r. Each population has a normal distribution of growth rates, $r \sim N(\mu_{pop}, \sigma^2_{pop})$, which translates to a log-normal distribution of lambdas, $\lambda \sim LN(\mu_{pop}, \sigma^2_{pop})$. A stable population would have a mean r of zero, or a mean λ of one. The growth rate distributions simulate a stochastic exponential model with process error. To allow adequate control of dataset parameters, the time series model had to be kept simple and flexible; therefore, growth rates are not serially correlated (although population sizes are), and carrying capacity is not modelled. Population mean growth rates (r) are drawn from a normal distribution for the species, $\mu_{pop} \sim N(\mu_{spec}, \sigma^2_{spec})$. Species mean growth rates are in turn drawn from a normal distribution representing the whole dataset, $\mu_{spec} \sim N(\mu_{ds}, \sigma^2_{ds})$.

Process error, η, is represented by the mean of the population standard deviations of r, with $\sigma_{pop} \sim \text{Exp}(\lambda_{\sigma_{pop}})$. An exponential distribution has only one parameter, the rate (λ), but here I instead describe the 'mean' for ease of interpretation and to avoid confusion with the annual growth rate, which is also λ. The mean, μ_η, is the inverse of the rate (not the annual growth rate):

$$\mu_\eta = 1/\lambda_{\sigma_{pop}} \quad (4)$$

Each dataset was constructed as follows: a normal distribution of μ_{spec} was generated from μ_{ds} and σ_{ds}, and an exponential distribution of σ_{pop} was generated from μ_η. Each species was randomly assigned a mean growth rate from the normal distribution and a standard deviation from a uniform distribution, $\sigma_{spec} \sim U(0, \mu_\eta)$. These species parameters were used to generate normal distributions of population means for each species. Population growth rates must be expressed as lambdas, λ, to enable calculation of population sizes. Therefore, a log-normal distribution, $\lambda \sim LN(\mu_{pop}, \sigma^2_{pop})$, was created for each population using μ_{pop} randomly selected from the normal distribution for the species it was assigned to and σ_{pop} randomly selected from the exponential distribution. This method ensured that variance in growth rates within a population would be lower than the variance in growth rates within

the species it was assigned to, and that population mean growth rates would be more similar within species than between species. This models ecologically significant relationships between populations within species and species within taxa by assuming that growth rates of related populations and species will respond to ecological processes more similarly than unrelated ones. Growth for each population was modelled for 50 years, starting at a population size of 100, with lambdas randomly assigned each year from a log-normal distribution with parameters drawn from a species-level normal distribution of growth rates. Populations were assigned to species by randomly sampling from a pool of 1000 species IDs, with replacement, resulting in a normal distribution of populations per species, $pps \sim N(\mu_{pps}, \sigma^2_{pps})$, with μ_{pps} = 10 and σ_{pps} = 3.1. While populations are unlikely to be normally distributed across species in the real world (one would expect more rare species than common species), simulations confirmed that my modelling approach is robust against distributional assumptions for this parameter (see Fig. S3.2 in Appendix 2).

3.3.2. Observation error

The variation in lambdas modelled above assumes all variation is due to process error. However, time series in the LPD are based on population estimates, which can be assumed to include some level of observation error due to e.g., species misidentification, non-detection, and counting errors. This observation error is not accounted for in the LPI but may affect trend reliability. Observation error, ε, can be calculated using the coefficient of variation (cv), defined as

$$cv_{\varepsilon} = \frac{\sigma_{ab}}{\mu_{ab}} \quad (5)$$

where μ_{ab} and σ_{ab} are the mean and standard deviation (respectively) of the abundance values. Since data in the LPD was collected using a variety of methods, and ε is not recorded in the database, I chose a range of ε consistent with values reported for other vertebrate surveys (Fryxell et al., 2014; Westcott et al., 2012; Zylstra et al., 2010). I determined through simulations that there is no effect of increasing observation error on trend accuracy (Fig. S3.4 in Appendix 2), therefore an approximate range of ε should suffice. For each time series, ε was randomly selected from a normal distribution with μ_{ε} = 0.15 and σ_{ε} = *0.1*. I simulated observed versions of each time series, modeled as

$$Z_t = X_t + \phi_t \ , \qquad Z_t \geq 0 \tag{6}$$

where Z_t is a simulated observation, X_t is a simulated value from a time series at time t, and ϕ_t is a normally distributed variable, $\phi_t \sim N(0, \sigma_{obs}^2)$, with

$$\sigma_{obs} = X_t * \mu_\varepsilon \tag{7}$$

where σ_{obs} is the standard deviation of ϕ_t. A value for μ_ε of 0.1 (10%) would result in approximately 68.2% of observations falling within 10% of their corresponding simulated values and 99.7% of simulated observations falling within 30%.

3.3.3. Data degradation

Observed versions of the datasets were then randomly degraded to resemble the varied quality of sampled real-world data present within the LPD. The length (number of years from first to final observation) for each degraded time series within a dataset was randomly chosen by sampling from a Poisson distribution. I determined through simulations that varying the number of observations does not affect trend accuracy at a given time series length, so I fixed the mean number of observations at half of the mean time series length (rounded up). The starting years for each time series were assigned randomly. Time series were then cut to their assigned length, and half of the remaining observations were randomly removed.

3.3.4. Sampling

Populations were randomly sampled from each dataset, without replacement. This was repeated to obtain 20 random samples of the same size for each dataset. Values for four of the six dataset parameters described in Section 3.3.1 may be different for samples than for the dataset they are selected from, and may also vary between samples: the mean number of populations per species (μ_{pps}), the mean and standard deviation of population mean growth rates (μ_{ds} and σ_{ds}, respectively), and the mean of population standard deviations of the growth rate (μ_η).

3.3.5. Calculation of sampled trends

Non-linear index trends were calculated from each sample, following the LPI method described in McRae et al. (2017). First, time series with six or more data points were modelled using a Generalized Additive Model (GAM), as described in Collen et al. (2009), with a Gaussian (normal) distribution, smoothed by a thin plate regression spline, with the number of knots set to half the number of observations (rounded down). The model fit was checked by applying a GAM to the residuals, this time smoothed by a shrinkage version of a cubic regression spline, with the number of knots set to the full number of observations of the time series (before the GAM model was applied) and gamma set to 1.4. If the sum of the estimated degrees of freedom from the modeled residuals was close to one (greater than 0.99 and less than 1.01), the population GAM was considered a good fit. Time series that did not pass the model fit test, or that had fewer than six data points, were interpolated using the chain method (Loh et al., 2005), as described in Collen et al. (2009). The chain method imputes missing values using log-linear interpolation by

$$N_i = N_p\left(\frac{N_s}{N_p}\right)^{[i-p/s-p]} \quad (8)$$

where *N* is the population estimate, *i* is the year for which the value is to be interpolated, *p* is preceding year with an observed value, and *s* is the subsequent year with an observed value. For all populations, whether interpolated or modeled by a GAM, species indices were formed by a three-step process. First, population sizes were converted to growth rates by

$$r = log_{10}\frac{N_t}{N_{t-1}} \quad (9)$$

where *N* is the population estimate and *t* is the year. Second, average growth rates were calculated for each species by

$$\overline{r_t} = \frac{1}{n}\sum_{i=1}^{n_t} r_{it} \quad (10)$$

where n_t is the number of populations in a given species, r_{it} is the growth rate for population *i* at year *t*, and $\overline{r_t}$ is the average growth rate at year *t*. Growth rates were capped at [-1:1]. Finally, index values were calculated by

$$I_t = I_{t-1} * 10^{\bar{r}_t}, \qquad I_0 = 1 \tag{11}$$

where *I* is the index value and *t* is the year.

3.3.6. Calculation of the 'true' trend

A non-linear index trend was calculated for each complete, undegraded dataset (without observation error), following McRae et al. (2017), as for the sampled trends. However, the undegraded datasets have no missing values, therefore modeling each time series using the chain method or a GAM was unnecessary, and that step was skipped.

3.3.7. Comparison of trends

I used the process described in Chapter 2 to determine appropriate distance measures to compare sampled trends with 'true' trends. Of the distance measures deemed appropriate, I chose the Jaccard distance because it uses a 0-1 scale, making it easier to interpret. The Jaccard distance is calculated as

$$d_{Jaccard} = \frac{\sum_{t=1}^{n}(P_t - Q_t)^2}{\sum_{t=1}^{n} P_t^2 + \sum_{t=1}^{n} Q_t^2 - \sum_{t=1}^{n} P_t Q_t} \tag{12}$$

(from Cha, 2007), where P_t and Q_t are index values from two trends P and Q at time point *t*, and *n* is the number of time points. From here on, any value calculated by applying the Jaccard distance to compare sampled vs 'true' trends will be referred to as a trend deviation value, or TDV.

I use TDV here as a measure of trend accuracy, but it is in fact the complement of accuracy (a perfectly accurate trend would yield a TDV of zero); lower TDV means higher accuracy. Furthermore, when referring to TDVs of simulated trends, I use the term 'trend accuracy,' but when referring to TDVs of LPI trends, I use the term 'trend reliability.' This is because TDVs for simulated trends are measured, while TDVs for LPI trends are estimated based on a model. So, a 'reliable trend' is expected to be accurate but may not be; likewise, an 'unreliable trend' is expected to be inaccurate but may not be.

3.3.8. Generation of datasets

I generated 3,000 datasets (each consisting of 1,000 species and 10,000 populations), with each dataset sampled 20 times, resulting in 60,000 samples. Values for mean time series length, μ_{ds}, σ_{ds}, and μ_{η} were randomly selected from uniform distributions, while sample size was randomly selected from a log-uniform distribution, ln(SS) ~ U(ln(a), ln(b)), where SS is sample size and a and b are the minimum and maximum values, respectively (log-uniform was chosen to ensure the model would be robust at small sample sizes, as most datasets in the LPD are small). Ranges for the distributions were chosen to ensure that parameter ranges in the samples would be broader than the ranges present in the LPD (Table 3.1). Regional taxonomic groups from the LPD with fewer than 20 populations were excluded from parameter range calculations to avoid extreme outliers. I set the minimum sample size to 50 because smaller samples rarely generated a complete trend, and the maximum to 10,000 to improve predictions of the effects of sample size increases.

Table 3.1. Parameters with value ranges for simulated datasets, degraded samples, and the LPD.

Independent Variable	Range in Datasets	Range in Samples	Range in LPD
Sample Size	–	50 – 9975	2 – 3000
Mean Length of Time Series	6.0 - 38	5.5 – 39	6.0 – 39
Mean of Pop. Mean Growth Rates, μ_{ds}	-0.13 – 0.12	-0.25 – 0.31	-0.19 – 0.16
St. Dev. of Pop. Mean Growth Rates, σ_{ds}	0.074 – 0.59	0.097 – 0.83	0.12 – 0.63
***Mean of Pop. Growth Rate St. Dev., μ_{η}**	0.049 – 1.17	0.13 – 1.06	0.16 – 0.89

*This parameter is modelled as process error in the simulated datasets, but in the degraded samples it represents process error and observation error combined.

3.3.9. Multiple regression model

I built a multiple linear regression model to understand how variables in the data determine trend accuracy (TDV). First, I removed all datasets in which the mean of the sample parameter values fell outside of LPD parameter ranges (individual replicates were allowed to fall outside of LPD ranges), leaving 2,361 datasets, or 47,220 samples. I then randomly selected 67% of the remaining datasets (1,581 datasets) to train the model. The other 33% (780 datasets) I set aside for testing the model.

3.3.10. Model validation

The residuals of the combined data used to train the model were approximately normally distributed. Likewise, the residuals appeared homoscedastic when plotted against fitted

values. I compared the actual TDV of each sample in the testing datasets to the predicted TDV for that sample calculated by the model, then calculated the RRMSEP (relative root mean squared error of prediction), defined as

$$RRMSEP = RMSE/SD \quad (13)$$

where RMSE is the root mean squared error and SD is the standard deviation of the actual TDVs, and

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \quad (14)$$

where y_i is the ith actual TDV, $\hat{y}$ is the predicted TDV, and n is the number of samples.

3.3.11. Maximum trend deviation value

I set a maximum predicted TDV as a threshold that regional taxonomic group trends within the LPI should not exceed to be considered reliable. First, I built a linear regression model of the square root of TDV from my training datasets, with the natural log of sample size as the predictor variable, since sample size is the only user-controlled variable within the LPD. Every regional taxonomic group within the LPD represents a single sample from the real world; therefore, I was not interested in the mean TDV achieved by each dataset, but in the range of possible TDV values, especially the upper part of the range (the least accurate sample trends from each dataset).

I used 10,000 bootstrap estimations of the mean of the TDV from each dataset to calculate the 90% confidence intervals using the bias corrected and accelerated bootstrap interval (BCa) method, also known as the adjusted bootstrap percentile method. The BCa method is a non-parametric method that does not assume the data is normally distributed (the TDV values have a beta distribution) and corrects for bias and skewness in the distribution of the mean estimates. I plotted the curve of the sqrt-log model of the upper 90% confidence interval of TDV in relation to sample size on a (non-log) graph of TDV vs sample size (Fig. 3.2).

To choose a maximum TDV, I used a method called the concordance probability method (CZ) (Liu, 2012). I borrowed CZ from the field of biomedical research, where it is often necessary to specify a cut-off value to discriminate between positive and negative results from screening or diagnostic tests (Liu, 2012). First, a receiver operating characteristic (ROC) curve is built, plotting the rate of true positives (sensitivity) against the rate of false positives (1 - specificity). The idea is to find the point on the curve that maximises both sensitivity and specificity. The CZ method simply finds the point where their product is maximized.

By considering the sqrt-log model of the upper 90% confidence interval of TDV vs sample size (Fig. 3.2) as equivalent to an ROC curve, I applied the CZ method to find the point on the curve where TDV and sample size are minimized. This is the point where the data should provide maximum value. Further right along the curve, increasing the sample size would give a smaller improvement in trend reliability and is therefore not cost or resource effective. Since an ROC curve is intended for binary classification, the CZ method assumes that both sensitivity and specificity are on a 0-1 scale. TDV already ranges from 0-1, so I set sensitivity as 1 - TDV. I normalized sample size to a 0-1 scale by converting it to a proportion of the complete dataset (dividing by the total number of time series in the dataset). Since all datasets were the same size, the relationship between TDV and sample size was not altered by the conversion to a proportion. Specificity was then 1 - sample proportion. The optimal cut-point on the curve is defined as

$$\max(CZ), \qquad CZ(c) = Se(c) * Sp(c) \tag{15}$$

where Se is sensitivity, Sp is specificity, and c is any cut-point.

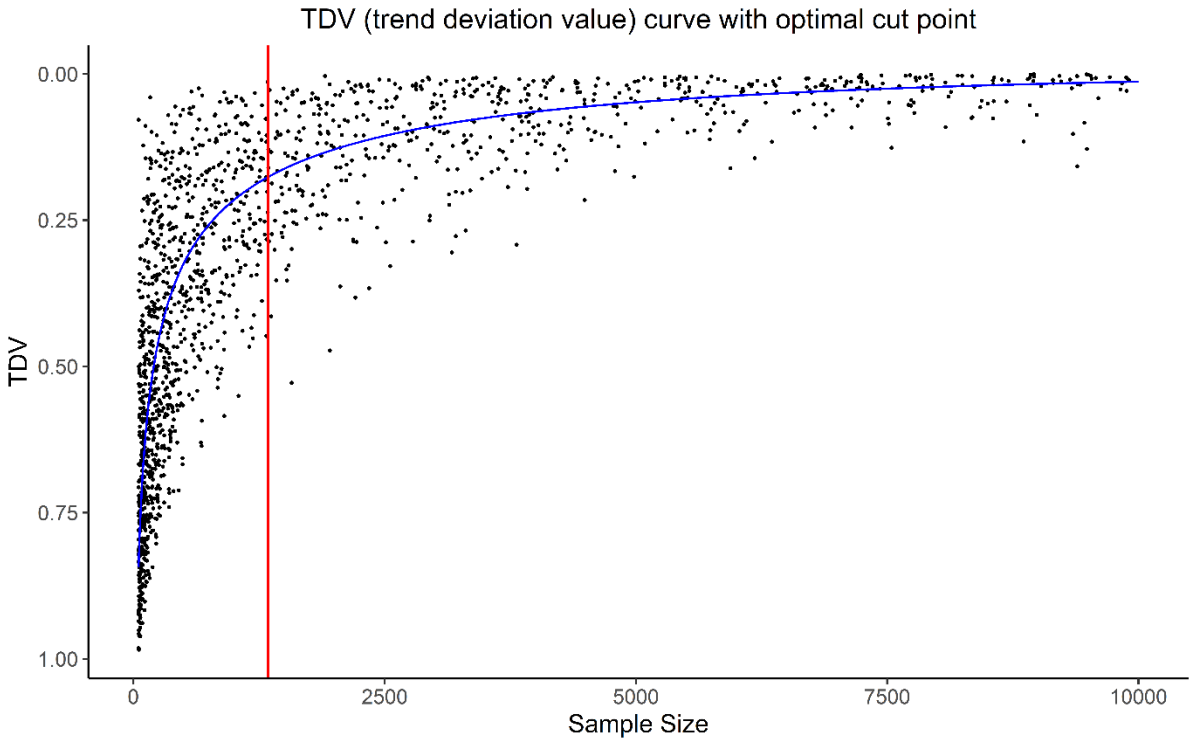


Figure 3.2. Sqrt-log model of trend deviation value (TDV) vs sample size with optimal cut point. This plot includes only the upper 90% confidence interval of TDV from each simulated dataset. The curved blue line is the sqrt-log model of the plotted values. The vertical red line intersects the sqrt-log curve at the optimal cut-point.

3.3.12. Minimum sample size for regional taxonomic groups

Minimum sample size was calculated by rearranging the formula for the multiple regression model to solve for sample size and replacing the TDV variable in the formula with the cut-off value determined above. Values for the other variables in the formula were determined separately for each regional taxonomic group from the LPD, as follows: any populations with less than two data points were removed, missing data was interpolated using the chain method (Collen et al., 2009), then the mean growth rate, μ_{pop}, was calculated for each population. Growth rates were capped at [-1:1] before taking the mean, as in the LPI (McRae et al., 2017). Next, μ_{ds}, σ_{ds}, and μ_{η} were calculated. The mean time series length was calculated by dividing the total number of observations (after interpolation) by the total number of populations (excluding those with less than two data points). The calculated values were then placed into the model formula to determine minimum sample size.

3.3.13. Assigning reliability ratings to regional taxonomic groups

The actual number of populations in each regional taxonomic group was divided by the minimum sample size and multiplied by 100 to determine the percentage of the minimum sample size met by each group. Groups achieving 100% or greater were designated as reliable, those achieving between 50% and 100% were designated as data deficient, and those achieving less than 50% were designated as severely data deficient.

3.3.14. Correlations between reliability rating and LPI relative weighting

The Pearson's product moment correlation coefficient test was performed to determine if there was any significant correlation between percentage of the minimum sample size achieved for each regional taxonomic group and the assigned relative weightings in the LPI for each group. The test was performed on the combined dataset as well as each individual system.

3.3.15. Modelling potential solutions

I used the model to simulate three different methods of improving trend reliability in the LPD: A) tracking unstudied populations for ten years, B) resampling previously studied populations, and C) gathering more time series from existing studies. First, I generated 50 control datasets with a sample size of 200 and mean time series length of 14 (similar to the median values for regional taxonomic groups in the LPI of 180 and 13, respectively). I set μ_{ds} to zero, σ_{ds} to 0.25, and μ_{η} to 0.30. Using the same parameters, I then generated groups of 50 datasets with each of the following changes: group A had an extra 200 populations (total sample size: 400), but with observations only for the final ten years, to simulate tracking additional populations for ten years; group B had the final observation revealed on every sampled, degraded time series (total sample size: 200) to simulate resampling previously-studied populations; group C had an extra 200 randomly sampled populations (total sample size: 400) to simulate adding existing data to the LPD.

3.3.16. Coding and data

All trends for synthetic data were produced using original code designed to reproduce the functionality of the rlpi package (Freeman et al., 2021). All coding was done in R (R Core

Team, 2021) using RStudio (RStudio Team, 2022). Fig. 3.1 and parts of Fig. 3.4 were produced using Inkscape (Inkscape Project, 2020). All other figures were produced in R (R Core Team, 2021) using the ggplot2 package (Wickham, 2016). Population time series used to evaluate reliability of LPI trends are from the LPD (McRae et al., 2016). All original code is available on GitHub at https://github.com/shawndove/DD_LPI.

3.4. Results

3.4.1. Regression model

The regression model contains five independent variables (Tables 3.1 & 3.2). Together they describe 62% of the variation (adjusted r-squared: 0.6223) in the TDV associated with sampled trends. The model is significant, with $F(5, 29385) = 9{,}686$, $p < .001$. All independent variables are significant predictors, with $p < 0.001$. Interaction terms were significant but did not increase the adjusted r-squared of the model, so I left them out. RRMSEP is 0.231. Sample size is the most important variable affecting trend accuracy, with differences in importance between the other three variables comparatively small. Much of the unexplained variance from the model is due to random sampling. I confirmed this by remaking the model using the sample means, which resulted in an adjusted r-squared of 0.8706. Using the square root of TDV instead of the log further increased the adjusted r-squared to 0.9343. This was not the case for the model using the individual samples, where the log resulted in a higher adjusted r-squared than the square root.

Table 3.2. Multiple regression model of ln(TDV).

coefficient	estimate	standard error	beta coefficient	t value	p value
(Intercept)	3.957	0.04406	–	89.81	< .001
ln(Sample size)	-0.8460	0.004441	-0.6860	-190.5	< .001
ln(St. dev. of mean growth rate, σ_{ds})	0.7569	0.01630	0.1672	46.42	< .001
Mean growth rate, μ_{ds}	8.057	0.1454	0.1989	55.42	< .001
Mean of population st. dev., μ_{η}	1.503	0.02224	0.2426	67.57	< .001
Mean time series length	-0.03890	0.0007336	-0.1917	-53.02	< .001

3.4.2. Maximum trend deviation value

Using the concordance probability (CZ) method to select a cut point on the sqrt-log model of the 90% upper confidence interval of TDV vs sample size, I found a maximum TDV value of

0.176. After placing this value into the model equation and reorganizing to solve for sample size, I applied the model to the LPI to find the minimum number of populations needed for each regional taxonomic group.

3.4.3. Minimum sample size

The number of populations needed to achieve the TDV threshold for a reliable trend varies across taxonomic groups and realms (Table 3.3), but weakly across systems, with medians of 210, 259, and 233 for terrestrial, freshwater, and marine systems, respectively. Fewer populations are needed in the global north (median: 201) than in the global south (median: 259). Birds show the highest variability, having both the smallest number of populations needed for any group (freshwater Nearctic birds: 32), and the largest (freshwater Afrotropic birds: 6,768). Mammals have the smallest sample size requirements, with a median of 170, while fishes have the largest, with a median of 472.

Table 3.3. Estimated TDV and number of populations needed to meet the threshold for all regional taxonomic groups in the LPD. Note that because the trend deviation values here were estimated using the model formula, they occasionally fall outside of the 0-1 range of the Jaccard distance used to determine TDVs for simulated data.

System	Realm	Taxon	TDV	Current Sample Size	Minimum Sample Size	Additional Pops Needed
Terrestrial	Afrotropic	Birds	0.497	166	566	400
		Mammals	0.031	916	116	0
		Reptiles & Amphibians	0.425	56	159	103
	IndoPacific	Birds	0.148	466	378	0
		Mammals	0.123	279	182	0
		Reptiles & Amphibians	0.841	84	533	449
	Palearctic	Birds	0.018	1724	119	0
		Mammals	0.019	2153	157	0
		Reptiles & Amphibians	0.370	57	137	80
	Neotropic	Birds	0.148	375	305	0
		Mammals	0.298	210	391	181
		Reptiles & Amphibians	0.166	225	210	0
	Nearctic	Birds	0.005	2564	38	0
		Mammals	0.095	775	374	0
		Reptiles & Amphibians	0.577	127	516	389
Freshwater	Afrotropic	Birds	4.603	143	6768	6625
		Mammals	0.477	18	58	40

		Reptiles & Amphibians	0.729	18	96	78
		Fishes	0.808	180	1090	910
	IndoPacific	Birds	0.175	267	264	0
		Mammals	0.336	22	47	25
		Reptiles & Amphibians	0.242	141	206	65
		Fishes	0.395	231	600	369
	Palearctic	Birds	0.039	1527	255	0
		Mammals	0.255	179	277	98
		Reptiles & Amphibians	0.317	100	201	101
		Fishes	0.171	601	581	0
	Neotropic	Birds	0.348	88	197	109
		Mammals	3.146	13	392	379
		Reptiles & Amphibians	0.353	94	214	120
		Fishes	0.364	295	696	401
	Nearctic	Birds	0.012	736	32	0
		Mammals	0.310	30	58	28
		Reptiles & Amphibians	0.260	307	485	178
		Fishes	0.081	821	326	0
Marine	Temperate Atlantic	Birds	0.038	783	128	0
		Mammals	0.214	196	247	51
		Reptiles & Amphibians	0.836	57	359	302
		Fishes	0.039	2826	472	0
	Tropical Atlantic	Birds	0.218	174	223	49
		Mammals	2.082	20	371	351
		Reptiles & Amphibians	0.887	113	764	651
		Fishes	0.041	3037	547	0
	Arctic	Birds	0.153	175	149	0
		Mammals	0.300	56	105	49
		Fishes	1.192	36	345	309
	South temperate	Birds	0.045	510	101	0
		Mammals	0.482	27	89	62
		Fishes	0.175	246	244	0
	IndoPacific	Birds	0.218	197	254	57
		Mammals	0.276	68	116	48
		Reptiles & Amphibians	0.578	86	351	265
		Fishes	0.067	1103	349	0
	Pacific temperate	Birds	0.084	245	102	0
		Mammals	0.235	154	216	62
		Reptiles & Amphibians	6.528	2	143	141
		Fishes	0.054	798	199	0

3.4.4. Trend reliability

Reliability varies strongly across realms, taxonomic groups, and systems (Figs 3.3-3.4). Terrestrial trends are the most reliable and freshwater trends the least. Terrestrial and

freshwater trends are more reliable in the global north than in the global south, with the exception of terrestrial reptiles and amphibians. Marine bird trends are more reliable in temperate areas than the tropics, while marine fish trends are more reliable in warm waters than cold. Globally, bird trends are the most reliable, but are nonetheless poor in the tropics, especially Africa. Reptile and amphibian trends are data deficient everywhere except the terrestrial Neotropical realm, and aquatic mammal trends are data deficient everywhere.

The groups with the greatest potential to affect the reliability of aggregated LPI trends are exclusively tropical (Fig. 3.5), due to a combination of high relative weighting and low reliability scores. The nine groups of greatest concern include six freshwater and three terrestrial groups, but no marine groups. All are from the tropics. Fishes, birds, and reptiles and amphibians are represented, with mammals absent. Overall, the reliability scores of regional taxonomic groups did not show a statistically significant correlation with their relative weightings in the LPI, $r(55) = -0.042$, $t = -0.31$, $p = 0.76$. Likewise, there were no statistically significant correlations for terrestrial and freshwater systems, with terrestrial $r(13) = -0.34$, $t = -1.31$, $p = 0.21$; and freshwater $r(18) = -0.19$, $t = -0.81$, $p = 0.43$. The marine system showed a moderate positive correlation between reliability and relative weightings: $r(20) = 0.43$, $t = 2.15$, $p = 0.044$.

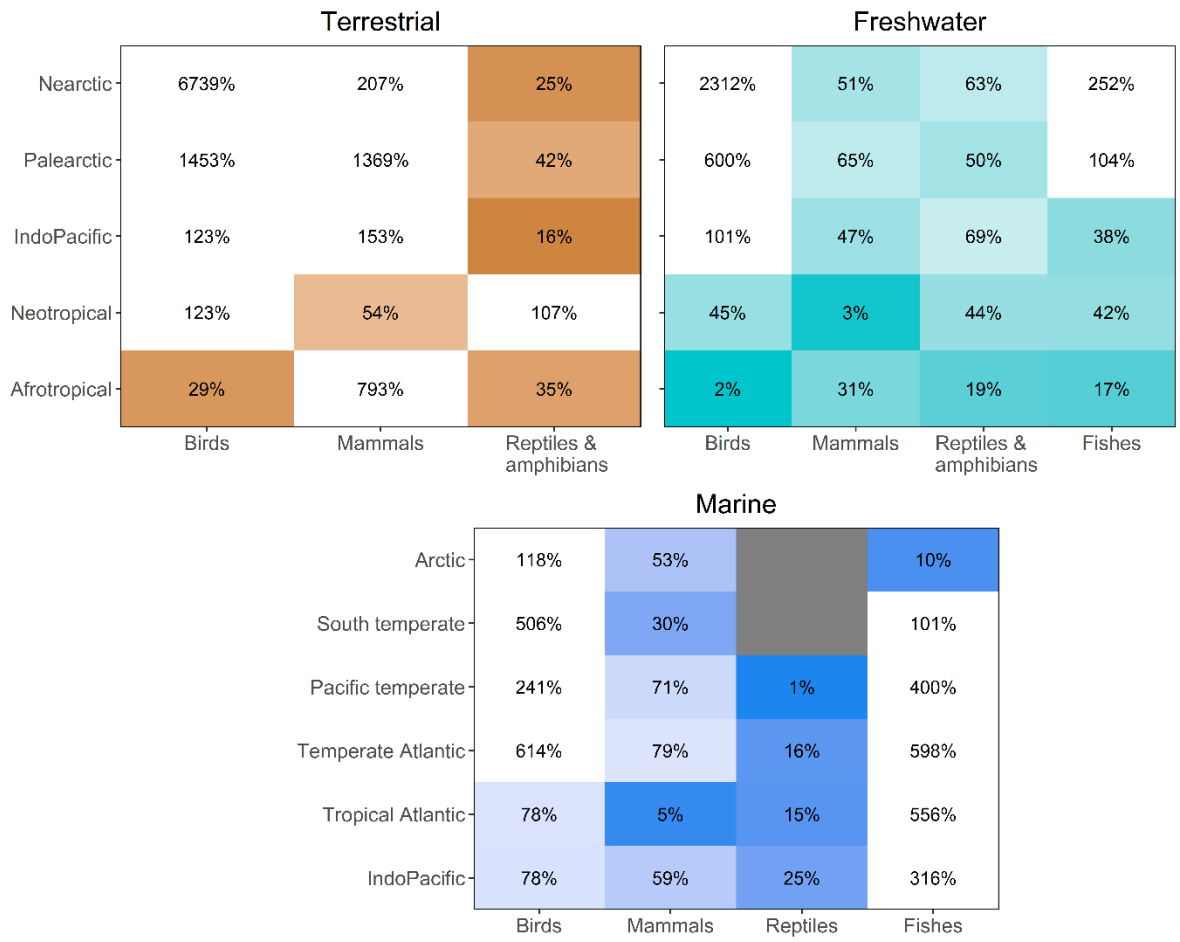


Figure 3.3. Proportion of the total amount of time series data needed to achieve the trend reliability threshold that each regional taxonomic group in the LPD currently contains. A score of 100% or greater means that group already has enough data to produce a reliable trend. A grey box refers either to a group that could not be evaluated because there was too little data (South temperate marine reptiles) or to an invalid realm-taxon combination (there are no marine reptiles in the Arctic).

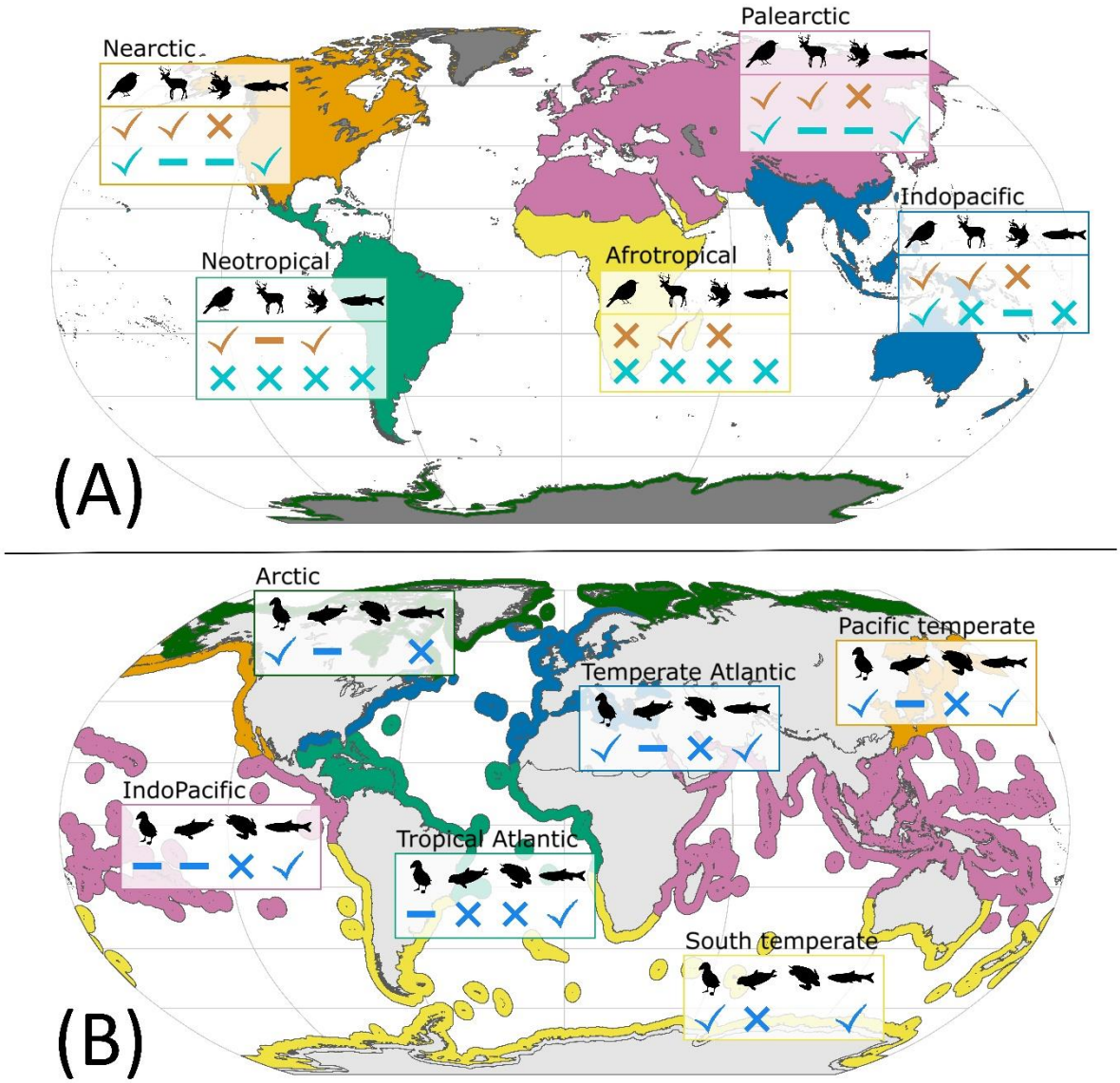


Figure 3.4. Reliability of regional taxonomic group trends in the LPI, grouped by system, realm, and taxon. Map A shows the terrestrial (top) and freshwater (bottom) results. Map B shows the marine results. Reliability scores are binned into three categories, according to the number of time series in the LPD relative to the minimum sample size needed to achieve the TDV threshold. A check mark means that group has at least 100% of the minimum sample size and is considered reliable, a dash means it is data deficient (50-99%) and considered unreliable, and an X mark means it is severely data deficient (< 50%) and considered very unreliable.

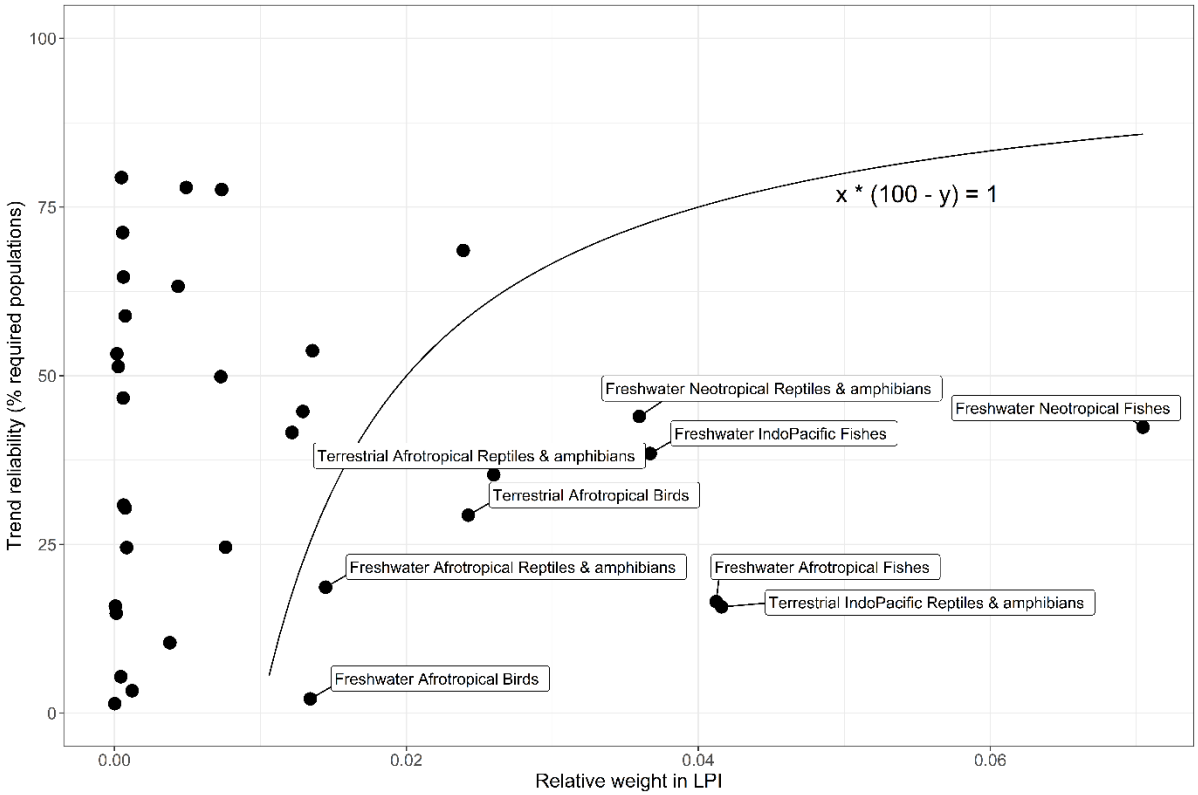


Figure 3.5. Trend reliability of regional taxonomic groups in the LPD vs the relative weighting applied to each group when calculating aggregated LPI trends. Trend reliability is measured as the percentage of populations in the LPD relative to the number required to achieve the TDV threshold. Only groups with reliability ratings below the threshold (less than 100%) are included here. To determine the groups having the strongest negative effect on the reliability of aggregated LPI trends, I calculated relative weight * (100 – reliability) and labelled the groups with a value higher than one.

3.4.5. Modelling potential solutions

Revealing the final year observation (equivalent to resampling previously studied populations) for every population improved the median TDV by 6.5%, while adding 200 additional time series to the sample with observations only in the final ten years (equivalent to tracking 200 unstudied populations for ten years) improved the mean TDV by 11% (Fig. 3.6). By contrast, simply doubling the sample size (equivalent to randomly adding 200 existing time series to the LPD) improved the median TDV by a significant 43%. This solution showed a significant improvement in trend accuracy compared to the control group ($p < 0.001$).

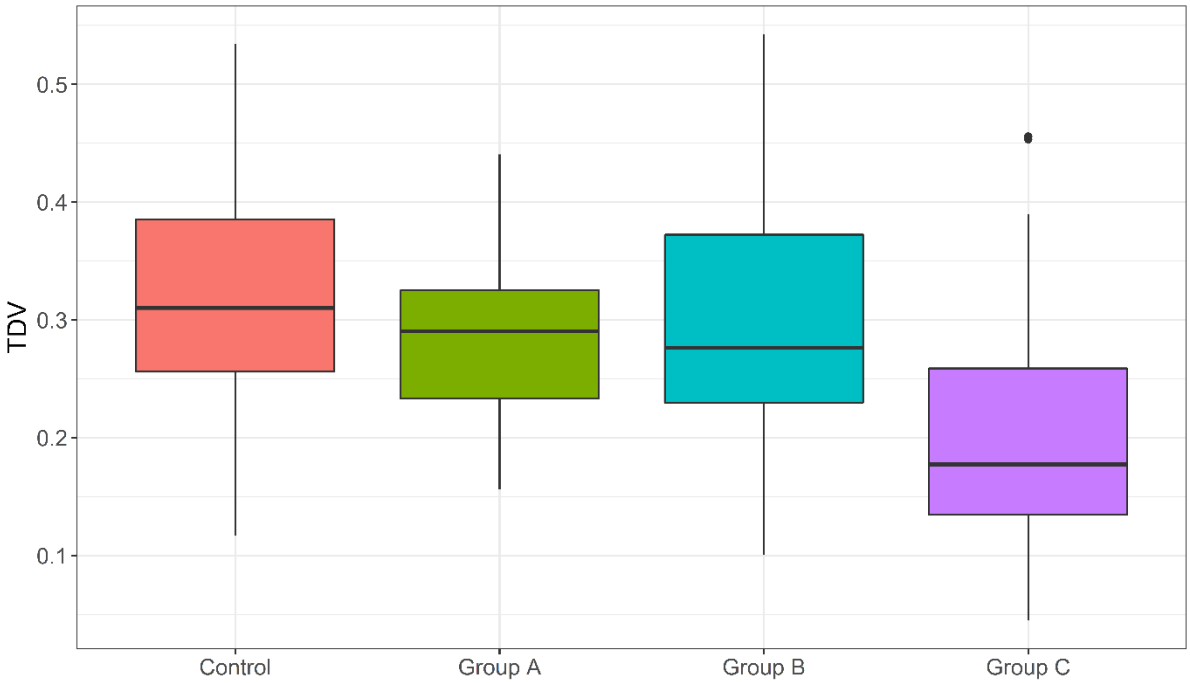


Figure 3.6. The effect on trend accuracy of potential solutions to data deficiency in LP regional taxonomic groups. The control group has a sample size of 200 and mean time series length of 14. Group A has an additional 200 time series with observations only in the final ten years of the index to simulate a ten-year data blitz. In group B, the final observation has been added back in for every time series to simulate resampling of previously studied populations. Group C is like the control group, but the sample size has been doubled to 400 to simulate adding additional pre-existing studies to the LPI.

3.5. Discussion

Understanding the changing global state of biodiversity is crucial to making good policy and conservation decisions to 'bend the curve' of biodiversity loss. Acquiring accurate and comprehensive data is crucial, but the first step is to answer the question: what do we actually know? The present study quantifies the reliability of trends for each regional taxonomic group in the Living Planet Index and estimates the number of population time series needed to meet a standard of expected accuracy.

I used synthetic population time series datasets to construct a multiple regression model of trend accuracy by comparing trends of degraded samples with the trends of the full, undegraded datasets using the Jaccard distance metric (Fig. 3.1). I applied the model to regional taxonomic groups in the Living Planet Database to reveal that the majority need additional data for their trends to be considered reliable. Data deficiency is a problem

globally but is more pronounced in the tropics. This is consistent with the analysis of geographical representativeness in McRae et al. (2017), which tested proportional representativeness of biodiversity compared to the global dataset and found that species groups in tropical realms are underrepresented. Bird trends are the most reliable and reptiles and amphibians the least. This is consistent with the picture of species representation in the LPD presented in McRae et al. (2017) and is unsurprising given that monitoring and data collection for birds is more extensive than for reptiles and amphibians (Oliver et al., 2021; Scheele et al., 2019), especially with the rise of citizen science (Oliver et al., 2021). However, many of my reliability scores differ from what would be expected given McRae et al. (2017)'s analysis of taxonomic representativeness. McRae et al. (2017) found that all Nearctic taxonomic groups are overrepresented, yet in my analysis Nearctic terrestrial and freshwater reptiles and amphibians, as well as Nearctic freshwater mammals, score as data deficient. The starkest differences occur in the marine system, where mammals and marine reptiles are overrepresented by species in all realms (except South temperate reptiles, which are not represented at all) but which I found to be data deficient in all realms. By contrast, marine fishes are underrepresented by species numbers (McRae et al., 2017), but I found that in all except the Arctic realm marine fishes are data-rich enough to produce reliable trends. These results strongly suggest that the percentage of species represented does not tell the whole story.

Geographical and taxonomic biases in the distribution of data in the LPI are known (McRae et al., 2017), and reflect underlying biases in the availability of data (Boakes et al., 2010; Collen et al., 2008; Yesson et al., 2007). In 2017, McRae et al. introduced a weighting system to the LPI, which accounts for the estimated number of species in each regional taxonomic group to reduce representational bias. One problem with this is that the majority of the world's vertebrate species are located in the tropics (Collen et al., 2008; McRae et al., 2017), which are underrepresented in the LPD (McRae et al., 2017). My concern was that if trends from these areas are the most unreliable due to data deficiency, then the LPI could have simply replaced one problem, representation bias, with another: overreliance on unreliable trends. Indeed, my analysis shows that all regional taxonomic groups with a high relative weight and low reliability (bottom right of Fig. 3.5) are tropical. Surprisingly, though, I did not find a statistically significant negative correlation between reliability of trends and their

relative weights in the LPI. This also holds true for the terrestrial and freshwater systems when considered separately (the marine system shows a slight positive correlation), and is consistent with Nori et al. (2020), who found that species richness and knowledge gaps are not always correlated.

According to my model, the size of a dataset, i.e., the number of species or populations existing in the real world for any regional taxonomic group, is unimportant to the calculation of trend reliability for a given sample, as long as the sample represents less than half of the time series in the dataset (see Fig. S3.1 in Appendix 2). In other words, it is the absolute number of populations represented in the sample that matters, regardless of whether that sample represents 1% or 50% of the total populations in a regional taxonomic group. There are two principles working to cause this seemingly counterintuitive effect. First, the relationship between population size and the sample size needed to reach a desired level of precision is logarithmic and becomes more extreme at lower levels of precision (Israel, 1992). This means that a small sample size should be able to estimate a large population almost as well as it can estimate a small population. Second, there are limitations to the level of trend accuracy that can be achieved, regardless of sample size, because most time series in my simulated samples (and in the LPD) are much shorter than the length of the trend being estimated. Short time series tend to produce more extreme trends (Leung et al., 2020) and are less likely to accurately reflect long-term trends for individual populations (Wauchope et al., 2019). They also reduce the number of observations used for the calculation of group trends. For example, even if the mean time series length was 50% of the length of a trend (mean time series lengths for all regional taxonomic groups in the LPD are much shorter than that), if those time series were randomly distributed in time, only about 4% of them would begin at the first year and about 4% would end at the final year. Thus, the crucial early and final years of the trend would depend on only a fraction of the observations that the sample size indicates. This randomized distribution of time series across the trend results in less accurate trends than would be possible if observations were evenly distributed across time points (confirmed through simulations – see Fig. S3.5 in Appendix 2). This issue is slightly complicated in the LPD. On one hand, the database begins 20 years earlier than the index, giving time for the number of observations to increase before measuring the trend. But on the other hand, there is a delay in getting recent studies

into the LPD (McRae et al., 2017), reducing the number of observations in the final years even more than a random distribution would suggest (see Fig. S3.6 in Appendix 2).

This dramatic fall-off of observations suggests that the LPI may not reliably reflect changes in the status of global vertebrate biodiversity over the past decade. More data is needed, and while a reduction in the delay involved in getting new studies into the LPD might help, increasing the number of populations in the LPD is only possible to the extent that the necessary data exists. Therefore, I simulated two potential ways of generating new data to improve trend reliability: A) a global data blitz, with researchers coordinating to track as many unstudied populations as possible for ten years to generate new time series, and B) resampling already-studied populations to uncover recent changes and lengthen existing time series (Fig. 3.6). Both solutions had a slight but non-significant positive effect on trend accuracy but were far less effective than adding existing data. It is likely that both solutions have a greater effect on the accuracy of the final portion of the trend than on the overall trend, but further study would be required to be certain. Either way, resampling would be more efficient than a data blitz, as the same improvement could be achieved in one year instead of ten. In the long term, tracking additional populations is essential to completing our picture of biodiversity change. But natural stochasticity means that short time series are of limited value in generating reliable trends (Wauchope et al., 2019), so tracking additional populations takes time to pay dividends.

There is another problem underlying the LPI, which cannot be solved by generating new data. All trends in the LPI begin in the year 1970, which is set as the base year for calculating the index values. Past trends can only be determined by existing data; therefore, while there may be some currently inaccessible data that either could be shared or made available for confidential storage in the LPD (Saha et al., 2018), there are likely to be severe limitations to relieving data deficiency for this time period. However, two other potential solutions could be examined in future studies. One would be to begin the index at a later year in which there is more data available (e.g., 1990). Another would be to change the base year for calculating the index to a more data-rich year, thus increasing the uncertainty around the early years of LPI trends (Gregory et al., 2019). The downside is that the interpretation of trends would be different. The LPI would no longer measure change in global vertebrate

biodiversity relative to 1970, but relative to another year. Much of the change currently recorded in the index would have already occurred before the base year.

My modelling approach to quantifying trend reliability is subject to several limitations. Certain aspects of the underlying data, such as the distribution of observations and biases in which populations or species are tracked, are too complex to be included as factors in the model, but nonetheless may play a significant role in determining trend reliability. For example, monitoring efforts tend to focus on species at higher risk of extinction (Scheele et al., 2019). Many amphibian populations in the LPD were tracked because they were declining due to the devastating disease *chytridiomycosis*. This could negatively bias trends and falsely reduce variance in growth rates, leading the model to overestimate reliability because it assumes that tracked populations are randomly selected. On the other hand, Murali et al. (2022) found that population coverage in the LPD is biased towards protected areas, where species are less likely to be threatened, therefore potentially causing a positive bias in LPI trends. Another common phenomenon in the LPD is that time series are non-randomly distributed across time and/or space. For example, while some biodiversity hotspots (e.g., tropical Africa) are poorly known, others, especially islands (e.g., Madagascar), are well-studied (Nori et al., 2020), and this may bias entire realms. In the Afrotropical realm, only 12.5% of terrestrial reptile and amphibian populations in the LPD are from mainland Africa, while 20% are from a tiny uninhabited island near Mauritius, and more than half are from a single study that took place at a reserve in Madagascar over a nine-year period. In this case, the model likely severely underestimates the amount of data needed to get a reliable trend. While this is an extreme example, it makes the point that there are important underlying aspects of the data that cannot be assessed by a model. Fortunately, these issues tend to diminish when more data is present, and thus should not have a large effect on trends assessed as reliable.

The model also assumes that adding additional time series to the LPD will maintain the parameters of the regional taxonomic group to which they are added (e.g., the mean time series length and the level of variance in population mean growth rates will not change). This results in the model occasionally suggesting that unrealistically large numbers of populations are required to achieve a reliable trend. For example, it suggested that 6,768 populations of freshwater Afrotropical birds are needed. This likely occurred due to

problems with the existing data. Although there are 143 freshwater Afrotropical bird populations in the LPD, most of them are short and/or sporadically observed (the mean number of observations is 4.1), and observations are clustered in the 90's and 2000's, with only a single time series containing observations past 2009. However, this is an issue only for small or exceptionally poor-quality samples, and if more and better time series are added to the LPD, the model should improve its estimates.

Another limitation of my modelling approach is that I could not correct for the sizes of the real-world datasets (the number of populations that exist) that the LPD samples are drawing from, and therefore may overestimate the sample size needed to achieve a reliable trend for very small datasets. Although there are estimates of the number of species for each regional taxonomic group, my model uses populations as the base unit to measure sample size. I chose to base sample size on populations rather than species for two reasons. First, I found that mean growth rates within the LPD vary almost as much between populations within a species as they do between species. Therefore, I cannot assume that the trend of a population represents the trend of the species it belongs to any better than it represents the trend of its entire regional taxonomic group. Second, localized threats such as land-use change and habitat destruction are likely to affect some populations within a species disproportionately. Population extinctions also occur much more frequently than species extinctions and may serve as a prelude (Ceballos et al., 2017). However, a population is not a well-defined unit, and there are no estimates of how many populations each species or regional taxonomic group is composed of. While my testing suggested the number of existing populations can be assumed to be unimportant in determining trend reliability, this assumption breaks down when the sample comprises a large percentage of the dataset. It is unlikely that any regional taxonomic groups currently approach this level of representation within the LPD, but it is nonetheless an important caveat to be aware of.

3.6. Conclusion

The results of this study reveal the strengths and weaknesses in our understanding of global vertebrate biodiversity, highlighting the regional taxonomic groups for which we have enough data to make responsible decisions, as well as those on which future data gathering

and collation efforts should focus. Some underlying aspects of the data create biases that are not taken into account by my modelling approach, and more fine-scale studies on gaps in population trends should be performed to better understand these biases and where to divert scientific resources. I show that revisiting previously studied populations is a quick and efficient way to improve trend reliability for data deficient groups until more long-term studies can be completed and made available. The modelling approach I use to quantify trend reliability can also be generalized to assess other global and/or regional biodiversity indices that utilize population time series data. We are facing an urgent global biodiversity crisis made worse by biased and deficient data, but through careful study and cooperative global efforts we can solve the data problem and begin to 'bend the curve' of biodiversity toward a positive trend.

3.7. Bibliography

Arkilanian, A. A., Clements, C. F., Ozgul, A., & Baruah, G. (2020). Effect of time series length and resolution on abundance- and trait-based early warning signals of population declines. *Ecology*, *101*(7). https://doi.org/10.1002/ecy.3040

Baillie, J. E. M., Collen, B., Amin, R., Akcakaya, H. R., Butchart, S. H. M., Brummitt, N., Meagher, T. R., Ram, M., Hilton-Taylor, C., & Mace, G. M. (2008). Toward monitoring global biodiversity. *Conservation Letters*, *1*(1), 18–26. https://doi.org/10.1111/j.1755-263X.2008.00009.x

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-Qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6). https://doi.org/10.1371/journal.pbio.1000385

Butchart, S. H. M., Stattersfield, A. J., Baillie, J., Bennun, L. A., Stuart, S. N., Akçakaya, H. R., Hilton-Taylor, C., & Mace, G. M. (2005). Using Red List Indices to measure progress towards the 2010 target and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 255–268. https://doi.org/10.1098/rstb.2004.1583

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *4*(1), 300–307.

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring Change in Vertebrate Abundance: the Living Planet Index. *Conservation Biology*, *23*(2), 317–327. https://doi.org/10.1111/j.1523-1739.2008.01117.x

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The Tropical Biodiversity Data Gap: Addressing Disparity in Global Monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Freeman, R., McRae, L., Deinet, S., Amin, R., & Collen, B. (2021). *rlpi: Tools for calculating indices using the Living Planet Index method. R package version 0.1.0.* https://github.com/Zoological-Society-of-London/living_planet_index

Fryxell, J. M., Sinclair, A. R. E., & Caughley, G. (2014). *Wildlife Ecology, Conservation, and Management* (3rd ed.). Wiley Blackwell.

Gregory, R. D., Skorpilova, J., Vorisek, P., & Butler, S. (2019). An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe. *Ecological Indicators*, *103*, 676–687. https://doi.org/10.1016/j.ecolind.2019.04.064

Henriques, S., Böhm, M., Collen, B., Luedtke, J., Hoffmann, M., Hilton-Taylor, C., Cardoso, P., Butchart, S. H. M., & Freeman, R. (2020). Accelerating the monitoring of global biodiversity: Revisiting the sampled approach to generating Red List Indices. *Conservation Letters*, *13*(3). https://doi.org/10.1111/conl.12703

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Israel, G. D. (1992). Determining Sample Size. In *Fact Sheet PEOD-6*. Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida.

IUCN Species Survival Commission. (2012). *IUCN Red List Categories and Criteria: Version 3.1* (2nd ed.). IUCN.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology and Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J. M., Heurich, M., Jung, A., Klenke, R., Neumann, C., Pause, M., Rocchini, D., Schaepman, M. E., Schmidtlein, S., Schulz, K., Selsam, P., Settele, J., Skidmore, A. K., & Cord, A. F. (2016). Linking Earth Observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecological Indicators*, *70*, 317–339. https://doi.org/10.1016/j.ecolind.2016.06.022

Leung, B., Hargreaves, A. L., Greenberg, D. A., McGill, B., Dornelas, M., & Freeman, R. (2020). Clustered versus catastrophic global vertebrate declines. *Nature*, *588*(7837), 267–271. https://doi.org/10.1038/s41586-020-2920-6

Liu, X. (2012). Classification accuracy and cut pointselection. *Statistics in Medicine*, *31*(23), 2676–2686. https://doi.org/10.1002/sim.4509

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., & Randers, J. (2005). The Living Planet Index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 289–295. https://doi.org/10.1098/rstb.2004.1584

Mace, G. M., & Baillie, J. E. M. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, *21*(6), 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

McRae, L., Deinet, S., & Freeman, R. (2016). Data from: The diversity-weighted Living Planet Index: controlling for taxonomic bias in a global biodiversity indicator. In *Drya, Dataset*.

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1), 1–20. https://doi.org/10.1371/journal.pone.0169156

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*(8221). https://doi.org/10.1038/ncomms9221

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*(8). https://doi.org/10.1371/journal.pbio.1001127

Murali, G., de Oliveira Caetano, G. H., Barki, G., Meiri, S., & Roll, U. (2022). Emphasizing declining populations in the Living Planet Report. *Nature*, *601*, E20–E22. https://doi.org/10.1038/s41586-021-04165-z

Nichols, J. D., O'Connell, A. F., & Karanth, K. U. (2011). Camera traps in animal ecology and conservation: What's next? In *Camera Traps in Animal Ecology: Methods and Analyses* (pp. 253–263). Springer Japan. https://doi.org/10.1007/978-4-431-99495-4_14

Nori, J., Loyola, R., & Villalobos, F. (2020). Priority areas for conservation of and research focused on terrestrial vertebrates. *Conservation Biology*, *34*(5), 1281–1291. https://doi.org/10.1111/cobi.13476

Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K., & Jetz, W. (2021). Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biology*, *19*(8). https://doi.org/10.1371/journal.pbio.3001336

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R. D., Honrado, J. P., Jürgens, N., Opige, M., Schmeller, D. S., Tiago, P., & van Swaay, C. A. M. (2017). Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. *Biological Conservation*, *213*, 256–263. https://doi.org/10.1016/j.biocon.2016.07.014

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rose, R. A., Byler, D., Eastman, J. R., Fleishman, E., Geller, G., Goetz, S., Guild, L., Hamilton, H., Hansen, M., Headley, R., Hewson, J., Horning, N., Kaplin, B. A., Laporte, N., Leidner, A., Leimgruber, P., Morisette, J., Musinsky, J., Pintea, L., … Wilson, C. (2015). Ten ways remote sensing can contribute to conservation. *Conservation Biology*, *29*(2), 350–359. https://doi.org/10.1111/cobi.12397

RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com

Saha, A., McRae, L., Dodd, C. K., Gadsden, H., Hare, K. M., Lukoschek, V., & Böhm, M. (2018). Tracking Global Population Trends: Population Time-Series Data and a Living Planet Index for Reptiles. *Journal of Herpetology*, *52*(3), 259–268. https://doi.org/10.1670/17-076

Scheele, B. C., Legge, S., Blanchard, W., Garnett, S., Geyle, H., Gillespie, G., Harrison, P., Lindenmayer, D., Lintermans, M., Robinson, N., & Woinarski, J. (2019). Continental-scale assessment reveals inadequate monitoring for threatened vertebrates in a megadiverse country. *Biological Conservation*, *235*, 273–278. https://doi.org/10.1016/j.biocon.2019.04.023

Secretariat of the Convention on Biological Diversity. (2006). *Global Biodiversity Outlook 2*. https://www.cbd.int/gbo2/

Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., Finlayson, C. M., Ferrier, S., Freyhof, J., Hermoso, V., Juffe-Bignoli, D., Linke, S., Nel, J., Patricio, H. C., Pittock, J., Raghavan, R., Revenga, C., Simaika, J. P., & de Wever, A. (2017). Essential Biodiversity Variables for measuring change in global freshwater biodiversity. *Biological Conservation*, *213*, 272–279. https://doi.org/10.1016/j.biocon.2016.09.005

Wauchope, H. S., Amano, T., Sutherland, W. J., & Johnston, A. (2019). When can we trust population trends? A method for quantifying the effects of sampling interval and duration. *Methods in Ecology and Evolution*, *10*(12), 2067–2078. https://doi.org/10.1111/2041-210X.13302

Westcott, D. A., Fletcher, C. S., Mckeown, A., & Murphy, H. T. (2012). Assessment of monitoring power for highly mobile vertebrates. *Ecological Applications*, *22*(1), 374–383.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the global biodiversity information facility? *PLoS ONE*, *2*(11). https://doi.org/10.1371/journal.pone.0001124

Zylstra, E. R., Steidl, R. J., & Swann, D. E. (2010). Evaluating Survey Methods for Monitoring a Rare Vertebrate, the Sonoran Desert Tortoise. *Journal of Wildlife Management*, *74*(6), 1311–1318. https://doi.org/10.2193/2009-331

Chapter 4: Accounting for sampling and measurement error in aggregated abundance-based biodiversity indicators

4.1. Abstract

Multi-species biodiversity indicators are used to track biodiversity trends and make conservation and policy decisions. Therefore, it is important that they include reliable measurements of uncertainty. Precise confidence intervals can be calculated for indicators that track select groups of species, as they typically use systematic monitoring protocols and high-quality site-based data. However, for indicators that use aggregated time series data collected from multiple sources, such as the Living Planet Index and the Priority Species Indicator, existing methods of calculating confidence intervals do not account for sampling and measurement error at the population level. Proposed alternatives are untenable for these indicators due to a lack of site-based data, a high level of missing observations, and strong differences between population trends within species.

I developed the GAM-resampled rank envelope method to account for sampling and measurement error in aggregated abundance-based biodiversity indicators without requiring site-based data or similarity between population trends; this method is also highly robust against missing data. Here, I use synthetic time series data to compare my method to that used in the Living Planet Index and show that my method generates more accurate and precise confidence intervals across a wide range of parameters. I show that my method not only accounts for multiple levels of uncertainty in the confidence intervals, but also reduces the influence of outlier growth rates in the data that lead to inappropriately wide confidence intervals when measurement or sampling error are high.

4.2. Abbreviations

CI – confidence interval

CP – capture percentage

GO – GAM only

GAM – Generalized Additive Model

GC – GAM + Chain

GRRE – GAM-resampled rank envelope

LPI – Living Planet Index

LPD – Living Planet Database

PPS – populations per species

RE – rank envelope

TDV – trend deviation value

4.3. Introduction

Reliability is essential for multi-species biodiversity indicators (Soldaat et al., 2017). They compile complex scientific information into simple, user-friendly indices, and are therefore frequently used as information sources by policymakers and conservation decision-makers (Buckland et al., 2012; Jetz et al., 2019; Jones et al., 2011; Mace & Baillie, 2007; Mcowen et al., 2016; Nicholson et al., 2012; Rochette et al., 2019; Watermeyer et al., 2021). Biodiversity indicators are used to track biodiversity trends, make policy and conservation decisions, and evaluate the success of existing policies and programs (Mace & Baillie, 2007; Nicholson et al., 2012; Rowland et al., 2021). Indicators that monitor trends are calculated from samples or estimates and have some level of uncertainty, which is generally reported as intervals around the index values of the trend (e.g., Butchart et al., 2004; Eaton et al., 2015; Freeman et al., 2001; Gregory et al., 2004, 2005; Gregory & van Strien, 2010; McRae et al., 2017; Wotton et al., 2020). Intervals can account for variability in the data, quantify statistical precision of sampled or subsetted data, or show uncertainty from sampling or measurement errors (Rowland et al., 2021). When the data are used to generalize trends based on a sample, as in the Living Planet Index (LPI; McRae et al., 2017), confidence intervals (CIs) produced by resampling methods are considered an appropriate way to estimate precision in the index (Rowland et al., 2021). Many biodiversity indicators, especially at the national level, are constructed from standardized, systematically collected data from repeat sampling of the same sites over many years, referred to here as site-level data; others, especially at the global level where systematically collected data do not exist, are constructed from collated data from multiple sources and data types (e.g., The Living Planet Index: Loh et al., 2005; Large mammal trends in African protected areas: Craigie et al., 2010;

The Priority Species Indicator: Eaton et al., 2015). In the latter case, it is common to apply bootstrapping to the interannual variation of the species indices, considering species trends as replicates of the multi-species index (Craigie et al., 2010; Eaton et al., 2015; Gregory et al., 2019; Loh et al., 2005). However, this method neglects to take sampling and measurement error into account (Gregory et al., 2019; Soldaat et al., 2017).

Sampling error constitutes the uncertainty caused by sampling only part of a population, while measurement error is caused by e.g., species misidentification and non-detection, errors in counting, and inaccurate plot area measurements (Elphick, 2008; Holdaway et al., 2014). Because counting all individuals in a species or population is often expensive and inaccurate, sampling strategies are typically used to estimate population abundance (Fryxell et al., 2014). Sampling may also occur at each level of data collation. Sample-based counts represent population sizes; samples of population trends represent species trends; and samples of species trends represent taxonomic group trends. Each occurrence of sampling introduces uncertainty to the index. Unless this uncertainty is accounted for, CIs may appear artificially narrow, leading to a false sense of precision in the index.

Soldaat et al. (2017) proposed a method of calculating confidence intervals that accounts for sampling error at the population level using a Monte Carlo simulation. Their method has been applied to several bird and insect indices (Dennis et al., 2019; Gregory et al., 2019; Kamp et al., 2021; Wotton et al., 2020). However, the Monte Carlo simulation requires species indices with standard errors. Soldaat et al. (2017) suggested using the TRIM (Trends and Indices for Monitoring data) software (Pannekoek & van Strien, 2005) to produce species indices with standard errors. TRIM uses Poisson regression to estimate indices and trends from multiple counts and can impute missing counts based on other counts for the same year (Pannekoek & van Strien, 2005). TRIM is intended for use with site-level data. Gregory et al. (2005) and van Strien et al. (2001) employed a solution for European bird indicators; they used TRIM to estimate supranational year totals from country year totals by grouping countries by region and assuming that species showed similar population trends within regions. However, this solution assumes that grouped countries have similar trends and that at least one country-level count exists for each year (van Strien et al., 2001). Because the Living Planet Index (LPI) and similarly constructed indices are formed from

aggregated studies, they not only lack site-level data, but population-level trends are often dramatically different within species, with many years having no observations and many populations entirely missing. Therefore, an alternative approach is needed.

In this study, I present an alternative method of calculating CIs for multi-species biodiversity indicators based on aggregated time series data. My method not only accounts for measurement error and sampling error at the survey level as in Soldaat et al. (2017), but goes a step further by accounting for the additional sampling error introduced by sampling at the population and species levels. I use Generalized Additive Models (GAMs) to model population trends, then resample repeatedly from a multivariate normal distribution calculated from the means and covariance matrices of the coefficients of the GAMs, thus utilizing the variation inherent in the model as an estimate of sampling and measurement error at the survey level. The resampled population trends are pooled and used to propagate variation through each step of index calculation. Population trend variants are chosen for each species by random sampling with replacement, thus accounting for sampling error at the population level. Species trend variants are chosen in the same way as population trend variants, thus accounting for sampling error at the species level. The final index is formed by taking the mean of multi-species trend variants. Confidence intervals are calculated from the multi-species trend variants using the rank envelope method. The rank envelope method is a Monte Carlo global envelope test developed for hypothesis testing on spatial data (Myllymäki et al., 2017). It has also been shown to be useful in testing disparity through time curves to detect non-random bursts of evolution (Murrell, 2018). The rank envelope method has the advantage of reducing false positives (type 1 errors) in hypothesis testing by treating a curve or set of points as a whole instead of as independent points; in the context of calculating CIs, that means it will take serial correlation, an inherent property of time series, into account, which current methods do not.

I first test my GAM-resampled rank-envelope (GRRE) method using simulated time series datasets. I sample from the datasets, introduce error and remove some observations to emulate regional taxonomic groups in the Living Planet Database (LPD; McRae et al., 2016), then compare the GRRE method with the LPI method in terms of the percentage of the 'true' trend ('true' meaning unsampled, undegraded, and without introduced error)

captured by the CIs, the width of the CIs, and the accuracy of the sampled trend. Since the GRRE method introduces an important change in how time series are modelled, namely it avoids the use of the chain method which the LPI uses for time series with fewer than six data points or where the GAM fit is poor, I also reproduce the LPI method with all time series modelled by GAMs and use it as a control. Finally, I apply all three methods to the LPD itself.

4.4. Material and Methods

4.4.1. Synthetic data generation, observation error, data degradation, and sampling

I began by creating simulated datasets, each constructed from 1,000 population time series generated by a stochastic exponential model to represent real-world regional taxonomic groups. I created an observed version of each time series by modelling error, which I will refer to as observation error, or ε, using a normal distribution of the coefficient of variation (cv). Observation error ε can be assumed to include both sampling and measurement error, since I did not model them separately. The observed time series were then degraded by removing observations to resemble time series in the LPD. Finally, I randomly sampled time series from each dataset, without replacement, 20 times for each dataset. The process of generating datasets, adding error, degrading the time series, and sampling followed the methods detailed in Sections 3.3.1-3.3.4 of Chapter 3 of this thesis. Sampled parameters are referenced using the subscript 'samp' instead of 'ds' (e.g., μ_{samp} and σ_{samp}) and were calculated by first taking the means of the replicates for each dataset, then the overall mean (i.e., a mean of means). This was done to avoid issues of pseudoreplication, as samples from the same dataset are not independent.

I varied ε across a range of cv (see Table 4.1), but when other parameters varied, ε was fixed at μ_ε = 0.3 and σ_ε = 0.2. These values are higher than I used in Chapter 3 for two reasons. First, the range of ε chosen in Chapter 3 may have been an underestimate (e.g., de Valpine, 2003; Dunham et al., 2001; Viljugrein et al., 2005; G. Wang et al., 2006). Second, while the simulations I did in Chapter 3 showed that ε has no effect on trend accuracy (Fig. S3.4 in Appendix 2), increasing ε does increase within-population variance in growth rates (Fig. S4.1 in Appendix 3), which may affect confidence intervals. In samples, process noise, η, and ε

cannot be measured separately, so I refer to them as combined error, $\eta\varepsilon$, except in specific cases where observation error was set to zero.

4.4.2. Calculation of sampled trends for GRRE method

Non-linear index trends were calculated for each sample, following the LPI method described in McRae et al. (2017) and Section 3.3.5 in Chapter 3 of this thesis, but with some deviations. As in the LPI, population sizes were converted to growth rates by

$$r_t = log_{10} \frac{N_t}{N_{t-1}} \quad (1)$$

All time series were modeled using a Generalized Additive Model (GAM), as described in Collen et al. (2009). For each modelled time series, a normal distribution of 100 time series variants was drawn from the means and covariance matrix of the coefficients of the GAM. This technique uses the variance inherent in the GAM to represent observation error, ε. Since variance cannot be extracted this way from time series modelled using the chain method, I did not employ the chain method for any time series; nor did I check GAM fit, as there was no alternative in the case of poor fit.

Species trends were produced as follows. First, all time series variants for a given species were pooled and a sample equal in size to the number of populations in the sampled dataset belonging to that species was randomly selected from the pool, with replacement. For a species consisting of 10 populations, 10 time series variants would be selected from a pool of 1,000 (10 populations * 100 variants = 1,000). This step accounts for sampling error involved in selecting populations, as not all populations from the dataset can be in each sample. Second, average growth rates were calculated by

$$\overline{r_t} = \frac{1}{n} \sum_{i=1}^{n_t} r_{it} \quad (2)$$

where n_t is the number of time series variants selected from the pool, r_{it} is the growth rate for time series *i* at year *t*, and $\overline{r_t}$ is the average growth rate at year *t*. Growth rates were capped at [-1:1]. Finally, index values were calculated by

$$I_t = I_{t-1} * 10^{\overline{r_t}}, \qquad I_0 = 1 \quad (3)$$

where *I* is the index value and *t* is the year. This process was repeated 3,000 times for each species to generate 3,000 variants of each species trend. A large number of variants is important for generating CIs using the rank envelope method (see Section 4.4.6 below).

Species trends were then pooled and a sample size equal to the number of species in the sampled dataset was randomly selected from the pool, with replacement. This step accounts for sampling error involved in selecting species, as it is likely that not all species from the dataset will be represented in each sample. The sample was averaged using the geometric mean, as with the population variants. This was repeated 3,000 times to provide 3,000 multi-species trend variants. The final index was calculated as the arithmetic mean of the multi-species trend variants.

4.4.3. Calculation of sampled trends using the LPI (GC) method

Non-linear index trends were calculated from each sample exactly as described in Section 3.3.5 of Chapter 3. Since this method uses both GAMs and the chain method to model populations, it will henceforth be referred to as the GC (GAM + Chain) method.

4.4.4. Calculation of sampled trends using a modified LPI (GO) method

Non-linear index trends were calculated from each sample as described in Section 3.3.5 of Chapter 3, except that every time series was modelled using a GAM. Model fit was not tested, and the chain method was not employed for any time series. This was to make results directly comparable to the GRRE method, as the smoothing effect of a GAM removes outlier growth rates that can lead to wider CIs. Since this method uses only GAMs, it will henceforth be referred to as the GO (GAM only) method.

4.4.5. Calculation of the 'true' trend & trend comparison

An index was calculated for each unsampled original dataset (without observation error or degradation), following McRae et al. (2017), but without using a GAM or the chain method to model the time series (because there were no missing values).

Sampled trends calculated by all three methods were then compared with the 'true' trend using the Jaccard distance metric, with the resulting value referred to as a trend deviation

value, or TDV, as described in Section 3.3.7 of Chapter 3. The TDV represents the complement of trend accuracy, therefore lower is better.

4.4.6. Confidence intervals for sampled trends using the GRRE method

Confidence intervals for sampled trends were determined using the rank envelope (RE) method adapted from Murrell (2018) and Myllymäki et al. (2017). The rank envelope test is a method widely used in spatial statistics to evaluate the range of output values of a function which leads to rejection of the null hypothesis (Myllymäki et al., 2017). Unlike the LPI method of bootstrapping the species rates of change, the GRRE method accounts for serial correlation by not treating time points within trends as independent. Instead, each trend is assigned a ranking based on its most extreme deviation from the median at any given time point and then removing the most extreme trends to form confidence intervals. The GRRE method employs a non-parametric rank test to determine CIs, and Myllymäki et al. (2017) recommends at least 2,500 simulated variants. Here I used 3,000 multi-species trend variants, calculated as described in Section 4.4.2, and assigned two ranks for every time point, based on its index value at that time point relative to all other variants. One rank was in ascending order of index value, the other in descending order. A global rank for each variant was then determined according to the maximum rank (ascending or descending) of that variant at *any* time point. The CIs for the final index were defined by the highest and lowest index values at each time point within the 95% globally lowest-ranked multi-species trend variants. In other words, I excluded the most extreme 5% of variants and then took the highest and lowest values at each time point from the remaining variants as the CIs.

4.4.7. Confidence intervals for sampled trends by bootstrapping the species rates of change (GC and GO methods)

For both the GC method and the GO method, confidence intervals for sampled trends were determined according to Collen et al. (2009). For each year *t*, 3,000 variants of the multi-species trend were calculated from n randomly selected species annual rates of change (with replacement), where n is the number of species with rates of change for that year. The bounds of the central 2,850 values for each year formed the 95% CIs.

4.4.8. Percentage of 'true' trend captured within confidence interval of sampled trends (capture percentage)

For each sampled multi-species index, I calculated the percentage of time points of the 'true' trend that fell between the values of the upper and lower 95% CIs of the index. I then averaged across samples to determine the mean percentage of the trend that was captured by the 95% confidence intervals for that dataset. This will be referred to as mean capture percentage, or CP.

4.4.9. Mean normalized width of confidence intervals of sampled trends

The mean difference between the upper and lower CIs of the index was calculated, excluding the first year where the difference was always zero, and this was then averaged across samples to determine mean CI width for a given dataset. Since CI width is naturally higher for increasing trends than for decreasing trends due to differences in index values, the mean CI width was then divided by the mean index value of the trend to normalize.

4.4.10. Comparison of confidence interval methods

I applied the GC, GO, and GRRE methods described above to simulated datasets to test performance at different parameter settings, including mean time series length, sample size, mean number of populations assigned to each species (populations per species, or PPS), mean of population mean growth rates (overall slope of the dataset; μ_{ds}), standard deviation in population mean growth rates (variation in growth rates among species; σ_{ds}), mean of the population standard deviations (process noise, which is the inherent, or 'real', stochasticity in a time series, unrelated to measurement or observation; μ_{η}), and mean coefficient of variation (observation error; the standard deviation was set to the same as the mean; cv_{ε}). Values for these parameters are provided in Table 4.1. For each parameter, I calculated the mean percentage of the 'true' trend captured, the mean confidence interval width, and the mean trend deviation for the GO, GC and GRRE methods.

I then applied all three methods to regional taxonomic groups from the Living Planet Database and plotted the results side-by-side to visually demonstrate the differences.

Table 4.1. Parameters with values for simulated datasets and degraded samples.

Independent Variable	Values in Datasets	Values in Samples
Sample Size (out of 1,000 time series)	-	50, 70, 100, 150, 200, 300, 500, 800
Mean Length of Time Series	5.5, 8, 13, 18, 23, 31, 39	6.1, 8.3, 13, 18, 23, 31, 39
Mean of Pop. Mean Growth Rates, μ_{ds}	-0.08, -0.04, -0.02, 0, 0.02, 0.04, 0.08	-0.1, -0.05, -0.01, 0.01, 0.05, 0.1
St. Dev. of Pop. Mean Growth Rates, σ_{ds}	0.05, 0.15, 0.25, 0.4, 0.55, 0.6, 0.8, 1	0.1, 0.2, 0.3, 0.5, 0.7, 0.9, 1.1
***Mean of Pop. Std. Deviations, μ_{η}**	0.1, 0.2, 0.3, 0.4, 0.55, 0.7, 0.85, 1	0.32, 0.37, 0.41, 0.47, 0.52, 0.6, 0.65, 0.7
Mean Number of Pops / Species (PPS)	5, 10, 15, 20, 30, 50, 100, 200	1.6, 2.3, 3.2, 4.1, 6.0, 10, 20, 40
Mean Observation Error, cv_{ε}	-	0.05, 0.25, 0.45, 0.65, 0.85, 1, 2, 4

*This parameter represents process noise in simulated datasets, but in degraded samples it represents process noise and observation error combined.

4.5. Results

Trend deviation values (TDV, lower is better) were similar across methods for all levels of mean time series length, sample size, mean PPS_{samp}, μ_{samp}, σ_{samp}, or $\eta\varepsilon$ (Fig. 4.1), but TDVs were higher for the GC method than GO or GRRE at all but the lowest level of observation error (Fig. 4.1F). All three methods showed lower TDV when mean time series length increased (Fig. 4.1A), and higher TDV when μ_{samp}, σ_{samp}, or $\eta\varepsilon$ increased (Fig. 4.1D, E, G). The mean and range of TDV decreased for all methods as sample size increased (Fig. 4.1B). There was no clear effect on TDV when altering the mean number of populations per species for any method (Fig. 4.1C). Increasing observation error caused an increase in TDV for the GC method; however, the GO and GRRE methods showed an increased range of TDV but no clear effect on the mean (Fig. 4.1F). Mean TDV remained comparable between all three methods when process noise (the 'real' stochasticity in a time series) increased without the presence of observation error (Fig. 4.1H).

Capture percentage (CP) was consistently higher for the GRRE method than the GC or GO methods across all tested levels of every parameter (Fig. 4.2). The GC method consistently had a higher CP than the GO method, except when σ_{samp} was very high (Fig. 4.2). Capture percentage increased for the GO and GC methods at both low and high mean time series lengths, while the GRRE method was unaffected (Fig. 4.2A). All methods showed an increase in CP as mean PPS_{samp} increased until it surpassed 4.1 (GRRE method) or 6 (GC and GO methods) (Fig. 4.2C). The same relationship was evident with increasing sample size, except when sample size was very low (all methods) or greater than 200 (GRRE method only) (Fig.

4.2B). The GC method showed lower CP as σ_{samp} increased but held steady at σ_{samp} above 0.5; there was no apparent effect of σ_{samp} on GO or GRRE methods (Fig. 4.2E). The parameters μ_{samp} and $\eta\varepsilon$ did not seem to have an effect on CP for any of the methods (Fig. 4.2D, G). The GC method showed an increase in CP as observation error increased up to 45%, then maintained the higher CP as observation error increased to the maximum tested level of 400% (Fig. 4.2F). When only process noise was present, capture percentage increased slightly for the GC method, but the effect was weak (Fig. 4.2H).

Mean normalized CI width decreased for all methods with increasing mean time series length or sample size, and increased for all methods with increasing μ_{samp}, σ_{samp}, or $\eta\varepsilon$ (Fig. 4.3). Confidence interval width increased for all methods with increasing PPS_{samp} when mean PPS_{samp} was above 6 (Fig. 4.3C). The range of CI widths increased with increasing σ_{samp} or $\eta\varepsilon$ (Fig. 4.3E, G). The GO method resulted in narrower CIs than the other two methods across all levels of all parameters; the GRRE method resulted in wider CIs than the other two methods across all levels of all parameters except observation error (Fig. 4.3). The GO method showed no effect of observation error on CI width; CI width increased for the GRRE and GC methods, but the increase was much faster for the GC method, with the GC method having the widest CIs for all values of observation error above 45% (Fig. 4.3F). When no observation error was present, results were similar as for combined error, with increasing process noise resulted in wider CIs for all methods, the strongest increases occurring for the GRRE method, and the weakest for the GO method (Fig. 4.3H).

In contrast to the results from the simulated datasets, regional taxonomic groups in the LPD typically had wider CIs with the GC method than the GRRE method (e.g., Fig. 4.4; for full results see Figs S4.2-S4.18 in Appendix 3). The GO method results were more consistent with the simulated datasets, often showing narrower CIs than the GRRE and GC methods (Fig. 4.4), but not in all cases (e.g., marine South Temperate birds in Fig. 4.4). When sections of trends were represented by one or more time series belonging to a single species, CIs were calculated for the GRRE method but could not be calculated for the GO and GC methods, resulting in wider overall CIs for the GRRE method for those trends (e.g., freshwater Afrotropical herps in Fig. 4.4).

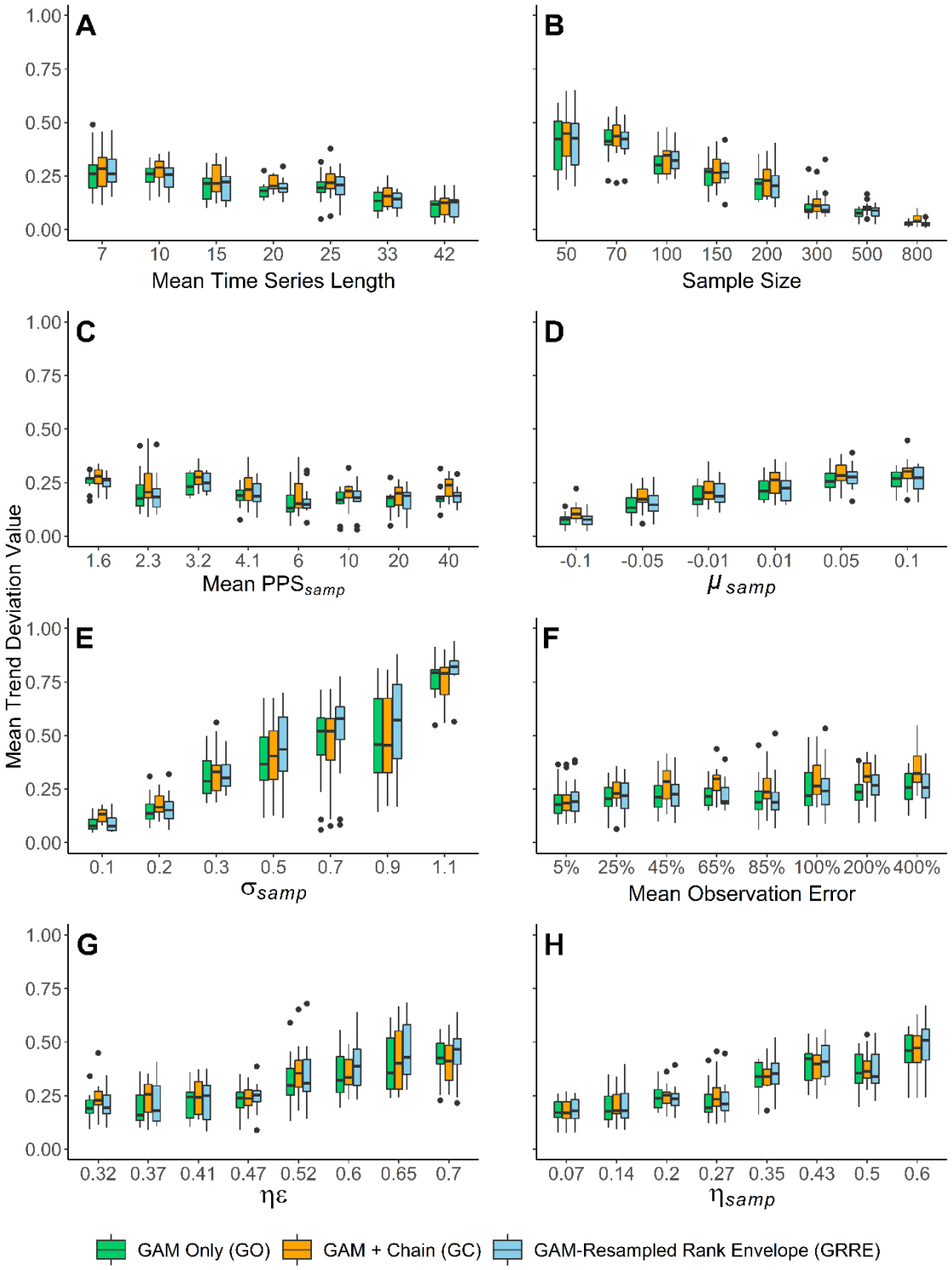


Figure 4.1. Mean trend deviation value of sampled trends. The following eight parameters were varied: A) time series length, B) sample size, C) number of populations per species, D) population mean growth rates, E) variance in population mean growth rates, F) observation error, G) process noise with fixed observation error, and H) process noise without observation error. Reported values are from the samples. Fixed values were standardized as follows – dataset size: 1000; sample size: 200; PPS_{ds}: 20; μ_{ds}: 0; σ_{ds}: 0.2; η_{ds}: 0.2; ε: 0.3. Twelve datasets were simulated at each parameter value, with 20 samples collected per dataset.

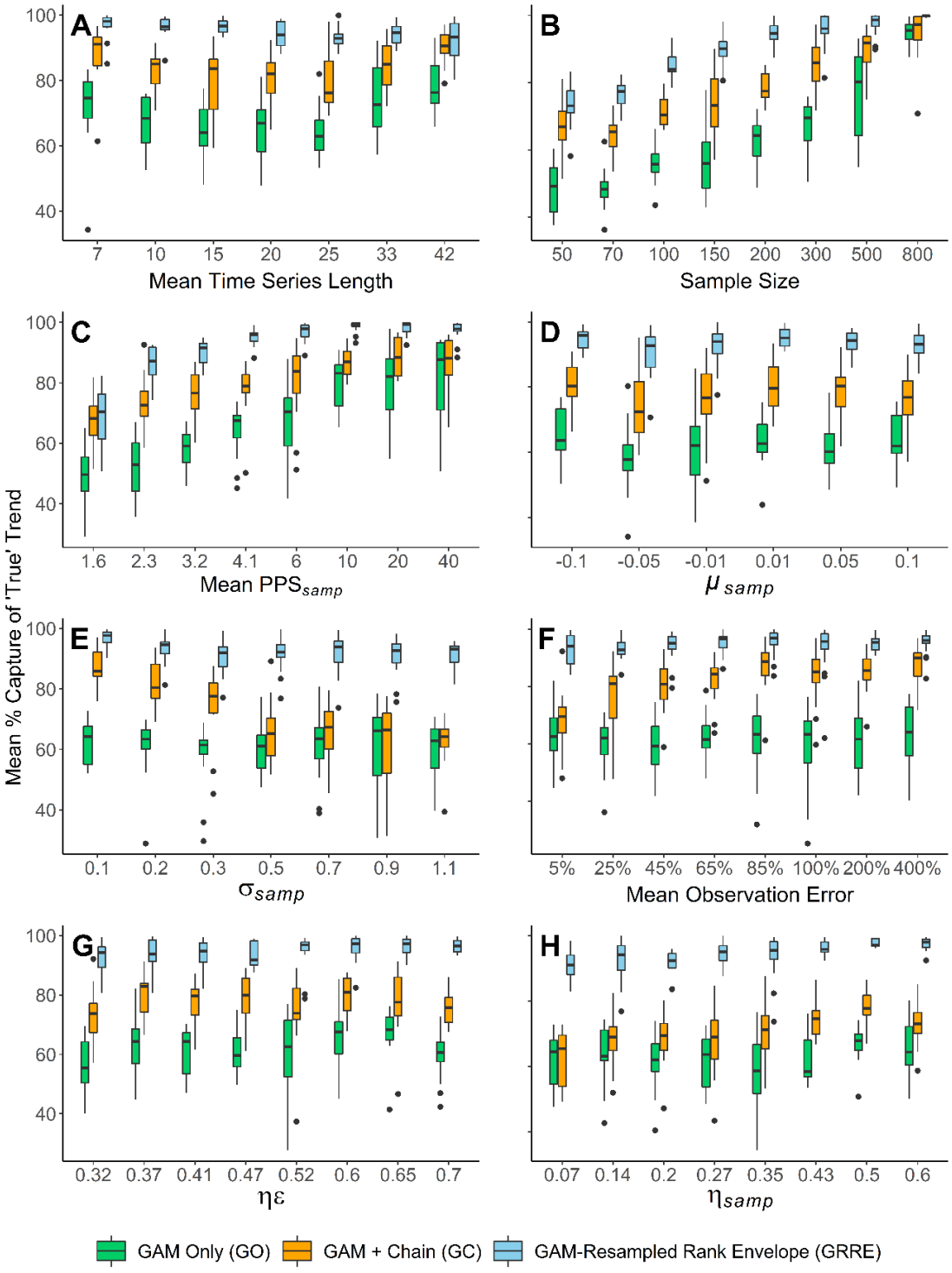


Figure 4.2. Percentage of the 'true' trend captured within sample confidence intervals. Eight parameters were varied: A) time series length, B) sample size, C) populations per species, D) population mean growth rates, E) variance in population mean growth rates, F) observation error, G) process noise with fixed observation error, and H) process noise without observation error. Reported values are from the samples. Fixed values were standardized as follows – dataset size: 1000; sample size: 200; PPS$_{ds}$: 20; μ_{ds}: 0; σ_{ds}: 0.2; η_{ds}: 0.2; ε: 0.3. Twelve datasets were simulated at each parameter value, with 20 samples collected per dataset.

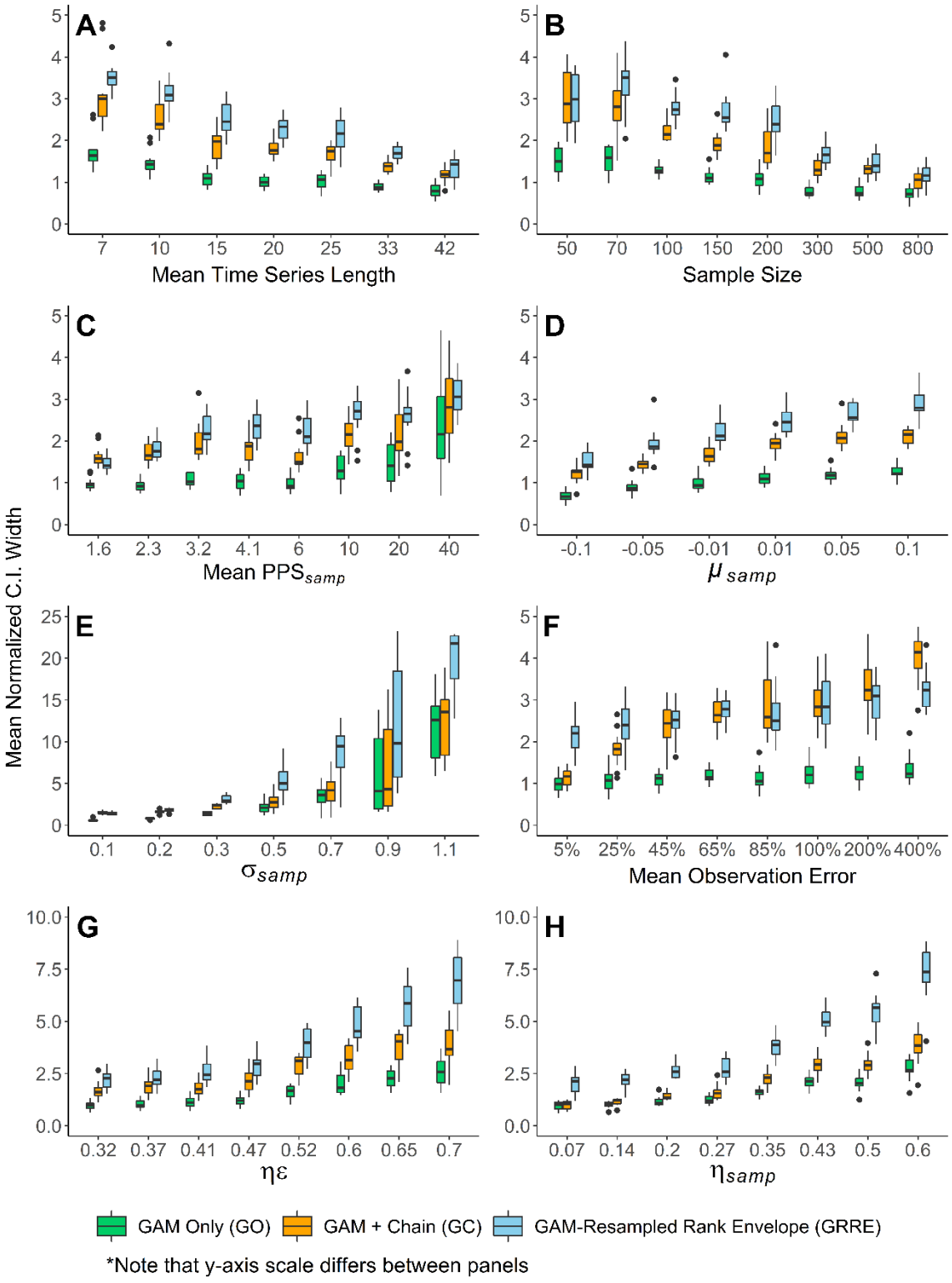


Figure 4.3. Normalized width of sampled confidence intervals. Eight parameters were varied: A) time series length, B) sample size, C) populations per species, D) population mean growth rates, E) variance in population mean growth rates, F) observation error, G) process noise with fixed observation error, and H) process noise without observation error. Reported values are from the samples. Fixed values were standardized as follows – dataset size: 1000; sample size: 200; PPS$_{ds}$: 20; μ_{ds}: 0; σ_{ds}: 0.2; η_{ds}: 0.2; ε: 0.3. Twelve datasets were simulated at each parameter value, with 20 samples collected per dataset.

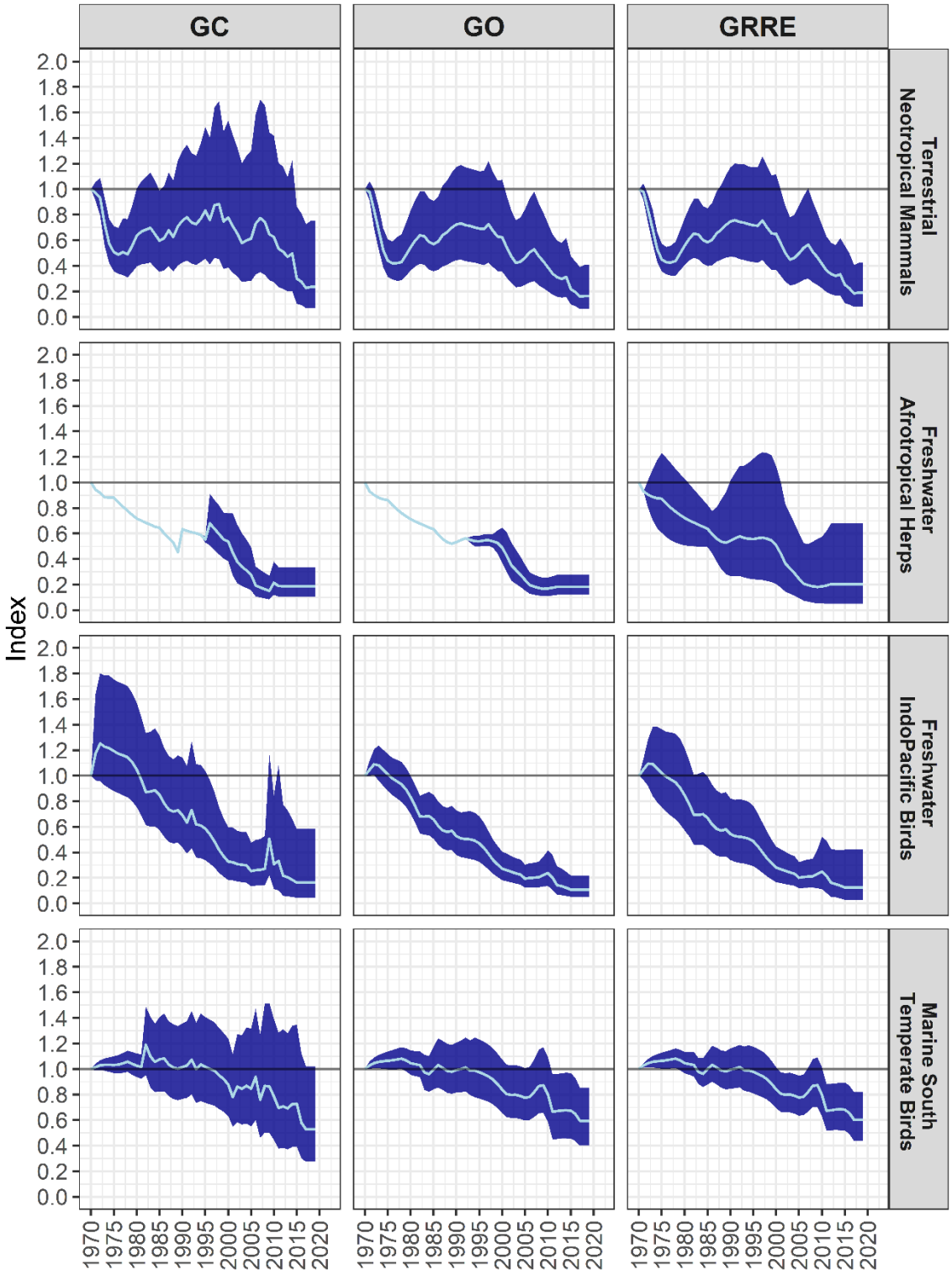


Figure 4.4. LPI trends for four regional taxonomic groups. From left to right, the columns were calculated using the GAM + Chain (GC) method, which is the method used for the LPI; the GAM Only (GO) method; and the GAM-Resampled Rank Envelope (GRRE) method, respectively. The GC and GO methods were unable calculate CIs for the first half of the freshwater Afrotropical herps trend because the group is represented in the LPD by a single species during those years.

4.6. Discussion

Existing methods of calculating confidence intervals (CIs) for biodiversity indicators that are based on aggregated time series data collected from multiple sources rather than through systematic monitoring protocols fail to account for sampling and measurement error at the population level (Soldaat et al., 2017) and serial correlation. I tested an alternative method, the GAM-resampled rank envelope (GRRE) method, against the method used in the LPI (here called the GC method) as well as a control method identical to the GC method except that all time series are modelled using GAMs (the GO method). I found the GRRE method maintained the highest capture percentage (CP) across all parameter ranges, while maintaining similar trend accuracy (TDV) to the GC and GO methods. It also produced wider CIs in simulations, with one exception; when observation error was high, CI width for the GC method exceeded the GRRE method. However, even when the GC method produced wider CIs, it was unable to match the GRRE method for CP. At the tested sample size of 200 time series, the GRRE method maintained a CP of approximately 95% across most parameter settings, while the CP for the GC and GO methods was consistently too low, only reaching 95% at a sample size of 800 (80% of the dataset), four times higher than the GRRE method. This suggests that the GRRE method is better at accounting for error and noise and produces more efficient and more accurate CIs.

The wider CIs produced by the GRRE method when using simulated datasets reflect the uncertainty introduced from three sources: a combination of measurement and sampling error introduced when estimating population abundances or densities (here termed observation error), sampling error introduced when selecting only certain populations within each species, and sampling error introduced when selecting only certain species within each dataset (regional taxonomic group). The GC method used to calculate LPI CIs makes no attempt to account for observation error nor the sampling error related to population selection; instead, CIs in the LPI reflect only the sampling error related to species selection (Gregory et al., 2019; Soldaat et al., 2017). Confidence interval width increases more strongly in response to increased observation error for the GC method than it does for the GRRE method. This is because the GC method models some time series by the chain method, which uses log-linear interpolation, and the additional variance introduced to those

time series by randomized observation error is not smoothed away. This results in outlier growth rates propagating to species trends, and the growth rates of the species trends are then bootstrapped to calculate CIs. Final CIs around the multi-species index are calculated by cumulative product, therefore wide CIs around any year will increase CI width in later years. However, the increased width makes these CIs inefficient. Capture percentage increases but remains below that of the GRRE method, likely because the wider CIs are offset by reduced trend accuracy caused by the propagation of outlier growth rates. The GC and GO methods not only ignore sampling error but calculate CIs separately for each year and thus do not account for serial correlation; even when producing wider CIs than the GRRE method, it is clear that the GC method does not work as intended.

When all time series are modelled by GAMs (GO method), increased observation error does not change CI width, capture percentage, nor trend accuracy, because the error is random and is largely smoothed away; outlier growth rates are not propagated to species trends. Observation error does cause wider CIs for the GRRE method, but without any change in trend accuracy or capture percentage. While the GRRE method models all time series by GAMs and therefore smooths away the outlier growth rates, it also captures the increased variance from the covariance matrices of the GAMs and propagates this variance through to the CIs, increasing their width. However, given that TDV does not increase, the lack of corresponding improvement in capture percentage suggests that increased observation error reduces CI efficiency for the GRRE method as well as for the GC method.

The differences in response to process noise vs observation error in the GRRE and GO methods is interesting, given that process noise and observation error are both captured in the samples as within-population variance in the growth rates. However, process noise is present in the unsampled dataset while observation error is not. Increasing process noise also increases the variance in mean growth rates *between* populations in the samples (but not in the original dataset), while increasing observation error does not (see Fig. S4.1 in Appendix 3). Since all methods respond much more strongly to increased variance between populations than to increased variance within populations, this is likely a major reason for the difference. Process noise is applied to the entire dataset, so samples likely capture a

similar range of mean growth rates in a smaller number of populations, thus increasing the standard deviation. Observation error, on the other hand, is present only in the samples.

Capture percentage is independent of time series length, but strongly depends on sample size for all methods. As time series length is reduced, TDV increases, and CI width increases to compensate. The same occurs when sample size is reduced; however, compensation is limited because when sample size is low outliers are less likely to be sampled and the range of growth rates represented in the sample reduces. The mean PPS in the samples also reduces at low sample sizes, and low PPS can reduce CP. When sample size (the number of populations in a sample) is set, sample PPS is the inverse of species richness. Raising PPS (reducing richness) increases CI width and CP for all methods without affecting trend accuracy. When there are few species, if other parameters stay the same there will be higher variance in population mean growth rates within each species. Due to randomized resampling with replacement, the GRRE method should produce a wider variation in trends for each species, and a wider variation in multi-species trends, which will in turn lead to wider CIs. While this is not the case for the GC and GO methods, they do use random sampling of species growth rates, with replacement, to form CIs; fewer species trends mean a higher frequency of extreme growth rates being selected multiple times, and thus more extreme CIs when those growth rates are averaged. If sampling were performed without replacement, I would not expect the same relationship between CI width and PPS or CP and PPS. As sample size decreases, PPS drops faster than richness because the likelihood of randomly selecting populations from different species is higher than from the same species. This drop in PPS likely tempers the rise in CI width, further dropping CP in addition to the effect of the reduced representation of the range of growth rates.

Increasing the mean population growth rates does not affect CP for any method, although it decreases trend accuracy and increases normalized CI width for all methods. Increased CI width due to more extreme trends balances decreased trend accuracy to maintain capture rates. More positive growth rates result in larger index values, which amplifies differences in trends caused by small differences in growth rates, while more negative growth rates minimize differences due to a smaller potential range of index values. This is partially compensated for by normalization of CI widths and the use of a normalizing distance

measure for comparing trends. However, relative differences are still higher because index values cannot go below zero but are unbounded on the positive side.

When all three confidence interval methods were applied to LPI regional taxonomic groups, the GC method consistently resulted in the widest CIs. The difference is likely due to two types of growth rate outliers. The first type of outlier occurs in individual growth rates within LPI time series, which may occur due to high levels of observation error. Typical observation error estimates for vertebrates are within the standard range I modelled for my tests (Fryxell et al., 2014; X. Wang et al., 2013; Westcott et al., 2012; Zylstra et al., 2010), but it has been shown that observation error may depend on density, reaching 400% for voles when burrow density was low but less than 10% when density and sample size were high (Lisická et al., 2007). A deeper look into the LPD could shed light on the matter by revealing whether outlying growth rates occur more frequently when abundance or density is low. It would be difficult to determine whether process noise or observation error is responsible; however, with my simulated datasets CI widths for the GC method surpassed the GRRE method when only observation error was high but did not show the same effect from high process noise. The second type of outlier occurs in the mean growth rates for population time series, and may also explain why even the GO method produced wider CIs than the GRRE method for some regional taxonomic group trends, as this type of outlier is not smoothed away by GAMs. The GC and GO methods calculate CIs by bootstrapping species growth rates. Each species has equal power to influence CIs regardless of the number of populations associated with it, so species represented by a single population, which are more likely to be outliers due to the lack of moderating effect from other populations contributing to the mean, have as much power to influence CIs as species represented by many populations. Therefore, some outlier populations may be highly influential, especially when sample sizes are small. This does not occur with the GRRE method, which calculates CIs from multi-species trend variants rather than species growth rates and therefore gives individual populations less influence.

My implementation of the GRRE method employs successive trend calculation steps that involve random sampling with replacement to account for sampling error. This method assumes that sampling of populations and species is random. While this assumption would

be entirely false for some indicators that use carefully selected species, it is closer to the truth for indicators like the LPI that are based on multiple and varied data sources. Time series in the LPD are selected according to availability of survey data, and surveyed populations are chosen for various reasons, such as protected status, commercial or popular interest, or suspected or known declines (Scheele et al., 2019). A weighting system applied to LPI trends accounts for biases in the number of populations available in the LPD for particular taxa and regions (McRae et al., 2017), but biases in the reasons those populations are studied are not accounted for (see Discussion in Chapter 3). However, since selection biases are varied, with some favouring positive trends and others negative trends, it is reasonable to assume that selection is random.

4.7. Conclusion

My analysis suggests that the current method of calculating confidence intervals for the Living Planet Index overestimates their width, likely due to high levels of observation error and a failure to account for serial correlation in the time series, which may also reduce trend accuracy. If the observation error is random, modelling all populations with GAMs and calculating the multi-species index and confidence intervals using the GRRE method I present here would not only make the confidence intervals narrower, but likely improve capture percentage and trend accuracy as well. The GRRE method proved robust against missing observations (short time series), extreme growth rates, and high levels of process noise and observation error. It improved mean capture of the 'true' trend by sampled confidence intervals under all tested parameter ranges, and maintained a capture percentage of approximately 95% across most parameter settings. Therefore, it could also be used to improve confidence interval efficiency for other indicators that are based on aggregated population time series data. While current methods are not viable because they fail to account for serial correlation or sampling and measurement error, the GRRE method overcomes both issues and produces accurate confidence intervals.

4.8. Bibliography

Buckland, S. T., Baillie, S. R., Dick, J. M. P., Elston, D. A., Magurran, A. E., Scott, E. M., Smith, R. I., Somerfield, P. J., Studeny, A. C., & Watt, A. (2012). How should regional biodiversity be monitored? *Environmental and Ecological Statistics*, *19*(4), 601–626. https://doi.org/10.1007/s10651-012-0202-7

Butchart, S. H. M., Stattersfield, A. J., Bennun, L. A., Shutes, S. M., Akçakaya, H. R., Baillie, J. E. M., Stuart, S. N., Hilton-Taylor, C., & Mace, G. M. (2004). Measuring global trends in the status of biodiversity: Red list indices for birds. *PLoS Biology*, *2*(12). https://doi.org/10.1371/journal.pbio.0020383

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring Change in Vertebrate Abundance: the Living Planet Index. *Conservation Biology*, *23*(2), 317–327. https://doi.org/10.1111/j.1523-1739.2008.01117.x

Craigie, I. D., Baillie, J. E. M., Balmford, A., Carbone, C., Collen, B., Green, R. E., & Hutton, J. M. (2010). Large mammal population declines in Africa's protected areas. *Biological Conservation*, *143*(9), 2221–2228. https://doi.org/10.1016/j.biocon.2010.06.007

de Valpine, P. (2003). Better inferences from population-dynamics experiments using Monte Carlo state-space likelihood methods. *Ecology*, *84*(11), 3064–3077.

Dennis, E. B., Brereton, T. M., Morgan, B. J. T., Fox, R., Shortall, C. R., Prescott, T., & Foster, S. (2019). Trends and indicators for quantifying moth abundance and occupancy in Scotland. *Journal of Insect Conservation*, *23*(2), 369–380. https://doi.org/10.1007/s10841-019-00135-z

Dunham, J., Rieman, B., & Davis, K. (2001). Sources and Magnitude of Sampling Error in Redd Counts for Bull Trout. *North American Journal of Fisheries Management*, *21*(2), 343–352. https://doi.org/10.1577/1548-8675(2001)021<0343:samose>2.0.co;2

Eaton, M. A., Burns, F., Isaac, N. J. B., Gregory, R. D., August, T. A., Barlow, K. E., Brereton, T., Brooks, D. R., al Fulaij, N., Haysom, K. A., Noble, D. G., Outhwaite, C., Powney, G. D., Procter, D., & Williams, J. (2015). The priority species indicator: measuring the trends in threatened species in the UK. *Biodiversity*, *16*(2–3), 108–119. https://doi.org/10.1080/14888386.2015.1068222

Elphick, C. S. (2008). How you count counts: The importance of methods research in applied ecology. *Journal of Applied Ecology*, *45*(5), 1313–1320. https://doi.org/10.1111/j.1365-2664.2008.01545.x

Freeman, S. N., Baillie, S. R., & Gregory, R. D. (2001). *Statistical analysis of an indicator of population trends in farmland birds*. British trust for Ornithology.

Fryxell, J. M., Sinclair, A. R. E., & Caughley, G. (2014). *Wildlife Ecology, Conservation, and Management* (3rd ed.). Wiley Blackwell.

Gregory, R. D., Noble, D. G., & Custance, J. (2004). The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. *Ibis*, *146*(Suppl. 2), 1–13.

Gregory, R. D., Skorpilova, J., Vorisek, P., & Butler, S. (2019). An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe. *Ecological Indicators*, *103*, 676–687. https://doi.org/10.1016/j.ecolind.2019.04.064

Gregory, R. D., & van Strien, A. (2010). Wild bird indicators: using composite population trends of birds as measures of environmental health. *Ornithological Science*, *9*, 3–22.

Gregory, R. D., van Strien, A., Vorisek, P., Meyling, A. W. G., Noble, D. G., Foppen, R. P. B., & Gibbons, D. W. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 269–288. https://doi.org/10.1098/rstb.2004.1602

Holdaway, R. J., McNeill, S. J., Mason, N. W. H., & Carswell, F. E. (2014). Propagating Uncertainty in Plot-based Estimates of Forest Carbon Stock and Carbon Stock Change. *Ecosystems*, *17*(4), 627–640. https://doi.org/10.1007/s10021-014-9749-5

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology and Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Jones, J. P. G., Collen, B., Atkinson, G., Baxter, P. W. J., Bubb, P., Illian, J. B., Katzner, T. E., Keane, A., Loh, J., Mcdonald-Madden, E., Nicholson, E., Pereira, H. M., Possingham, H. P., Pullin, A. S., Rodrigues, A. S. L., Ruiz-Gutierrez, V., Sommerville, M., & Milner-Gulland, E. J. (2011). The Why, What, and How of Global Biodiversity Indicators Beyond the 2010 Target. *Conservation Biology*, *25*(3), 450–457. https://doi.org/10.1111/j.1523-1739.2010.01605.x

Kamp, J., Frank, C., Trautmann, S., Busch, M., Dröschmeister, R., Flade, M., Gerlach, B., Karthäuser, J., Kunz, F., Mitschke, A., Schwarz, J., & Sudfeldt, C. (2021). Population trends of common breeding birds in Germany 1990–2018. *Journal of Ornithology*, *162*(1), 1–15. https://doi.org/10.1007/s10336-020-01830-4

Lisická, L., Losík, J., Zejda, J., Heroldová, M., Nesvadbová, J., & Tkadlec, E. (2007). Measurement error in a burrow index to monitor relative population size in the common vole. *Folia Zoologica*, *56*(2), 169–176.

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., & Randers, J. (2005). The Living Planet Index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 289–295. https://doi.org/10.1098/rstb.2004.1584

Mace, G. M., & Baillie, J. E. M. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, *21*(6), 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

Mcowen, C. J., Ivory, S., Dixon, M. J. R., Regan, E. C., Obrecht, A., Tittensor, D. P., Teller, A., & Chenery, A. M. (2016). Sufficiency and Suitability of Global Biodiversity Indicators for Monitoring Progress to 2020 Targets. *Conservation Letters*, *9*(6), 489–494. https://doi.org/10.1111/conl.12329

McRae, L., Deinet, S., & Freeman, R. (2016). Data from: The diversity-weighted Living Planet Index: controlling for taxonomic bias in a global biodiversity indicator. In *Drya, Dataset*.

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0169156

Murrell, D. J. (2018). A global envelope test to detect non-random bursts of trait evolution. *Methods in Ecology and Evolution*, *9*(7), 1739–1748. https://doi.org/10.1111/2041-210X.13006

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., & Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *79*(2), 381–404. https://doi.org/10.1111/rssb.12172

Nicholson, E., Collen, B., Barausse, A., Blanchard, J. L., Costelloe, B. T., Sullivan, K. M. E., Underwood, F. M., Burn, R. W., Fritz, S., Jones, J. P. G., McRae, L., Possingham, H. P., & Milner-Gulland, E. J. (2012). Making robust policy decisions using global biodiversity indicators. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041128

Pannekoek, J., & van Strien, A. (2005). *TRIM 3 Manual (TRends & Indices for Monitoring data)*.

Rochette, A. J., Akpona, J. D. T., Akpona, H. A., Akouehou, G. S., Kwezi, B. M., Djagoun, C. A. M. S., Habonimana, B., Idohou, R., Legba, I. S., Nzigidahera, B. T., Matilo, A. O., Taleb, M. S., Bamoninga, B. T., Ivory, S., de Bisthoven, L. J., & Vanhove, M. P. M. (2019). Developing policy-relevant biodiversity indicators: Lessons learnt from case studies in Africa. *Environmental Research Letters*, *14*(3). https://doi.org/10.1088/1748-9326/aaf495

Rowland, J. A., Bland, L. M., James, S., & Nicholson, E. (2021). A guide to representing variability and uncertainty in biodiversity indicators. *Conservation Biology*, *35*(5), 1669–1682. https://doi.org/10.1111/cobi.13699

Scheele, B. C., Legge, S., Blanchard, W., Garnett, S., Geyle, H., Gillespie, G., Harrison, P., Lindenmayer, D., Lintermans, M., Robinson, N., & Woinarski, J. (2019). Continental-scale assessment reveals inadequate monitoring for threatened vertebrates in a megadiverse country. *Biological Conservation*, *235*, 273–278. https://doi.org/10.1016/j.biocon.2019.04.023

Soldaat, L. L., Pannekoek, J., Verweij, R. J. T., van Turnhout, C. A. M., & van Strien, A. J. (2017). A Monte Carlo method to account for sampling error in multi-species indicators. *Ecological Indicators*, *81*, 340–347. https://doi.org/10.1016/j.ecolind.2017.05.033

van Strien, A. J., Pannekoek, J., & Gibbons, D. W. (2001). Indexing european bird population trends using results of national monitoring schemes: A trial of a new method. *Bird Study*, *48*(2), 200–213. https://doi.org/10.1080/00063650109461219

Viljugrein, H., Stenseth, N. C., Smith, G. W., & Steinbakk, G. H. (2005). Density dependence in North American ducks. *Ecology*, *86*(1), 245–254.

Wang, G., Thompson Hobbs, N., Boone, R. B., Illius, A. W., Gordon, I. J., Gross, J. E., & Hamlin, K. L. (2006). Spatial and temporal variability modify density dependence in populations of large herbivores. *Ecology*, *87*(1), 95–102.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5

Watermeyer, K. E., Guillera-Arroita, G., Bal, P., Burgass, M. J., Bland, L. M., Collen, B., Hallam, C., Kelly, L. T., McCarthy, M. A., Regan, T. J., Stevenson, S., Wintle, B. A., & Nicholson, E. (2021). Using decision science to evaluate global biodiversity indices. *Conservation Biology*, *35*(2), 492–501. https://doi.org/10.1111/cobi.13574

Westcott, D. A., Fletcher, C. S., Mckeown, A., & Murphy, H. T. (2012). Assessment of monitoring power for highly mobile vertebrates. *Ecological Applications*, *22*(1), 374–383.

Wotton, S. R., Eaton, M. A., Sheehan, D., Munyekenye, F. B., Burfield, I. J., Butchart, S. H. M., Moleofi, K., Nalwanga-Wabwire, D., Ndang'ang'a, P. K., Pomeroy, D., Senyatso, K. J., & Gregory, R. D. (2020). Developing biodiversity indicators for african birds. *ORYX*, *54*(1), 62–73. https://doi.org/10.1017/S0030605317001181

Zylstra, E. R., Steidl, R. J., & Swann, D. E. (2010). Evaluating Survey Methods for Monitoring a Rare Vertebrate, the Sonoran Desert Tortoise. *Journal of Wildlife Management*, *74*(6), 1311–1318. https://doi.org/10.2193/2009-331

Chapter 5: Discussion and Synthesis

Global biodiversity is facing a crisis, which must be solved through effective policies and on-the-ground conservation. But governments, NGOs, and scientists need reliable indicators to guide research, conservation actions, and policy decisions. Developing reliable indicators is challenging because the data underlying those tools is incomplete and biased. For example, the Living Planet Index (LPI) tracks the changing status of global vertebrate biodiversity, but gaps, biases and quality issues plague the aggregated data used to calculate trends. Large-scale scientific surveys to overcome data deficiency are important but are an expensive and long-term solution. In the meantime, new methods are needed to determine the limitations of existing data, target data-gathering efforts, and maximize the reliability and robustness of indicators.

In this thesis, I presented a set of methods to quantify the reliability and robustness of LPI trends and their confidence intervals, and provided quantitative recommendations for targeted data-gathering efforts that would allow all LPI trends to meet a reliability standard. First, I explored the properties of distance measures in relation to comparing time series and trends in ecology, and developed a framework for choosing an appropriate measure for any time-series comparison task (Chapter 2). Importantly, I demonstrated that distance measures can be used to directly compare not only stochastic time series, but smoothed trends. I then developed a model of trend reliability, using simulated datasets as stand-ins for the real world, degraded samples as stand-ins for LPD datasets, and the Jaccard distance, chosen using the described framework, to quantify reliability (Chapter 3). The model revealed that many trends in the LPI are too data-poor to be considered reliable, particularly across the global south. I set a trend reliability standard and showed which regions and taxa are most in need of additional data, as well as how much data is needed to bring each regional taxonomic group to meet a reliability standard. A key result was that taxonomic representativeness as measured by McRae et al. (2017) is not always indicative of trend reliability; a regional taxonomic group can be data deficient despite having a comparatively high percentage of species represented in the Living Planet Database (LPD). My results showed that the number and quality of time series is more important than the number of species. Finally, I demonstrated a method to account for sampling and measurement error

in the confidence intervals of LPI trends that resulted in more accurate and robust confidence intervals (Chapter 4). The results of Chapter 4 revealed that current confidence intervals in the LPI are larger than necessary and therefore inefficient, likely due to high levels of measurement and/or sampling error; the GRRE (GAM-resampled rank envelope) method can account for measurement error and multiple levels of sampling error while reducing the width of confidence intervals by smoothing away outlier growth rates.

In this chapter, I will discuss my work in a broader context. I will explain the importance of my research on distance measures, highlight the strengths and weaknesses of my modelling approach, and describe its usefulness in indicator research. I will outline ongoing issues, limitations, complications, and repercussions related to my research and the LPI. Finally, I will discuss how the work I present in this thesis can move the field of biodiversity indicator research forward, and give my perspective on future research directions.

5.1. Distance measures

Time series are not only the basic components of abundance-based biodiversity indicators but are ubiquitous in ecology and conservation science to track changes in populations and environments over time. Often time series need to be compared, e.g., to detect anomalies, or for classification or clustering tasks, and distance measures are common and highly diverse tools for such tasks. Hundreds of distance measures have been described in the literature, but guidance on selection is largely limited to comparisons of mean classification accuracy (e.g., Bagnall et al., 2017; Paparrizos et al., 2020; Pree et al., 2014; Wang et al., 2013), without accounting for differences between datasets or tasks. Properties-based selection has been studied for certain tasks (Kocher & Savoy, 2017) or data types (Lhermitte et al., 2011). Mori et al. (2016) provided an automated selection process for clustering based on quantifiable properties of datasets; however, they ignored task-based differences and included only five distance measures. The research I presented in Chapter 2 is novel in two important ways. First, it provides a selection process generalized to cover any task or dataset; it covers 42 distance measures and 16 properties, and because the process is user-directed and the code is open-source, it can be expanded. The work goes far beyond any previous studies to fill an important gap in the literature. Although the selection process can be utilized by a general audience, it is tailored to ecologists, providing contextual examples

to guide them in the use of tools for which most existing literature is highly technical and aimed at computer or information scientists. The second novel aspect of my work in Chapter 2 is that I used distance measures to compare non-linear trends. I showed that distance measures can match the results of other comparison methods, both for stochastic time series and for smoothed trends. But the advantage of using distance measures is not that they can emulate the results of other methods, but that they can solve problems those other methods cannot. Distance measures are highly varied and compare different aspects of time series, but generally include some level of temporal information. The number and variety of distance measures described in the literature means that there should be an appropriate measure available for most comparison tasks. However, since many are only described within computer or information science literature, and very little guidance has been provided on appropriate usage or selection, many scientists are unaware that distance measures may be useful for a particular task, and have no idea how to choose one that is appropriate. The work I've done in this thesis not only introduces many distance measures to ecologists but provides context and a framework for selecting one that is fit for purpose. This should increase the scope for the application of distance measures in ecology. Furthermore, I introduced distance measures as a way of comparing trends, a task for which I am not aware of any previous application, and which provides the foundation the rest of my thesis rests upon. Therefore, the method I presented in Chapter 2 expands the possibilities for ecological research and is fundamental to the modelling approach I developed in this thesis.

5.2. Modelling approach

The modelling approach I developed greatly improves our ability to evaluate biodiversity indicators. Simulation models are used for indicator testing when comprehensive data are not available because models can provide a known 'truth' as a basis for comparison (Rowland et al., 2020). Typically, these models are parameterized to represent specific real-world datasets and/or scenarios to test indicator responses (e.g., Halouani et al., 2019; Nicholson et al., 2012; Rowland et al., 2020). I applied a different approach in this thesis; instead of modelling specific scenarios or datasets, I built a generalized and highly flexible model to represent regional taxonomic groups in the Living Planet Database (LPD). My

approach sacrificed ecological accuracy for flexibility, enabling me to test the LPI across a wide range and combination of parameter settings. I then used the results to create something novel: a predictive of model of trend reliability. While other studies are concerned with indicator responses to specific modelled scenarios or datasets, often for policy reasons, my generalized modelling approach allowed me to predict the reliability of all existing indicator trends, as well as modelled scenarios and datasets. My approach can be used to test whether new methods or tools impart actual improvements in trend accuracy. It can also be adapted for other indicators using the modular open-source code I created. I have already adapted part of the code for an online-accessible, real-time version of the Species Awareness Index (SAI) (online version not yet published; for the offline SAI, see Millard et al., 2021). There are two key innovations to my approach that have the potential to change the way that biodiversity indicators are tested. The first is the ability to evaluate indicators based not just on *how* they respond to simulations, but on how *accurately* they respond. The second is the ability to *predict* accuracy (referred to as reliability because it is predicted rather than measured) for unsimulated data.

There are three important caveats to my approach. First, while I was able to simulate the parameters of LPD datasets, I did not accurately simulate individual time series. I developed a stochastic exponential time series model, but I did not model carrying capacity; exponential growth of individual populations is limited only by stochasticity and the growth rate parameters (mean and standard deviation) assigned to them. Second, the model implicitly assumes closed populations. While stochasticity could be assumed to include some element of migration, it is not modelled as such, and therefore there is no intraspecies dependence in size fluctuations across populations. Third, growth rates are not serially correlated; they are randomly distributed across time series (in the real world, temporal variations in the environment caused lagged responses in life history traits, leading to serially correlated growth rates - Tuljapurkar et al., 2009). The same is not true of abundances (modelled as index values), which are serially correlated because they are calculated from growth rates using the cumulative product function. The lack of carrying capacity, migration, and serial correlation of growth rates were necessary trade-offs, as increasing the complexity of the time series model reduces control over the parameters of

the dataset. However, it could be a valuable next step to test the performance of the reliability model using a more complex time series model that can simulate real-world taxa.

A lot of biodiversity indicator research is designed to achieve specific policy-related goals, which can result in simplified methods but also bias the results and limit their application. The sampled approach to the Red List Index (sRLI) was designed to achieve a sample size large enough to ensure that extinction risk trends did not appear positive when they were in fact negative (Baillie et al., 2008). This approach assumes that it is more important to detect increased extinction risk than reduced extinction risk, which is a biased value judgement. Dr. Brian Leung, who along with his colleagues published a Nature paper (Leung et al., 2020) that has caused considerable controversy around the LPI, recently presented a method to determine the number of populations needed to detect positive change in specific LPI regional taxonomic group trends (Leung, 2022). While the idea has some similarities to the work I have presented in this thesis, the goal is aimed at satisfying a specific policy goal, namely tracking the reversal of biodiversity declines for 2030 CBD (Convention on Biological Diversity) milestones. Just like the sRLI, this results in biased methodology, with negative change not considered. While it is important to support policy goals aimed at reversing biodiversity loss, the goal of the research I presented in this thesis is to improve our overall understanding of biodiversity change. That results in my approach being necessarily more general and more complex, but the results will be more widely applicable, potentially leading to an increase in the relevance of the LPI and other similar indicators for a wide range of policy goals. Analysis targeted to specific policy goals tends to be overly specific and provides results with limited predictive power. The work of Baillie et al. (2008) was based on only two taxonomic groups and ignored the underlying factors that led to their results, so they were only able to offer a general prescription to be applied to all taxonomic groups. Likewise, the work of Leung (2022) is currently based on ten regional taxonomic groups that have suffered catastrophic declines and seems to offer a generalized prescription only for those groups (the work is unfinished, so this may change). By contrast, my modelling approach takes into account the underlying factors that lead to particular results, and is based on thousands of synthetic datasets that cover a wide range of parameter values; it is therefore able to make tailored predictions and prescriptions for each regional taxonomic group in the LPI, and can be easily applied to similar indicators.

5.3. Populations vs species

My reliability model revealed that populations are more relevant than species as sampling units in the LPD; however, this revelation is potentially problematic because populations are poorly defined in the LPD. It is important to be careful about making inferences from a model before considering the real-world context. Many biodiversity indicators are localized and/or specialized to single taxa or even a group of species within a taxon, such as European grassland butterflies (Van Swaay et al., 2019; Van Swaay & van Strien, 2005) or UK farmland birds (Freeman et al., 2001; Gregory et al., 2004). These indicators use carefully chosen species with comprehensive survey data that typically includes yearly transect counts or abundance estimates across multiple sites (Gregory et al., 2019; Soldaat et al., 2017; Van Swaay et al., 2019). Global biodiversity indicators cannot rely on the availability of such comprehensive data because it is unavailable for many taxa and geographical areas (Collen et al., 2009; Gregory et al., 2019; McRae et al., 2017; Soldaat et al., 2017). Species trends in the LPI are constructed from time series units referred to as populations, which may be aggregated from different countries or regions, different studies with different methodologies, and even different units of measurement. A population might be represented by transect-sampled individuals from a precisely defined 250 m^2 location (e.g., Cole et al., 2014) or the biomass caught by the marine fishing fleets of an entire country (e.g., Carvalho et al., 2014). Yet populations are equally weighted when averaged to form species trends; this is by necessity, as even though there is no standardization between studies in the database, the calculation of the index must be standardized, else it would not be feasible to aggregate thousands of time series into a single index. Diversity weighting was introduced to the LPI to account for biases in representativeness but is based only on the geographic and taxonomic distributions of species richness. My results suggest that species representativeness may not be a good indicator of taxonomic or geographical representativeness since the number of populations is more important than the number of species. This is because mean growth rates in the LPD vary almost as widely between populations within the same species as they do between populations belonging to different species. Species are traditionally considered as discrete fundamental units of both taxonomy and biodiversity (Hey & Pinho, 2012), with associated life history traits, genetic and phenotypic profiles, and ecological niches, while populations (often referred to as

subpopulations) can be thought of as subunits sharing the traits and profiles of the species they belong to but confined to a more limited geographical area. However, as different environments can result in heritable epigenetic and phenotypic differences between populations belonging to the same species (Bossdorf et al., 2008), it may be reasonable to think of populations as ecological units. This ecological meaning may be partially undermined by the ill-defined nature of populations, but the concept and delimitation of species are also major sources of confusion and disagreement between biologists and across disciplines (de Queiroz, 2007; Tobias et al., 2010). There is no clearly defined boundary between species and populations; they can be said to sit at different ends of a continuum (de Queiroz, 2007), with species having slightly more evolutionary independence due to a lower distribution of migration rates and higher distribution of separation times compared to populations, but with a lot of overlap (Hey & Pinho, 2012). Nonetheless, the number of populations is problematic as a measure of representativeness because there are no estimates of total numbers of populations for comparison. Nor is the number of populations consistent across species; rare species may exist as a single population, while abundant or widely distributed species may consist of hundreds or thousands of populations, and some species may have greater variance between population trends than others. Population trends may also depend on the location of the population within the species distribution range, with more extreme or fluctuating trends likely to be found at the margins (e.g., European terrestrial breeding bird growth rates were found to be lowest at species thermal maxima and highest at thermal minima — Jiguet et al., 2010). My modelling approach defined sample size in terms of populations, and showed it to be the most important factor in determining the reliability of LPI trends as well as their associated confidence intervals, while the number of species was not relevant. If populations were selected at random for research, that might be true. But, as selection is never truly random due to intentional and unintentional research biases, species representativeness should not be ignored when adding populations to the LPD to meet the trend reliability threshold.

In the context of my results, it may be worth considering whether the use of species trends as an intermediate step in calculation of the LPI remains relevant. Currently, confidence intervals in the LPI are generated by resampling species rates of change. However, I showed that it is more robust and accurate to resample from GAMs of the population trends. While

the GRRE method still employs resampling of the species trends to account for sampling error in species selection, Soldaat et al. (2017) suggested that this may be unjustified because it is based on the idea that species are randomly sampled when in fact they are deliberately selected. This is only partially true for aggregated abundance indicators; however, if we are to consider species as irrelevant for the purpose of meeting the trend reliability threshold (which I argue above should not be done, contrary to the results of my own research), then there is no reason to account for sampling error in a unit we have assigned no meaning to. Another reason for calculating species trends as an intermediate step in the LPI is to reduce the influence of species which are overrepresented (many populations present from the same species) in the LPD. But if we consider the number of species to be irrelevant, then overrepresentation should not be a concern. Therefore, it would be an interesting avenue of further research to rebuild the model without calculating species indices and compare the results. What would change? Is it possible that the number of species is only irrelevant to the model *because* species indices are preventing overrepresentation, and thus would become relevant if their indices were *not* calculated?

5.4. Challenges, remaining questions, and the future

In Chapter 3, I showed that due to an uneven distribution of time series data by year in the LPD, and a steep drop in the number of observations from the late 2000s until the present due to a lag in acquiring new data, the LPI may not reflect a successful reversal of biodiversity loss until a decade or more after it occurs. Although my analysis assigned a single reliability rating to each trend in the LPI, the accuracy of a trend is likely to vary over its length, with the first and final parts of each trend being the least reliable because that is where the fewest observations occur. Inaccuracy at the beginning of a trend leads to larger disparities in final index values and therefore a poorer picture of how population abundances in the present compare to abundances in 1970. Inaccuracy at the end of a trend has a smaller effect on index values but is more concerning from a policy perspective because it hinders our ability to track and react appropriately to biodiversity change, to evaluate whether targets have been met, and to set effective future targets. There is no easy fix for this problem, and it is not confined to the LPI, but it is important to be aware of the issue so that it can be accounted for in scientific and policy discussions. Confidence

intervals provide an indication of uncertainty in index values, and my work contributes to making them more robust and reliable. However, two important issues remain. First, confidence intervals indicate uncertainty surrounding index values, but do not give a clear indication of the level of uncertainty in the direction of a trend at any given year along the index. Second, my work in this thesis considered confidence intervals only at the regional taxonomic level, but LPI trends are further aggregated to realm, system, and global levels. While there is no sampling error to account for at these levels, weighting is applied before averaging at each level, and there is a large but unknown amount of error inherent in the weighting system. It might be possible to account for uncertainty in estimates of species richness, but species richness is a poor indicator of representativeness and therefore the true uncertainty in the weightings is much higher than the uncertainty in richness estimates would indicate. That means each successive level of aggregation results in additional uncertainty that the LPI confidence intervals do not account for.

Such weaknesses are inherent to the production of global biodiversity indicators. The Red List Index is a global, standardized, and comprehensive indicator of extinction risk. It gets around some of the weaknesses of the LPI by sacrificing precision. Extinction risk is categorical, therefore expert assessment can fill in when quantitative data are lacking. Trends are linear as they are assessed at multi-year intervals rather than yearly. Species, not populations or sites, are the base units. And because assessments are conducted for all species at once, trends can only be calculated to the most recent assessment year. Amphibians, for example, have not been assessed since 2004, so there is no indication of the extinction risk trend for amphibians over the last 18 years. There is no perfect method of tracking global biodiversity change, and every index has its strengths and weaknesses. The important thing is to keep probing and assessing those weaknesses and finding ways to minimize them. My thesis has extensively discussed weaknesses in the Living Planet Index and the state of biodiversity knowledge, and suggested ways to reduce those weaknesses. However, I chose to focus on the LPI because it is one of the most comprehensive and useful global biodiversity indicators; its 25-year development history (Ledger et al., 2022) attests to its continued relevance. The LPI is built on the foundation of data collected by thousands of researchers over a period of more than 70 years, and the extensive research and effort that

has gone (and continues to go) into developing and maintaining the LPI is what made this thesis possible.

Scientists, policymakers, and conservations need up-to-date information on biodiversity change to make objective decisions; therefore, biodiversity indicators should respond quickly to changes in biodiversity trends. Stevenson et al. (2021) classified biodiversity indicators as leading (changes before the fact; predictive), coincident (measures the target variable), or lagging (changes after the fact) for three target variables related to global species extinctions: changes in abundance (population declines), changes in distribution (population-level extinctions), and species extinctions. The LPI, along with other abundance-based indicators, was classified as a leading indicator for distribution change and extinction, and a coincident indicator of abundance change. But in light of my third-chapter analysis showing that the LPI is unlikely to reliably indicate changes in biodiversity trends until a decade after the fact, the LPI should be considered a lagging indicator of abundance change. Indicators like the LPI, that compile information from many sources, are essential to bridging the synthesis gap, a gap between science and policy through which policy-relevant scientific information is often lost due to the time and effort required to synthesize it into a form that can be communicated to policymakers (Westgate et al., 2018). However, the lagging response of the LPI means that there is still a large time gap between the research and its communication to policymakers. This lagging response is largely due to the time involved in getting new monitoring data into the LPD. There are at least three components involved in the lag. First, it takes time for new research to be published, especially if it involves multi-year sampling, which all studies in the LPD necessarily do. Second, once published, studies must be found, and their data extracted, compiled, and added to the LPD. Third, LPI trends must be recalculated and published, which currently happens every two years. All three of these components can be reduced to some extent. Publication times can be reduced by making new studies available on open-access archives such as bioRxiv. Finding and extracting data from new studies may be able to be automated or semi-automated (Cornford et al., 2022; Millard et al., 2020; Westgate et al., 2018), although the technology still has teething issues (Cornford et al., 2022), or the work involved could be reduced by the researchers themselves offering data directly (although it would still need to be reviewed by a human). The website for the LPI already contains a way for scientists to contribute data,

but many are likely unaware. Finally, the two-year time gap between LPI publications could be eliminated by calculating trends in real time through a web application, although there may be labour, resource, and security-related hurdles to implementation.

In the future, it may be possible to produce a true real-time global biodiversity indicator, with monitoring data automatically added directly to a time-series database as it occurs, and trends continuously updated in real-time to reflect the new data. Monitoring technology is improving; cheap, low-maintenance camera trap networks can be rapidly deployed (Blount et al., 2021; Chianucci et al., 2021), machine learning algorithms can identify some taxa from camera traps (Tabak et al., 2019), recorded sounds (Darras et al., 2019; Sugai et al., 2019), or satellite images (Duporge et al., 2021); drones can cover territory quickly, spotting animals from above (Edney & Wood, 2021; Petso et al., 2022); eDNA can be used to detect the presence of marine or freshwater species, even in subterranean locations, without the need for visual or audial detection or invasive sampling (Blattner et al., 2021; Bonfil et al., 2021; Pukk et al., 2021; Saccò et al., 2022); citizen scientists can identify species and upload images, GPS coordinates, and other information in real time (Echeverria et al., 2021). However, there are hurdles that need to be overcome first. Machine learning models can only detect species included in their training data, and tend to be poor at generalizing out-of-sample data, so manual verification is generally still used (Wäldchen & Mäder, 2018; Whytock et al., 2021); however, some scientists feel that manual verification is unnecessary if machine learning models are used with care (Whytock et al., 2021). Most available automated or citizen-science monitoring methods have strong detection biases towards charismatic animals, large animals, animals that move in the open, or highly vocal animals. Small, quiet animals, including many reptiles, remain difficult to detect without active searching by experts, so will likely be undercounted or ignored, leading to bias in the indicator. Insects are even harder to detect, given their size, vast numbers, and diversity, and how many remain unidentified. Microbial biodiversity is virtually unknown, with up to 99.999% of species still undiscovered (Locey & Lennon, 2016). Many of these hurdles can eventually be overcome, leading to the production of comprehensive global biodiversity indices that responds instantly to changes in biodiversity; but it will take time, resources, and effort to get there.

5.5. Conclusions

Hortal et al. (2015) conducted a review of shortfalls related to large-scale biodiversity knowledge. They concluded that it is important to assess the extent, quality, and representativeness of biodiversity data to identify knowledge gaps and biases and direct future research. They also stated that it is important to find new ways to represent and account for uncertainty to improve the robustness of conclusions obtained from biodiversity research. This thesis not only reveals weaknesses in our understanding of global vertebrate biodiversity, but quantifies the amount of data needed to shore up each weakness, and describes new methods to improve our global picture of vertebrate biodiversity using existing data. The second chapter presents an objective way of choosing an appropriate distance measure for any time-series comparison task by comparing how 42 different distance measures react to common differences between time series. It provides the foundation for using distance measures as a tool for testing the accuracy and reliability of biodiversity indicator trends. The third chapter quantitatively highlights weaknesses in the Living Planet Index by measuring the reliability of each regional taxonomic group trend against a target rating and calculating the number of time series needed to meet the target. It then describes and compares potential solutions to acquire the needed data. The fourth chapter improves our understanding of biodiversity trends and their associated uncertainty by defining confidence intervals in a more robust, efficient, and accurate way.

This thesis is primarily methodological; the work will form part of the foundational toolset that other scientists can use to aid their own research. My work on distance measures can lead to new ways of conducting ecological research, while my modelling framework will allow new methods, solutions, or other potential improvements for abundance-based biodiversity indicators to be tested before implementation. While a proposed change to the way an indicator is calculated might theoretically improve trend reliability or robustness, my modelling framework can be applied to quantitatively measure the effect on synthetic data. Furthermore, the methods I've presented here can be expanded upon to increase their utility. Additional distance measures can be tested for relevant properties. Time series can be simulated to emulate specific taxonomic groups to improve the model's precision for

those groups. Thus, the potential for my thesis work to move forward the fields of ecology and conservation biology is greater than the sum of its parts.

The most important issue that my work highlights is that reliability will remain limited by the availability of data. If there is not enough existing data to reach the reliability threshold, then more studies must be conducted, and it could take many years before time series from those studies contribute positively to trend reliability. By then, computational methods such as machine learning may reduce the data needed to reach threshold levels. Additionally, mass data gathering efforts using satellites, drones, camera traps, listening devices, and eDNA, combined with artificial intelligence or advanced machine learning methods, may allow for real-time indicators that present a more sensitive picture of the changing state of biodiversity. Regardless of current limitations of biodiversity indicators, we cannot wait around for things to improve. While research continues, important policy and conservation decisions must be made based on the limited and incomplete picture we have. Biodiversity loss will not wait for science to catch up.

5.6. Bibliography

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Baillie, J. E. M., Collen, B., Amin, R., Akcakaya, H. R., Butchart, S. H. M., Brummitt, N., Meagher, T. R., Ram, M., Hilton-Taylor, C., & Mace, G. M. (2008). Toward monitoring global biodiversity. *Conservation Letters*, *1*(1), 18–26. https://doi.org/10.1111/j.1755-263x.2008.00009.x

Blattner, L., Ebner, J. N., Zopfi, J., & von Fumetti, S. (2021). Targeted non-invasive bioindicator species detection in eDNA water samples to assess and monitor the integrity of vulnerable alpine freshwater environments. *Ecological Indicators*, *129*. https://doi.org/10.1016/j.ecolind.2021.107916

Blount, J. D., Chynoweth, M. W., Green, A. M., & Şekercioğlu, Ç. H. (2021). Review: COVID-19 highlights the importance of camera traps for wildlife conservation research and

management. *Biological Conservation*, *256*. https://doi.org/10.1016/j.biocon.2021.108984

Bonfil, R., Palacios-Barreto, P., Mendoza Vargas, O. U., Ricaño-Soriano, M., & Píndaro Díaz-Jaimes, ·. (2021). Detection of critically endangered marine species with dwindling populations in the wild using eDNA gives hope for sawfishes. *Marine Biology*, *168*(60). https://doi.org/10.1007/s00227-021-03862-7

Bossdorf, O., Richards, C. L., & Pigliucci, M. (2008). Epigenetics for ecologists. *Ecology Letters*, *11*(2), 106–115. https://doi.org/10.1111/j.1461-0248.2007.01130.x

Carvalho, F., Ahrens, R., Murie, D., Ponciano, J. M., Aires-da-Silva, A., Maunder, M. N., & Hazin, F. (2014). Incorporating specific change points in catchability in fisheries stock assessment models: An alternative approach applied to the blue shark (Prionace glauca) stock in the south Atlantic Ocean. *Fisheries Research*, *154*, 135–146. https://doi.org/10.1016/j.fishres.2014.01.022

Chianucci, F., Bajocco, S., & Ferrara, C. (2021). Continuous observations of forest canopy structure using low-cost digital camera traps. *Agricultural and Forest Meteorology*, *307*. https://doi.org/10.1016/j.agrformet.2021.108516

Cole, E. M., Bustamante, M. R., Almeida-Reinoso, D., & Funk, W. C. (2014). Spatial and temporal variation in population dynamics of Andean frogs: Effects of forest disturbance and evidence for declines. *Global Ecology and Conservation*, *1*, 60–70. https://doi.org/10.1016/j.gecco.2014.06.002

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring Change in Vertebrate Abundance: the Living Planet Index. *Conservation Biology*, *23*(2), 317–327. https://doi.org/10.1111/j.1523-1739.2008.01117.x

Cornford, R., Millard, J., González-Suárez, M., Freeman, R., & Johnson, T. F. (2022). Automated synthesis of biodiversity knowledge requires better tools and standardised research output. *Ecography*, *2022*(3). https://doi.org/10.1111/ecog.06068

Darras, K., Eter, P., Ary, B., Furnas, B. J., Grass, I., Mulyani, Y. A., & Tscharntke, T. (2019). *Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide*. https://doi.org/10.1002/eap.1954

de Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, *56*(6), 879–886. https://doi.org/10.1080/10635150701701083

Duporge, I., Isupova, O., Reece, S., Macdonald, D. W., Wang, T., & Buchanan, G. (2021). Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, *7*(3), 369–381. https://doi.org/10.1002/rse2.195

Echeverria, A., Ariz, I., Moreno, J., Peralta, J., & Gonzalez, E. M. (2021). Learning Plant Biodiversity in Nature: The Use of the Citizen-Science Platform iNaturalist as a Collaborative Tool in Secondary Education. *Sustainability*, *13*, 735. https://doi.org/10.3390/su13020735

Edney, A. J., & Wood, M. J. (2021). Applications of digital imaging and analysis in seabird monitoring and research. *Ibis*, *163*, 317–337. https://doi.org/10.1111/ibi.12871

Freeman, S. N., Baillie, S. R., & Gregory, R. D. (2001). *Statistical analysis of an indicator of population trends in farmland birds*. British trust for Ornithology.

Gregory, R. D., Noble, D. G., & Custance, J. (2004). The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. *Ibis*, *146*(Suppl. 2), 1–13.

Gregory, R. D., Skorpilova, J., Vorisek, P., & Butler, S. (2019). An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe. *Ecological Indicators*, *103*, 676–687. https://doi.org/10.1016/j.ecolind.2019.04.064

Halouani, G., le Loc'h, F., Shin, Y. J., Velez, L., Hattab, T., Romdhane, M. S., & ben Rais Lasram, F. (2019). An end-to-end model to evaluate the sensitivity of ecosystem indicators to track fishing impacts. *Ecological Indicators*, *98*, 121–130. https://doi.org/10.1016/j.ecolind.2018.10.061

Hey, J., & Pinho, C. (2012). Population genetics and objectivity in species diagnosis. *Evolution*, *66*(5), 1413–1429. https://doi.org/10.1111/j.1558-5646.2011.01542.x

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Jiguet, F., Devictor, V., Ottvall, R., van Turnhout, C., van der Jeugd, H., & Lindström, Å. ˚. (2010). Bird population trends are linearly affected by climate change along species

thermal ranges. *Proceedings of the Royal Society B*, *277*(1700), 3601–3608. https://doi.org/10.1098/rspb.2010.0796

Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. *Information Processing and Management*, *53*(5), 1103–1119. https://doi.org/10.1016/j.ipm.2017.04.004

Ledger, S. E. H., McRae, L., Loh, J., Almond, R., Böhm, M., Currie, J., Deinet, S., Galewski, T., Grooten, M., Jenkins, M., Marconi, V., Painter, B., Scott-Gatty, K., Young, L., & Hoffmann, M. (2022). Past, present, and future of the Living Planet Index. *BioRxiv*. https://doi.org/10.1101/2022.06.20.496803

Leung, B. (2022). Risk, biodiversity, power analysis, and sampling needs for global biodiversity monitoring. *[Conference Abstract.] Ecological Society of America 2022 Annual Meeting. Montréal, Québec, Canada.*

Leung, B., Hargreaves, A. L., Greenberg, D. A., McGill, B., Dornelas, M., & Freeman, R. (2020). Clustered versus catastrophic global vertebrate declines. *Nature*, *588*(7837), 267–271. https://doi.org/10.1038/s41586-020-2920-6

Lhermitte, S., Verbesselt, J., Verstraeten, W. W., & Coppin, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, *115*(12), 3129–3152. https://doi.org/10.1016/j.rse.2011.06.020

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(21), 5970–5975. https://doi.org/10.1073/pnas.1521291113

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0169156

Millard, J. W., Freeman, R., & Newbold, T. (2020). Text-analysis reveals taxonomic and geographic disparities in animal pollination literature. *Ecography*, *43*(1), 44–59. https://doi.org/10.1111/ecog.04532

Millard, J. W., Gregory, R. D., Jones, K. E., & Freeman, R. (2021). The species awareness index as a conservation culturomics metric for public biodiversity awareness. *Conservation Biology*, *35*(2), 472–482. https://doi.org/10.1111/cobi.13701

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Similarity Measure Selection for Clustering Time Series Databases. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 181–195. https://doi.org/10.1109/TKDE.2015.2462369

Nicholson, E., Collen, B., Barausse, A., Blanchard, J. L., Costelloe, B. T., Sullivan, K. M. E., Underwood, F. M., Burn, R. W., Fritz, S., Jones, J. P. G., McRae, L., Possingham, H. P., & Milner-Gulland, E. J. (2012). Making robust policy decisions using global biodiversity indicators. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041128

Paparrizos, J., Liu, C., Elmore, A. J., & Franklin, M. J. (2020). Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1887–1905. https://doi.org/10.1145/3318464.3389760

Petso, T., Jamisola Jr., R. S., & Mpoeleng, D. (2022). Review on methods used for wildlife species and individual identification. *European Journal of Wildlife Research*, *68*(3), 2–18. https://doi.org/10.1007/s10344-021-01549-4

Pree, H., Herwig, B., Gruber, T., Sick, B., David, K., & Lukowicz, P. (2014). On general purpose time series similarity measures and their use as kernel functions in support vector machines. *Information Sciences*, *281*, 478–495. https://doi.org/10.1016/j.ins.2014.05.025

Pukk, L., Kanefsky, J., Heathman, A. L., Weise, E. M., Nathan, L. R., Herbst, S. J., Sard, N. M., Scribner, K. T., & Robinson, J. D. (2021). eDNA metabarcoding in lakes to quantify influences of landscape features and human activity on aquatic invasive species prevalence and fish community diversity. *Diversity and Distributions*, *27*, 2016–2031. https://doi.org/10.1111/ddi.13370

Rowland, J. A., Lee, C. K. F., Bland, L. M., & Nicholson, E. (2020). Testing the performance of ecosystem indices for biodiversity monitoring. *Ecological Indicators*, *116*. https://doi.org/10.1016/j.ecolind.2020.106453

Saccò, M., Guzik, M. T., van der Heyde, M., Nevill, P., Cooper, S. J. B., Austin, A. D., Coates, P. J., Allentoft, M. E., & White, N. E. (2022). eDNA in subterranean ecosystems: Applications, technical aspects, and future prospects. *Science of the Total Environment*, *820*. https://doi.org/10.1016/j.scitotenv.2022.153223

Soldaat, L. L., Pannekoek, J., Verweij, R. J. T., van Turnhout, C. A. M., & van Strien, A. J. (2017). A Monte Carlo method to account for sampling error in multi-species

indicators. *Ecological Indicators*, *81*, 340–347. https://doi.org/10.1016/j.ecolind.2017.05.033

Stevenson, S. L., Watermeyer, K., Caggiano, G., Fulton, E. A., Ferrier, S., & Nicholson, E. (2021). Matching biodiversity indicators to policy needs. *Conservation Biology*, *35*(2), 522–532. https://doi.org/10.1111/cobi.13575

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr., J. W., & Llusia, D. (2019). Overview Articles Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, *69*(1), 15–25. https://doi.org/10.1093/biosci/biy147

Tabak, M. A., Norouzzadeh, M. S., David, |, Wolfson, W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Salvo, P. A. di, Lewis, J. S., White, M. D., Teton, B., James, |, Beasley, C., Peter, |, Schlichting, E., Boughton, R. K., Wight, B., Newkirk, E. S., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods Ecol Evol*, *10*. https://doi.org/10.1111/2041-210X.13120

Tobias, J. A., Seddon, N., Spottiswoode, C. N., Pilgrim, J. D., Fishpool, L. D. C., & Collar, N. J. (2010). Quantitative criteria for species delimitation. *Ibis*, *152*, 724–746. https://doi.org/10.1111/j.1474-919X.2010.01051.x

Tuljapurkar, S., Gaillard, J.-M., & Coulson, T. (2009). From stochastic environments to life histories and back. *Philosophical Transactions of the Royal Society B*, *364*(1523), 1499–1509. https://doi.org/10.1098/rstb.2009.0021

van Swaay, C. A. M., Dennis, E. B., Schmucki, R., Sevilleja, C., Balalalaikins, M., Botham, M., Bourn, N., Brereton, T., Cancela, J. P., Carlisle, B., Chambers, P., Collins, S., Dopagne, C., Escobés, R., Feldmann, R., Fernández-García, J. M., Fontaine, B., Gracianteparaluceta, A., Harrower, C., … Roy, D. B. (2019). *The EU Butterfly Indicator for Grassland species: 1990-2017*. www.butterfly-monitoring.net

van Swaay, C., & van Strien, A. (2005). Using butterfly monitoring data to develop a European grassland butterfly indicator. In E. Kühn, R. Feldmann, J. A. Thomas, & J. Settele (Eds.), *Studies on the Ecology and Conservation of Butterflies in Europe Vol. 1: General Concepts and Case Studies* (pp. 106–108). Pensoft.

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*, 2216–2225. https://doi.org/10.1111/2041-210X.13075

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5

Westgate, M. J., Haddaway, N. R., Cheng, S. H., McIntosh, E. J., Marshall, C., & Lindenmayer, D. B. (2018). Software support for environmental evidence synthesis. *Nature Ecology and Evolution*, *2*(4), 588–590. https://doi.org/10.1038/s41559-018-0502-x

Whytock, R. C., Świeżewski, J., Zwerts, J. A., Bara-Słupski, T., Flore Koumba Pambo, A., Rogala, M., Bahaa-el-din, L., Boekee, K., Brittain, S., Cardoso, A. W., Henschel, P., Lehmann, D., Momboua, B., Kiebou Opepa, C., Orbell, C., Pitman, R. T., Robinson, H. S., Abernethy, K. A., Franke, A., & Alberto Silva, C. (2021). Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, *12*, 1080–1092. https://doi.org/10.1111/2041-210X.13576

Appendices

This section contains supplementary figures, tables, and text from the three research chapters of this thesis. Due to the number of supplementary figures and tables (approximately 150), an exhaustive list has not been included.

Table of Contents for Appendices

Appendix 1: Supplementary materials for Chapter 2 160

S2.1. Descriptions and formulas for selected distance measures 160

S2.2. Distance measure properties 170

S2.3. Uncontrolled testing 172

S2.4. Metric test results 174

S2.5. Controlled test results 175

S2.5.1. Sensitivity tests 175

S2.5.2. Time-based invariances and other tests 176

S2.5.3. Pairwise correlations between distance measures 177

S2.6. Uncontrolled test results 178

S2.7. Example dataset results 184

S2.8. Speeding up DTW 185

S2.9. Plots of controlled test results for all distance measures 186

S2.10. Tables of controlled test results for all distance measures 228

S2.11. Plots of wading bird rankings for all distance measures 271

S2.12. Bibliography for Appendix 1 313

Appendix 2: Supplementary materials for Chapter 3 315

Appendix 3: Supplementary materials for Chapter 4 321

Appendix 1: Supplementary materials for Chapter 2

S2.1. Descriptions and formulas for selected distance measures

The **Euclidean distance** (Euclidean), also known as the L^2 distance, is the straight-line distance between a pair of points. It also forms the basis for some of the more complicated transformation-based and model-based metrics presented here. It is defined as:

$$d_{Euc} = \sqrt{\sum_{i=1}^{d} | P_i - Q_i |^2} \quad (3)$$

where P and Q are (time series) vectors and *d* is the length of the vectors.

The **Manhattan distance** (Manhattan), or L^1 distance, is the shortest distance between two points on a grid. Because it is not based on Euclidean geometry, there can be multiple paths with the same shortest distance. It is defined as:

$$d_{CB} = \sum_{i=1}^{d} | P_i - Q_i | \quad (4)$$

The **Chebyshev distance** (Chebyshev), or L^∞ distance, is the greatest of the differences between two points or vectors along any coordinate dimension. For example, if two points had the x,y coordinates (0,0) and (3,5), the Chebyshev distance would be five, the difference between the y coordinates of the two points, as this is greater than three, the difference between the x coordinates. The Chebyshev distance is defined as:

$$d_{Cheb} = \max_i | P_i - Q_i | \quad (5)$$

The **Complexity-Invariant Distance** (CID) applies a complexity correction factor to the Euclidean distance to increase the dissimilarity value between time series with different complexities (where complexity is the length of a time series if stretched into a straight line—more and greater peaks and valleys means more complexity). It is defined as:

$$d_{CID}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = CF\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) \cdot d\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) \tag{6}$$

where d is the Euclidean distance and CF is a complexity correction factor

$$CF\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \frac{\max\left\{CE\left(\boldsymbol{X}_T\right), CE\left(\boldsymbol{Y}_T\right)\right\}}{\min\left\{CE\left(\boldsymbol{X}_T\right), CE\left(\boldsymbol{Y}_T\right)\right\}} \tag{7}$$

where CE is a complexity estimator

$$CE\left(\boldsymbol{X}_T\right) = \sqrt{\sum_{t=1}^{T-1}\left(X_t - X_{t+1}\right)^2} \tag{8}$$

The **Dynamic Time Warping distance** (DTW) computes a warping path between two time series to align them in time. It can be defined as a man-dog distance, but instead of the shortest leash length, it measures the average leash length. This makes it more robust than the Fréchet distance, as it is less sensitive to outliers and short divergences. The DTW distance is defined as:

$$d_{DTW}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \min_{r \in M}\left(\sum_{i=1,..,m}\left|X_{a_i} - Y_{b_i}\right|\right) \tag{9}$$

The **Time Alignment Measurement distance** (TAM) is a derivative of the DTW distance that measures how well two time series align in time. Segments not in phase are penalized, while amplitude differences are not. A dissimilarity value of zero can occur for non-identical series that are perfectly aligned in time.

The **Normalized Compression Distance** (NCD) is based on the concept of Kolmogorov complexity, which is the minimum information needed to generate a string using an algorithm. The Kolmogorov complexity is a measure of randomness of the string. The smaller the value, the less randomness. The NCD applies the concept to a relationship between objects (time series) when a compression algorithm is applied. The greater the advantage in compression (reduction in randomness) gained by multiplying two time series

together, the more closely they are related, and therefore the smaller the dissimilarity between them. NCD is defined as:

$$d_{NCD}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \frac{C\left(\boldsymbol{X}_T \boldsymbol{Y}_T\right) - \min\left\{C(\boldsymbol{X}_T), C(\boldsymbol{Y}_T)\right\}}{\max\left\{C(\boldsymbol{X}_T), C(\boldsymbol{Y}_T)\right\}} \quad (10)$$

where C represents the compressed size.

The **Compression-based Dissimilarity Measure** (CDM) is a simplified version of the NCD, defined as:

$$d_{CDM}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \frac{C\left(\boldsymbol{X}_T \boldsymbol{Y}_T\right)}{C(\boldsymbol{X}_T) C(\boldsymbol{Y}_T)} \quad (11)$$

The **Edit Distance with Real Penalty** (ERP) is an edit distance, meaning it quantifies the number of insert, delete, or replace operations required to turn one string (time series) into another. ERP includes a penalty for gaps between matched substrings based on the gap length.

The **Edit Distance for Real Sequences** (EDR) is an edit distance refined for trajectories. It includes a quantization feature as well as the length-based gap penalty of ERP.

The **Fourier Coefficient-based distance** (Fourier) calculates the Euclidean distance between Discrete Fourier Transforms of a pair of time series. Fourier transforms extract frequency information by decomposing a signal (time series) into its frequency components (sine and cosine functions). While a time series is visualized as a single graph of amplitude vs time, its Fourier Transform consists of multiple sinusoidal waves, each with a specific, constant amplitude and frequency. Time information is lost. The Fourier Transform works well for stationary time series, as they have periodic repeating signals. However, the loss of time information presents problems for deconstructing non-stationary series, as they change randomly over time.

The **Autocorrelation-based dissimilarity** (ACF) calculates the Euclidean distance between estimated autocorrelation functions of time series. An autocorrelation function of a time series describes the correlation between two values of the time series at different times

with a specified lag (delay between the two values). In other words, it describes the correlation of a time series with a time-offset version of itself. It is defined as:

$$d_{ACF}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \sqrt{\left(\hat{\boldsymbol{\rho}}_{X_T} - \hat{\boldsymbol{\rho}}_{Y_T}\right)^\top \boldsymbol{\Omega} \left(\hat{\boldsymbol{\rho}}_{X_T} - \hat{\boldsymbol{\rho}}_{Y_T}\right)} \quad (12)$$

where Ω is a matrix of weights and *ρ*-hat refers to estimated autocorrelation vectors.

The **Partial Autocorrelation-based dissimilarity** (PACF) is identical to the ACF except that it uses the partial autocorrelation functions.

The **Periodogram-based dissimilarity** (Per) calculates the Euclidean distance between the periodograms of time series. A periodogram is a method of estimating the power spectrum of a time series, which is equivalent to the Fourier transform of the autocorrelation function. It describes how power is distributed over the frequency components of a time series.

The **Piccolo distance** (Piccolo) calculates the Euclidean distance between the AR(∞) operators, or autoregressive expansions, of invertible ARIMA models of time series. ARIMA is a time series forecasting method. ARIMA models work by describing autoregressive (AR) and moving average (MA) parameters. An autoregressive model explains a value in a time series by one or more previous values plus random error. It is generally written as AR(p), where p is the order of the model. An autoregressive expansion, AR(∞), is thus an AR model of infinite order. A moving average model—written as MA(q), where q is the order—explains a value in a time series by one or more past random errors as well as its own random error term. Invertible ARIMA models are those which can be written simply as autoregressive (AR) models. This is a necessary property to be able to forecast the dependent variable, and is important for the Piccolo distance, since only the AR aspect is used. ARIMA models can be applied to non-stationary time series, but they must first be converted to stationary time series by one or more differencing operations (subtracting each value from the one before it to remove stochastic trends).

The **Short Time Series distance** (STS) measures the difference between the slopes of time series defined as piecewise linear functions. It is intended to incorporate temporal information while ignoring absolute values, to overcome a weakness of many other

distances, including the Euclidean distance, which ignore the temporal order of points and the length of sampling intervals. The STS distance is defined as:

$$d_{STS}(X,Y) = \sqrt{\sum_{k=0}^{N-1} \left(\frac{y_{k+1} - y_k}{t_{k+1} - t_k} - \frac{x_{k+1} - x_k}{t'_{k+1} - t'_k} \right)} \tag{13}$$

Equations 3, 4, and 5 are copied from Cha (2007), equations 6 through 12 are copied from Montero and Vilar (2014), and equation 13 is copied from Mori *et al*. (2016).

Table S2.1: Distance measures included in the study. Parameters are provided as a range if one or more parameters are optional (e.g., some can be set to 'NULL,' while others are only sometimes relevant, depending on the input data). Note that one or more parameters may be listed for distance measures that are considered parameter-free if they require the user to choose a linked component, e.g., a compression algorithm. Note that names of distance measures are not necessarily given by the original authors, as some authors did not name their distance measures. References in the source column provide additional information on respective distance measures, but do not necessarily introduce them to the literature.

Distance Measure	Abbreviated Name	Type/Family	Category	Characteristics	Parameters	Source
Manhattan Distance[1,2]	Manhattan	Norm distance (L_p Minkowski family)	Lock-step Shape-based	Shortest distance between points on a grid	0	(Cha, 2007)
Euclidean Distance[1,2]	Euclidean	Norm distance (L_p Minkowski family)	Lock-step Shape-based	Shortest distance between points in Euclidean space	0	(Cha, 2007)
Chebyshev Distance[1,2]	Chebyshev	Norm distance (L_p Minkowski family)	Lock-step Shape-based	Takes maximum distance between point pairs (all other point pairs are ignored)	0	(Cha, 2007)
Lorentzian Distance[2]	Lorentz	L_1 family	Lock-step Shape-based		1	(Cha, 2007)
Gower Distance[2]	Gower	L_1 family	Lock-step Shape-based		0	(Cha, 2007)
Soergel Distance[2]	Soergel	L_1 family	Lock-step Shape-based		0	(Cha, 2007)
Kulczynski Distance[2]	Kulcz	L_1 family	Lock-step Shape-based		0	(Cha, 2007)

Canberra Distance[2]	Canberra	L_1 family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Squared Euclidean Distance[2]	SqEuclid	Squared L_2 family	Lock-step Shape-based		0	(Cha, 2007)
Divergence Squared Distance[2]	Diverge	Squared L_2 family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Squared Chi-Squared distance[2]	SqChi	Squared L_2 family	Lock-step Shape-based		0	(Cha, 2007)
Probabilistic Symmetric Chi-Squared Distance[2]	ProbSymm	Squared L_2 family	Lock-step Shape-based		0	(Cha, 2007)
Clark Squared Distance[2]	Clark	Squared L_2 family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Additive Symmetric Chi-Squared Distance[2]	Additive	Squared L_2 family	Lock-step Shape-based		0	(Cha, 2007)
Topsoe Distance[2]	Topsoe	Shannon's entropy family	Lock-step Shape-based	Symmetric form of K divergence	1	(Cha, 2007)
Kullback-Leibler (KL) Divergence[2]	Kullback	Shannon's entropy family	Lock-step Shape-based	Non-symmetric	1	(Cha, 2007)
K Divergence[2]	KDiv	Shannon's entropy family	Lock-step Shape-based	Non-symmetric	1	(Cha, 2007)
Jensen Difference[2]	Jensen	Shannon's entropy family	Lock-step Shape-based		1	(Cha, 2007)
Jeffreys Divergence[2]	Jeffreys	Shannon's entropy family	Lock-step Shape-based	Symmetric form of KL divergence	1	(Cha, 2007)

Squared-Chord Distance[2]	SqChord	Fidelity family	Lock-step Shape-based		0	(Cha, 2007)
Jaccard Distance[2]	Jaccard	Inner Product family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Dice Dissimilarity[2]	Dice	Inner Product family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Wave-Hedges Distance[2]	Wave	Intersection family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Czekanowski Distance[2]	Czek	Intersection family	Lock-step Shape-based	Normalizes	0	(Cha, 2007)
Time Alignment Measurement Distance[1]	TAM	Dog-man Distance	Elastic Shape-based	Warping path Time distortion penalty	0	(Folgado et al., 2018)
Edit Distance with Real Penalty[1]	ERP	Edit distance	Elastic Shape-based	Gap-length penalty	1-2	(Chen and Ng, 2004)
Dynamic Time Warping Distance[1]	DTW	Dog-man Distance	Elastic Shape-based	Warping path	0-4	(Sakoe and Chiba, 1978)
Edit Distance on Real Sequences[1]	EDR	Edit distance	Elastic Shape-based	Threshold parameter Gap-length penalty	1-2	(Chen et al., 2005)
Taneja Difference[2]	Taneja		Lock-step Shape-based	Uses both arithmetic and geometric means	1	(Cha, 2007)
Short Time Series Distance[1]	STS		Lock-step Shape-based	Captures temporal information	0-2	(Möller-Levet et al., 2003)

Kumar-Johnson Distance[2]	Kumar		Lock-step Shape-based	Uses symmetric chi-squared, arithmetic, and geometric means.	0	(Cha, 2007)
Dissimilarity Index Combining Temporal Correlation and Raw Value Behaviour[1]	Cort	Correction factor	Lock-step Shaped-based and Feature-based	Temporal correlation coefficient Adaptive tuning function	2	(Chouakria and Nagabhushan, 2007)
Complexity-Invariant Distance[1]	CID	Correction factor	Lock-step Shape-based	Invariant to complexity	1	(Batista et al., 2011)
Average(L_1,L_{inf})[2]	AVG		Lock-step Shape-based	Average of the Manhattan and Chebyshev distances	0	(Cha, 2007)
Periodogram Based Dissimilarity[1]	Per		Feature-based	Frequency domain	2	(Caiado et al., 2005)
Partial Autocorrelation-Based Dissimilarity[1]	PACF		Feature-based	Compares partial autocorrelation coefficients	1-2	(Montero and Vilar, 2014)
Integrated Periodogram Based Dissimilarity[1]	IntPer		Feature-based	Frequency domain	1	(Casado de Lucas, 2010)
Fourier Coefficient Based Distance[1]	Fourier		Feature-based	Frequency domain	1	(Agrawal et al., 1993)
Autocorrelation-Based Dissimilarity[1]	ACF		Feature-based	Compares autocorrelation coefficients	1-2	(D'Urso and Maharaj, 2009)
Piccolo Distance[1]	Piccolo		Model-based	ARIMA models	0-3	(Piccolo, 1990)
Normalized Compression Distance[1]	NCD	Compression distance	Compression-based	Normalization of differences Quasi-universality Choice of compression algorithm	1	(Cilibrasi and Vitanyi, 2005)

Compression-Based Dissimilarity Measure[1]	CDM	Compression distance	Compression-based	Compatible with symbolic representation Choice of compression algorithm	1	(Keogh et al., 2004)

[1]Available in the TSdist R package (Mori *et al*., 2016).

[2]Available in the philentropy R package (Drost, 2018).

S2.2. Distance measure properties

Here I have included some additional explanation for translation invariance, amplitude sensitivity and duration sensitivity.

Translation invariance: Translation invariance is a shape-preserving property, meaning that a distance measure with this property would treat two time series with identical shapes as equal, even if the mean values were different. The same effect can be achieved by a vertical shift transformation. For example, time series X can be transformed by adding the same real number q to each observation, f(X) = X + q, such that time series X and Y have the same starting value (if they already have the same starting value there is no need for transformation). It is a simple matter to apply this transformation to thousands of time series. Note, however, that translation invariance can be problematic. Consider two populations, with population A having a starting size of 100 and population B a starting size of 10,000. If both populations increase by 10 every year for 10 years, population A would now be 200, which means it doubled to twice its original size, while population B would be 10,100, an increase of only 1%. A distance measure with the property of translation invariance (or any distance measure after applying a vertical shift transformation to equalize the starting values) would treat these trends as equal.

An alternative way to deal with such comparisons would be a scale transformation, f(X) = X * q, multiplying each observation of time series x by the same real number q, such that time series X and Y have the same starting value. A scale transformation allows for shape deformation while preserving percentage change. If populations A and B both doubled by increasing linearly for 10 years from 100 to 200 and 10,000 to 20,000 respectively, a scale transformation would result in identical trends, although they did not originally have the same shape. Likewise, in the previous example where two populations of different sizes increase by the same amount but different percentages, a scale transformation would result in trends with very different shapes (slopes). Scale invariance, which is defined as d(X*q, Y) = d(X, Y), with q greater than 0 (Batyrshin *et al*., 2016), is a rare property of distance measures.

Note that translation invariance is a special case of translation insensitivity, where d(X + q, Y) is independent of q. Defining its opposite, translation sensitivity, gives d(X + q, Y) > d(X, Y), with d(X + q, Y) increasing with q. In other words, adding a number to all values of X causes the dissimilarity between X and Y to increase, and the greater the number added to X, the greater the increase in dissimilarity. This is a useful property and one that can be measured in relative terms.

Amplitude sensitivity: Amplitude sensitivity is particularly relevant when comparing time series which have been scale transformed or vertical shift transformed to have the same starting value. But some distance measures, especially among edit distances, are insensitive to amplitude. For example, the Edit Distance for Real Sequences (EDR) has a threshold value that can be set. Only differences that exceed the threshold are counted. This can be useful when looking for aberrations. For example, time series of maximum daily temperatures could be ranked according to how often they exceed a baseline by a set number of degrees.

Duration sensitivity: Some distance measures, such as Dynamic Time Warping (DTW) or the Short Time Series Distance (STS), may rank time series with more differences as more dissimilar *only if those differences are separated by similarities*. Consider a time series A, with five points, t_0, t_1, ..., t_4. Some transformation f(t) is applied only to point t_1 to form time series B (B thus differs from A by a single point), to both t_1 and t_2 to form time series C, and to both t_1 and t_3 to form time series D (thus C and D each differ from A by the same value at two points, but in C those points are consecutive while in D they are not). For distance measures which are sensitive to *both* frequency and duration, d(D, A) > d(C, A) > d(B, A), but for distance measures which are sensitive to frequency but *not* duration, d(D, A) > d(B, A), while d(C, A) = d(B, A). This is because a distance that is invariant to duration will treat a difference that occurs over multiple consecutive time points as a single difference.

S2.3. Uncontrolled testing

I created a function for each property to be tested, which applies a transformation to one or more time points of a real-world time series given as input. Each function accepts a value q, the purpose of which varies depending on the function (see below for details). For example, the translation function adds a real number q to every value t_i of a time series T. The transformed time series is returned as output and compared against its unaltered counterpart. I applied the functions to a range of q in increments, then graphed the results as response curves (see Figs S2.2-S2.5). I did not compare them against a reference or assign sensitivity ratings, as they were intended only as a confirmatory check against the results of controlled testing.

Functions:

Translation sensitivity: Add q to every data point of a time series T.

White noise sensitivity: Create a normal distribution with mean q and standard deviation 0.3 times q (the latter is arbitrary). Randomly select half of the data points of a time series T and add randomly selected values from the normal distribution to the selected points. Finally, subtract randomly selected values from the normal distribution from the points that were not selected. This function scales q by $\frac{q}{\max(q)}$ to avoid the noise being too large.

Biased noise sensitivity: Proceed exactly as with white noise sensitivity but skip the final step (the points that were not selected remain untransformed). This function scales q by $\frac{q}{0.5*\max(q)}$ (the 0.5 is because the function is only applied to half of the time points).

Outlier sensitivity: Add q to one randomly selected point of a time series T (excluding the first and final points, which can cause unintended behaviour in some distance measures).

Phase invariance: Shift the first q time points of a time series T to the end of the time series.

Warping invariance: Randomly select a single value from a time series T and extend the time series by repeating the chosen value q times.

Uniform time scaling invariance: Stretch a time series T along the x-axis by a factor of q. The y-axis values of the first and final points remain unchanged, but the final point is shifted

along the x-axis and all points in between are recomputed. For this function, q is scaled: $\frac{q}{\max(q)} + 1$. Thus, time series T will be stretched to a maximum of twice its original length.

S2.4. Metric test results

In some cases, results depended on input values or settings. Eight of the lock-step shape-based distance measures passed the triangle inequality test and/or non-negativity test when inputs were constrained to non-negative real numbers but failed when negative numbers were included. EDR behaved as a metric when the threshold setting, ε, was set near zero, but failed the triangle inequality test when ε was set at five. The Normalized Compression Distance (NCD) and the Compression-based Dissimilarity Measure (CDM) both failed uniqueness and symmetry tests and thus qualified as non-metrics, although NCD is stated by its authors to be a metric (Cilibrasi and Vitányi, 2005). However, this is qualified with respect to the compression algorithm paired with it, with none quite reaching the definition the metric behaviour depends on. NCD should approach closer to true metric behaviour the longer the time series (Cilibrasi and Vitányi, 2005). I tested it here with very short time series, and therefore it would not be expected to behave as a metric. Additional testing (not included) showed NCD came closer to passing the uniqueness and symmetry tests, although as the time series reached a length of one million, it was still failing. Beyond that length, running the tests was too slow to be practical. CDM, on the other hand, is not considered to be a metric (Keogh *et al*., 2004), nor did it approach closer to metric behaviour when tested with longer time series.

S2.5. Controlled test results

S2.5.1. Sensitivity tests

Results for EDR depended on the value of the threshold setting, ε. I reported the results with ε at 0.1. However, when ε was set high, EDR was invariant to all seven of these properties. When ε was set within the range of the input values, results were less predictable.

The two compression-based distances I tested, the Normalized Compression Distance (NCD) and the Compression-based Dissimilarity Measure (CDM), showed insensitivity to translation and outliers. However, uncontrolled test results did not confirm this. It is not clear why this difference occurred, but keep in mind that compression-based distances may behave differently for short time series than for long ones (e.g., they do not behave as metrics when comparing short time series).

S2.5.2. Time-based invariances and other tests

Again, results for EDR depended on the value of ε. I reported the results with ε at 0.1., but when ε was set high, EDR was invariant to phase, and sensitive to both warping and time scaling. When ε was set within the range of the input values, it responded unpredictably to phase shift, but remained sensitive to both warping and time scaling.

The results for the Autocorrelation-based dissimilarity (ACF) and the Partial Autocorrelation-based dissimilarity (PACF) were 'n/a' for both warping and time scaling, suggesting that these distance measures are unable to deal with unequal-length time series. However, this is not the case. The problem is that these measures require an equal number of autocorrelation coefficients, which the short time series I used for controlled testing did not satisfy. However, ACF and PACF did provide results for warping and time scaling in uncontrolled testing (Fig. S2.2 and S2.4).

S2.5.3. Pairwise correlations between distance measures

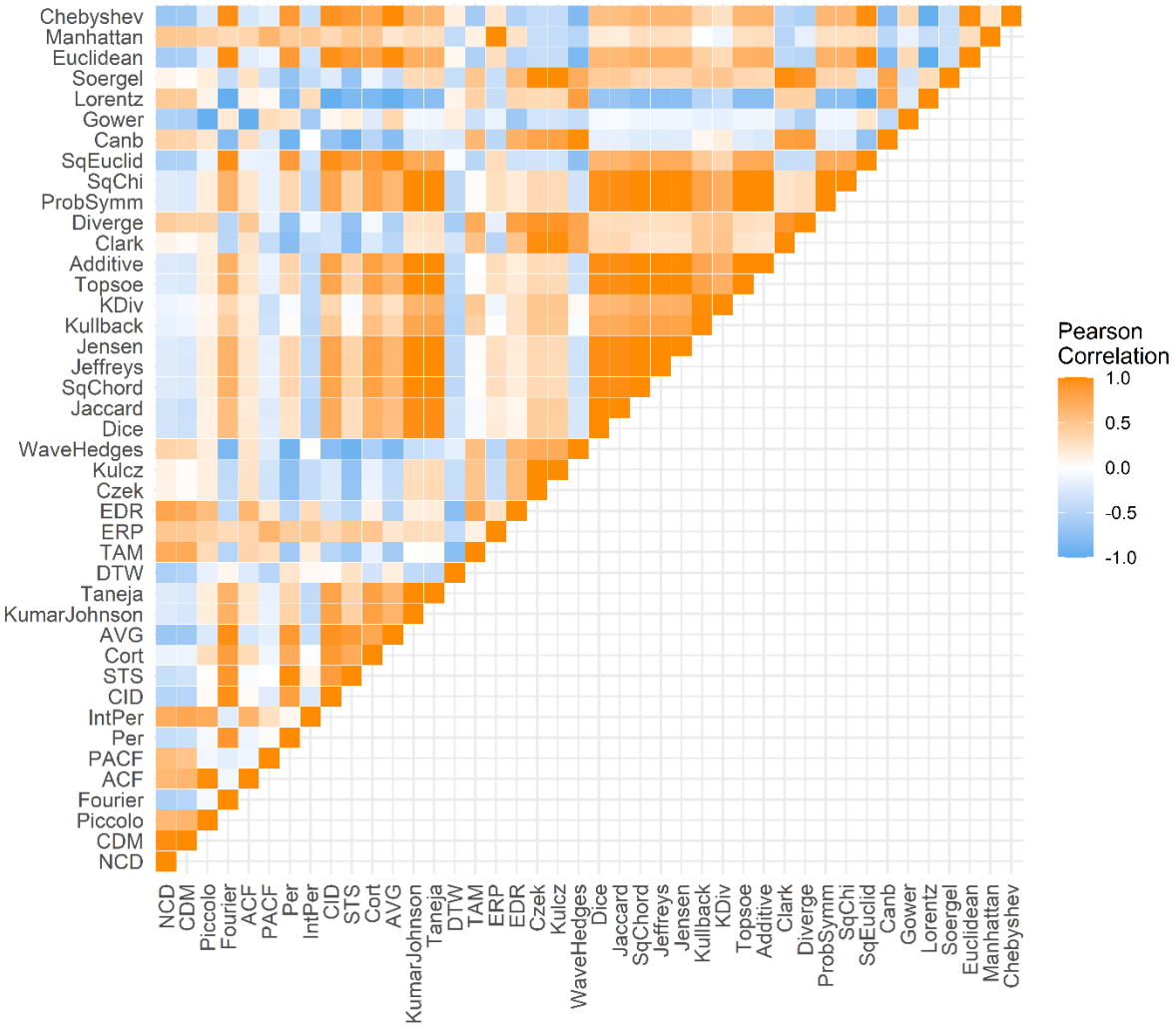


Figure S2.1. Pairwise Pearson correlation of all tested distance measures, based on the unbinned results of controlled sensitivities testing for translation, amplitude, duration, frequency, white noise, biased noise, and outliers. Distance measures are organized by family on the plot.

S2.6. Uncontrolled test results

Figures S2.2-S2.5 show the results of uncontrolled testing of distance measure properties using two real-world time series (main text: Fig. 2.3) from the UCR Time-Series Classification Archive (Dau *et al.,* 2019), an archive of 128 time-series datasets intended for testing of classification algorithms. All dissimilarity values in the test results have been rescaled to a range of [0,1] using Min-Max scaling. This was done to facilitate placing response curves for different types of transformations on the same plot while still allowing the shape of each response curve to be seen regardless of the strength of the response. For controlled testing, time series were carefully constructed to allow comparison of response strength across different properties. That is a far more difficult problem when working with real-world time series, so I opted instead to exchange strength information for better shape resolution.

For those distance measures that are sensitive to a tested property, the dissimilarity value shows a response curve as the size of the transformation value q increases. The sensitivity response curve may be linear or not but should be described by a function. Invariances show as horizontal lines at a dissimilarity value of zero, while insensitivities show as horizontal lines at some non-zero value. The response curves for some properties differ in shape between time series, especially for elastic distance measures and those designed for stationary time series. Despite this, results are largely consistent with the controlled testing results shown in Figs 2.5-2.6 (main text).

There are a few exceptions, however. Both compression-based distances I tested, NCD and CDM, registered as insensitive to translation and outliers in controlled testing, while showing unpredictability in uncontrolled testing. Two feature-based distances, ACF and PACF, showed unpredictability for warping invariance and uniform time scaling invariance in uncontrolled testing but failed to give results in controlled testing. This was because these distance measures require the time series being compared to have an equal number of autocorrelation coefficients, a requirement which was met when extending the real-world time series, but not when extending the short time series that I created for controlled testing. Finally, the Time Alignment Measurement distance, TAM, showed unpredictability to outliers in controlled testing, but was insensitive in uncontrolled testing. The raw dissimilarity values from the controlled testing showed a sudden increase from a

dissimilarity value of 0 to 0.33 as the value of q increased from 2 to 3. Given that I used the same starting value of q (1) and the same increment size (also 1) for both controlled and uncontrolled testing, the threshold is presumably not determined simply by the value of the outlier, q, but by a more complex calculation.

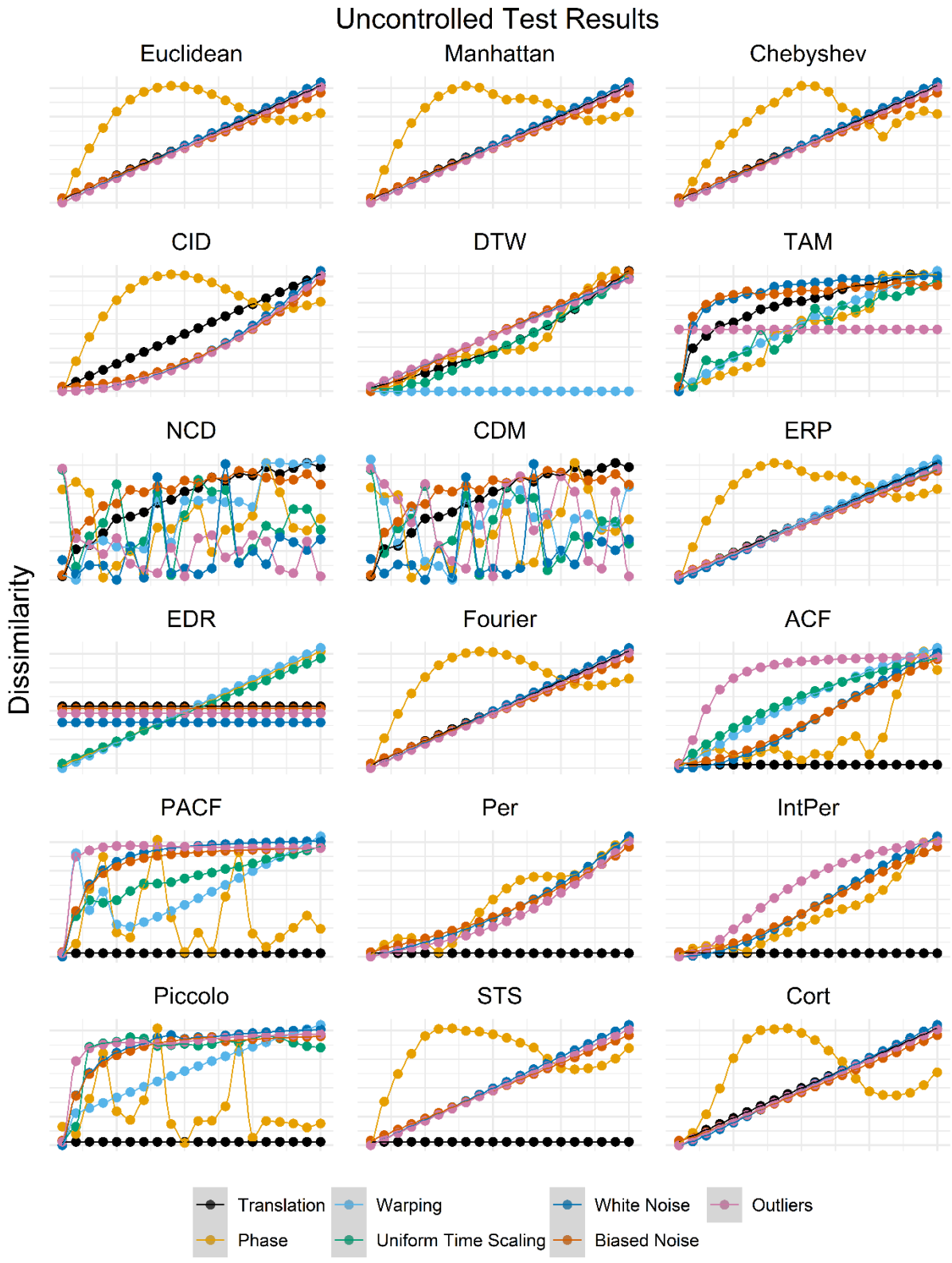


Figure S2.2. Dissimilarity measurements from 17 distance measures of the TSdist package after applying transformations to a randomly selected time series from the Yoga dataset of the UCR Time-Series Archive. The x-axis depicts the transformation value q across a range of 1 to 200 in increments of 10. Dissimilarity values were rescaled using Min-Max scaling to a range of [0,1] to ensure that the shape of each response curve would be visible regardless of the strength of the response.

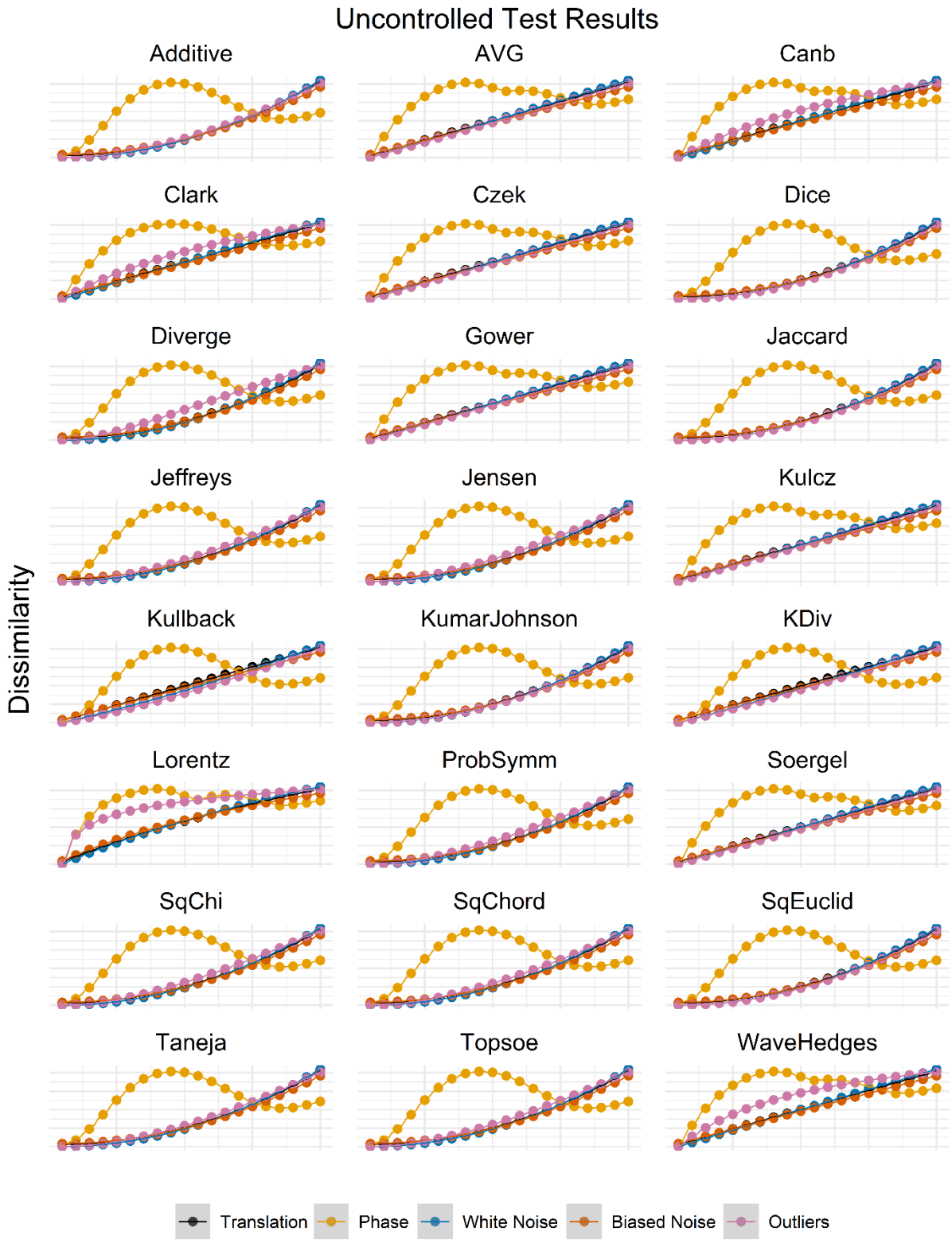


Figure S2.3. Dissimilarity measurements from 24 distance measures of the philentropy package after applying transformations to a randomly selected time series from the Yoga dataset of the UCR Time-Series Archive. The x-axis depicts the transformation value q across a range of 1 to 200 in increments of 10. Dissimilarity values were rescaled using Min-Max scaling to a range of [0,1] to ensure that the shape of each response curve would be visible regardless of the strength of the response.

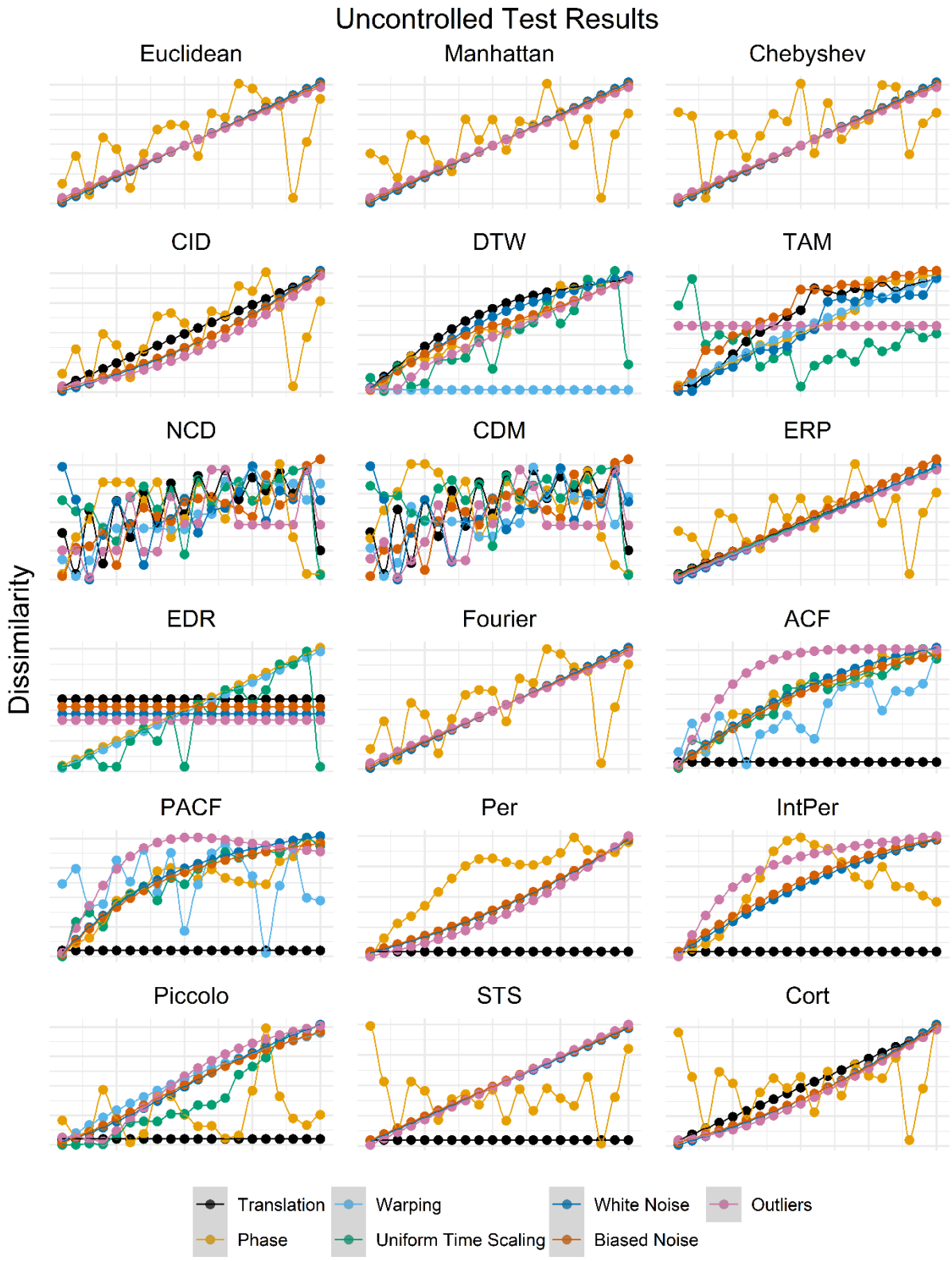


Figure S2.4. Dissimilarity measurements from 17 distance measures of the TSdist package after applying transformations to a randomly selected time series from the Synthetic Control dataset of the UCR Time-Series Archive. The x-axis depicts the transformation value q across a range of 1 to 20 in increments of 1. Dissimilarity values were rescaled using Min-Max scaling to a range of [0,1] to ensure that the shape of each response curve would be visible regardless of the strength of the response.

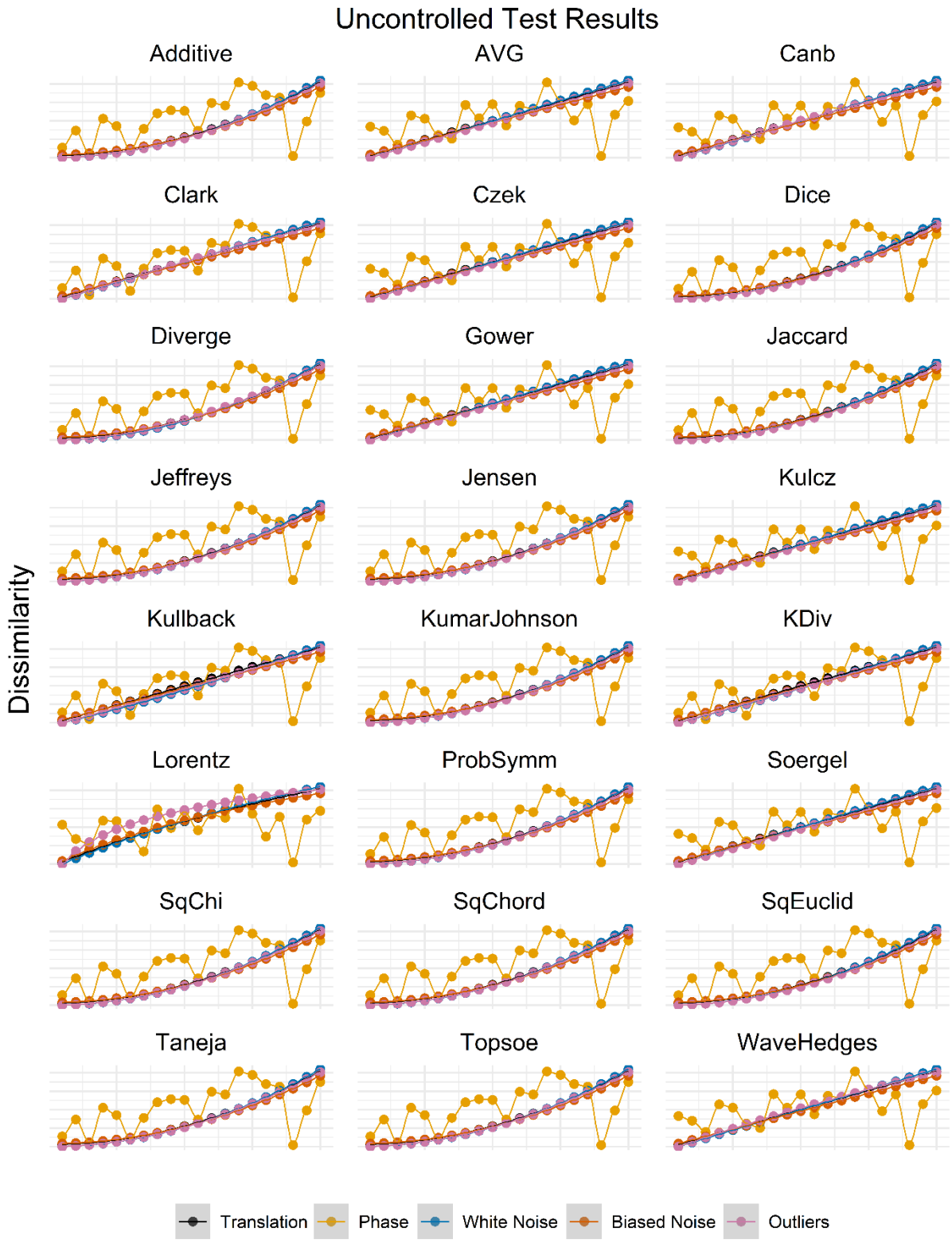


Figure S2.5. Dissimilarity measurements from 24 distance measures of the philentropy package after applying transformations to a randomly selected time series from the Synthetic Control dataset of the UCR Time-Series Archive. The x-axis depicts the transformation value q across a range of 1 to 20 in increments of 1. Dissimilarity values were rescaled using Min-Max scaling to a range of [0,1] to ensure that the shape of each response curve would be visible regardless of the strength of the response.

S2.7. Example dataset results

When comparing unsmoothed time series, the two selected distance measures gave identical results to percent improvement from Jellesmark *et al*. (2021) (main text: Fig. 2.9). Among the 40 unselected distance measures, only two, EDR and TAM, gave the same rankings as the selected distance measures and percent improvement (main text: Fig. 2.9). Another 30 agreed with Jellesmark *et al.* (2021) in ranking Redshank first, but beyond that the results differed strongly, with 28 ranking Yellow Wagtail second and 23 ranking Curlew last (main text: Fig. 2.9). None of the distance measures returned the same results as the t-test.

In the smoothed time series comparison, all seven of the selected distance measures agreed with each other but differed slightly from the percent improvement results of Jellesmark *et al*. (2021) by placing Snipe ahead of Lapwing (main text: Fig. 2.10). Of the 35 unselected distance measure, four gave the same rankings as the selected distance measures, while 11 agreed with percent improvement and five agreed with the t-test (main text: Fig. 2.10).

S2.8. Speeding up DTW

For matching problems, such as content queries and classification, the slowness of DTW can be avoided by indexing, which severely reduces the number of time series that need to be compared to find the best match. For the Euclidean Distance, indexing is straightforward to accomplish. However, as DTW does not satisfy the triangle inequality (main text: Fig. 2.4), it presents more of a challenge. Keogh and Ratanamahatana (2005) solved this problem using a tight 'lower-bounding' measure, which is included in the TSdist package (Mori *et al.,* 2016) as LBKeoghDistance. For an explanation of lower bounding and the indexing process with respect to DTW, refer to Keogh and Ratanamahatana (2005). The lower-bounding technique does not apply to clustering, where some real-world problems can take weeks or even months (Zhu *et al*., 2012). However, Zhu *et al*. (2012) solved this problem for clustering by creating an interactive 'anytime algorithm', which uses a fast approximation of DTW to give a best available answer that improves over time as exact DTW calculations are performed and can be paused or terminated at any time.

S2.9. Plots of controlled test results for all distance measures

This section contains plots of controlled testing results for all 42 distance measures I tested. Each figure includes all time-based and values-based properties for which that distance measure gave results. Distance measures are presented in alphabetical order.

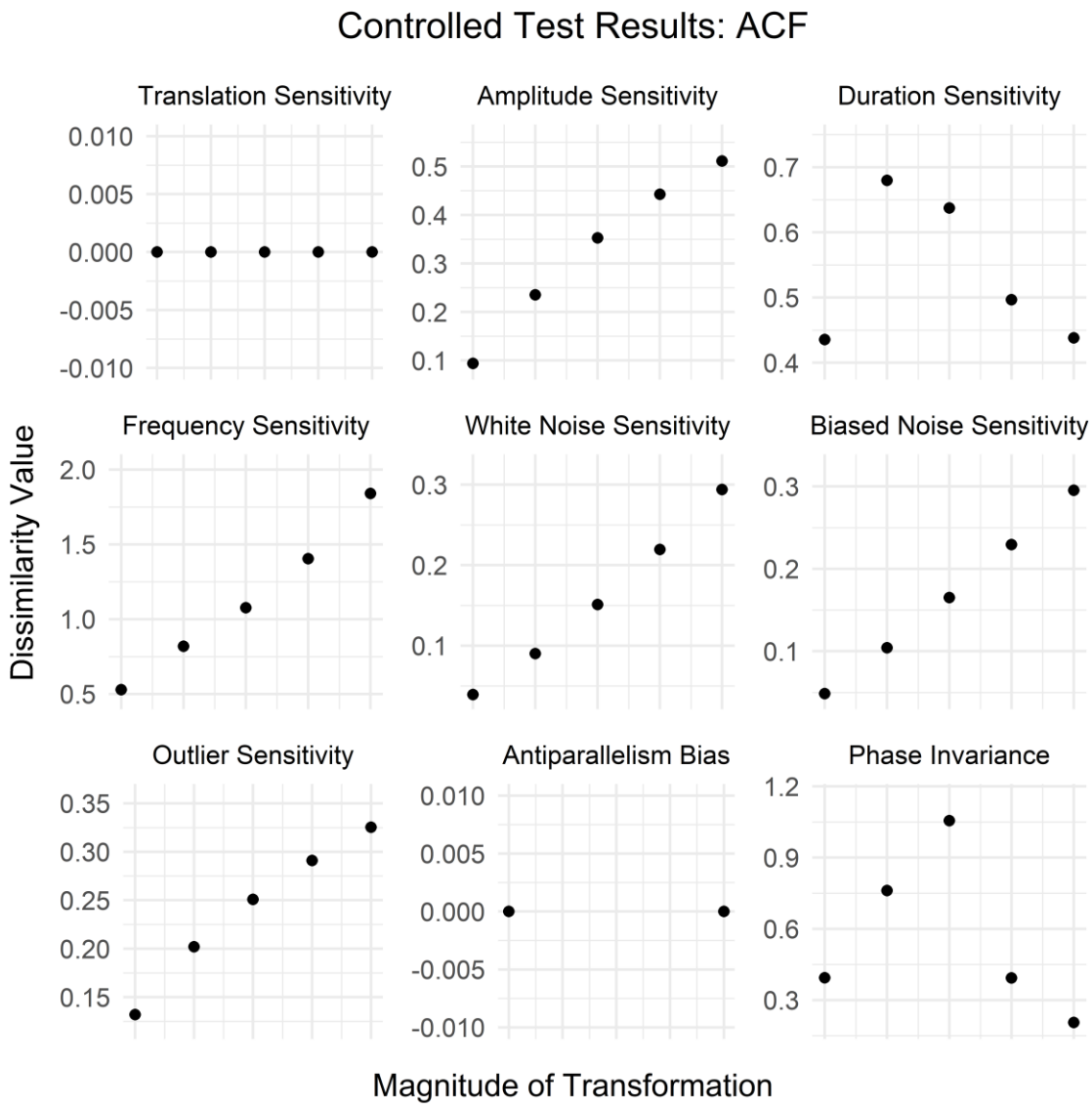


Figure S2.6. Controlled testing results for the Autocorrelation-Based Dissimilarity. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

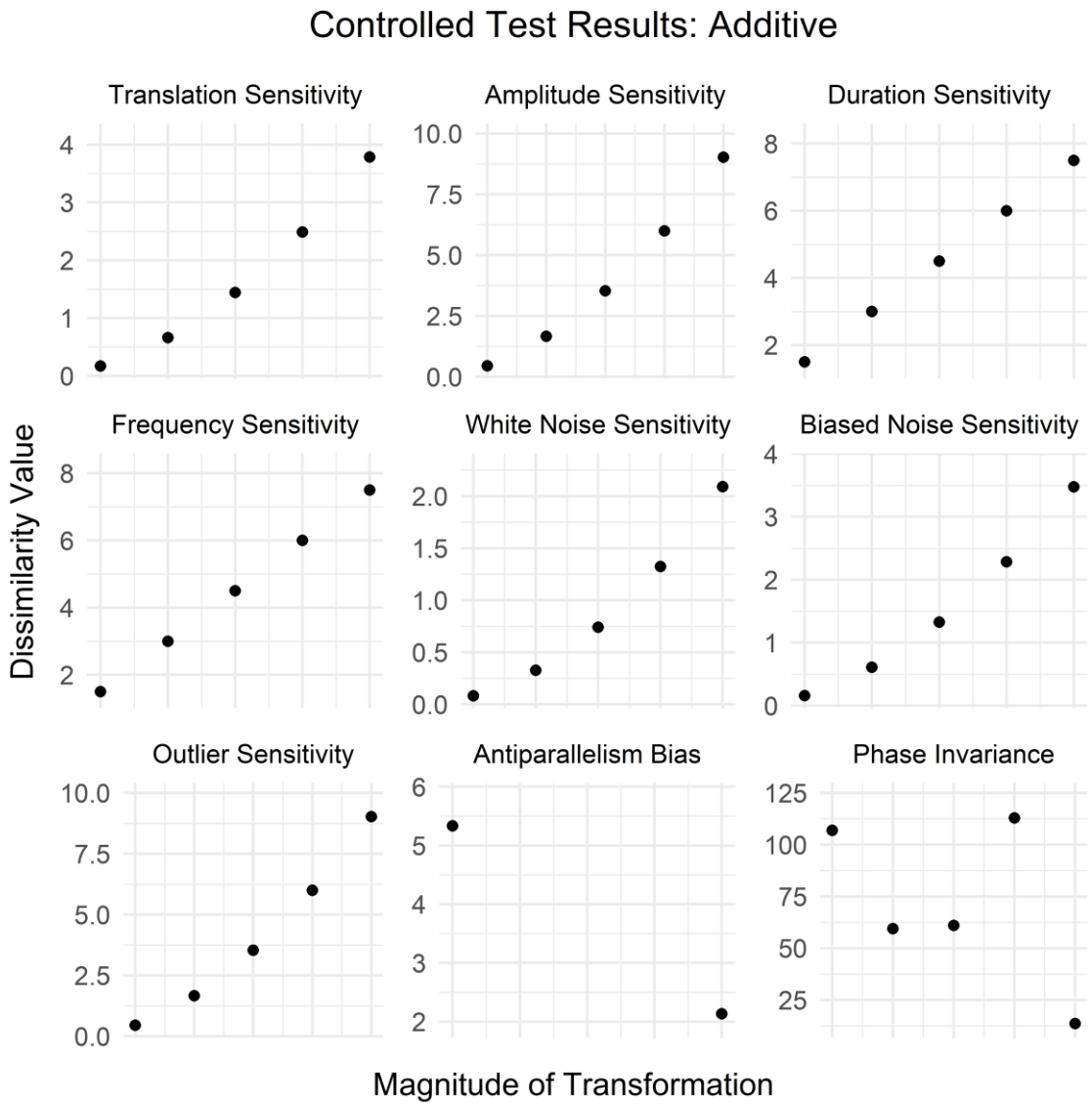


Figure S2.7. Controlled testing results for the Additive Symmetric Chi-Squared Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

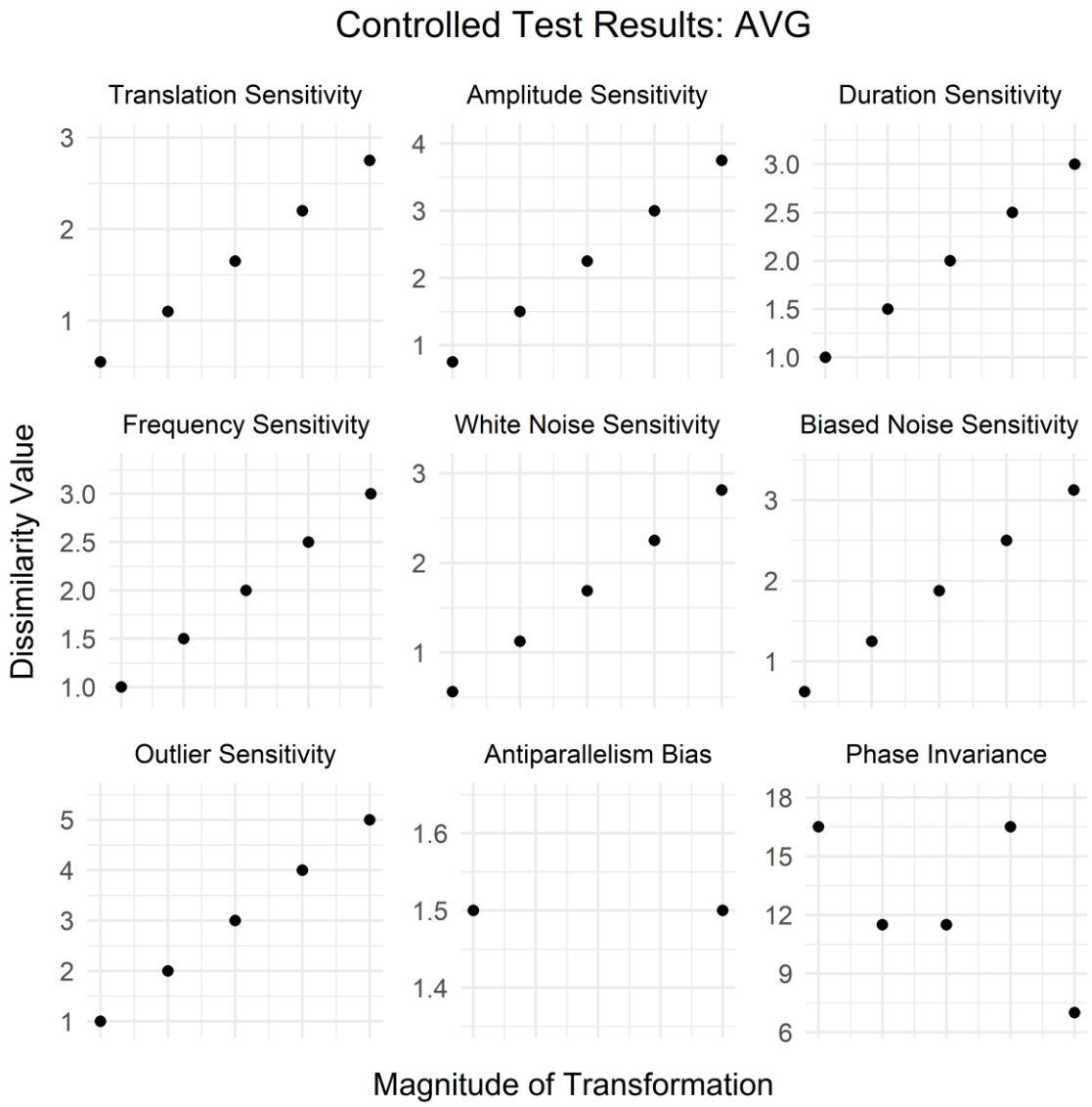


Figure S2.8. Controlled testing results for the Average Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

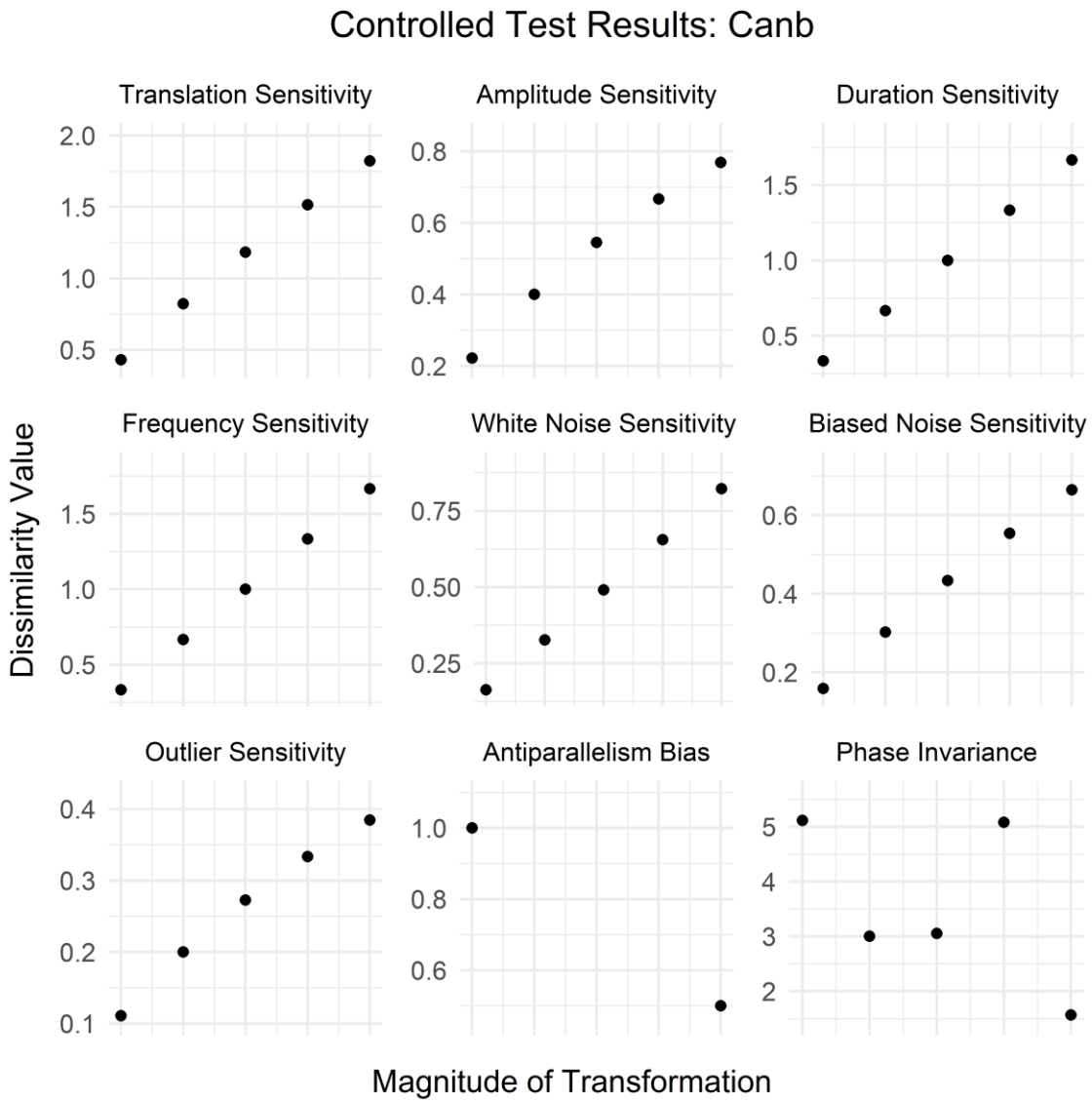


Figure S2.9. Controlled testing results for the Canberra Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

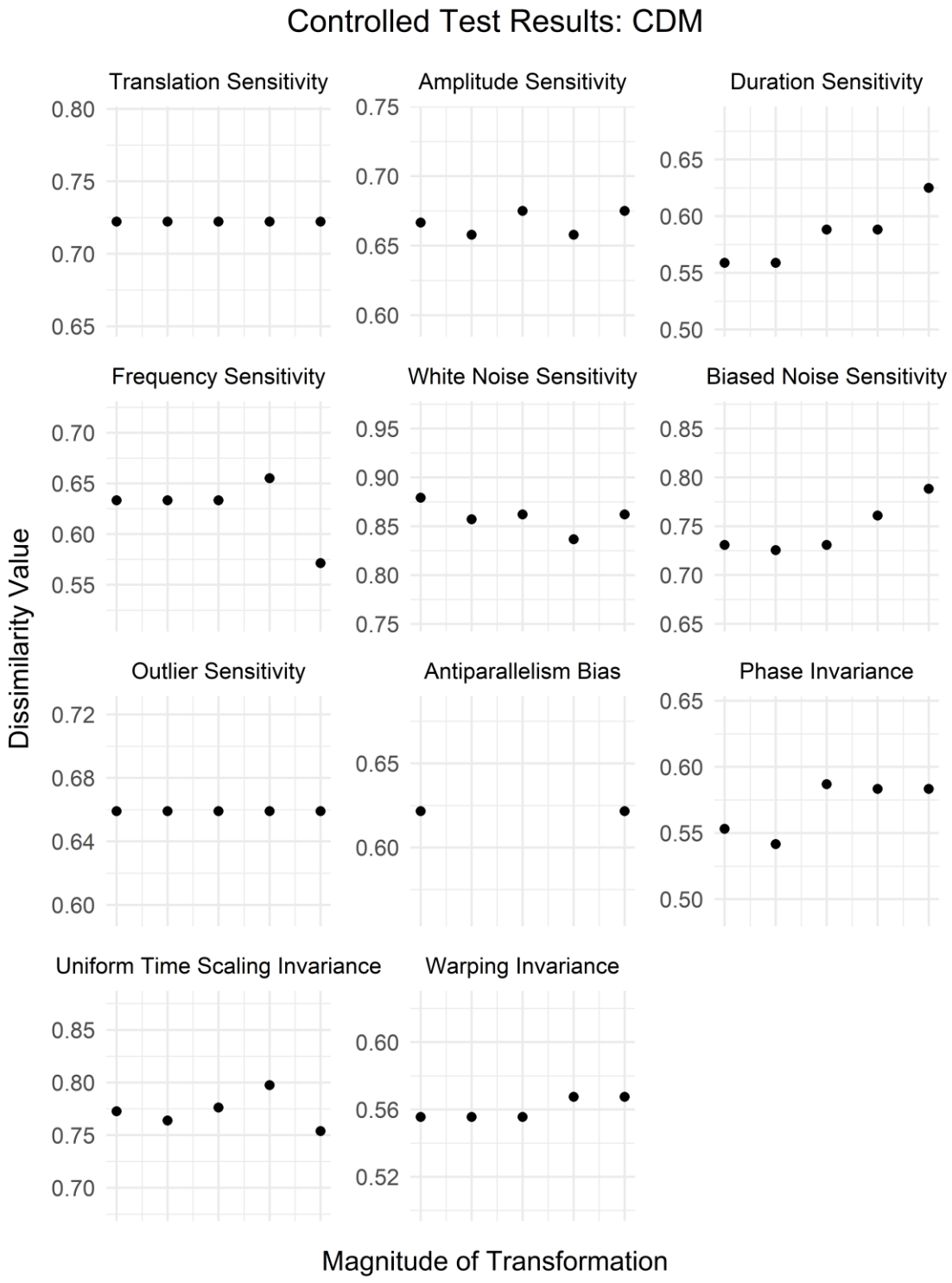


Figure S2.10. Controlled testing results for the Compression-Based Dissimilarity Measure. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

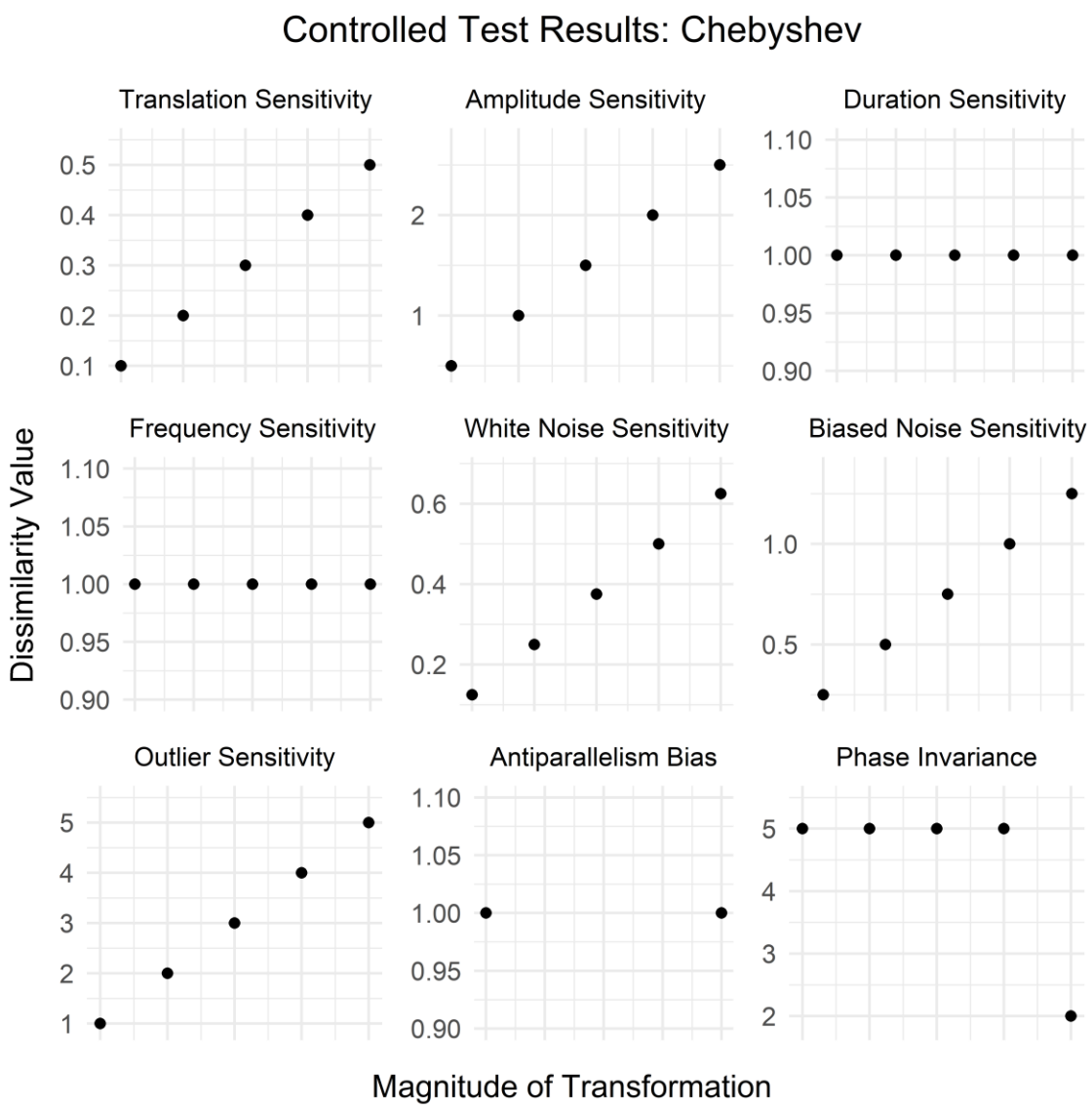


Figure S2.11. Controlled testing results for the Chebyshev Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

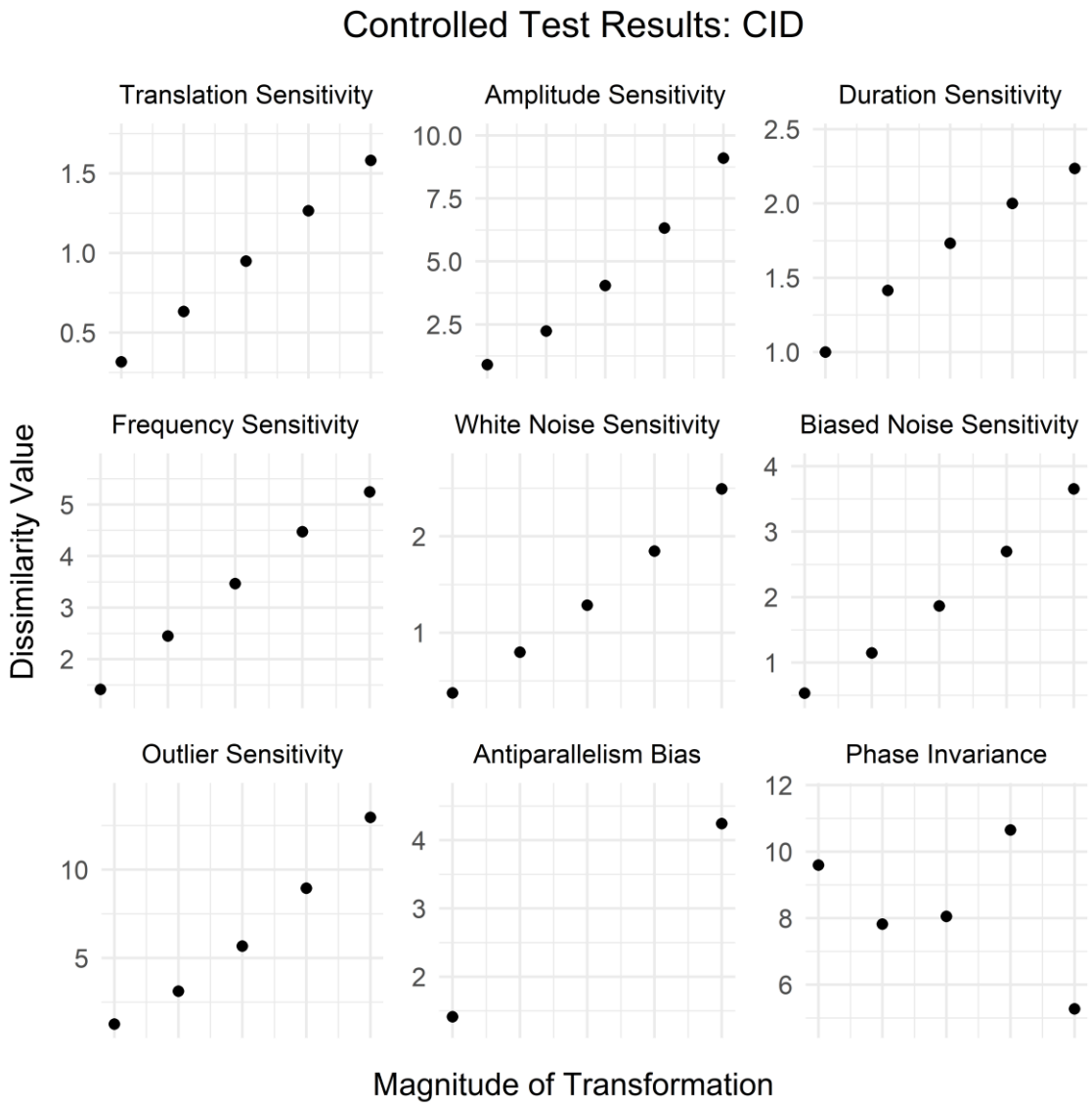


Figure S2.12. Controlled testing results for the Complexity-Invariant Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

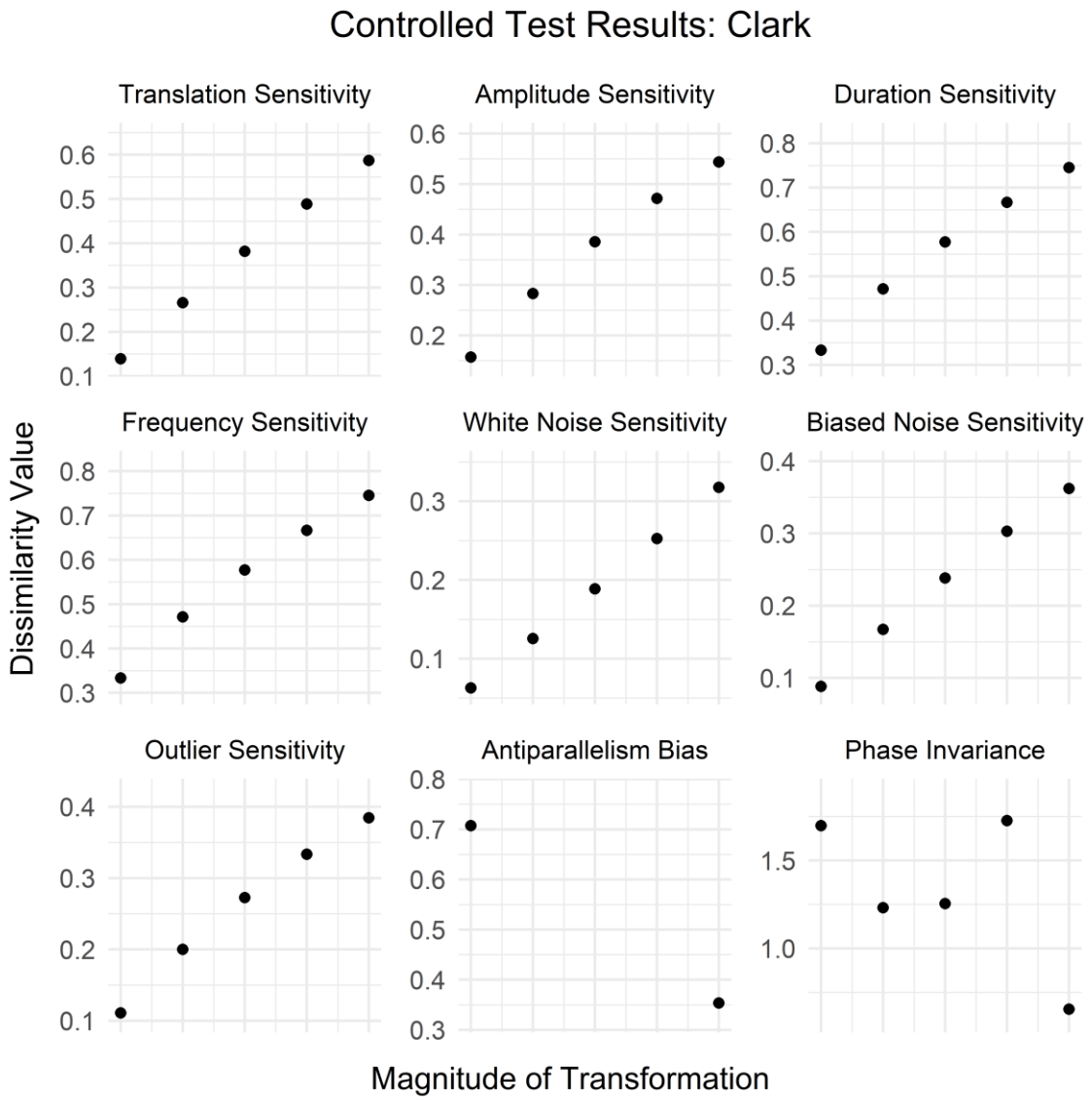


Figure S2.13. Controlled testing results for the Clark Squared Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

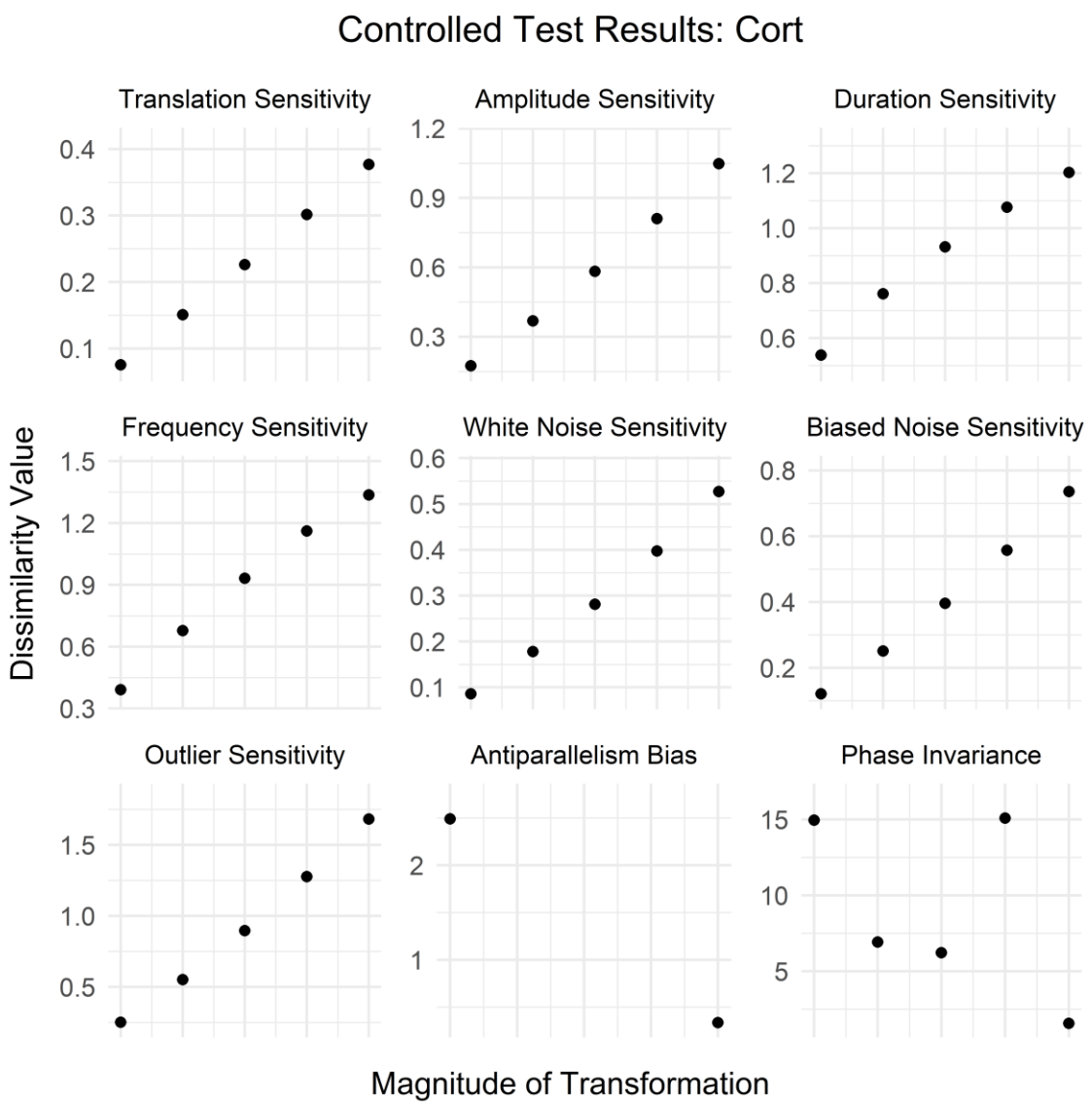


Figure S2.14. Controlled testing results for the Dissimilarity Index Combining Temporal Correlation and Raw Value Behaviour. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

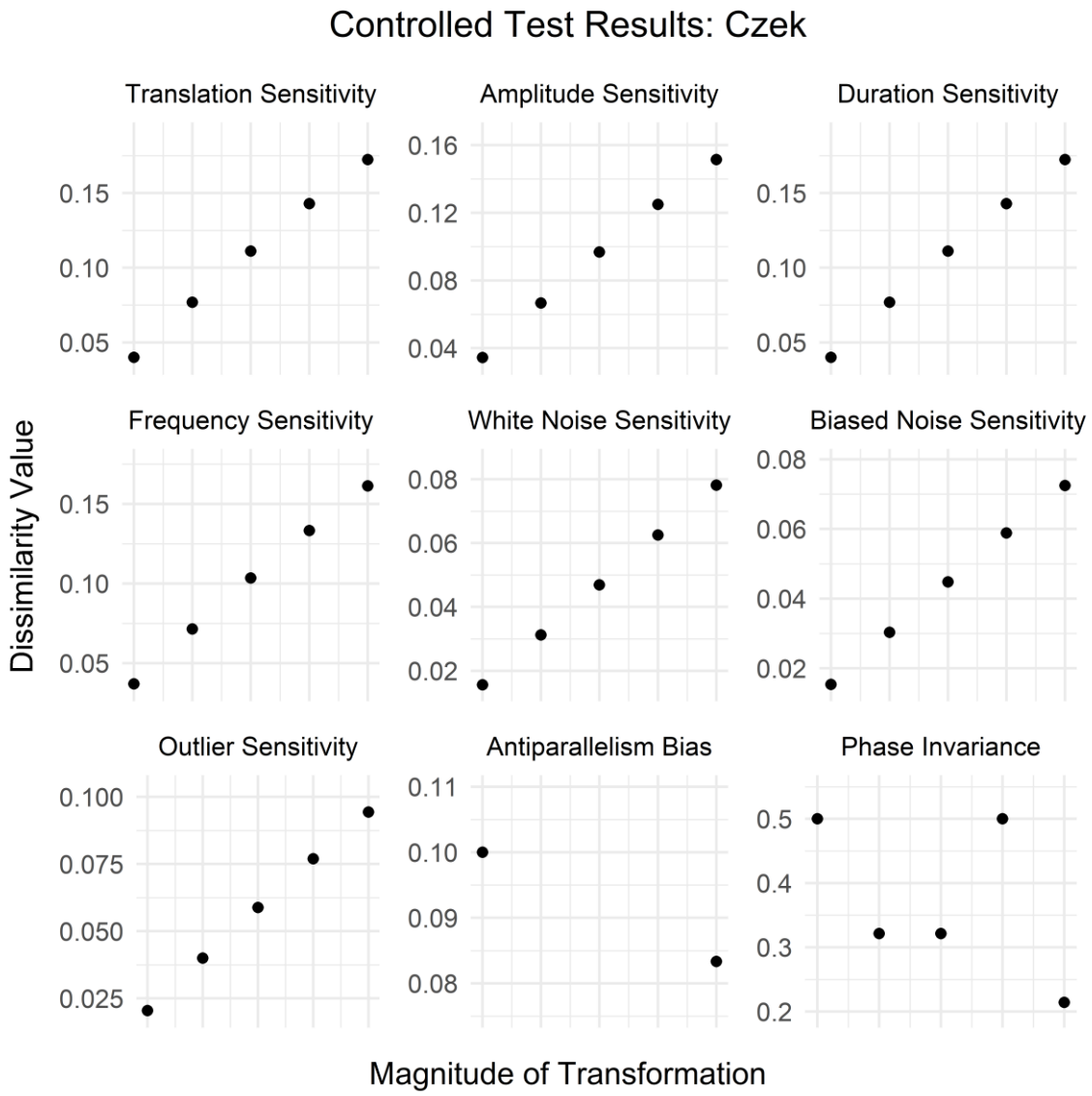


Figure S2.15. Controlled testing results for the Czekanowski Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

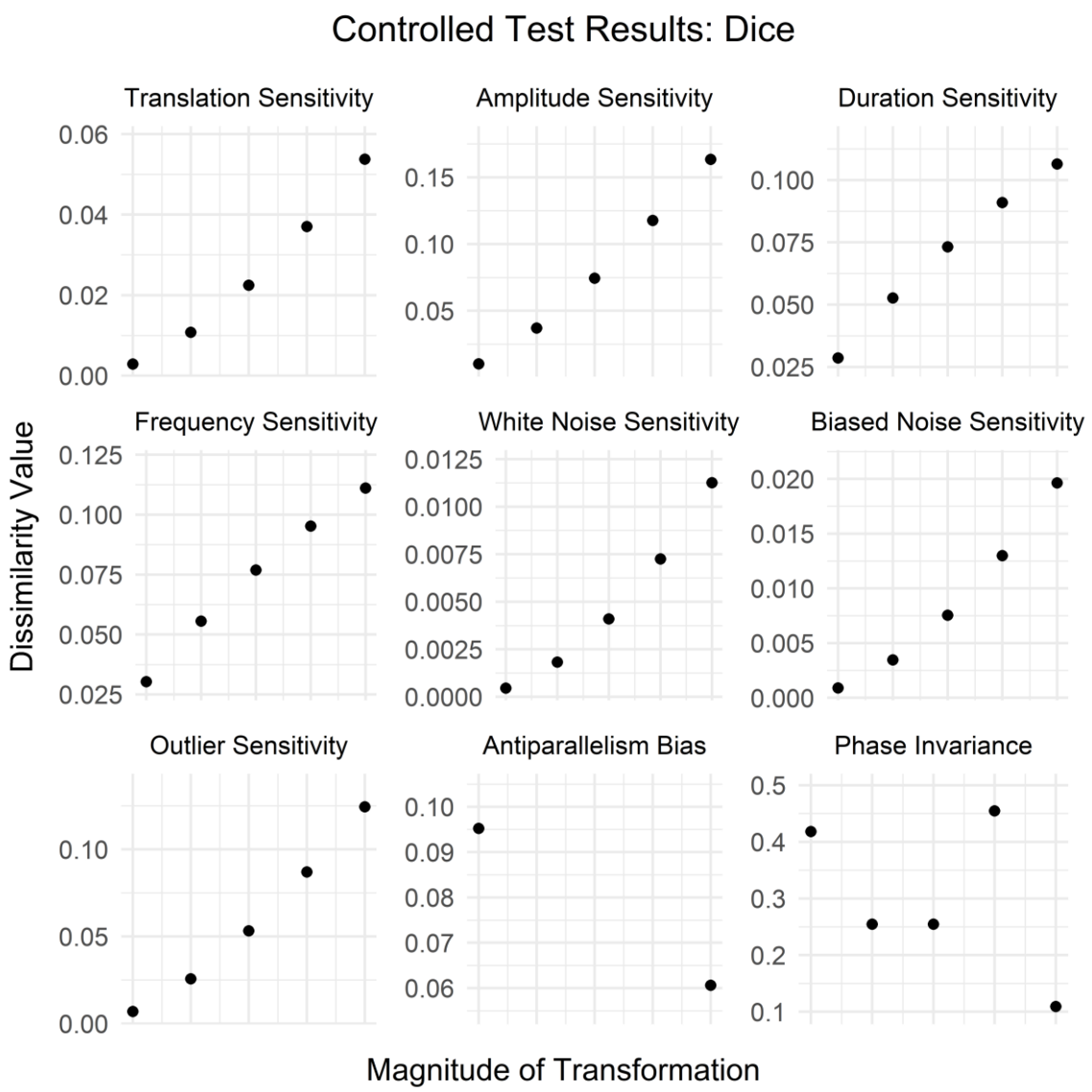


Figure S2.16. Controlled testing results for the Dice Dissimilarity. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

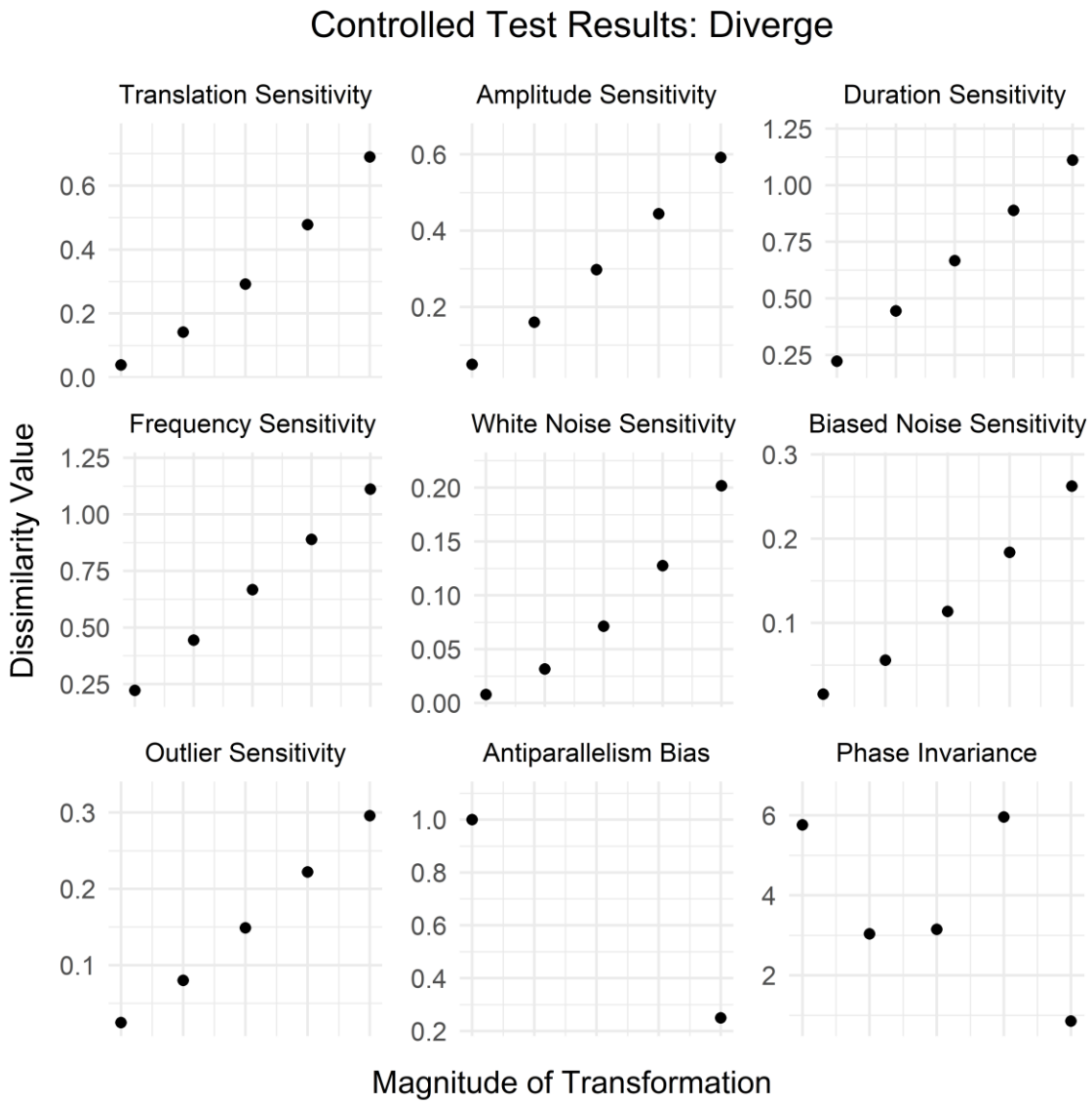


Figure S2.17. Controlled testing results for the Divergence Squared distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

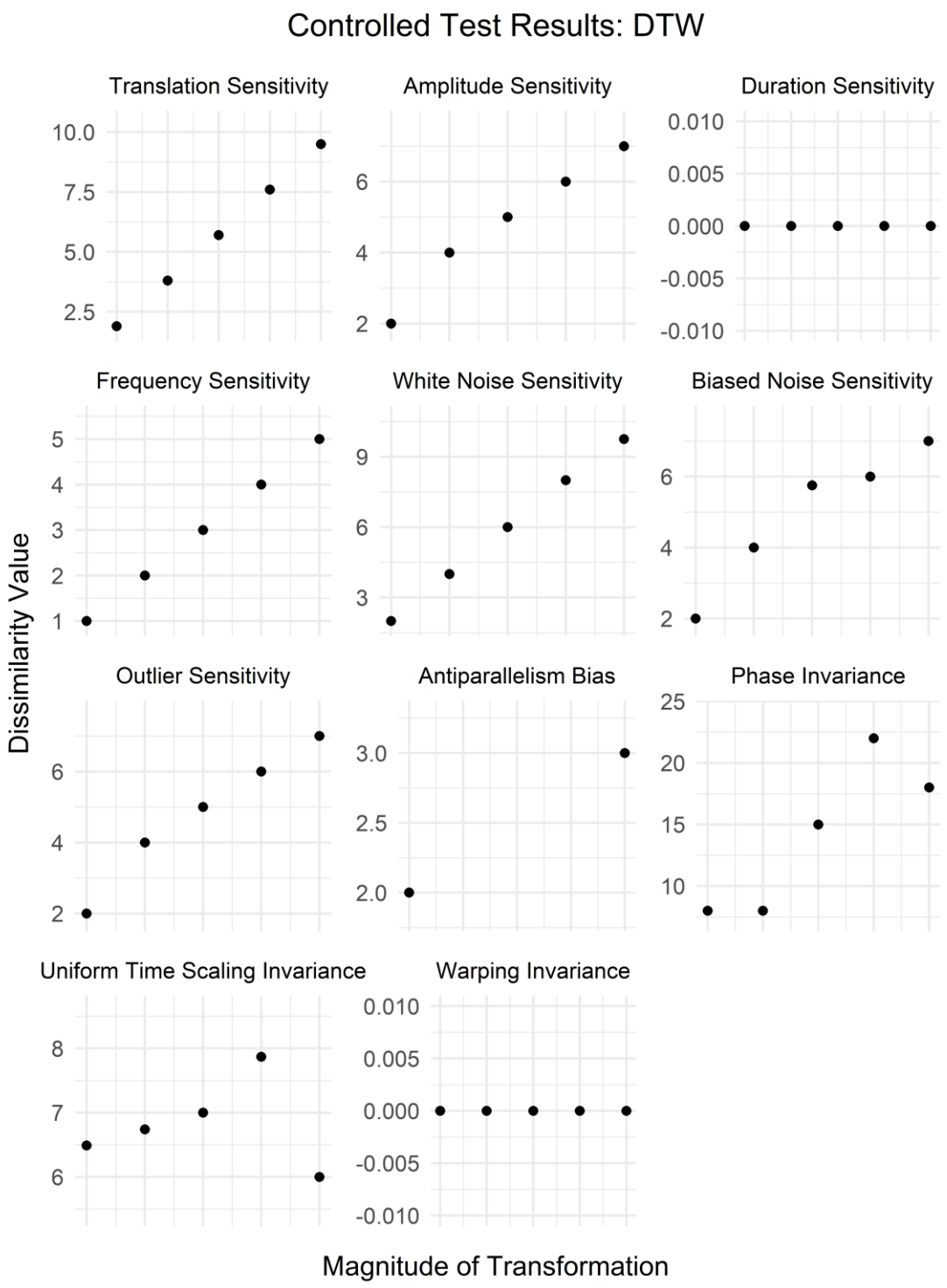


Figure S2.18. Controlled testing results for the Dynamic Time Warping Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

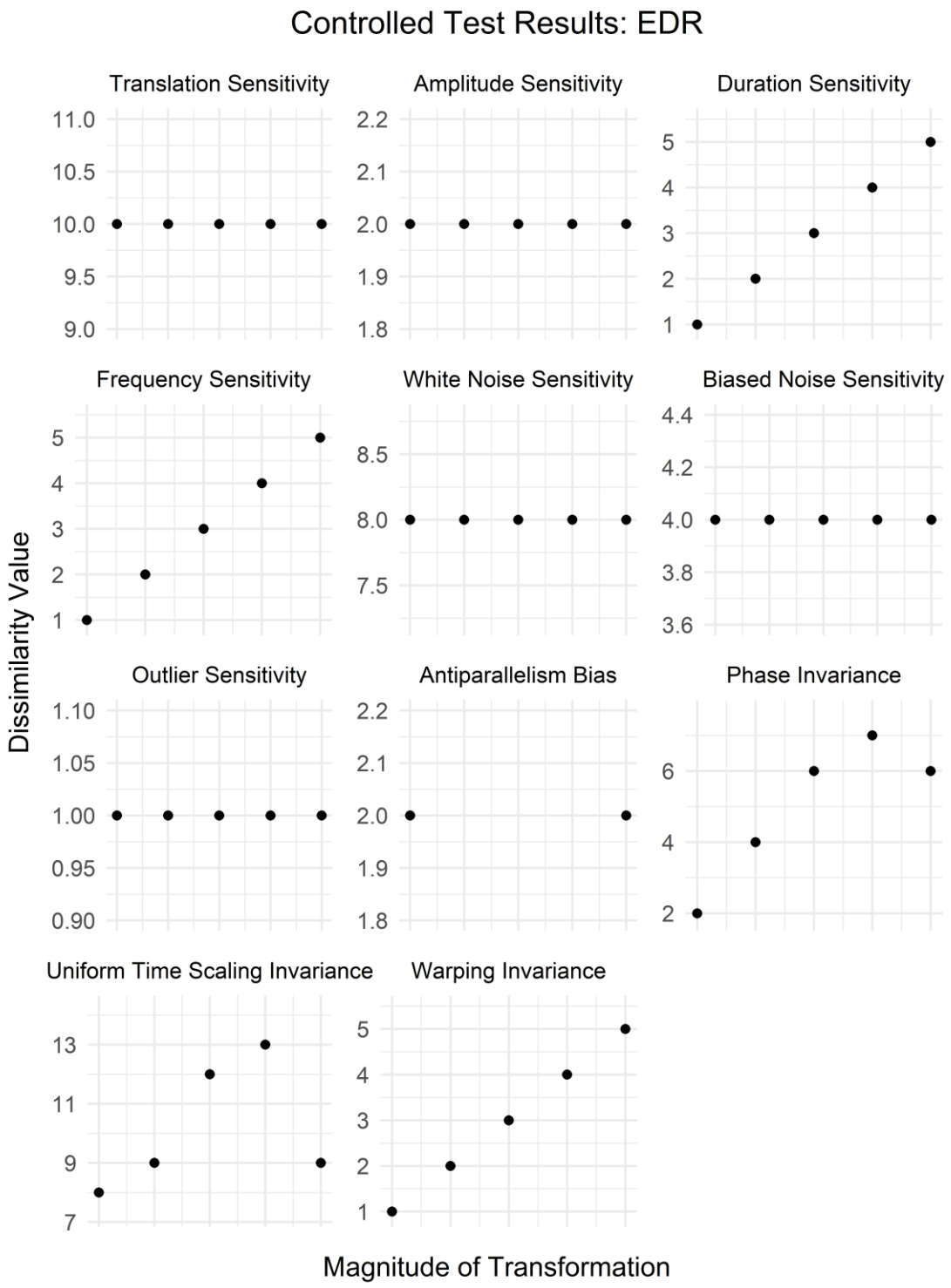


Figure S2.19. Controlled testing results for the Edit Distance on Real Sequences. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

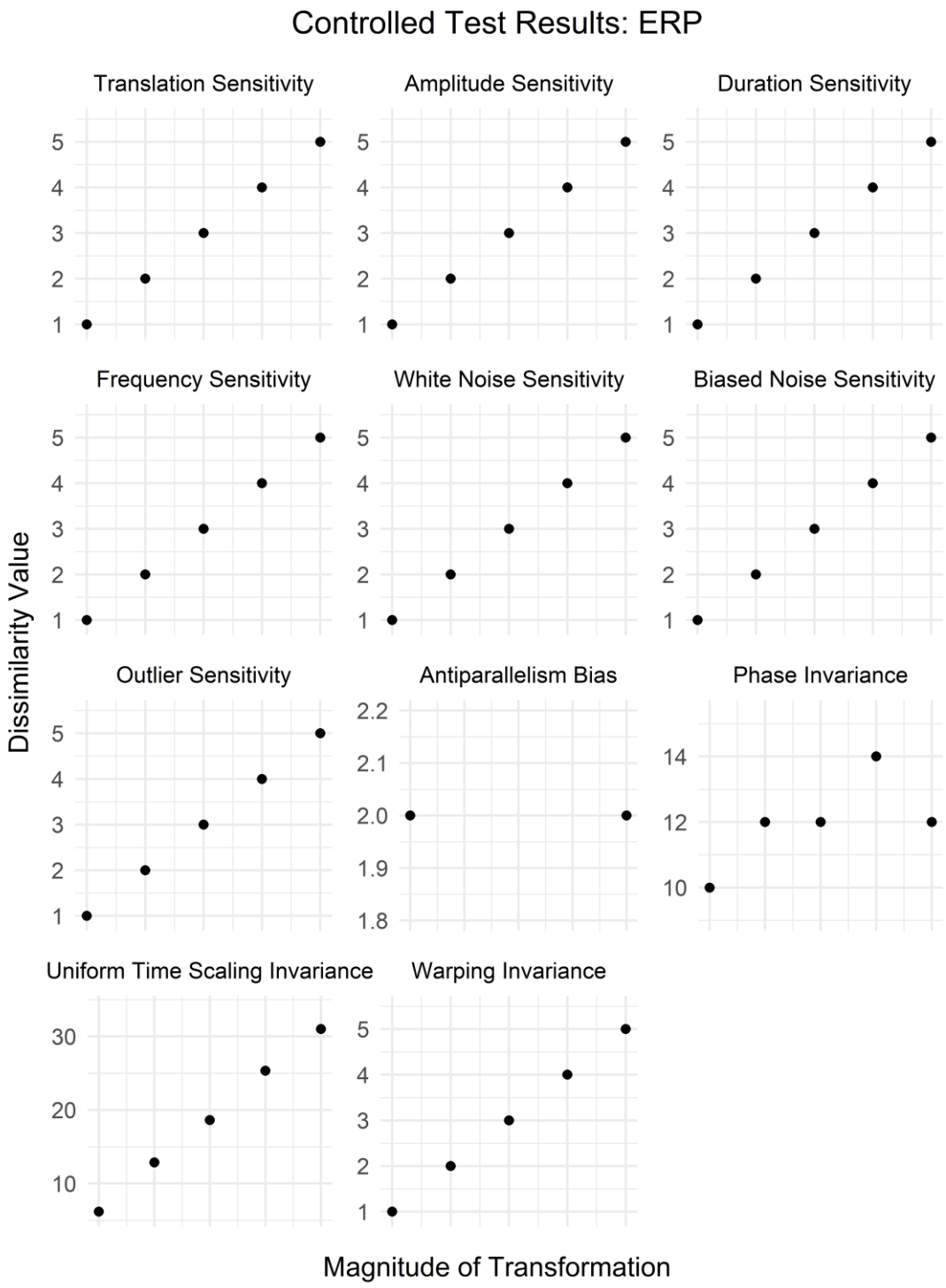


Figure S2.20. Controlled testing results for the Edit Distance with Real Penalty. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

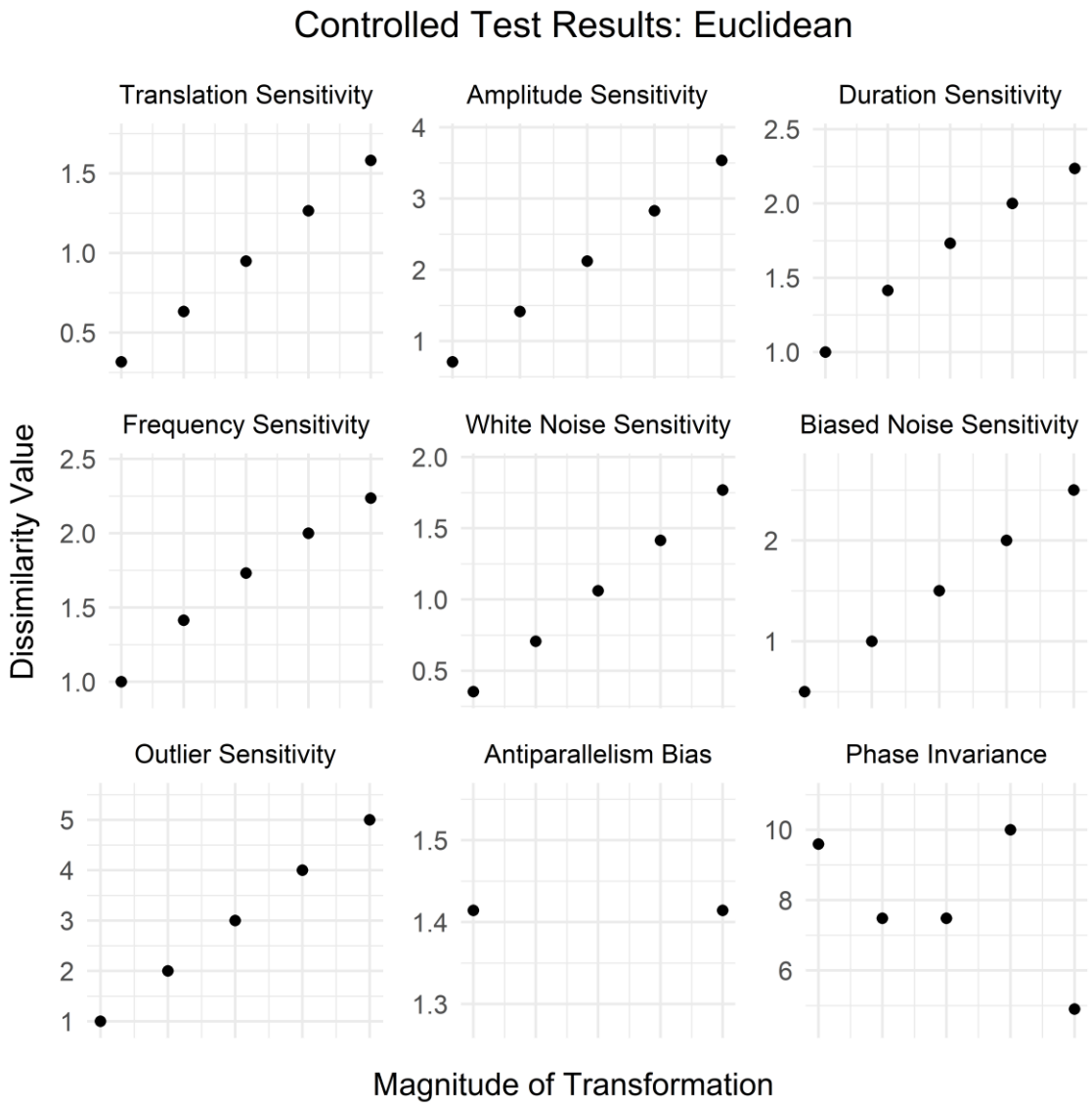


Figure S2.21. Controlled testing results for the Euclidean Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

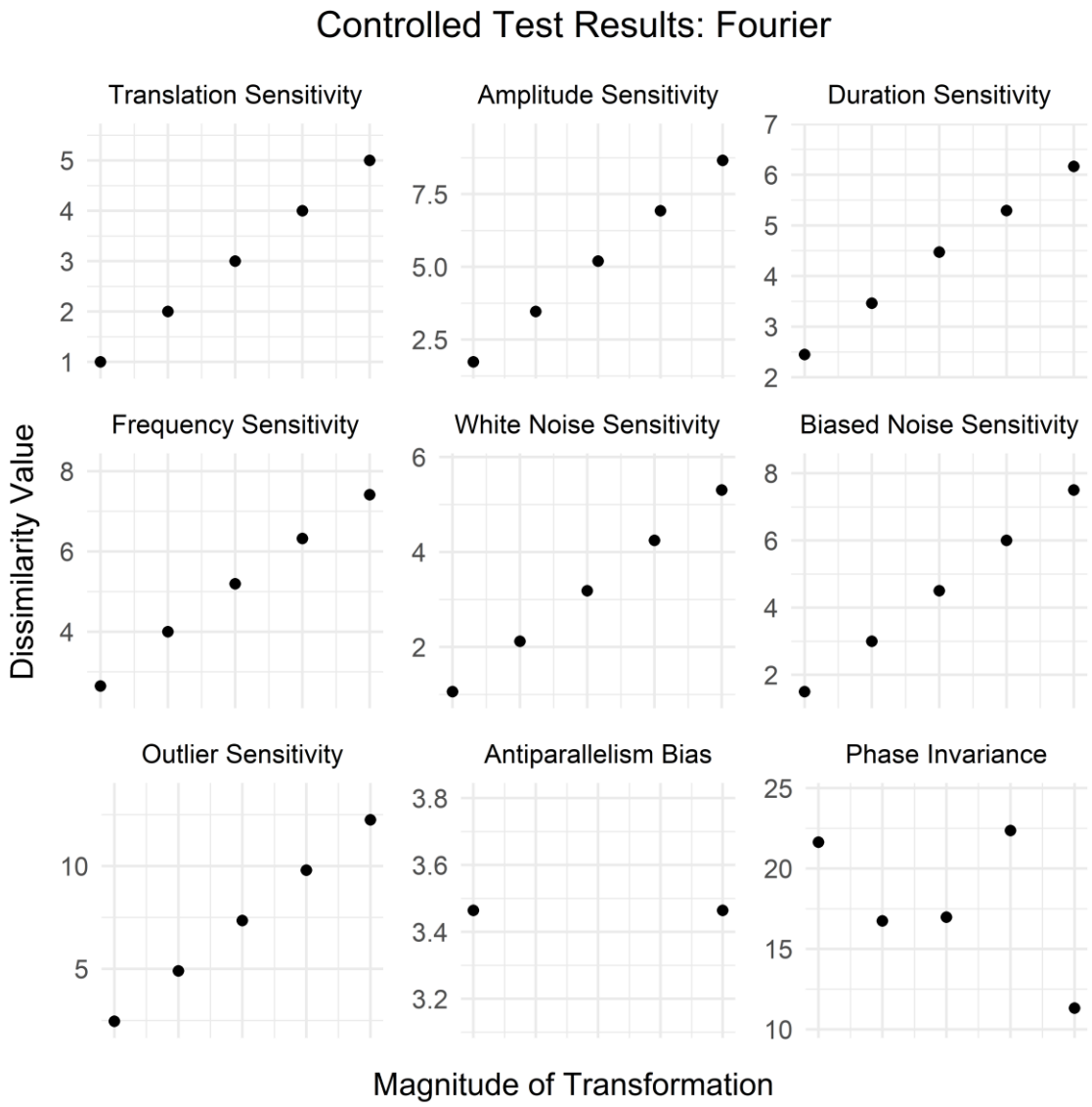


Figure S2.22. Controlled testing results for the Fourier Coefficient-Based Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

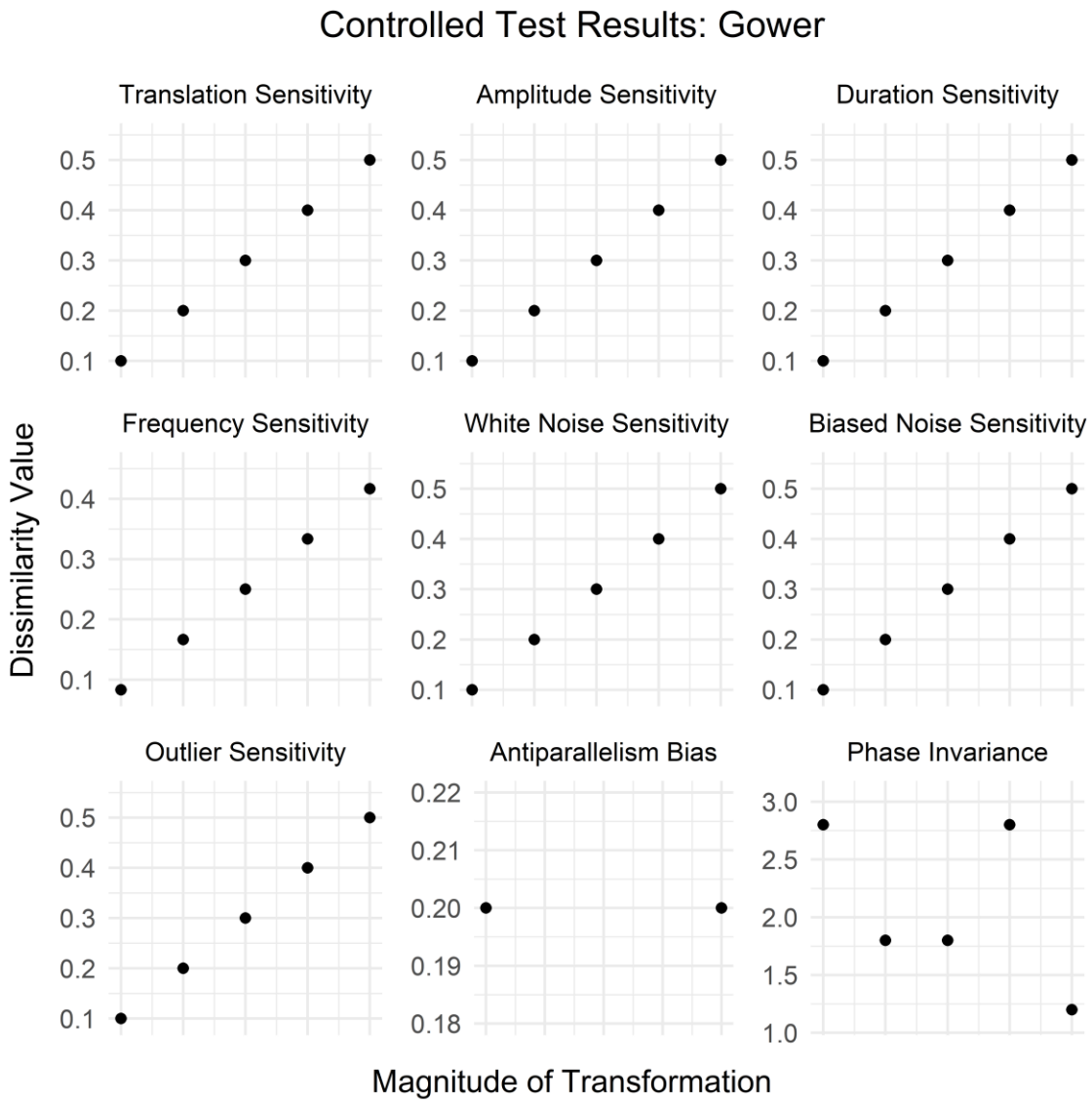


Figure S2.23. Controlled testing results for the Gower Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

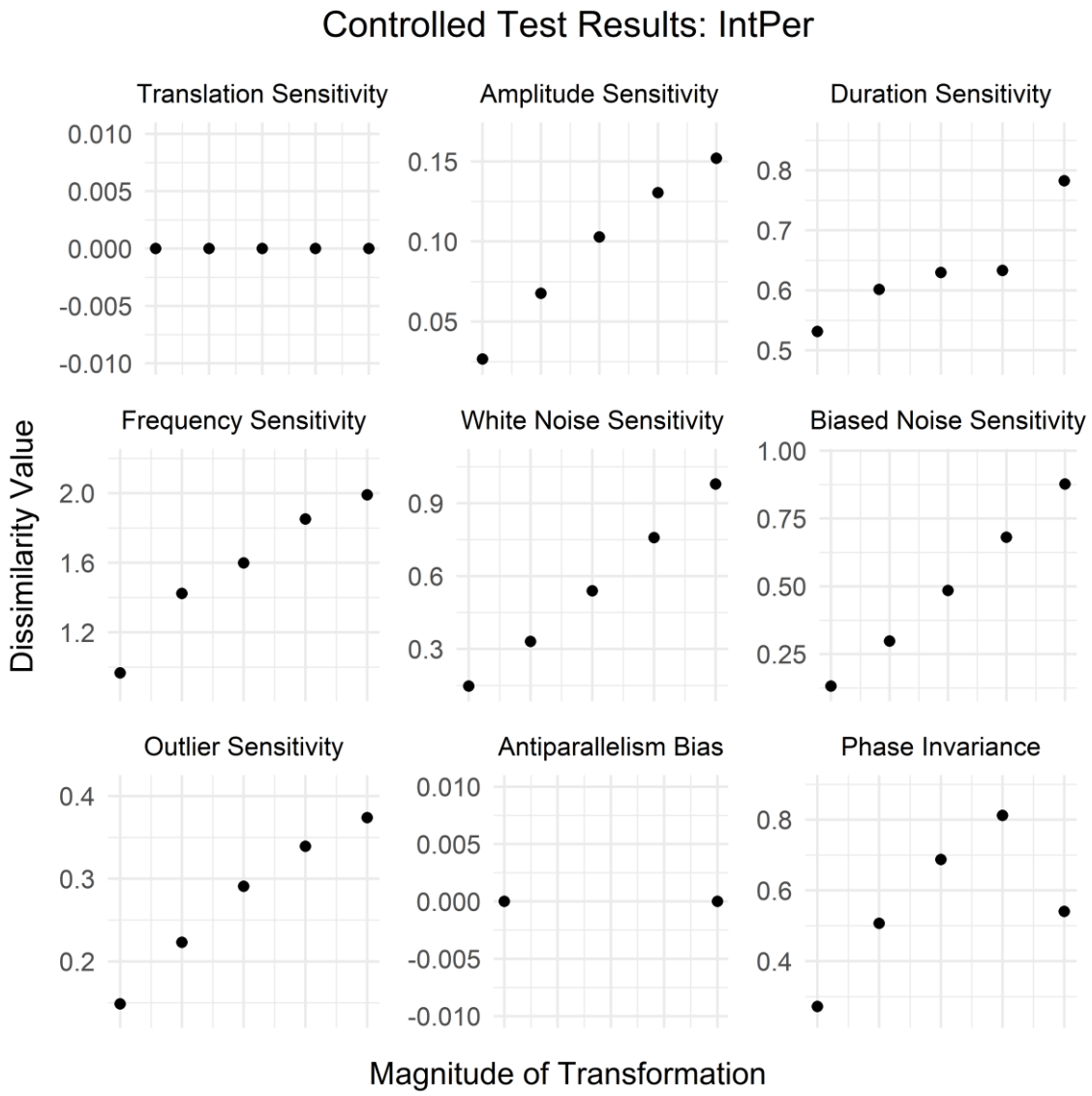


Figure S2.24. Controlled testing results for the Integrated Periodogram Based Dissimilarity. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

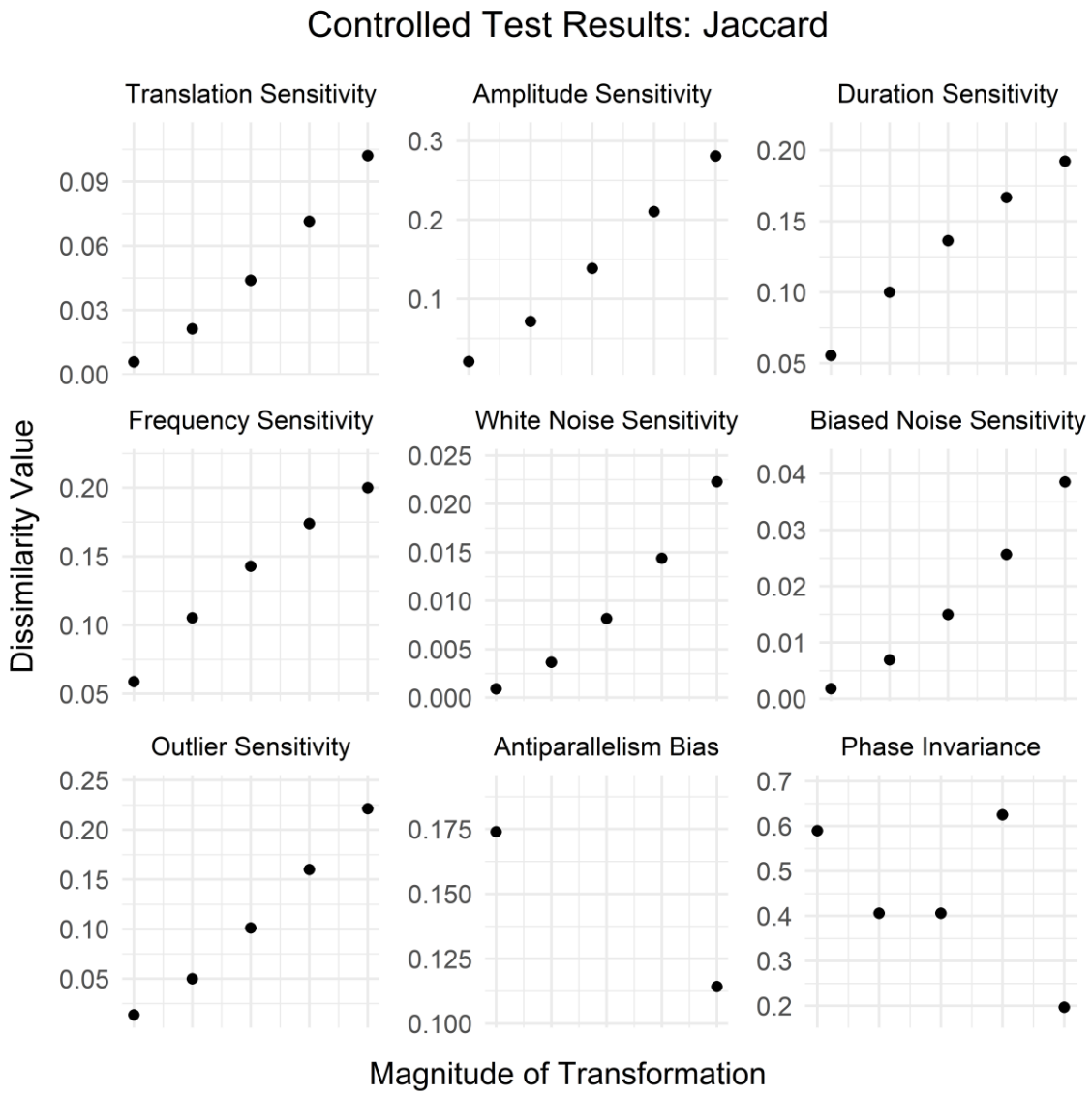


Figure S2.25. Controlled testing results for the Jaccard Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

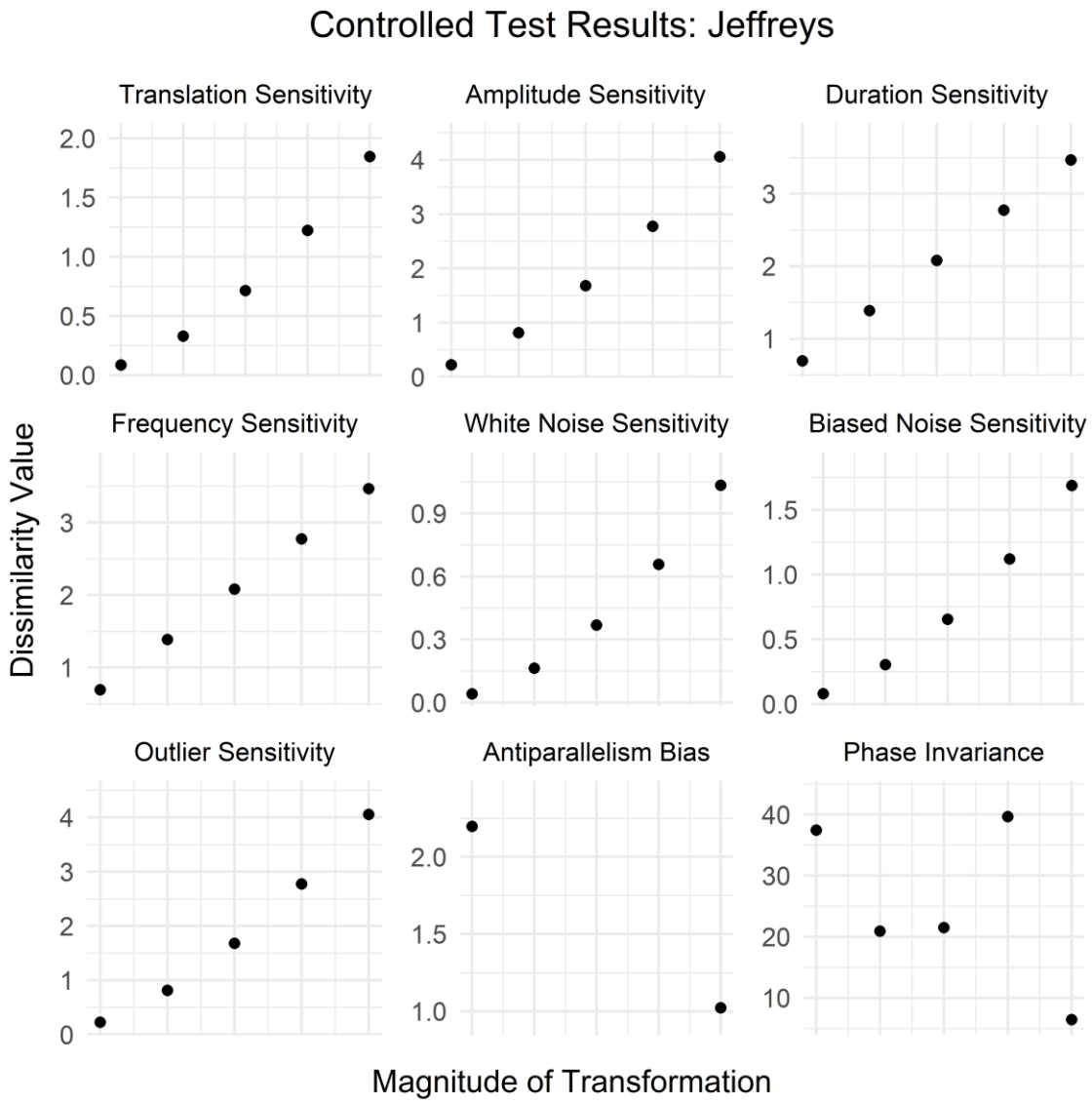


Figure S2.26. Controlled testing results for the Jeffreys Divergence. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

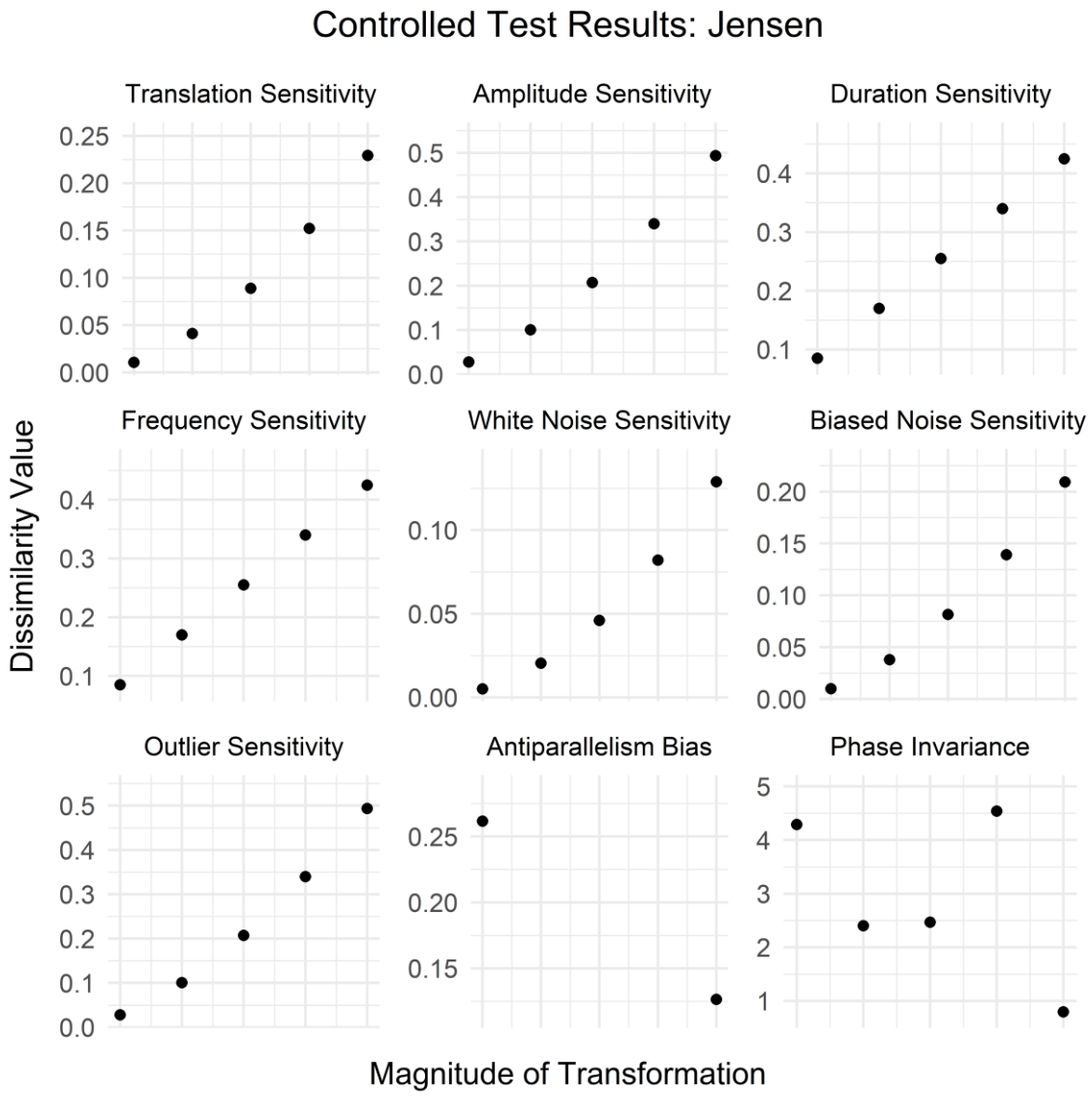


Figure S2.27. Controlled testing results for the Jensen Difference. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

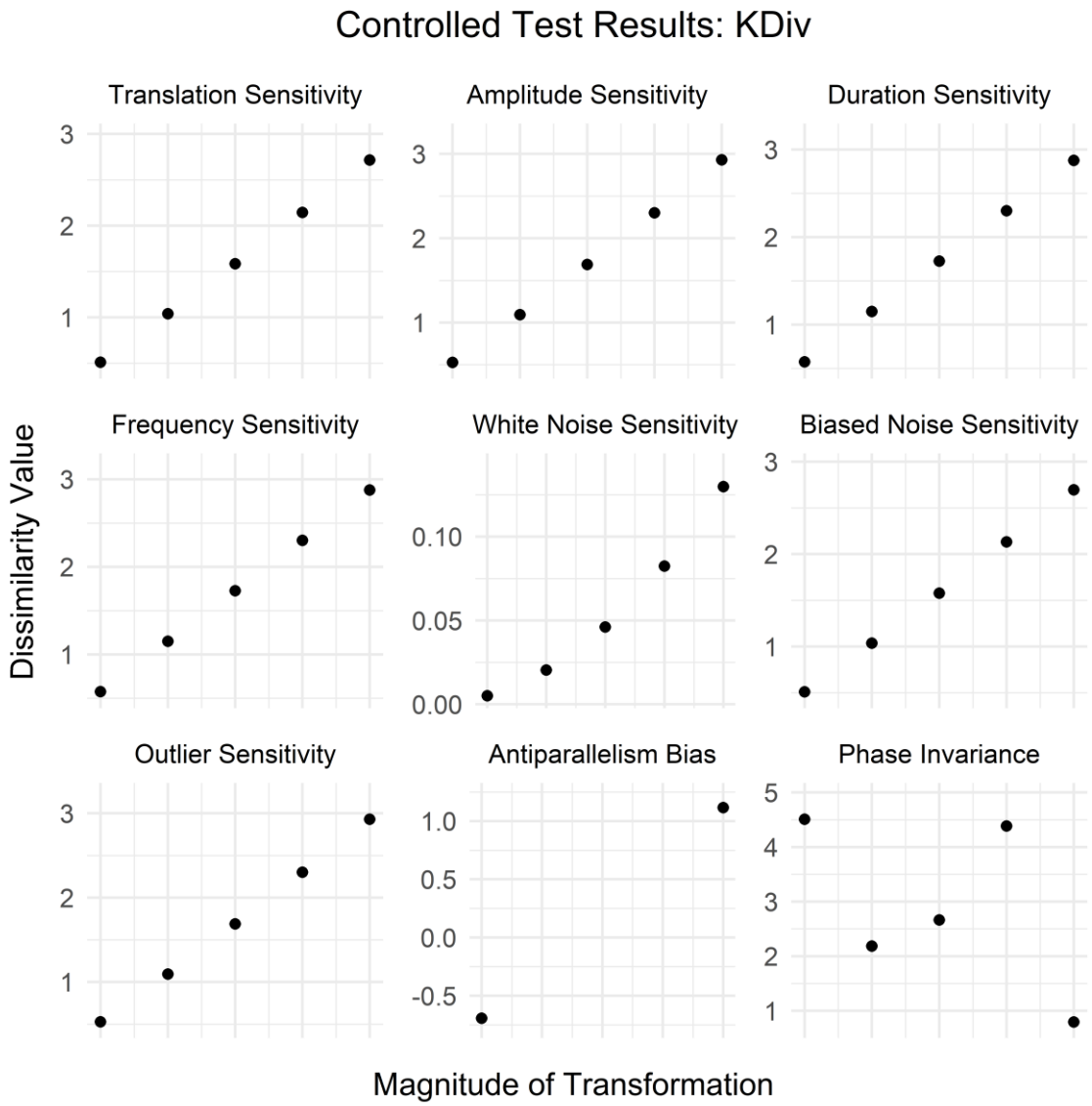


Figure S2.28. Controlled testing results for the K Divergence. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

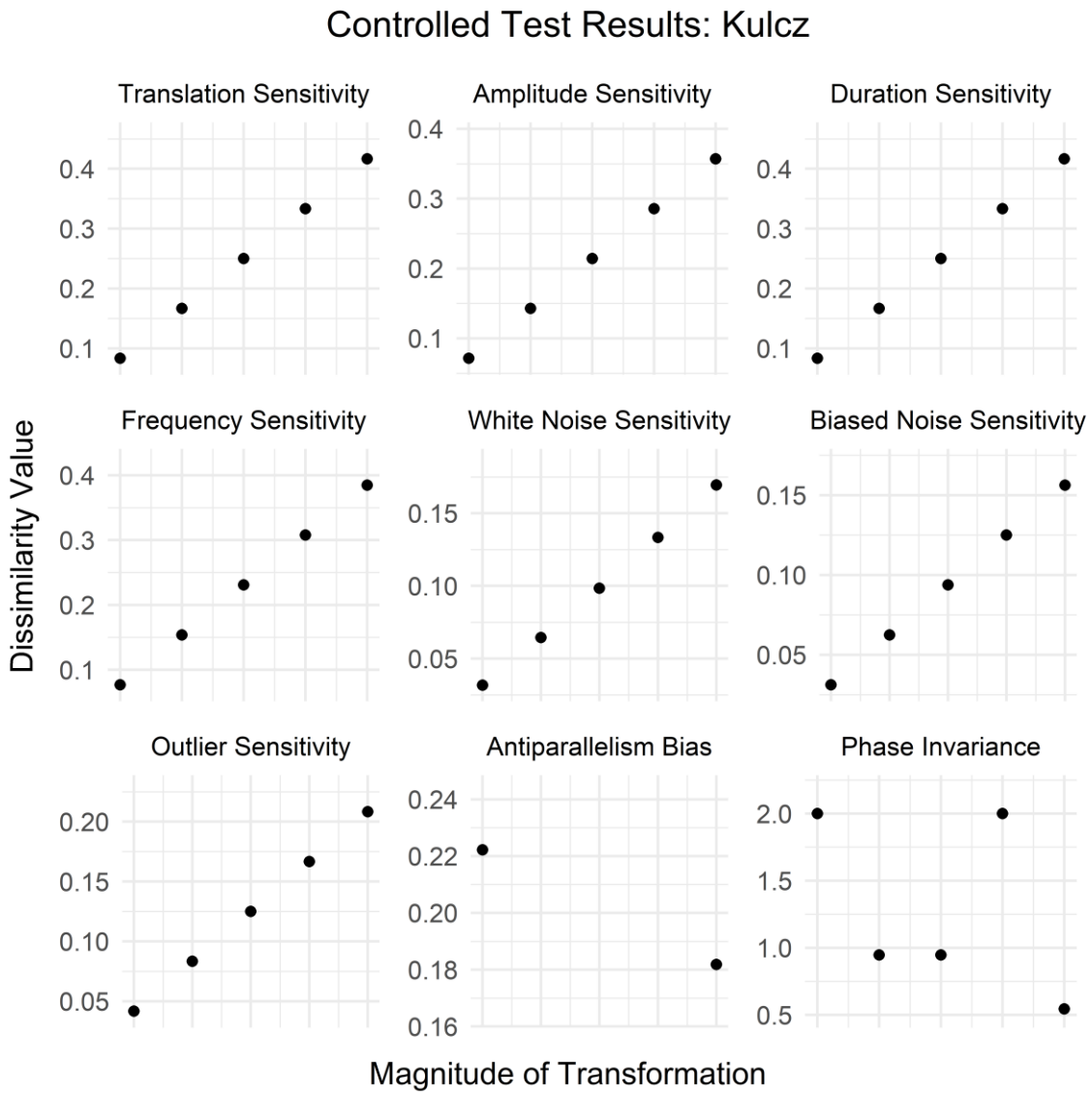


Figure S2.29. Controlled testing results for the Kulczynski Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

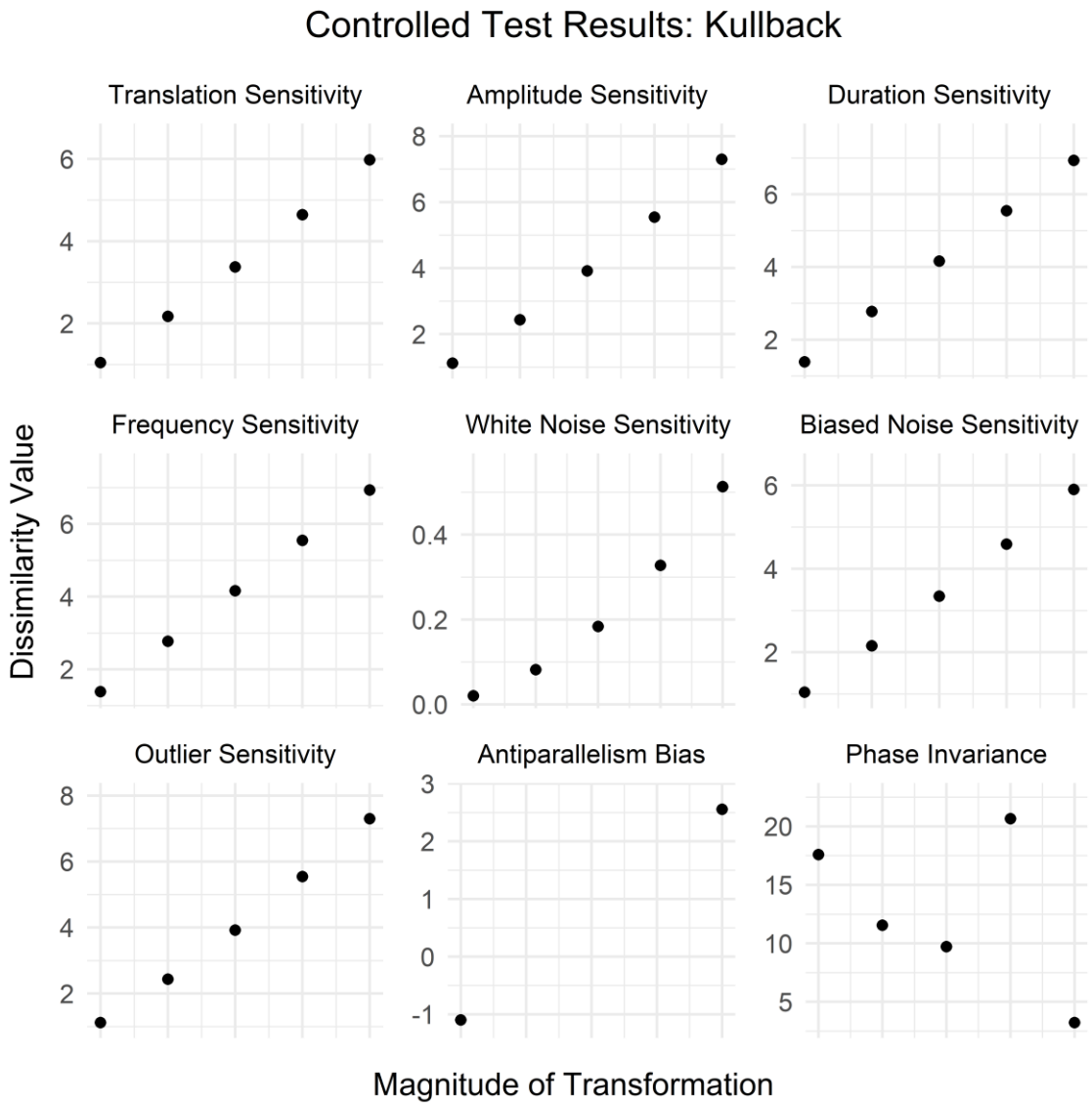


Figure S2.30. Controlled testing results for the Kullback-Leibler Divergence. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

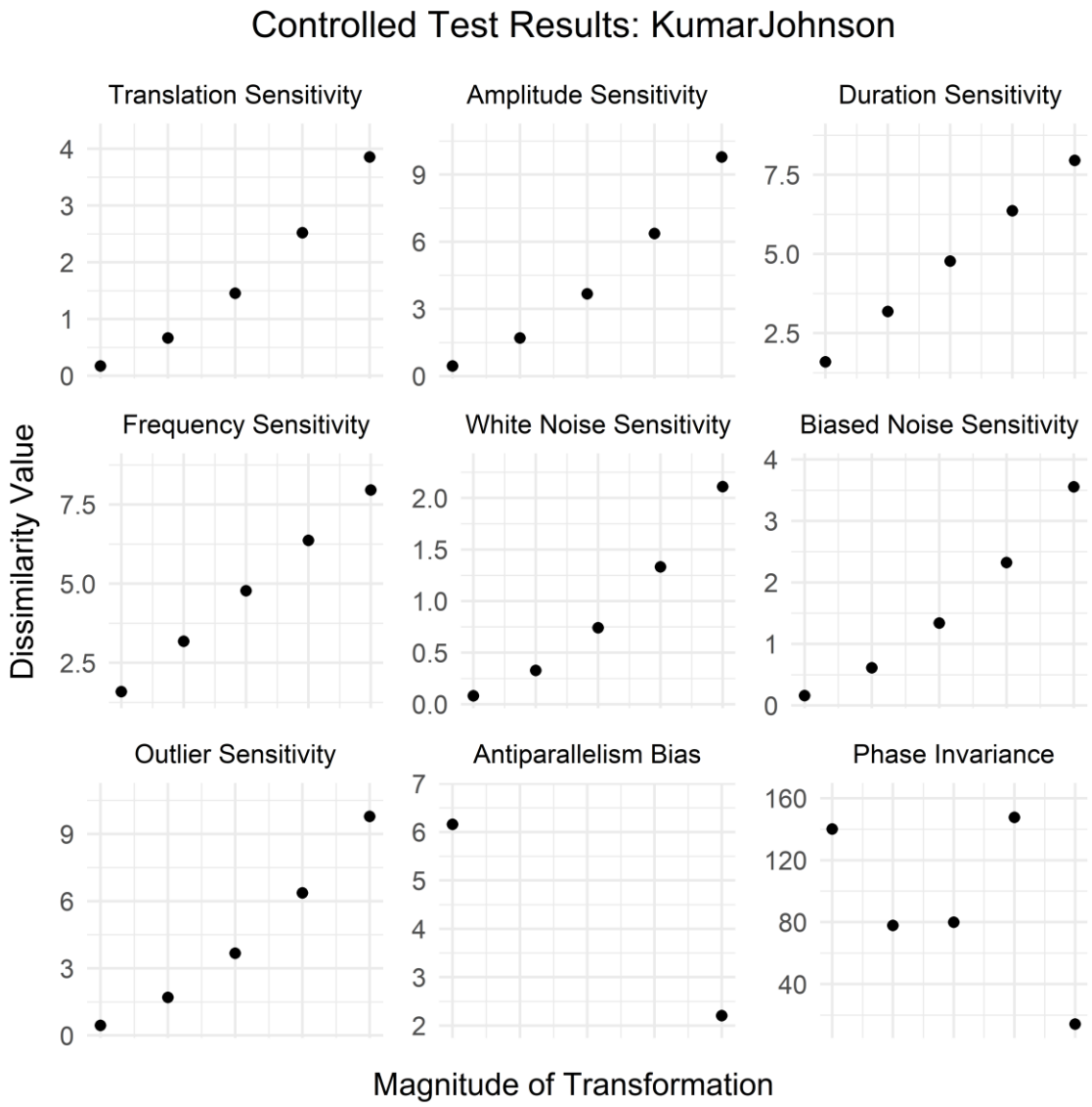


Figure S2.31. Controlled testing results for the Kumar-Johnson Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

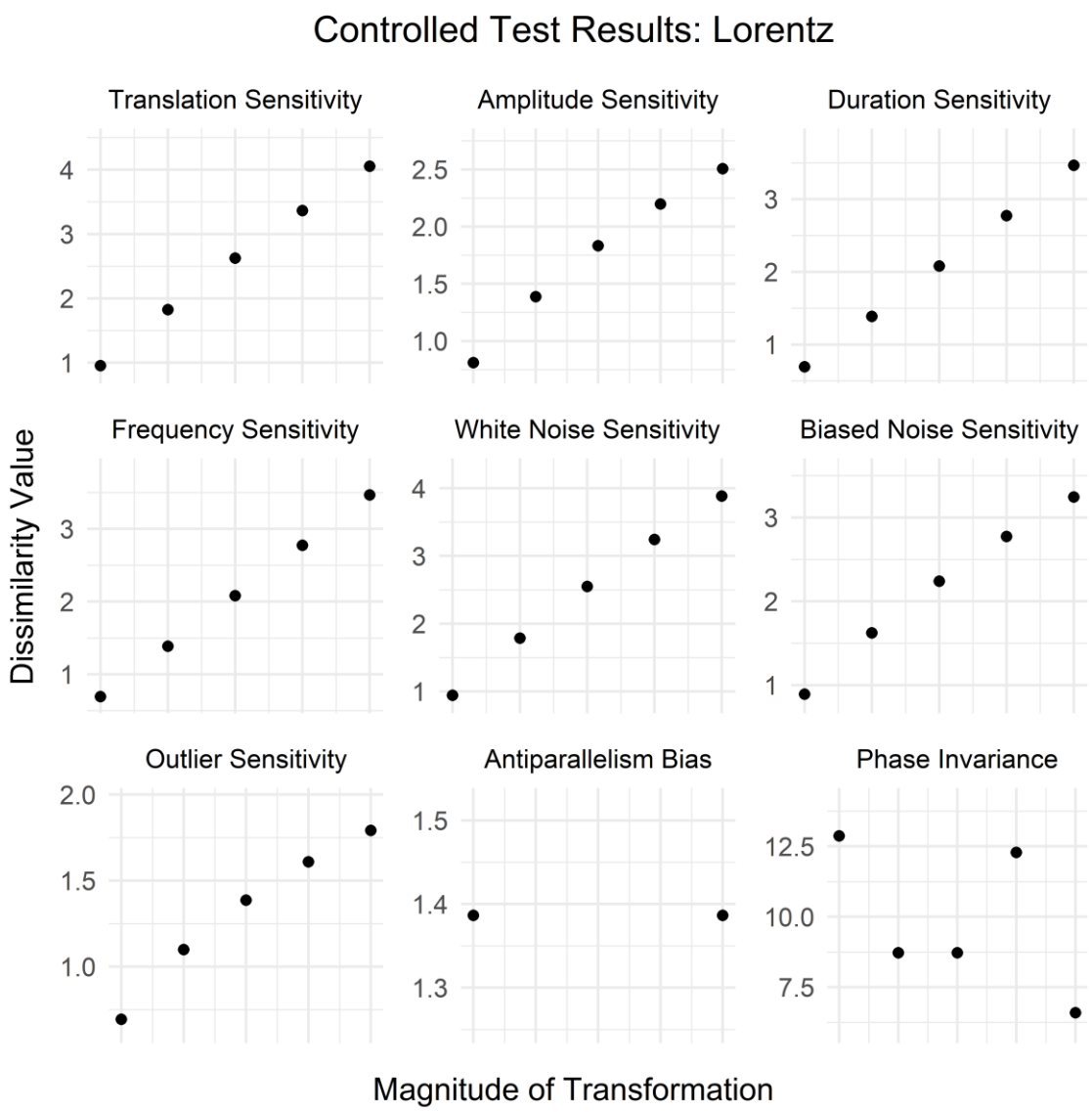


Figure S2.32. Controlled testing results for the Lorentzian Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

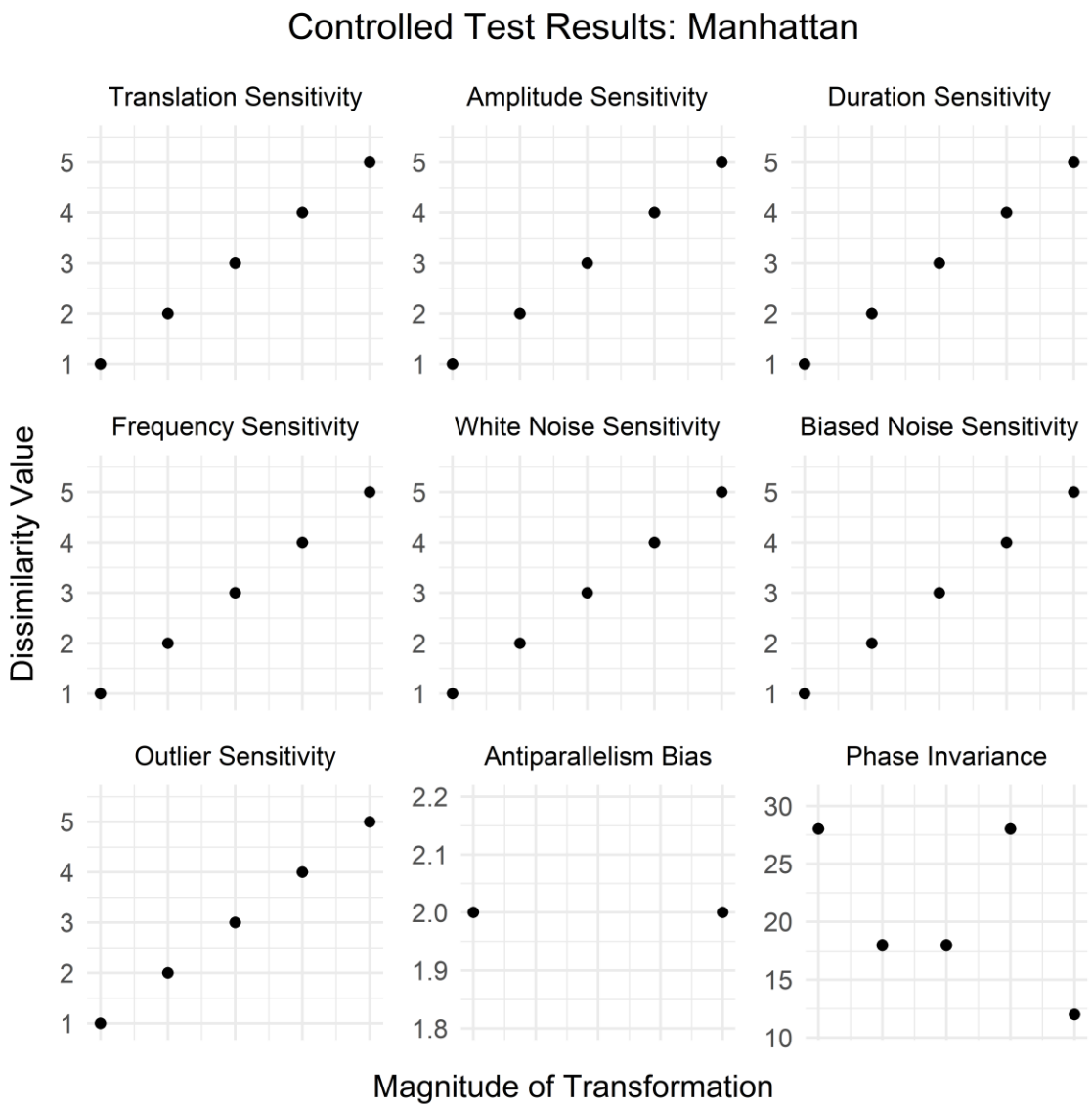


Figure S2.33. Controlled testing results for the Manhattan Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

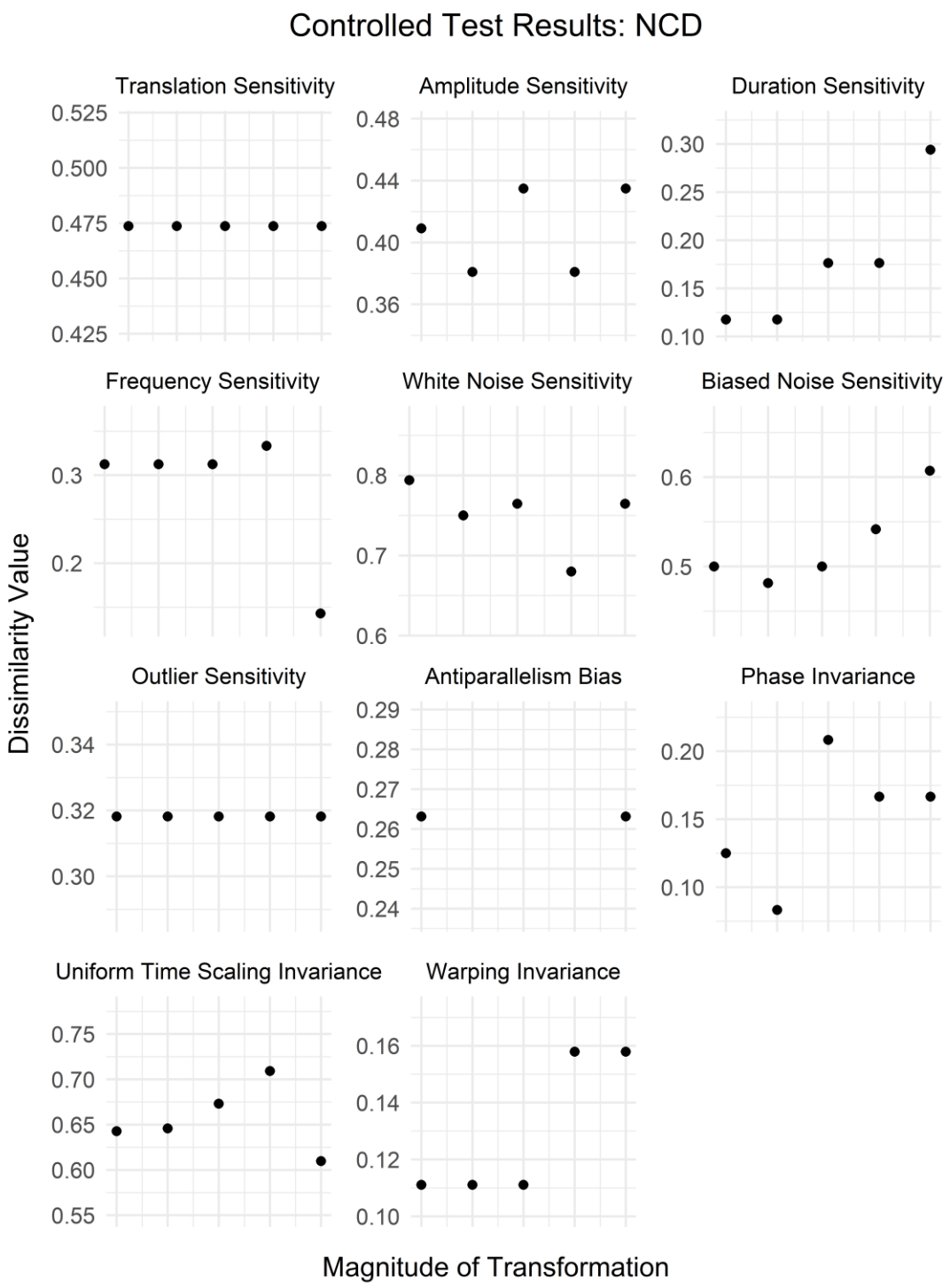


Figure S2.34. Controlled testing results for the Normalized Compression Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

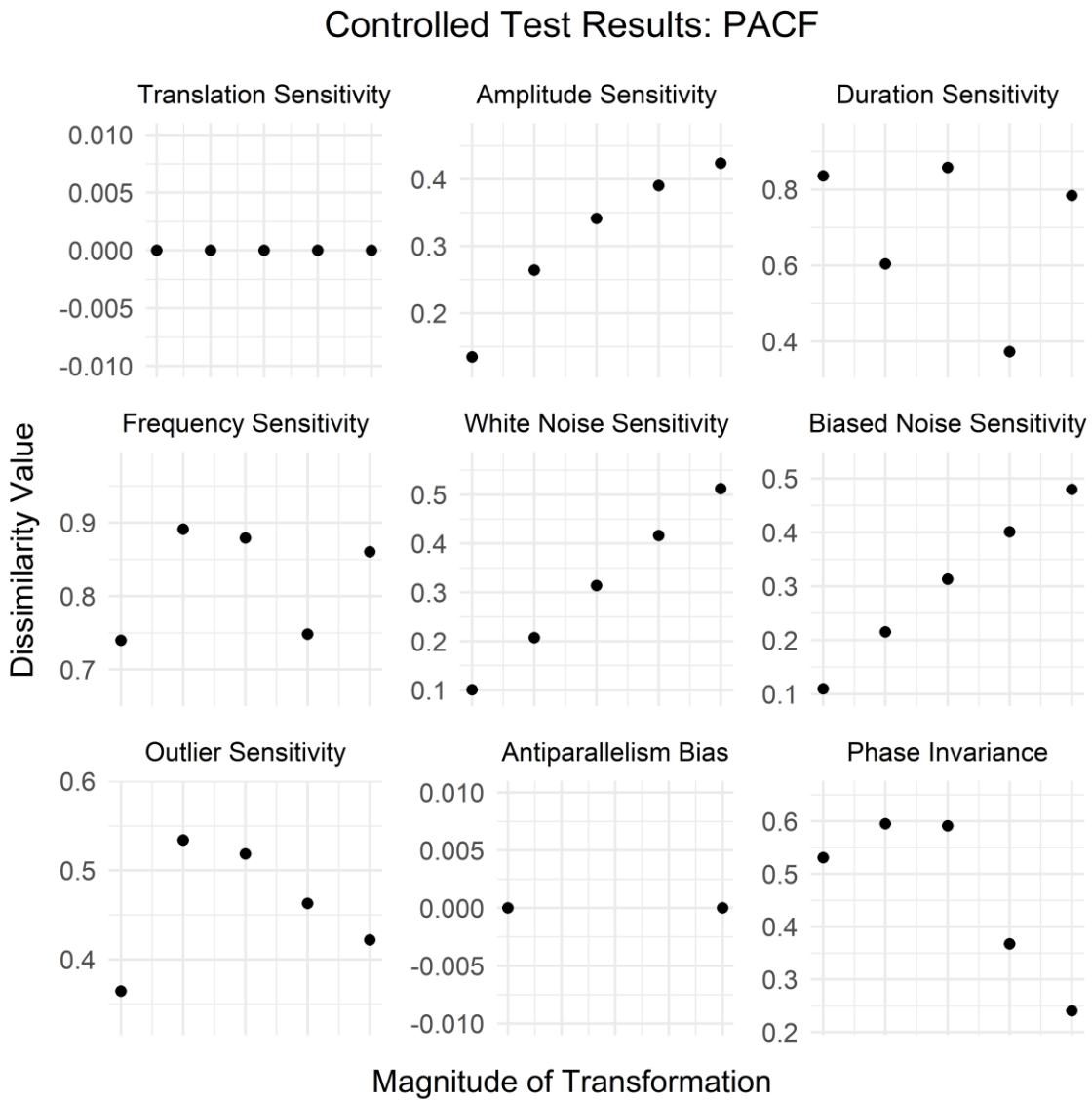


Figure S2.35. Controlled testing results for the Partial Autocorrelation-Based Dissimilarity. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

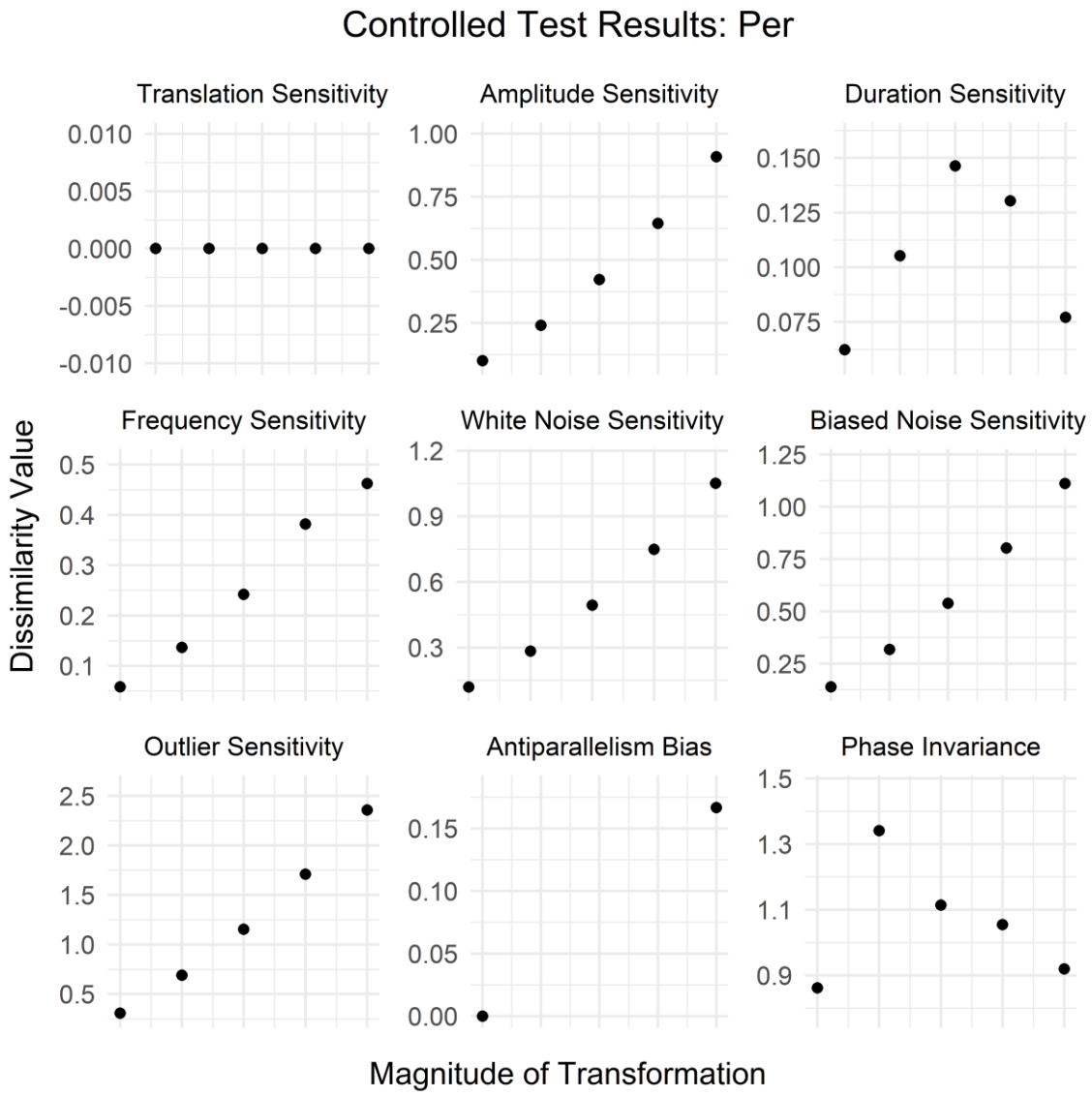


Figure S2.36. Controlled testing results for the Periodogram-Based Dissimilarity. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

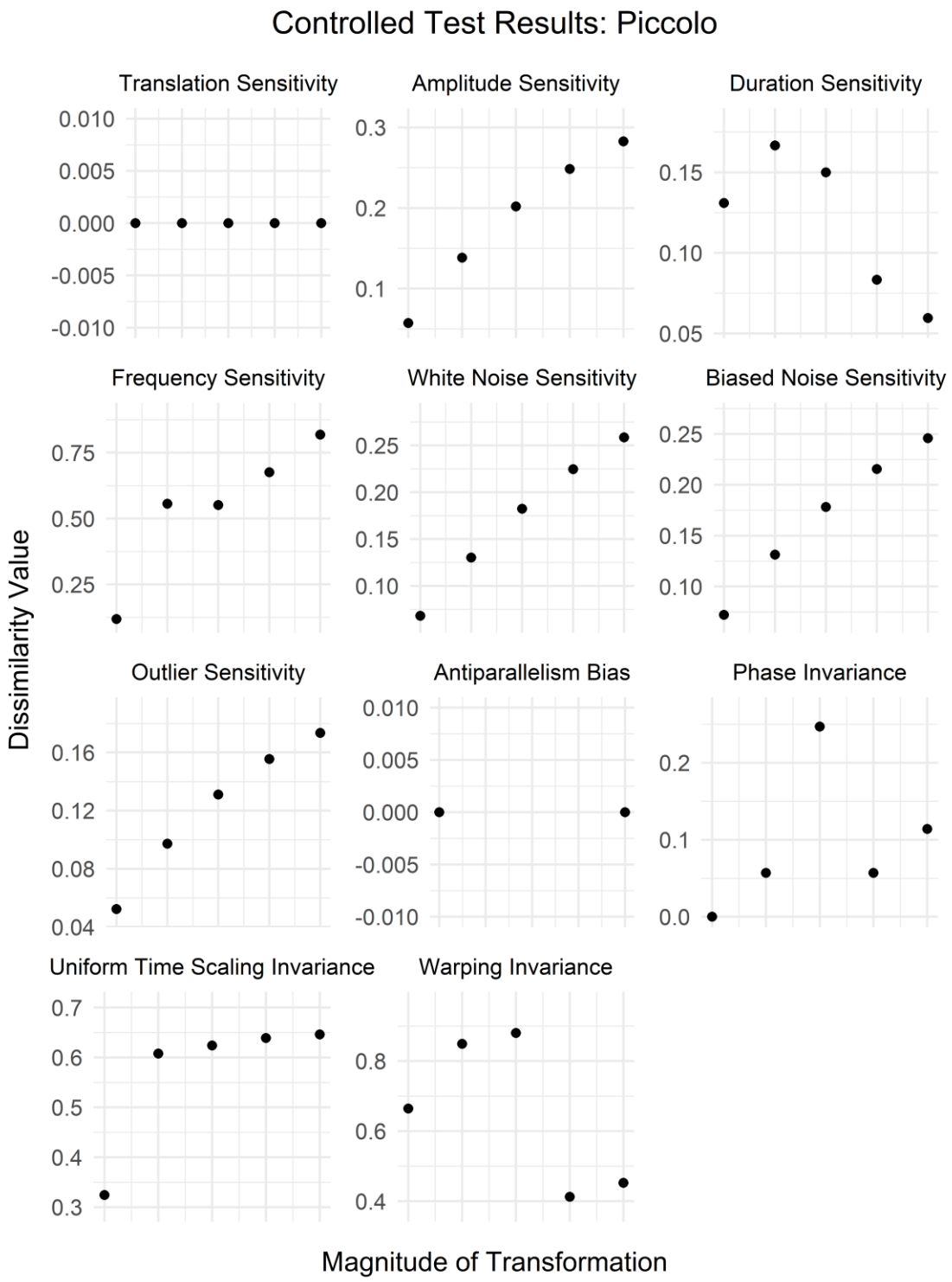


Figure S2.37. Controlled testing results for the Piccolo Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

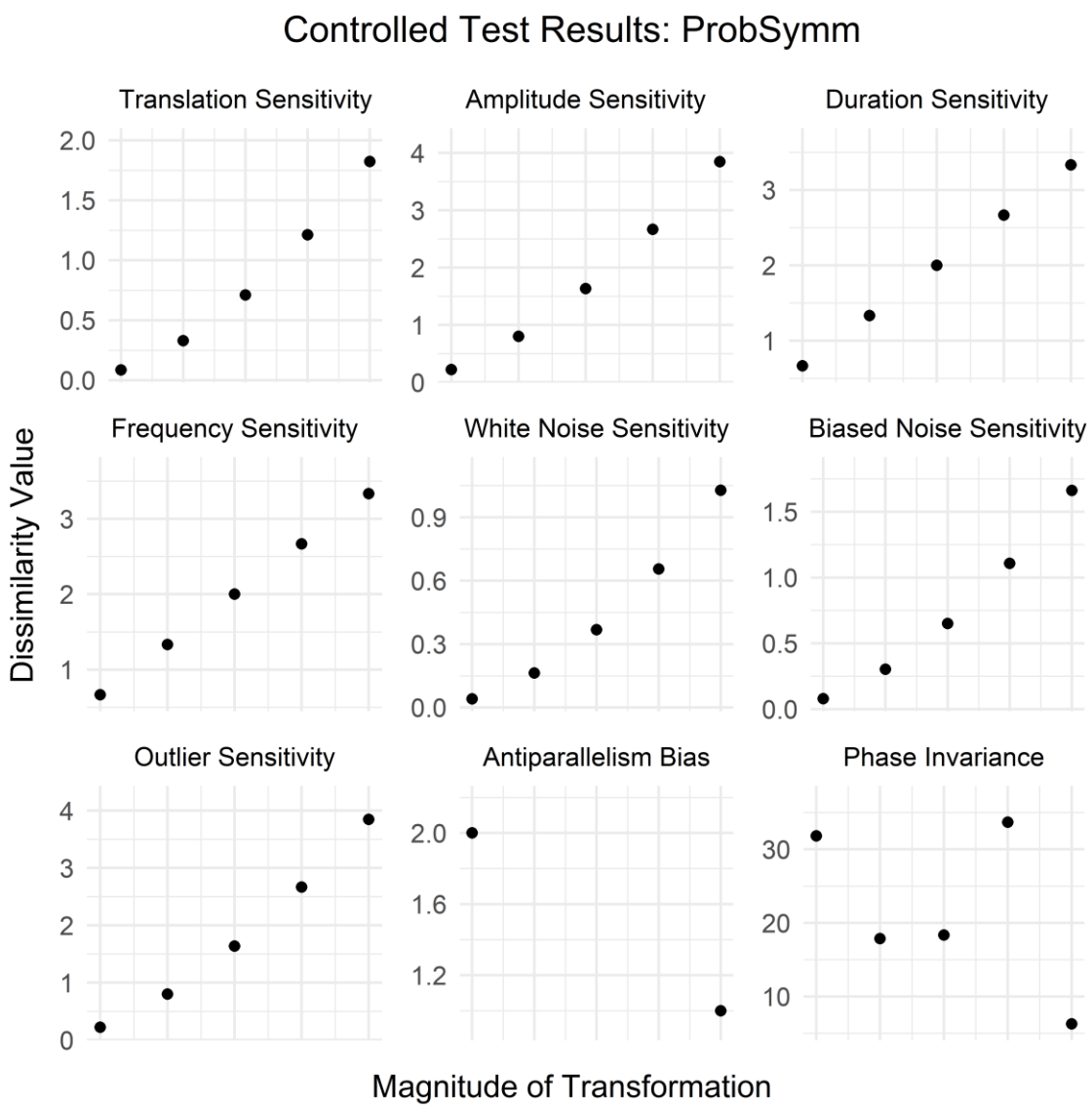


Figure S2.38. Controlled testing results for the Probabilistic Symmetric Chi-Squared Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

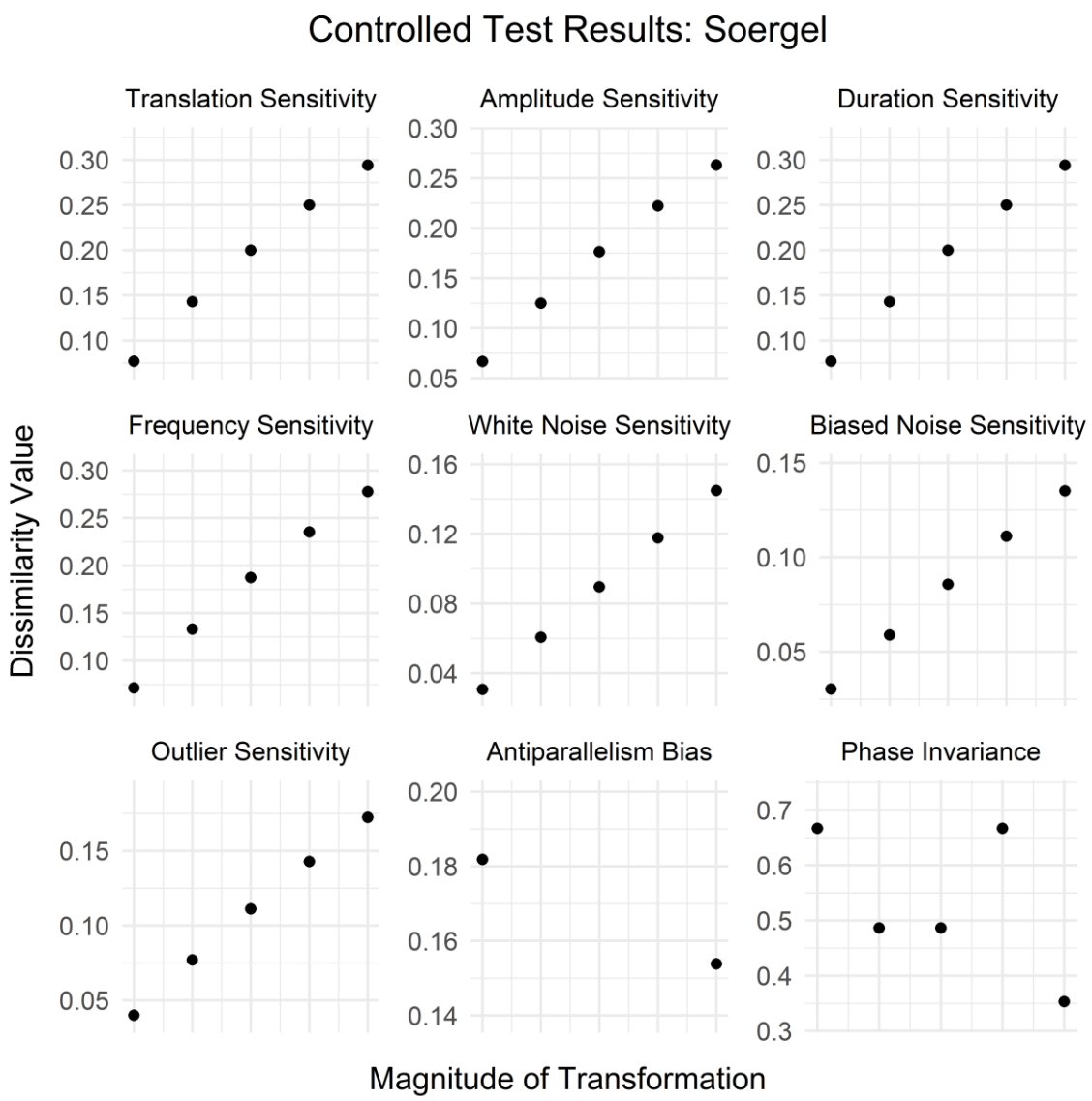


Figure S2.39. Controlled testing results for the Soergel Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

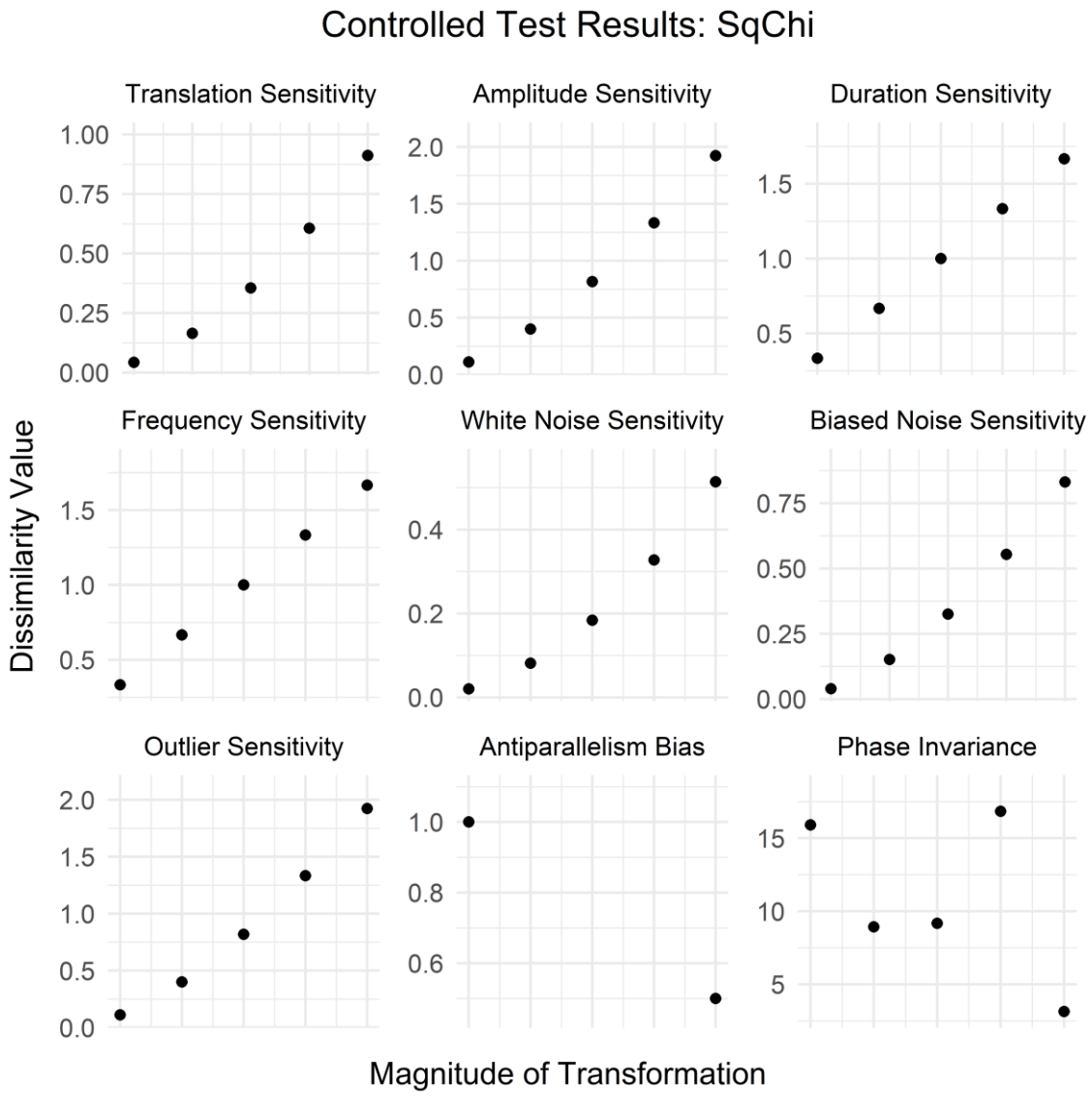


Figure S2.40. Controlled testing results for the Squared Chi-Squared Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

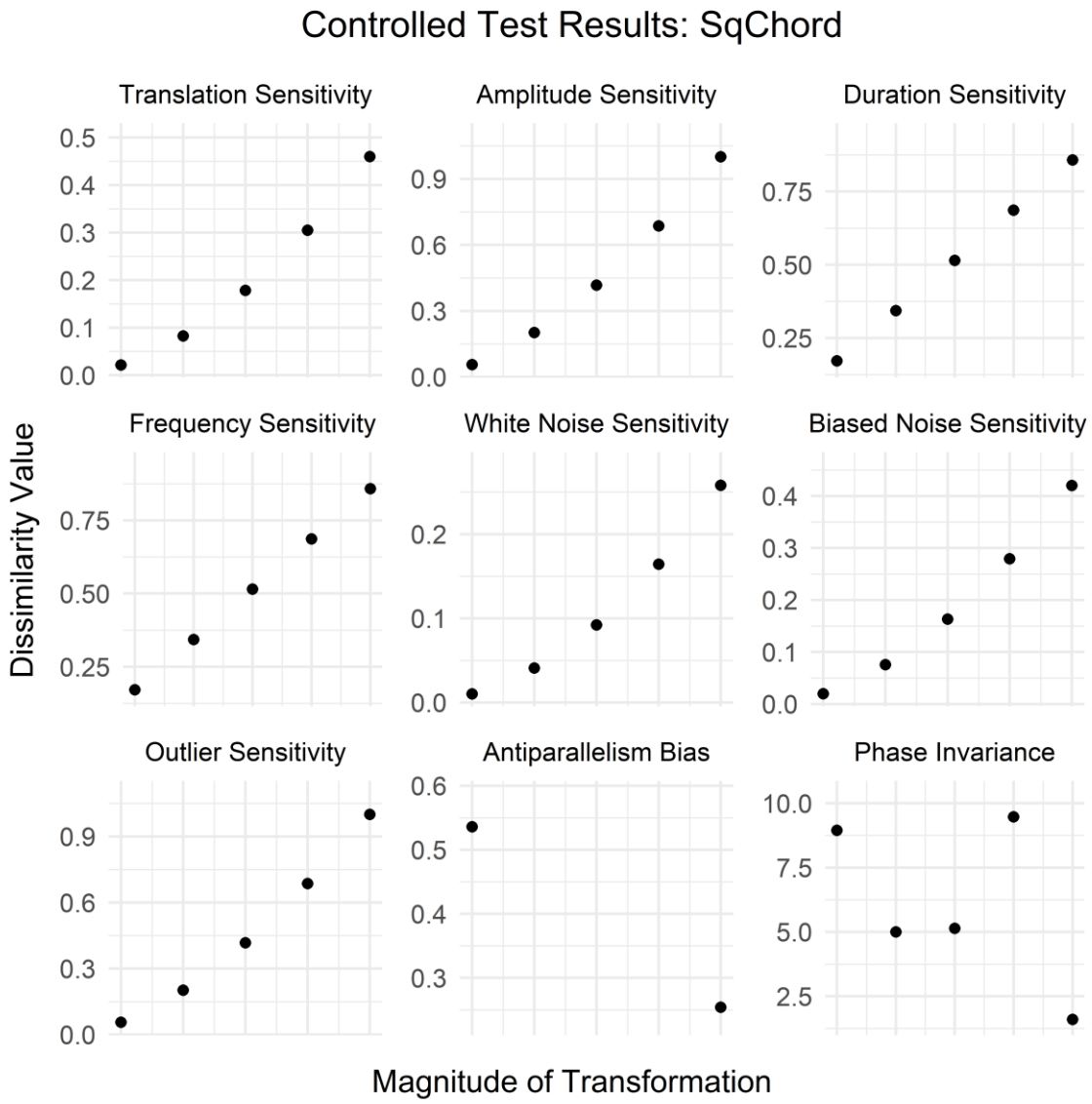


Figure S2.41. Controlled testing results for the Squared-Chord Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

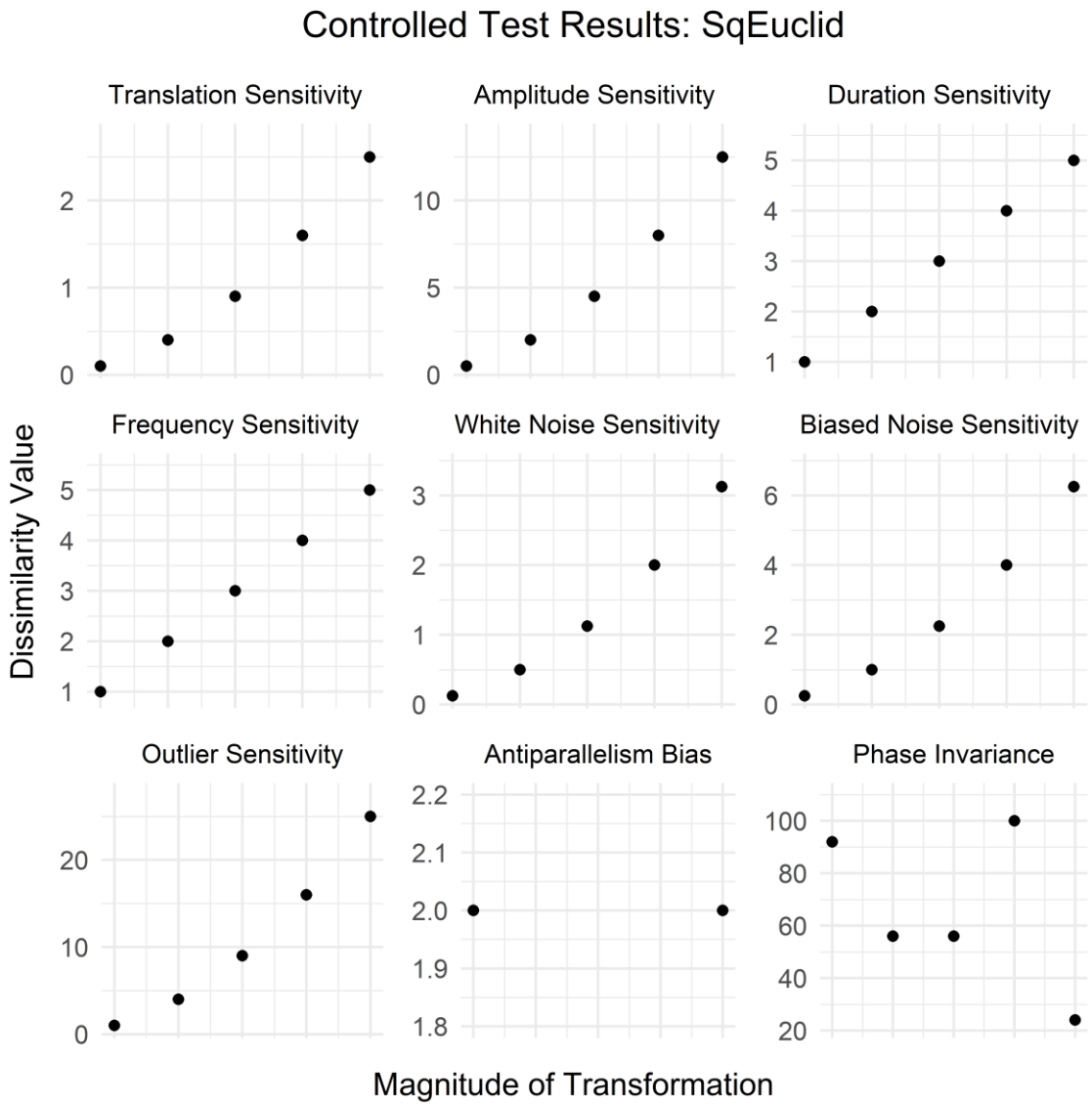


Figure S2.42. Controlled testing results for the Squared Euclidean Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

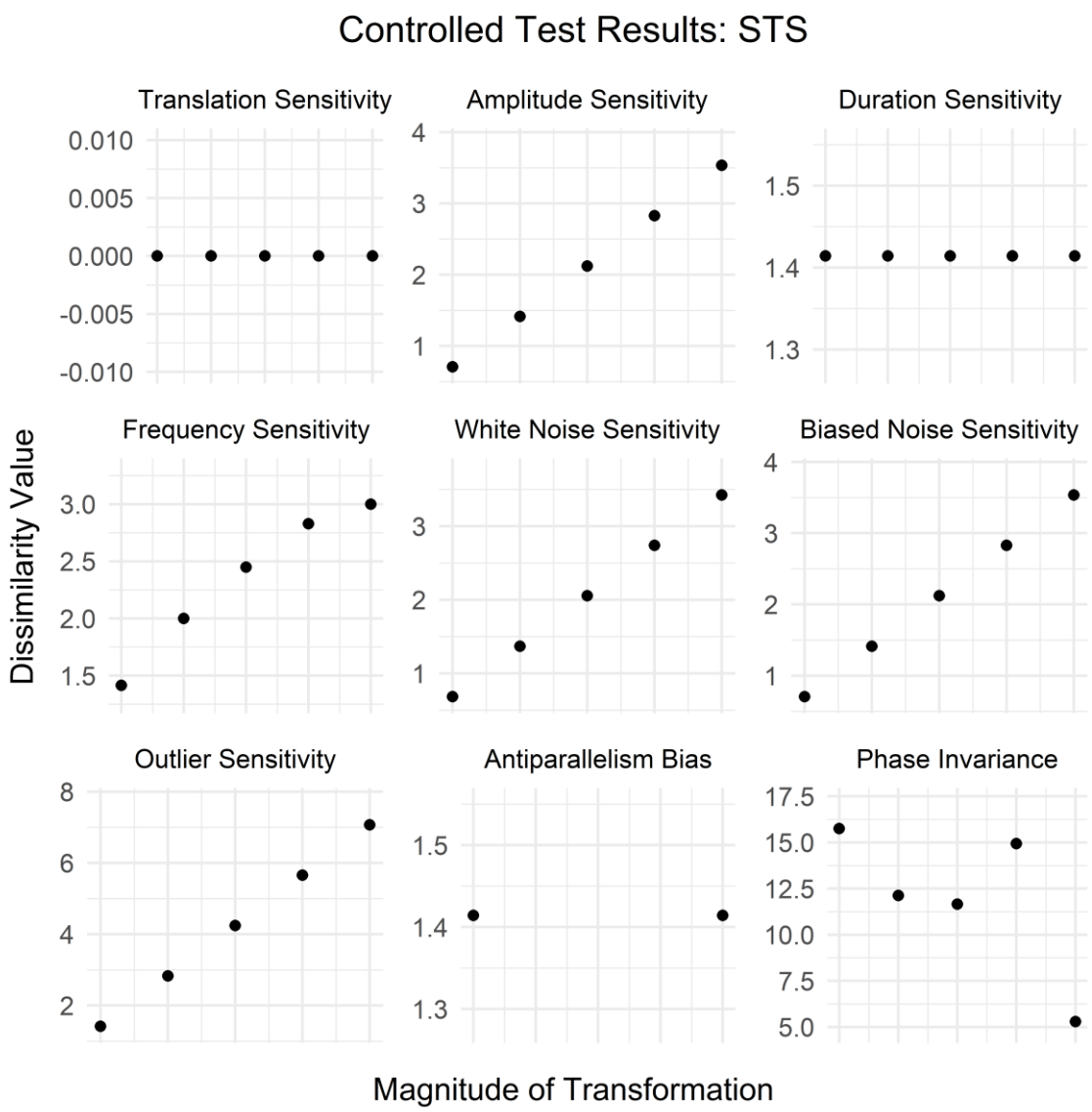


Figure S2.43. Controlled testing results for the Short Time Series Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

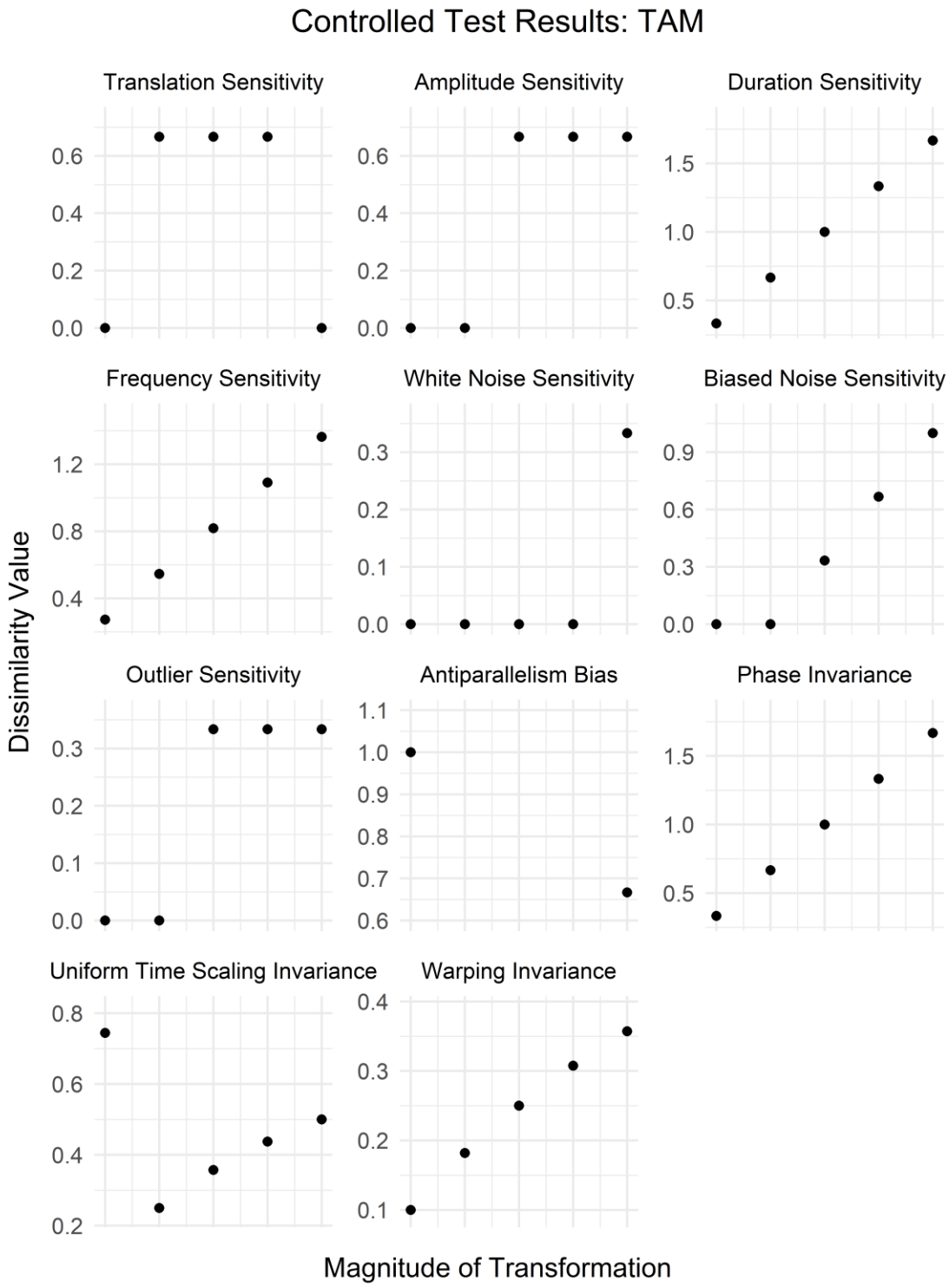


Figure S2.44. Controlled testing results for the Time Alignment Measurement Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance were tested by stretching time series, or parts of time series, respectively, by different amounts.

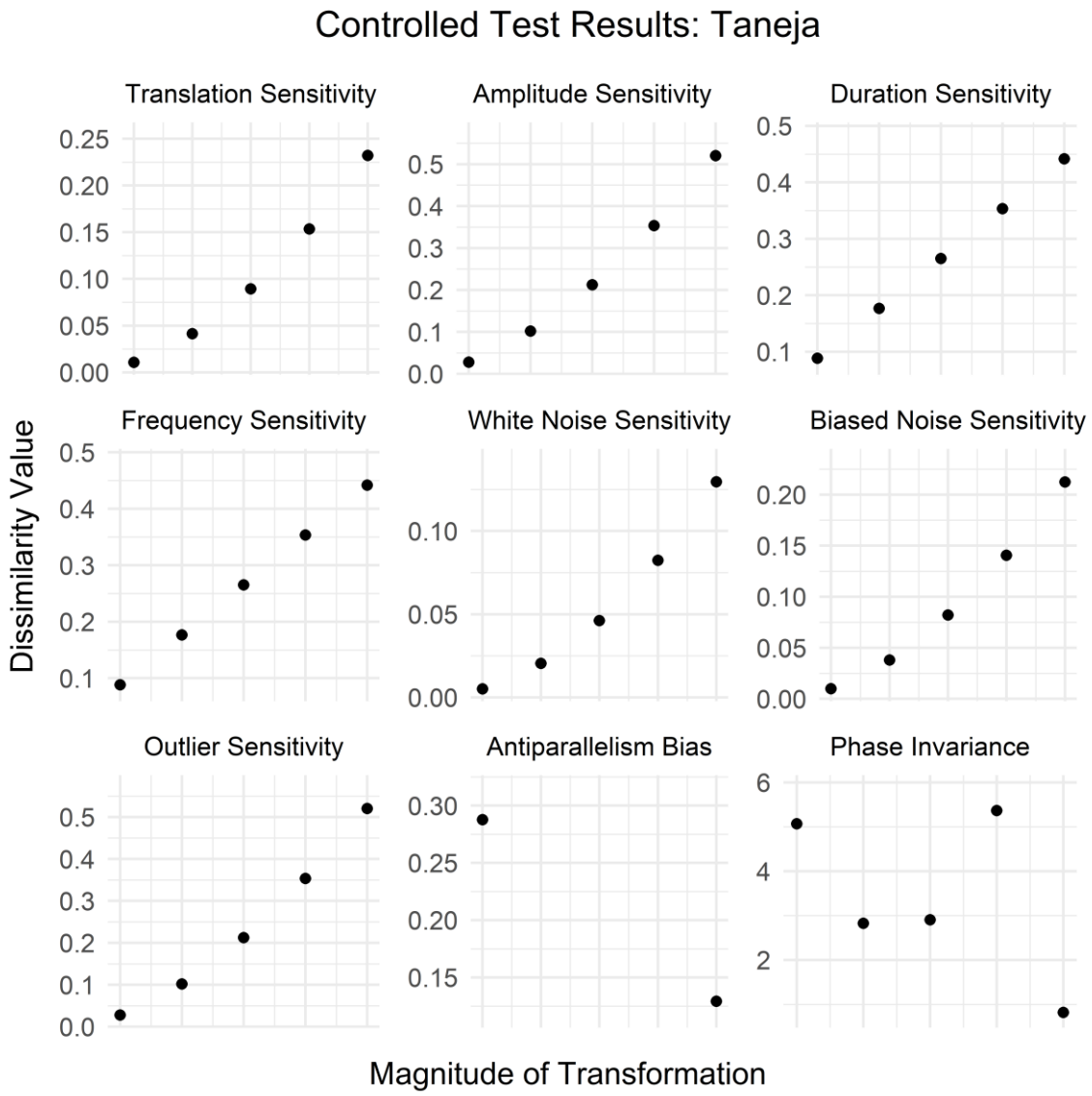


Figure S2.45. Controlled testing results for the Taneja Difference. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

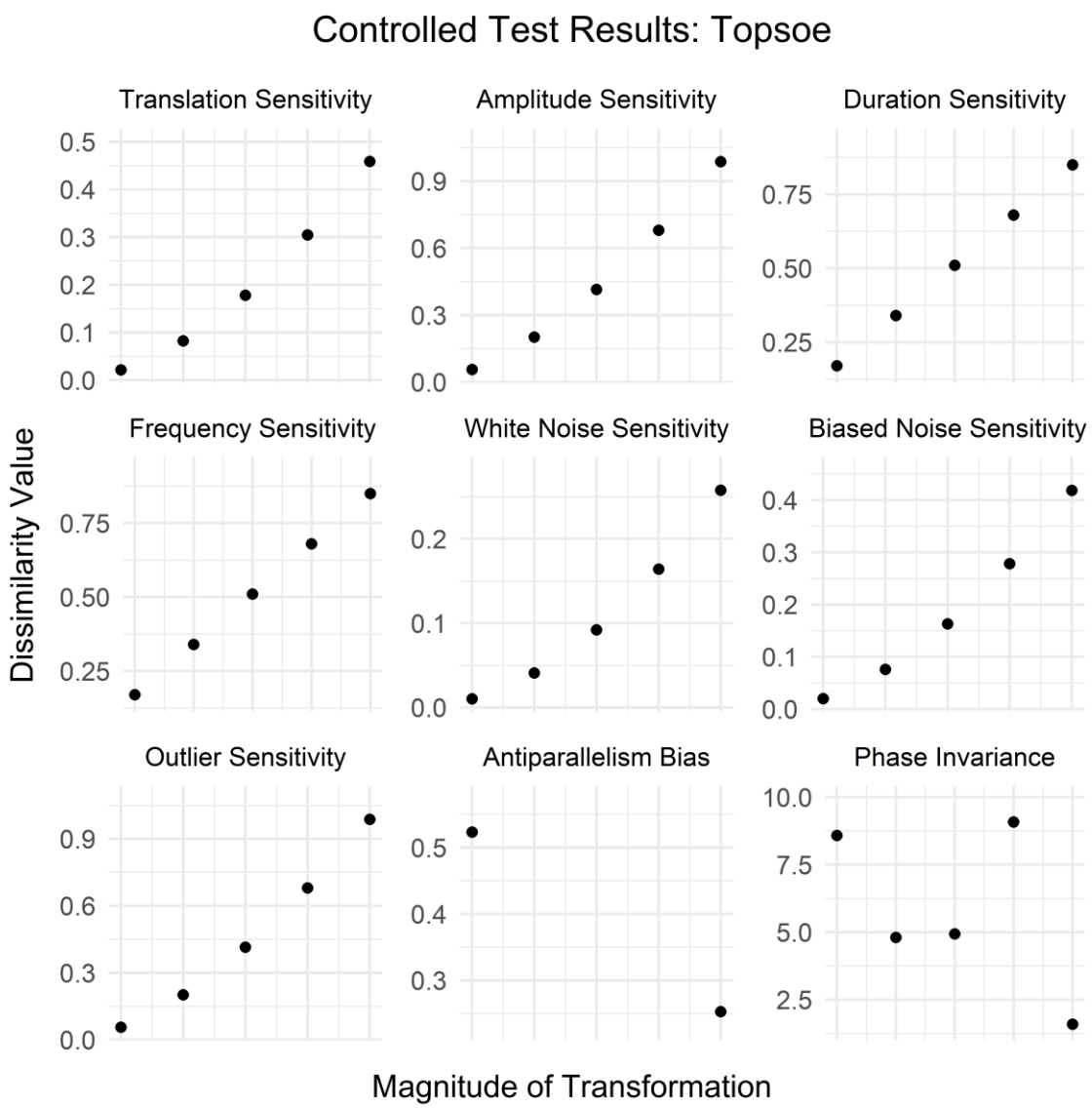


Figure S2.46. Controlled testing results for the Topsoe Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

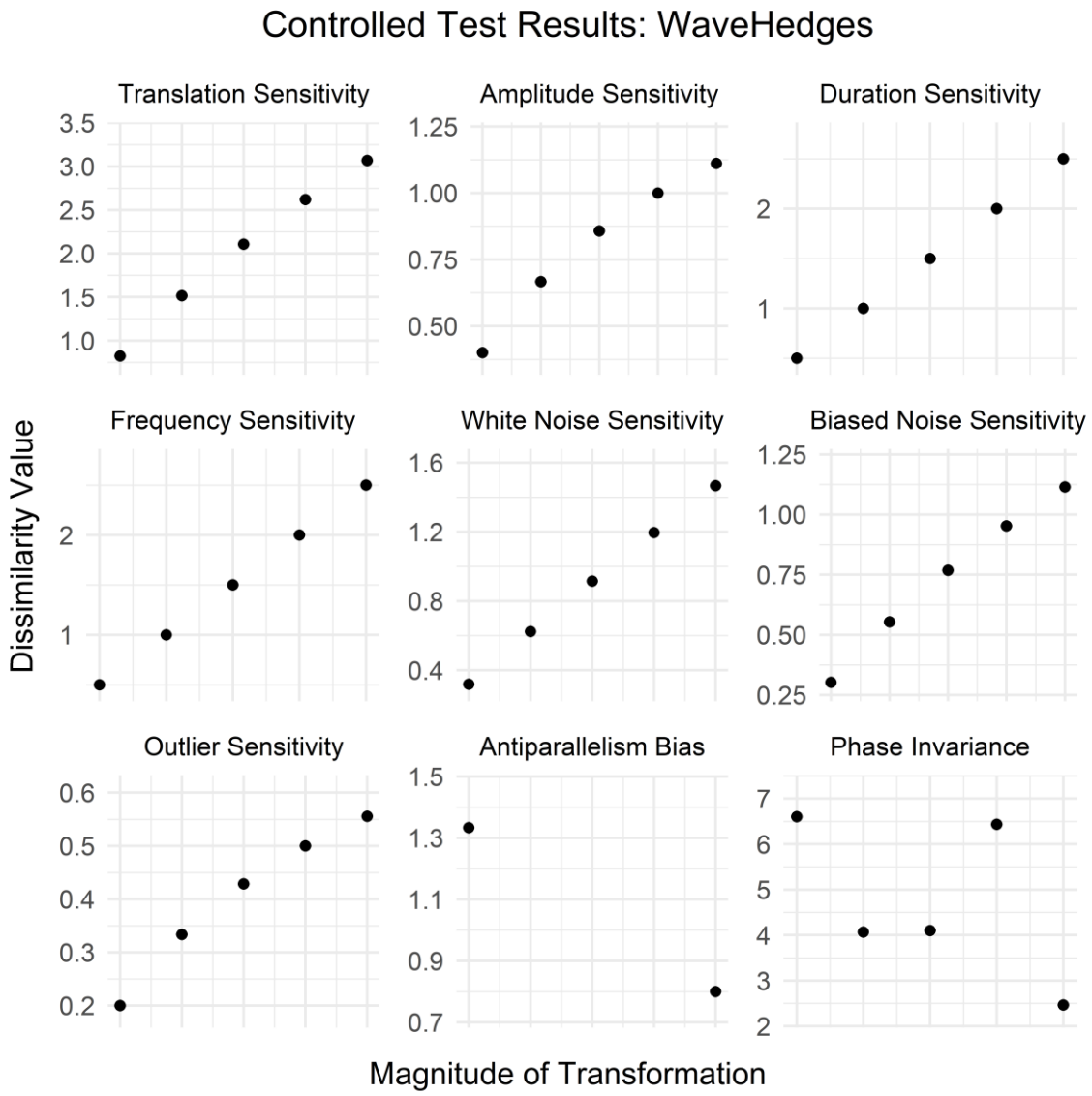


Figure S2.47. Controlled testing results for the Wave-Hedges Distance. Sensitivities and phase invariance were tested by comparing time series with linearly increasing differences in summed y-axis values (or phase), against a reference time series. Antiparallelism bias was tested by comparing pairs of time series that differed by the same relative amount in different directions. The bias is neutral if the two values are identical, negative if the value on the left is higher, and positive if the value on the right is higher. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

S2.10. Tables of controlled test results for all distance measures

This section contains tables of controlled testing results for all 42 distance measures I tested. Each figure includes all time-based and values-based properties for which that distance measure gave results. Distance measures are presented in alphabetical order.

Table S2.2. Raw controlled testing results for the Autocorrelation-based Dissimilarity. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: ACF

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.9611	0.9611					
Translation Sensitivity	0.0000	0.0000	0.0000	0.0000	0.0000		
Amplitude Sensitivity	0.0941	0.2352	0.3529	0.4428	0.5114		
Duration Sensitivity	0.4355	0.6796	0.6374	0.4965	0.4381		
Frequency Sensitivity	0.5277	0.8189	1.0761	1.4041	1.8403		
White Noise Sensitivity	0.0393	0.0902	0.1510	0.2196	0.2939		
Biased Noise Sensitivity	0.0485	0.1042	0.1651	0.2294	0.2953		
Outlier Sensitivity	0.1320	0.2021	0.2508	0.2910	0.3254		
Antiparallelism Bias	0.0000	0.0000					
Phase Invariance	0.3938	0.7610	1.0553	0.3934	0.2057		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.6900	0.6900	0.8552				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	1.0946	0.6401	3.4429	0.6678	0.6473	0.5073

Table S2.3. Raw controlled testing results for the Additive Symmetric Chi-Squared Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Additive

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.00						
Symmetry	42.74	42.74					
Translation Sensitivity	0.17	0.66	1.44	2.49	3.78		
Amplitude Sensitivity	0.45	1.67	3.54	6.00	9.03		
Duration Sensitivity	1.50	3.00	4.50	6.00	7.50		
Frequency Sensitivity	1.50	3.00	4.50	6.00	7.50		
White Noise Sensitivity	0.08	0.33	0.74	1.32	2.09		
Biased Noise Sensitivity	0.16	0.61	1.33	2.29	3.48		
Outlier Sensitivity	0.45	1.67	3.54	6.00	9.03		
Antiparallelism Bias	5.33	2.13					
Phase Invariance	106.90	59.40	61.00	112.90	13.60		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	3,999,998.00	0.00	-4.50				
Non-negativity	0.00						
Triangle Inequality	0.00						
Relative Sensitivity Ranges	0.66	1.58	1.10	1.10	0.37	0.61	1.58

Table S2.4. Raw controlled testing results for the Average Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: AVG

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	7.900	7.900					
Translation Sensitivity	0.550	1.100	1.650	2.200	2.750		
Amplitude Sensitivity	0.750	1.500	2.250	3.000	3.750		
Duration Sensitivity	1.000	1.500	2.000	2.500	3.000		
Frequency Sensitivity	1.000	1.500	2.000	2.500	3.000		
White Noise Sensitivity	0.562	1.125	1.688	2.250	2.812		
Biased Noise Sensitivity	0.625	1.250	1.875	2.500	3.125		
Outlier Sensitivity	1.000	2.000	3.000	4.000	5.000		
Antiparallelism Bias	1.500	1.500					
Phase Invariance	16.500	11.500	11.500	16.500	7.000		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.000	2.000	3.000				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	0.858	1.170	0.780	0.780	0.877	0.975	1.560

Table S2.5. Raw controlled testing results for the Canberra Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Canb

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	4.6030	4.6030					
Translation Sensitivity	0.4297	0.8225	1.1830	1.5152	1.8222		
Amplitude Sensitivity	0.2222	0.4000	0.5455	0.6667	0.7692		
Duration Sensitivity	0.3333	0.6667	1.0000	1.3333	1.6667		
Frequency Sensitivity	0.3333	0.6667	1.0000	1.3333	1.6667		
White Noise Sensitivity	0.1633	0.3265	0.4903	0.6553	0.8222		
Biased Noise Sensitivity	0.1585	0.3022	0.4334	0.5538	0.6649		
Outlier Sensitivity	0.1111	0.2000	0.2727	0.3333	0.3846		
Antiparallelism Bias	1.0000	0.5000					
Phase Invariance	5.1143	3.0032	3.0528	5.0810	1.5667		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.0000	1.0000	3.0000				
Non-negativity	0.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.6125	0.6334	1.5440	1.5440	0.7630	0.5864	0.3167

Table S2.6. Raw controlled testing results for the Compression-Based Dissimilarity Measure. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: CDM

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.5588						
Symmetry	0.7073	0.6829					
Translation Sensitivity	0.7222	0.7222	0.7222	0.7222	0.7222		
Amplitude Sensitivity	0.6667	0.6579	0.6750	0.6579	0.6750		
Duration Sensitivity	0.5588	0.5588	0.5882	0.5882	0.6250		
Frequency Sensitivity	0.6333	0.6333	0.6333	0.6552	0.5714		
White Noise Sensitivity	0.8793	0.8571	0.8621	0.8367	0.8621		
Biased Noise Sensitivity	0.7308	0.7255	0.7308	0.7609	0.7885		
Outlier Sensitivity	0.6591	0.6591	0.6591	0.6591	0.6591		
Antiparallelism Bias	0.6216	0.6216					
Phase Invariance	0.5532	0.5417	0.5870	0.5833	0.5833		
Uniform Time Scaling Invariance	0.7727	0.7639	0.7763	0.7975	0.7538		
Warping Invariance	0.5556	0.5556	0.5556	0.5676	0.5676		
Non-positive Value Handling	0.6923	0.6571	0.6571				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	0.3138	1.2139	1.5362	0.7810	1.1551	0.0000

Table S2.7. Raw controlled testing results for the Complexity-Invariant Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: CID

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	18.512	18.512					
Translation Sensitivity	0.316	0.632	0.949	1.265	1.581		
Amplitude Sensitivity	0.901	2.236	4.039	6.325	9.100		
Duration Sensitivity	1.000	1.414	1.732	2.000	2.236		
Frequency Sensitivity	1.414	2.449	3.464	4.472	5.244		
White Noise Sensitivity	0.373	0.796	1.284	1.846	2.492		
Biased Noise Sensitivity	0.533	1.148	1.865	2.697	3.654		
Outlier Sensitivity	1.254	3.117	5.669	8.944	12.956		
Antiparallelism Bias	1.414	4.243					
Phase Invariance	9.592	7.820	8.052	10.650	5.272		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	3.464	3.464	7.036				
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.281	1.824	0.275	0.852	0.471	0.694	2.603

Table S2.8. Raw controlled testing results for the Clark Squared Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Clark

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	1.5951	1.5951					
Translation Sensitivity	0.1390	0.2658	0.3819	0.4886	0.5871		
Amplitude Sensitivity	0.1571	0.2828	0.3857	0.4714	0.5439		
Duration Sensitivity	0.3333	0.4714	0.5774	0.6667	0.7454		
Frequency Sensitivity	0.3333	0.4714	0.5774	0.6667	0.7454		
White Noise Sensitivity	0.0629	0.1256	0.1887	0.2525	0.3176		
Biased Noise Sensitivity	0.0881	0.1670	0.2383	0.3030	0.3622		
Outlier Sensitivity	0.1111	0.2000	0.2727	0.3333	0.3846		
Antiparallelism Bias	0.7071	0.3536					
Phase Invariance	1.6974	1.2317	1.2546	1.7263	0.6536		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.0000	1.0000	3.0000				
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.2743	1.1001	1.1718	1.1718	0.7244	0.7797	0.7779

Table S2.9. Raw controlled testing results for the Dissimilarity Index Combining Temporal Correlation and Raw Value Behaviour. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Cort

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	2.566	2.566					
Translation Sensitivity	0.075	0.151	0.226	0.302	0.377		
Amplitude Sensitivity	0.174	0.369	0.582	0.810	1.048		
Duration Sensitivity	0.538	0.761	0.932	1.076	1.203		
Frequency Sensitivity	0.391	0.678	0.932	1.161	1.336		
White Noise Sensitivity	0.086	0.178	0.281	0.398	0.527		
Biased Noise Sensitivity	0.121	0.251	0.396	0.558	0.736		
Outlier Sensitivity	0.252	0.551	0.897	1.277	1.681		
Antiparallelism Bias	2.491	0.337					
Phase Invariance	14.955	6.919	6.215	15.084	1.561		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.959	0.959	1.793				
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.400	1.160	0.883	1.255	0.587	0.816	1.898

Table S2.10. Raw controlled testing results for the Czekanowski Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Czek

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.4103	0.4103					
Translation Sensitivity	0.0400	0.0769	0.1111	0.1429	0.1724		
Amplitude Sensitivity	0.0345	0.0667	0.0968	0.1250	0.1515		
Duration Sensitivity	0.0400	0.0769	0.1111	0.1429	0.1724		
Frequency Sensitivity	0.0370	0.0714	0.1034	0.1333	0.1613		
White Noise Sensitivity	0.0156	0.0312	0.0469	0.0625	0.0781		
Biased Noise Sensitivity	0.0154	0.0303	0.0448	0.0588	0.0725		
Outlier Sensitivity	0.0204	0.0400	0.0588	0.0769	0.0943		
Antiparallelism Bias	0.1000	0.0833					
Phase Invariance	0.5000	0.3214	0.3214	0.5000	0.2143		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.0769	0.0769	0.1200				
Non-negativity	0.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.3248	1.1710	1.3248	1.2432	0.6253	0.5711	0.7397

Table S2.11. Raw controlled testing results for the Dice Dissimilarity. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Dice

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.3077	0.3077					
Translation Sensitivity	0.0029	0.0108	0.0224	0.0370	0.0538		
Amplitude Sensitivity	0.0103	0.0370	0.0744	0.1176	0.1634		
Duration Sensitivity	0.0286	0.0526	0.0732	0.0909	0.1064		
Frequency Sensitivity	0.0303	0.0556	0.0769	0.0952	0.1111		
White Noise Sensitivity	0.0005	0.0018	0.0041	0.0072	0.0113		
Biased Noise Sensitivity	0.0009	0.0035	0.0075	0.0130	0.0196		
Outlier Sensitivity	0.0069	0.0256	0.0533	0.0870	0.1244		
Antiparallelism Bias	0.0952	0.0606					
Phase Invariance	0.4182	0.2545	0.2545	0.4545	0.1091		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.0909	0.0909	0.2000				
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.6987	2.1029	1.0688	1.1100	0.1483	0.2575	1.6138

Table S2.12. Raw controlled testing results for the Divergence Squared Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Diverge

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	5.089	5.089					
Translation Sensitivity	0.039	0.141	0.292	0.478	0.689		
Amplitude Sensitivity	0.049	0.160	0.298	0.444	0.592		
Duration Sensitivity	0.222	0.444	0.667	0.889	1.111		
Frequency Sensitivity	0.222	0.444	0.667	0.889	1.111		
White Noise Sensitivity	0.008	0.032	0.071	0.128	0.202		
Biased Noise Sensitivity	0.016	0.056	0.114	0.184	0.262		
Outlier Sensitivity	0.025	0.080	0.149	0.222	0.296		
Antiparallelism Bias	1.000	0.250					
Phase Invariance	5.763	3.034	3.148	5.960	0.854		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.000	2.000	18.000				
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	1.237	1.031	1.690	1.690	0.368	0.469	0.515

Table S2.13. Raw controlled testing results for the Dynamic Time Warping Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: DTW

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	13.800	13.800					
Translation Sensitivity	1.900	3.800	5.700	7.600	9.500		
Amplitude Sensitivity	2.000	4.000	5.000	6.000	7.000		
Duration Sensitivity	0.000	0.000	0.000	0.000	0.000		
Frequency Sensitivity	1.000	2.000	3.000	4.000	5.000		
White Noise Sensitivity	2.000	4.000	6.000	8.000	9.750		
Biased Noise Sensitivity	2.000	4.000	5.750	6.000	7.000		
Outlier Sensitivity	2.000	4.000	5.000	6.000	7.000		
Antiparallelism Bias	2.000	3.000					
Phase Invariance	8.000	8.000	15.000	22.000	18.000		
Uniform Time Scaling Invariance	6.490	6.740	7.000	7.870	6.000		
Warping Invariance	0.000	0.000	0.000	0.000	0.000		
Non-positive Value Handling	3.000	3.000	4.000				
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	1.328	0.873	0.000	0.699	1.354	0.873	0.873

Table S2.14. Raw controlled testing results for the Edit Distance on Real Sequences. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: EDR

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	10.000	10.000					
Translation Sensitivity	10.000	10.000	10.000	10.000	10.000		
Amplitude Sensitivity	2.000	2.000	2.000	2.000	2.000		
Duration Sensitivity	1.000	2.000	3.000	4.000	5.000		
Frequency Sensitivity	1.000	2.000	3.000	4.000	5.000		
White Noise Sensitivity	8.000	8.000	8.000	8.000	8.000		
Biased Noise Sensitivity	4.000	4.000	4.000	4.000	4.000		
Outlier Sensitivity	1.000	1.000	1.000	1.000	1.000		
Antiparallelism Bias	2.000	2.000					
Phase Invariance	2.000	4.000	6.000	7.000	6.000		
Uniform Time Scaling Invariance	8.000	9.000	12.000	13.000	9.000		
Warping Invariance	1.000	2.000	3.000	4.000	5.000		
Non-positive Value Handling	1.000	1.000	1.000				
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.000	0.000	1.000	1.000	0.000	0.000	0.000

Table S2.15. Raw controlled testing results for the Edit Distance with Real Penalty. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: ERP

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	11.200	11.200					
Translation Sensitivity	1.000	2.000	3.000	4.000	5.000		
Amplitude Sensitivity	1.000	2.000	3.000	4.000	5.000		
Duration Sensitivity	1.000	2.000	3.000	4.000	5.000		
Frequency Sensitivity	1.000	2.000	3.000	4.000	5.000		
White Noise Sensitivity	1.000	2.000	3.000	4.000	5.000		
Biased Noise Sensitivity	1.000	2.000	3.000	4.000	5.000		
Outlier Sensitivity	1.000	2.000	3.000	4.000	5.000		
Antiparallelism Bias	2.000	2.000					
Phase Invariance	10.000	12.000	12.000	14.000	12.000		
Uniform Time Scaling Invariance	6.170	12.860	18.625	25.340	31.000		
Warping Invariance	1.000	2.000	3.000	4.000	5.000		
Non-positive Value Handling	2.000	2.000	3.000				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table S2.16. Raw controlled testing results for the Euclidean Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Euclidean

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	4.541	4.541					
Translation Sensitivity	0.316	0.632	0.949	1.265	1.581		
Amplitude Sensitivity	0.707	1.414	2.121	2.828	3.536		
Duration Sensitivity	1.000	1.414	1.732	2.000	2.236		
Frequency Sensitivity	1.000	1.414	1.732	2.000	2.236		
White Noise Sensitivity	0.354	0.707	1.061	1.414	1.768		
Biased Noise Sensitivity	0.500	1.000	1.500	2.000	2.500		
Outlier Sensitivity	1.000	2.000	3.000	4.000	5.000		
Antiparallelism Bias	1.414	1.414					
Phase Invariance	9.592	7.483	7.483	10.000	4.899		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.000	2.000	3.000				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	0.633	1.416	0.619	0.619	0.708	1.001	2.003

Table S2.17. Raw controlled testing results for the Fourier Coefficient-Based Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Fourier

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	11.381	11.381					
Translation Sensitivity	1.000	2.000	3.000	4.000	5.000		
Amplitude Sensitivity	1.732	3.464	5.196	6.928	8.660		
Duration Sensitivity	2.449	3.464	4.472	5.292	6.164		
Frequency Sensitivity	2.646	4.000	5.196	6.325	7.416		
White Noise Sensitivity	1.061	2.121	3.182	4.243	5.303		
Biased Noise Sensitivity	1.500	3.000	4.500	6.000	7.500		
Outlier Sensitivity	2.449	4.899	7.348	9.798	12.247		
Antiparallelism Bias	3.464	3.464					
Phase Invariance	21.633	16.733	16.971	22.361	11.314		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	4.899	4.899	7.348				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	0.710	1.229	0.659	0.846	0.753	1.065	1.738

Table S2.18. Raw controlled testing results for the Gower Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Gower

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	1.2800	1.2800					
Translation Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Amplitude Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Duration Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Frequency Sensitivity	0.0833	0.1667	0.2500	0.3333	0.4167		
White Noise Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Biased Noise Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Outlier Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Antiparallelism Bias	0.2000	0.2000					
Phase Invariance	2.8000	1.8000	1.8000	2.8000	1.2000		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.2000	0.2000	0.3000				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	1.0244	1.0244	1.0244	0.8537	1.0244	1.0244	1.0244

Table S2.19. Raw controlled testing results for the Chebyshev Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Chebyshev

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	3.0000	3.0000					
Translation Sensitivity	0.1000	0.2000	0.3000	0.4000	0.5000		
Amplitude Sensitivity	0.5000	1.0000	1.5000	2.0000	2.5000		
Duration Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000		
Frequency Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000		
White Noise Sensitivity	0.1250	0.2500	0.3750	0.5000	0.6250		
Biased Noise Sensitivity	0.2500	0.5000	0.7500	1.0000	1.2500		
Outlier Sensitivity	1.0000	2.0000	3.0000	4.0000	5.0000		
Antiparallelism Bias	1.0000	1.0000					
Phase Invariance	5.0000	5.0000	5.0000	5.0000	2.0000		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.0000	2.0000	3.0000				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.2532	1.2658	0.0000	0.0000	0.3165	0.6329	2.5316

Table S2.20. Raw controlled testing results for the Integrated Periodogram-Based Dissimilarity. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: IntPer

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.8095	0.8095					
Translation Sensitivity	0.0000	0.0000	0.0000	0.0000	0.0000		
Amplitude Sensitivity	0.0266	0.0676	0.1029	0.1305	0.1520		
Duration Sensitivity	0.5316	0.6016	0.6298	0.6331	0.7826		
Frequency Sensitivity	0.9671	1.4235	1.5994	1.8517	1.9900		
White Noise Sensitivity	0.1465	0.3306	0.5389	0.7586	0.9791		
Biased Noise Sensitivity	0.1315	0.2976	0.4847	0.6810	0.8772		
Outlier Sensitivity	0.1486	0.2231	0.2909	0.3391	0.3739		
Antiparallelism Bias	0.0000	0.0000					
Phase Invariance	0.2718	0.5069	0.6876	0.8120	0.5408		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.1307	1.1307	1.2898				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	0.2742	0.5485	2.2354	1.8197	1.6298	0.4924

Table S2.21. Raw controlled testing results for the Jaccard Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Jaccard

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.4706	0.4706					
Translation Sensitivity	0.0058	0.0213	0.0439	0.0714	0.1020		
Amplitude Sensitivity	0.0204	0.0714	0.1385	0.2105	0.2809		
Duration Sensitivity	0.0556	0.1000	0.1364	0.1667	0.1923		
Frequency Sensitivity	0.0588	0.1053	0.1429	0.1739	0.2000		
White Noise Sensitivity	0.0009	0.0036	0.0082	0.0144	0.0223		
Biased Noise Sensitivity	0.0018	0.0069	0.0150	0.0256	0.0385		
Outlier Sensitivity	0.0137	0.0500	0.1011	0.1600	0.2212		
Antiparallelism Bias	0.1739	0.1143					
Phase Invariance	0.5897	0.4058	0.4058	0.6250	0.1967		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.1667	0.1667	0.3333				
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.7484	2.0254	1.0633	1.0977	0.1660	0.2856	1.6137

Table S2.22. Raw controlled testing results for the Jeffreys Divergence. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Jeffreys

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	14.372	14.372					
Translation Sensitivity	0.086	0.330	0.714	1.223	1.845		
Amplitude Sensitivity	0.223	0.811	1.679	2.773	4.055		
Duration Sensitivity	0.693	1.386	2.079	2.773	3.466		
Frequency Sensitivity	0.693	1.386	2.079	2.773	3.466		
White Noise Sensitivity	0.041	0.163	0.368	0.658	1.034		
Biased Noise Sensitivity	0.079	0.303	0.654	1.119	1.687		
Outlier Sensitivity	0.223	0.811	1.679	2.773	4.055		
Antiparallelism Bias	2.197	1.022					
Phase Invariance	37.437	20.901	21.476	39.634	6.438		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	29.017	23.026					
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.701	1.527	1.105	1.105	0.396	0.641	1.527

Table S2.23. Raw controlled testing results for the Jaccard Difference. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Jensen

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	1.6459	1.6459					
Translation Sensitivity	0.0107	0.0412	0.0890	0.1522	0.2292		
Amplitude Sensitivity	0.0278	0.1007	0.2072	0.3398	0.4934		
Duration Sensitivity	0.0849	0.1699	0.2548	0.3398	0.4247		
Frequency Sensitivity	0.0849	0.1699	0.2548	0.3398	0.4247		
White Noise Sensitivity	0.0051	0.0204	0.0460	0.0821	0.1288		
Biased Noise Sensitivity	0.0099	0.0378	0.0815	0.1392	0.2093		
Outlier Sensitivity	0.0278	0.1007	0.2072	0.3398	0.4934		
Antiparallelism Bias	0.2616	0.1263					
Phase Invariance	4.2886	2.4001	2.4658	4.5413	0.7938		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.6931	0.6931					
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.7105	1.5142	1.1052	1.1052	0.4024	0.6485	1.5142

Table S2.24. Raw controlled testing results for the Kulczynski Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Kulcz

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	1.3913	1.3913					
Translation Sensitivity	0.0833	0.1667	0.2500	0.3333	0.4167		
Amplitude Sensitivity	0.0714	0.1429	0.2143	0.2857	0.3571		
Duration Sensitivity	0.0833	0.1667	0.2500	0.3333	0.4167		
Frequency Sensitivity	0.0769	0.1538	0.2308	0.3077	0.3846		
White Noise Sensitivity	0.0317	0.0645	0.0984	0.1333	0.1695		
Biased Noise Sensitivity	0.0312	0.0625	0.0938	0.1250	0.1562		
Outlier Sensitivity	0.0417	0.0833	0.1250	0.1667	0.2083		
Antiparallelism Bias	0.2222	0.1818					
Phase Invariance	2.0000	0.9474	0.9474	2.0000	0.5455		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.1667	0.1667	0.2727				
Non-negativity	0.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.3811	1.1838	1.3811	1.2749	0.5707	0.5179	0.6905

Table S2.25. Raw controlled testing results for the Kullback-Leibler Divergence. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Kullback

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	14.449	-0.077					
Translation Sensitivity	1.044	2.170	3.371	4.644	5.981		
Amplitude Sensitivity	1.116	2.433	3.917	5.545	7.298		
Duration Sensitivity	1.386	2.773	4.159	5.545	6.931		
Frequency Sensitivity	1.386	2.773	4.159	5.545	6.931		
White Noise Sensitivity	0.020	0.082	0.184	0.327	0.513		
Biased Noise Sensitivity	1.040	2.156	3.342	4.591	5.900		
Outlier Sensitivity	1.116	2.433	3.917	5.545	7.298		
Antiparallelism Bias	-1.099	2.554					
Phase Invariance	17.594	11.549	9.716	20.654	3.219		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	-0.000	0.000					
Non-negativity	0.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	1.024	1.283	1.150	1.150	0.102	1.008	1.283

Table S2.26. Raw controlled testing results for the Kumar-Johnson Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: KumarJohnson

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.00						
Symmetry	60.0	60.0					
Translation Sensitivity	0.17	0.67	1.45	2.52	3.85		
Amplitude Sensitivity	0.45	1.70	3.68	6.36	9.78		
Duration Sensitivity	1.59	3.18	4.77	6.36	7.95		
Frequency Sensitivity	1.59	3.18	4.77	6.36	7.95		
White Noise Sensitivity	0.08	0.33	0.74	1.33	2.11		
Biased Noise Sensitivity	0.16	0.61	1.34	2.32	3.55		
Outlier Sensitivity	0.45	1.70	3.68	6.36	9.78		
Antiparallelism Bias	6.16	2.20					
Phase Invariance	140	77.9	79.9	148	14.2		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.82×10^9	1.60×10^6					
Non-negativity	1.00						
Triangle Inequality	0.00						
Relative Sensitivity Ranges	0.64	1.61	1.10	1.10	0.35	0.59	1.61

Table S2.27. Raw controlled testing results for the K Divergence. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: KDiv

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	5.1984	-1.9065					
Translation Sensitivity	0.5106	1.0400	1.5853	2.1441	2.7147		
Amplitude Sensitivity	0.5268	1.0939	1.6881	2.3015	2.9288		
Duration Sensitivity	0.5754	1.1507	1.7261	2.3015	2.8768		
Frequency Sensitivity	0.5754	1.1507	1.7261	2.3015	2.8768		
White Noise Sensitivity	0.0051	0.0204	0.0461	0.0824	0.1298		
Biased Noise Sensitivity	0.5097	1.0367	1.5780	2.1314	2.6954		
Outlier Sensitivity	0.5268	1.0939	1.6881	2.3015	2.9288		
Antiparallelism Bias	-0.6931	1.1157					
Phase Invariance	4.5097	2.1852	2.6634	4.3847	0.7938		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	-0.0000						
Non-negativity	0.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.1083	1.2078	1.1572	1.1572	0.0627	1.0990	1.2078

Table S2.28. Raw controlled testing results for the Lorentzian Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Lorentz

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	7.927	7.927					
Translation Sensitivity	0.953	1.823	2.624	3.365	4.055		
Amplitude Sensitivity	0.811	1.386	1.833	2.197	2.506		
Duration Sensitivity	0.693	1.386	2.079	2.773	3.466		
Frequency Sensitivity	0.693	1.386	2.079	2.773	3.466		
White Noise Sensitivity	0.942	1.785	2.548	3.244	3.884		
Biased Noise Sensitivity	0.893	1.622	2.238	2.773	3.244		
Outlier Sensitivity	0.693	1.099	1.386	1.609	1.792		
Antiparallelism Bias	1.386	1.386					
Phase Invariance	12.871	8.723	8.723	12.283	6.592		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.099	1.099	1.386				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	1.297	0.709	1.160	1.160	1.231	0.984	0.460

Table S2.29. Raw controlled testing results for the Manhattan Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Manhattan

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	12.800	12.800					
Translation Sensitivity	1.000	2.000	3.000	4.000	5.000		
Amplitude Sensitivity	1.000	2.000	3.000	4.000	5.000		
Duration Sensitivity	1.000	2.000	3.000	4.000	5.000		
Frequency Sensitivity	1.000	2.000	3.000	4.000	5.000		
White Noise Sensitivity	1.000	2.000	3.000	4.000	5.000		
Biased Noise Sensitivity	1.000	2.000	3.000	4.000	5.000		
Outlier Sensitivity	1.000	2.000	3.000	4.000	5.000		
Antiparallelism Bias	2.000	2.000					
Phase Invariance	28.000	18.000	18.000	28.000	12.000		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.000	2.000	3.000				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table S2.30. Raw controlled testing results for the Normalized Compression Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: NCD

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.1176						
Symmetry	0.5000	0.4583					
Translation Sensitivity	0.4737	0.4737	0.4737	0.4737	0.4737		
Amplitude Sensitivity	0.4091	0.3810	0.4348	0.3810	0.4348		
Duration Sensitivity	0.1176	0.1176	0.1765	0.1765	0.2941		
Frequency Sensitivity	0.3125	0.3125	0.3125	0.3333	0.1429		
White Noise Sensitivity	0.7941	0.7500	0.7647	0.6800	0.7647		
Biased Noise Sensitivity	0.5000	0.4815	0.5000	0.5417	0.6071		
Outlier Sensitivity	0.3182	0.3182	0.3182	0.3182	0.3182		
Antiparallelism Bias	0.2632	0.2632					
Phase Invariance	0.1250	0.0833	0.2083	0.1667	0.1667		
Uniform Time Scaling Invariance	0.6429	0.6458	0.6731	0.7091	0.6098		
Warping Invariance	0.1111	0.1111	0.1111	0.1579	0.1579		
Non-positive Value Handling	0.4545	0.3333	0.3333				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	0.4075	1.3358	1.4418	0.8638	0.9512	0.0000

Table S2.31. Raw controlled testing results for the Partial Autocorrelation-Based Dissimilarity. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: PACF

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.6472	0.6472					
Translation Sensitivity	0.0000	0.0000	0.0000	0.0000	0.0000		
Amplitude Sensitivity	0.1346	0.2640	0.3414	0.3904	0.4238		
Duration Sensitivity	0.8358	0.6037	0.8581	0.3727	0.7841		
Frequency Sensitivity	0.7398	0.8911	0.8791	0.7482	0.8603		
White Noise Sensitivity	0.1006	0.2073	0.3138	0.4162	0.5122		
Biased Noise Sensitivity	0.1096	0.2155	0.3131	0.4011	0.4796		
Outlier Sensitivity	0.3645	0.5340	0.5184	0.4628	0.4217		
Antiparallelism Bias	0.0000	0.0000					
Phase Invariance	0.5308	0.5950	0.5913	0.3674	0.2403		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.8325	0.8325	0.9658				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	1.0786	1.8102	0.5642	1.5350	1.3798	0.6321

Table S2.32. Raw controlled testing results for the Periodogram-Based Dissimilarity. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Per

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.8701	0.8701					
Translation Sensitivity	0.0000	0.0000	0.0000	0.0000	0.0000		
Amplitude Sensitivity	0.1000	0.2404	0.4218	0.6445	0.9086		
Duration Sensitivity	0.0623	0.1053	0.1463	0.1303	0.0771		
Frequency Sensitivity	0.0582	0.1368	0.2421	0.3819	0.4624		
White Noise Sensitivity	0.1191	0.2836	0.4936	0.7494	1.0511		
Biased Noise Sensitivity	0.1389	0.3176	0.5385	0.8029	1.1115		
Outlier Sensitivity	0.3066	0.6883	1.1539	1.7090	2.3572		
Antiparallelism Bias	0.0000	0.1666					
Phase Invariance	0.8618	1.3408	1.1143	1.0548	0.9195		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.1963	0.1963	0.4163				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	1.0778	0.1120	0.5386	1.2421	1.2963	2.7332

Table S2.33. Raw controlled testing results for the Piccolo Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: Piccolo

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.5655	0.5655					
Translation Sensitivity	0.0000	0.0000	0.0000	0.0000	0.0000		
Amplitude Sensitivity	0.0571	0.1382	0.2019	0.2484	0.2827		
Duration Sensitivity	0.1310	0.1667	0.1500	0.0833	0.0595		
Frequency Sensitivity	0.1182	0.5564	0.5511	0.6753	0.8182		
White Noise Sensitivity	0.0680	0.1302	0.1824	0.2246	0.2586		
Biased Noise Sensitivity	0.0721	0.1312	0.1782	0.2155	0.2459		
Outlier Sensitivity	0.0523	0.0973	0.1310	0.1555	0.1734		
Antiparallelism Bias	0.0000	0.0000					
Phase Invariance	0.0000	0.0570	0.2468	0.0570	0.1139		
Uniform Time Scaling Invariance	0.3241	0.6077	0.6239	0.6389	0.6461		
Warping Invariance	0.6642	0.8488	0.8799	0.4125	0.4522		
Non-positive Value Handling	0.9883	0.9883	1.1271				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	0.0000	1.0399	0.4940	3.2275	0.8785	0.8014	0.5587

Table S2.34. Raw controlled testing results for the Probabilistic Symmetric Chi-Squared Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: ProbSymm

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	12.279	12.279					
Translation Sensitivity	0.086	0.329	0.710	1.212	1.822		
Amplitude Sensitivity	0.222	0.800	1.636	2.667	3.846		
Duration Sensitivity	0.667	1.333	2.000	2.667	3.333		
Frequency Sensitivity	0.667	1.333	2.000	2.667	3.333		
White Noise Sensitivity	0.041	0.163	0.368	0.655	1.028		
Biased Noise Sensitivity	0.079	0.302	0.650	1.108	1.662		
Outlier Sensitivity	0.222	0.800	1.636	2.667	3.846		
Antiparallelism Bias	2.000	1.000					
Phase Invariance	31.810	17.854	18.334	33.676	6.267		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	4.000	4.000	18.000				
Non-negativity	0.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.720	1.502	1.105	1.105	0.409	0.656	1.502

Table S2.35. Raw controlled testing results for the Soergel Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Soergel

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	0.5818	0.5818					
Translation Sensitivity	0.0769	0.1429	0.2000	0.2500	0.2941		
Amplitude Sensitivity	0.0667	0.1250	0.1765	0.2222	0.2632		
Duration Sensitivity	0.0769	0.1429	0.2000	0.2500	0.2941		
Frequency Sensitivity	0.0714	0.1333	0.1875	0.2353	0.2778		
White Noise Sensitivity	0.0308	0.0606	0.0896	0.1176	0.1449		
Biased Noise Sensitivity	0.0303	0.0588	0.0857	0.1111	0.1351		
Outlier Sensitivity	0.0400	0.0769	0.1111	0.1429	0.1724		
Antiparallelism Bias	0.1818	0.1538					
Phase Invariance	0.6667	0.4865	0.4865	0.6667	0.3529		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	0.1429	0.1429	0.2143				
Non-negativity	1.0000						
Triangle Inequality	1.0000						
Relative Sensitivity Ranges	1.2791	1.1572	1.2791	1.2152	0.6723	0.6174	0.7798

Table S2.36. Raw controlled testing results for the Squared Chi-Squared Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: SqChi

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	6.139	6.139					
Translation Sensitivity	0.043	0.165	0.355	0.606	0.911		
Amplitude Sensitivity	0.111	0.400	0.818	1.333	1.923		
Duration Sensitivity	0.333	0.667	1.000	1.333	1.667		
Frequency Sensitivity	0.333	0.667	1.000	1.333	1.667		
White Noise Sensitivity	0.020	0.082	0.184	0.328	0.514		
Biased Noise Sensitivity	0.040	0.151	0.325	0.554	0.831		
Outlier Sensitivity	0.111	0.400	0.818	1.333	1.923		
Antiparallelism Bias	1.000	0.500					
Phase Invariance	15.905	8.927	9.167	16.838	3.133		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.000	2.000	9.000				
Non-negativity	0.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.720	1.502	1.105	1.105	0.409	0.656	1.502

Table S2.37. Raw controlled testing results for the Squared-Chord Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: SqChord

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	3.4299	3.4299					
Translation Sensitivity	0.0215	0.0824	0.1781	0.3050	0.4598		
Amplitude Sensitivity	0.0557	0.2020	0.4170	0.6863	1.0000		
Duration Sensitivity	0.1716	0.3431	0.5147	0.6863	0.8579		
Frequency Sensitivity	0.1716	0.3431	0.5147	0.6863	0.8579		
White Noise Sensitivity	0.0102	0.0408	0.0921	0.1643	0.2580		
Biased Noise Sensitivity	0.0198	0.0757	0.1633	0.2791	0.4201		
Outlier Sensitivity	0.0557	0.2020	0.4170	0.6863	1.0000		
Antiparallelism Bias	0.5359	0.2540					
Phase Invariance	8.9453	4.9998	5.1372	9.4719	1.5984		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.9972	2.0000					
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.7057	1.5204	1.1050	1.1050	0.3990	0.6446	1.5204

Table S2.38. Raw controlled testing results for the Squared Euclidean Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: SqEuclid

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.00						
Symmetry	20.62	20.62					
Translation Sensitivity	0.10	0.40	0.90	1.60	2.50		
Amplitude Sensitivity	0.50	2.00	4.50	8.00	12.50		
Duration Sensitivity	1.00	2.00	3.00	4.00	5.00		
Frequency Sensitivity	1.00	2.00	3.00	4.00	5.00		
White Noise Sensitivity	0.12	0.50	1.12	2.00	3.12		
Biased Noise Sensitivity	0.25	1.00	2.25	4.00	6.25		
Outlier Sensitivity	1.00	4.00	9.00	16.00	25.00		
Antiparallelism Bias	2.00	2.00					
Phase Invariance	92.00	56.00	56.00	100.00	24.00		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	4.00	4.00	9.00				
Non-negativity	1.00						
Triangle Inequality	0.00						
Relative Sensitivity Ranges	0.30	1.52	0.51	0.51	0.38	0.76	3.03

Table S2.39. Raw controlled testing results for the Short Time Series Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: STS

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	5.257	5.257					
Translation Sensitivity	0.000	0.000	0.000	0.000	0.000		
Amplitude Sensitivity	0.707	1.414	2.121	2.828	3.536		
Duration Sensitivity	1.414	1.414	1.414	1.414	1.414		
Frequency Sensitivity	1.414	2.000	2.449	2.828	3.000		
White Noise Sensitivity	0.685	1.369	2.054	2.739	3.423		
Biased Noise Sensitivity	0.707	1.414	2.121	2.828	3.536		
Outlier Sensitivity	1.414	2.828	4.243	5.657	7.071		
Antiparallelism Bias	1.414	1.414					
Phase Invariance	15.748	12.124	11.662	14.933	5.292		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	2.828	2.828	4.243				
Non-negativity	1.000						
Triangle Inequality	1.000						
Relative Sensitivity Ranges	0.000	1.085	0.000	0.608	1.051	1.085	2.170

Table S2.40. Raw controlled testing results for the Time Alignment Measurement Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4.

Controlled Test Results: TAM

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	2.0000	2.0000					
Translation Sensitivity	0.0000	0.6667	0.6667	0.6667	0.0000		
Amplitude Sensitivity	0.0000	0.0000	0.6667	0.6667	0.6667		
Duration Sensitivity	0.3333	0.6667	1.0000	1.3333	1.6667		
Frequency Sensitivity	0.2727	0.5455	0.8182	1.0909	1.3636		
White Noise Sensitivity	0.0000	0.0000	0.0000	0.0000	0.3333		
Biased Noise Sensitivity	0.0000	0.0000	0.3333	0.6667	1.0000		
Outlier Sensitivity	0.0000	0.0000	0.3333	0.3333	0.3333		
Antiparallelism Bias	1.0000	0.6667					
Phase Invariance	0.3333	0.6667	1.0000	1.3333	1.6667		
Uniform Time Scaling Invariance	0.7444	0.2500	0.3571	0.4375	0.5000		
Warping Invariance	0.1000	0.1818	0.2500	0.3077	0.3571		
Non-positive Value Handling	0.3333	0.3333	0.3333				
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.8603	0.8603	1.7207	1.4078	0.4302	1.2905	0.4302

Table S2.41. Raw controlled testing results for the Taneja Difference. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Taneja

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.000						
Symmetry	1.947	1.947					
Translation Sensitivity	0.011	0.041	0.089	0.153	0.232		
Amplitude Sensitivity	0.028	0.102	0.213	0.353	0.520		
Duration Sensitivity	0.088	0.177	0.265	0.353	0.442		
Frequency Sensitivity	0.088	0.177	0.265	0.353	0.442		
White Noise Sensitivity	0.005	0.020	0.046	0.082	0.130		
Biased Noise Sensitivity	0.010	0.038	0.082	0.141	0.212		
Outlier Sensitivity	0.028	0.102	0.213	0.353	0.520		
Antiparallelism Bias	0.288	0.129					
Phase Invariance	5.071	2.825	2.903	5.367	0.816		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	6.561	12.206					
Non-negativity	1.000						
Triangle Inequality	0.000						
Relative Sensitivity Ranges	0.692	1.539	1.104	1.104	0.389	0.633	1.539

Table S2.42. Raw controlled testing results for the Topsoe Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: Topsoe

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	3.2919	3.2919					
Translation Sensitivity	0.0215	0.0824	0.1779	0.3043	0.4584		
Amplitude Sensitivity	0.0557	0.2014	0.4143	0.6796	0.9868		
Duration Sensitivity	0.1699	0.3398	0.5097	0.6796	0.8495		
Frequency Sensitivity	0.1699	0.3398	0.5097	0.6796	0.8495		
White Noise Sensitivity	0.0102	0.0408	0.0920	0.1641	0.2577		
Biased Noise Sensitivity	0.0198	0.0757	0.1631	0.2784	0.4186		
Outlier Sensitivity	0.0557	0.2014	0.4143	0.6796	0.9868		
Antiparallelism Bias	0.5232	0.2527					
Phase Invariance	8.5772	4.8003	4.9316	9.0826	1.5876		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.3863						
Non-negativity	1.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	0.7105	1.5142	1.1052	1.1052	0.4024	0.6485	1.5142

Table S2.43. Raw controlled testing results for the Wave-Hedges Distance. For details on how testing was performed and how relative sensitivity ranges were calculated, see supplementary text S2.4. Uniform time scaling invariance and warping invariance could not be tested for this distance measure, as it cannot measure unequal-length time series.

Controlled Test Results: WaveHedges

Test	Res.1	Res.2	Res.3	Res.4	Res.5	Res.6	Res.7
Uniqueness	0.0000						
Symmetry	6.0500	6.0500					
Translation Sensitivity	0.8225	1.5152	2.1070	2.6190	3.0667		
Amplitude Sensitivity	0.4000	0.6667	0.8571	1.0000	1.1111		
Duration Sensitivity	0.5000	1.0000	1.5000	2.0000	2.5000		
Frequency Sensitivity	0.5000	1.0000	1.5000	2.0000	2.5000		
White Noise Sensitivity	0.3189	0.6231	0.9147	1.1955	1.4670		
Biased Noise Sensitivity	0.3022	0.5538	0.7677	0.9524	1.1141		
Outlier Sensitivity	0.2000	0.3333	0.4286	0.5000	0.5556		
Antiparallelism Bias	1.3333	0.8000					
Phase Invariance	6.6000	4.0667	4.1000	6.4333	2.4667		
Uniform Time Scaling Invariance							
Warping Invariance							
Non-positive Value Handling	1.0000	1.0000	1.5000				
Non-negativity	0.0000						
Triangle Inequality	0.0000						
Relative Sensitivity Ranges	1.6945	0.5369	1.5101	1.5101	0.8669	0.6130	0.2685

S2.11. Plots of wading bird rankings for all distance measures

This section contains plots of wading bird dissimilarity results for all 42 distance measures I tested. Each figure shows dissimilarity results for both smoothed and unsmoothed indices of all five wading birds. Distance measures are presented in alphabetical order.

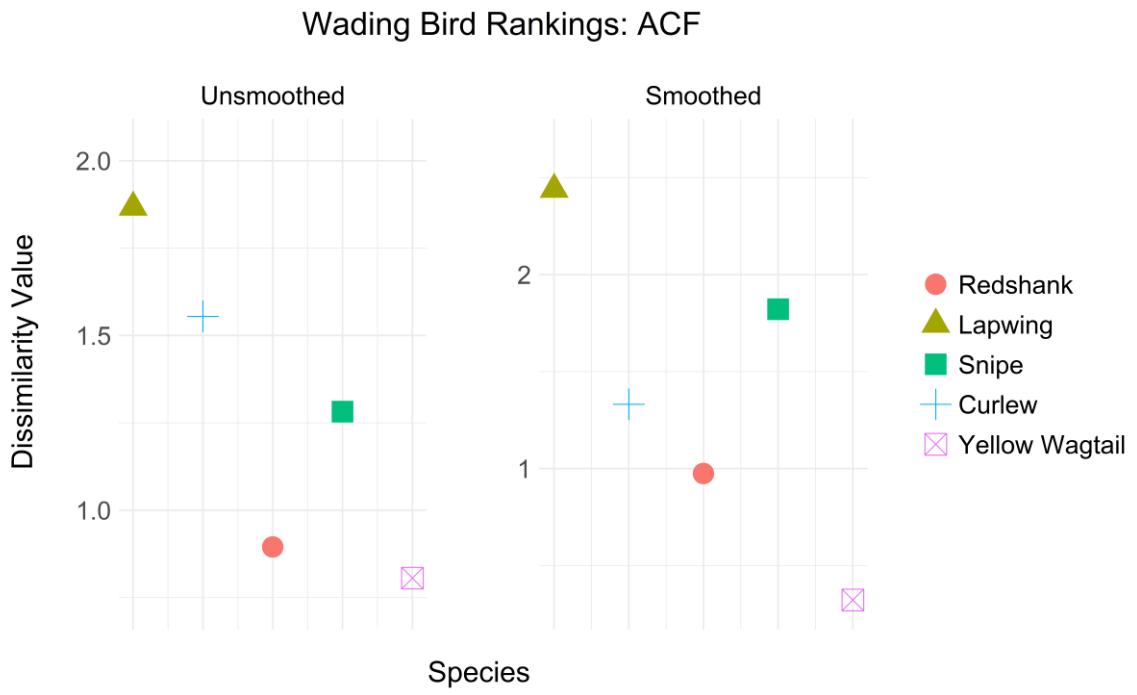


Figure S2.48. Dissimilarity values for trend comparisons of five wading bird species using the Autocorrelation-Based Dissimilarity. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

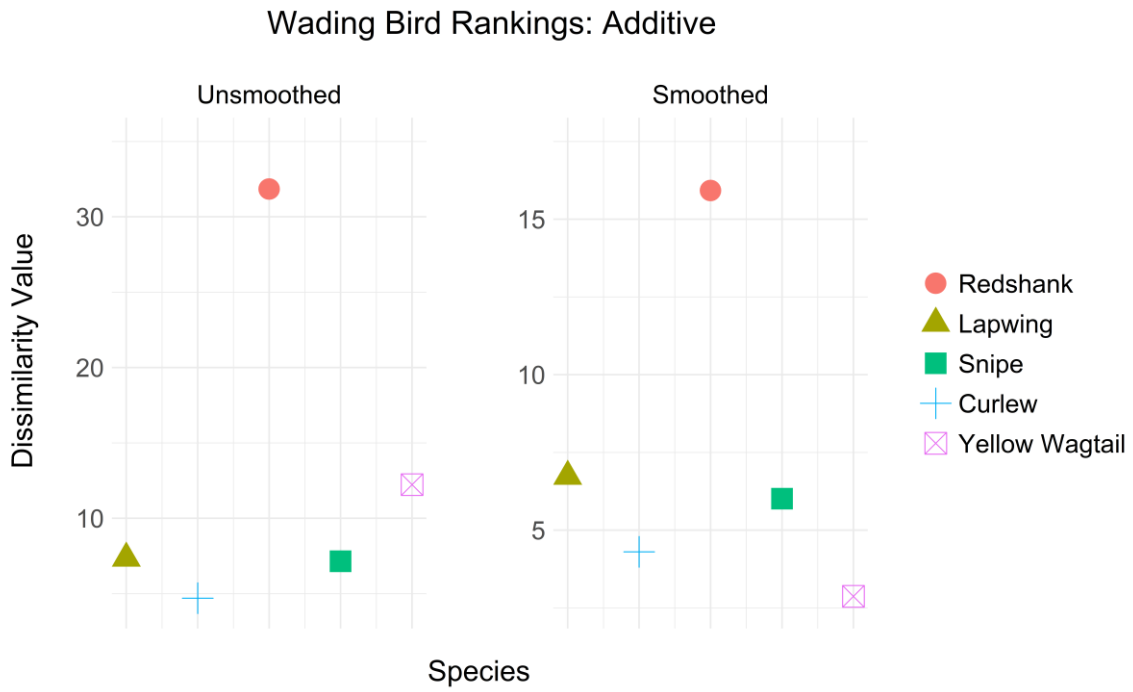


Figure S2.49. Dissimilarity values for trend comparisons of five wading bird species using the Additive Symmetric Chi-Squared Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

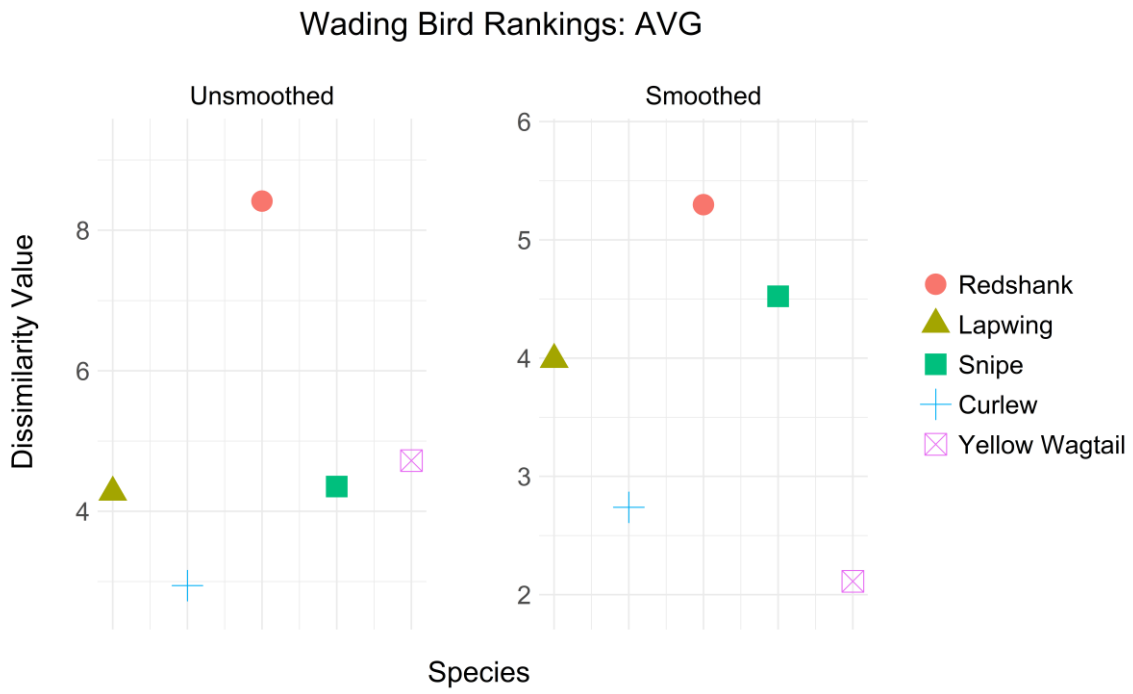


Figure S2.50. Dissimilarity values for trend comparisons of five wading bird species using the Average Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

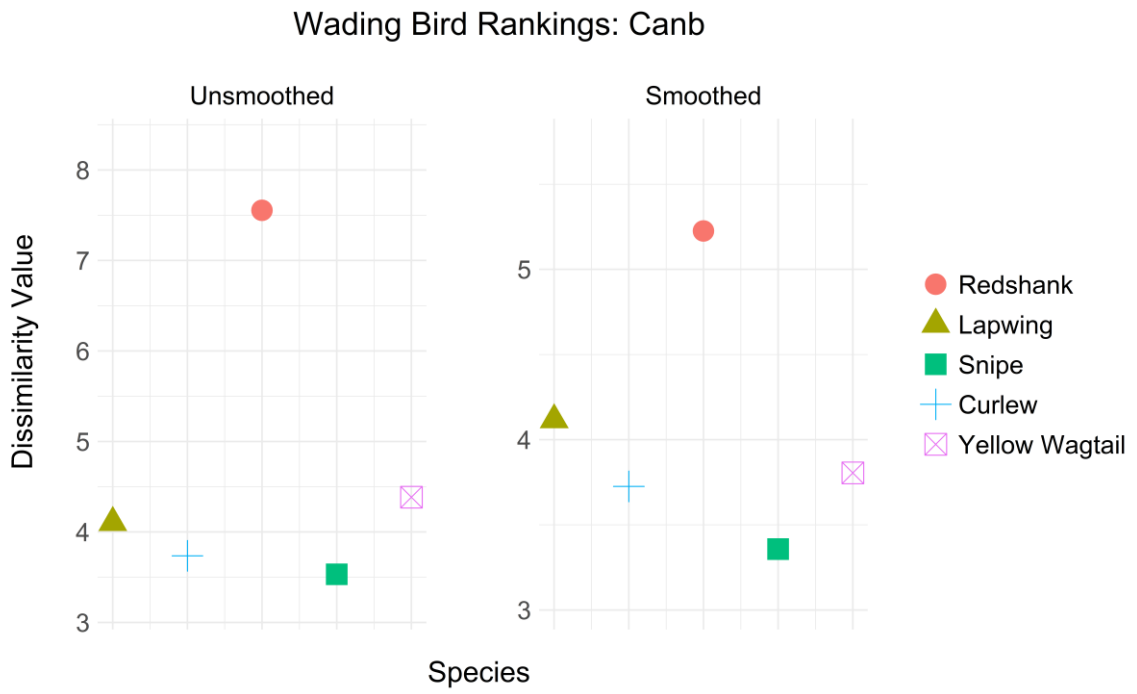


Figure S2.51. Dissimilarity values for trend comparisons of five wading bird species using the Canberra Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

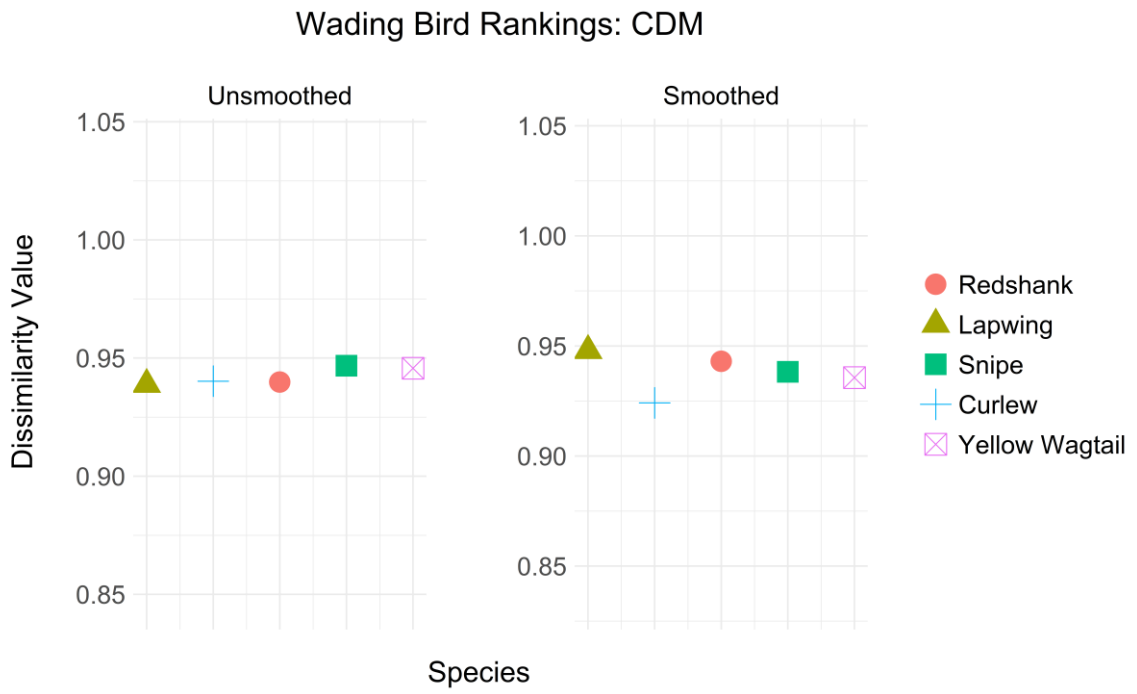


Figure S2.52. Dissimilarity values for trend comparisons of five wading bird species using the Compression-Based Dissimilarity Measure. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

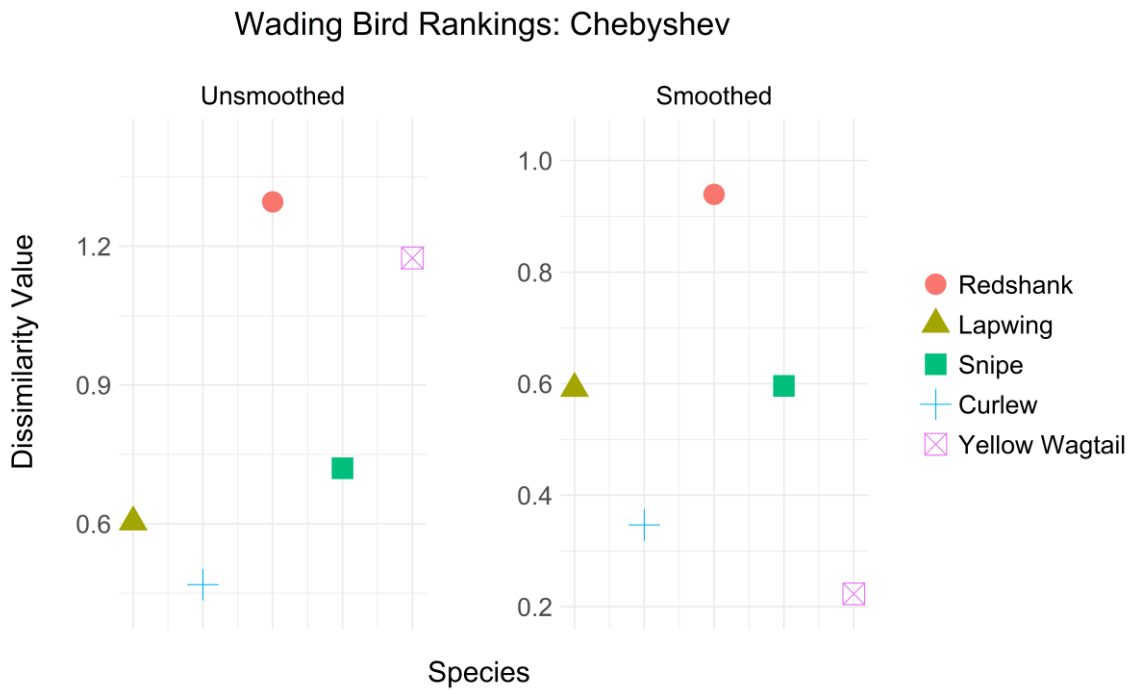


Figure S2.53. Dissimilarity values for trend comparisons of five wading bird species using the Chebyshev Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

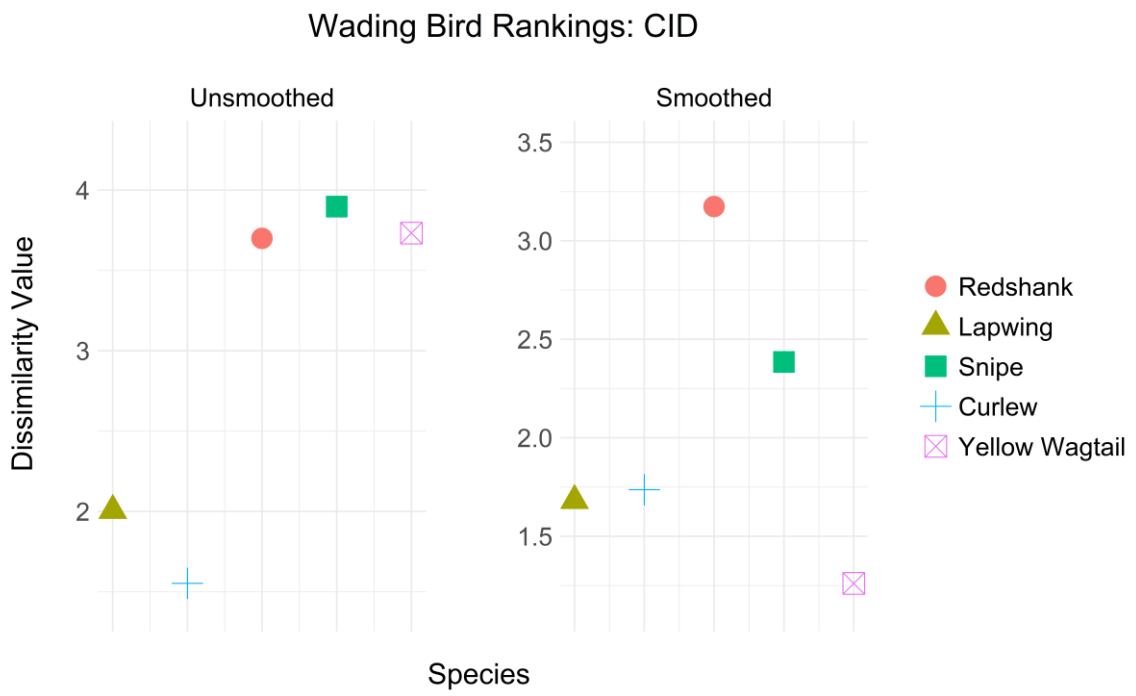


Figure S2.54. Dissimilarity values for trend comparisons of five wading bird species using the Complexity-Invariant Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

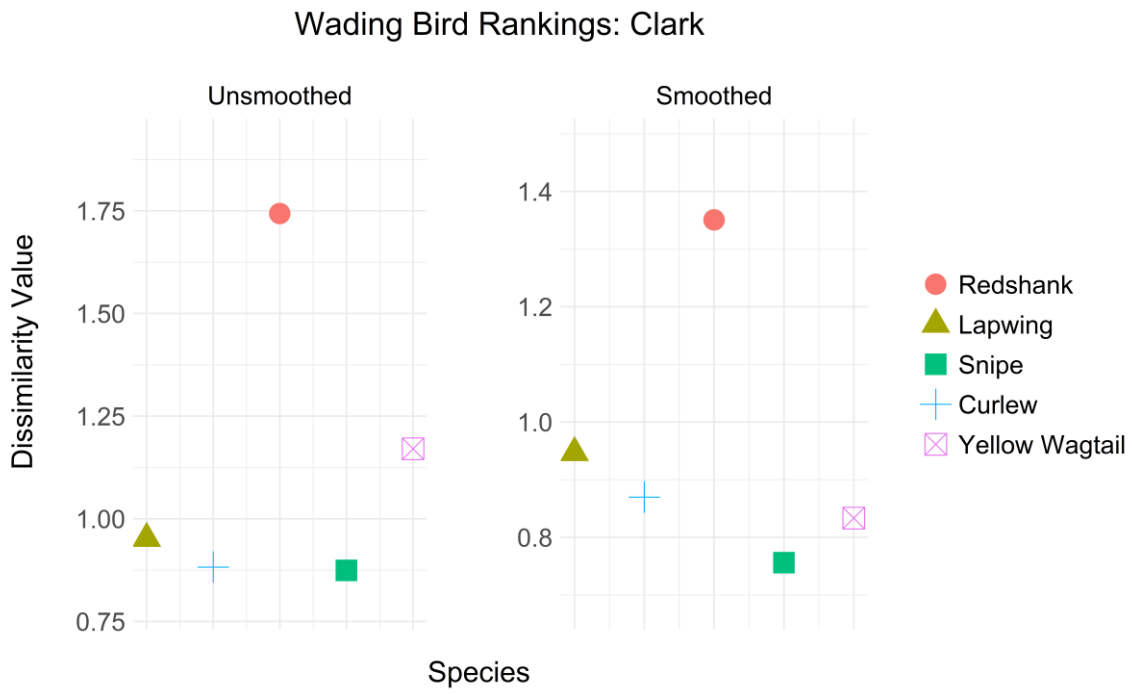


Figure S2.55. Dissimilarity values for trend comparisons of five wading bird species using the Clark Squared Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

Figure S2.56. Dissimilarity values for trend comparisons of five wading bird species using the Dissimilarity Index Combining Temporal Correlation and Raw Value Behaviour. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

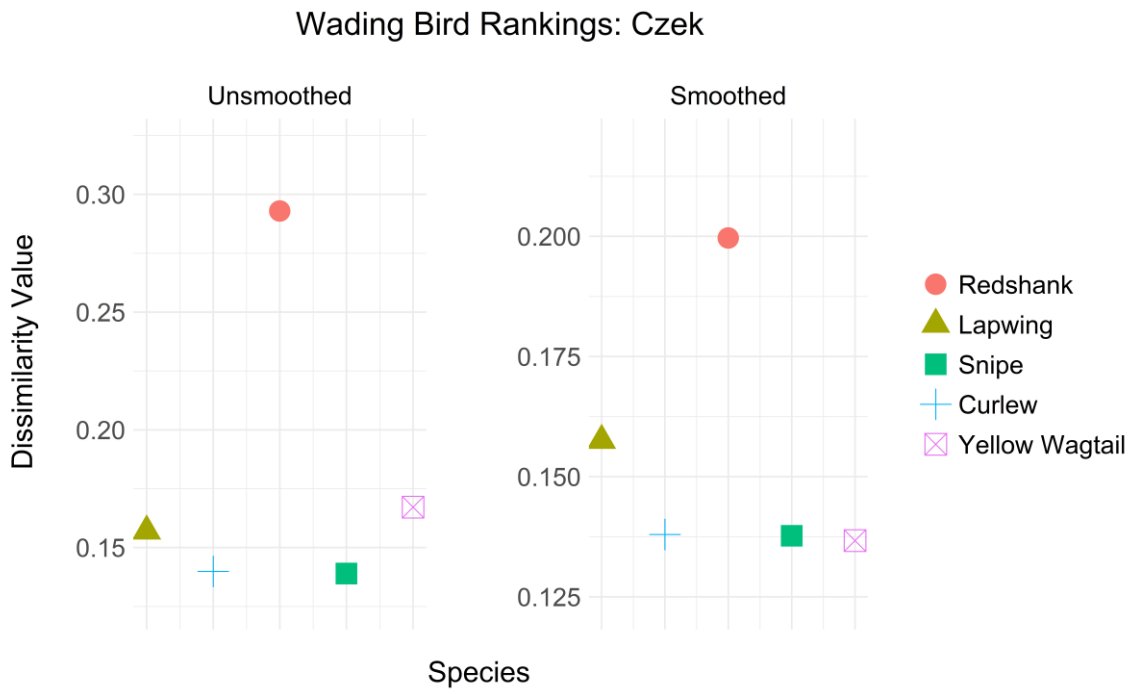


Figure S2.57. Dissimilarity values for trend comparisons of five wading bird species using the Czekanowski Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

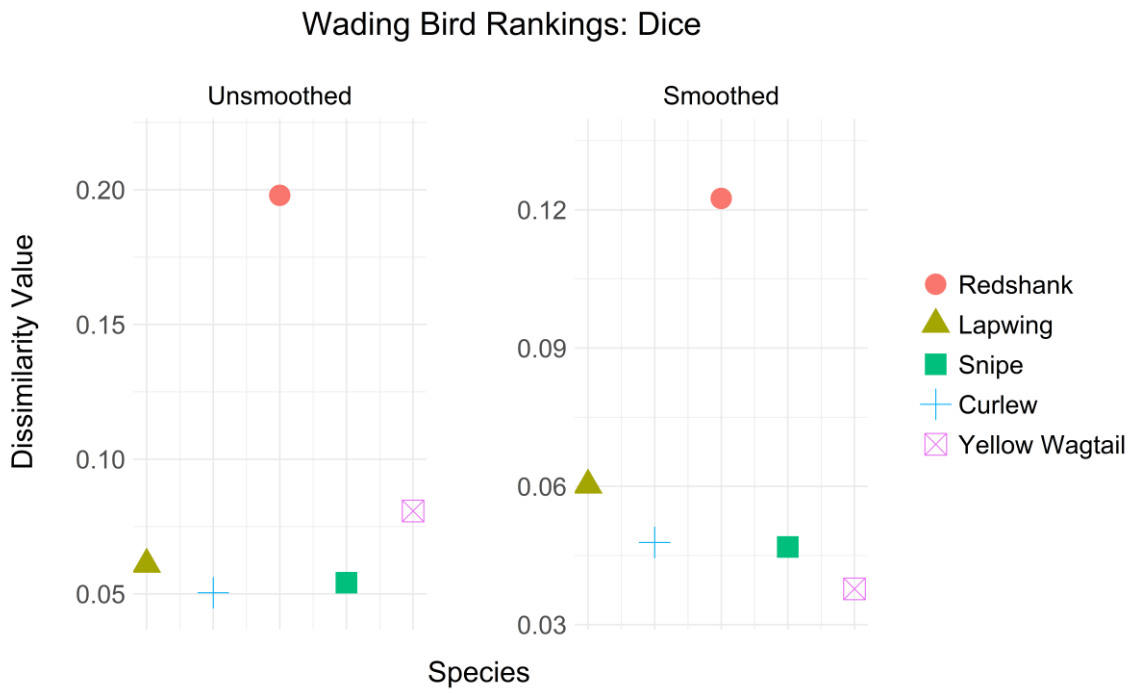


Figure S2.58. Dissimilarity values for trend comparisons of five wading bird species using the Dice Dissimilarity. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

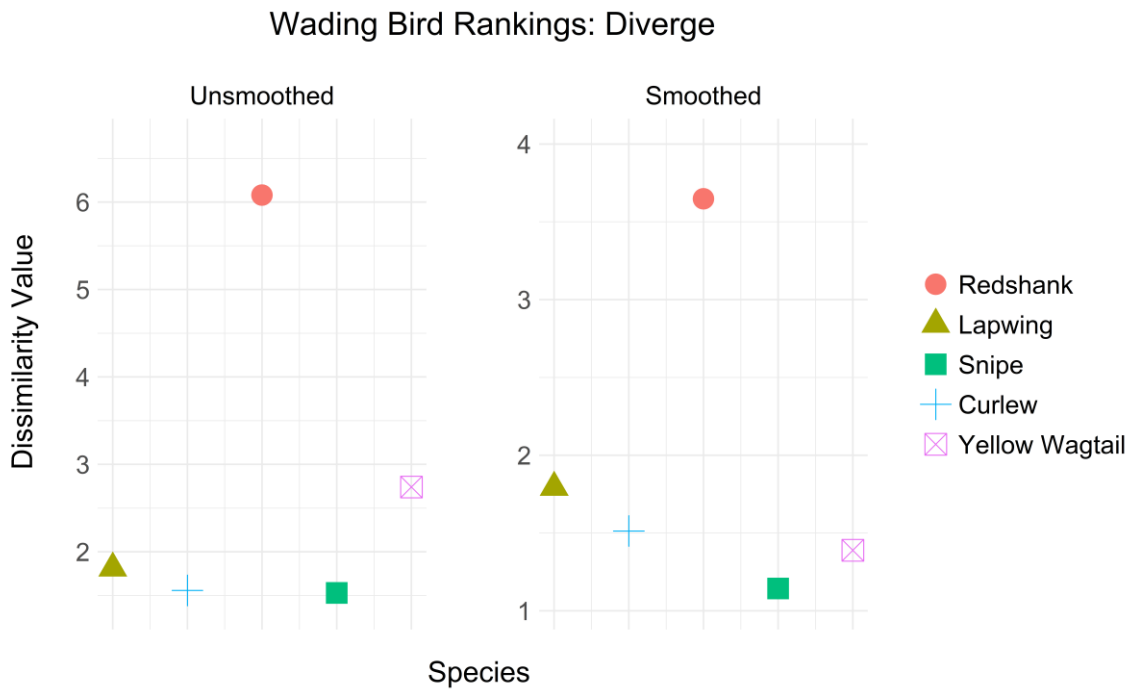


Figure S2.59. Dissimilarity values for trend comparisons of five wading bird species using the Divergence Squared Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

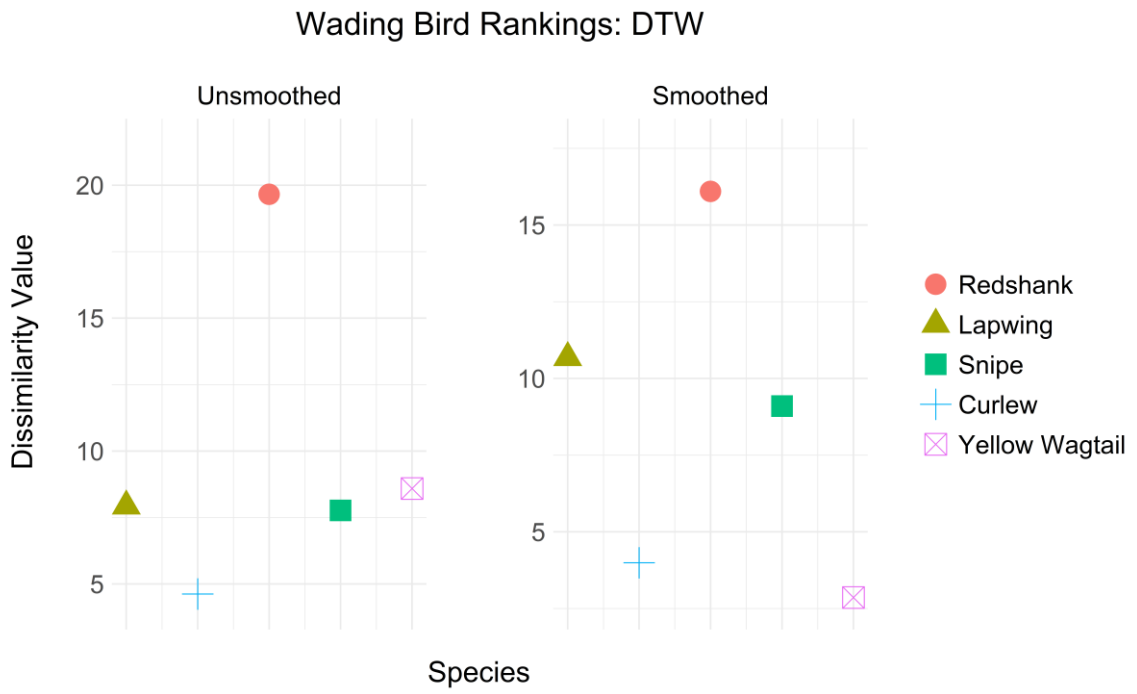


Figure S2.60. Dissimilarity values for trend comparisons of five wading bird species using the Dynamic Time Warping Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

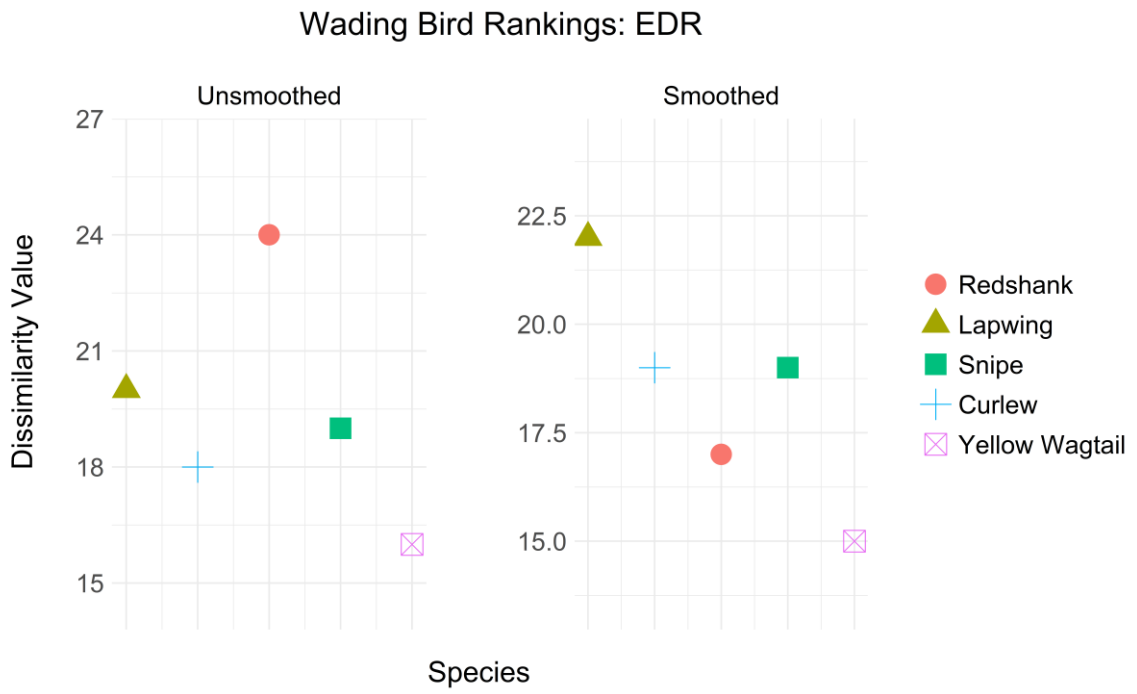


Figure S2.61. Dissimilarity values for trend comparisons of five wading bird species using the Edit Distance on Real Sequences. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

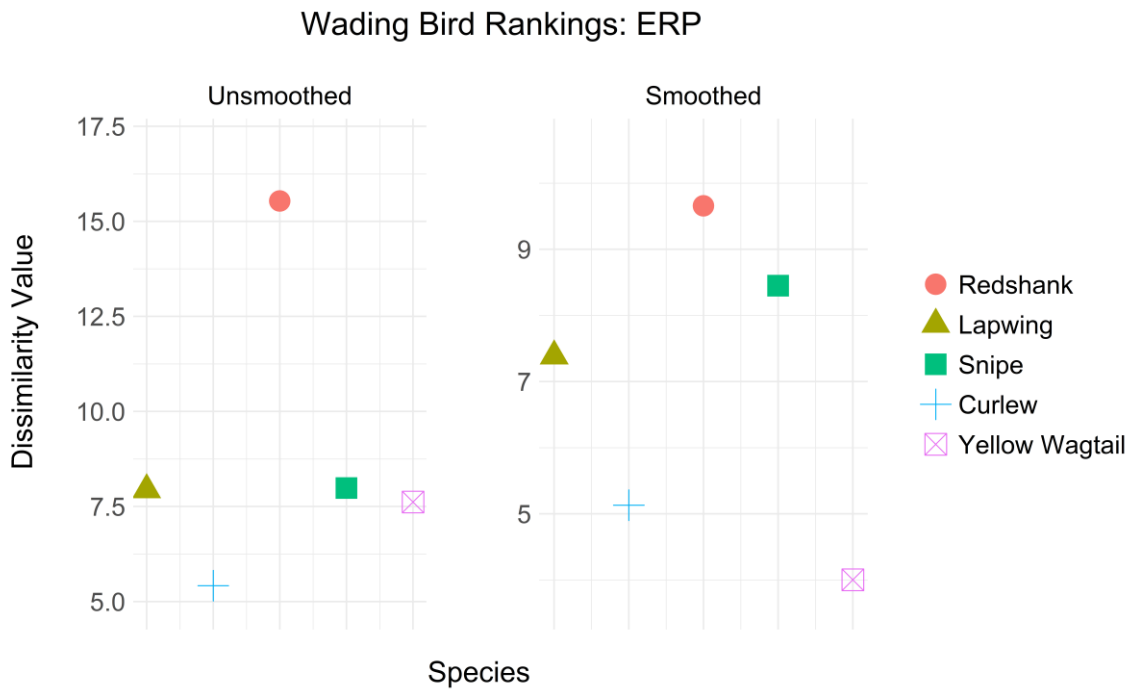


Figure S2.62. Dissimilarity values for trend comparisons of five wading bird species using the Edit Distance with Real Penalty. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

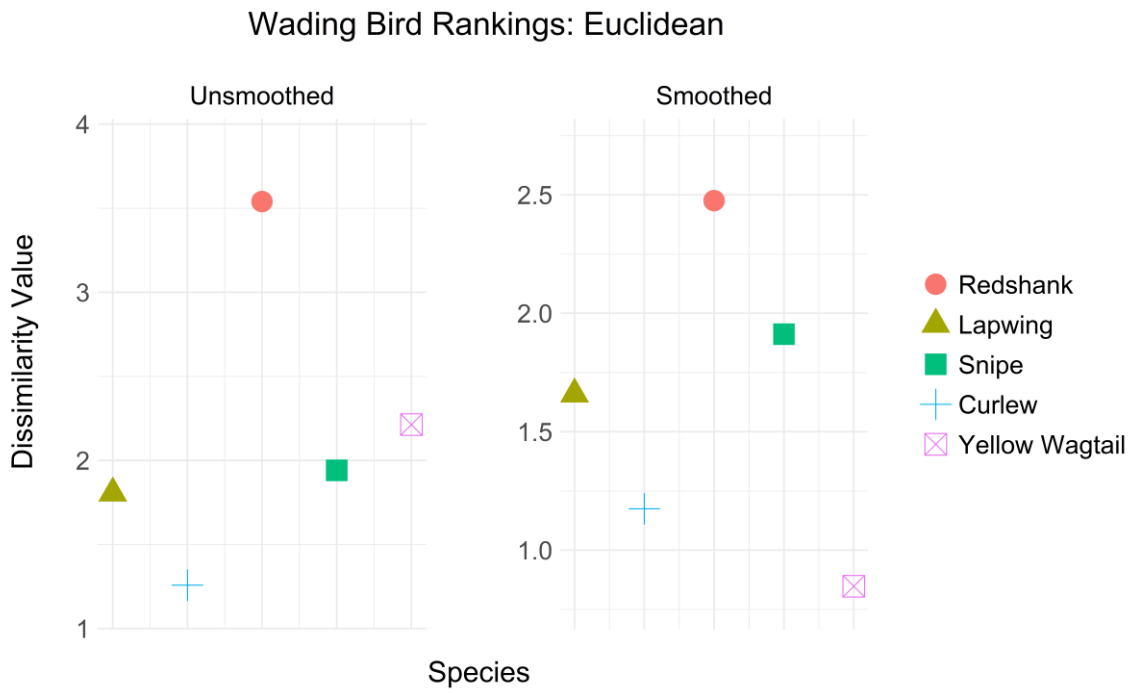


Figure S2.63. Dissimilarity values for trend comparisons of five wading bird species using the Euclidean Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

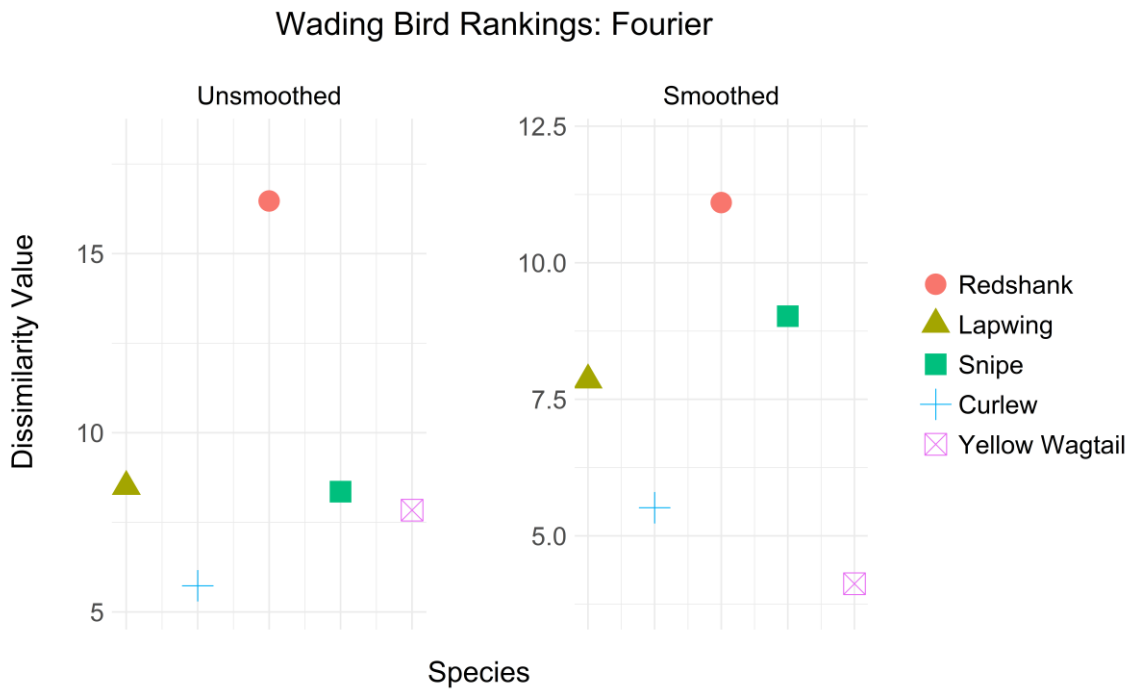


Figure S2.64. Dissimilarity values for trend comparisons of five wading bird species using the Fourier Coefficient-Based Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

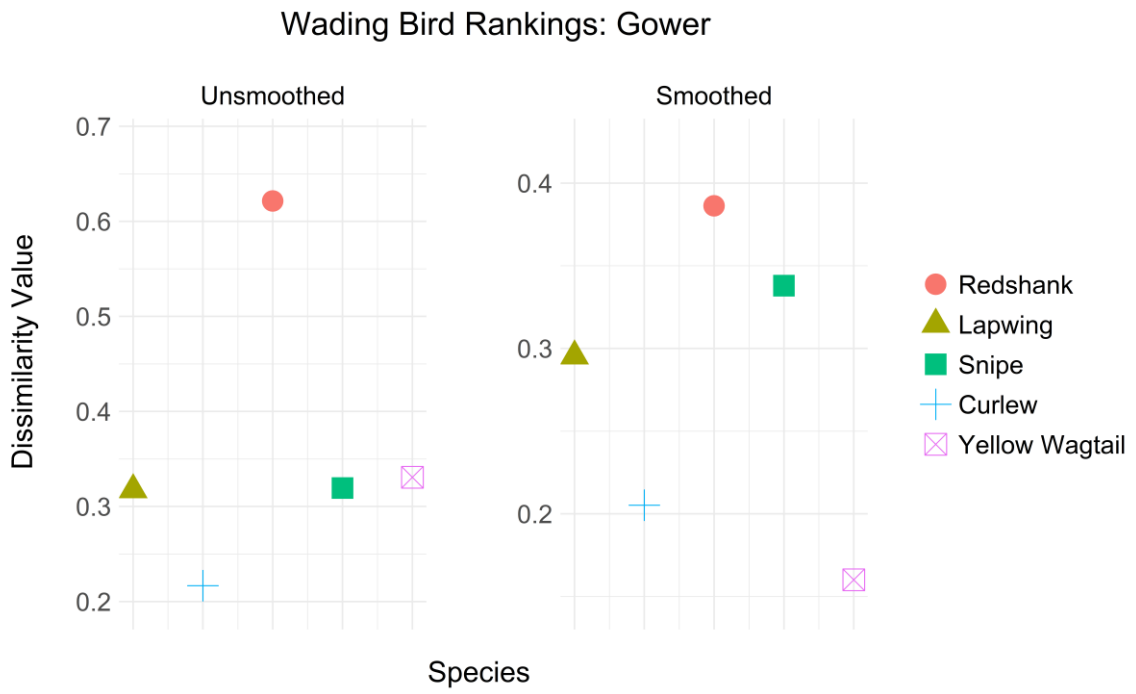


Figure S2.65. Dissimilarity values for trend comparisons of five wading bird species using the Gower Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

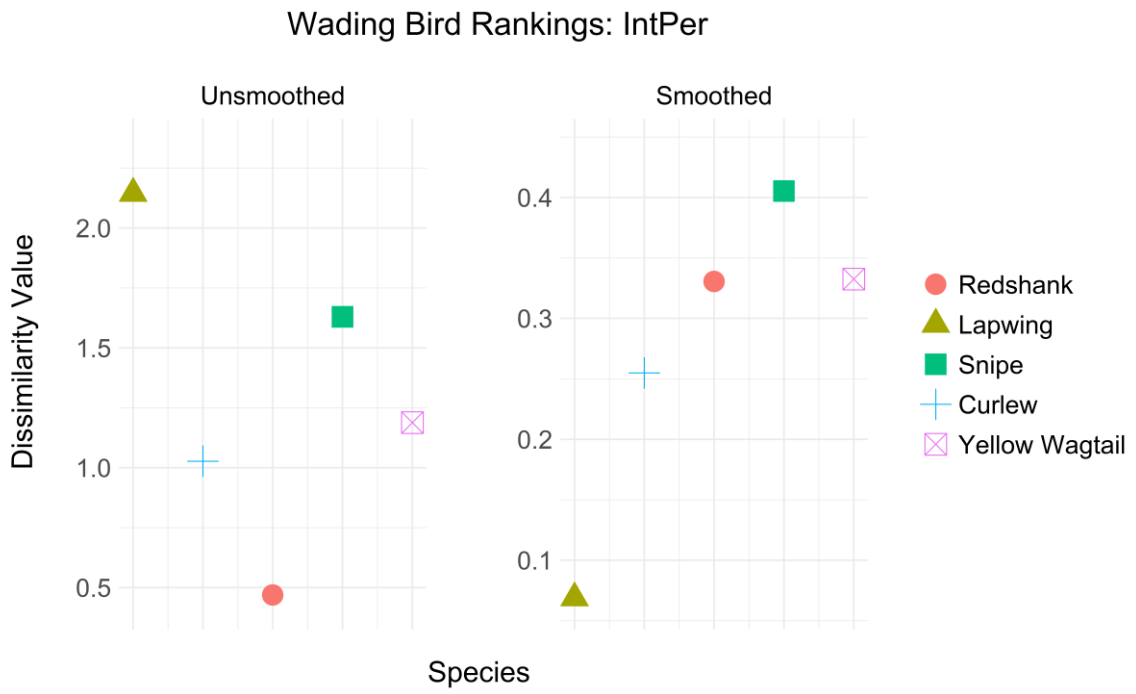


Figure S2.66. Dissimilarity values for trend comparisons of five wading bird species using the Integrated Periodogram-Based Dissimilarity. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

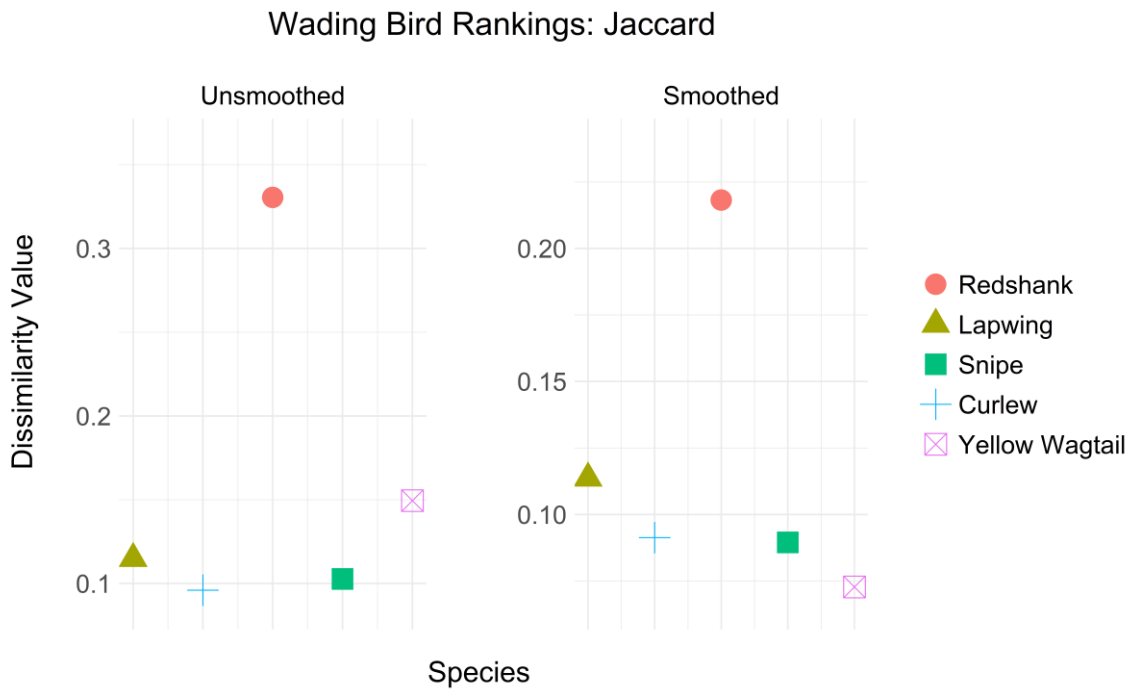


Figure S2.67. Dissimilarity values for trend comparisons of five wading bird species using the Jaccard Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

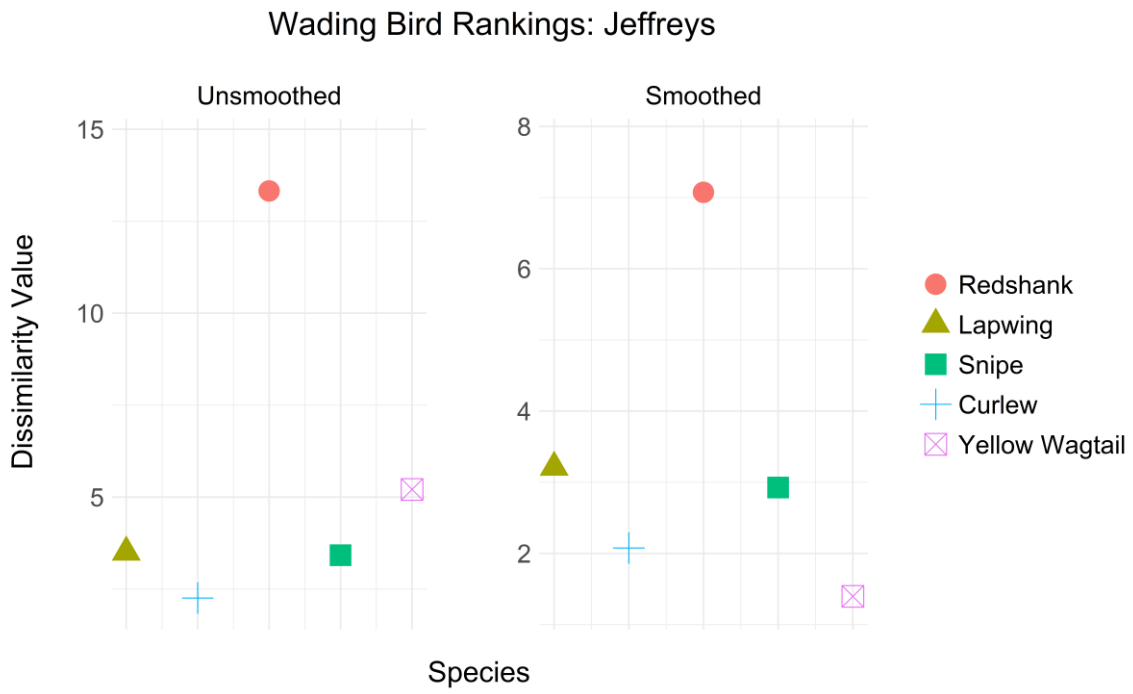


Figure S2.68. Dissimilarity values for trend comparisons of five wading bird species using the Jeffreys Divergence. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

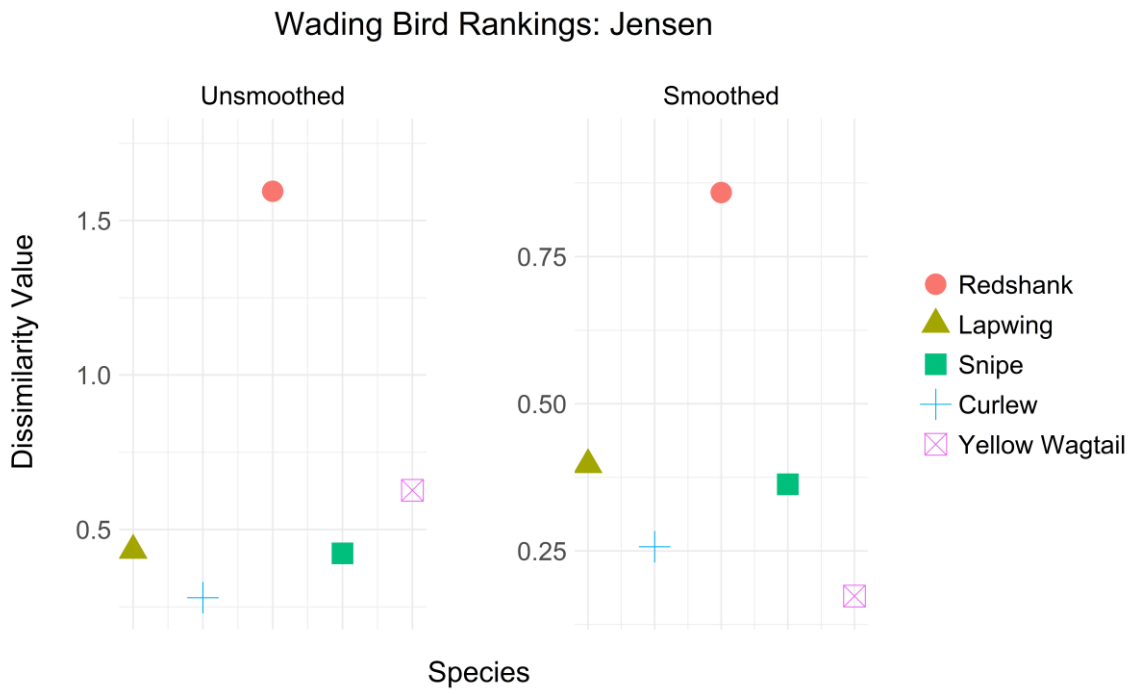


Figure S2.69. Dissimilarity values for trend comparisons of five wading bird species using the Jensen Difference. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

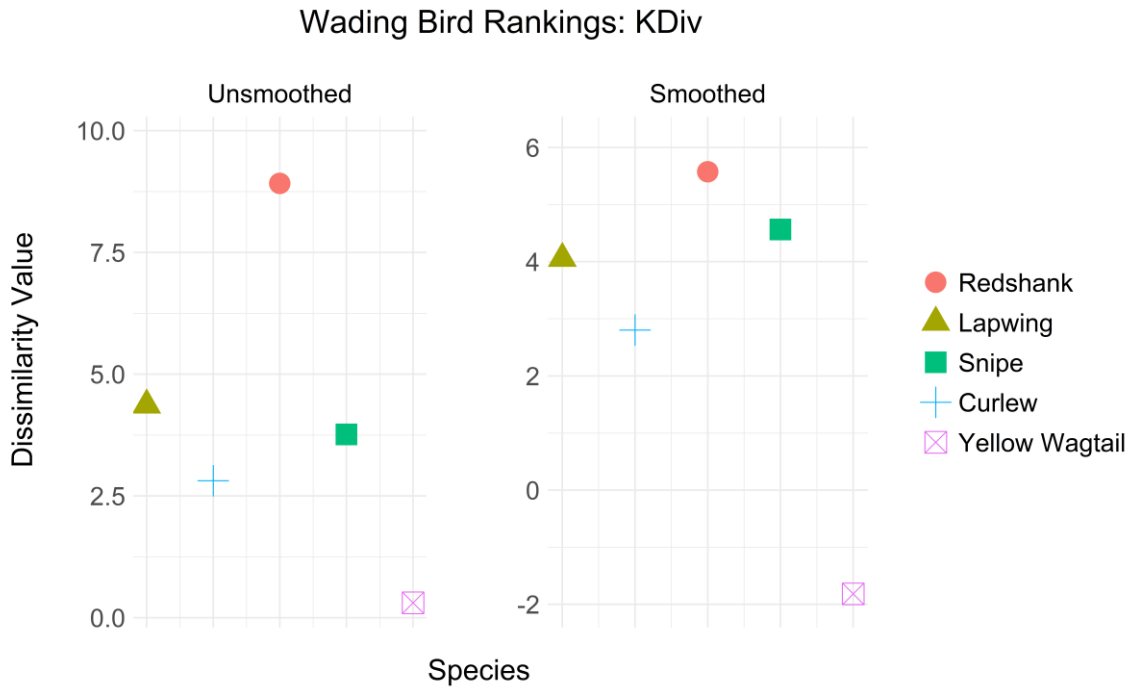


Figure S2.70. Dissimilarity values for trend comparisons of five wading bird species using the K Divergence. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

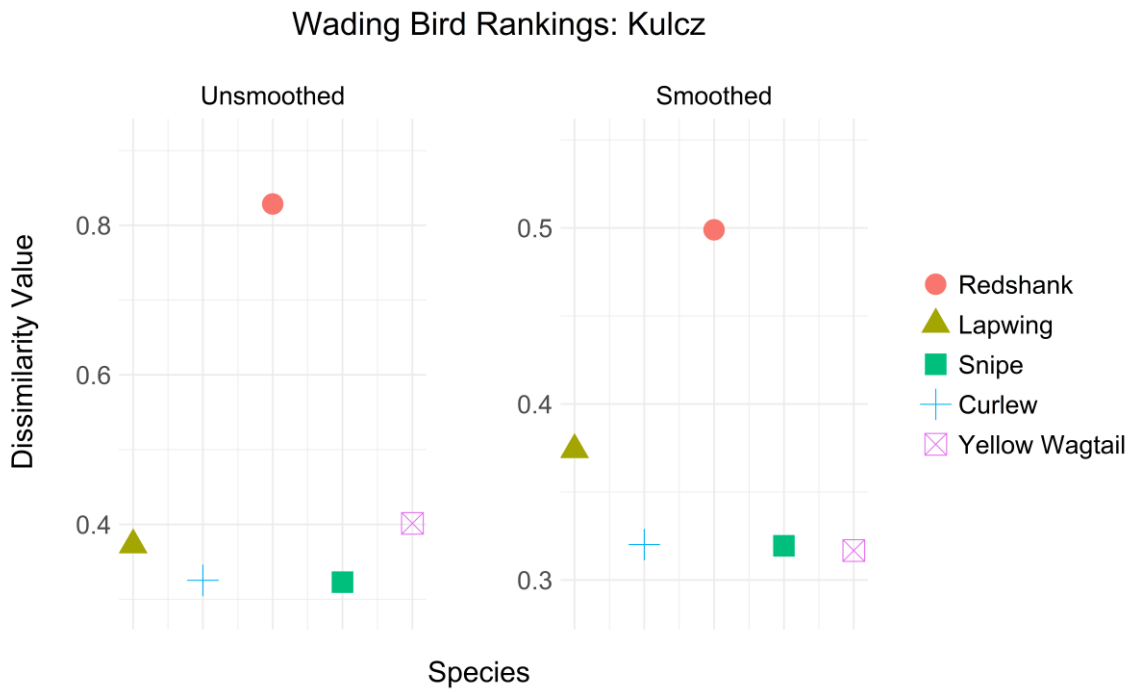


Figure S2.71. Dissimilarity values for trend comparisons of five wading bird species using the Kulczynski Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

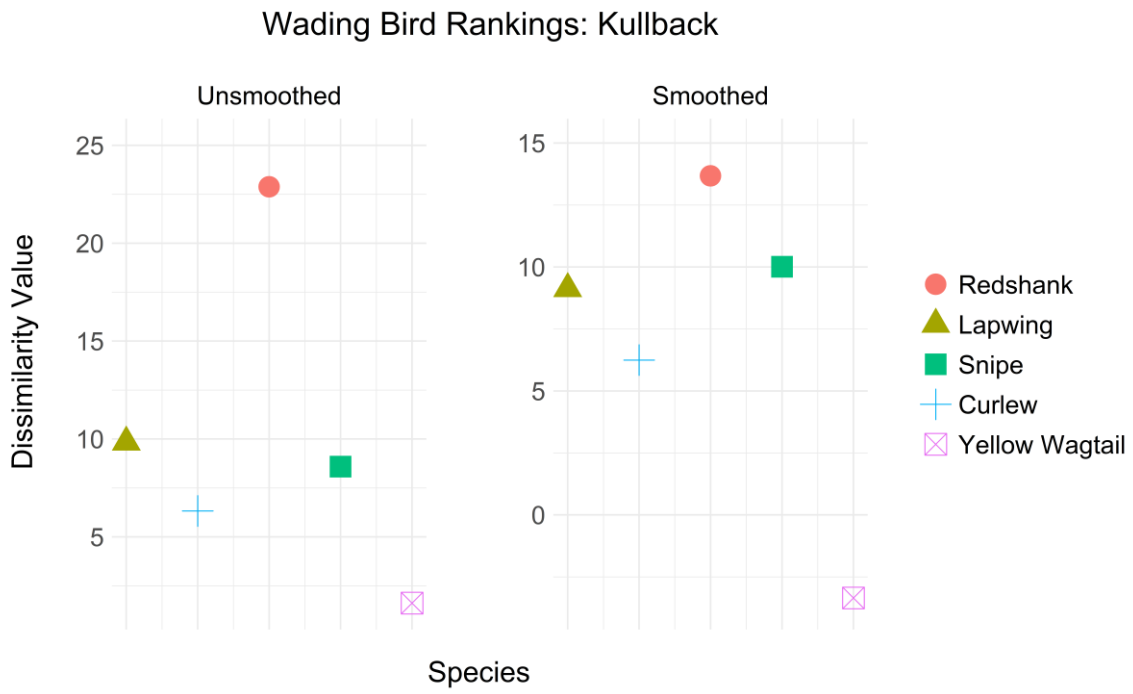


Figure S2.72. Dissimilarity values for trend comparisons of five wading bird species using the Kullback-Leibler Divergence. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

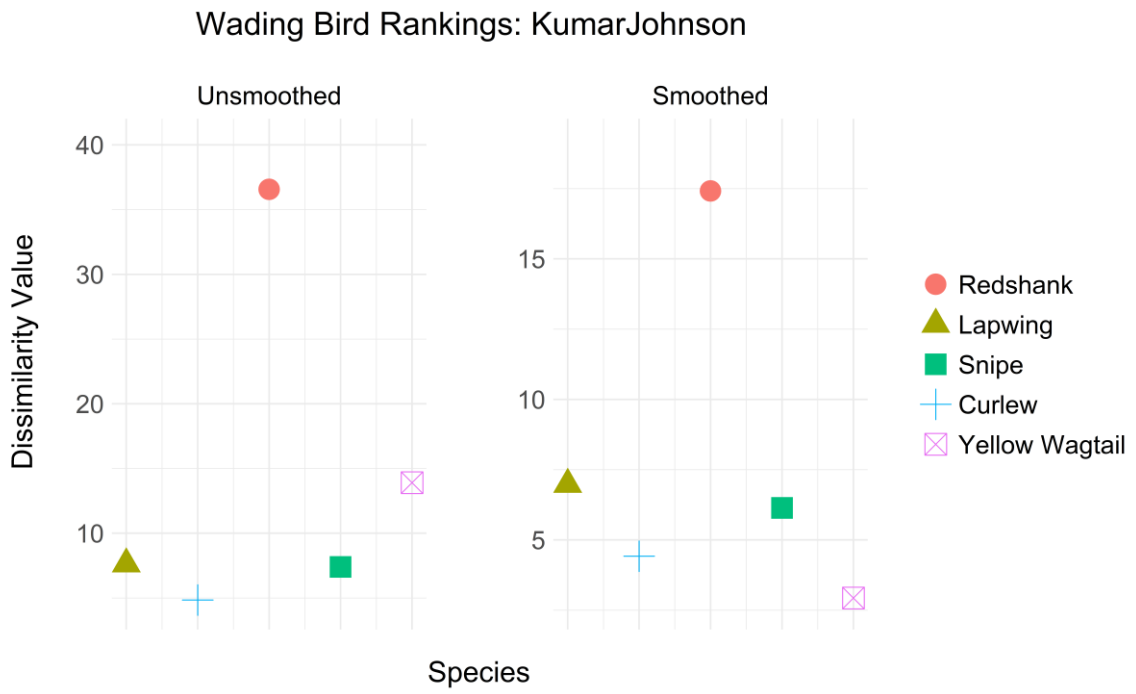


Figure S2.73. Dissimilarity values for trend comparisons of five wading bird species using the Kumar-Johnson Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

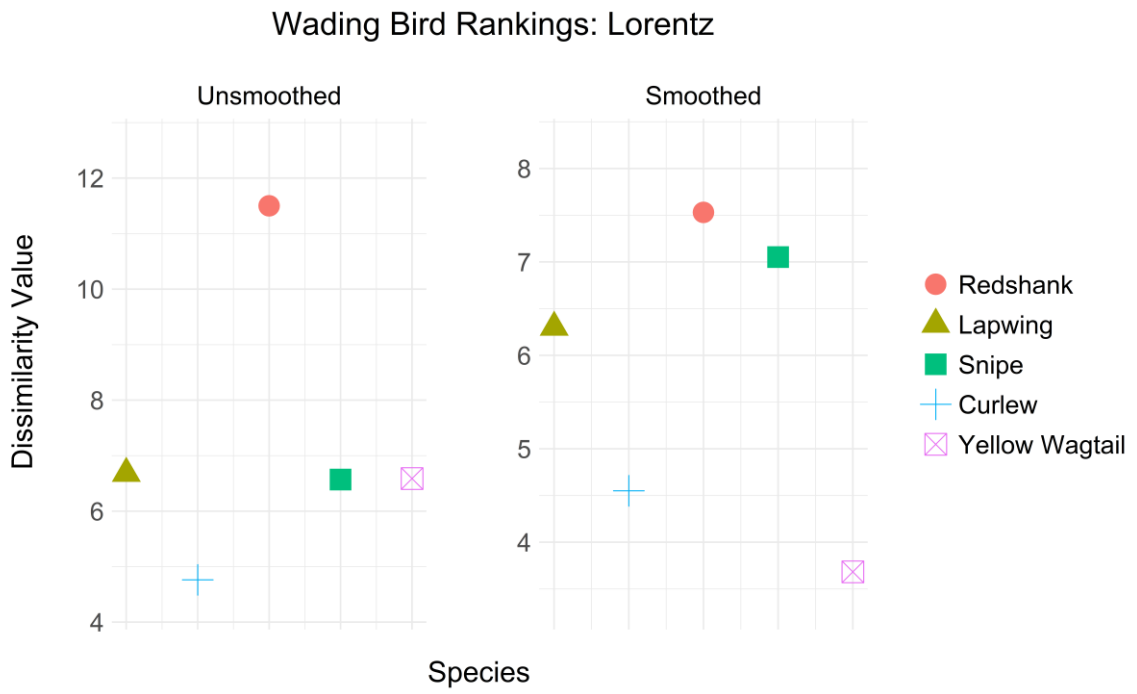


Figure S2.74. Dissimilarity values for trend comparisons of five wading bird species using the Lorentzian Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

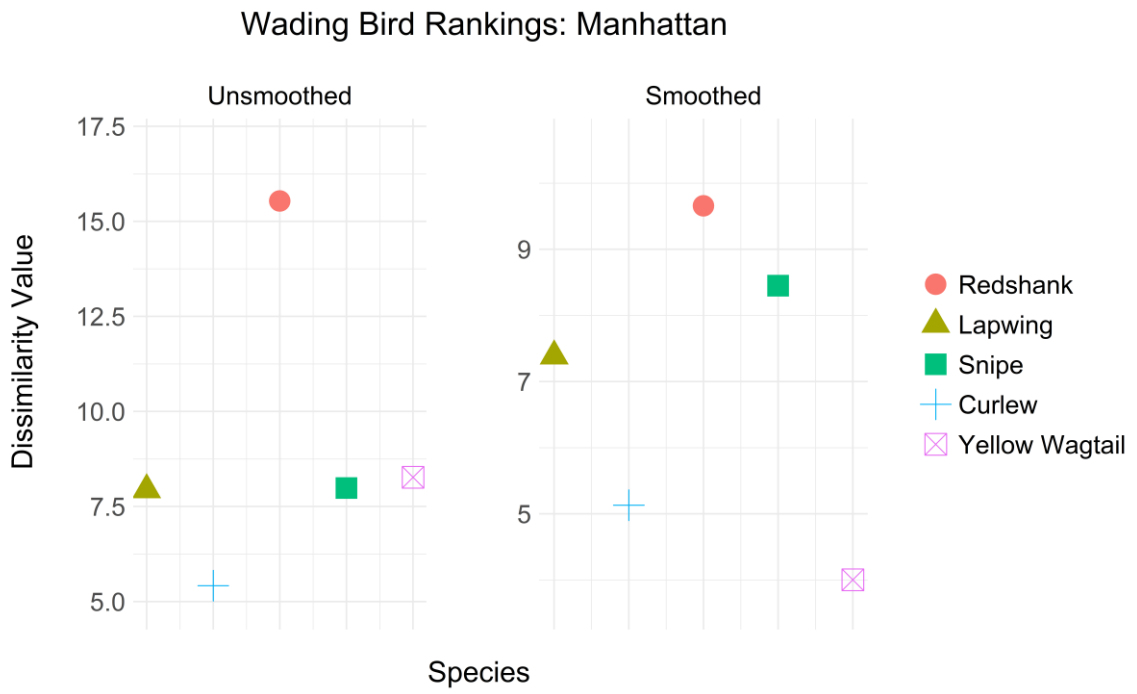


Figure S2.75. Dissimilarity values for trend comparisons of five wading bird species using the Manhattan Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

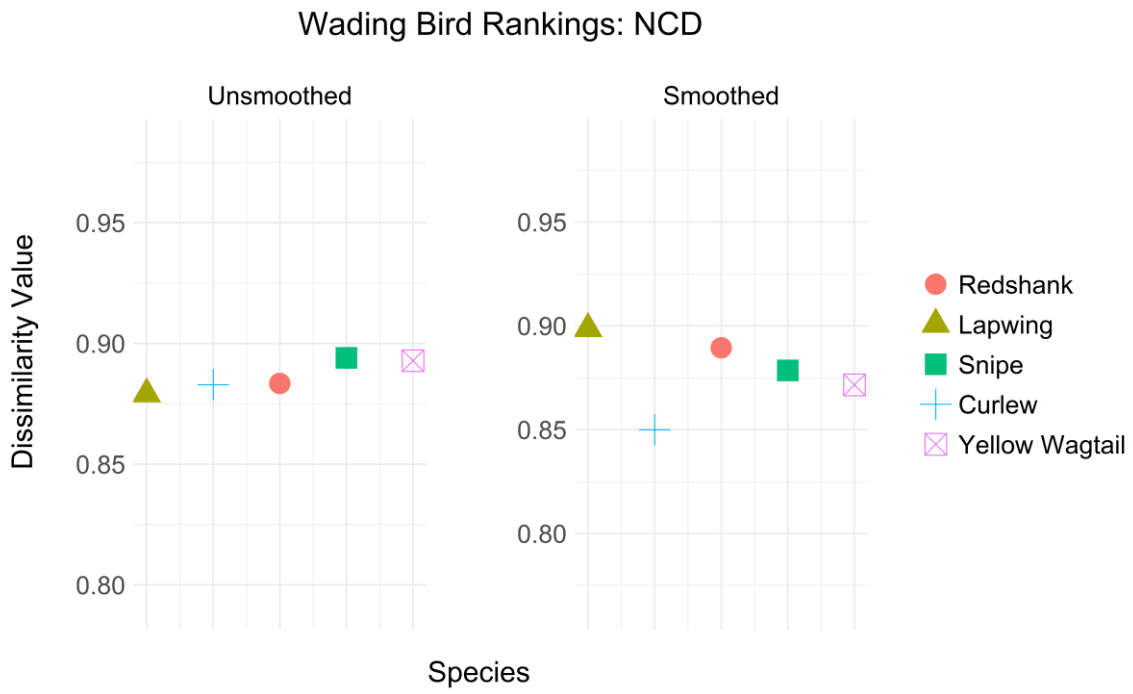


Figure S2.76. Dissimilarity values for trend comparisons of five wading bird species using the Normalized Compression Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

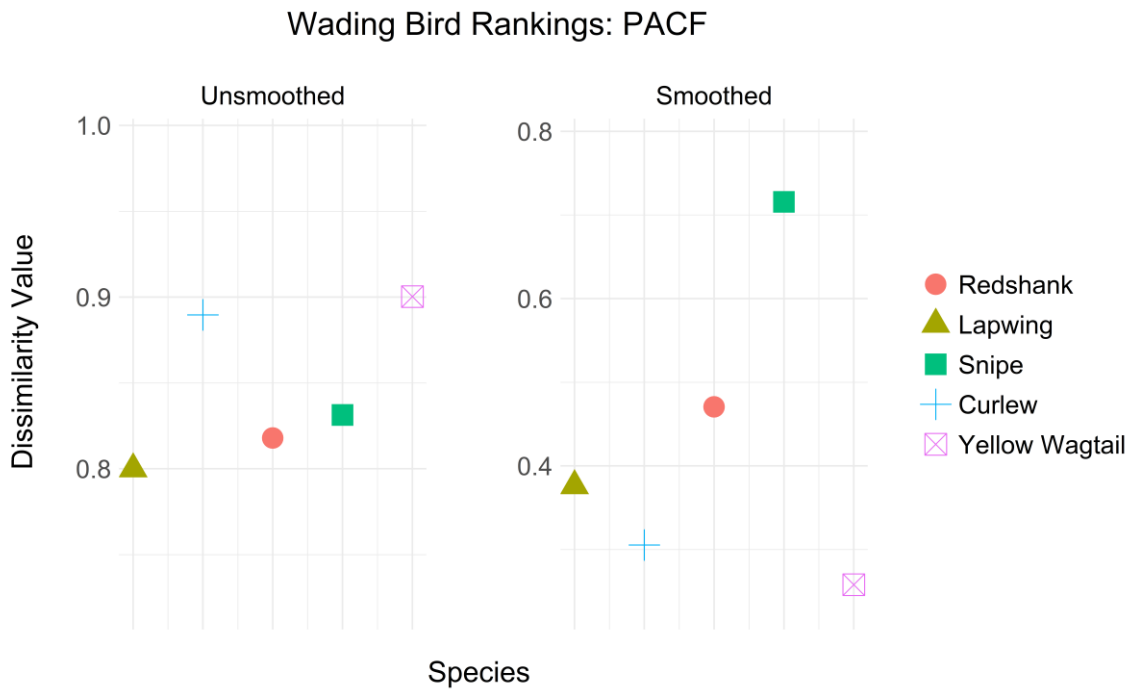


Figure S2.77. Dissimilarity values for trend comparisons of five wading bird species using the Partial Autocorrelation-Based Dissimilarity. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

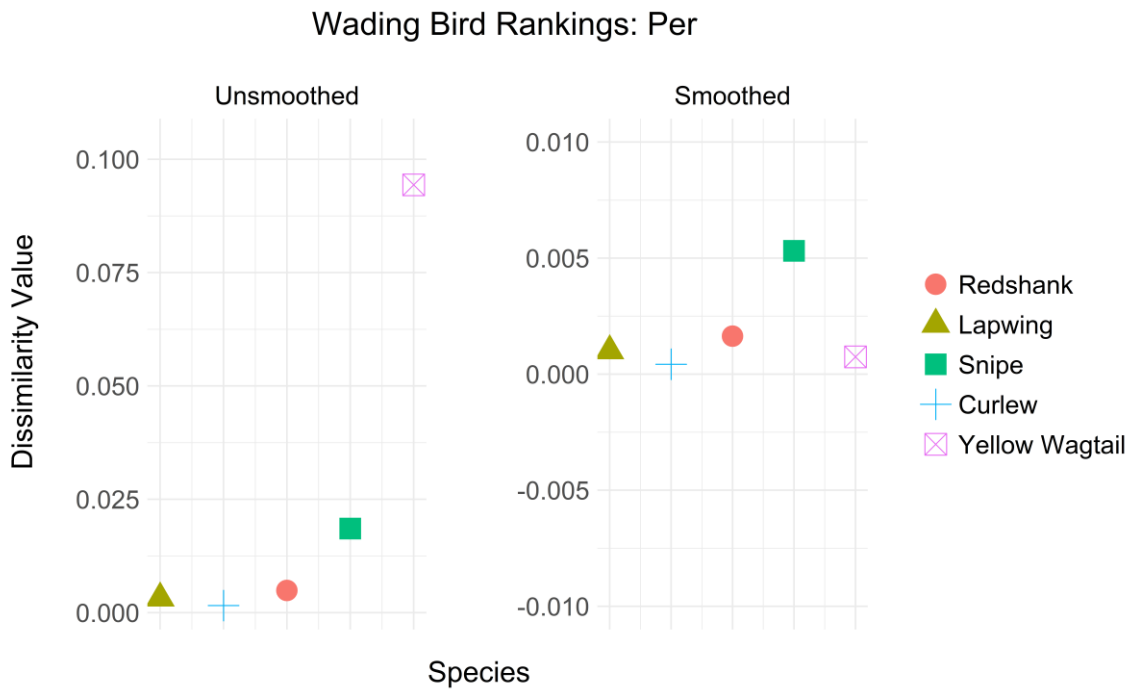


Figure S2.78. Dissimilarity values for trend comparisons of five wading bird species using the Periodogram-Based Dissimilarity. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

Figure S2.79. Dissimilarity values for trend comparisons of five wading bird species using the Piccolo Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

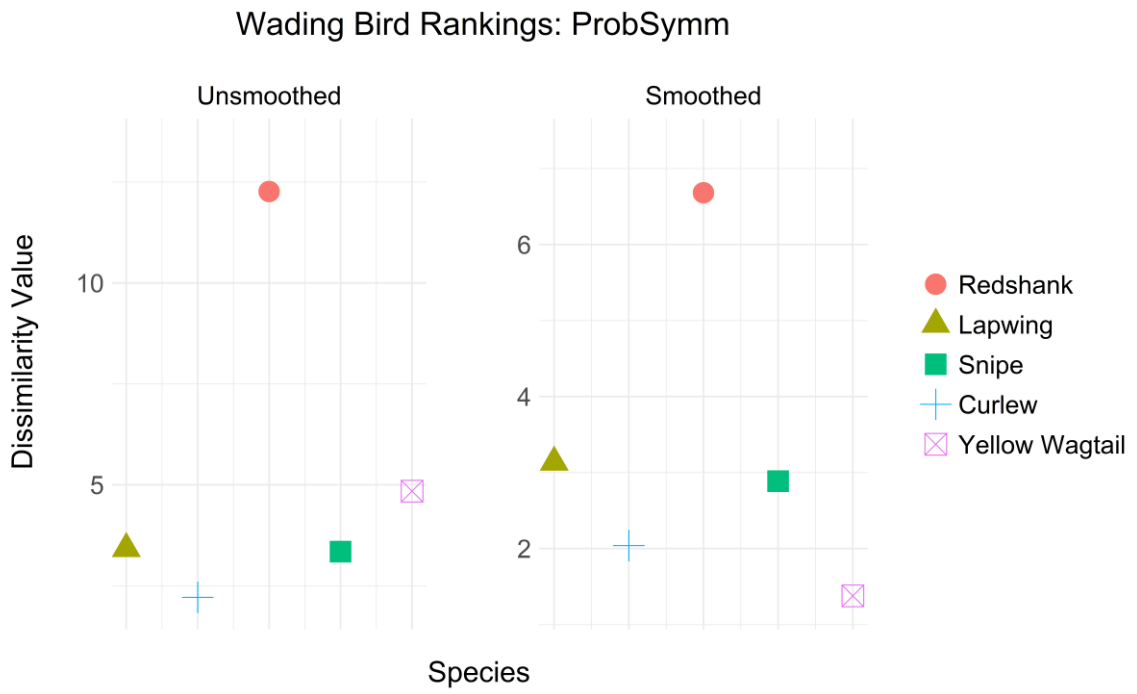


Figure S2.80. Dissimilarity values for trend comparisons of five wading bird species using the Probabilistic Symmetric Chi-Squared Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

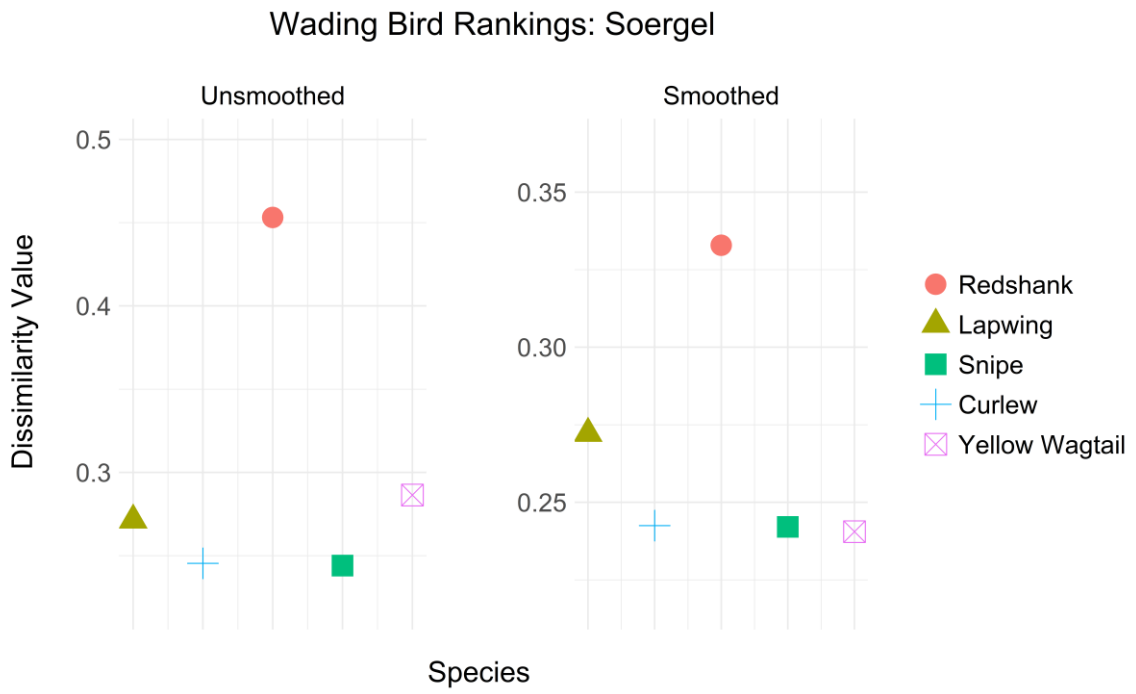


Figure S2.81. Dissimilarity values for trend comparisons of five wading bird species using the Soergel Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

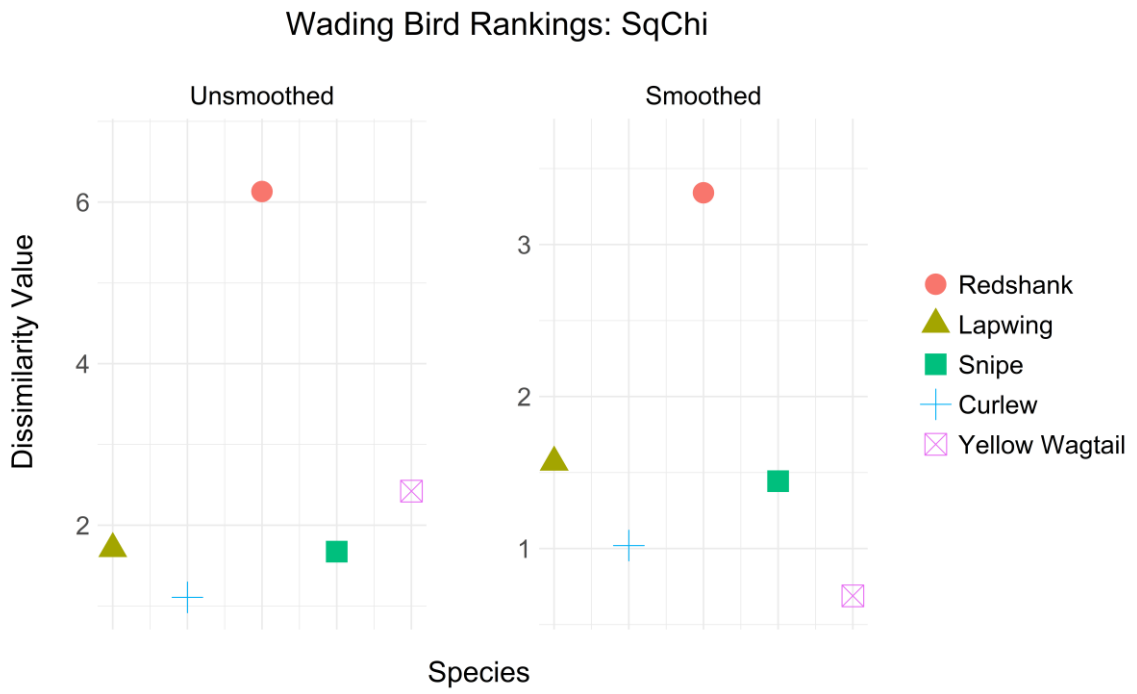


Figure S2.82. Dissimilarity values for trend comparisons of five wading bird species using the Squared Chi-Squared Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

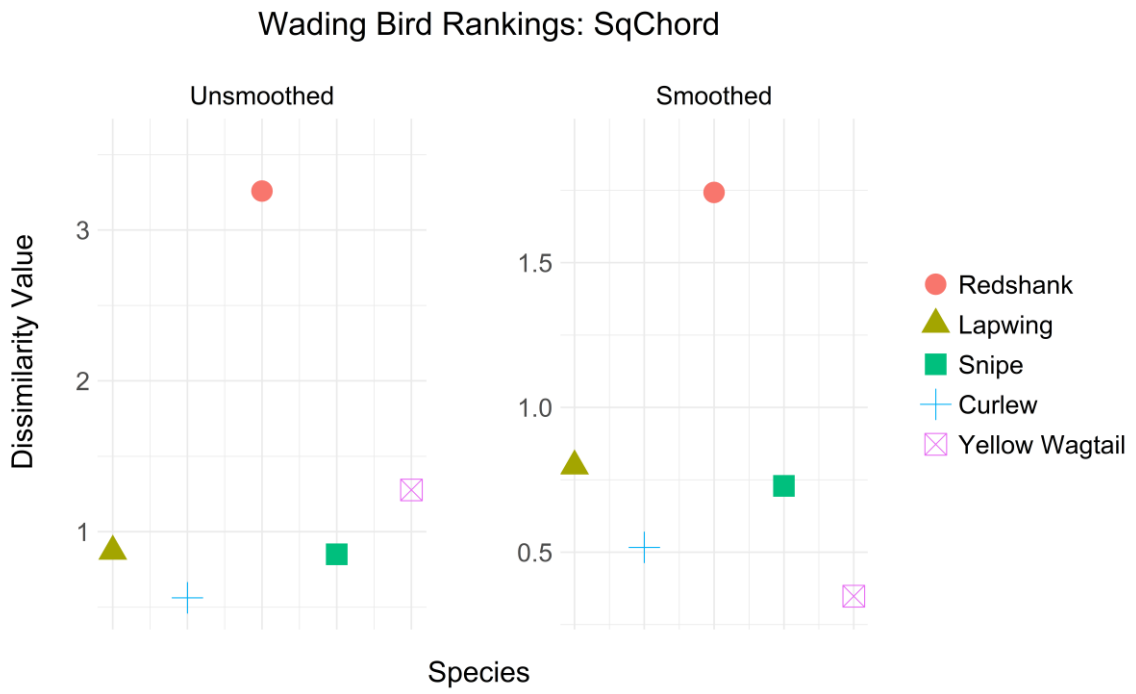


Figure S2.83. Dissimilarity values for trend comparisons of five wading bird species using the Squared-Chord Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

Figure S2.84. Dissimilarity values for trend comparisons of five wading bird species using the Squared Euclidean Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

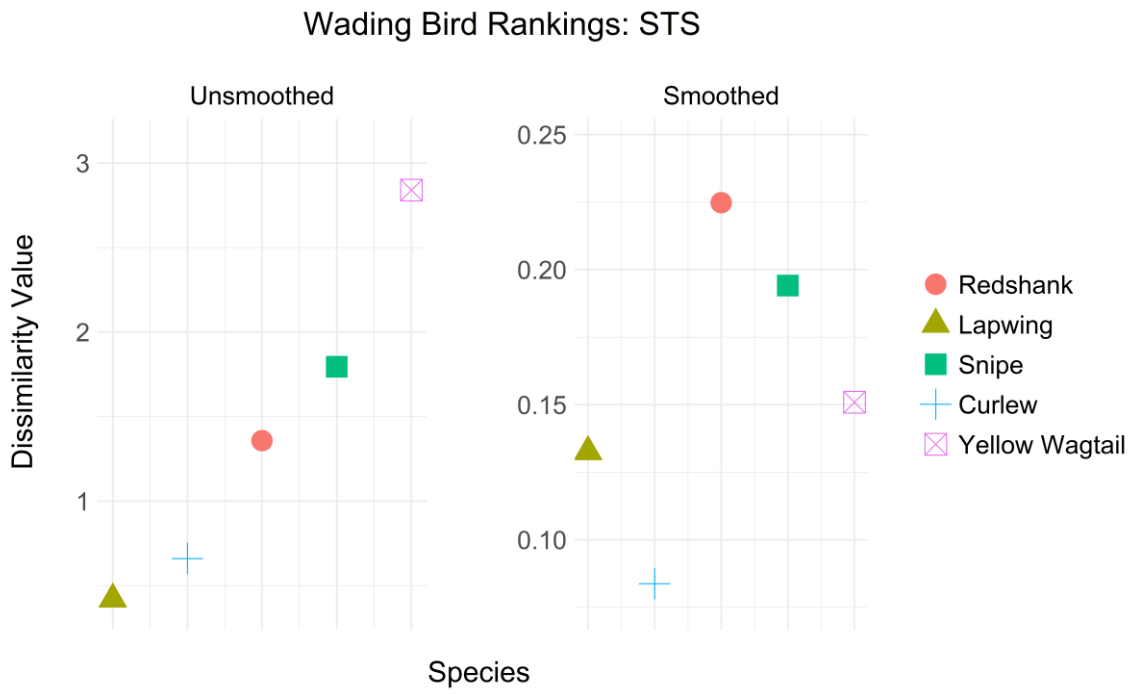


Figure S2.85. Dissimilarity values for trend comparisons of five wading bird species using the Short Time Series Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

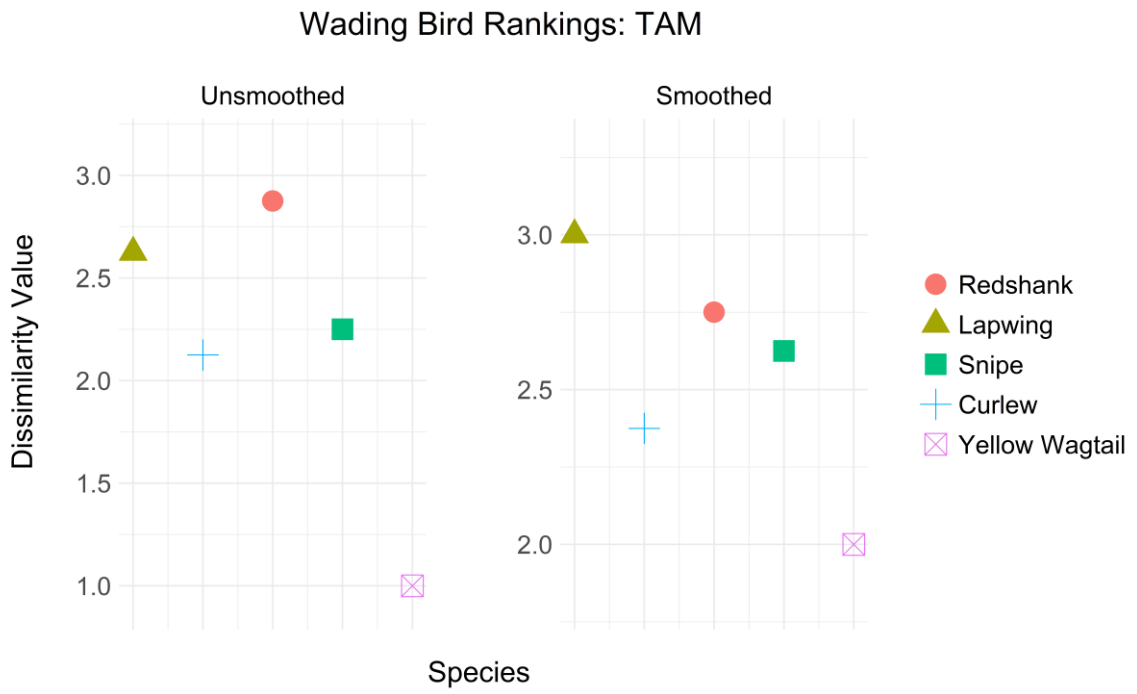


Figure S2.86. Dissimilarity values for trend comparisons of five wading bird species using the Time Alignment Measurement Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

Figure S2.87. Dissimilarity values for trend comparisons of five wading bird species using the Taneja Difference. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

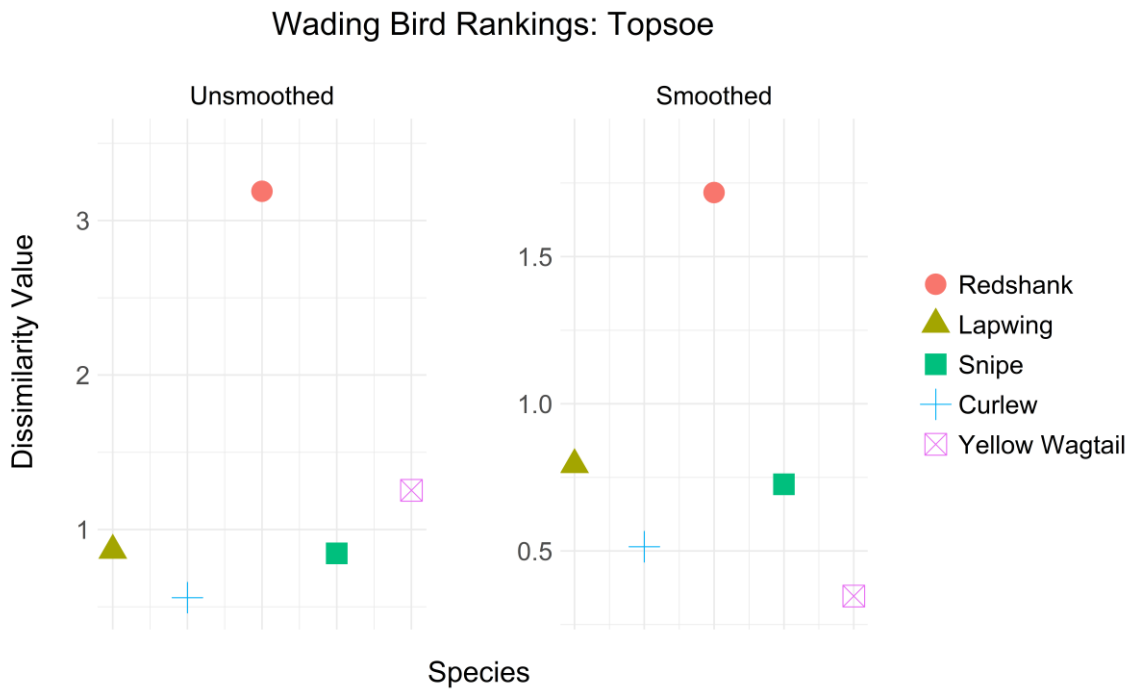


Figure S2.88. Dissimilarity values for trend comparisons of five wading bird species using the Topsoe Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

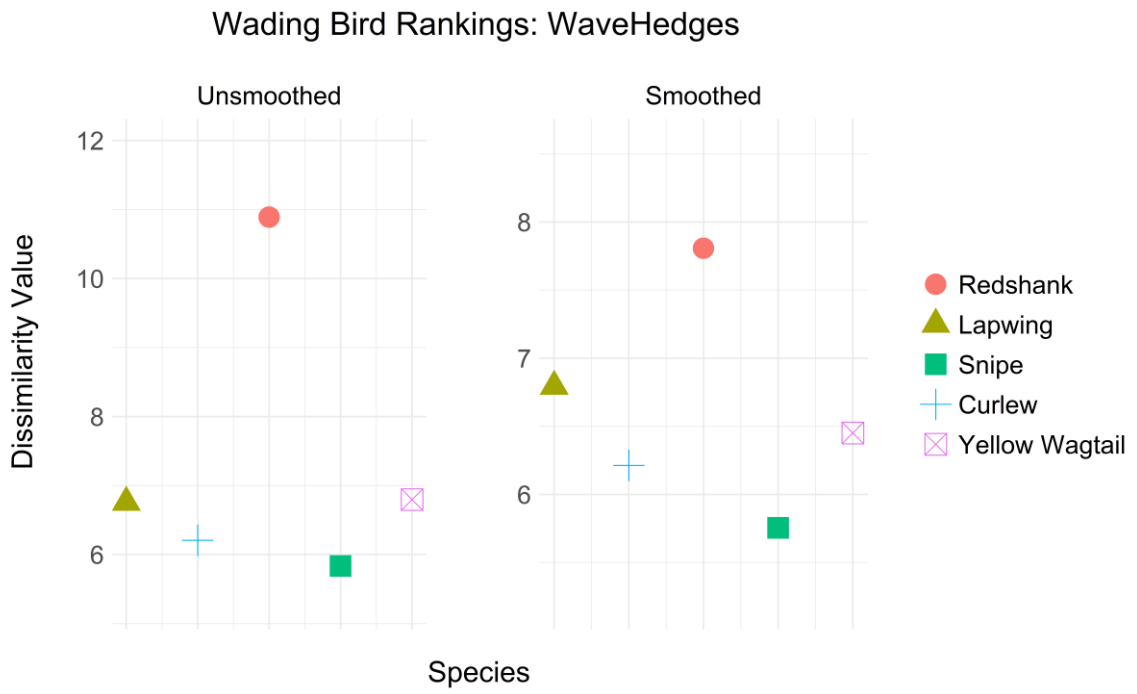


Figure S2.89. Dissimilarity values for trend comparisons of five wading bird species using the Wave-Hedges Distance. Trends within reserves were compared with counterfactual trends from outside of reserves. Values on the left were from comparisons of the unsmoothed trends, while values on the right were calculated after applying LOESS smoothing with a span setting of 0.75. The dataset is from a study of conservation impact of wet grassland reserves on breeding birds in the UK (Jellesmark et al., 2021).

S2.12. Bibliography for Appendix 1

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. In *International conference on foundations of data organization and algorithms* (pp. 69-84). Springer, Berlin, Heidelberg. DOI: 10.1007/3-540-57301-1_5

Batista, G. E., Wang, X., & Keogh, E. J. (2011). A complexity-invariant distance measure for time series. In *Proceedings of the 2011 SIAM international conference on data mining* (pp. 699-710). Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611972818.60

Brandmaier, A. M. (2015). pdc: An R package for complexity-based clustering of time series. *Journal of Statistical Software*, *67*(5), 1-23. DOI: 10.18637/jss.v067.i05

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, *50*(10), 2668-2684. DOI: 10.1016/j.csda.2005.04.012

Casado de Lucas, D. (2010). *Classification techniques for time series and functional data* (Doctoral dissertation, Universidad Carlos III de Madrid).

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1

Chen, L., & Ng, R. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 792-803).

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 491-502). DOI: 10.1145/1066157.1066213

Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification, 1*(1), 5-21. DOI: 10.1007/s11634-006-0004-6

Cilibrasi, R., & Vitányi, P. M. (2005). Clustering by compression. *IEEE Transactions on Information theory*, *51*(4), 1523-1545. DOI: 10.1109/TIT.2005.844059

D'Urso, P., & Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, *160*(24), 3565-3589. DOI: 10.1016/j.fss.2009.04.013

Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S. , Ratanamahatana, C.A., Chen, Y., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., & Hexagon, M.L. (2019). *The UCR Time Series Classification Archive.* DOI: 10.1109/JAS.2019.1911747 URL: https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

Díaz, S. P., & Vilar, J. A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of classification*, *27*(3), 333-362. DOI: 10.1007/s00357-010-9064-6

Folgado, D., Barandas, M., Matias, R., Martins, R., Carvalho, M., & Gamboa, H. (2018). Time alignment measurement for time series. *Pattern Recognition*, *81*, 268-279. DOI: 10.1016/j.patcog.2018.04.003

Frentzos, E., Gratsias, K., & Theodoridis, Y. (2007, April). Index-based most similar trajectory search. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 816-825). IEEE. DOI: 10.1109/ICDE.2007.367927

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., & Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for FMRI. *Magnetic resonance in medicine*, *40*(2), 249-260.

Keogh, E., Lonardi, S., & Ratanamahatana, C. A. (2004). Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 206-215). DOI: 10.1145/1014052.1014077

Keogh, E., & Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and information systems*, *7*(3), 358-386. DOI: 10.1007/s10115-004-0154-9

Möller-Levet, C. S., Klawonn, F., Cho, K. H., & Wolkenhauer, O. (2003, August). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International symposium on intelligent data analysis* (pp. 330-340). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-45231-7_31

Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, *62*(1), 1-43.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance Measures for Time Series in R: The TSdist Package. *R J.*, *8*(2), 451-459. https://journal.r-project.org/archive/2016/RJ-2016-058/index.html

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, *11*(2), 153-164. DOI: 10.1111/j.1467-9892.1990.tb00048.x

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*(1), 43-49. DOI: 10.1109/TASSP.1978.1163055

Appendix 2: Supplementary materials for Chapter 3

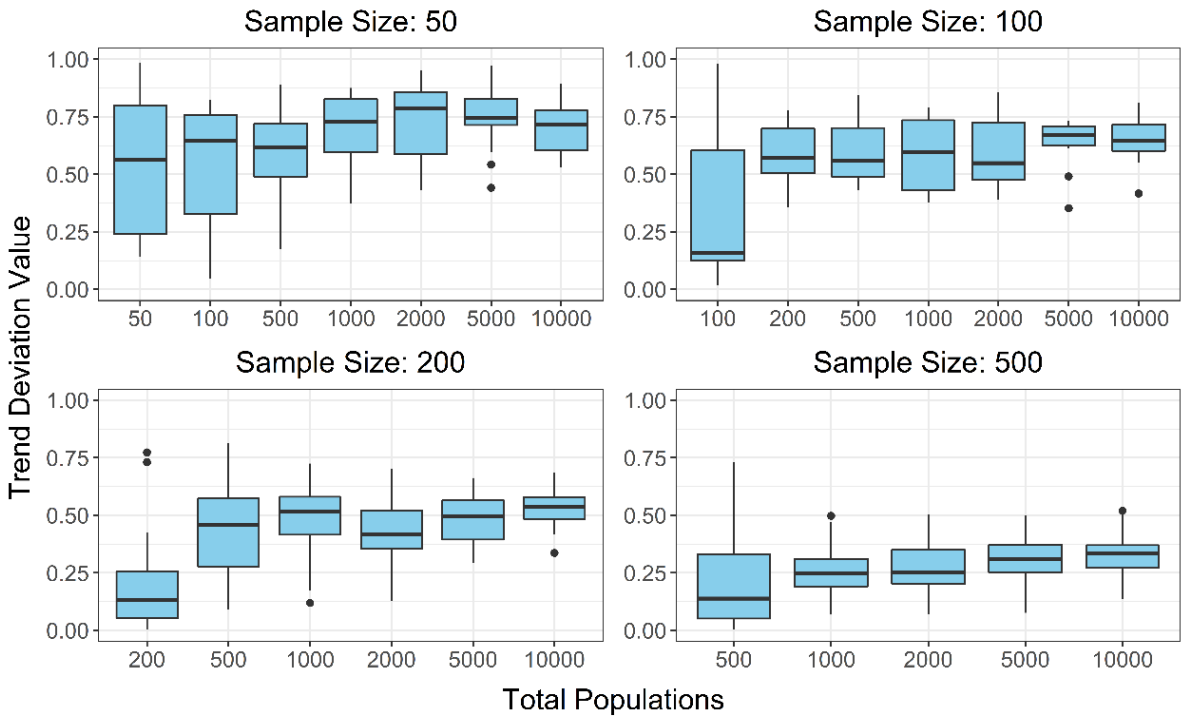


Figure S3.1. Trend deviation value (TDV) vs total number of populations in dataset, at different sample sizes. All other parameters are fixed – μ_{ds}: 0; σ_{ds}: 0.4; μ_{η}: 0.6; populations per species: 20; mean time series length: 20; trend length: 50; μ_{ε}: 0; σ_{ε}: 0. Each box represents the mean values of 10 datasets, with 20 samples per dataset. There is no clear effect of dataset size on TDV. While TDV is lower when the dataset is equal to the sample size (TDV is non-zero due to degradation of the samples), there is no clear trend in median TDVs as the number of populations in the dataset increases from twice the size of the sample to 10,000 populations.

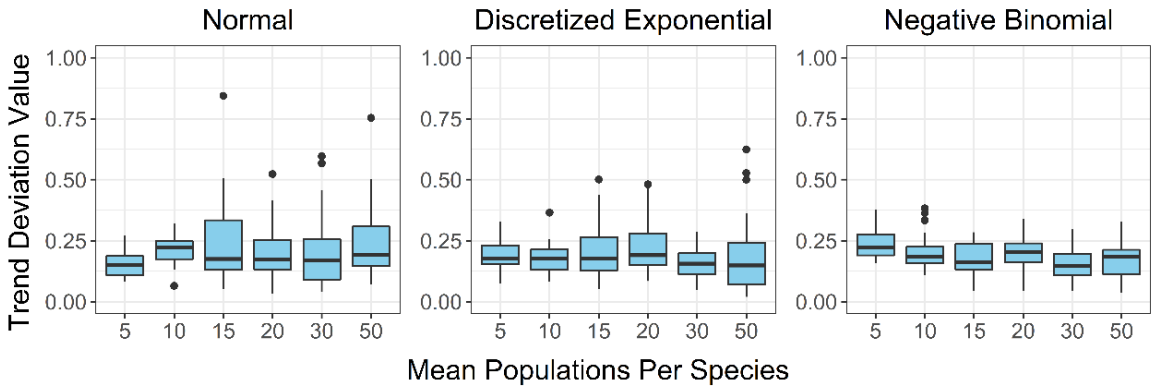


Figure S3.2. Trend deviation value vs mean number of populations assigned to each species, using three different distributions: a normal distribution; an exponential distribution discretized by rounding to the nearest whole number; and a negative binomial distribution with zeros removed by adding one to every value (the resulting increase of the mean was accounted for). All other parameters are fixed – dataset size: 1000; sample size: 200; μ_{ds}: 0; σ_{ds}: 0.2; μ_{η}: 0.2; mean time series length: 20; trend length: 50; μ_{ε}: 0.15; σ_{ε}: 0.1. Each box represents the mean values of 20 datasets, with 20 samples per dataset. Neither the mean number of populations assigned to each species, nor the distribution used to assign them, shows any effect on trend accuracy.

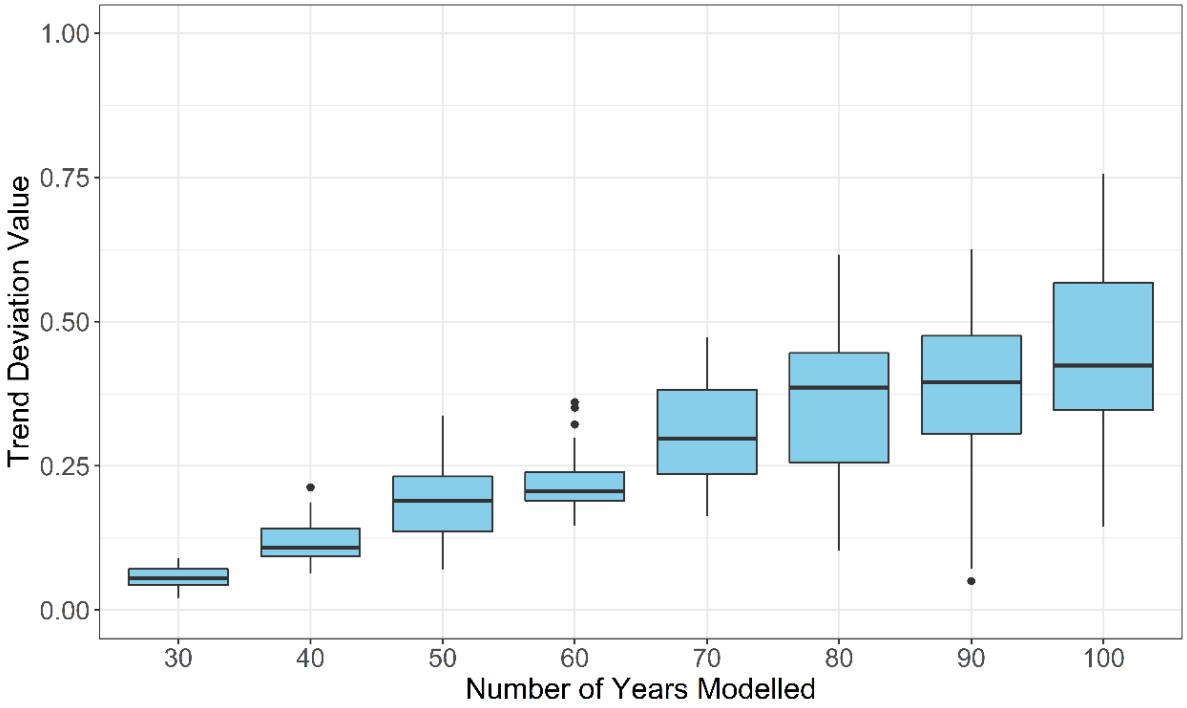


Figure S3.3. Trend deviation value vs length of trend. All other parameters are fixed – dataset size: 1000; sample size: 200; μ_{ds}: 0; σ_{ds}: 0.2; μ_{η}: 0.2; populations per species: 20; mean time series length: 20; μ_{ε}: 0.15; σ_{ε}: 0.1. Each box represents the mean values of 20 datasets, with 20 samples per dataset. The median and the range of the TDV increase as trend length increases; this likely occurs because the mean time series length and sample size are static, resulting in fewer observations per year.

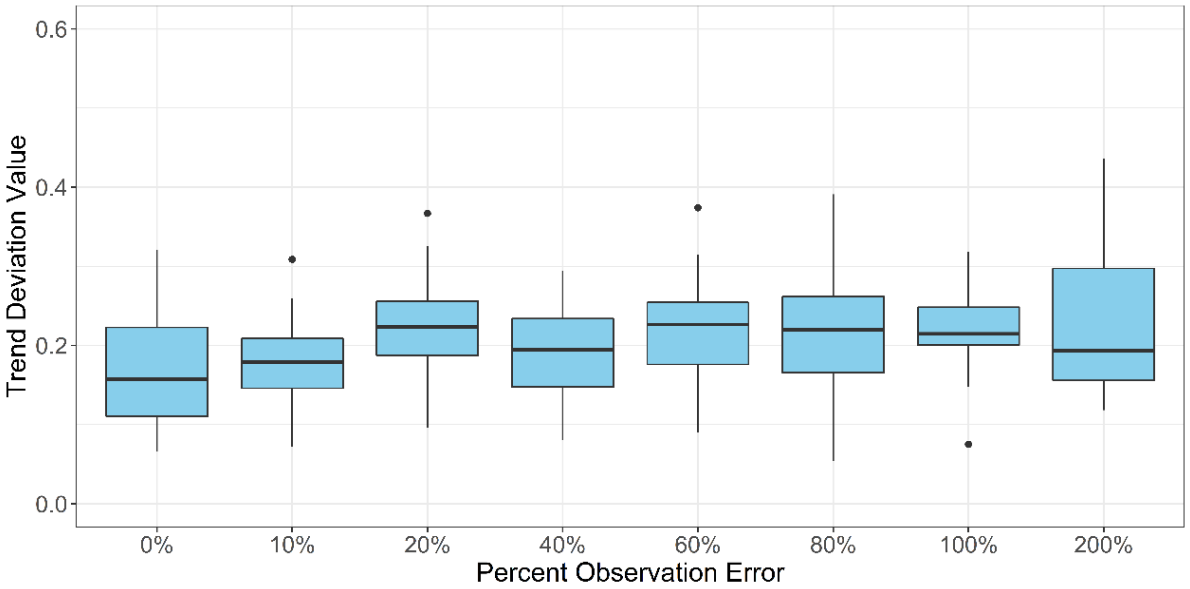


Figure S3.4. Trend deviation value vs percent observation error. All other parameters are fixed – dataset size: 1000; sample size: 200; μ_{ds}: 0; σ_{ds}: 0.2; μ_{η}: 0.2; populations per species: 20; mean time series length: 20; trend length: 50. Each box represents the mean values of 20 datasets, with 20 samples per dataset. The percentage of observation error has no effect on trend accuracy.

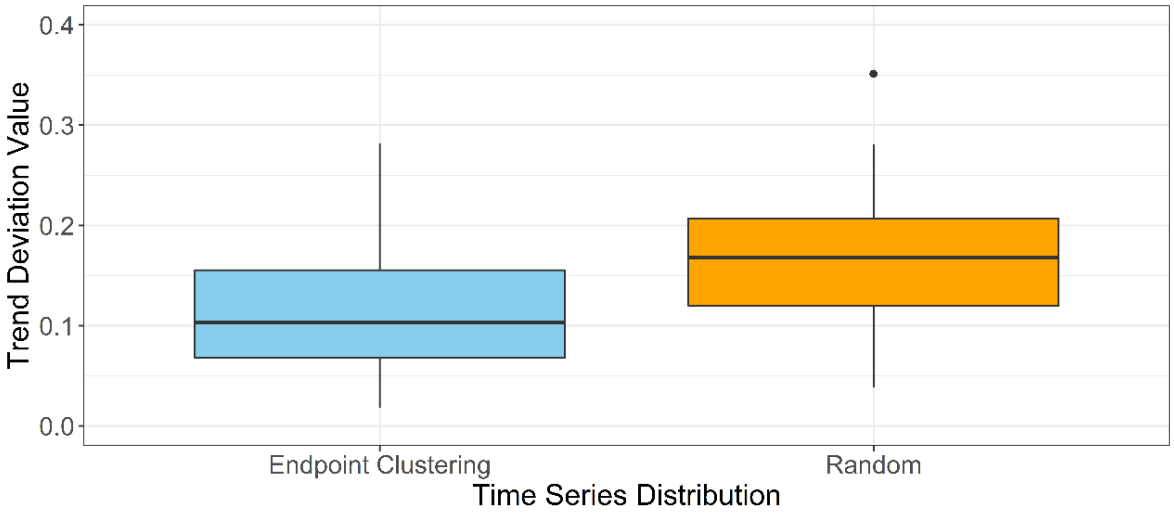


Figure S3.5. Trend deviation value vs time series distribution. Datasets assigned to 'Endpoint Clustering' had time series randomly assigned to begin at the first year of the trend or at end at the final year of the trend, while datasets assigned to 'Random' had time series randomly distributed across the simulated years of the trend. All other parameters are fixed – dataset size: 1000; sample size: 200; μ_{ds}: 0; σ_{ds}: 0.2; μ_{η}: 0.2; populations per species: 20; mean time series length: 25; trend length: 50; μ_{ε}: 0.15; σ_{ε}: 0.1. Each box represents the mean values of 40 datasets, with 20 samples per dataset.

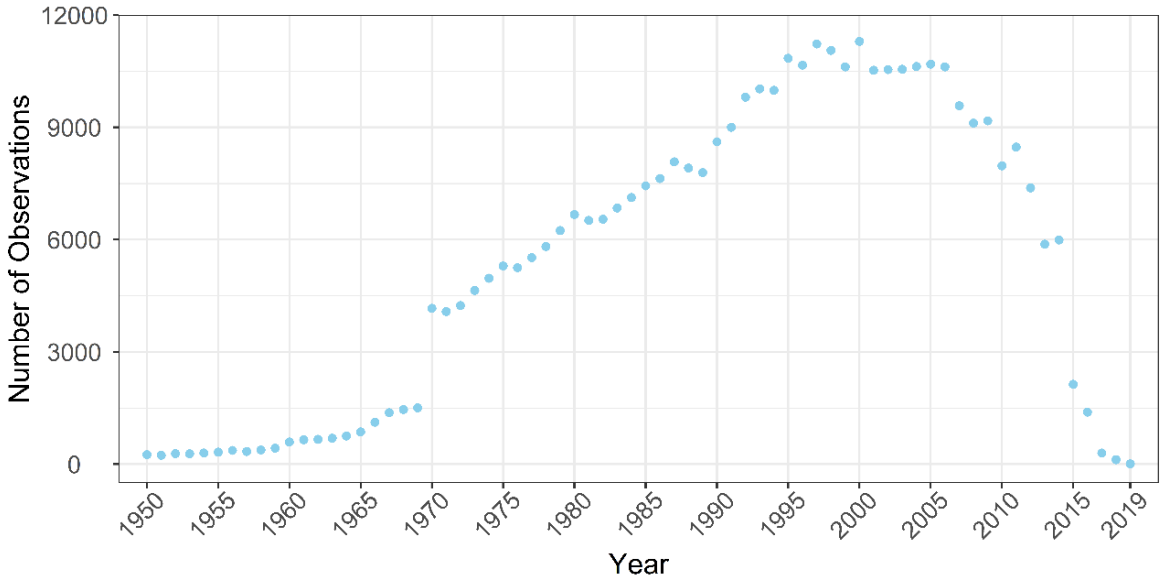


Figure S3.6. Total number of observations recorded in the Living Planet Database for each year from the start of the database until 2019 (the most recent observations in the version of the database I used for my analysis). There are 250 observations for the year 1950 and 9 observations for the year 2019. The year 2000 has the highest number of observations, at 11,297.

Appendix 3: Supplementary materials for Chapter 4

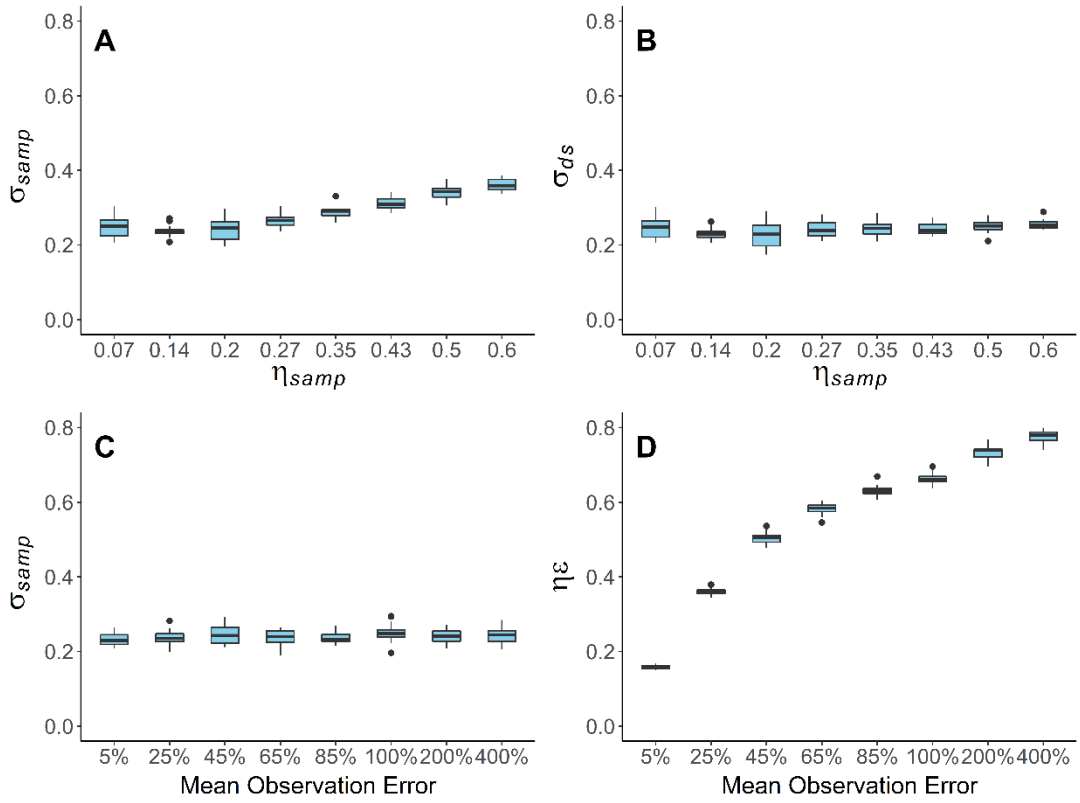


Figure S4.1. Variance in population growth rates vs process error and observation error. Panel A shows that between-population variance in the samples increases as process error increases; panel B shows that between-population variance in the unsampled datasets is unaffected by process error; panel C shows that within-population variance in the samples is unaffected by observation error; panel D shows that within-population variance in the sample increases as observation error increases. For all panels, parameters were set as follows – size of dataset: 1000; sample size: 200; length of trend: 50 years; populations per species: 20; mean growth rate: 0; variance in mean growth rate: 0.2; mean time series length: 20. For panels A and B, observation error was set to 0. For panels C and D, the mean of the growth rate standard deviations was set to 0.2.

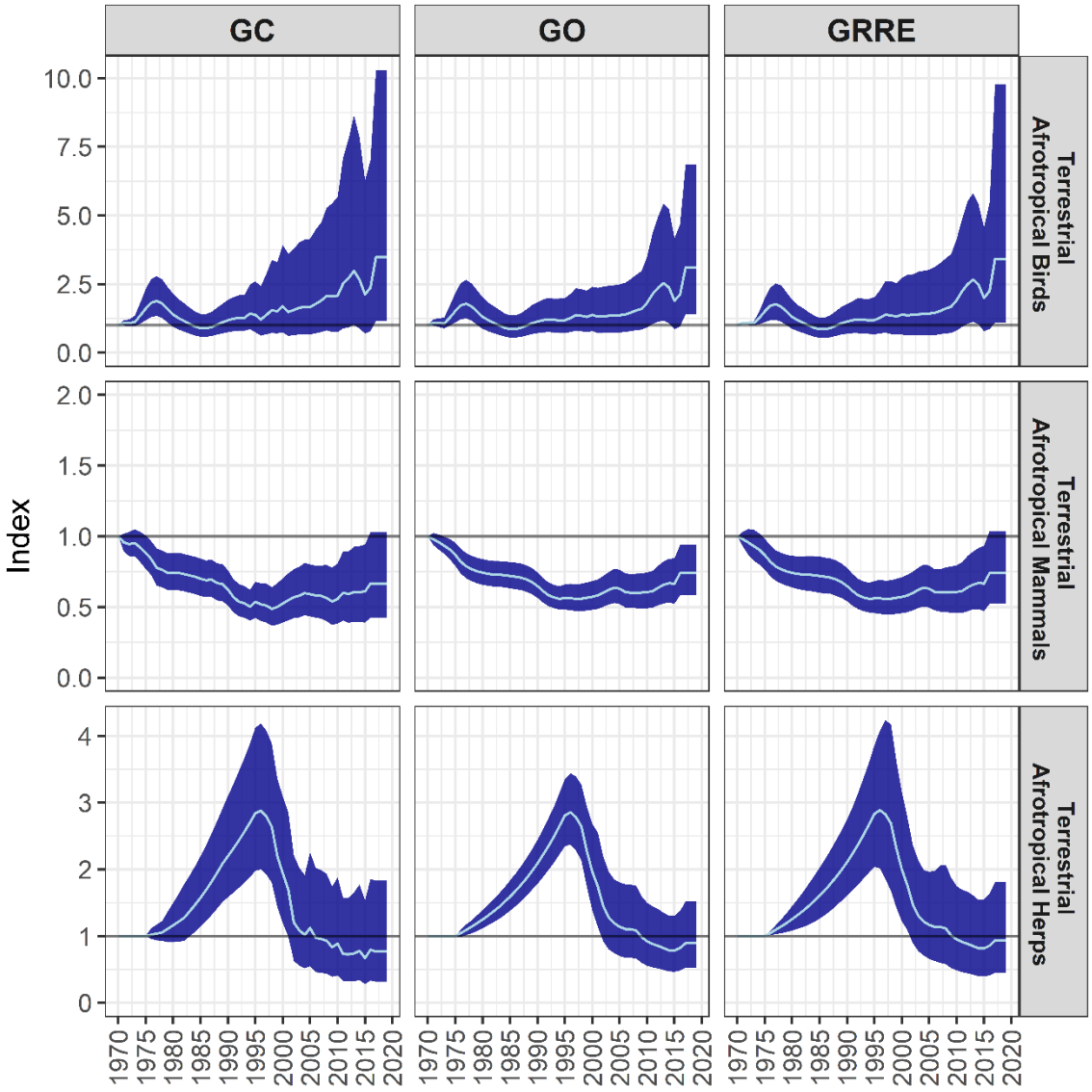


Figure S4.2. LPI trends for terrestrial Afrotropical species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

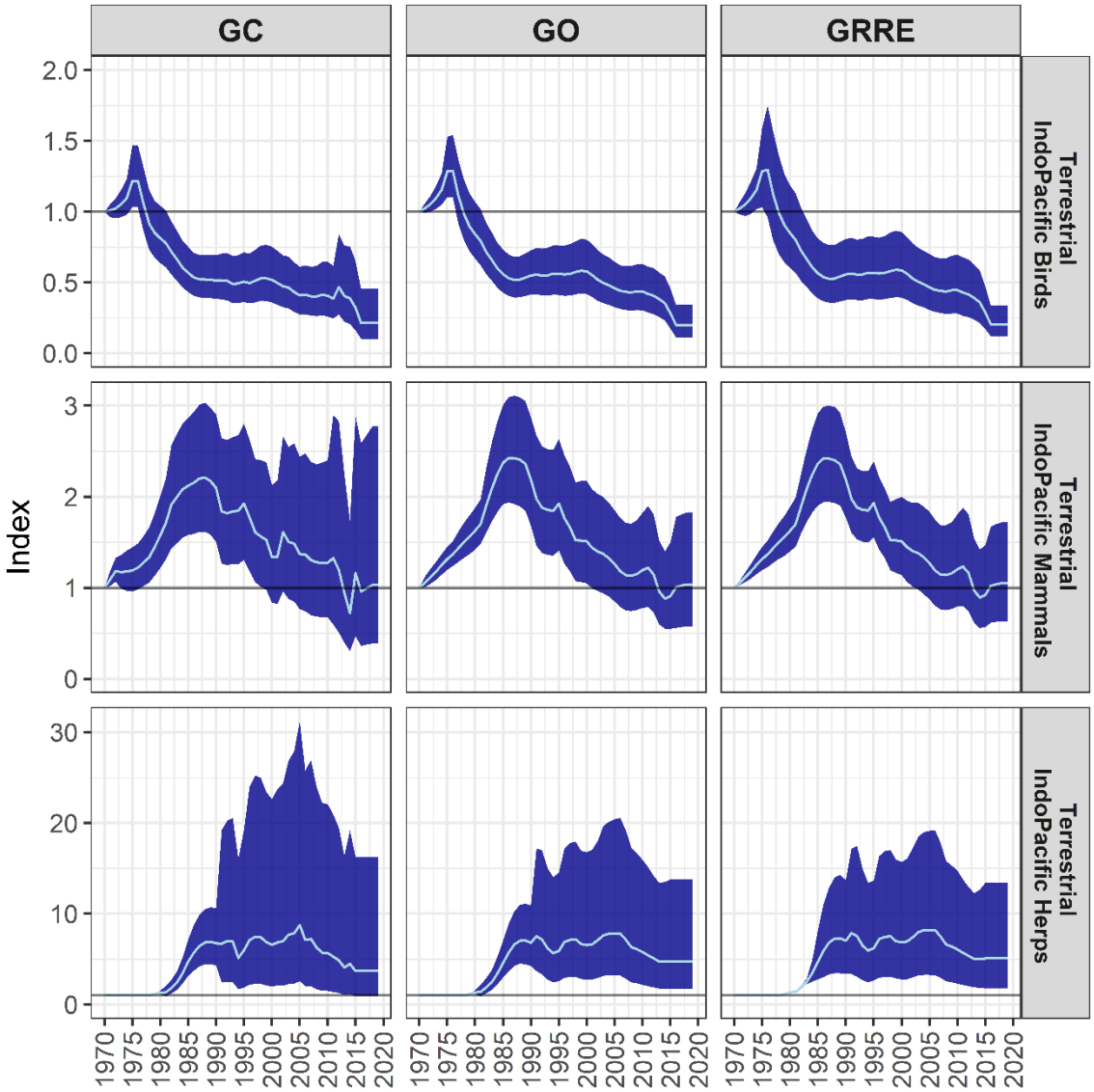


Figure S4.3. LPI trends for terrestrial IndoPacific species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

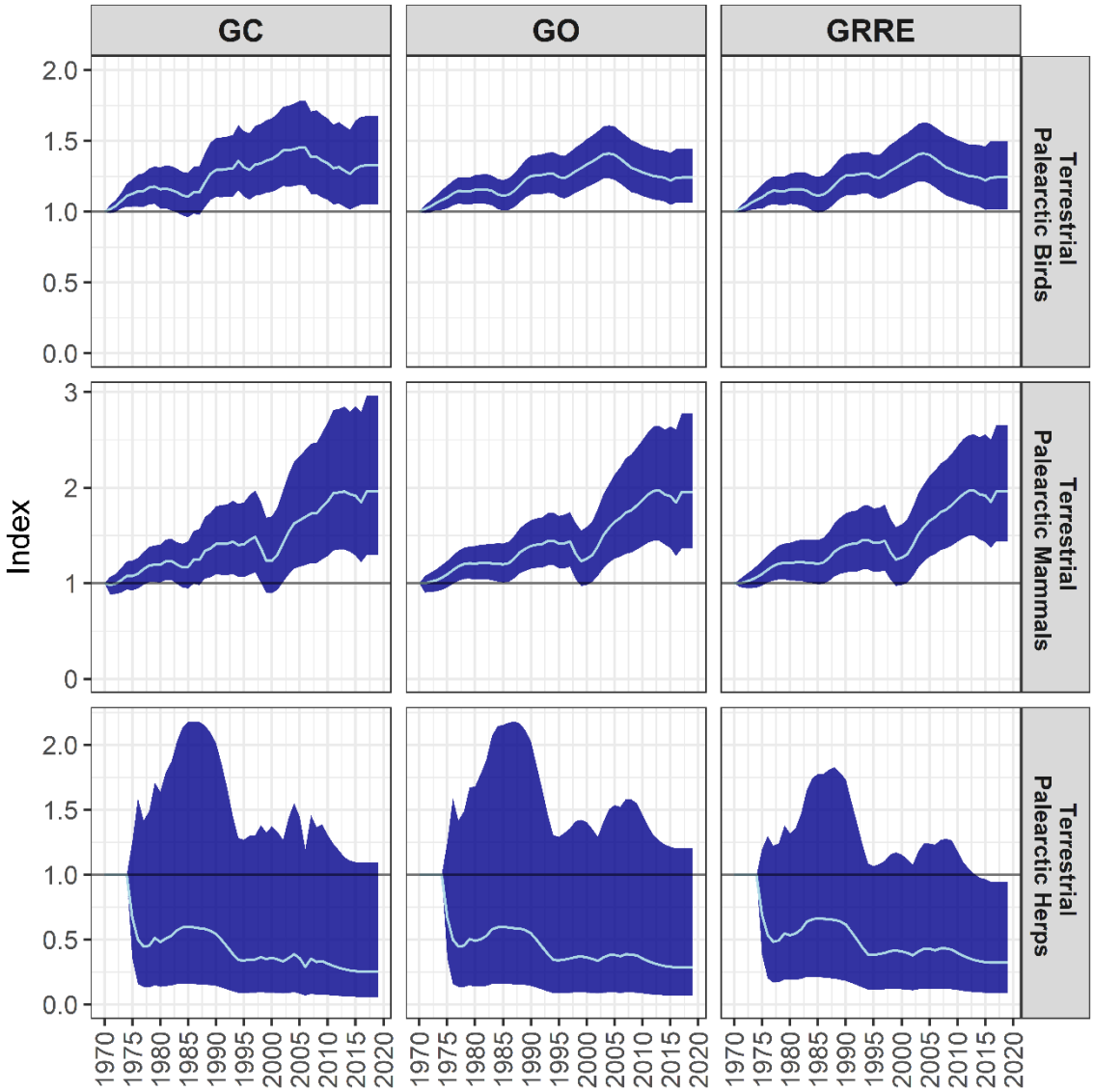


Figure S4.4. LPI trends for terrestrial Palearctic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

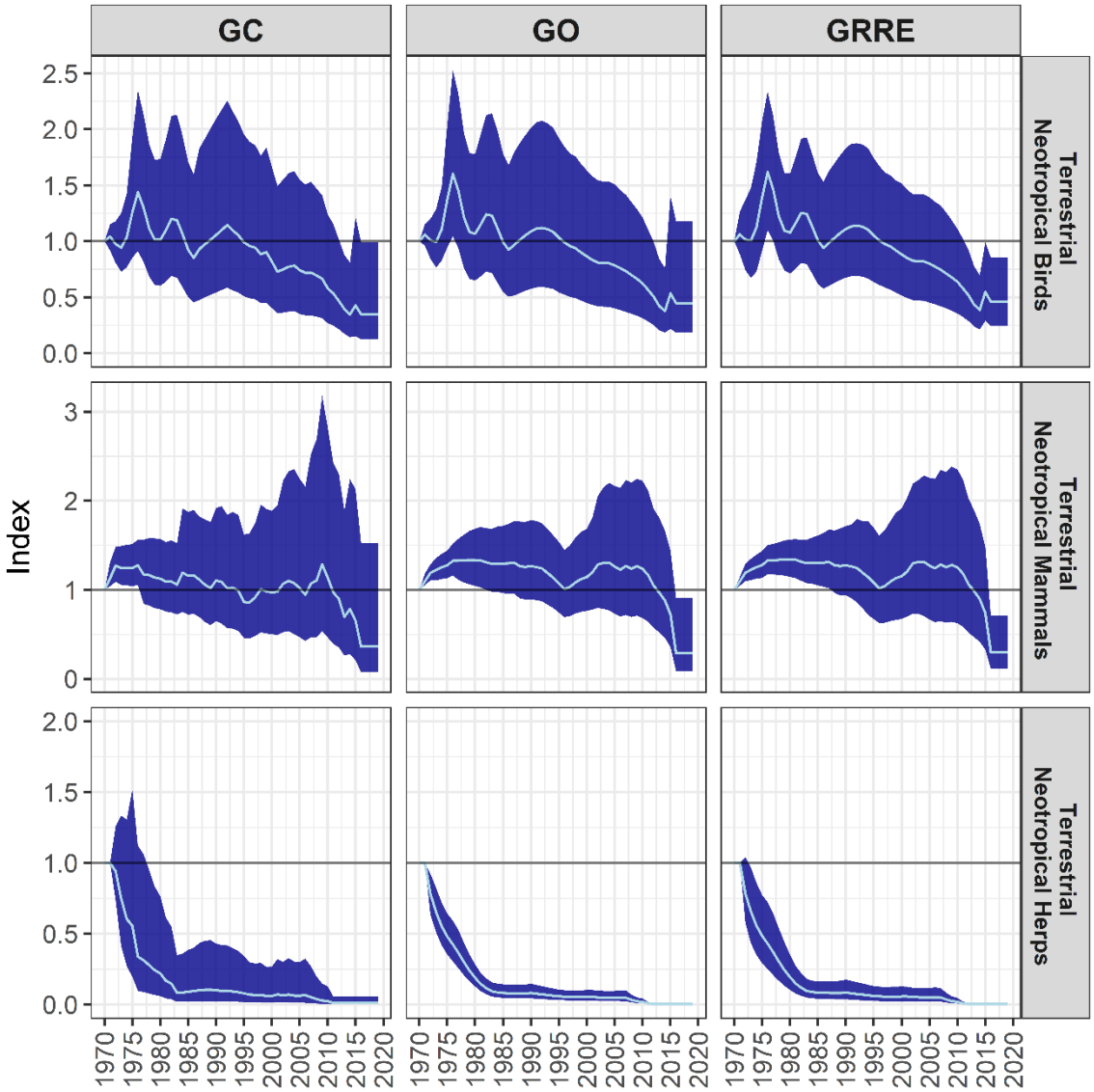


Figure S4.5. LPI trends for terrestrial Neotropical species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

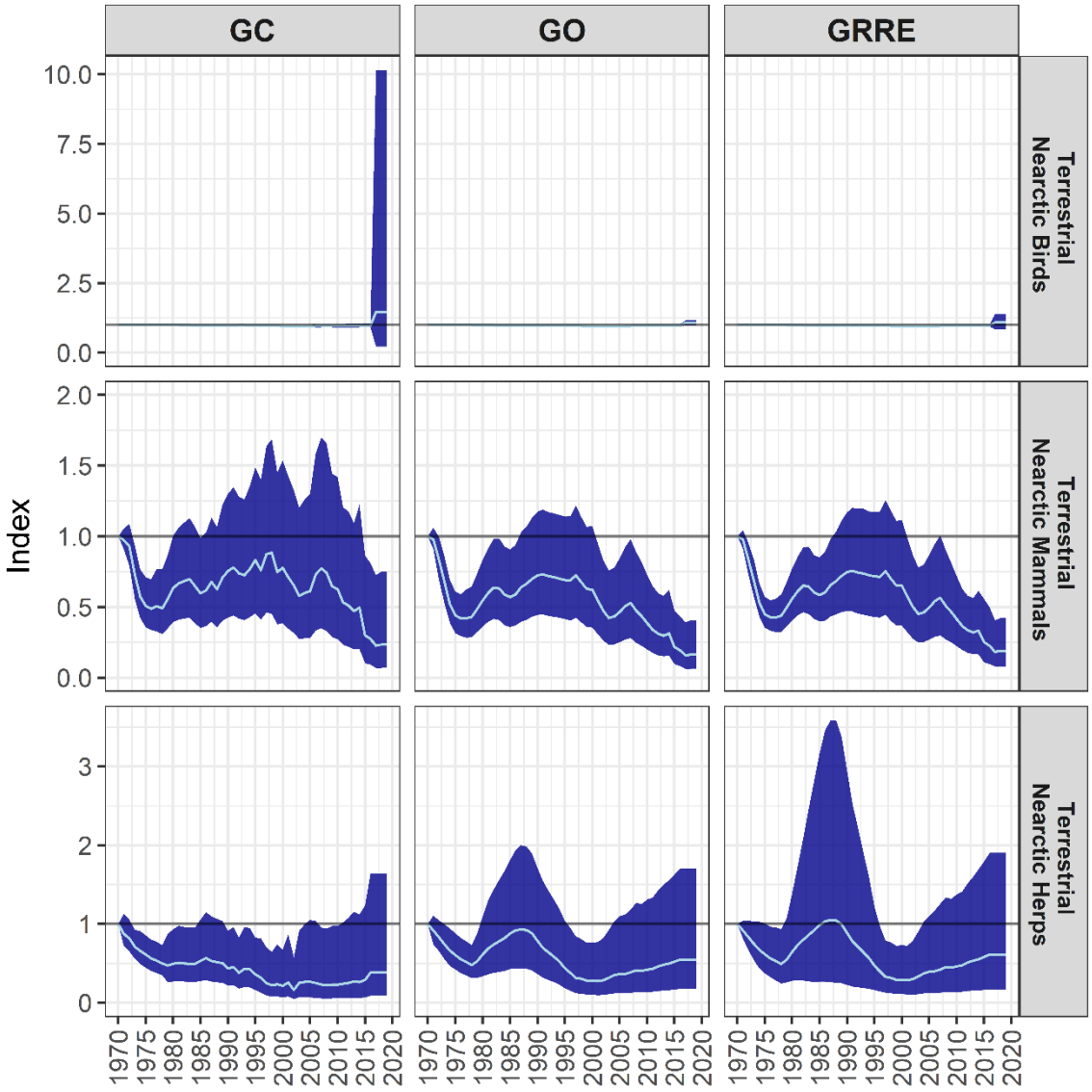


Figure S4.6. LPI trends for terrestrial Nearctic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

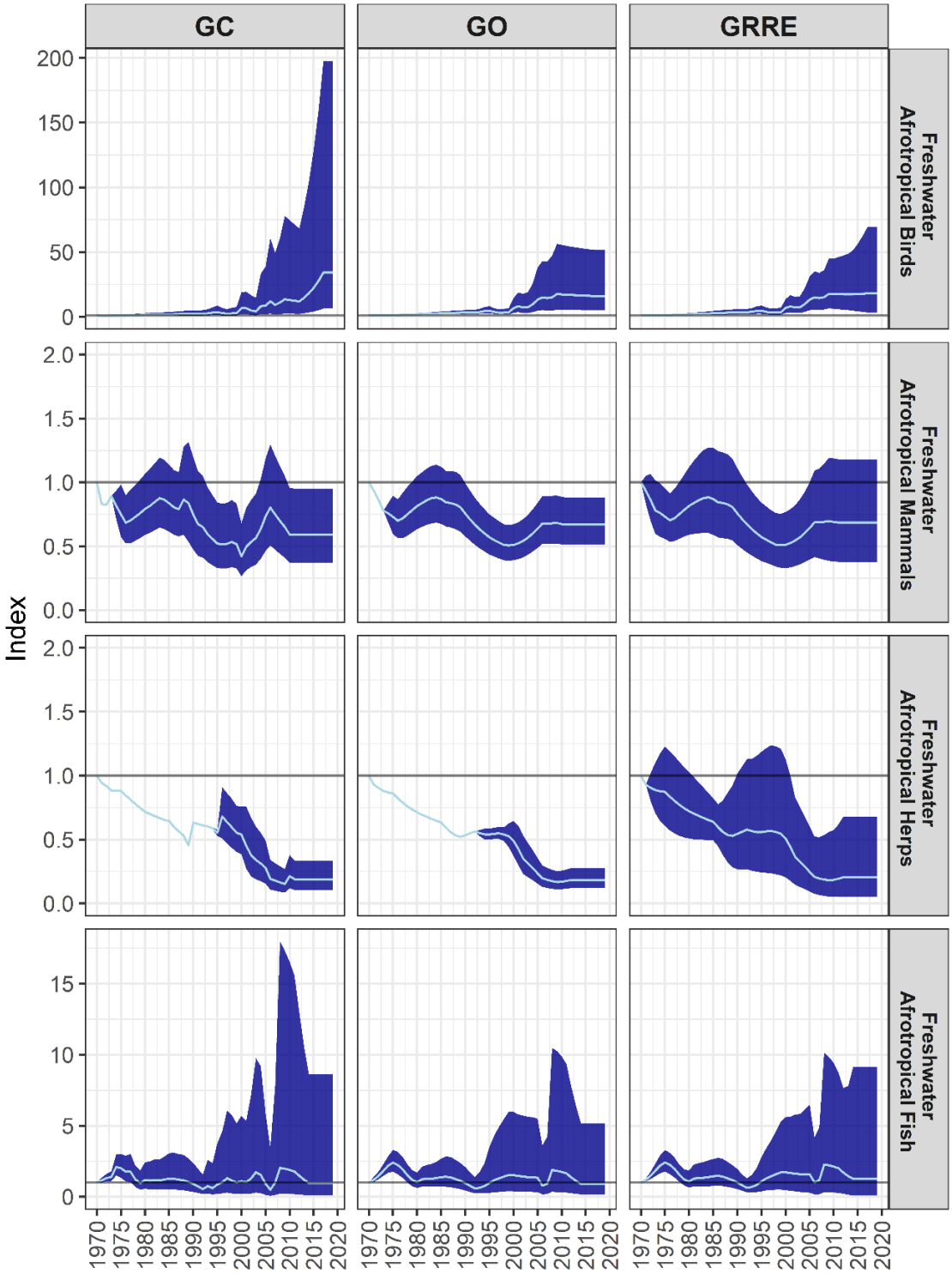


Figure S4.7. LPI trends for freshwater Afrotropical species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

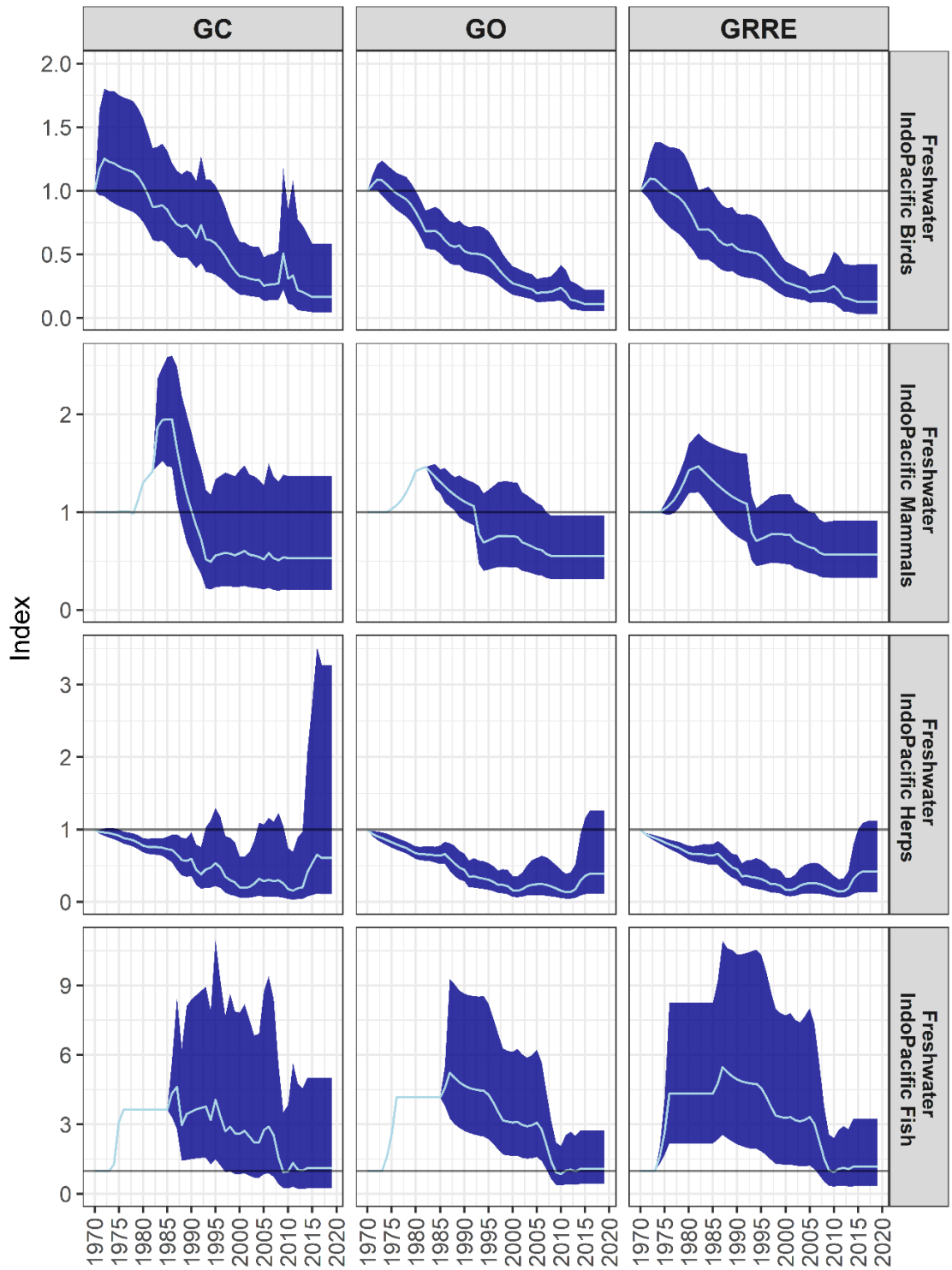


Figure S4.8. LPI trends for IndoPacific species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

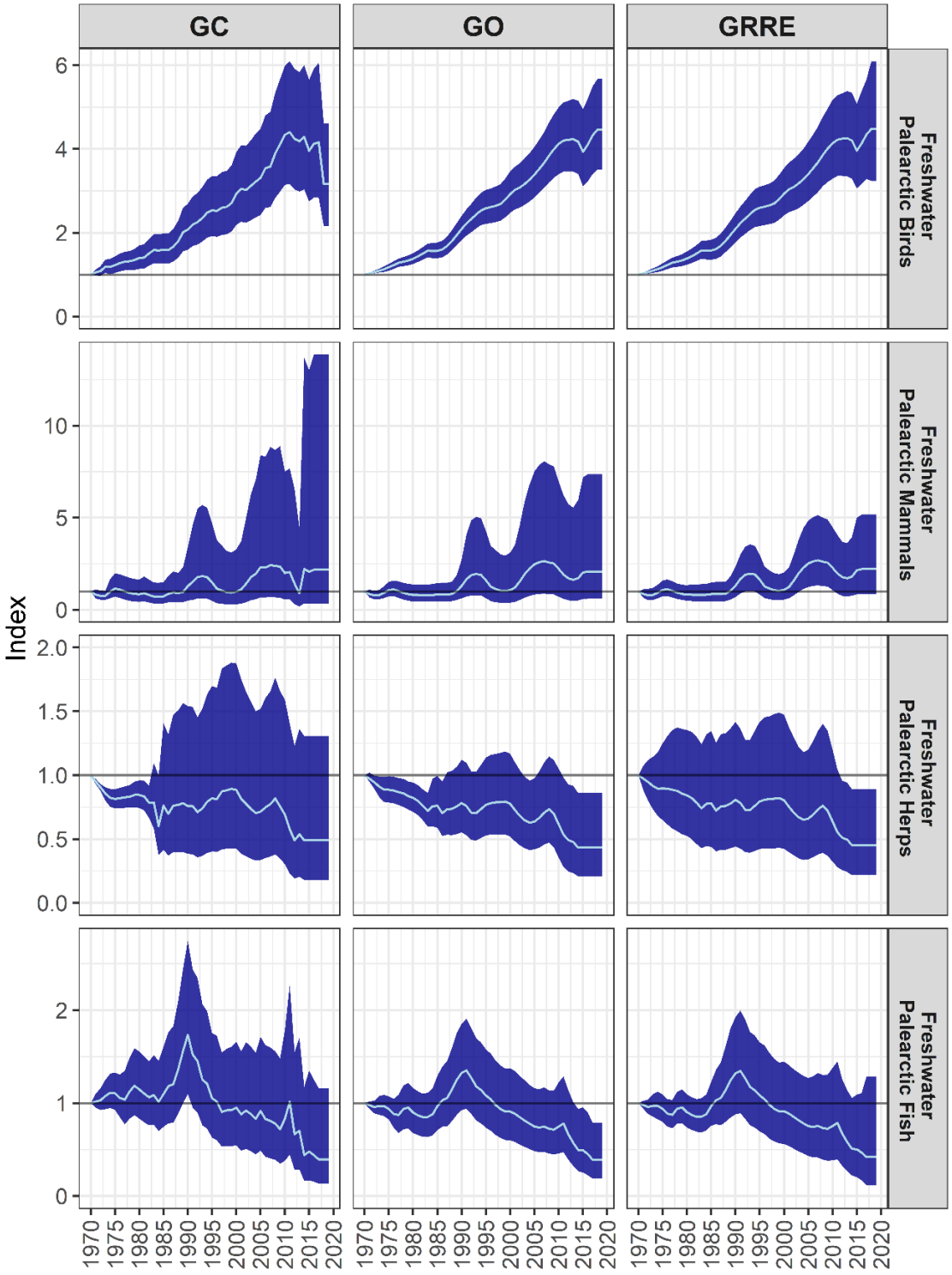


Figure S4.9. LPI trends for freshwater Palearctic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

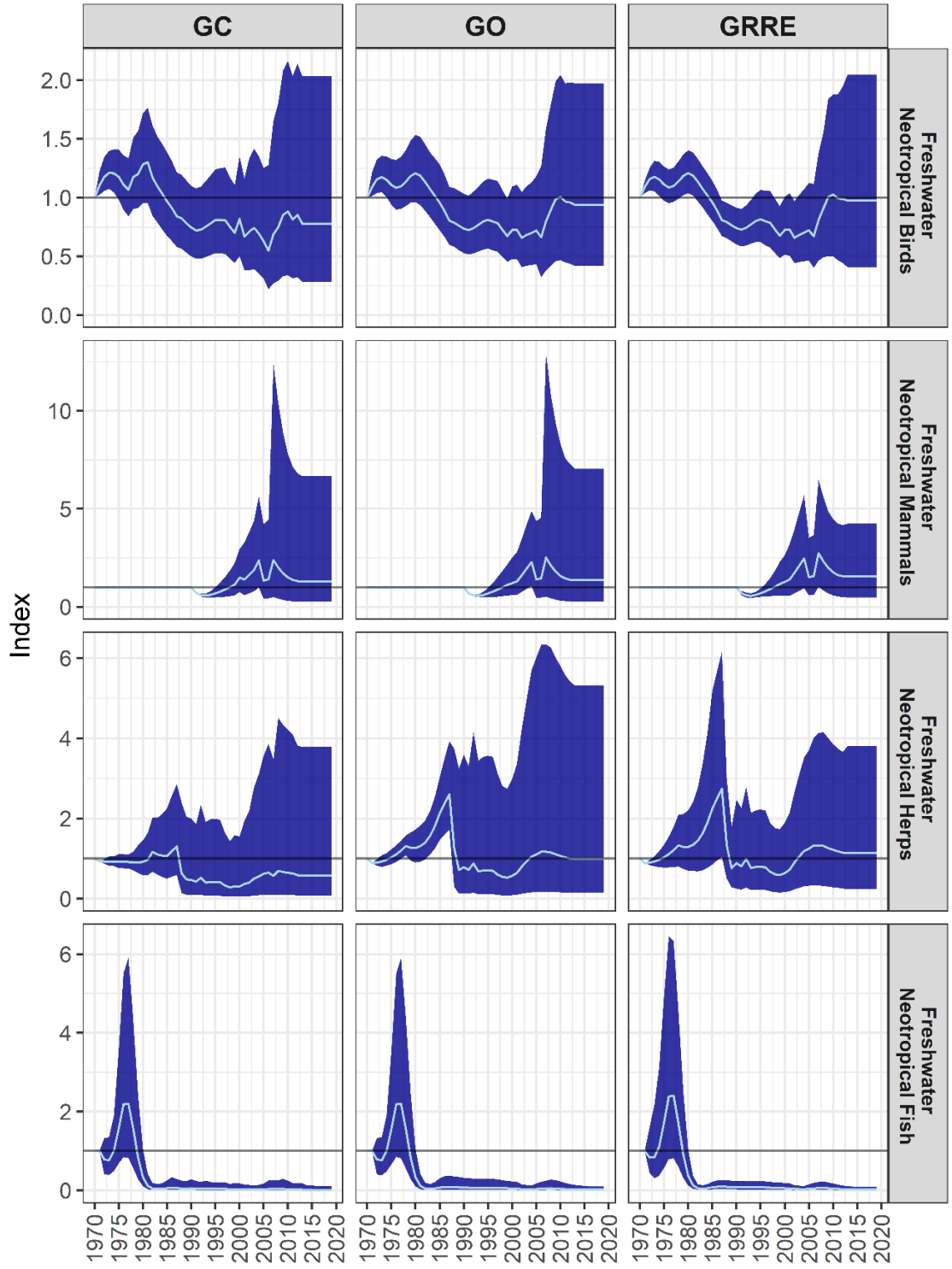


Figure S4.10. LPI trends for freshwater Neotropical species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

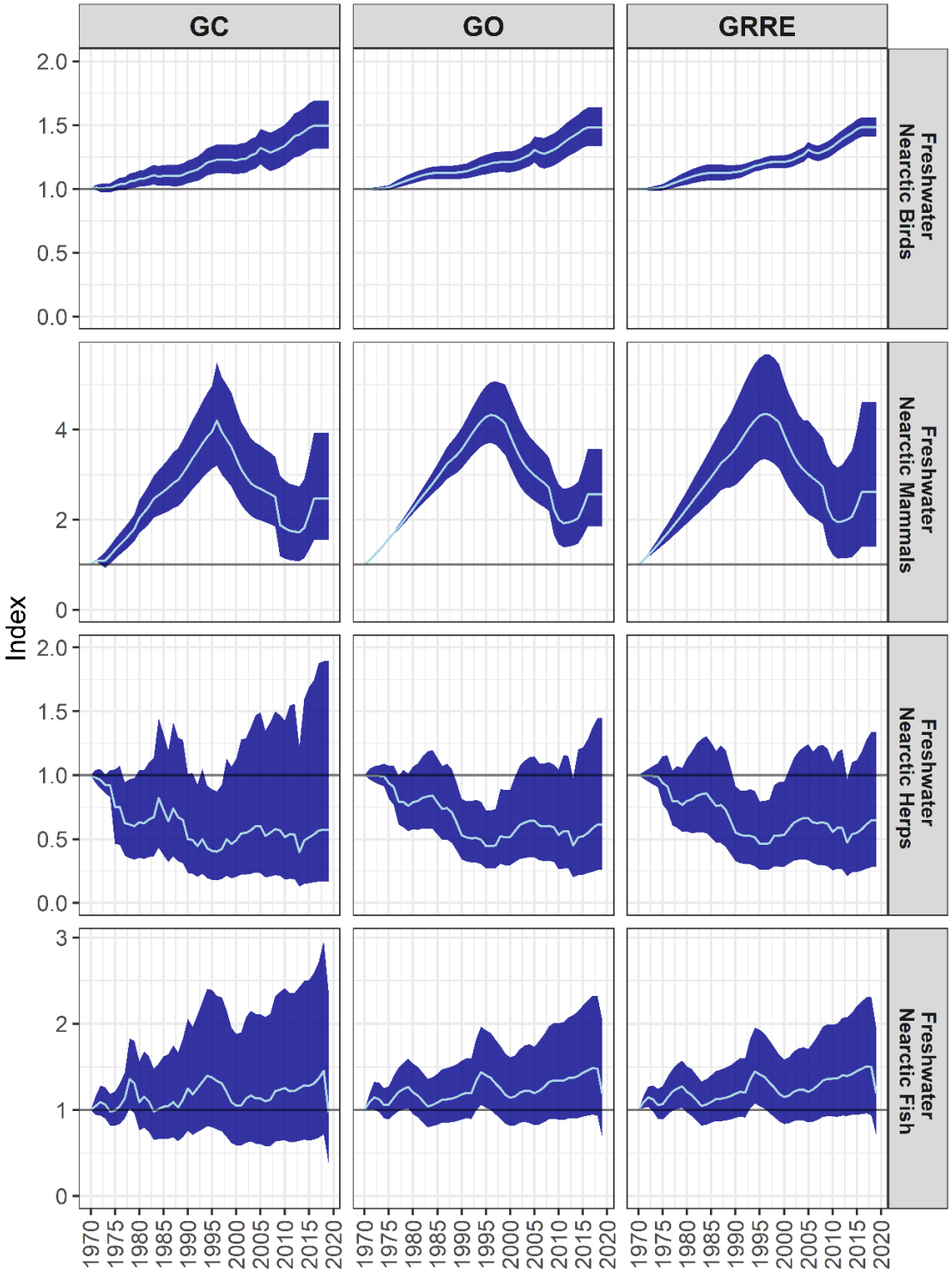


Figure S4.11. LPI trends for freshwater Nearctic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

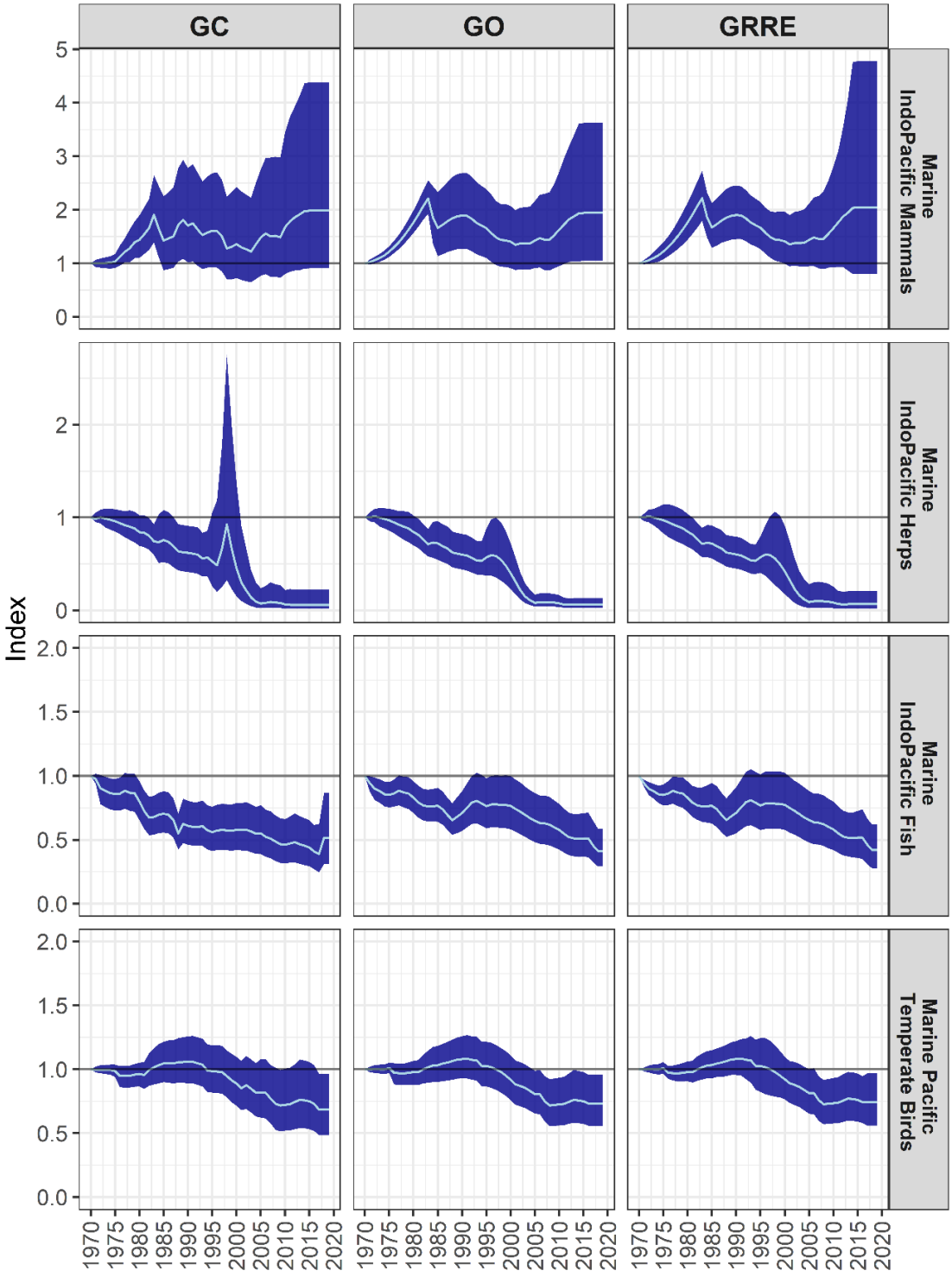


Figure S4.12. LPI trends for marine Pacific Temperate species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

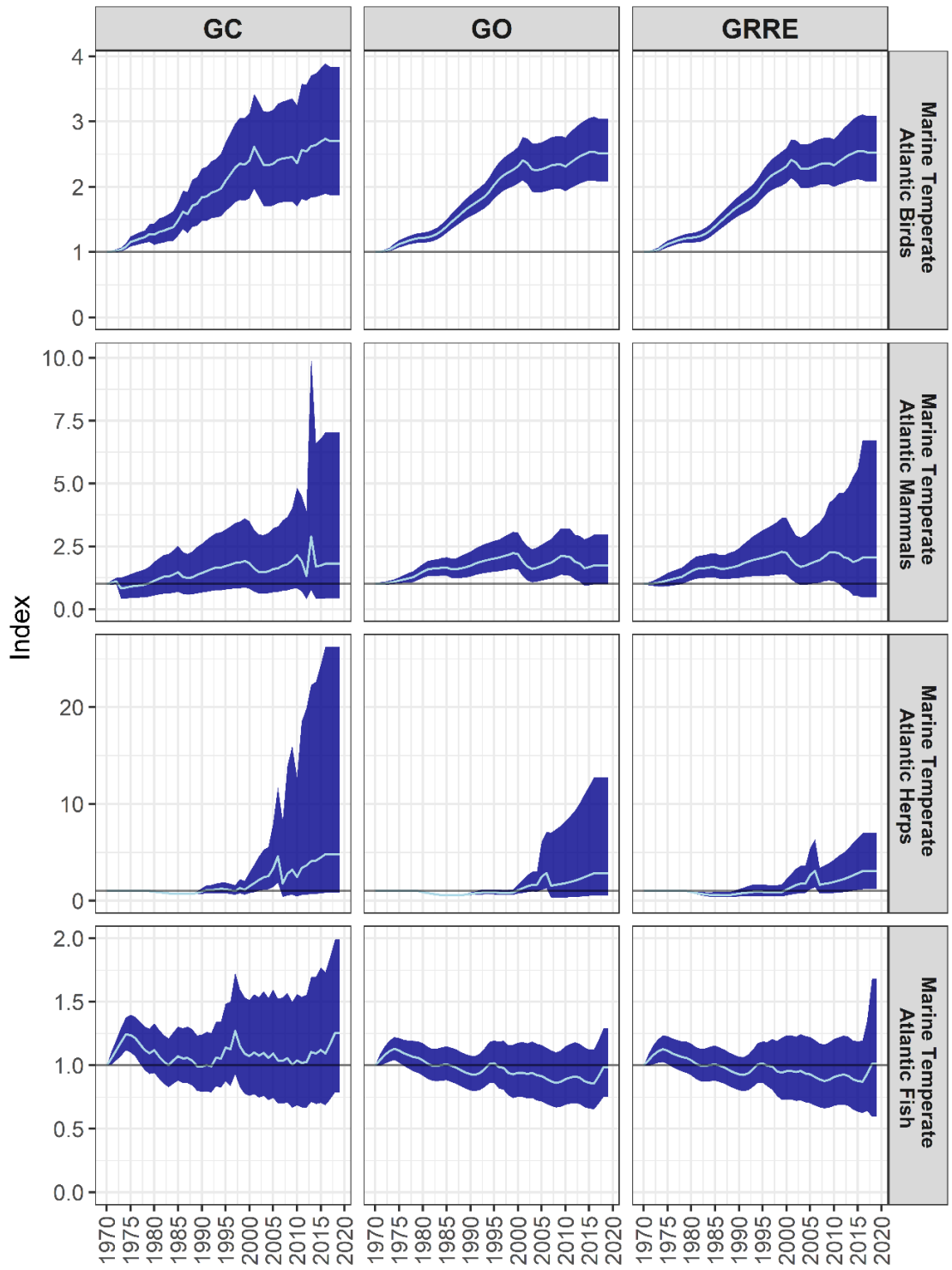


Figure S4.13. LPI trends for marine Temperate Atlantic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

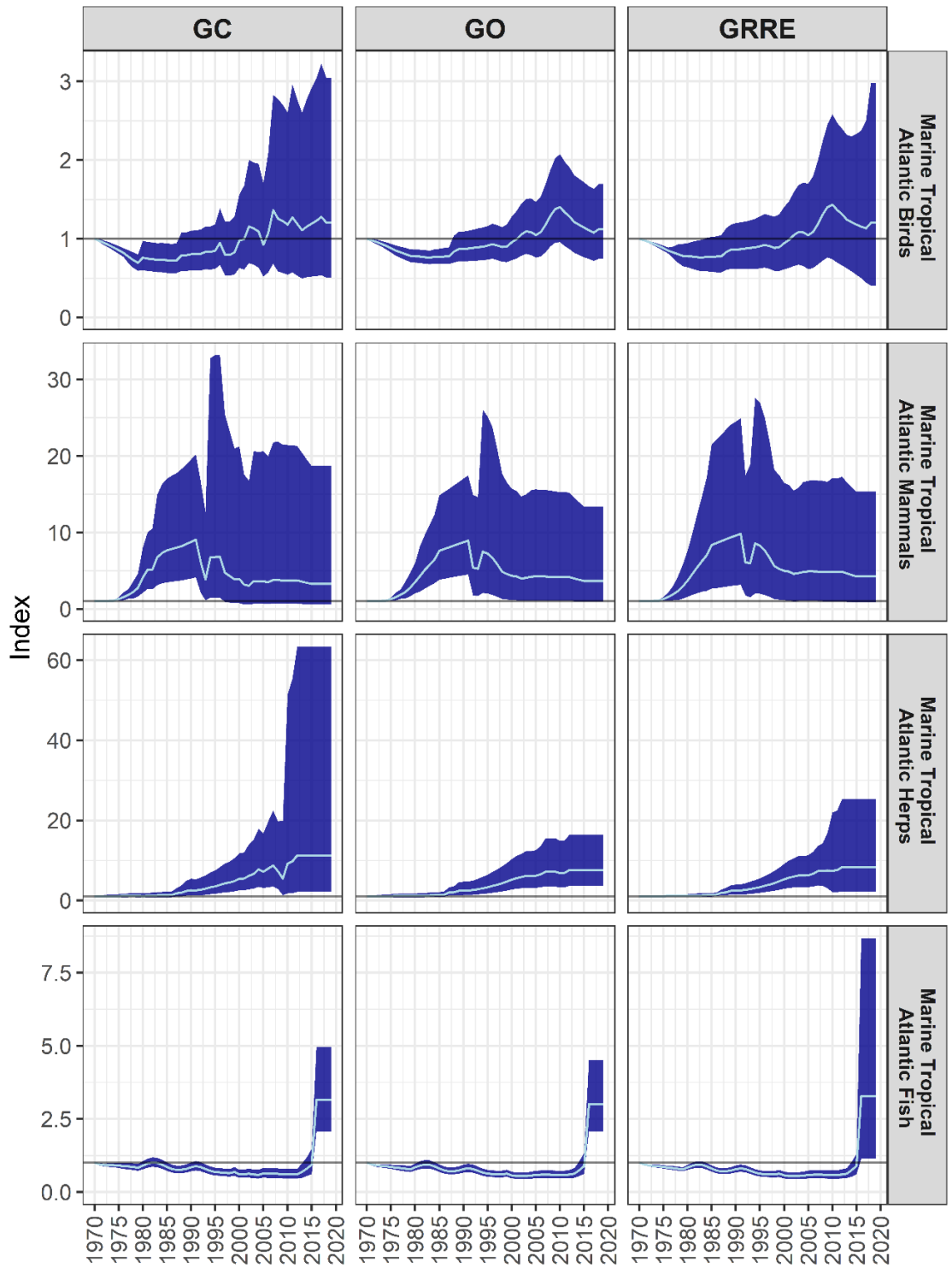


Figure S4.14. LPI trends for marine Tropical Atlantic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

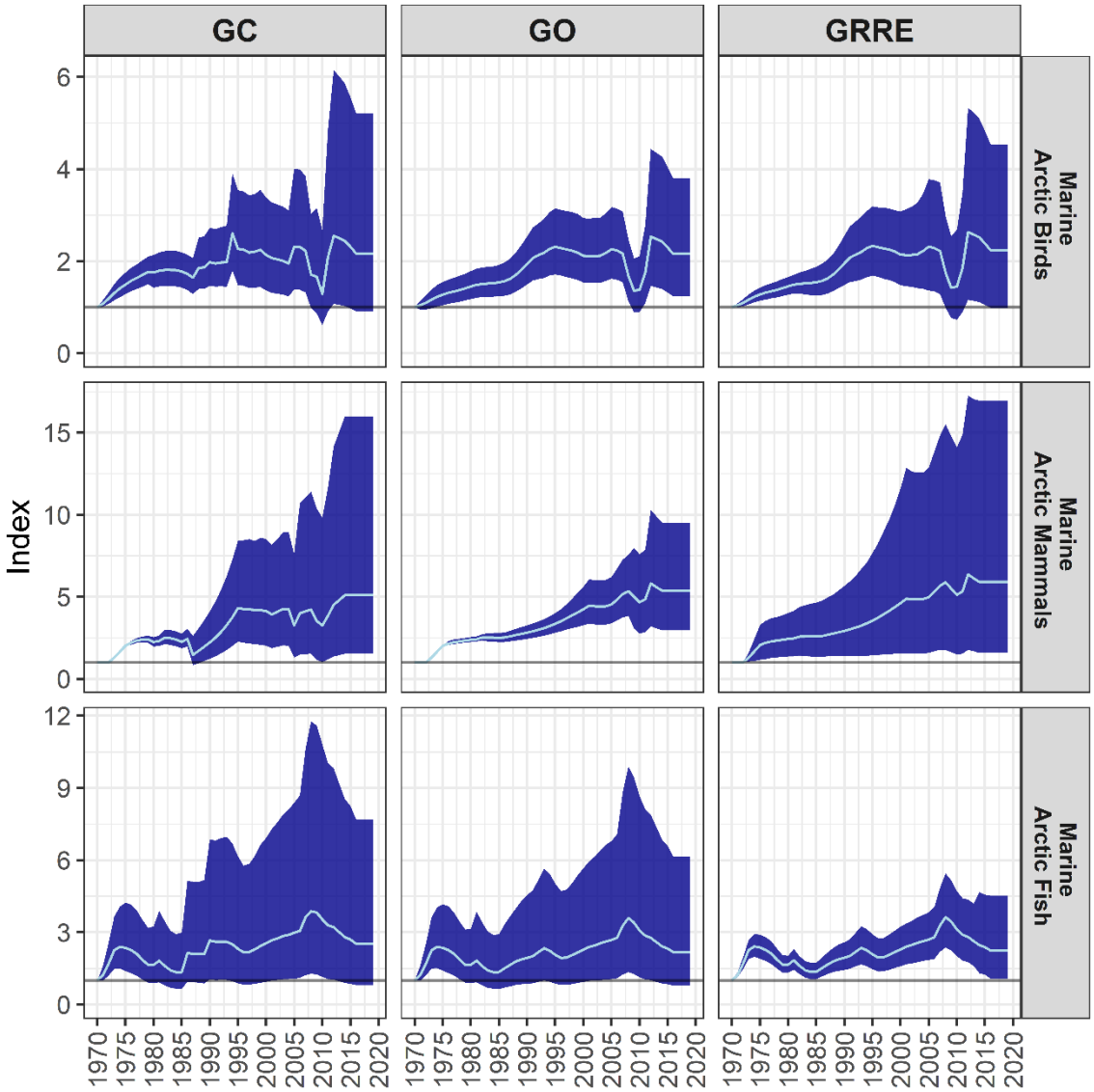


Figure S4.15. LPI trends for marine Arctic species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

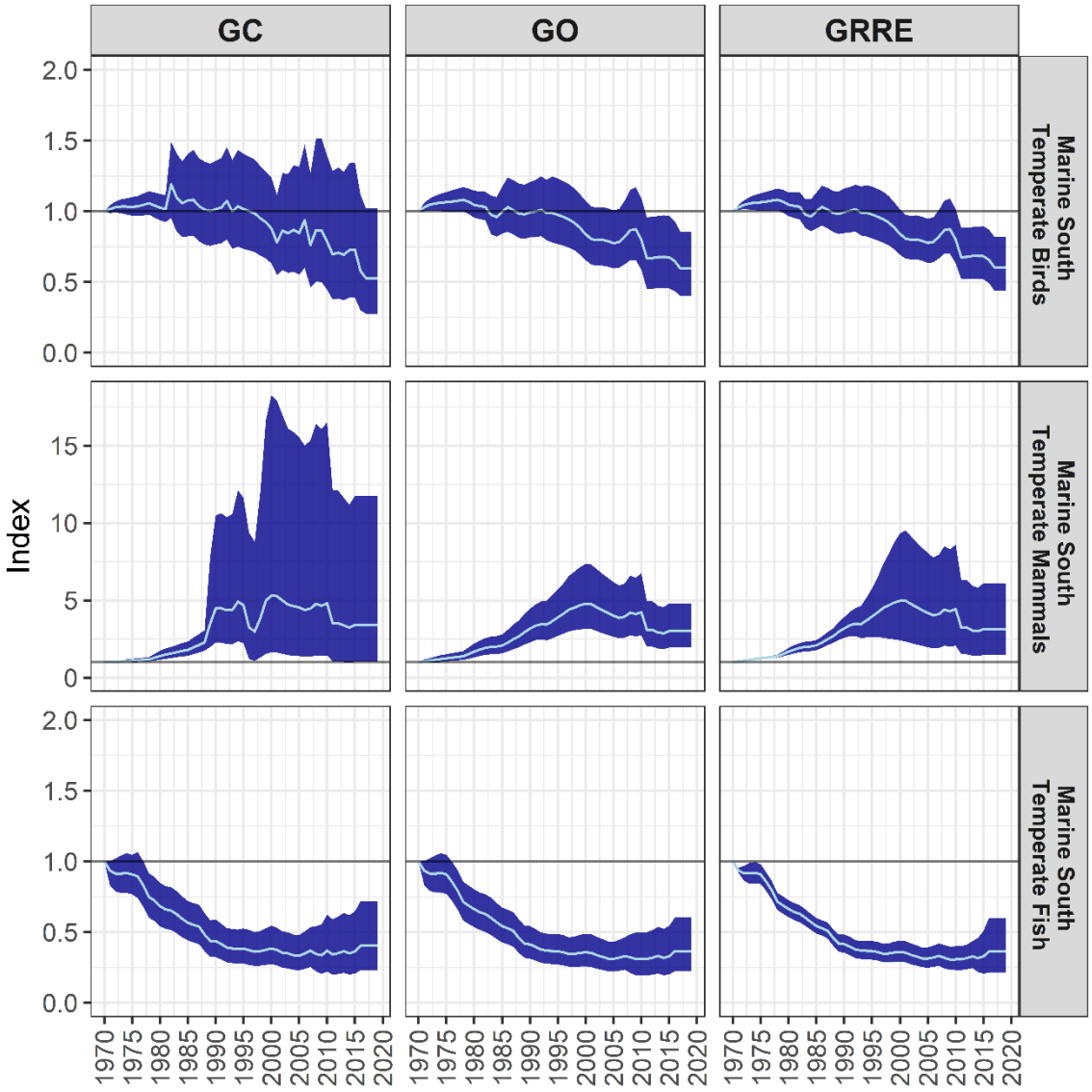


Figure S4.16. LPI trends for marine South Temperate species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

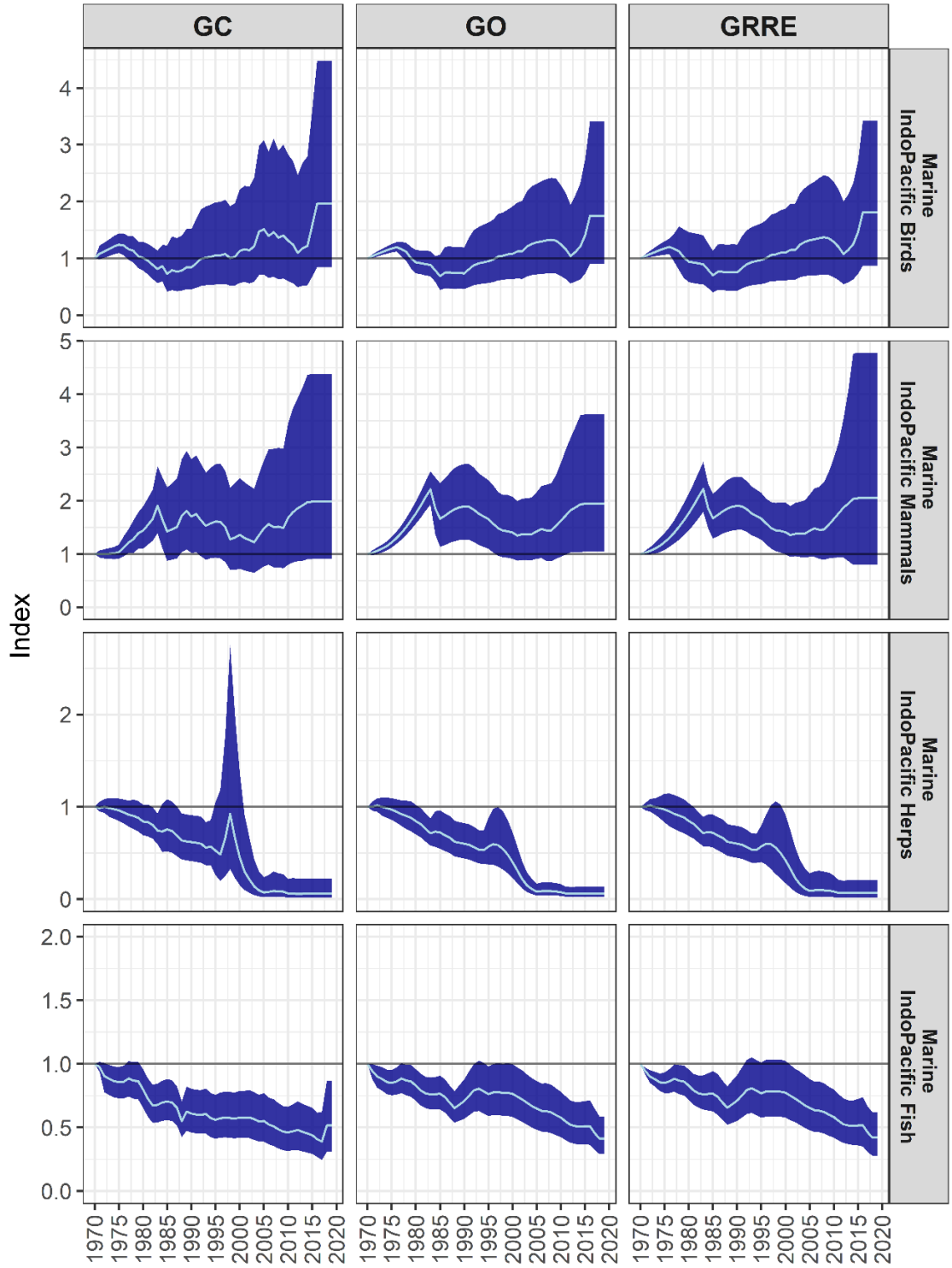


Figure S4.17. LPI trends for marine IndoPacific species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.

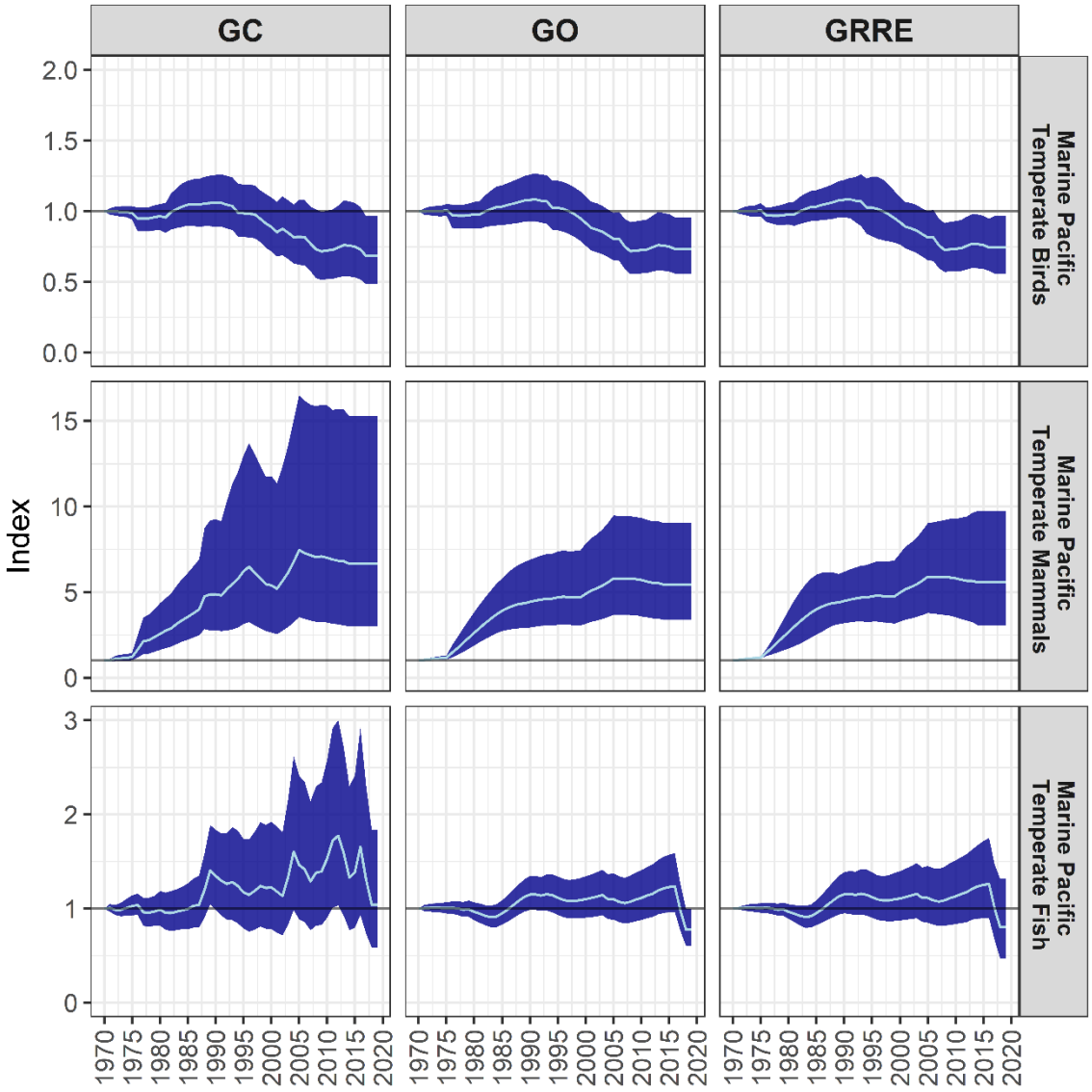


Figure S4.18. LPI trends for marine Pacific Temperate species groups, with confidence intervals. From left to right, the columns were calculated using the GAM + Chain (GC) method, which calculates confidence intervals by bootstrapping the species growth rates, and is the method used for the LPI; the GAM Only (GO) method, which models every population with a GAM but produces confidence intervals in the same way as the GC method; and the GAM-Resampled Rank Envelope (GRRE) method, which accounts for sampling error through repeated resampling of all populations from GAMs, and uses the rank envelope method to produce confidence intervals from multi-species trend variants.