Scientific methods for integrating expert knowledge in Bayesian models

Perepolkin, Dmytro

2023

Document Version:
Publisher's PDF, also known as Version of record

Scientific methods for integrating expert knowledge in Bayesian models

Scientific methods for integrating expert knowledge in Bayesian models

by Dmytro Perepolkin

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Dr. Ullrika Sahlin, Prof. Johan Elmberg, Prof. Erik Lindström
Faculty opponent: Prof. John Quigley

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in the Blue hall (Blå hallen) at the Centre for Environmental and Climate Science on Sunday, the 23rd of January 2023 at 1:00 pm.

Organization	Document name
LUND UNIVERSITY Centre for Environmental and Climate Science Box 188 SE–221 00 LUND, Sweden	**DOCTORAL DISSERTATION** Date of disputation 2024-1-23 Sponsoring organization
Author(s) Dmytro Perepolkin	

Title and subtitle

Scientific methods for integrating expert knowledge in Bayesian models

Abstract

Generating scientific advice to environmental management involves assessments with complex models, sparse data, and challenging empirical experiments, necessitating the integration of expert judgment with data into scientific models. To integrate expert judgement, assessors might elicit judgement by experts as quantiles, find a probability distribution that matches the quantiles, and add this information to the model. Data is then integrated into the model by Bayesian inference to learn parameters or make predictions. This thesis aims to simplify such integration of expert judgment, and introduce the use of Quantile-Parameterized Distributions (QPDs) into Bayesian models. Key questions addressed include identifying suitable QPDs for encoding expert judgment, and conditions for using QPDs as priors or likelihoods in Bayesian inference. The creation of new QPDs through quantile function transformation is explored, providing a methodological advancement. The use of the proposed methodology is demonstrated on expert-informed bias-adjustment of citizen science data in a Species Distribution Model for conservation assessment.

Key words

Bayesian inference, expert judgement, quantile parameterized distribution, quantile function

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language
	English
ISSN and key title	ISBN
	978-91-8039-915-9 (print) 978-91-8039-914-2 (pdf)
Recipient's notes	Number of pages: 131 / Price
	Security classification

Scientific methods for integrating expert knowledge in Bayesian models

by Dmytro Perepolkin

LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Cover illustration front: Generated with AI in response to a prompt: "hybrid elicitation and quantile-parameterized likelihood" (Credits: Microsoft Bing Image Creator, December 2, 2023 at 5:55 PM).

Funding information: The thesis work was financially supported by the Strategic Research Area "Biodiversity and Ecosystem Services in a Changing Climate", BECC, funded by the Swedish government.

Faculty of Science, Centre for Environmental and Climate Science

ISBN: 978-91-8039-915-9 (print)
ISBN: 978-91-8039-914-2 (pdf)

Printed in Sweden by Tryckeriet, E-Huset, Lund University, Lund 2023

A problem well put is half solved.
John Dewey

Contents

List of publications ii
Acknowledgements iii
Populärvetenskaplig sammanfattning iv
Abbrevations vi

Scientific methods for integrating expert knowledge in Bayesian models **1**

1 Introduction 1
2 Research aim and questions 4
3 Methodological approach to the thesis 5
4 Environmental relevance 6
5 Results and application 7
6 Discussion 10
7 Conclusion and future outlook 18
8 Bibliography 19

Scientific publications **27**

Author contributions 27
Paper I: Quantile-parameterized distributions for expert knowledge elicitation. 29
Paper II: The tenets of quantile-based inference in Bayesian models. . 59
Paper III: Hybrid elicitation and quantile-parametrized likelihood. . . 77
Paper IV: Observer-adjusted species distribution models using presence-only data. 95
Doctoral theses published in Environmental Science, Lund University 114

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Quantile-parameterized distributions for expert knowledge elicitation.**
D. Perepolkin, E. Lindström, U. Sahlin
In review

II **The tenets of quantile-based inference in Bayesian models.**
D. Perepolkin, B. Goodrich, U. Sahlin
Computational Statistics & Data Analysis, 187, 107795. https://doi.org/10.1016/j.csda.2023.107795

III **Hybrid elicitation and quantile-parametrized likelihood.**
D. Perepolkin, B. Goodrich, U. Sahlin
Statistics and Computing, 34(1), 11. https://doi.org/10.1007/s11222-023-10325-0

IV **Observer-adjusted species distribution models using presence-only data.**
D. Perepolkin, J. Elmberg, F. Lindgren, U. Sahlin
Manuscript

Acknowledgements

First of all, I want to thank God for giving me the opportunity to embark on this remarkable journey. I am convinced that none of this would have been possible if stars did not line up every time I needed direction. I feel incredibly blessed to have had the luxury of spending four years indulging in the activities I like the most - reading, researching, writing, and engaging in thoughtful debates.

A special thanks to my advisor, Ullrika Sahlin, whose kindness and support made this PhD project not only intellectually stimulating but also immensely enjoyable. Your infectious energy and a continuous stream of brilliant ideas added immeasurable joy to this academic endeavor. This period of academic exploration has been a true privilege, and I am profoundly thankful for your guidance that has shaped this meaningful journey.

In remembrance and gratitude, I extend my thanks to the late Prof. Warren G. Gilchrist (1932-2015) for imparting a new perspective on statistics. The profound impact of his teachings has shaped my understanding in ways beyond measure.

I am indebted to Prof. N. Unnikrishnan Nair, whose pioneering work has encouraged me to pursue the fascinating realm of my studies. I feel both grateful and humbled to stand on the shoulders of giants like yourself.

To my beloved wife, Iryna – my heart, my soulmate – your unwavering dedication to my success has been the wind beneath my wings.

Populärvetenskaplig sammanfattning

Vetenskapligt baserade lösningar på miljöproblem kan kräva metoder att använda komplexa modeller, hantera begränsade data eller ersätta av experiment som är oetiska eller svåra att utföra. Ett exempel på en sådan metod är att integrera expertbedömningar som komplement till empiriska observationer när man bygger vetenskapliga modeller. Expertkunskap och data kan kombineras genom att beskriva experters osäkerhet och slumpmässighet i data med sannolikhetsmodeller, och genom att tillämpa sannolikhetsregler (Bayesiansk inferens) för att dra slutsatser.

Expertkunskaper spelar en stor roll i processer för att ta fram vetenskapliga råd och att fatta beslut. De kan användas för att bestämma beslutsfattares värderingar och preferenser, eller för att formulera vetenskapliga modeller. Vetenskapliga experter har kunskaper som genererar vetenskapliga hypoteser eller en mekanistisk förståelse, vilket ger struktur till vetenskapliga modeller. Experter kan också bidra med kunskap om fakta eller möjliga värden på parametrar i vetenskapliga modeller. Det kan vara om en art finns på en viss plats, hur snabbt en population växer under vissa förhållanden, eller storlek på systematiska fel i observationer som rapporteras in av fågelskådare. Slutligen, kan experter bidra med bedömningar om framtida händelser eller observationer under olika beslut.

Expertbedömningar är subjektiva, men det är inte ett hinder för att använda dem i vetenskapliga modeller. Det finns forskning som visar att människors bedömningar påverkas av vad det är man frågar efter, hur frågor ställs och vad andra säger om man är i en grupp. Vetenskapliga experter är inget undantag. För att minska eventuella skevheter och missförstånd i experters bedömningar, och öka transparens, och därmed vetenskaplig tillförlitlighet i resultat, bör expertbedömningar följa en strukturerad process.

Ett exempel på en sådan process är Sheffield-metoden, där man ber experter att komma överens om par mellan kvantiler och sannolikheter för att representera deras osäkerhet om en parameter inom en vetenskaplig modell. En kvantil är ett värde som delar in en sannolikhetsfördelning för en kontinuerlig variabel i två delar, där det är en viss sannolikhet att variabeln är lägre än detta värde. Nästa steg i Sheffield-metoden är att anpassa en parametrisk sannolikhetsfördelning till kvantil-sannolikhetsparen. Denna fördelning kan sedan användas för att beskriva osäkerhet.

De flesta har hört talas om normalfördelningen. Den är ett exempel på en parametrisk sannolikhetsfördelning som definieras utifrån en täthetsfunktion. Det är nog få som känner till Metalog, Myerson eller J-QPD, vilket är exem-

pel på sannolikhetsfördelningar som definieras utifrån en kvantilfunktion, och som dessutom har kvantiler som parametrar. Kvantil-parametriserade sannolikhetsfördelningar är flexibla men används idag endast i begränsad omfattning för att beskriva experters bedömningar. Detta kan bero på att dessa sannolikhetsfördelningar är ovanliga och att det inte är känt hur dessa kan användas i Bayesiansk modellering. Jag har i denna avhandling gjort en sammanställning av kvantildefinerade fördelningar, med syfte att visa hur kvantil-parametriserade sannolikhetsfördelningar skulle kunna underlätta integrering av expertbedömningar i vetenskapliga modeller.

Vetenskapliga råd som stöder beslut om biologisk mångfald och bevarande tar i bästa fall hänsyn till inverkan på artpopulationers dynamik i rum och tid, vilket kräver tillförlitliga uppskattningar av arters förekomst. Artutbredningsmodeller förlitar sig ofta på endast närvarodata, insamlade av medlemmar av allmänheten utan att följa något särskilt observationsprotokoll. Även om det finns gott om data från medborgarforskning, har de visat sig innehålla en betydande del av systematiska fel på grund av hur data samlas in.

Den största databasen med artdata från medborgarforskning är Global Biodiversity Information Facility (GBIF). Förutom informationen om individuella observationer, såsom artnamn och antal, registrerar GBIF en del metadata som beskriver observationshändelsen (exempelvis tid, plats, observatörs-ID). Jag har använt denna information för att generera data på icke-närvaro och för att justera för systematiska fel som beror på olika intentioner hos observatörer i en artutbredningsmodell för gäss i nordöstra Skåne.

Abbrevations

CDF	Culuative Distribution Function
CSW	Chalaby Scott and Wuertz
FKML	Freimer Kollia Mudholkar Lin
GBIF	Global Biodiversity Information Facility
GLD	Generalized Lambda Distribution
INLA	Integrated Nested Laplace Approximation
IQR	Inter-Quartile Range
PDF	Probability Density Function
QF	Quantile Function
QPD	Quantile-Parameterized Distribution
SMD	Species Distribution Models
MCMC	Markov Chain Monte Carlo
HMC	Hamiltonian Monte Carlo

Scientific methods for integrating expert knowledge in Bayesian models

1 Introduction

Environmental Management is an area where models are complex, data are sparse (expensive to collect) and experiments are difficult to set up (Roberts et al., 2018). Therefore, scientific assessors producing scientific advice have to rely on experts to complement the evidence collected by observation (Choy et al., 2009). Bayesian methods allow for integration of expert judgment with data for inference on relevant quantities of interest (Gelman et al., 2013).

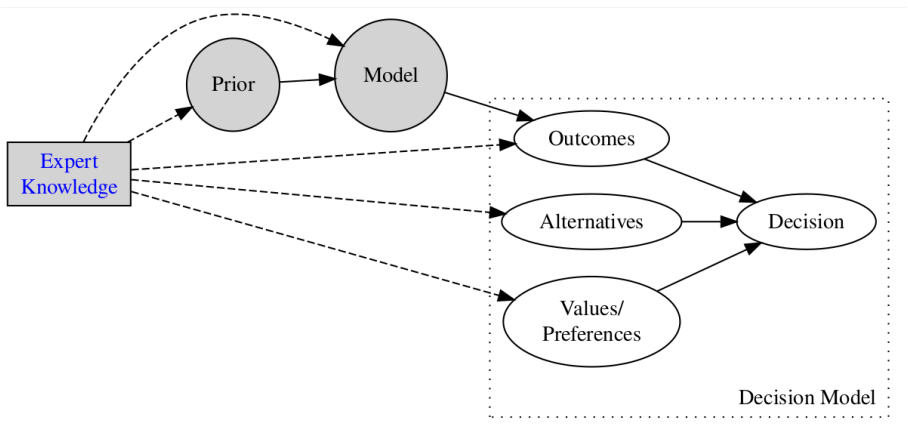


Figure 1: Expert knowledge elicitation in Environmental Management

Expert knowledge elicitation plays a crucial role in various facets of the en-

vironmental management decision-making process (Figure 1). It can be used for determining decision makers' values and preferences when setting population targets (Johnson et al., 2021), hypothesis generation and model structure elicitation for understanding population dynamics (Madsen et al., 2017), prior elicitation for the parameters of population models (Johnson et al., 2017), and predictive elicitation for anticipated decision outcomes (observables) (Tulloch et al., 2017).

1.1 Opportunistic environmental data

Scientific advice supporting biodiversity and conservation decisions considers at best impact on dynamics of populations in space and time, requiring reliable estimates of species prevalence. Species Distribution Models (SDMs) often rely on presence-only data, collected by members of the public without following any particular observation protocol. Although citizen science data are abundant, they have been shown to contain a significant amount of bias due to irregularities in the observation process (Hastie and Fithian, 2013; Grimmett et al., 2020; Stolar and Nielsen, 2015; Isaac and Pocock, 2015). Opportunistically collected wildlife observations may be prone to one or more of the following challenges:

- Absences are typically not recorded. The data is often referred to as *presence-only*.
- The observations are collected at non-random locations. This is referred to as the *spatial preferential sampling*
- Only selected species are recorded, implying *taxonomic preferential sampling*
- The time it took to collect observations varies from one event to another; also, the spatial extent explored while making observations varies between events. These two effects are jointly referred to as *varying observation effort*.

The largest database accumulating citizen science data is the Global Biodiversity Information Facility (GBIF) (GBIF, 2023). In addition to the key information about individual observations, such as the species taxon and the quantity, GBIF records some metadata describing the observation event (time, location, observer ID, etc). One way to compensate the deficiency of presence-only datasets is to consider this metadata and place individual observation of focal species in the context of the rest of the sightings made by an observer during a trip. The

list of all observations made by an observer on a single trip is called an "event species list" or just "species list" (Szabo et al., 2010). The absence can then be interpreted as the absence of the focal species on the species list, provided that at least one other species was registered. This observer-oriented approach, which focuses on the analysis of the list length (Ruete et al., 2017, 2020) have been shown to outperform randomly sampled pseudo-absences in de-biasing the citizen science data used in species distribution modelling (Di Cecco et al., 2021; Milanesi et al., 2020).

The bias related to preferential sampling can be adjusted by employing a thinning process with a parameter responsible for the influence of bias on the likelihood. Expert judgment is required to define the thinning function and provide an informed prior for the parameter responsible for shrinkage (Sicacha-Parada et al., 2021).

Not all observers are contributing equally to citizen science databases. The majority of observations are recorded by a few "super-observers", employed by institutions or otherwise committed to repeatedly recording the wildlife observations (Cretois, 2021; Di Cecco et al., 2021). Therefore, taxonomic preferential sampling can be addressed by including the species list length explicitly in the model.

1.2 Structured expert elicitation

Expert judgment constitutes a valuable source of knowledge, particularly in scenarios where evidence is limited or difficult to obtain. However, behavioral research has revealed inherent human biases that can impact the accuracy (Kahneman, 2012) and consistency (Kahneman et al., 2021) of judgments. To address these challenges, a substantial body of knowledge has emerged regarding the selection of experts. It emphasizes the importance of experts being qualified, well-calibrated, and diverse in their background knowledge and expertise (Dias et al., 2018; European Food Safety Authority, 2014; Hanea et al., 2021a; Morgan, 2014).

To mitigate cognitive biases in expert judgment and to bolster transparency and scientific rigor, it is crucial to adhere to a structured elicitation process (Burgman, 2015; Hanea et al., 2021b; O'Hagan, 2019). One such process, the Sheffield protocol (Gosling, 2018), asks experts to agree on quantiles-probability pairs to represent their uncertainty about a fixed quantity (usually a parameter within a model). Subsequently, a parametric distribution, chosen from a set of familiar and practical distributions, is fitted to these elicited quantiles. This fitting

process can introduce ambiguity with regards to the choice of the distribution, especially when the selected distribution does not fit the elicited quantiles exactly.

1.3 Quantiles: from elicitation to distribution

Sarma and Kay (2020) identified three target objectives when choosing a prior for a Bayesian model: centrality matching (mean, median), interval matching (IQR, variance) and probability mass allocation. Probabilistic judgments are commonly expressed as triplets of quantile-probability pairs, comprising a median and upper and lower quantiles (e.g., 25th/75th or 5th/95th) (Johnson et al., 2017; O'Hagan et al., 2006). These pairs represent points from an unknown distribution function $F(x)$.

Figure 2 shows a cumulative distribution function (CDF) with the median and four quantiles marked by dashed lines. The slope of the CDF is called a probability density function (PDF). In the PDF chart, the probabilities correspond to the area under the curve. Finally, the inter-quantile range can be illustrated by a boxplot.

Several specialized distributions have been developed to facilitate the smooth interpolation of probabilistic judgments. Quantile-Parameterized Distributions (QPDs) is a flexible class of probability distributions parameterized by the quantile-probability pairs. These distributions might fulfill all of the objectives identified by Sarma and Kay (2020) by matching the median, an IQR and allocating the probability mass as required. QPDs enable the precise capture of expert knowledge while maintaining a high level of flexibility in modeling (Hadlock, 2017; Keelin and Powley, 2011; Powley, 2013).

2 Research aim and questions

The aim of this thesis two-fold: 1) to simplify the integration of expert judgement in scientific models and 2) to integrate quantile-parameterized distributions into Bayesian models.

This aim is addressed by attacking the following questions:

- What could be suitable distributions for encoding expert judgment on parameters versus observable (data) levels?

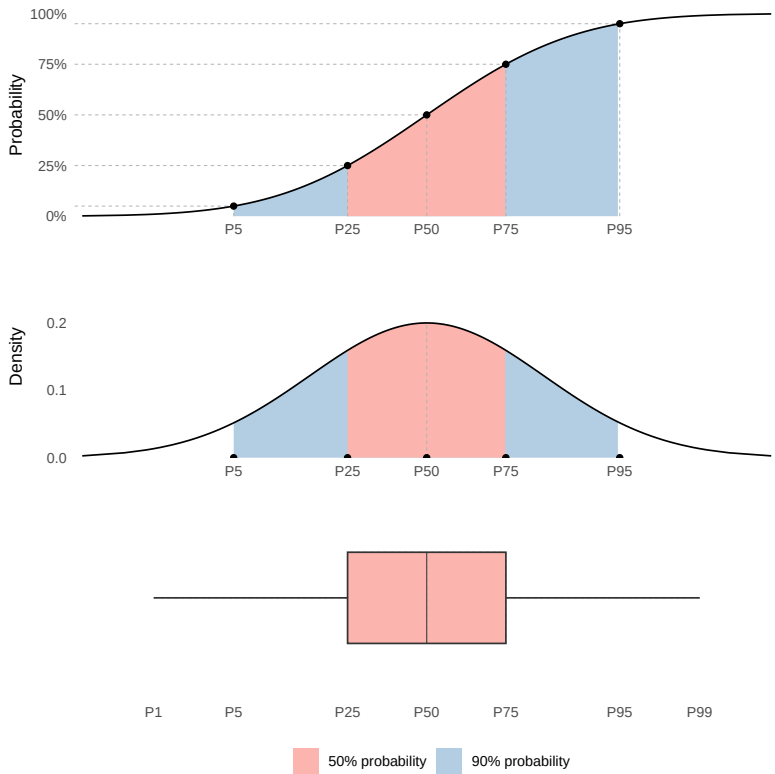


Figure 2: CDF, PDF and boxplot of an unknown distribution F(x)

- Under what conditions can a quantile-parameterized distribution be used as a prior or as a likelihood in a Bayesian model, and how can Bayesian inference be performed for QPDs?
- Will the expert-informed bias adjustment for preferential sampling improve the predictive performance of species distribution model based on opportunitistically collected data?

3 Methodological approach to the thesis

Different methodological approaches were used throughout the thesis.

- The thesis includes a comprehensive *literature review*, synthesizing the existing knowledge about quantile-parameterized distributions (**Paper I**) and likelihood-based inference with quantile functions (**Paper II**).

- I made a *theoretical development* and proposed generalization of Myerson distribution (**Paper I**) and extended Bayesian inference to parametric quantile regression (**Paper II**).
- As an example of *methodological development*, I proposed a prior for a quantile-based model and developed a new method of hybrid elicitation (**Paper III**).
- The theoretical findings were confirmed by a *simulation study* (**Paper II** and **Paper III**).
- Finally, the thesis includes a *case study*, which was used to justify the need for theoretical development and to apply the concepts explored to an environmental assessment problem (**Paper IV**). The case study is related to management of waterfowl in rural and urban landscapes of Southern Sweden.

The details of the methodological approaches are provided in the Results and Application section.

4 Environmental relevance

Expert knowledge elicitation is a method of integrating subjective knowledge into statistical or decision models with high significance in Environmental Science. The use of expert judgment complements the existing methods, addressing complexity of environmental systems and enhancing the robustness of scientific analyses. The methodological contribution of this thesis can contribute to the assessment and management of environmental problems.

The case study in waterfowl management illustrates the application of concepts developed in the thesis to support the responsible stewardship of a diverse family of water birds, including ducks, swans, and geese, categorized under the Order *Anseriformes* and the Family *Anatidae*. The primary objectives of waterfowl management involve sustaining these avian species for their nutritional, recreational, and conservation values. This involves the preservation of ecosystems crucial for supporting thriving populations while mitigating the negative impacts, such as damage to human interests, caused by the abundance of these bird species (Figure 3).

Waterfowl management programs in North America and Europe combine comprehensive monitoring, population modeling, and expert knowledge in conser-

Figure 3: Barnacle Goose by the pond in Lomma, Sweden on 2023-08-02, 17:36. Credits: Dmytro Perepolkin

vation and biology to create flyway-scale species management plans for growing populations of water birds (Madsen et al., 2017). Due to the large number of actors and national decision makers involved, the international waterfowl management bodies are constantly dealing with incomplete information and challenges with labor-intensive data collection. Therefore, a lot of attention has been dedicated to the value of information gained by the routine monitoring programs (Runge et al., 2011; Tulloch et al., 2017; Johnson et al., 2014, 2017; Roberts et al., 2018). To tackle this problem, European Goose Management Platform adopted expert-informed Bayesian population models, which explicitly account for data gaps and input uncertainties (Johnson and Koffijberg, 2021; Johnson et al., 2023).

The focus of this thesis is to further develop methods for integrating expert judgment in scientific models, and apply this theory to species distribution modeling of waterfowl in North-Eastern part of Skåne region in Sweden.

5 Results and application

The following section summarizes the results from the research work presented in the papers included in this thesis.

5.1 Expert-specified priors

Paper I proposes the use of quantile-parameterized distributions (QPDs) as a tool for translating the probabilistic judgments elicited from experts, into probability distribution for parameters. The paper includes a comprehensive review of the existing literature on QPDs and proposes a generalized version of one of the simplest QPDs, the Myerson distribution. **Paper I** also examines various methods for constructing QPDs and extending the univariate QPDs to the multivariate setting. Additionally, we introduce a multivariate extension for our Generalized Myerson distribution.

5.2 Bayesian inference with quantile functions

The QPDs discussed in **Paper I** are constructed using the inverse of the cumulative distribution function, known as the quantile function (QF). In should be noted that many of the distributions composed using the quantile function lack a closed form CDF and PDF. The challenge arises when incorporating them into Bayesian models, where likelihood is calculated as the ratio of densities.

Paper II systematically introduces and illustrates Bayesian inference in the models where prior or likelihood are expressed using a quantile function. We show that the quantile-based Bayesian inference leads to the same posterior beliefs as the conventional density-based inference (Figure 4). We validate quantile-based models using a simulation study and apply the principles of quantile-based inference to Bayesian updating of parameters in both univariate and regression settings, using flexible and extensible quantile sampling distributions.

5.3 Quantile-parameterized models and hybirid elicitation

The *predictive approach* to expert elicitation focuses on gathering information about observable quantities, possibly conditional on covariates (Kadane, 1980; Winkler, 1980). This approach is advantageous as it aligns with the intuitive understanding of experts. However, a drawback is that the predictive distribution does not distinguish between the randomness explained by the model and the uncertainty about the parameters. This lack of distinction makes it challenging to update the prior beliefs in light of new observations. On the other hand, *the structural approach* attempts to elicit the model parameters directly. However, experts may find it challenging to express their judgment in the parameter

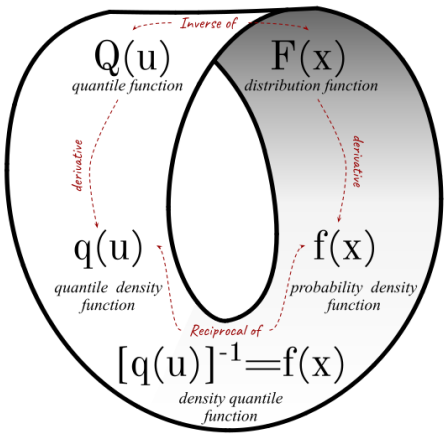


Figure 4: Moebius strip of probability functions (Perepolkin et al., 2023a)

space, due to the abstract and, at times, counter-intuitive nature of parameters within scientific models (Mikkola et al., 2023).

The *hybrid elicitation* approach introduced in **Paper III** combines the best of both worlds. By eliciting information about observable quantities along with associated uncertainty, the hybrid method integrates predictive and structural elements. The purpose of hybrid elicitation is to define a prior for a model expressed by a quantile-parameterized distribution. In this paper, we propose a variant of Dirichlet prior that effectively captures uncertainty in quantile parameters. The hybrid approach facilitates the expert elicitation and Bayesian updating of variable quantities with minimal assumptions about the underlying model structure.

5.4 Expert judgment in waterfowl species distribution modeling

One of the distinct features of environmental problems is that they are typically spatially referenced. This aspect of the environmental data requires specialized approaches, which can handle spatial autocorrelation explicitly. Integrated Laplace Approximations implemented in INLA (Rue et al., 2009; Lindgren and Rue, 2015) offer a quick and efficient method of fitting Bayesian models with latent Gaussian fields responsible for estimating the spatial extent of the effect of interest. In this thesis we used `inlabru`, a user-friendly wrapper over INLA for working with ecological survey data (Bachl et al., 2019). Recently a new functionality for specifying custom priors was added to `inlabru`: it is

now possible to specify the prior distribution for a parameter of interest using a quantile function and its derivative (quantile-based approach). In **Paper IV** we specified the prior for the scale parameter in the detection function using the quantile function of the Myerson distribuition.

6 Discussion

6.1 Data-centric vs model-centric view

There's an important tension in science, which goes across the divide between the primacy of data vs theory. In the world of machine learning it is common to advocate for "letting the evidence speak for itself". This shifts the burden of explanation to the available data and portrays data science as an objective enterprise of merely observing and impartially recording the facts. But as Gomez-Marin (2023) brilliantly summarises:

> Data does not speak for itself, we articulate its meaning with our interpretations. These, in turn, depend on our initial presuppositions, cognitive biases and philosophical committments. If data are "given" ("datum" in Latin), facts are "made" (factum). And understanding comes later, on developing a bases on which to "under" "stand" the facts that are made.

The distinction between the data-centric and the model-centric view of science can be traced to the way we define probability distributions. In the data-centric approach the probability distribution for the observable x is defined via the distribution function $F(x|\theta)$. Compare it to the distribution defined by the quantile function $Q(u|\theta)$ of the depth u (**Paper II**). The former is naturally descriptive, while the latter is naturally generative; one puts forth the data, while the other puts forth the model[1].

In the data-centric world, we are in search of a perfect F, "where the data is coming from". It is easy to believe that the distribution F is objective, that it is "out there" in the world. With quantile function Q we are forced to ask: "which quantile function Q should I pick so that the data I generate looks more or less

[1]The value $x = 0.6744898$ could be a measurement of a temperature or a distance, but the depth $u = 0.75$ is by itself not a measure of anything. However, in the context of a model (such as the standard normal), the meaning of the depth $u = 0.75$ and the data $x = 0.6744898$ is equivalent.

like what I observe in the world?". The quantile function puts a model between the observer and observables and makes it clear that we are merely attempting to match the real and messy x with our idealized $Q(u)$.

The model is always a means for one of three ends: descriptive, predictive, or causal (Carlin and Moreno-Betancur, 2023). Any inference, prediction or intervention suggested in research is always conditional on the parameterized scientific model. And the presence of parameters imply existence of prior knowledge (Gelman et al., 2017; Clayton, 2021). Even when such prior knowledge is based on a known (previously observed) distribution, an expert makes an act of personal judgment, deeming the observed relative frequencies as *relevant and sufficient* for characterizing the probability of obtaining a specific value from a randomly drawn outcome. This constitutes a shift from a "frequency" to a "personal probability" domain, based on the assumption of exchangeability (Jaynes, 2003).

6.2 Quantile-parameterized distributions

When eliciting the quantitative judgment of experts, the assessor should not inquire them about the statistical moments (mean, standard deviation), but rather about probabilities and quantiles (Kadane and Wolfson, 1998). The characterization of prior knowledge with quantile-parameterized distribution has a distinct advantage: once quantile judgments are elicited, QPDs do not require any additional steps of fitting or encoding. The elicited quantiles can become parameters of the prior distribution directly.

However, as much as QPDs are useful, they posses some important limitations.

Unintuitive tails

Even though quantile-parameterized priors guarantee that the parameterizing quantile-probability pairs will be matched exactly (**Paper I**), the chosen distribution may imply some counterintuitive values of quantiles in the tails. All symmetrically-parameterized distributions have 3-5 parameters, which means that they are more flexible than the kernel distributions they are based on. Perepolkin et al. (2023b) discuss certain distributional kernels which produce unexpected shape of the bounded tail in a QPD. Figure 5 shows J-QPD-B and Beta distribution fitted to the same quantile-probability pairs. Although J-QPD fits the parameterizing quantiles exactly, it is very unlikely that this particular

shape of the J-QPD-B matches the intuition of the expert. Such prior may also have quite dramatic effect when used in MCMC for Bayesian computation.

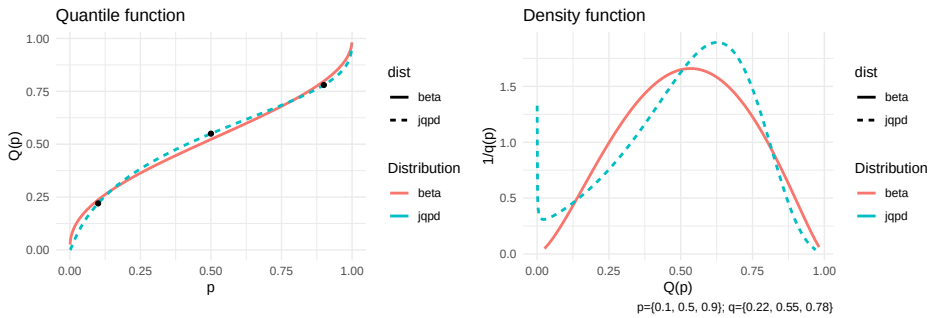


Figure 5: J-QPD-B vs Beta distribution

Implied bounds

Some quantile-parameterized distributions have explicitly defined bounds (e.g. J-QPD), while the bounds of other distributions are defined implicitly by other parameters (e.g. Myerson, GLD CSW). Implicitly defined bounds may not always correspond to the expert's knowledge about the valid domain for the quantity of interest. Therefore the assessor needs to make the implied bounds clear to the expert to confirm that those are, in fact, reasonable.

Implicit bounds have some important bearing on MCMC sampling as the values outside the bounds can not be explored by the sampler. Stan is the popular programming environment for building probabilistic Bayesian models, implementing Hamiltonian Monte Carlo (HMC) method (Gabry and Češnovar, 2022). In Stan the limits on parameters have to be explicitly defined to limit the ability of the sampler to explore the part of the posterior where density is not defined. Therefore, the implicit limits of quantile function may need to be computed ahead of time and explicitly defined as the parameter bounds to prevent the MCMC/HMC transitions from diverging.

Truncation

Another limitation of using quantile priors is related to truncation. Truncation is often used for defining distributions with a particular shape, e.g. for defining prior for variance parameters (Gelman, 2006). When a distribution is truncated

at values a, b, so that $P(X \leq x | a < X < b) = G(x)$ the CDF of a truncated distribution is

$$G_t(x) = \frac{F(\max(\min(x, b), a)) - F(a)}{F(b) - F(a)}$$

and the corresponding quantile function is

$$Q_t(u) = Q(F(a) + u[F(b) - F(a)])$$

where u is the depth corresponding to the random variable x (Nadarajah and Kotz, 2006). It can be observed that the truncated quantile function requires computing the depths $u_a = F(a)$ and $u_b = F(b)$, which for non-invertible quantile functions, may need to be computed numerically.

Infeasibility

Despite the fact that all QPDs are developed using the quantile function, the methods of reparameterization used to map parameters to quantile-probability pairs does not always guarantee the feasibility the resulting distribution (**Paper I**). The distributions created following the Gilchrist transformations (Gilchrist, 2000) are guaranteed to stay feasible for all valid ranges of parameters. However, it is precisely the transformations that violate the Gilchrist rules that create interesting, flexible and unique distribution, such as GLD (Freimer et al., 1988), g-and-h, g-and-k (Rayner and MacGillivray, 2002) and metalog (Keelin, 2016). These distributions need the "guard rails" of parameter feasibility conditions, which ensure that the quantile function stays monotonic. Some attempts were made to use reparameterization in order to move the parameter boundaries to the values that "make sense". For example Chalabi et al. (2012) parameterization of GLD (CSW GLD) repackages (Freimer et al., 1988) (FKML) GLD mapping the location and scale parameters to the median and IQR, and restricting the other two parameters (asymmetry and steepness) to $[-1, 1]$ and $[0, 1]$ ranges, respectively. This makes valid parameterization easier, but also creates a potential problem of rapid changes in the shape of the distribution with only a minor change in parameters, due to the fact that CSW GLD is a combination of four distributions (see Equation 23 in Perepolkin et al. (2023b)).

There's a need for developing new quantile distributions which are flexible yet feasible. Quantile mixtures is a promising new direction in quantile function

research (Peng et al., 2023). Quantile mixtures are formed by a linear combination of quantile functions, providing a high level of flexibility while staying feasible by construction. In comparison to density mixtures, quantile mixtures offer a distinct advantage: they are always unimodal, unless at least one of the components in the mixture is explicitly multi-modal (Gilchrist, 2000). As the weights in the linear combination of quantile functions must be non-negative, the fitting method requires constrained optimization. The resulting mixture can be mapped back to quantiles using the implicit function method (Perepolkin et al., 2023b), under the condition that all coefficients, except for the intercept, are positive.

The thesis includes the paper on metalog likelihood (**Paper III**), for which we developed a novel approach to elicitation. The model based on metalog distribution has to carefully handle the parameter feasibility condition for the metalog quantile function. The infeasible combination of quantiles should be rejected as soon as the computed quantile density function is found to be negative. The model in **Paper III** can be further improved and extended by replacing the metalog distribution with a properly constructed quantile mixture (Peng et al., 2023) with guaranteed feasibilty.

6.3 Bayesian inference with quantile functions

There could be several reasons for resorting to a likelihood for data defined by a quantile function. First of all there could be some domain-specific reasons, i.e. particular knowledge about the world that implies an unconventional data-generative process. For example, Wakeby distribution has been specifically created for modeling flood flows, and extreme rainfall (Rahman et al., 2015), but has since been adopted for modeling citation counts (Katchanov and Markova, 2015). Wakeby distribution is defined by a quantile function that does not have a closed-form inverse. Applying the principles of quantile-based inference makes the likelihood-based inference with Wakeby distribution possible.

Another reason for using the quantile-based likelihood could be parametric quantile regression (Gilchrist, 2008). Parametric quantile regression is useful when the goal is not to predict the mean, but to produce a probabilistic judgment about the *quantiles* of the quantity of interest. The resulting conditional quantiles should, of course, be subject to axioms of probability, i.e. they should be non-crossing (Gilchrist, 2007).

Quantile function methods are primarily used in relation to continuous variables. The quantile functions for discrete random variables require more rigorous defin-

ition to avoid ambiguity related to mapping the real values of probabilities to integers (Gilchrist, 2000), since CDF (and QF) of a discrete random variable is a step function. Therefore, the method in which the regression equation is represented by a quantile function will naturally have some limitations when used to model the discrete response variable.

Quantile-based likelihood models require numerical inversion of a quantile function, which is rather trivial for univariate models, but becomes computationally challenging in the presence of covariates. The numerical cost of inverting quantile functions has been studied in Perepolkin et al. (2023a). For univariate case the inversion algorithm can be optimized by ordering the observations and, therefore, incrementally reducing the root-finding interval, but in case of parametric quantile regression ordering of observations may be more difficult, because the regression quantile function includes covariates.

6.4 Expert judgment and hybrid elicitation

Mikkola et al. (2023) propose a set of useful distinctions which divide the field of prior elicitation into sub-categories based on the hypercube of dimensions dealing with the expert, the model, and the elicitation process itself. One of the dimensions the authors discuss is *elicitation space*, where they distinguish between the parameter and observable space. The elicitation literature strongly favors querying the experts about the quantities of interest in the observable space (Kadane and Wolfson, 1998). In **Paper III** we propose *hybrid elicitaion* approach, in which the experts are asked to describe only observable quantities and their subjective uncertainty about those quantities. The key idea of the method is to translate the uncertainty about cumulative probabilities corresponding to quantiles of an observable quantity into the proportions of a hypothetical sample, thus switching from the relative frequency to the natural frequency frame (Gigerenzer, 2011).

Figure 6 shows the (predictive) 10th, 50th, and 90th quantiles of fish weights elicited from an expert corresponding to the weights of 4, 9 and 17 lbs (interpolated with a 3-term metalog). We treat the quantile values as fixed and consider the "probability band" widths (highlighted by different colors on the chart) as varying. The hybrid elicitation method amounts to inquiring the expert about their personal uncertainty about the widths of these probability bands.

Note, that the widths of the probability bands will always sum to 1. Therefore, we propose a prior based on Dirichlet distribution, although the same elicited quantities can be used to fit a more flexible Connor-Mosimann (general-

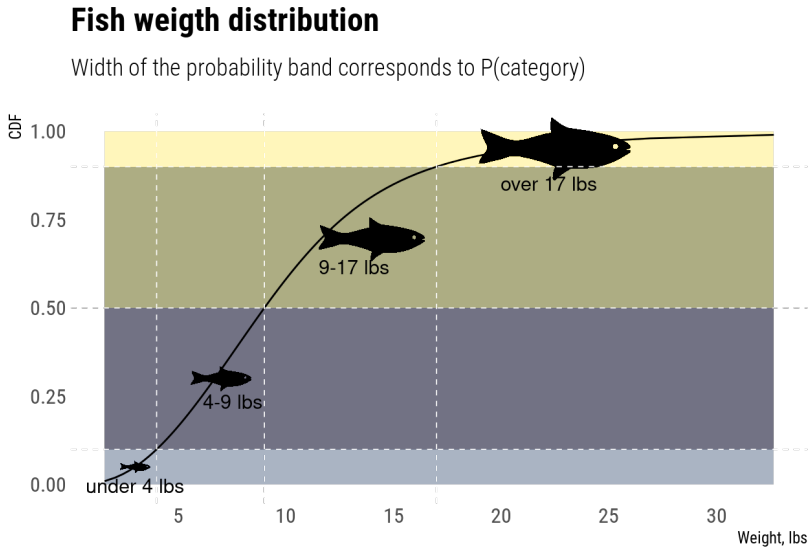


Figure 6: Elicited quantiles of steelhead trout weights, from Perepolkin et al. (2023a)

ized Dirichlet) distribution (Elfadaly and Garthwaite, 2013). The choice of the multivariate distribution can represent a source of ambiguity by itself. Besides, when estimating the conditional beta distributions, from which the multivariate prior is constructed, the elicited quartiles are fitted using the improved Pratt et al. (1995)'s algorithm, which also carries some approximation error.

The elicited QDirichlet prior (together with a QPD model) constitute a probabilistic package which can be unfolded into 2D Monte Carlo (2DMC) samples (Simon et al., 2015; Michiels and Geeraerd, 2022), where the aleatory and epistemic uncertainty have been separated through elicitation. More research is needed to test the perceptions of experts towards this method of encoding expert judgment and to investigate the potential limitations of the method when applied to real-life problems.

6.5 Observer-specific adjustments to species distribution models

Citizen science databases, such as GBIF, provide a rich source of wildlife observational data which is difficult to make a good use of. They represent a perfect

example of a principle attributed to Hal Stern by Gelman (2021): "The most important thing is what data you use, not what you do with the data".

In *Paper IV* we address this challenge by identifying a subset of the GBIF data with a higher signal-to-noise ratio. Specifically, we focus on two aspects:

- Some contributors to GBIF aim to document every species they encounter, but distinguishing these records from selectively observed ones remains a challenge. For instance, observations recorded in eBird (Sullivan et al., 2009) system allow users to mark their checklists as *complete*, indicating a more systematic approach to wildlife observations, which is closer to the one adopted by the Swedish Bird Survey (Fågeltaxering, 2023).
- Some users report numerous species from a single wildlife visit, while others report only a few (or even a single observation). We hypothesize that those reporting longer species lists may follow checklists or just diligently record all observations. This suggests potentially greater comprehensiveness of their observations compared to users registering only rare or surprising species sightings.

The core concept of *Paper IV* revolves around these two key insights.

We connect the eBird dataset to the GBIF sample and mark the observations belonging to complete checklists as assumed to be less biased, setting them aside as holdout data for predictive testing. We also consider the lengths of reported species lists as a proxy for observation effort (Szabo et al., 2010).

Our approach aligns with the distance sampling method, commonly used to address spatial preferential sampling bias. Following the distance sampling methodology, we define the 'ideal' or 'unbiased' state (i.e. the situation where all available species are registered), and introduce a metric, Species List Shortage, to quantify the deviation of a specific species list from this ideal. We explore different methods of incorporating Species List Shortage into the model, either as a covariate with a diffuse prior or as a thinner with an expert-informed prior encoded by a Quantile-Parameterized Distribution (QPD).

It is essential to acknowledge several limitations in our approach. Firstly, we aggregated multiple species into a single response count for prediction, focusing on the presence and abundance of *all goose species* rather than individual ones. This decision was driven by the need to deal with the inherent sparsity and overdispersion in the data.

Species distribution models, aiming to estimate both presence and abundance, exhibit sensitivity to mesh construction and the specification of priors, as highlighted in recent research (Dambly et al., 2023). Our study focuses on the prior for the scale parameter of the thinning factor within the species distribution model. Results demonstrate a discernible impact on the posterior.

Our method retains simplicity through the adoption of a parametric Bayesian model with only a few covariates. This simplicity is complemented by the versatile spatial modeling capabilities offered by INLA (Lindgren and Rue, 2015), providing a powerful analytical framework. Our findings underscore the potential of species list lengths to enhance predictive performance of models based on presence-only data. We advocate for further exploration to determine the optimal utilization of this approach within the context of the abundant citizen science data.

7 Conclusion and future outlook

Expert knowledge is an essential part of solving complex problems. Human judgment is required in all stages of environmental decision making from specifying objectives to providing priors in scientific models. While inherently subjective, the expression of uncertainty by experts can achieve scientific rigor through structured elicitation processes and effective integration into scientific models (O'Hagan, 2019). Adopting the language of uncertainty, facilitated by appropriate probabilistic methods, is paramount for a robust scientific process (Spiegelhalter, 2014).

Genuine scientific inquiry commences not with raw data but by articulating the model and making prior knowledge and uncertainty explicit (Gelman and Hennig, 2017). The case study included in the thesis emphasizes the importance of not merely accumulating more data but also critically assessing the data collection process, as it fundamentally shapes the utility of the collected data.

This thesis contributes to the advancement of distribution theory, particularly in the application of quantile methods for statistical inference. I hope this work inspires new research in quantile functions, Bayesian species distribution models and expert knowledge elicitation for making a positive and lasting impact in the world.

8 Bibliography

Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., 2019. Inlabru: An R package for Bayesian spatial modelling from ecological survey data. Methods in Ecology and Evolution 10, 760–766. doi:10.1111/2041-210X.13168.

Burgman, M.A., 2015. Trusting Judgements: How to Get the Best out of Experts. 1 ed., Cambridge University Press. doi:10.1017/CBO9781316282472.

Carlin, J.B., Moreno-Betancur, M., 2023. On the uses and abuses of regression models: A call for reform of statistical practice and teaching. doi:10.48550/arXiv.2309.06668, arXiv:2309.06668.

Chalabi, Y., Scott, D.J., Wuertz, D., 2012. Flexible Distribution Modeling with the Generalized Lambda Distribution. Working Paper MPRA Paper No. 43333,. ETH. Zurich, Switzerland.

Choy, S.L., O'Leary, R., Mengersen, K., 2009. Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. Ecology 90, 265–277. doi:10.1890/07-1886.1.

Clayton, A., 2021. Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science. Columbia University Press, New York.

Cretois, B., 2021. Transforming the Use of Citizen Science Data for Biodiversity Conservation at Different Scales. Ph.D. thesis. NTNU. Trondheim, Norway.

Dambly, L.I., Isaac, N.J.B., Jones, K.E., Boughey, K.L., O'Hara, R.B., 2023. Integrated species distribution models fitted in INLA are sensitive to mesh parameterisation. Ecography 2023, e06391. doi:10.1111/ecog.06391.

Di Cecco, G.J., Barve, V., Belitz, M.W., Stucky, B.J., Guralnick, R.P., Hurlbert, A.H., 2021. Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. BioScience 71, 1179–1188. doi:10.1093/biosci/biab093.

Dias, L.C., Morton, A., Quigley, J. (Eds.), 2018. Elicitation: The Science and Art of Structuring Judgement. volume 261 of *International Series in Operations Research & Management Science*. Springer International Publishing, Cham. doi:10.1007/978-3-319-65052-4.

Elfadaly, F.G., Garthwaite, P.H., 2013. Eliciting Dirichlet and Connor–Mosimann prior distributions for multinomial models. TEST 22, 628–646. doi:10.1007/s11749-013-0336-4.

European Food Safety Authority, 2014. Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA Journal 12, 3734.

Fågeltaxering, 2023. Metoder — Svensk Fågeltaxering. http://www.fageltaxering.lu.se/inventera/metoder.

Freimer, M., Kollia, G., Mudholkar, G.S., Lin, C.T., 1988. A study of the generalized Tukey lambda family. Communications in Statistics-Theory and Methods 17, 3547–3567. doi:10.1080/03610928808829820.

Gabry, J., Češnovar, R., 2022. Cmdstanr: R Interface to 'CmdStan'.

GBIF, 2023. What is GBIF? https://www.gbif.org/what-is-gbif.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1, 515–534. doi:10.1214/06-BA117A.

Gelman, A., 2021. Reflections on Breiman's Two Cultures of Statistical Modeling. Observational Studies 7, 95–98. doi:10.1353/obs.2021.0025.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. CRC press.

Gelman, A., Hennig, C., 2017. Beyond subjective and objective in statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180, 967–1033. doi:10.1111/rssa.12276.

Gelman, A., Simpson, D., Betancourt, M., 2017. The Prior Can Often Only Be Understood in the Context of the Likelihood. Entropy 19, 555. doi:10.3390/e19100555, arXiv:1708.07487.

Gigerenzer, G., 2011. What are natural frequencies? BMJ 343, d6386–d6386. doi:10/cnzvf7.

Gilchrist, W., 2000. Statistical Modelling with Quantile Functions. Chapman & Hall/CRC, Boca Raton.

Gilchrist, W., 2008. Regression Revisited. International Statistical Review 76, 401–418. doi:10.1111/j.1751-5823.2008.00053.x.

Gilchrist, W.G., 2007. Modeling and Fitting Quantile Distributions and Regressions. American Journal of Mathematical and Management Sciences 27, 401–439. doi:10.1080/01966324.2007.10737707.

Gomez-Marin, A., 2023. The Consciousness of Neuroscience. eneuro 10, ENEURO.0434–23.2023. doi:10.1523/ENEURO.0434-23.2023.

Gosling, J.P., 2018. SHELF: The Sheffield elicitation framework, in: Elicitation. Springer, pp. 61–93.

Grimmett, L., Whitsed, R., Horta, A., 2020. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. Ecological Modelling 431, 109194. doi:10.1016/j.ecolmodel.2020.109194.

Hadlock, C.C., 2017. Quantile-Parameterized Methods for Quantifying Uncertainty in Decision Analysis. Ph.D. thesis. University of Texas. Austin, TX. doi:10.15781/T2F18SX41.

Hanea, A.M., Hemming, V., Nane, G.F., 2021a. Uncertainty Quantification with Experts: Present Status and Research Needs. Risk Analysis , risa.13718doi:10.1111/risa.13718.

Hanea, A.M., Nane, G.F., Bedford, T., French, S. (Eds.), 2021b. Expert Judgement in Risk and Decision Analysis. volume 293 of *International Series in Operations Research & Management Science*. Springer International Publishing, Cham. doi:10.1007/978-3-030-46474-5.

Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy. Ecography 36, 864–867. doi:10.1111/j.1600-0587.2013.00321.x.

Isaac, N.J.B., Pocock, M.J.O., 2015. Bias and information in biological records. Biological Journal of the Linnean Society 115, 522–531. doi:10.1111/bij.12532.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. 1 ed., Cambridge University Press. doi:10.1017/CBO9780511790423.

Johnson, F.A., Heldbjerg, H., Nagy, S., Madsen, J., 2021. Setting population-size targets for geese causing socio-economic conflicts. Ambio doi:10.1007/s13280-021-01539-5.

Johnson, F.A., Jensen, G.H., Madsen, J., Williams, B.K., 2014. Uncertainty, robustness, and the value of information in managing an expanding Arctic goose population. Ecological Modelling 273, 186–199. doi:10.1016/j.ecolmodel.2013.10.031.

Johnson, F.A., Koffijberg, K., 2021. Biased monitoring data and an info-gap model for regulating the offtake of greylag geese in Europe. Wildlife Biology 2021, wlb.00803. doi:10.2981/wlb.00803.

Johnson, F.A., Madsen, J., Clausen, K.K., Frederiksen, M., Jensen, G.H., 2023. Assessing the value of monitoring to biological inference and expected management performance for a European goose population. Journal of Applied Ecology 60, 132–145. doi:10.1111/1365-2664.14313.

Johnson, F.A., Smith, B.J., Bonneau, M., Martin, J., Romagosa, C., Mazzotti, F., Waddle, H., Reed, R.N., Eckles, J.K., Vitt, L.J., 2017. Expert Elicitation, Uncertainty, and the Value of Information in Controlling Invasive Species. Ecological Economics 137, 83–90. doi:10.1016/j.ecolecon.2017.03.004.

Kadane, J., Wolfson, L.J., 1998. Experiences in elicitation. Journal of the Royal Statistical Society: Series D (The Statistician) 47, 3–19. doi:10/cvdn73.

Kadane, J.B., 1980. Predictive and structural methods for eliciting prior distributions, in: Zellner, A. (Ed.), Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys. North Holland Publishing Company, Amsterdam, pp. 89–93.

Kahneman, D., 2012. Thinking, Fast and Slow. Penguin Psychology, Penguin Books, London.

Kahneman, D., Sibony, O., Sunstein, C.R., 2021. Noise: A Flaw in Human Judgment. Little, Brown.

Katchanov, Y.L., Markova, Y.V., 2015. On a heuristic point of view concerning the citation distribution: Introducing the Wakeby distribution. SpringerPlus 4, 94. doi:10.1186/s40064-015-0821-1.

Keelin, T.W., 2016. The Metalog Distributions. Decision Analysis 13, 243–277. doi:10.1287/deca.2016.0338.

Keelin, T.W., Powley, B.W., 2011. Quantile-Parameterized Distributions. Decision Analysis 8, 206–219. doi:10.1287/deca.1110.0213.

Lindgren, F., Rue, H., 2015. Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software 63, 1–25. doi:10.18637/jss.v063.i19.

Madsen, J., Williams, J.H., Johnson, F.A., Tombre, I.M., Dereliev, S., Kuijken, E., 2017. Implementation of the first adaptive management plan for a European migratory waterbird population: The case of the Svalbard pink-footed goose Anser brachyrhynchus. Ambio 46, 275–289. doi:10.1007/s13280-016-0888-0.

Michiels, F., Geeraerd, A., 2022. Two-dimensional Monte Carlo simulations in LCA: An innovative approach to guide the choice for the environmentally preferable option. The International Journal of Life Cycle Assessment 27, 505–523. doi:10.1007/s11367-022-02041-0.

Mikkola, P., Martin, O.A., Chandramouli, S., Hartmann, M., Pla, O.A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.C., Klami, A., 2023. Prior Knowledge Elicitation: The Past, Present, and Future. Bayesian Analysis -1, 1–33. doi:10.1214/23-BA1381.

Milanesi, P., Mori, E., Menchetti, M., 2020. Observer-oriented approach improves species distribution models from citizen science data. Ecology and Evolution 10, 12104–12114. doi:10.1002/ece3.6832.

Morgan, M.G., 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National Academy of Sciences 111, 7176–7184. doi:10.1073/pnas.1319946111.

Nadarajah, S., Kotz, S., 2006. R Programs for Truncated Distributions. Journal of Statistical Software 16, 1–8. doi:10.18637/jss.v016.c02.

O'Hagan, A., 2019. Expert Knowledge Elicitation: Subjective but Scientific. The American Statistician 73, 69–81. doi:10.1080/00031305.2018.1518265.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities: O'Hagan/Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons, Ltd, Chichester, UK. doi:10.1002/0470033312.

Peng, C., Li, Y., Uryasev, S., 2023. Mixture Quantiles Estimated by Constrained Linear Regression. doi:arXiv:2305.00081[stat], arXiv:2305.00081.

Perepolkin, D., Goodrich, B., Sahlin, U., 2023a. The tenets of quantile-based inference in Bayesian models. Computational Statistics & Data Analysis 187, 107795. doi:10.1016/j.csda.2023.107795.

Perepolkin, D., Lindström, E., Sahlin, U., 2023b. Quantile-parameterized distributions for expert knowledge elicitation. doi:10.31219/osf.io/tq3an.

Powley, B.W., 2013. Quantile Function Methods for Decision Analysis. Ph.D. thesis. Stanford University. Paolo Alto, CA.

Pratt, J.W., Raiffa, H., Schlaifer, R., 1995. Introduction to Statistical Decision Theory. MIT press.

Rahman, A., Zaman, M.A., Haddad, K., El Adlouni, S., Zhang, C., 2015. Applicability of Wakeby distribution in flood frequency analysis: A case study for eastern Australia. Hydrological Processes 29, 602–614. doi:10/f6wzmh.

Rayner, G.D., MacGillivray, H.L., 2002. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Statistics and Computing 12, 57–75. doi:10.1023/a:1013120305780.

Roberts, A., Eadie, J.M., Howerter, D.W., Johnson, F.A., Nichols, J.D., Runge, M.C., Vrtiska, M.P., Williams, B.K., 2018. Strengthening links between waterfowl research and management. The Journal of Wildlife Management 82, 260–265. doi:10.1002/jwmg.21333.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology 71, 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Ruete, A., Arlt, D., Berg, Å., Knape, J., Żmihorski, M., Pärt, T., 2020. Cannot see the diversity for all the species: Evaluating inclusion criteria for local species lists when using abundant citizen science data. Ecology and Evolution 10, 10057–10065. doi:10.1002/ece3.6665.

Ruete, A., Pärt, T., Berg, Å., Knape, J., 2017. Exploiting opportunistic observations to estimate changes in seasonal site use: An example with wetland birds. Ecology and Evolution 7, 5632–5644. doi:10.1002/ece3.3100.

Runge, M.C., Converse, S.J., Lyons, J.E., 2011. Which uncertainty? Using expert elicitation and expected value of information to design an adaptive program. Biological Conservation 144, 1214–1223. doi:10.1016/j.biocon.2010.12.020.

Sarma, A., Kay, M., 2020. Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–12. doi:10.1145/3313831.3376377.

Sicacha-Parada, J., Steinsland, I., Cretois, B., Borgelt, J., 2021. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. Spatial Statistics 42, 100446. doi:10.1016/j.spasta.2020.100446.

Simon, S.L., Hoffman, F.O., Hofer, E., 2015. The Two-Dimensional Monte Carlo: A New Methodologic Paradigm for Dose Reconstruction for Epidemiological Studies. Radiation research 183, 27–41. doi:10.1667/RR13729.1.

Spiegelhalter, D.J., 2014. The future lies in uncertainty. Science 345, 264–265. doi:10.1126/science.1251122.

Stolar, J., Nielsen, S.E., 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Diversity and Distributions 21, 595–608. doi:10.1111/ddi.12279.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142, 2282–2292. doi:10.1016/j.biocon.2009.05.006.

Szabo, J.K., Vesk, P.A., Baxter, P.W.J., Possingham, H.P., 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. Ecological Applications 20, 2157–2169. doi:10.1890/09-0877.1, arXiv:29779611.

Tulloch, A.I.T., Nicol, S., Bunnefeld, N., 2017. Quantifying the expected value of uncertain management choices for over-abundant Greylag Geese. Biological Conservation 214, 147–155. doi:10.1016/j.biocon.2017.08.013.

Winkler, R.L., 1980. Prior information, predictive distributions, and Bayesian model-building. Bayesian Analysis in Econometrics and Statistics. North-Holland Publishing Company , 95–109.

Scientific publications

Author contributions

Co-authors are abbreviated as follows: Erik Lindström (EL), Benjamin Goodrish (BG), Johan Elmberg (JE), Finn Lindgren (FL), Ullrika Sahlin (US).

Paper I: Quantile-parameterized distributions for expert knowledge elicitation.

I conceived the idea, wrote the first draft of the manuscript and incorporated the suggestions by US and EL. All authors reviewed and contributed to the paper.

Paper II: The tenets of quantile-based inference in Bayesian models.

I conceived the idea and wrote the first draft of the manuscript. BG provided input on implementation of models in Stan. All authors reviewed and contributed to the paper.

Paper III: Hybrid elicitation and quantile-parametrized likelihood.

I conceived the idea. US contributed a suggestion for the form of the prior. BG provided input on implementation of models in Stan. All authors reviewed and contributed to the paper.

Paper IV: Observer-adjusted species distribution models using presence-only data.

I and JE conceived the idea. I wrote the first draft of the manuscript. FL contributed to model implementation in `inlabru`. US and JE provided feedback on the manuscript. US provided valuable contribution to the interpretation of the results.

Paper I

Quantile-parameterized distributions for expert knowledge elicitation

Dmytro Perepolkin[a,1,*], Erik Lindström[b], Ullrika Sahlin[a]

[a] *Centre for Environmental and Climate Science Solvegatan 37 223 62 Lund Sweden*
[b] *Centre for Mathematical Sciences Lund University Sweden*

Abstract

This paper presents a comprehensive overview of quantile-parameterized distributions (QPDs) as a tool for capturing expert predictions and parametric judgments. We survey various types of QPDs covered in the literature and focus on the Myerson distribution as the simplest method of parameterizing a distribution by a set of quantile-probability pairs. We propose the generalization of the Myerson distribution to increase the flexibility of its tails. Additionally, we explore the extension of QPDs to the multivariate setting, discussing methods for constructing bivariate distributions with quantile-parameterized margins.

Keywords: quantile-parameterized distributions, quantile functions, expert knowledge elicitation

1. Introduction

Judgment plays a crucial role in transforming raw data into meaningful insights. For judgment to be useful, it needs to be translated into the language of mathematical models and assumptions. These models are designed to capture the expert's understanding of the world, including the causal links between relevant entities. The models serve as a representation of this understanding, while also accounting for any limitations in knowledge, which are treated as uncertainties. The process of elicitation involves translating the qualitative understanding of the problem at hand into quantitative models that can provide valuable insights.

Most of the expert elicitation protocols described in the literature (Hanea et al., 2021; Gosling, 2018; O'Hagan et al., 2006; Hemming et al., 2018; Morgan, 2014; Welsh and Begg, 2018; Spetzler and Staël Von Holstein, 1975) encode expert judgments about the parameter or quantity of interest as an ordered set of quantiles with corresponding probabilities. This typically includes measures such as the median and the upper and lower quartiles. Assessors are then encouraged to select a probability distribution that reasonably fits the elicited quantile-probability pairs and validate the choice with the expert (Gosling, 2018). A distribution is selected from a predefined set of "simple and convenient" distributions with boundedness that accounts for the nature of the elicited quantity (O'Hagan et al., 2006).

Several specialized distributions have been developed to simplify and streamline the process of smooth interpolation of probabilistic assessments. These distributions, parameterized by quantile-probability pairs, ensure that the elicited QPPs are exactly preserved (Keelin and Powley, 2011; Powley, 2013; Hadlock, 2017). Quantile-parameterized distributions are particularly valuable thanks to the interpretability of their parameters. By leveraging the elicited quantiles, these distributions enable precise capturing of expert knowlegde while maintaining a high level of flexibility in modeling.

In this paper, we conduct a comprehensive review of the existing literature on quantile-parameterized distributions (QPDs) and propose a generalized version of one of the simplest QPDs, namely the Myerson distribution. Our investigation encompasses an examination of various methods for constructing QPDs and

*Corresponding author
Email addresses: dmytro.perepolkin@cec.lu.se (Dmytro Perepolkin), erik.lindstrom@matstat.lu.se (Erik Lindström), ullrika.sahlin@cec.lu.se (Ullrika Sahlin)

[1] The article includes online Supplementary Materials.

extending univariate QPDs to the multivariate setting. Additionally, we put forth a multivariate extension for our Generalized Myerson distribution.

We believe that the primary utility of QPDs lies in their ability to simplify the specification of probability distributions for model parameters, known as *prior elicitation* (Mikkola et al., 2021). However, these same distributions can also be employed to describe an expert's predictions for the next observation, referred to as *predictive elicitation* (Winkler, 1980; Kadane, 1980; Akbarov, 2009; Hartmann et al., 2020), or to capture both uncertainty and variability through a two-dimensional probability distribution in *hybrid elicitation* (Perepolkin et al., 2021a). Through our comprehensive review and identification of research gaps, we aim to contribute to the development of flexible and extensible distributions that can effectively capture expert knowledge.

The paper is structured as follows:

In Section 2, we revisit the approaches to quantile parameterization of probability distributions and explore how QPDs can effectively describe expert beliefs regarding model parameters or predictions.

Moving on to Section 3, we conduct a comprehensive review and comparison of various continuous univariate QPDs found in the literature. Specifically, we focus on the Myerson distribution and propose its generalization to accommodate different tail thicknesses. To assess the flexibility and behavior of the QPDs, we compare their robust moments. This comparative analysis can guide the selection of an appropriate distribution to characterize the quantity of interest.

In Section 4, we delve into several methods for extending QPDs to a multivariate setting. These methods include the utilization of standard multivariate distributions (Drovandi and Pettitt, 2011), copulas (Hoff, 2007), and bivariate quantiles (Nair and Vineshkumar, 2023). We apply these techniques to develop the bivariate version of the Generalized Myerson distribution and demonstrate its application in parametric and predictive elicitation.

Finally, in Section 5, we discuss future research directions and potential applications of QPDs in Bayesian analysis.

2. Quantile parameterization of probability distributions

In Bayesian data analysis, a fundamental principle is that learning from data requires more than just formulating hypotheses and models. It necessitates the articulation of prior beliefs, expressing existing knowledge in a mathematical form and translating it into a probability distribution for the model parameters.

To accurately translate knowledge into the language of statistical models the encoding distribution needs to be flexible, the process should be transparent, and the results must be interpretable. For continuous distributions, elicitation often consists of capturing a series of quantile-probability pairs (QPPs) (Kadane and Wolfson, 1998; Morgan, 2014), and then fitting a distribution to these pairs (O'Hagan, 2019). However, in practice, the choice of a parametric distribution to fit the elicited QPPs is often influenced by concerns about conjugacy with the selected statistical model that represents the data-generative process (the likelihood) and/or the availability of required distribution functions and fitting algorithms in the software employed. Frequently, the selected distribution possesses fewer parameters than the number of elicited QPPs, which can result in a less-than-perfect fit (O'Hagan, 2019). For instance, it is common to elicit three quantiles (the median along with an upper and lower quartile) and subsequently attempt to fit a normal or lognormal distribution (which features two parameters) to these points.

An alternative approach to characterizing the distribution of predictions or parameters is through quantile-parameterized distributions (QPDs). These distributions are parameterized by the QPPs, allowing the elicited values to directly define the distribution, thereby ensuring a good fit and interpretability of the parameters. The QPDs examined in this paper can accommodate a wide range of shapes and boundedness, making them valuable for accurately representing experts' prior beliefs.

Parameterizing distributions using a vector of quantiles is not a novel concept in the scientific community. The earliest mention can be traced back to the *substitution likelihood* proposed by Jeffreys (1939), which outlines a non-parametric procedure for inferring the median using a set of sample quantiles. Subsequently, similar ideas were further developed by Boos and Monahan (1986), Lavine (1995), and Dunson and Taylor (2005).

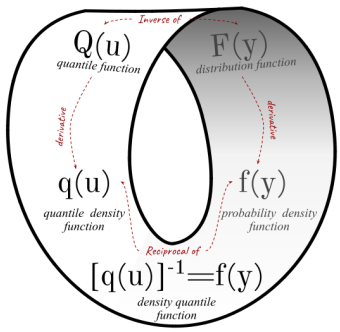


Figure 1: Moebius strip of probability functions (Perepolkin et al., 2021b).

All the QPDs found in the literature are constructed using the *quantile function*. These distributions are built either by transforming simpler quantile functions or by fitting the quantiles, as described below.

Let Y be a random variable with a (cumulative) distribution function (CDF) denoted as $F_Y(y|\theta)$. The quantile function (QF) $Q_Y(u|\theta)$ for Y is defined as

$$Q_Y(u|\theta) = \inf\{y : F_Y(y|\theta) \geq u\},\ u \in [0, 1] \tag{1}$$

Here, θ represents the distribution parameter, and the subscript $_Y$ indicates that the depth u corresponds to the random variable Y.

Both the CDF and the QF are considered equally valid ways of defining a distribution (Tukey, 1965). For a quantile function that is right-continuous and strictly increasing over the support of Y, the quantile function $Q_Y(u)$ is simply the inverse of the distribution function, denoted as $Q_Y(u|\theta) = F_Y^{-1}(u|\theta)$. Therefore, the quantile function is often referred to as the *inverse CDF*.

The derivative of the quantile function, known as the *quantile density function* (QDF), is denoted as $q(u) = \frac{dQ(u)}{du}$. It is reciprocally related to the probability density function (PDF) $f(x)$, such that $f(Q(u))q(u) = 1$. The quantity $f_Y(Q_Y(u|\theta)) = [q_Y(u|\theta)]^{-1}$ is referred to as the *density quantile* function (Parzen, 1979) or *p-pdf* (Gilchrist, 2000). The relationships between these functions are concisely illustrated in the probability function Möbius loop (Figure 1), as described in Perepolkin et al. (2021b).

Although many of the distributions discussed in Section 3 have closed-form cumulative distribution functions (CDFs) and probability density functions (PDFs), the functional form of the quantile function (QF) is often simpler and can be reasoned about in terms of other quantile functions, following *Gilchrist's QF transformation rules* summarized in Table 1 (Gilchrist, 2000). This table presents the addition, linear combination, and multiplication rules, which involve two quantile functions Q_1 and Q_2. We will refer to these three rules as *Gilchrist combinations*, as they represent valid ways to combine quantile functions to create new quantile functions.

The quantile-parameterized distributions described in this paper can be categorized into two groups based on their construction method. The first group comprises distributions that are *directly* parameterized by the quantile-probability pairs (QPPs). This group includes the Myerson distribution (Myerson, 2005), and the Johnson Quantile-Parameterized Distribution (Hadlock and Bickel, 2017, 2019). These distributions are constructed by reparameterizing or transforming existing distributions, following Gilchrist rules (Table 1). The transformations used to construct them are detailed in the next section.

The other group of distributions is *indirectly* parameterized by the QPPs. They require a fitting step where the quantile-probability pairs are translated into distribution parameters, usually through optimization

Table 1: Gilchrist's quantile function transformation rules (Gilchrist, 2000)

Original QF	Rule	Resulting QF	Resulting variable
$Q_Y(u)$	Reflection rule	$-Q(1-u)$	QF of -Y
$Q_Y(u)$	Reciprocal rule	$1/Q(1-u)$	QF of $1/Y$
$Q_1(u), Q_2(u)$	Addition rule	$Q_1(u)+Q_2(u)$	valid QF
$Q_1(u), Q_2(u)$	Linear combination rule	$aQ_1(u)+bQ_2(u)$	valid QF for $a, b > 0$
$Q_1(u), Q_2(u) > 0$	Multiplication rule	$Q_1(u)Q_2(u)$	valid QF
$Q_Y(u)$	Q-transformation	$T(Q_Y(u))$	QF of $T(Y)$, $T(Y)$ non-decreasing
$Q_Y(u)$	p-transformation	$Q_Y(H(u))$	p-transformation of $Q_Y(u)$, $H(u)$ non-decreasing

or least-squares methods. This group includes the Simple Q-Normal (Keelin and Powley, 2011), Metalog (Keelin, 2016), quantile mixtures (Peng et al., 2023), the variant of the Generalized Lambda Distribution (GLD) by Chalabi et al. (2012), and the quantile-parameterized Triangular (Two-Sided Power) distribution by Kotz and Van Dorp (2004). The fitting methods for each of these distributions is described in the respective sub-sections below.

3. Univariate quantile-parameterized distributions

In this section, we review various continuous univariate distributions that are parameterized by quantile-probability pairs found in the literature. We also introduce the generalized form for one of these distributions, namely the Myerson distribution. For each distribution, we present its quantile function and discuss the parameterization and feasibility conditions. The derivative and inverse of each distribution can be found in Appendix A.

3.1. Myerson distribution

One of the earliest examples of a distribution parameterized by quantiles is the generalized log-normal distribution proposed by Myerson (2005). It relies on a transformation of the normal quantile function.

The Myerson distribution (Myerson, 2005) is parameterized by three quantile values $\{q_1, q_2, q_3\}$, which correspond to the cumulative probabilities $\{\alpha, 0.5, 1-\alpha\}$. These quantiles are symmetrical around the median and are defined by the tail parameter $0 < \alpha < 0.5$. This type of parameterization is known as the symmetric percentile triplet (α-level SPT or α-SPT) and is also used in several other quantile-parameterized distributions that we will describe below. The Myerson quantile function is

$$\begin{aligned} \rho = q_3 - q_2;\ \beta = \frac{\rho}{q_2 - q_1};\ \kappa(u) = \frac{S(u)}{S(1-\alpha)} \\ Q_Y(u|q_1, q_2, q_3, \alpha) = \begin{cases} q_2 + \rho\frac{\beta^{\kappa(u)}-1}{\beta-1}, & \beta \neq 1 \\ q_2 + \rho\kappa(u), & \beta = 1 \end{cases} \end{aligned} \tag{2}$$

Here, u represents the depth of the observations of the random variable Y given the parameterizing α-SPT $\{q_1, q_2, q_3, \alpha\}$, with $0 < \alpha < 0.5$. The parameter ρ is the *upper p-difference*, and β is the ratio of the inter-percentile ranges, known as the *skewness ratio* (Gilchrist, 2000, p.72). The *kernel* quantile function $S(u)$ is equal to the quantile function of the standard normal distribution, also referred to as the probit, defined as $S(u) = \Phi^{-1}(u)$. The formulas for the derivative and the inverse quantile function of the Myerson QPD can be found in Appendix A.

It is important to note that while the Myerson distribution includes the normal distribution as a special case when the skewness parameter $\beta = 1$, it can exhibit right-skewness or left-skewness for other values of β. In

the symmetrical case, the range of the quantile function is $(-\infty, \infty)$. For the right-skewed distribution ($\beta > 1$), the range is $(q_2 - \frac{\rho}{\beta-1}, \infty)$, and for the left-skewed distribution ($0 < \beta < 1$), the range is $(-\infty, q_2 - \frac{\rho}{\beta-1})$. The limiting case of the skewed Myerson distribution $\lim_{u\to 0} Q_Y(u|\theta)$ for $\beta > 1$ (and the other limit for $0 < \beta < 1$) possesses some important properties that we discuss in Section 3.3 below.

The basic quantile function (Gilchrist, 2000; Lampasi, 2008) underlying the Myerson distribution is a simple probit function, denoted as $S(u) = \Phi^{-1}(u)$, transformed using the exponentiation function $T(x) = \beta^x$, where $\beta > 0$ represents the skewness ratio (Gilchrist, 2000). The quantile parameterization is facilitated by $\kappa(u)$, which takes values $\{-1, 0, 1\}$ for the three quantiles $\{q_1, q_2, q_3\}$, such that $Q(\alpha) = q_1$, $Q(0.5) = q_2$, and $Q(1-\alpha) = q_3$.

3.2. Johnson Quantile-Parameterized Distribution

Hadlock (2017) reviewed the existing quantile-parameterized distributions and proposed the quantile parameterization of the Johnson SU family of distributions (Johnson et al., 1994). In their paper, Hadlock and Bickel (2017) presented two versions of the distribution: the bounded (J-QPD-B) and the semi-bounded (J-QPD-S), both parameterized by an SPT $\{q_1, q_2, q_3, \alpha\}$ and the bound(s).

The J-QPD-B distribution is obtained by applying the inverse-probit transformation to the Johnson SU quantile function $Q_{SU}(u) = \xi + \lambda \sinh(\delta(S(u) + \gamma))$, where δ and γ are two shape parameters. This function is then rescaled to the interval $[l_b, u_b]$. The J-QPD-B quantile funciton is

$$Q_B(u|q_1, q_2, q_3, \alpha) = \begin{cases} l + (u_b - l_b) S^{-1}(\xi + \lambda \sinh(\delta(S(u) + nc))), & n \neq 0 \\ l + (u_b - l_b) S^{-1}\left(B + \left(\frac{H-L}{2c}\right) S(u)\right), & n = 0 \end{cases} \tag{3}$$

where

$$\begin{gathered} S(u) = \Phi^{-1}(u); \quad c = S(1-\alpha); \\ L = S\left(\frac{q_1 - l_b}{u_b - l_b}\right); \quad B = S\left(\frac{q_2 - l_b}{u_b - l_b}\right); \\ H = S\left(\frac{q_3 - l_b}{u_b - l_b}\right); \quad n = \operatorname{sgn}(L + H - 2B) \\ \xi = \begin{cases} L, & n = 1, \\ B, & n = 0, \\ H, & n = -1, \end{cases} \\ \delta = \frac{1}{c} \cosh^{-1}\left(\frac{H - L}{2\min(B - L, H - B)}\right) \\ \lambda = \frac{H - L}{\sinh(2\delta c)} \end{gathered} \tag{4}$$

The left panel in Figure 2 showcases the J-QPD-B quantile function, which is parameterized using 0.25-SPT and 0.01-SPT assessments of the proportion of fruit infested with *Citripestis sagittiferella*, as elicited by EFSA et al. (2023). The dashed line represents the Beta distribution fitted by the authors. The J-QPD-B, being parameterized by an SPT, effectively captures three of the five parameterizing quantiles, while the Beta distribution only provides an approximation. Furthermore, determining the parameters of the Beta distribution necessitates an optimization step.

The J-QPD-S distribution is a semi-bounded variant of the distribution that employs exponentiated hyperbolic arcsine transformations of the Johnson's SU quantile function (Hadlock and Bickel, 2017)

$$Q_S(u|q_1, q_2, q_3, \alpha) = \begin{cases} l_b + \theta \exp\left(\lambda \sinh\left(\sinh^{-1}(\delta S(u)) + \sinh^{-1}(nc\delta)\right)\right), & n \neq 0 \\ l_b + \theta \exp\left(\lambda \delta S(u)\right), & n = 0 \end{cases} \tag{5}$$

where

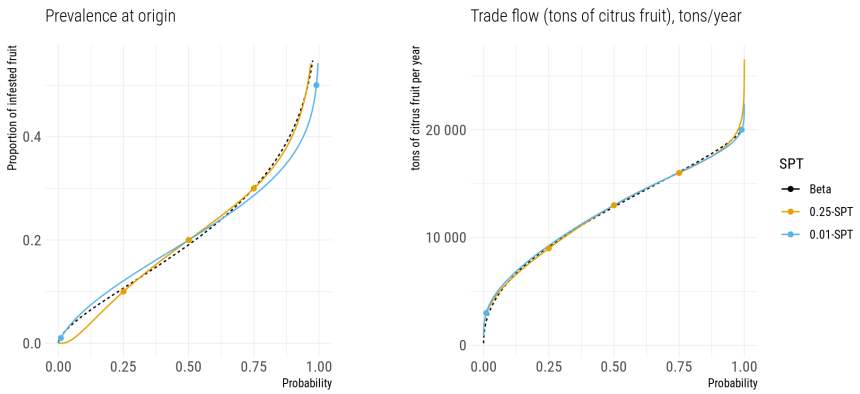


Figure 2: Fitted J-QPD-B (left) and J-QPD-S (right) distribution for prevalence at origin and total trade flow, respectively

$$
\begin{aligned}
S(u) = \Phi^{-1}(u); \quad c = S(1-\alpha); \\
L = \ln(q_1 - l_b); \quad B = \ln(q_2 - l_b); \\
H = \ln(q_3 - l_b); \quad n = \operatorname{sgn}(L + H - 2B) \\
\theta = \begin{cases} q_1 - l_b, & n = 1, \\ q_2 - l_b, & n = 0, \\ q_3 - l_b, & n = -1, \end{cases} \\
\delta = \frac{1}{c} \sinh\left(\cosh^{-1}\left(\frac{H - L}{2\min(B - L, H - B)}\right)\right) \\
\lambda = \frac{1}{\delta c} \min(H - B, B - L)
\end{aligned}
\tag{6}
$$

When $n = \operatorname{sgn}(L + H - 2B)$ evaluates to zero, he resulting distribution is a lognormal distribution with parameters $\mu = \ln(\theta) = \ln(q_2 - l_b)$ and $\sigma = \lambda\delta = (H - B)/c$. This distribution has support on the interval $[l_b, \infty]$.

The right panel in Figure 2 depicts the J-QPD-S quantile function, which is parameterized using 0.25-SPT and 0.01-SPT assessments of the total trade flow for citrus fruit imported by the EU from Indonesia, Malaysia, Thailand, and Vietnam in tons/year (EFSA et al., 2023).

3.3. Generalisations of QPDs

3.3.1. Generalized Johnson Quantile-Parameterized Distribution

Hadlock and Bickel (2019) introduced the *generalized* version of the Johnson Quantile-Parameterized distribution system, denoted as G-QPD, by replacing the Normal distribution in the core of the Johnson SU quantile function with the quantile functions of the logistic and Cauchy distributions.

The generalized quantile function (QF) shares similarities with the probit-based distribution described earlier, with $S(u)$ defined as the quantile function of either the logistic or Cauchy distribution.

The standard quantile function and distribution function of the logistic distribution are given by:

$$S(u) = \ln\left(\frac{u}{1-u}\right); \quad S^{-1}(y) = [\exp(-y)+1]^{-1} \tag{7}$$

The standard quantile function and distribution function of the Cauchy distribution are given by:

$$S(u) = \tan\left[\pi\left(u - \frac{1}{2}\right)\right]; \quad S^{-1}(y) = \frac{1}{\pi}\arctan(y) + \frac{1}{2} \tag{8}$$

Hadlock and Bickel (2019) show that the *kernel* quantile function $S(u)$ can be any standardized ($S(0.5) = 0$), symmetrical ($s(u) = s(1-u)$), and unbounded ($S(u) \in (-\infty;\infty)$) quantile function with a smooth quantile density $dS(u)/du = s(u)$. The authors further showed that if $S(u)$ and $S^{-1}(y)$ are expressible in closed-form, the quantile function and distribution function of G-QPD will also be closed-form.

For the *logistic* kernel, the G-QPD-S represents the generalized log-logistic distribution, characterized by two shape parameters, λ and δ. For the Cauchy kernel, the G-QPD-S corresponds to the shifted log-Cauchy distribution (Hadlock and Bickel, 2019).

3.3.2. Generalized Myerson distributions

Drawing inspiration from Hadlock and Bickel (2019), we can extend the Myerson distribution by substituting the Normal kernel quantile function $S(u) = \Phi^{-1}(u)$ with an alternative symmetrical quantile function based on the depth u. Below, we present the proposed kernels and the resulting distributions:

Logit-Myerson distribution - employs the standard logistic quantile function:

$$S(u) = \ln\left(\frac{u}{1-u}\right) \tag{9}$$

There are several reasons why one might prefer the logit function over the probit function (Berkson, 1951). We discovered that the Logit-Myerson distribution exhibits greater numerical stability due to its simple closed-form quantile function, which does not rely on numerical approximation during sampling. This distribution displays slightly heavier tails compared to the standard probit-based Myerson distribution (Figure 3).

Sech-Myerson distribution - employs the hyperbolic secant quantile function:

$$S(u) = \ln\left[\tan\left(\frac{\pi}{2}u\right)\right] \tag{10}$$

The Sech-Myerson distribution possesses thicker tails than the Logit-Myerson distribution for the same parameterizing SPT $\{-5, 4, 16, 0.25\}$ (Figure 3). In Section 3.6, we conduct a comparative analysis of different variations of the Generalized Myerson distribution alongside their parametric counterparts and other quantile distributions.

Theoretically, there is an infinite range of quantile function (QF) kernels that can be utilized to generate new variations of the Generalized Myerson distribution. These candidate kernel distributions can even include shape parameters, as long as the resulting $S(u)$ remains standardized, symmetrical, and unbounded, as specified above. For instance, it is possible to incorporate the basic QF of the Tukey Lambda distribution $S(u|\lambda) = u^\lambda - (1-u)^\lambda$ for a fixed $\lambda \neq 0$, or the Cauchy distribution $S(u) = \tan[\pi(u-0.5)]$, as employed by Hadlock and Bickel (2019). However, it is important to note that not all standard quantile functions are created equal. To illustrate the issue of unreliable kernels, let us consider Myerson distributions based on the Cauchy and Tukey Lambda quantile functions (for $\lambda = -0.5$), as depicted in Figure 4.

While all right-skewed Generalized Myerson distributions are bounded on the left at $\lim_{u\to 0} Q(u|\theta) = q_2 - \rho\frac{1}{\beta-1}$ regardless of the kernel used, the quantile density at the left limit $\lim_{u\to 0}[q(u|\theta)]^{-1}$ is not independent of the kernel. Although we can assume that $q(0) = \infty$, the lower tail of the density quantile function $[q(u)]^{-1}$ may exhibit a curling effect for certain kernels, resulting in an increase in density for lower values of u. This effect is caused by the non-monotonic behavior of the quantile convexity function $c(u) = dq(u)/du$. This can be easily verified by taking the second derivative of $\beta^{S(u)}$ for $\beta > 0$. While such kernels are mathematically valid and yield a non-decreasing Generalized Myerson QF, we believe that they

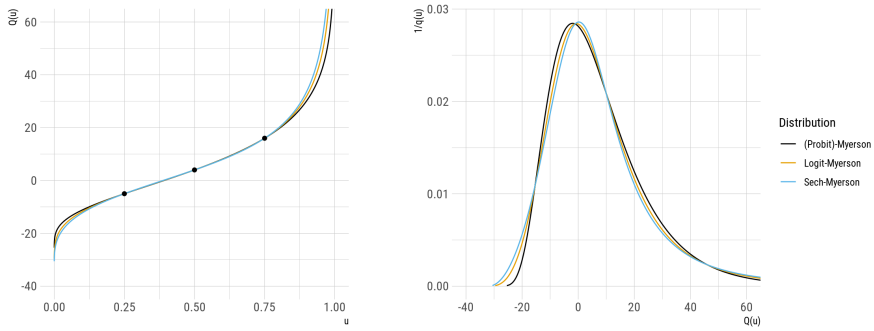


Figure 3: Quantile function and quantile density of Generalized Myerson Distributions

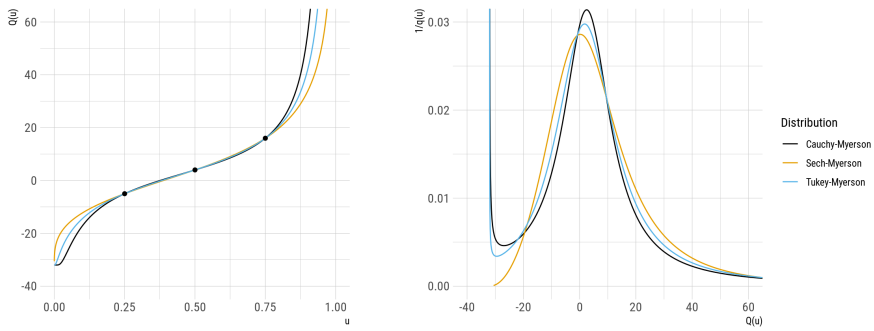


Figure 4: Quantile function and quantile density of Generalized Myerson Distributions with unreliable kernels

may be less useful due to the counter-intuitive concentration of density in the bounded tail. Consequently, we do not recommend using Cauchy or Tukey Lambda kernels in practical applications.

3.4. Simple Q-Normal, Metalog and quantile mixtures

An alternative system of quantile-parameterized distributions was proposed by Keelin and Powley (2011) and Powley (2013). This approach relies on the finite Taylor expansion of parameters in the standardized quantile functions. Within this framework, two distributions were introduced: the Simple Q-Normal distribution and the Metalog distribution.

The Simple Q-Normal (SQN) distribution was developed by expanding the parameters in the normal quantile function. Keelin et al. (2011) used this method to express the parameters of the normal quantile function $Q(u|\mu,\sigma) = \mu + \sigma z(u)$ as linear functions of the depth u. Specifically, $\mu(u) = a_1 + a_4 u$ and $\sigma(u) = a_2 + a_3 u$, where $z(u) = \Phi^{-1}(u)$ denotes the standard normal quantile function. Therefore, the quantile function of the SQN distribution can be expressed as follows:

$$Q(u) = a_1 + a_2 z(u) + a_3 u z(u) + a_4 u \tag{11}$$

where $z(u) = \Phi^{-1}(u)$, and $a = \{a_1, a_2, a_3, a_4\}$ represents a vector of parameters. It should be noted that the SQN quantile function is the product of the normal and uniform quantile functions.

Consider a quantile-probability tuple of size 4, denoted as $\{\mathbf{p}, \mathbf{q}\}_4$, which consists of an ordered vector of cumulative probabilities $\mathbf{p} = \{p_1, p_2, p_3, p_4\}$ and an ordered vector of corresponding quantiles $\mathbf{q} = \{q_1, q_2, q_3, q_4\}$. Substituting these vectors into the SQN quantile function for u and $Q(u)$, respectively, we obtain the following matrix equation:

$$\mathbf{q} = \mathbb{P} a \tag{12}$$

where

$$\mathbb{P} = \begin{bmatrix} 1 & z(p_1) & p_1 z(p_1) & p_1 \\ 1 & z(p_2) & p_2 z(p_2) & p_2 \\ 1 & z(p_3) & p_3 z(p_3) & p_3 \\ 1 & z(p_4) & p_4 z(p_4) & p_4 \end{bmatrix} \tag{13}$$

and $a = \{a_1, a_2, a_3, a_4\}$ represents the parameter vector of the SQN distribution.

The parameter vector a can be obtained by solving the matrix Equation (12), given the 4-element quantile-probability tuple $\{\mathbf{p}, \mathbf{q}\}_4$ (Keelin and Powley, 2011; Perepolkin et al., 2021a).

The same approach was later employed by Keelin (2016) in creating the metalog (meta-logistic) distribution. Starting with the quantile function of the logistic distribution $Q(u|\mu, s) = \mu + s\text{logit}(u)$, where μ corresponds to the mean and s is proportional to the standard deviation $\sigma = s\pi/\sqrt{3}$, Keelin (2016) expanded the parameters μ and s using a finite Taylor series centered at 0.5. Specifically, $\mu(u) = a_1 + a_4(u - 0.5) + a_5(u - 0.5)^2 + \ldots$ and $s(u) = a_2 + a_3(u - 0.5) + a_6(u - 0.5)^2 + \ldots$, where $a_i,\ i = \{1, 2, \ldots, n\}$ are real constants.

Therefore, the metalog quantile function is:

$$Q(u) = a_1 + a_2\text{logit}(u) + a_3(u - 0.5)\text{logit}(u) + a_4(u - 0.5) + a_5(u - 0.5)^2 \cdots, \tag{14}$$

Given a QPT of size m denoted by $\{\mathbf{p}, \mathbf{q}\}_m$, where $\mathbf{p}$ and $\mathbf{q}$ are ordered vectors of cumulative probabilities and corresponding quantiles, respectively, the vector of coefficients $\mathbf{a} = a_1, \ldots, a_m$ can be determined by solving the matrix equation $\mathbf{q} = \mathbb{P}\mathbf{a}$, where $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{a}$ are column vectors, and $\mathbb{P}$ is an $m \times n$ matrix:

$$\mathbb{P} = \begin{bmatrix} 1 & \text{logit}(p_1) & (p_1 - 0.5)\text{logit}(p_1) & (p_1 - 0.5) & \cdots \\ 1 & \text{logit}(p_2) & (p_2 - 0.5)\text{logit}(p_2) & (p_2 - 0.5) & \cdots \\ & & \vdots & & \\ 1 & \text{logit}(p_m) & (p_m - 0.5)\text{logit}(p_m) & (p_m - 0.5) & \cdots \end{bmatrix} \tag{15}$$

The vector of coefficients $\mathbf{a}$ can be determined as $\mathbf{a} = [\mathbb{P}^T\mathbb{P}]^{-1}\mathbb{P}^T\mathbf{q}$. If $\mathbb{P}$ is a square matrix, meaning the number of terms n is equal to the size of the parameterizing QPT m, the equation can be further simplified to $\mathbf{a} = \mathbb{P}^{-1}\mathbf{q}$. Metalog is said to be *approximated* when the number of quantile-probability pairs used for parameterization exceeds the number of terms in the metalog QF (Keelin, 2016; Perepolkin et al., 2021a).

The SQN and Metalog distributions are families of extended distributions that, in theory, can have an arbitrary number of terms. Keelin (2016) demonstrated the flexibility of the metalog distribution and its ability to approximate arbitrarily complex probability density functions with high precision, given enough terms in the metalog specification. In practice, 10-15 terms are sufficient to approximate the distributional shapes of virtually any complexity (Keelin and Howard, 2021).

In Keelin (2016), semi-bounded and bounded versions of metalog distributions are offered, which use the *log* and *logit* Q-transformations.

However, not all combinations of parameters $\mathbf{a}$ in metalog and SQN distributions result in a feasible (non-decreasing) quantile function. To ensure that the quantile function is feasible on $u \in [0, 1]$, the transformations used to construct it should follow Gilchrist's QF transformation rules (Table 1). The SQN and metalog QFs violate the *multiplication rule*. The product of the base quantile function (normal for SQN and logistic for metalog) and the uniform quantile function $Q(u) = u$ is not strictly positive. The problem is more severe in the metalog distribution, where the uniform QF is centered at 0.5 and, therefore, spans 0. The shifted uniform distribution is further Q-transformed using the power operator $(u - 0.5)^k$, which preserves the negative values for the odd powers of k. Therefore, quantile functions of metalog and SQN distributions are not maximally feasible for all values of $\mathbf{a}$.

Recently Peng et al. (2023) proposed a novel framework for extended quantile-parameterized distributions based on quantile mixtures (not to be confused with CDF/PDF mixtures, Gilchrist (2000), p. 107). They introduced a formulation where a QPD QF is expressed as a linear combination of I standardized quantile functions, following Gilchrist's *linear combination rule* (Table 1):

$$G(u|\theta) = \sum_{i=0}^{I} \theta_i Q_i(u) \tag{16}$$

Here, $Q_i(u)$ represent basis quantile functions for random variable Y with $Q_0(u) = 1$, and $\boldsymbol{\theta} = \{\theta_0, \theta_1, \ldots, \theta_I\}$ is a non-negative parameter vector that determines the contribution of each QF component in the quantile mixture. To compute the coefficients $\boldsymbol{\theta}$, the system of equations is solved

$$\mathbf{q} = \mathbb{Q}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{17}$$

where $\mathbf{q} = \{q_1, q_2, \ldots, q_j\}$ is an ordered vector of J parameterizing quantiles, corresponding to an ordered vector of cumulative probabilities $\mathbf{p} = \{p_1, p_2, \ldots, p_j\}$, $\boldsymbol{\theta}$ is a non-negative vector of $I + 1$ parameters, $\boldsymbol{\epsilon}$ is a J-size vector of errors to be minimized, and $\mathbb{Q}$ is a $J \times (I + 1)$ matrix of regression factors

$$\mathbb{Q} = \begin{bmatrix} 1 & Q_1(p_1) & Q_2(p_1) & \cdots & Q_I(p_1) \\ 1 & Q_1(p_2) & Q_2(p_2) & \cdots & Q_I(p_2) \\ & \vdots & \vdots & \ddots & \\ 1 & Q_1(p_J) & Q_2(p_J) & \cdots & Q_I(p_J) \end{bmatrix} \tag{18}$$

By ensuring non-negativity of weights ($\theta_i \geq 0$), the solution guarantees a proper non-decreasing quantile function. For estimating the values of the vector $\boldsymbol{\theta} \in \Theta$, the authors suggest to use constrained weighted least squares regression with optional regularization. The authors demonstrated that the estimator $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left(\frac{1}{J} \sum_{j=1}^{J} w_j \mathcal{E}_q(y_j - Q_j\boldsymbol{\theta}) \right)^{\frac{1}{q}}$, $\mathcal{E}_q(x) = |x|^q$, $w_j > 0$, is asymptotically a q-Wasserstein distance estimator, which converges in distribution to a Normal distribution. The Peng et al. (2023) paper includes the application of the quantile mixture model using a large number of asymmetric t-distributions, and a quantile mixture of Generalized Beta II distributions.

3.5. Other distributions

3.5.1. Triangular and Two-Sided Power distributions

Several other distributions with at least some parameters mapped to quantiles were proposed, including the reparameterization of the Generalized Lambda Distribution by Chalabi et al. (2012) and the quantile-parameterized triangular (two-sided power) distribution by Kotz and Van Dorp (2004).

Kotz and Van Dorp (2004) describe the quantile-parameterized version of the triangular distribution (Johnson, 1997). This bounded distribution is widely used in the finance and insurance industry and is popularized by the @Risk software package, developed by Palisade (Palisade Corporation, 2009). The triangular distribution is parameterized by the two quantiles q_a and q_b, and the mode m, subject to the constraint that $a \leq q_a \leq m \leq q_b \leq b$, where a and b represent the lower and upper bounds, respectively. The standard quantile function for the triangular distribution is expressed in terms of the bounds a, b, and the mode m.

$$Q(u|a,m,b) = \begin{cases} a + \sqrt{u(m-a)(b-a)}, & \text{for } 0 \leq u \leq \frac{m-a}{b-a} \\ b - \sqrt{(1-u)(b-m)(b-a)}, & \text{for } \frac{m-a}{b-a} \leq u \leq 1 \end{cases} \tag{19}$$

Kotz and Van Dorp (2004) show that given the two parameterizing quantile-probability pairs q_a, p_a and q_b, p_b and the mode value m, there exists a unique value of depth $p_a < p < p_b$ corresponding to the root of the function

$$g(p) = \frac{(m-q_a)(1-\sqrt{\frac{1-p_b}{1-p}})}{(q_b-m)(1-\sqrt{\frac{p_a}{p}}) + (m-q_a)(1-\sqrt{\frac{1-p_b}{1-p}})} - p \tag{20}$$

The root value $p \in (p_a, p_b)$ of the function $g(p)$ can be found using any of the bracketing root-finding algorithms (Perepolkin et al., 2021b). It can then be substituted into the following expressions to find the lower a and upper b limit parameters of the triangular distribution:

$$\begin{aligned} a(p) &\equiv \frac{q_a - m\sqrt{\frac{p_a}{p}}}{1-\sqrt{\frac{p_a}{p}}}, \quad a(p) < q_a \\ b(p) &\equiv \frac{q_b - m\sqrt{\frac{1-p_b}{1-p}}}{1-\sqrt{\frac{1-p_b}{1-p}}}, \quad b(p) > q_b \end{aligned} \tag{21}$$

Kotz and Van Dorp (2004) provide an algorithm for fitting a four-parameter generalization of the triangular distribution called the Two-Sided Power Distribution (TSP), using three quantile-probability pairs and a mode value. For more information on fitting the Quantile-Parameterized TSP Distribution by quantiles, refer to Section 4.3.3 of Kotz and Van Dorp (2004).

3.5.2. Generalized Lambda Distribution

Chalabi et al. (2012) (CSW) proposed an asymmetry-steepness reparameterization of the FKML Generalized Lambda Distribution (GLD) (Freimer et al., 1988) with four parameters. This reparameterization involves mapping the location to the median and the scale to the interquartile range (IQR), which corresponds to the first and second robust moments (Kim and White, 2004; Moors, 1988).

The reparameterized Generalized Lambda Distribution by Chalabi, Scott and Würtz (CSW GLD) has a quantile function given by

$$Q(u|\tilde{\mu},\tilde{\sigma},\chi,\xi) = \tilde{\mu} + \tilde{\sigma}\frac{S(u|\chi,\xi) - S(\frac{1}{2}|\chi,\xi)}{S(\frac{3}{4}|\chi,\xi) - S(\frac{1}{4}|\chi,\xi)} \tag{22}$$

where $\tilde{\mu}, \tilde{\sigma}, \chi, \xi$ represent the location, scale, asymmetry, and steepness parameters, respectively. The specific form of the basic function $S(u)$ depends on the values of the parameters χ and ξ

$$S(u|\chi,\xi) = \begin{cases} \ln(u) - \ln(1-u), & \text{if } \chi = 0, \xi = 0.5 \\ \ln(u) - \dfrac{1}{2\alpha}\left[(1-u)^{2\alpha} - 1\right], & \text{if } \chi \neq 0, \xi = \dfrac{1}{2}(1+\chi) \\ \dfrac{1}{2\beta}\left[u^{2\beta} - 1\right] - \ln(1-u), & \text{if } \chi \neq 0, \xi = \dfrac{1}{2}(1-\chi) \\ \dfrac{1}{\alpha+\beta}\left[u^{\alpha+\beta} - 1\right] - \dfrac{1}{\alpha-\beta}\left[(1-u)^{\alpha-\beta} - 1\right], & \text{otherwise} \end{cases} \tag{23}$$

where $\alpha = 0.5\frac{0.5-\xi}{\sqrt{\xi(1-\xi)}}$ and $\beta = 0.5\frac{\chi}{\sqrt{1-\chi^2}}$. The bounds of the distribution are given by

$$\begin{aligned} S(0|\chi,\xi) &= \begin{cases} -\dfrac{1}{\alpha+\beta}, & \text{if } \xi < \dfrac{1}{2}(1+\chi) \\ -\infty, & \text{otherwise} \end{cases} \\ S(1|\chi,\xi) &= \begin{cases} \dfrac{1}{\alpha-\beta}, & \text{if } \xi < \dfrac{1}{2}(1-\chi) \\ \infty, & \text{otherwise} \end{cases} \end{aligned} \tag{24}$$

The CSW GLD can have unbounded, bounded, and semi-bounded support, accommodating a wide range of shapes, including unimodal, monotone, U-shaped, and S-shaped densities (Chalabi et al., 2012). Although the CSW GLD is not strictly parameterized by quantiles, the mapping of the location and scale parameters to the median and IQR makes it a suitable candidate for expert-informed distribution specification.

Several specialized methods have been developed for fitting the GLD to samples (Karian and Dudewicz, 2003). The parameterization of the CSW GLD simplifies the fitting process because two of the four parameters can be directly calculated from the sample: the location parameter is equal to the sample median, and the scale parameter is equal to the interquartile range. The remaining parameters can be estimated using various methods, including robust moment matching, quantile matching, trimmed L-moments, distributional least squares/absolutes, as well as maximum likelihood estimation (Chalabi et al., 2012; Gilchrist, 2000). The range of feasible values for the steepness and asymmetry parameters can be further reduced with the shape conditions specified in Section 3.5 of Chalabi et al. (2012).

Recently, Dedduwakumara et al. (2021) proposed a new method of matching the shape of the GLD distribution to data using the probability density quantile (pdQ) function (Staudte, 2017). For the quantile function $Q(v),\ v \in [0,1]$ and the corresponding density quantile function $f(Q(v)) = [q(v)]^{-1}$, the pdQ is defined as

$$f^*(v) = \frac{f(Q(v))}{E\left[f(Q(v))\right]} \tag{25}$$

The probability density quantile function is defined on the unit square and is independent of the location and scale parameters.

Since integrating the GLD density quantile function is difficult, Staudte (2017), in Section 2.2, proposed using the kernel density method to estimate the empirical QDF and, thus, an empirical pdQ for samples from continuous distributions. Fitting the CSW GLD to a sample can be reduced to finding the asymmetry and steepness parameters that minimize

$$\underset{\chi,\xi}{\operatorname{argmin}} \int_0^1 \left[f^*(v,\chi,\xi) - f_e^*(v)\right]^2 du \tag{26}$$

where $f^*(v,\chi,\xi)$ is the pdQ of the CSW GLD, and $f_e^*(v)$ is the empirical pdQ of the sample. Dedduwakumara et al. (2021) suggest approximating the integral by a discrete set of depths v, replacing the integral with a sum.

3.6. Choosing quantile-parameterized distribution

A common approach to assess the properties of probability distributions is through central moments, denoted by $\mu_k = \mathbb{E}[(Y - \mu)^k]$, where μ represents the expected value of Y. Karl Pearson introduced a classification system for distributions using moment ratios associated with skewness and kurtosis (Fiori and Zenga, 2009):

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} \tag{27}$$

While computing moments using the quantile function is straightforward (the n-th raw moment is $\mu_k = \int_0^1 Q(u)^k du$), it may not be possible to calculate higher-order moments for certain distributions.

Alternatively, robust alternatives to moments can be utilized, such as the sample median μ_r, the interquartile range σ_r, the quartile-based robust coefficient of skewness s_r (Kim and White, 2004), also known as Bowley's skewness (Bowley, 1920) or Galton's skewness (Gilchrist, 2000), and the octile-based robust coefficient of kurtosis κ_r, also known as Moors' kurtosis (Moors, 1988).

$$\begin{aligned} \mu_r &= Q(1/2) \\ \sigma_r &= Q(3/4) - Q(1/4) \\ s_r &= \frac{Q(3/4) + Q(1/4) - 2Q(1/2)}{\sigma_r} \\ \kappa_r &= \frac{Q(7/8) - Q(5/8) + Q(3/8) - Q(1/8)}{\sigma_r} \end{aligned} \tag{28}$$

Kim and White (2004) and Arachchige et al. (2022) have proposed to standardize robust moments to facilitate their comparison with the corresponding robust moments of the standard normal distribution. Groeneveld (1998) and Jones et al. (2011) have introduced generalizations of robust moments to other quantiles.

Unlike moments, quantiles always exist, and since QPDs are parameterized by quantile-probability pairs, quantile-based robust moments can sometimes be directly computed from the parameters. For instance, if the basic quantile function $S(u)$ in $Q(u) = \mu + \sigma S(u)$ is standardized (such that $S(0.5) = 0$), where μ and σ are the location and scale parameters of $Q(u)$ respectively, then $\mu_r = \mu$. Moreover, σ_r is always independent of location, and s_r and κ_r are independent of both location and scale.

Figures 5, 6, and 7 resemble the Cullen and Frey (Cullen et al., 1999) plots, also known as Pearson plots. However, instead of using central moments, they employ quartile/octile-based robust metrics of skewness s_r and kurtosis κ_r to compare the quantile-parameterized distributions to some of their parametric counterparts. In Figure 5, Metalog3 and Metalog4 refer to 3- and 4-term metalog distributions, respectively, which are also log- or logit-transformed for the semibounded case (Figure 6) and bounded case (Figure 7) cases. GLDcsw refers to Chalabi et al. (2012) parameterization of GLD.

4. Multivariate quantile-parameterized distributions

Quantile-parameterized distributions can serve as marginal distributions in multivariate models, where the dependency structure is captured by a standard (parametric) multivariate distribution, a copula, or described by bivariate quantiles. However, the marginal distributions alone are insufficient to determine the corresponding bivariate distribution, resulting in an infinite number of bivariate distributions with the same margins (Gumbel, 1960, 1961). In this section, we describe several methods for extending the distributions parameterized by the quantile-probability pairs to become Multivariate Quantile-Parameterized Distributions (MQPDs).

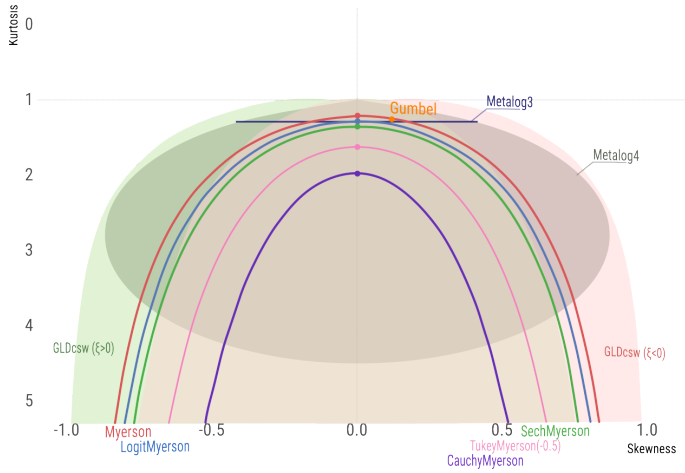


Figure 5: Robust skewness vs robust kurtosis for some unbounded distributions

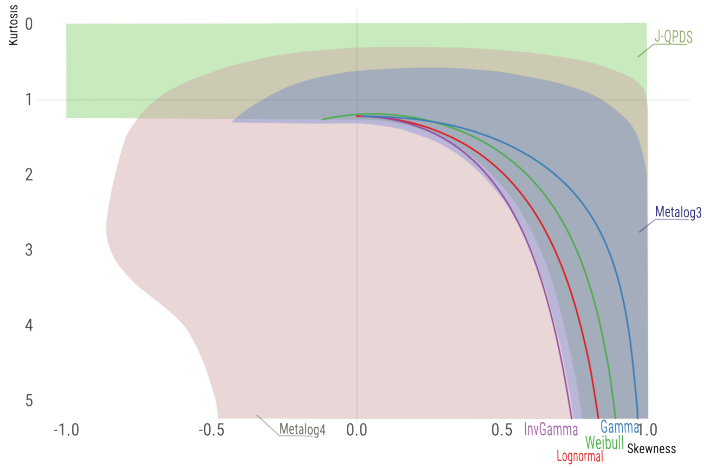


Figure 6: Robust skewness vs robust kurtosis for some left-bounded distributions

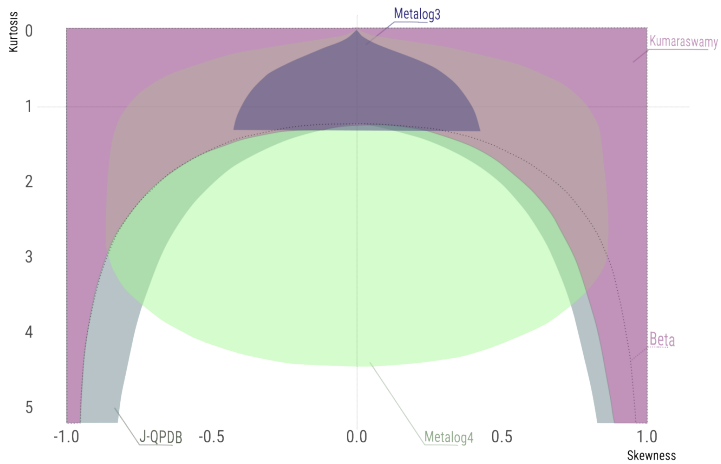


Figure 7: Robust skewness vs robust kurtosis for some bounded distributions

4.1. MQPDs based on standard multivariate distributions

In the simplest case, multivariate Quantile-Parameterized Distributions (MQPDs) can be created by using the multivariate normal distribution, following the approach of Hoff (2007). The Myerson, J-QPD, and SQN quantile functions are Q-transformations of the probit $Q(z(u)|\theta)$, where $z(u) = \Phi^{-1}(u)$ represents the standard normal quantile function. The multivariate versions of these distributions can be viewed as the Q-transformations of the multivariate normal distribution. To extend these QPDs to J dimensions using the multivariate normal distribution, we employ the method outlined in Drovandi and Pettitt (2011).

The i-th component of a single observation y_i can be described by the quantile function:

$$y_i = Q(z(u_i)|\theta_i), \text{ for } i = 1, \ldots, J \tag{29}$$

where θ_i represents the set of parameters for component i (e.g., $\{q_1, q_2, q_3, \alpha\}_i$) for Myerson or J-QPD distributions). The vector $(z(u_1), \ldots, z(u_j))^T \sim N(0, \Sigma)$, where Σ denotes the covariance matrix.

For invertible distributions, the inverse quantile function is the cumulative distribution function (CDF) $Q^{-1}(y_i|\theta) = F(y_i|\theta)$, otherwise, the inverse can be computed numerically as $\widehat{F}(y_i|\theta) = \widehat{Q^{-1}}(y_i|\theta)$ (Perepolkin et al., 2021b).

Drovandi and Pettitt (2011) show that the joint density of a single (multivariate) observation $(y_i, \ldots, y_J)$ can be expressed as:

$$f(y_1, \ldots, y_J|\theta) = \varphi(z(Q^{-1}(y_1|\theta_1)), \ldots, z(Q^{-1}(y_J|\theta_J)); \Sigma) \prod_{i=1}^{J} \frac{dQ^{-1}(y_i|\theta_i)}{dy_i} \tag{30}$$

where $z(Q^{-1}(y_i|\theta_i)) = z_i$, $\varphi(z_1, \ldots, z_J; \Sigma)$ represents the multivariate normal density with a mean of zero and a covariance matrix of Σ, and $\frac{dQ^{-1}(y_i)}{dy_i} = f(y_i)$ is the probability density function (PDF) of the QPD (refer to Appendix A).

For distributions without a PDF, the same joint density can be expressed as a joint density quantile function

$$[q(u_1,\ldots,u_j)]^{-1} = \varphi(z(u_1),\ldots,z(u_J);\Sigma)\prod_{i=1}^{J}[q(u_i|\theta_i)]^{-1} \tag{31}$$

since $Q^{-1}(y_i|\theta_i) = u_i$ and $f(y_i|\theta_i) = [q(u_i|\theta_i)]^{-1}$ (Gilchrist, 2000).

It's worth noting that this method of creating multivariate distributions does not require every component to follow the same distributional form. As illustrated earlier, it is entirely possible to combine several different QPDs using the multivariate Gaussian distribution (Drovandi and Pettitt, 2011).

To use the MQPD for the prior, both the density of the multivariate normal and the marginal densities need to be explicitly added to the log-likelihood. This is possible when the marginal QPDs used to define the multivariate prior are invertible, such as Myerson and J-QPD, as both the CDF ($Q^{-1}(y_i|\theta_i)$) and PDF ($dQ^{-1}(y_i|\theta_i)/dy_i$) are required.

When a quantile-based prior specification is used, only the multivariate normal log-density needs to be added because the Jacobian for the marginal QF transformation is reciprocal to the DQF of the prior (Perepolkin et al., 2021b).

The same approach of joining the marginal QPDs can be applied by using the base quantile functions of other distributions. For instance, the Logit-Myerson distribution discussed above is based on the logistic quantile function. Two Logit-Myerson distributions can be connected using the bivariate logistic distribution. Gumbel (1961) proposed three different formulations for the bivariate logistic distribution. The Type II distribution from the Morgenstern Family (Sajeevkumar and Irshad, 2014; Basikhasteh et al., 2021) has the following joint distribution and density functions:

$$\begin{aligned} F(y_1,y_2|\beta) =& F_1(y_1)F_2(y_2)[1+\beta(1-F_1(y_1))(1-F_2(y_2))] \\ f(y_1,y_2|\beta) =& f_1(y_1)f_2(y_2)[1+\beta(1-2F_1(y_1))(1-2F_2(y_2))] \end{aligned} \tag{32}$$

where $F_i(y_i)$ and $f_i(y_i)$ for $i \in \{1,2\}$ refer to the univariate logistic distribution and density funcitons, respectively and $-1 \leq \beta \leq 1$. Since $y_i = Q_i(u_i)$ we can express the bivariate density in the quantile form

$$\begin{aligned} f(Q(u_1),Q(u_2)|\beta) =& f_1(Q(u_1))f_2(Q(u_2))[1+\beta(1-2F_1(Q_1(u_1)))(1-2F_2(Q_2(u_2)))] \\ [q(u_1,u_2|\beta)]^{-1} =& [q_1(u_1)]^{-1}[q_2(u_2)]^{-1}\left[1+\beta(1-2u_1)(1-2u_2)\right] \end{aligned} \tag{33}$$

For logistic distribution $Q(u) = \ln(u) - \ln(1-u)$ and $[q(u)]^{-1} = u(1-u)$. Therefore, the bivariate logistic density quantile function can be expressed as

$$[q_L(u_1,u_2|\beta)]^{-1} = u_1(1-u_1)u_2(1-u_2)\left[1+\beta(1-2u_1)(1-2u_2)\right] \tag{34}$$

If we combine the QPD marginals, the result is the joint quantile-based density for the bivariate logistic-based QPD, where the dependency is captured by the bivariate logistic distribution with the coupling parameter β, and the margins are QPDs. The joint density quantile function is given by:

$$[q_{MQPD}(u_1,u_2|\theta_1,\theta_2,\beta)]^{-1} = u_1(1-u_1)u_2(1-u_2)\left[1+\beta(1-2u_1)(1-2u_2)\right][q_1(u_1|\theta_1)]^{-1}[q_2(u_2|\theta_2)]^{-1} \tag{35}$$

Here, $[q_i(u_i|\theta_i)]^{-1}$, for $i = 1,2$, represents the marginal QPD density quantile functions, such as the density quantile function (DQF) of the Logit-Myerson distribution (see Appendix A).

Figure 8 presents the Bivariate Logit-Myerson Distribution, parameterized by $\Theta = \{\theta_1,\theta_2,\rho\}$, where the marginal Myerson distributions are given by $y_{ij} = Q_j(z(u_{ij}),\theta_j)$ for $j = 1,2$, with parameter vectors $\theta_1 = \{3,7,10;0.25\}$, $\theta_2 = \{1,10,20;0.1\}$, and the dependence parameter $\beta = 0.6$.

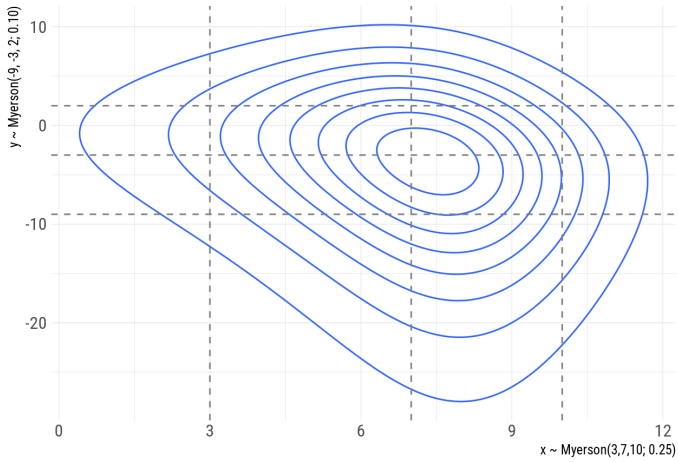


Figure 8: Density of Generalized Myerson distributions joined by Type II bivariate logistic distribution

4.2. Copula-based MQPDs

The approach we have used so far is similar to constructing the joint distribution using the Gaussian copula (Hoff, 2007). Copulas provide a more general approach to modeling joint distributions, aiming to separate the influence of bivariate dependence from the effects of marginal distributions (Kurowicka and Cooke, 2006). The literature describes a wide range of copulas (Genest and Favre, 2007; Smith, 2013; Kurowicka and Joe, 2011), and new copulas can be created using generator functions (Durrleman et al., 2000). When a copula is used to connect QPDs, the joint density is calculated as follows:

$$f_{MQPD}(y_1, y_2|\theta_1, \theta_2, \Xi) = c(F(y_1|\theta_1), F(y_2|\theta_2)|\Xi) f_1\left(y_1|\theta_1\right) f_2\left(y_2|\theta_2\right) \tag{36}$$

where c represents the copula density function with parameter Ξ, and $F(y_i|\theta_i)$ and $f_i(y_i|\theta_i)$ are the CDF and PDF of the marginal quantile-parameterized distributions, respectively.

The same density can be expressed in quantile-based form (Perepolkin et al., 2021b):

$$[q_{MQPD}(u_1, u_2|\theta, \Xi)]^{-1} = c\left(u_1, u_2|\Xi\right)[q_1(u_1|\theta_1)]^{-1}[q_2(u_2|\theta)]^{-1} \tag{37}$$

where c is the copula density function with parameter Ξ, and $[q_i(u_i|\theta_i)]^{-1}$, for $i = 1, 2$, are the marginal DQFs of QPDs. Figure 9 presents 10,000 samples from the bivariate Myerson distribution joined by the Joe copula with $\theta = 3$.

Elicitation of multivariate distributions may require a specialized approach (Elfadaly and Garthwaite, 2017; Wilson et al., 2021). For examples of expert-specified multivariate distributions encoded with copulas, we refer to Wilson (2018), Holzhauer et al. (2022), Sharma and Das (2018), and Aas et al. (2009). When fitting copulas to empirical observations, the "blanket" goodness of fit measure (Wang and Wells, 2000) based on Kendall's transform (Genest et al., 2006, 2009) can be used.

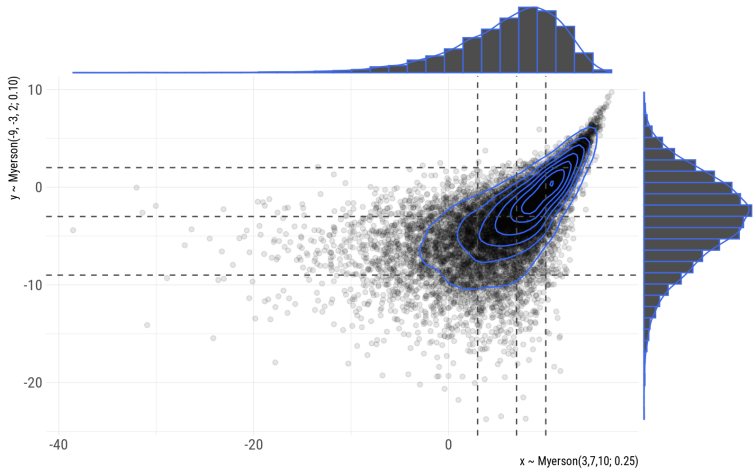


Figure 9: Samples from the bivariate Myerson distribution joined by the Joe copula ($\theta = 3$)

4.3. Bivariate quantiles

The formal definition of bivariate quantile functions and the method for constructing bivariate quantile distributions using marginal and conditional quantile functions are provided by Nair and Vineshkumar (2023) and Vineshkumar and Nair (2019). They define the bivarate quantile function (bQF) of (X_1, X_2) as the pair $Q(u_1, u_2) = (Q_1(u_1), Q_{21}(u_2|u_1))$, where $Q_1(u_1) = \inf\{x_1 : F_1(x_1) \geq u_1\}$, $u_1 \in [0, 1]$ and $Q_{21}(u_2|u_1) = \inf\{x_2 : F_{21}(Q_1, x_2) \geq u_2\}$.

The conditional quantile function $Q_{21}(u_2|u_1)$ can be obtained by inverting the conditional distribution function $F_{21}(x_1, x_2)$, which is computed from the factorization of the joint survival function. The joint survival function is defined as $\bar{F}(x_1, x_2) = P(X_1 > x_1)P(X_2 > x_2|X_1 > x_1) = \bar{F}(x_1)\bar{F}_{21}(x_1, x_2)$. Note that the joint survival function $\bar{F}(x_1, x_2) = 1 - F_1(x_1) - F_2(x_2) + F(x_1, x_2)$, and the conditional survival function $\bar{F}_{21}(x_1, x_2) = 1 - F_{21}(x_1, x_2)$.

Another approach for creating bivariate quantile functions is through Gilchrist's QF transformation rules (Gilchrist, 2000), which can be generalized to bivariate quantile functions. According to Nair and Vineshkumar (2023) (Property 6), the conditional QF can be constructed as a sum of two univariate QFs: $Q_{21}(u_2|u_1) = Q_1(u_1) + Q_2(u_2)$. This means that the pair $(Q_1(u_1), ; Q_1(u_1) + Q_2(u_2))$ is a valid bivariate quantile function, which generalizes Gilchrist's *addition rule* (Table 1). The addition rule also works for quantile density functions (Property 7). If Q_1 is left-bounded at zero, i.e., $Q_1(0) = 0$, then the margins of such a bQF are $X_1 = Q_1(u_1)$ and $X_2 = Q_2(u_2)$. Otherwise, the marginal distribution of X_2 will be $\lim_{u_1 \to 0} Q_{21}(u_2|u_1)$, which in many cases is not tractable.

If $Q_1(u_1)$ and $Q_2(u_2)$ are positive on $u_i \in [0, 1]$, then their product is also a valid conditional QF (Property 8), generalizing Gilchrist's "product rule". Finally, Property 9 generalizes the "Q-transformation rule," stating that for every increasing transformation functions T_1 and T_2, $(T_1(Q_1(u_1)), T_1(Q_1(u_1)) + T_2(Q_2(u_2)))$ is also a valid bQF.

Therefore, valid bivariate quantile-parameterized QFs can be created by constructing the conditional quantile functions as Gilchrist combinations of univariate quantile-parameterized QFs. Figure 10 shows 1000 samples from the bivariate distribution created by adding together two Myerson distributions. Note that in this case, only the marginal distribution of $x_1 = Q_1(u_1)$ is available in closed form.

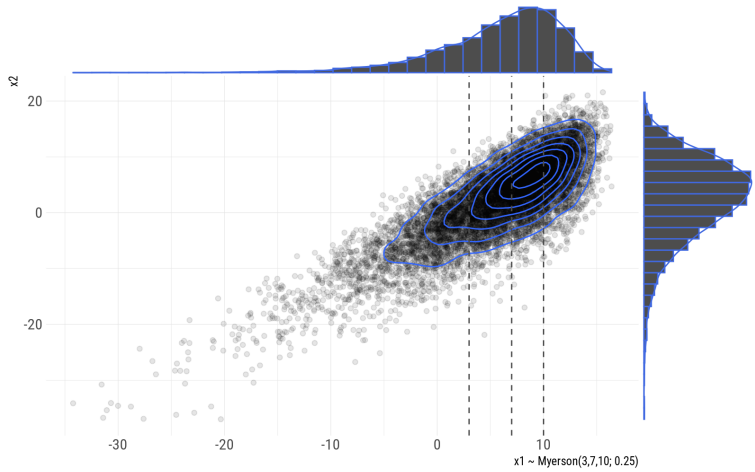


Figure 10: Samples from the Bivariate Myerson quantile function

$$
\begin{aligned}
(u_1, u_2) &\overset{X_1, X_2}{\backsim} (Q_1(u_1), Q_1(u_1) + Q_2(u_2)) \\
Q_1(u_1) &\sim \text{Myerson}(3, 7, 10; 0.1) \\
Q_2(u_2) &\sim \text{Myerson}(-9, -3, 2; 0.25)
\end{aligned} \tag{38}
$$

This bQF is easy to elicit and interpret, since $Q_2(u_2)$ can be thought of as a random adjustment to the value of $Q_1(u_1)$. In fact, the conditional quantile function $Q_{21}(u_2|u_1)$ can be thought of as having the classical form $Q_{21}(u_2|u_1) = \mu(u_1) + \sigma Q_2(u_2)$ (Gilchrist, 2000), where the location is randomly varying with $\mu(u_1) = Q_1(u_1)$ and the scale parameter $\sigma = 1$. First, the marginal distribution $Q_1(u_1)$ is elicited, and then the difference between the values x_1 and x_2 can be elicited as a QPT and encoded as $Q_2(u_2)$.

5. Discussion

Quantile-based distributions have garnered significant attention in the research community. Several distributions, such as the Generalized Lambda Distribution (GLD) (Freimer et al., 1988; Ramberg and Schmeiser, 1974), the g-and-k distribution (Haynes et al., 1997; Haynes and Mengersen, 2005; Jacob, 2017; Prangle, 2017), the g-and-h distribution (Field and Genton, 2006; Mac Gillivray, 1992; Rayner and MacGillivray, 2002), and the Wakeby distribution (Jeong-Soo, 2005; Rahman et al., 2015; Tarsitano, 2005b), have been extensively studied and documented in the literature. These distributions are defined by non-invertible quantile functions (Perepolkin et al., 2021b). However, the research on quantile-parameterized distributions remains relatively unexplored. These distributions offer interpretable parameters that are defined on the same scale as the quantities of interest, simplifying the elicitation process for experts. Many popular elicitation protocols for both predictive and parametric elicitation rely on the assessment of quantile-probability pairs (QPPs). Instead of fitting a parametric distribution to the elicited QPPs (Best et al., 2020; O'Hagan, 2019), assessors could directly use the elicited QPPs as inputs into one of the QPD quantile functions, which can be easily employed in both quantile-parameterized and parametric models.

Provided that the expert and the elicitor agree on the scientific model to be used for representing the expert's understanding of the world (Burgman et al., 2021), several types of inputs may be required to inform the model. Among those are the expert's judgement about the model *parameters* (Mikkola

et al., 2021; O'Hagan, 2019) and their *predictions* of the next observation (Akbarov, 2009; Kadane and Wolfson, 1998; Winkler, 1980). Both parametric and predictive judgments should be captured together with corresponding uncertainties to reflect the expert's state of knowledge. Quantile-parameterized distributions offer distinct advantages as high-fidelity priors that precisely capture expert assessments. These distributions are particularly beneficial for domain experts who may not be well-versed in statistics, as they provide high flexibility while retaining parameter interpretability. As a result, QPDs can faithfully represent an expert's beliefs without compromising convenience or precision.

Different quantile-parameterized distributions fitted to the same set of quantile-probability pairs may exhibit slight variations in shape. However, given the diverse range of QPDs proposed in the literature a knowledgeable assessor should be able to select an appropriate distribution and validate the choice with the expert, taking into account the thickness of the distribution tails.

Most QPDs we reviewed are parameterized by a symmetric percentile triplet (SPT). These distributions rely on the symmetric property of underlying *kernel* distributions and can be generalized by swapping the distribution with another one that exhibits different tail shapes. Hadlock and Bickel (2019) utilized this method to generalize Johnson Quantile Parameterized distributions (J-QPDs). In our study, we applied a similar approach to generalize the Myerson distribution, which has not yet been explored in the statistical literature.

The distributions discussed in this paper are defined using the quantile function and, therefore, they can be considered *quantile-based* quantile-parameterized distributions. Myerson, J-QPD, and several other quantile-parameterized distributions cleverly reparameterize conventional distributions, utilizing Gilchrist's Quantile Function (QF) transformations (Gilchrist, 2000).

Perepolkin et al. (2021b) demonstrated that the distributions defined by the quantile function can be used both as prior and as likelihood in Bayesian models. Priors defined by the quantile function eliminate the need to compute prior density. The quantile function acts as a non-linear transformation of a uniform degenerate random variate with the resulting Jacobian adjustment reciprocal to the density quantile function. Therefore, both the Jacobian and the density quantile function are omitted from the Bayesian updating equation (Perepolkin et al., 2021b). When using quantile-based QPDs as likelihood, special care needs to be taken with regards to the suitable prior for the QPP parameters. Perepolkin et al. (2021a) used the Dirichet-based prior for the metalog likelihood model and descibed the *hybrid* elicitation process for encoding the expert judgments into the two-dimensional prior distribution implied by the model.

Not all QPDs are equally reliable in approximating the underlying distributions. Violating the QF transformation rules imposes additional constraints on the feasibility of parameters, as certain combinations of parameters may result in locally decreasing quantile functions (Keelin, 2016; Hadlock, 2017). We discussed this limitation in relation to SQN and metalog distributions, but the same challenges affect other distributions with QF violating Gilchrist QF transformation rules. In this regard, the quantile-parameterized model, which relies on Gilchrist combination of basic quantile functions, proposed by Peng et al. (2023), represents a highly promising advancement. Weighted constrained optimization algorithm ensuring that the quantile mixture weights remain non-negative opens new possibilities for other QPDs using monotonic transformations of quantile functions. The estimator proposed by Peng et al. (2023) is asymptotically a q-Wasserstein distance, which has also been used for parameter estimation in Approximate Bayesian Computation (Bernton et al., 2019).

The feasibility conditions for the Generalized Lambda Distribution (GLD) have been a focal point of numerous research endeavors in the past (Dean, 2013; Fournier et al., 2007; Karian and Dudewicz, 2019; King and MacGillivray, 2007; Tarsitano, 2005a, etc). Various reparameterizations have been explored to enhance parameter identifiability (Ramberg and Schmeiser, 1974). Recently, Chalabi et al. (2012) proposed a novel asymmetry-skewness reparameterization (CSW GLD) for the previously popular FKML GLD (Freimer et al., 1988), wherein two of the four parameters are mapped to robust quantile-based moments, namely the median and Interquartile Range (IQR). This reduction in the number of parameters required for data fitting simplifies the previously computationally intensive fitting algorithms. As demonstrated in the plot of robust moments (Figure 5) GLD remains one of the most flexible unbounded distributions, capable of accommodating a wide range of shapes. Dedduwakumara et al. (2021) described a two-step method for fitting FKML GLD using the probability density quantile function (Staudte, 2017). However, when applying

their method to fitting the CSW GLD, the second step becomes unnecessary as the location and scale can be directly mapped to the empirical first and second robust moments.

CSW GLD represents a prime example of clever reparameterization aiming at alleviating the deficiencies of QF construction through setting consistent parameter boundaries and defining fall-back cases for an impossible combination of parameters. This degree of reparameterization is difficult for QPDs because the objective is to retain the mapping of parameters to the valid set of quantile-probability pairs. Therefore, for improperly constructed QPDs the feasibility conditions will have to be expressed as ratios of quantiles. Keelin (2017) provide approximate feasibility conditions for 3- and 4-term metalog. Further research on the general metalog feasibility is necessary to promote wider adoption of this distribution.

Quantile-parameterized distributions can be readily extended to the multivariate setting by leveraging traditional multivariate distributions. The combination of quantile-based marginal distributions joined by the multivariate normal has been previously discussed in the literature (Drovandi and Pettitt, 2011; Hoff, 2007). Building on this approach, we proposed the use of Gumbel's bivariate logistic distribution (Gumbel, 1961) to combine quantile-parameterized Logit-Myerson distributions.

Copulas offer a natural extension of univariate QPDs into the multivariate domain. Bivariate copulas can be assembled into more complex structures using vine copulas (Czado, 2019; Kurowicka and Joe, 2011; Wilson, 2018). Flexible QPDs serve as a viable alternative to empirical copulas, where the margins are represented by kernel density estimation (KDE) or other non-parametric approaches. Poorly fitted marginal distributions mean *less-than-ideal* starting point for copula modeling, because of deviations from uniformality of the copula margins.

Quantile-parameterized distributions defined by the quantile function are particularly well-suited for constructing new distributions using bivariate quantiles (Nair and Vineshkumar, 2023; Vineshkumar and Nair, 2019). The ability to construct a conditional quantile function as a Gilchrist combination of univariate quantile functions offers a convenient and interpretable approach to defining bivariate distributions, especially when the univariate quantile functions are parameterized by quantiles. These distributions are easy to sample from and construct. However, fitting these distributions to data or posterior samples can be challenging. As shown by Castillo et al. (1997) the fitting process requires all marginal and conditional quantile functions to be available in closed form, which is often unattainable.

There appears to be a limited availability of unbounded quantile-parameterized distributions in the current literature. Among the distributions we examined, only the metalog distribution and quantile mixtures can extend across the entire real line. The G-QPD system provides clear distributional bounds explicitly defined by the expert during elicitation. In contrast, the (Generalized) Myerson distribution system relies on implicit bounds that need to be communicated to the expert. Most of the distributions we reviewed are characterized by a symmetrical percentile triplet (SPT), as they rely on the symmetrical property of their kernels. However, there may be situations where an arbitrary (non-symmetrical) quantile parameterization could prove valuable (as shown by Perepolkin et al., 2021a). The development of flexible quantile-parameterized distributions defined by an arbitrary set of quantile-probability pairs using quantile mixtures (Peng et al., 2023) can enhance versatility of QPDs and facilitate their broader adoption.

In conclusion, quantile-parameterized distributions offer a valuable framework for capturing expert assessments and incorporating them into statistical models. They provide high flexibility and parameter interpretability, making them particularly beneficial for domain experts. The diverse range of quantile-parameterized distributions explored in the literature allows for customized modeling approaches that align with the expert's beliefs and uncertainties. By embracing these innovative distributions, researchers and practitioners can enhance the accuracy and reliability of their statistical models while leveraging expert knowledge effectively.

Appendix A. Distribution functions

Myerson Distribution

The derivative of the quantile function with respect to the depth u is the Quantile Density Function, which for Myerson distribution has the following form

$$q(u|q_1,q_2,q_3,\alpha) = \begin{cases} \rho\frac{\beta^\kappa \ln(\beta)}{(\beta-1)}\frac{q_{norm}(u)}{\Phi^{-1}(1-\alpha)}, & \beta \neq 1 \\ \rho\frac{q_{norm}(u)}{\Phi^{-1}(1-\alpha)}, & \beta = 1 \end{cases} \tag{39}$$

where $q_{norm} = \frac{d\Phi^{-1}(u)}{du}$ is the quantile density function for the standard normal distribution.

The Myerson distribution is invertible. The distribution function of random variable X has the form

$$\begin{aligned} \psi &= \Phi^{-1}(1-\alpha)\left(\frac{\ln\left(1+\frac{(x-q_2)(\beta-1)}{\rho}\right)}{\ln(\beta)}\right) \\ F(x|q_1,q_2,q_3,\alpha) &= \begin{cases} \Phi(\psi), & \beta \neq 1 \\ F_{normal}(x|q_2,\rho/\Phi^{-1}(1-\alpha)), & \beta = 1 \end{cases} \end{aligned} \tag{40}$$

where $\Phi()$ is the CDF of the standard normal distribution and $\Phi^{-1}()$ is its inverse. $F_{normal}(x|q_2,\rho/\Phi^{-1}(1-\alpha))$ is the CDF of the normal distribution with mean $\mu = q_2$ and standard deviation $\sigma = \rho/\Phi^{-1}(1-\alpha)$.

The derivative of the distribution function with respect to the random variable X is the probability density function, which for the Myerson distribution takes the following form

$$f(x|q_1,q_2,q_3,\alpha) = \begin{cases} \frac{\Phi^{-1}(1-\alpha)(\beta-1)}{(\rho+(x-q_2)(\beta-1))\ln(\beta)}\varphi(\psi), & \beta \neq 1 \\ f_{normal}(x|q_2,\rho/\Phi^{-1}(1-\alpha)), & \beta = 1 \end{cases} \tag{41}$$

where $\varphi()$ is the probability density function of the standard normal distribution, $f_{normal}(x|q_2,\frac{\rho}{\Phi^{-1}(1-\alpha)})$ is the PDF of the normal distribution with the mean $\mu = q_2$ and standard deviation $\sigma = \rho/\Phi^{-1}(1-\alpha))$.

Generalized Myerson Distributions

The Quantile Density Function of Generalized Myerson Distribution for $u \neq 0, u \neq 1$ is

$$q_M(u|q_1,q_2,q_3,\alpha) = \begin{cases} \rho\frac{\beta^\kappa \ln(\beta)}{(\beta-1)}\frac{s(u)}{S(1-\alpha)}, & \beta \neq 1 \\ \rho\frac{s(u)}{S(1-\alpha)}, & \beta = 1 \end{cases} \tag{42}$$

where $S(u)$ is the quantile function and $s(u) = \frac{dS(u)}{du}$ is the quantile density function for the kernel distribution. When $u = 0$ or $u = 1$ the $q_M(u) = \infty$.

The Generalized Myerson distribution is invertible. The distribution function of random variable X has the form

$$\begin{aligned} \psi &= S(1-\alpha)\left(\frac{\ln\left(1+\frac{(x-q_2)(\beta-1)}{\rho}\right)}{\ln(\beta)}\right) \\ F_M(x|q_1,q_2,q_3,\alpha) &= \begin{cases} F(\psi), & \beta \neq 1 \\ q_2 + \frac{\rho}{S(1-\alpha)}F(x), & \beta = 1 \end{cases} \end{aligned} \tag{43}$$

where $F()$ is the standard CDF of the kenel distribution and $S()$ is its inverse.

The derivative of the distribution function with respect to the random variable X is the probability density function, which for the Myerson distribution takes the following form

$$f_M(x|q_1,q_2,q_3,\alpha) = \begin{cases} \frac{S(1-\alpha)(\beta-1)}{(\rho+(x-q_2)(\beta-1))\ln(\beta)}f(\psi), & \beta \neq 1 \\ f\left(\frac{x-q_2}{\rho/S(1-\alpha)}\right), & \beta = 1 \end{cases} \tag{44}$$

where $f()$ is the probability density function of the standard kernel distribution. Compare it to the simplicity of the Quantile Density Function above.

Johnson Quantile-Parameterized Distribution

The JQPD-B quantile density function can be computed as

$$q_B(p) = \begin{cases} (u_b - l_b)\varphi(\xi + \lambda \sinh(\delta(z(p) + nc))) \times \\ \quad \lambda \cosh(\sigma(z(p) + nc))\sigma q_{norm}(p), & n \neq 0 \\ (u_b - l_b)\varphi\left(B + \left(\frac{H-L}{2c}\right) z(p)\right) \times \left(\frac{H-L}{2c}\right) q_{norm}(p), & n = 0 \end{cases} \tag{45}$$

The JQPD-B distribution function

$$F_B(x) = \begin{cases} \Phi\left((2c/(H-L))(-B + z\left(\frac{x-l}{u-l}\right))\right), & n = 0 \\ \Phi\left(\frac{1}{\delta}\sinh^{-1}\left(\frac{1}{\lambda}\left(z\left(\frac{x-l}{u-l}\right) - \xi\right)\right) - nc\right), & n \neq 0 \end{cases} \tag{46}$$

The JQPD-B probability density function (PDF) is

$$f(x) = \begin{cases} \frac{2c}{(H-L)(u_b-l_b)} \frac{1}{\varphi\left(z\left(\frac{x-l_b}{u_b-l_b}\right)\right)} \varphi\left(\frac{2c}{H-L}\left(-B + z\left(\frac{x-l_b}{u-l_b}\right)\right)\right), & n = 0 \\ \frac{1}{\delta}\frac{1}{u_b-l_b}\varphi\left(-nc + \frac{1}{\delta}\sinh^{-1}\left(\frac{1}{\lambda}\left(-\xi + z\left(\frac{x-l_b}{u_b-l_b}\right)\right)\right)\right) \frac{1}{\varphi\left(z\left(\frac{x-l_b}{u_b-l_b}\right)\right)} \frac{1}{\sqrt{\lambda^2 + \left(-\xi + z\left(\frac{x-l_b}{u_b-l_b}\right)\right)^2}}, & n \neq 0 \end{cases} \tag{47}$$

J-QPD-S quantile density function

$$q_S(p) = \begin{cases} \theta \exp\left(\lambda\delta z(p)\right) \lambda\delta q_{norm}(p), & n = 0 \\ \theta \exp\left(\lambda \sinh^{-1}(\delta z(p)) + \sinh^{-1}(nc\delta)\right) \lambda \frac{1}{\sqrt{1+(\delta z(p))^2}} \delta q_{norm}(p), & n \neq 0 \end{cases} \tag{48}$$

J-QPD-S distribution function

$$F_S(x) = \begin{cases} F_{lnorm}(x - l_b|\ln(\theta), \frac{H-B}{c}), & n = 0 \\ \Phi\left(\frac{1}{\delta}\sinh\left(\sinh^{-1}\left(\frac{1}{\lambda}\ln\frac{x-l_b}{\theta}\right) - \sinh^{-1}(nc\delta)\right)\right), & n \neq 0 \end{cases} \tag{49}$$

J-QPD-S probability density function (PDF)

$$f_S(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - ln\xi)^2}{2\frac{(H-B)^2}{c^2}}\right), & n = 0 \\ \varphi\left(\frac{\sinh(\sinh^{-1}(cn\sigma) - \sinh^{-1}(\frac{1}{\lambda}\ln\frac{x-l_b}{\theta}))}{\delta}\right) \frac{\cosh(\sinh^{-1}(cn\delta) - \sinh^{-1}(\frac{1}{\lambda}\ln\frac{x-l_b}{\theta}))}{(x-l_b)\delta\lambda\sqrt{1+\left(\frac{\ln\frac{x-l_b}{\theta}}{\lambda}\right)^2}}, & n \neq 0 \end{cases} \tag{50}$$

where $\mu = \ln\xi$ and $\sigma = \frac{H-B}{c}$.

Metalog distribution

This section recapitulates ideas and formulas provided in Keelin (2016) with our own notation and minor reinterpretations.

Metalog distribution is created from the logistic quantile function $Q(p) = \mu + s\text{logit}(p)$, where μ is the mean, s is proportional to the standard deviation such that $\sigma = s\pi/\sqrt{3}$, p is the probability $p \in [0, 1]$. The metalog quantile function is built by substitution and series expansion of its parameters μ and s with the polynomial of the form:

$$\begin{aligned} \mu &= a_1 + a_4(p-0.5) + a_5(p-0.5)^2 + a_7(p-0.5)^3 + a_9(p-0.5)^4 + \ldots, \\ s &= a_2 + a_3(p-0.5) + a_6(p-0.5)^2 + a_8(p-0.5)^3 + a_{10}(p-0.5)^4 + \ldots, \end{aligned} \tag{51}$$

where $a_i,\ i \in (1 \dots n)$ are real constants. Given a size-m QPT $\{p, q\}_m$, where $p = \{p_1 \dots p_m\}$ and $q = \{q_1 \dots q_m\}$ the vector of coefficients $a = \{a_1 \dots a_m\}$ can be determined through the set of linear equations.

$$\begin{aligned} q_1 &= a_1 + a_2\text{logit}(p_1) + a_3(p_1 - 0.5)\text{logit}(p_1) + a_4(p_1 - 0.5) + \cdots, \\ q_2 &= a_1 + a_2\text{logit}(p_2) + a_3(p_2 - 0.5)\text{logit}(p_2) + a_4(p_2 - 0.5) + \cdots, \\ &\vdots \\ q_m &= a_1 + a_2\text{logit}(p_m) + a_3(p_m - 0.5)\text{logit}(p_m) + a_4(p_m - 0.5) + \cdots. \end{aligned} \tag{52}$$

In the matrix form, this system of equations is equivalent to $q = \mathbb{P}a$, where q and a are column vectors and $\mathbb{P}$ is a $m \times n$ matrix:

$$\mathbb{P} = \begin{bmatrix} 1 & \text{logit}(p_1) & (p_1 - 0.5)\text{logit}(p_1) & (p_1 - 0.5) & \cdots \\ 1 & \text{logit}(p_2) & (p_2 - 0.5)\text{logit}(p_2) & (p_2 - 0.5) & \cdots \\ & & \vdots & & \\ 1 & \text{logit}(p_m) & (p_m - 0.5)\text{logit}(p_m) & (p_m - 0.5) & \cdots \end{bmatrix} \tag{53}$$

If $m = n$ and $\mathbb{P}$ is invertible, then the vector of coefficients a of this *properly parameterized* metalog QPD can be uniquely determined by

$$a = \mathbb{P}^{-1}q \tag{54}$$

If $m > n$ and $\mathbb{P}$ has a rank of at least n, then the vector of coefficients a of the *approximated* metalog QPD, can be estimated using

$$a = [\mathbb{P}^T\mathbb{P}]^{-1}\mathbb{P}^T q \tag{55}$$

The matrix to be inverted is always $n \times n$ regardless of the size m of QPT used.

Metalog *quantile function* (QF) with n terms $Q_{M_n}(u|a)$ can be expressed as

$$Q_{M_n}(u|a) = \begin{cases} a_1 + a_2\text{logit}(u), & \text{for } n = 2, \\ a_1 + a_2\text{logit}(u) + a_3(u - 0.5)\text{logit}(u), & \text{for } n = 3, \\ a_1 + a_2\text{logit}(u) + a_3(u - 0.5)\text{logit}(u) + a_4(u - 0.5), & \text{for } n = 4, \\ Q_{M_{n-1}} + a_n(u - 0.5)^{(n-1)/2}, & \text{for odd } n \geq 5, \\ Q_{M_{n-1}} + a_n(u - 0.5)^{n/2-1}\text{logit}(u), & \text{for even } n \geq 6, \end{cases} \tag{56}$$

where $u \in [0, 1]$ is the cumulative probability and a is the size-n parameter vector of real constants $a = \{a_1 \dots a_n\}$.

The metalog *quantile density function* (QDF) can be found by differentiating the equations (56) with respect to u:

$$q_{M_n}(u|a) = \begin{cases} a_2\mathcal{I}(u), & \text{for } n = 2, \\ a_2\mathcal{I}(u) + a_3\left((u - 0.5)\mathcal{I}(u) + \text{logit}(u)\right), & \text{for } n = 3, \\ a_2\mathcal{I}(u) + a_3\left((u - 0.5)\mathcal{I}(u) + \text{logit}(u)\right) + a_4, & \text{for } n = 4, \\ q_{M_{n-1}} + 0.5a_n(n-1)(u - 0.5)^{(n-3)/2}, & \text{for odd } n \geq 5, \\ q_{M_{n-1}} + a_n\left((u - 0.5)^{n/2-1}\mathcal{I}(u) + (0.5n - 1)(u - 0.5)^{n/2-2}\text{logit}(u)\right), & \text{for even } n \geq 6, \end{cases} \tag{57}$$

where $\mathcal{I}(u) = [u(1 - u)]^{-1}$. The constants a are feasible iif $q_{M_n}(u|a) > 0,\ \forall u \in [0, 1]$.

Metalog *density quantile function* (DQF), referred to as the "metalog pdf" in Keelin (2016) can be obtained by $f(Q_{M_n}(u|a)) = [q_{M_n}(u|a)]^{-1}$.

Metalog *cumulative distribution function* (CDF) $F_{M_n}(x|a)$ does not have an explicit form because $Q_{M_n}(u|a)$ is not invertible (Keelin, 2016). It is, however, possible to approximate $\widehat{Q}^{-1}_{M_n}(x|a)$ using approximation.

Metalog distribution is defined for all $x \in \mathbb{R}$ on the real line. Keelin (2016) provides semi-bounded *log-metalog*, and the bounded *logit-metalog* variations of the metalog distribution. As the names suggest, this is achieved through the variable substitution with $z = \ln(x - b_l)$ or $z = -\ln(b_u - x)$ for the semi-bounded case, and $z = \ln((x - b_l)/(b_u - x))$ for the bounded case, where z is metalog-distributed and b_l, b_u are the lower and upper limits, respectively. Substituting one of the transformations into the QF and QDF functions above, yields semi-bounded or bounded metalog distribution. For the exact formulae of the log-metalog and logit-metalog refer to Keelin (2016).

CSW GLD

Quantile density function for the CSW GLD is provided in (Chalabi et al., 2012)

$$\begin{aligned} q(u|\tilde{\sigma}, \chi, \xi) &= \frac{\tilde{\sigma}}{S(0.75|\chi, \xi) - S(0.25|\chi, \xi)} s(u|\chi, \xi) \\ s(u|\chi, \xi) &= \frac{d}{du} S(u|\chi, \xi) = u^{\alpha+\beta-1} + (1-u)^{\alpha-\beta-1} \end{aligned} \tag{58}$$

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Akbarov, A. (2009). *Probability Elicitation: Predictive Approach*. PhD thesis, University of Salford.

Arachchige, C. N. P. G., Prendergast, L. A., and Staudte, R. G. (2022). Robust analogs to the coefficient of variation. *Journal of Applied Statistics*, 49(2):268–290.

Basikhasteh, M., Lak, F., and Tahmasebi, S. (2021). Bayesian Estimation of Morgenstern Type Bivariate Rayleigh Distribution Using Some Types of Ranked Set Sampling. *Revista Colombiana de Estadística*, 44(2):279–296.

Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, 7(4):327–339, 3001655.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Best, N., Dallow, N., and Montague, T. (2020). Prior elicitation. *Bayesian methods in pharmaceutical research*, pages 87–109.

Boos, D. D. and Monahan, J. F. (1986). Bootstrap Methods Using Prior Information. *Biometrika*, 73(1):77–83, 2336273.

Bowley, A. L. (1920). *Elements of Statistics*. Scribner's, New York, NY.

Burgman, M., Layman, H., and French, S. (2021). Eliciting Model Structures for Multivariate Probabilistic Risk Analysis. *Frontiers in Applied Mathematics and Statistics*, 7.

Castillo, E., Sarabia, J. M., and Hadi, A. S. (1997). Fitting continuous bivariate distributions to data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):355–369.

Chalabi, Y., Scott, D. J., and Wuertz, D. (2012). Flexible distribution modeling with the generalized lambda distribution. Working Paper MPRA Paper No. 43333,, ETH, Zurich, Switzerland.

Cullen, A. C., Frey, H. C., and Frey, C. H. (1999). *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Springer Science & Business Media.

Czado, C. (2019). *Analyzing Dependent Data with Vine Copulas*. Springer Berlin Heidelberg, New York, NY.

Dean, B. (2013). *Improved Estimation and Regression Techniques with the Generalised Lambda Distribution*. PhD thesis, University of Newcastle, Callaghan, Australia.

Dedduwakumara, D. S., Prendergast, L. A., and Staudte, R. G. (2021). An efficient estimator of the parameters of the generalized lambda distribution. *Journal of Statistical Computation and Simulation*, 91(1):197–215.

Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556.

Dunson, D. B. and Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17(3):385–400.

Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). A simple transformation of copulas. *Available at SSRN 1032543*.

EFSA, P. o. P. H., Bragard, C., Baptista, P., Chatzivassiliou, E., Di Serio, F., Gonthier, P., Jaques Miret, J. A., Justesen, A. F., MacLeod, A., Magnusson, C. S., Milonas, P., Navas-Cortes, J. A., Parnell, S., Potting, R., Reignault, P. L., Stefani, E., Thulke, H.-H., van der Werf, W., Yuen, J., Zappalà, L., Makowski, D., Maiorano, A., Mosbach-Schulz, O., Pautasso, M., and Vicent Civera, A. (2023). Risk assessment of Citripestis sagittiferella for the EU. *EFSA Journal*, 21(2):e07838.

Elfadaly, F. G. and Garthwaite, P. H. (2017). Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Statistics and Computing*, 27(2):449–467.

Field, C. and Genton, M. G. (2006). The Multivariate g-and-h Distribution. *Technometrics*, 48(1):104–111.

Fiori, A. M. and Zenga, M. (2009). Karl Pearson and the origin of kurtosis. *International Statistical Review*, 77(1):40–50.

Fournier, B., Rupin, N., Bigerelle, M., Najjar, D., Iost, A., and Wilcox, R. (2007). Estimating the parameters of a generalized lambda distribution. *Computational Statistics & Data Analysis*, 51(6):2813–2835.
Freimer, M., Kollia, G., Mudholkar, G. S., and Lin, C. T. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics-Theory and Methods*, 17(10):3547–3567.
Genest, C. and Favre, A.-C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 12(4):347–368.
Genest, C., Quessy, J.-F., and Rémillard, B. (2006). Goodness-of-Fit Procedures for Copula Models Based on the Probability Integral Transformation. *Scandinavian Journal of Statistics*, 33(2):337–366, 4616928.
Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213.
Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman & Hall/CRC, Boca Raton.
Gosling, J. P. (2018). SHELF: The Sheffield elicitation framework. In *Elicitation*, pages 61–93. Springer.
Groeneveld, R. A. (1998). A Class of Quantile Measures for Kurtosis. *The American Statistician*, 52(4):325–329.
Gumbel, E. J. (1960). Bivariate Exponential Distributions. *Journal of the American Statistical Association*, 55(292):698–707, 2281591.
Gumbel, E. J. (1961). Bivariate Logistic Distributions. *Journal of the American Statistical Association*, 56(294):335–349, 2282259.
Hadlock, C. C. (2017). *Quantile-Parameterized Methods for Quantifying Uncertainty in Decision Analysis*. PhD thesis, University of Texas, Austin, TX.
Hadlock, C. C. and Bickel, J. E. (2017). Johnson Quantile-Parameterized Distributions. *Decision Analysis*, 14(1):35–64.
Hadlock, C. C. and Bickel, J. E. (2019). The Generalized Johnson Quantile-Parameterized Distribution System. *Decision Analysis*, 16(1):67–85.
Hanea, A. M., Nane, G. F., Bedford, T., and French, S., editors (2021). *Expert Judgement in Risk and Decision Analysis*, volume 293 of *International Series in Operations Research & Management Science*. Springer International Publishing, Cham.
Hartmann, M., Agiashvili, G., Bürkner, P., and Klami, A. (2020). Flexible Prior Elicitation via the Prior Predictive Distribution. *arXiv:2002.09868 [stat]*, 2002.09868.
Haynes, M. and Mengersen, K. (2005). Bayesian estimation of g-and-k distributions using MCMC. *Computational Statistics*, 20(1):7–30.
Haynes, M. A., MacGillivray, H. L., and Mengersen, K. L. (1997). Robustness of ranking and selection rules using generalised g-and-k distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66.
Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., and Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1):169–180.
Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283.
Holzhauer, B., Hampson, L. V., Gosling, J. P., Bornkamp, B., Kahn, J., Lange, M. R., Luo, W.-L., Brindicci, C., Lawrence, D., Ballerstedt, S., and O'Hagan, A. (2022). Eliciting judgements about dependent quantities of interest: The SHeffield ELicitation Framework extension and copula methods illustrated using an asthma case study. *Pharmaceutical Statistics*, 21(5):1005–1021.
Jacob, P. (2017). Likelihood calculation for the g-and-k distribution.
Jeffreys, H. (1939). *The Theory of Probability*. OUP Oxford.
Jeong-Soo, P. (2005). Wakeby Distribution and the Maximum Likelihood Estimation Algorithm in Which Probability Density Function Is Not Explicitly Expressed. *Communications for Statistical Applications and Methods*, 12(2):443–451.
Johnson, D. (1997). The triangular distribution as a proxy for the beta distribution in risk analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):387–398.
Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 2nd ed edition.
Jones, M. C., Rosco, J. F., and Pewsey, A. (2011). Skewness-Invariant Measures of Kurtosis. *The American Statistician*, 65(2):89–95.
Kadane, J. and Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19.
Kadane, J. B. (1980). Predictive and structural methods for eliciting prior distributions. In Zellner, A., editor, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 89–93. North Holland Publishing Company, Amsterdam.
Karian, Z. A. and Dudewicz, E. J. (2003). Comparison of GLD Fitting Methods: Superiority of Percentile Fits to Moments in L$\hat{2}$ Norm. *Journal of The Iranian Statistical Society*, 2(2):171–187.
Karian, Z. A. and Dudewicz, E. J. (2019). *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*. CRC Press, S.l.
Keelin, T. W. (2016). The Metalog Distributions. *Decision Analysis*, 13(4):243–277.
Keelin, T. W. (2017). The Metalog Distributions - Feasibility.
Keelin, T. W. and Howard, R. A. (2021). The Metalog Distributions: Virtually Unlimited Shape Flexibility, Combining Expert Opinion in Closed Form, and Bayesian Updating in Closed Form. Preprint, OSF Preprints.
Keelin, T. W. and Powley, B. W. (2011). Quantile-Parameterized Distributions. *Decision Analysis*, 8(3):206–219.
Kim, T.-H. and White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters*, 1(1):56–73.
King, R. A. and MacGillivray, H. L. (2007). Fitting the Generalized Lambda Distribution with Location and Scale-Free Shape Functionals. *American Journal of Mathematical and Management Sciences*, 27(3-4):441–460.

Kotz, S. and Van Dorp, J. R. (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific, Singapore ; Hackensack, NJ.

Kurowicka, D. and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley Series in Probability and Statistics. Wiley, Chichester, England ; Hoboken, NJ.

Kurowicka, D. and Joe, H., editors (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapore.

Lampasi, D. (2008). An alternative approach to measurement based on quantile functions. *Measurement*, 41(9):994–1013.

Lavine, M. (1995). On an Approximate Likelihood for Quantiles. *Biometrika*, 82(1):220–222, 2337641.

Mac Gillivray, H. (1992). Shape properties of the g-and-h and johnson families. *Communications in Statistics - Theory and Methods*, 21(5):1233–1250.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). Prior knowledge elicitation: The past, present, and future. *arXiv:2112.01380 [stat]*, 2112.01380.

Moors, J. J. A. (1988). A Quantile Alternative for Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 37(1):25–32, 2348376.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20):7176–7184.

Myerson, R. B. (2005). *Probability Models for Economic Decisions*. Duxbury Applied Series. Thomson/Brooke/Cole, Belmont, CA.

Nair, N. U. and Vineshkumar, B. (2023). Properties of Bivariate Distributions Represented through Quantile Functions. *American Journal of Mathematical and Management Sciences*, 0(0):1–12.

O'Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 73(sup1):69–81.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities: O'Hagan/Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd, Chichester, UK.

Palisade Corporation (2009). Guide to using@ RISK.: Risk analysis and simulation add-in for Microsoft Excel.

Parzen, E. (1979). Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association*, 74(365):105–121.

Peng, C., Li, Y., and Uryasev, S. (2023). Mixture Quantiles Estimated by Constrained Linear Regression. 2305.00081.

Perepolkin, D., Goodrich, B., and Sahlin, U. (2021a). Hybrid elicitation and indirect Bayesian inference with quantile-parametrized likelihood.

Perepolkin, D., Goodrich, B., and Sahlin, U. (2021b). The tenets of quantile-based inference in Bayesian models. Preprint, Open Science Framework, https://osf.io/enzgs.

Powley, B. W. (2013). *Quantile Function Methods for Decision Analysis*. PhD thesis, Stanford University, Paolo Alto, CA.

Prangle, D. (2017). Gk: An R Package for the g-and-k and generalised g-and-h Distributions. *arXiv:1706.06889 [stat]*, 1706.06889.

Rahman, A., Zaman, M. A., Haddad, K., El Adlouni, S., and Zhang, C. (2015). Applicability of Wakeby distribution in flood frequency analysis: A case study for eastern Australia. *Hydrological Processes*, 29(4):602–614.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2):78–82.

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75.

Sajeevkumar, N. and Irshad, M. (2014). Estimation of A Parameter of Morgenstern Type Bivariate Logistic Distribution with Equal Coefficients of Variation By Concomitants of Order Statistics. *Calcutta Statistical Association Bulletin*, 66(3-4):213–228.

Sharma, R. and Das, S. (2018). Regularization and Variable Selection with Copula Prior. *arXiv:1709.05514 [stat]*, 1709.05514.

Smith, M. S. (2013). Bayesian Approaches to Copula Modelling. *arXiv:1112.4204 [stat]*, 1112.4204.

Spetzler, C. S. and Staël Von Holstein, C.-A. S. (1975). Probability Encoding in Decision Analysis. *Management Science*, 22(3):340–358.

Staudte, R. G. (2017). The Shapes of Things to Come: Probability Density Quantiles. *Statistics*, 51(4):782–800, 1605.00189.

Tarsitano, A. (2005a). Estimation of the Generalized Lambda Distribution Parameters for Grouped Data. *Communications in Statistics - Theory and Methods*, 34(8):1689–1709.

Tarsitano, A. (2005b). Fitting Wakeby model using maximum likelihood. In *Statistica e Ambiente*, volume 1, pages 253–256.

Tukey, J. W. (1965). Which Part of the Sample Contains the Information? *Proceedings of the National Academy of Sciences*, 53(1):127–134.

Vineshkumar, B. and Nair, N. U. (2019). Bivariate Quantile Functions and their Applications to Reliability Modelling. *Statistica*, 79(1):3–21.

Wang, W. and Wells, M. T. (2000). Model Selection and Semiparametric Inference for Bivariate Failure-Time Data. *Journal of the American Statistical Association*, 95(449):62–72, 2669523.

Welsh, M. B. and Begg, S. H. (2018). More-or-less elicitation (MOLE): Reducing bias in range estimation and forecasting. *EURO Journal on Decision Processes*, 6(1):171–212.

Wilson, K. J. (2018). Specification of Informative Prior Distributions for Multinomial Models Using Vine Copulas. *Bayesian Analysis*, 13(3):749–766.

Wilson, K. J., Elfadaly, F. G., Garthwaite, P. H., and Oakley, J. E. (2021). Recent Advances in the Elicitation of Uncertainty Distributions from Experts for Multinomial Probabilities. In Hanea, A. M., Nane, G. F., Bedford, T., and French, S., editors, *Expert Judgement in Risk and Decision Analysis*, International Series in Operations Research & Management Science, pages 19–51. Springer International Publishing, Cham.

Winkler, R. L. (1980). Prior information, predictive distributions, and Bayesian model-building. *Bayesian Analysis in*

Econometrics and Statistics. North-Holland Publishing Company, pages 95–109.

Paper II

Computational Statistics and Data Analysis 187 (2023) 107795

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

The tenets of quantile-based inference in Bayesian models

Dmytro Perepolkin [a,*,1], Benjamin Goodrich [b], Ullrika Sahlin [a]

[a] Centre for Environmental and Climate Science, Solvegatan 37, 223 62 Lund, Sweden
[b] Applied Statistics Center, Columbia University, New York, NY, USA

ARTICLE INFO

Article history:
Received 22 August 2022
Received in revised form 7 May 2023
Accepted 2 June 2023
Available online 9 June 2023

Dataset link: https://github.com/dmi3kno/qpd

Keywords:
Bayesian analysis
Quantile functions
Quantile-based inference
Parametric quantile regression

ABSTRACT

Bayesian inference can be extended to probability distributions defined in terms of their inverse distribution function, i.e. their quantile function. This applies to both prior and likelihood. *Quantile-based likelihood* is useful in models with sampling distributions which lack an explicit probability density function. *Quantile-based prior* allows for flexible distributions to express expert knowledge. The principle of *quantile-based* Bayesian inference is demonstrated in the univariate setting with a Govindarajulu likelihood, as well as in a *parametric quantile regression*, where the error term is described by a quantile function of a Flattened Skew-Logistic distribution.

1. Introduction

Most statistics courses and textbooks introduce continuous random variables via the (cumulative) *distribution function* (CDF) and the *probability density function* (PDF). "An equally adequate representation" (Tukey, 1965) of a random variable can be made using the *inverse CDF*, known as the *quantile function* (QF), and its derivative, the *quantile density function* (QDF), but the use of such *quantile distributions* is rare. Defining a distribution via its quantile function has several advantages, including that the distributions with explicit quantile functions are easy to sample from and more complex distributions can be crafted using the simpler quantile functions as the building blocks (Gilchrist, 2000; Parzen, 1980; Hadlock, 2017; Powley, 2013).

Some of the widely used probability distributions defined in terms of the CDF and PDF (*density-defined* distributions) are not easily invertible (e.g. normal or gamma) and, therefore, the numerical approximation of their QF is used. Similarly, there are other distributions defined in terms of their QF and QDF (*quantile* distributions), that are also not invertible, and thus, the numerical approximation of their CDF can be used.

Most of the knowledge and methods for Bayesian inference have been developed for the *density-defined* distributions. While there have been several published articles where *quantile* distributions were used in the context of the likelihood-free approximate Bayesian computation (Allingham et al., 2009; Drovandi and Pettitt, 2011; Karabatsos and Leisen, 2018; Fearnhead and Prangle, 2012; Bernton et al., 2019; McVinish, 2012), the likelihood-based application of the Bayesian infer-

* Corresponding author.
E-mail addresses: dmytro.perepolkin@cec.lu.se (D. Perepolkin), ullrika.sahlin@cec.lu.se (U. Sahlin).
[1] The article includes online Supplementary Materials.

https://doi.org/10.1016/j.csda.2023.107795

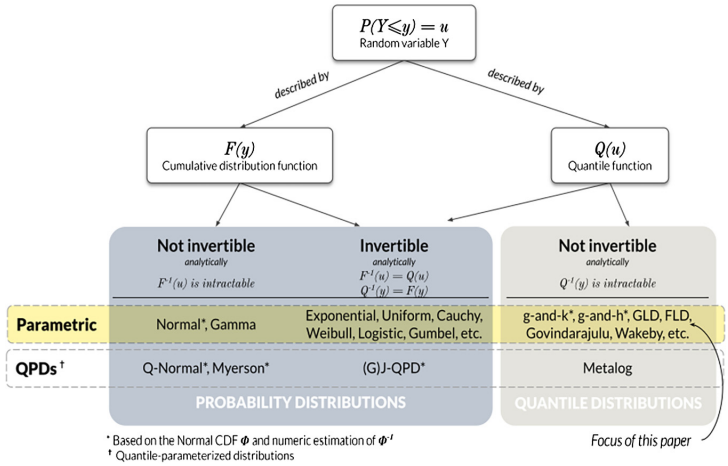


Fig. 1. Probability distributions, quantile distributions and parameterization.

ence for *quantile* distributions has been limited (Rayner and MacGillivray, 2002; Haynes and Mengersen, 2005; Nair et al., 2020).

This article builds on the ideas of Gilchrist (2000), Rayner and MacGillivray (2002), Nair et al. (2020) and systematically presents and illustrates the Bayesian inference using quantile functions. We refer to this method of inference as *quantile-based* because it deals with the inverse transformation of the *intermediate cumulative probabilities* (*depths*, indicating how "deep" an observation is into the distribution) corresponding to the observations given the parametrized model. We aim to show that the *quantile-based* Bayesian inference using the intermediate depths leads to the same posterior beliefs as the conventional density-based inference. We apply the principles of *quantile-based inference* to Bayesian updating of parameters in the univariate and regression settings using the flexible and extensible quantile sampling distributions.

Section 2 revisits the functions and identities for characterizing the distribution of a continuous random variable. Then in Section 3 we introduce the terms of *quantile-based likelihood* and *quantile-based prior* in Bayesian inference and show that the likelihood (and prior) can be expressed without the PDF. Section 4 discusses the computational aspects of the numerical inversion of quantile functions for approximating the intermediate depths in a quantile-based likelihood expressed by a quantile distribution. Section 5 discusses the applications of quantile-based inference in univariate and regression settings. We discuss the models and provide code examples implementing quantile-based likelihood in Stan (Gabry and Češnovar, 2022) and in R (R Core Team, 2021). The proposed models have been validated using the Simulation-Based Calibration (Modrák et al., 2022; Talts et al., 2020; Cook et al., 2006). The results of these simulation studies (provided in the Supplementary Materials) show successful recovery of model parameters for all widths of the posterior credible intervals. We conclude the paper with a discussion and summary of the results in Section 6.

Although the description of the *quantile-based likelihood* (Rayner and MacGillivray, 2002; King, 1999; Gilchrist, 2007; Nair et al., 2020) and prior (Nair et al., 2020) appeared in the literature before, they were presented in the context of specific distributions and not as a general principle of inference. In their recent work, Nair et al. (2020) presented Bayesian inference with quantile functions, but their presentation of what we describe here as *quantile-based prior* lacked the necessary adjustment due to the nonlinear transformation of the parameters involved (see Section 3.2 below).

In this paper, we apply the principles of *quantile-based inference* to implement the Bayesian version of the *parametric quantile regression* (Gilchrist, 2008) with the error term is described by a bespoke quantile function and estimate the regression parameters using MCMC.

2. Distribution specification

To set the scene for the discussion of density-based and quantile-based Bayesian inference we briefly review the different ways of specifying a probability distribution and discuss several examples of the distributions defined by a quantile function, found in the literature (Fig. 1).

2.1. Essential functions

Let Y be a continuous random variable. It can be expressed via the (cumulative) *distribution function* (CDF)

$$F_Y(y|\theta) = Pr(Y \leq y|\theta), \quad \theta \in \mathcal{A} \subset \mathbb{R}. \tag{1}$$

Table 1
Gilchrist's quantile function transformation rules.

Original QF	Rule	Resulting QF	Resulting variable
$Q_Y(u)$	Reflection rule	$-Q(1-u)$	QF of -Y
$Q_Y(u)$	Reciprocal rule	$1/Q(1-u)$	QF of $1/Y$
$Q_1(u)$, $Q_2(u)$	Addition rule	$Q_1(u)+Q_2(u)$	valid QF
$Q_1(u)$, $Q_2(u)$	Linear combination rule	$aQ_1(u)+bQ_2(u)$	valid QF for $a, b > 0$
$Q_1(u)$, $Q_2(u) > 0$	Multiplication rule	$Q_1(u)Q_2(u)$	valid QF
$Q_Y(u)$	Q-transformation	$T(Q_Y(u))$	QF of $T(Y)$, $T(Y)$ non-decreasing
$Q_Y(u)$	p-transformation	$Q_Y(H(u))$	p-transformation of $Q_Y(u)$, $H(u)$ non-decreasing

An alternative way of describing the random variable Y is via the *quantile function* (QF)

$$Q_Y(u|\theta) = \inf\{y : F_Y(y|\theta) \geq u\}, \quad 0 \leq u \leq 1. \tag{2}$$

The subscript $_Y$ is used to indicate the random variable that the depth u corresponds to.

New quantile functions can be easily created using Gilchrist's quantile function transformation rules (Gilchrist, 2000; Powley, 2013; Hadlock, 2017; Sharma and Chakrabarty, 2017) summarized in Table 1. We use these rules for crafting a bespoke quantile function for modeling the error term in a Bayesian parametric quantile regression in Section 5.

If $F_Y(y|\theta)$ is continuous and strictly monotonically increasing over the support of Y, then $Q_Y(u|\theta)$ is simply the inverse of $F_Y(y|\theta)$. Therefore, the quantile function is often referred to as the *inverse CDF*, i.e.

$$Q_Y(u|\theta) = F_Y^{-1}(y|\theta). \tag{3}$$

Not all QFs are analytically invertible (Fig. 1). A distribution whose quantile function $Q_Y(u|\theta)$ is not analytically invertible is called a *quantile distribution* (Gilchrist, 1997) or a *quantile-based distribution* (Sharma and Chakrabarty, 2020).

The derivative of the CDF is the *probability density function* (PDF) denoted by

$$f_Y(y|\theta) = \frac{dF_Y}{dy}. \tag{4}$$

Similarly, the derivative of the QF is the *quantile density function* (QDF) denoted by

$$q_Y(u|\theta) = \frac{dQ_Y}{du}, \quad 0 \leq u \leq 1. \tag{5}$$

The reciprocal of the QDF $[q_Y(u|\theta)]^{-1} = f(Q_Y(u|\theta))$ is referred to as the *density quantile function* (Parzen, 1980) or *p-pdf* (Gilchrist, 2000). Here and for the rest of the article, we will often omit the subscript $_Y$ and the conditioning on θ to simplify the notation.

$$f(Q(u)) = \frac{dF(Q(u))}{dQ(u)} = \frac{dF(Q(u))/du}{dQ(u)/du} = \frac{dF(F^{-1}(u))/du}{q(u)} = \frac{du/du}{q(u)} = [q(u)]^{-1}. \tag{6}$$

In Section 3 of this paper, we rely on the density quantile function (DQF) $[q(u)]^{-1}$, i.e. the density of a random variable expressed in terms of the cumulative distribution function (Perri and Tarsitano, 2007), to define the likelihood in a Bayesian model based on a quantile sampling distribution.

2.2. Derivatives of the inverses and the numerical approximation

Following the inverse function theorem (Price, 1984), for a function to be invertible in the neighborhood of a point, it should have a continuous non-zero derivative at that point. If the function is invertible, the derivative of the inverse is reciprocal to the function's derivative (Marsden et al., 1985). Formally, if dt/dy exists and $dt/dy \neq 0$, then dy/dt also exists and $dy/dt = [dt/dy]^{-1}$. Therefore, for a quantile function $Q(u) = y$, if a QDF $q(u)$ exists and $q(u) \neq 0$, then PDF $f(y)$ also exists and it is equal to $f(y) = [q(u)]^{-1}$.

Fig. 2 illustrates the relationship between the key probability functions. The distribution function (CDF) and the quantile function (QF) are depicted on the opposite sides of the Moebius strip. The derivatives of these functions (PDF and QDF, respectively) end up on the same side, due to the geometry of the Moebius surface. It means that the derivatives are no longer the inverses of each other, but rather the reciprocals, as stated in the equation at the bottom. The "do-it-yourself" copy of this probability function Moebius strip is included in the Supplementary Materials, along with the graphs of the five essential functions (CDF, PDF, QF, QDF, and DQF) for the common probability distributions (Normal, Logistic, Weibull, Exponential, and Kumaraswamy).

Even though the quantile distributions lack the closed-form CDF $F(y) = u$ in most cases the depths u can be approximated by numerically inverting the $Q(u)$. We denote the numerically inverted quantile function as $\widehat{Q^{-1}}(y)$ or $\widehat{F}(y)$. The

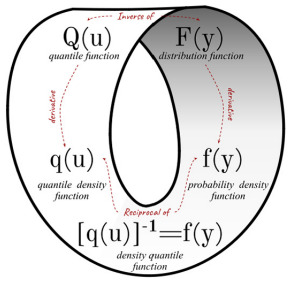


Fig. 2. Moebius strip of probability functions.

inverse of a quantile function $Q(u)$ at point u, corresponding to the observation y, is obtained by minimizing the difference between the actual observation y and $Q(u)$ by iteratively refining the depth u. The details of the numerical inversion algorithm are discussed in Section 4.

Figure provides examples of the CDF/PDF and QF/QDF for some common statistical distributions (normal, beta, lognormal, exponential and Weibull).

2.3. Quantile distributions

Statistical methods utilizing QF and QDF were pioneered by the seminal work of Parzen (1979). Some of the quantile distributions covered in the literature are generalized *g-and-h* and its sibling *g-and-k* distribution (Haynes and Mengersen, 2005; Jacob, 2017; Prangle, 2017; Rayner and MacGillivray, 2002), Tukey Lambda Distribution and its generalizations, known as GλD (Aldeni et al., 2017; Chalabi et al., 2012; Dedduwakumara et al., 2021; Ramberg and Schmeiser, 1974; Fournier et al., 2007; Freimer et al., 1988), Wakeby distribution (Rahman et al., 2015), flattened logistic distribution (Sharma and Chakrabarty, 2019) and Govindarajulu distribution (Nair et al., 2012, 2013).

The mathematical notation for describing probability distributions has been standardized and adopted across different domains. The first use of the tilde symbol $\sim$ to denote the CDF can be traced back to early 1960s. Olkin and Tate (1961) wrote: "$X \sim F(x)$ means that x is distributed according to the distribution function $F(x)$". Today this convention is adopted by the majority of Bayesian textbooks (Gelman et al., 2013; Johnson et al., 2022; O'Hagan et al., 2006; Lambert, 2018; Gelman et al., 2021; Koller and Friedman, 2009). For example, if a variable Y is normally distributed it is described as

$$Y \sim N(\mu, \sigma), \tag{7}$$

which means that the random variable Y has the distribution function $F_Y(y) = \Phi(y|\mu, \sigma)$, where Φ is the CDF of the normal distribution (Johnson et al., 1994).

When the distribution of a random variable Y is described by a non-invertible quantile function, such as, for example, the extensively researched Generalized Lambda Distribution (GLD) proposed by Ramberg and Schmeiser (1974)

$$Q_Y(u|\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{\lambda_2}\left[u^{\lambda_3} + (1-u)^{\lambda_4}\right], \tag{8}$$

stating $Y \sim \text{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is not strictly accurate, because the GLD quantile function is not invertible and its CDF can be computed only numerically as $F_Y(y) \approxeq \widehat{Q_Y^{-1}}(y)$.

Therefore, in this paper, we propose to denote this distribution as

$$u \overset{y}{\backsim} \text{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4), \tag{9}$$

where the *back-tilde* with the variable name overscript $\overset{y}{\backsim}$ should be read as "inversely distributed as" to indicate that the *depth* u is fully determined given the value of the random variable Y and the parameterized inverse distribution function indicated to the right of the *back-tilde* symbol (in this case, GLD). In situations where extra clarity is required, the *depth* variable name can also be mentioned in describing the density-defined distributions, e.g. $X \overset{p}{\sim} N(\mu, \sigma)$, where $\mu \overset{v}{\sim} N(\mu_0, \sigma_0)$.

Although in this paper we focus on the distributions with abstract parameters, the distributions parameterized by the quantile-probability pairs (*quantile-parameterized* distributions) are also worth a mention. The most prominent examples of the *quantile-parameterized density-defined* distributions are Myerson distribution (Myerson, 2005), Johnson QPD (J-QPD) with its generalizations (Hadlock and Bickel, 2017, 2019) and Simple Q-Normal distribution (Keelin and Powley, 2011). The group of the *quantile-parameterized quantile distributions* is represented by Metalog distribution (Keelin, 2016). Quantile-parameterized distributions play an important role in representing expert beliefs about variables, parameters, or quantities of interest (Dion et al., 2020; Gu et al., 2018; Larrain et al., 2021; Reinhardt et al., 2016; Baey et al., 2022), although they don't lend themselves easily as sampling distributions due to the special nature of their parameterization.

3. Bayesian inference with quantile functions

Gilchrist (2000), p. 209 used the term *quantile-based likelihood* while describing the method of maximum likelihood applied to a quantile distribution. Rayner and MacGillivray (2002) describe a three-step process of computing the log-likelihood for a quantile distribution and use it for the maximum likelihood estimation of parameters in *g-and-k* and generalized *g-and-h* distributions. Nair et al. (2020) performed quantile function substitution for both the observables $y_i = Q(u_i|\theta),\ i = \{1, 2, \dots n\}$, and parameters $\theta = Q_\Theta(v)$ and computed the Bayes estimator under the squared error loss for the Govindarajulu likelihood. In this section, we summarize this approach and use the terms *quantile-based* prior and *quantile-based* likelihood based on the identities and substitutions introduced in Section 2 to demonstrate the equivalence of the two ways of expressing the likelihood in Bayesian models.

3.1. Density-based and quantile-based likelihood

The traditional Bayesian inference formula can be restated using the substitutions involving quantile functions (Nair et al., 2020). Assume that the prior information about the scalar parameter θ can be summarized by the prior distribution over the parameter space Θ. Then, given a random sample of $\underline{y} = \{y_1, y_2, \dots y_n\}$, the posterior distribution of θ can be expressed as:

$$f(\theta|\underline{y}) \propto \mathcal{L}(\theta; \underline{y}) f(\theta), \tag{10}$$

where $f(\theta|\underline{y})$ is the posterior distribution of θ after having observed the sample $\underline{y}$, $f(\theta)$ is the prior distribution of θ, and $\mathcal{L}(\theta; \underline{y}) = \prod_{i=1}^{n} f(y_i|\theta)$ is the likelihood, which is a function of θ. We refer to this form of likelihood as *density-based*, because it is expressed using the density function (PDF) of the observable $\underline{y}$.

Given the random sample $\underline{y}$ and the value of the parameter θ, we can use the quantile function $Q_Y(u|\theta)$ to compute $\underline{Q} = \{Q(u_1), Q(u_2), \dots Q(u_n)\}$, such that $u_i = F(y_i|\theta),\ i = \{1, 2, \dots n\}$. The depths u_i are degenerate random variables because they are fully determined given the values of $\underline{y}$ and the parameter θ. Since $Q_Y(u_i|\theta) = y_i$ we can substitute $\underline{Q}$ for $\underline{y}$. Then the Bayesian inference formula (10) becomes

$$f(\theta|\underline{Q}) \propto \mathcal{L}(\theta; \underline{Q}) f(\theta). \tag{11}$$

We refer to the likelihood $\mathcal{L}(\theta; \underline{Q}) = \prod_{i=1}^{n} f(Q(u_i|\theta)) = \prod_{i=1}^{n} [q(u_i|\theta)]^{-1}$ as *quantile-based*, because it relies on computing the intermediate depths $u_i = F(y_i|\theta)$ corresponding to the observations $y_i,\ i = \{1, 2, \dots n\}$. The two forms of likelihood $\mathcal{L}(\theta; \underline{Q})$ and $\mathcal{L}(\theta; \underline{y})$ are equivalent to each other. Therefore, following the likelihood principle, the posterior beliefs about θ in both cases are identical.

Since the likelihood in the Equation (11) is expressed in terms of $\underline{Q}$, an additional transformation is required to arrive at $\underline{u} = F(\underline{y}|\theta)$. In case the CDF $F(\underline{y}|\theta)$ is not available, the numeric approximation of $\widehat{Q^{-1}}(\underline{y}|\theta)$ may be used. We discuss the details of the numerical approximation of the inverse quantile function in Section 4 of this paper.

3.2. Density-based and quantile-based prior

It is possible to extend the same logic of quantile function substitution to define *density-based* and *quantile-based* priors. In this section, we discuss the parameter transformation required for defining a quantile-based prior and show its connection to the inverse transform used for non-uniform sampling.

The Bayesian inference formula can be restated using the quantile form of the prior (Nair et al., 2020). Assume that the prior distribution of θ can be described using the invertible CDF $F_\Theta(\theta|\psi) = v$ with hyperparameter ψ, so that $Q_\Theta(v|\psi) = \theta$. Substituting the quantile values $Q_\Theta(v|\psi)$ for values of θ, prior beliefs about the parameter(s) of the distribution of θ can be expressed using the distribution of the *quantile values* corresponding to the random variate v, given hyperparameter ψ of the prior distribution as $f(Q_\Theta(v)|\psi) = [q_\Theta(v|\psi)]^{-1}$. We refer to the such formulation of the prior as *quantile-based* because it describes the prior distribution of the random variate v given hyperparameter ψ corresponding to the parameter $\theta = Q_\Theta(v|\psi)$ and not the distribution of the parameter θ itself.

Likewise, the likelihood $\mathcal{L}(Q_\Theta(v|\psi); \underline{y})$ will also rely on the parameter transformation $\theta = Q_\Theta(v|\psi)$. However, such non-linear parameter transformation requires a Jacobian adjustment (Andrilli and Hecker, 2010), which is equal to the absolute derivative of the transform, i.e. $J(Q_\Theta(v|\psi)) = |dQ_\Theta(v|\psi)/dv| = |q_\Theta(v|\psi)|$. Provided that the $Q_\Theta(v)$ is a valid (non-decreasing) quantile function, meaning that $q_\Theta(v|\psi)$ is non-negative on $v \in [0, 1]$, the density quantile term $[q_\Theta(v|\psi)]^{-1}$ representing the prior and the Jacobian adjustment $|q_\Theta(v|\psi)|$ can be dropped as they are reciprocal to each other. Therefore, the quantile-based posterior of the random variate v after observing the sample $\underline{y}$ can be expressed as

$$\begin{aligned} &[q_\Theta(v|\underline{y})]^{-1} \propto \mathcal{L}(Q_\Theta(v|\psi); \underline{y})[q_\Theta(v|\psi)]^{-1}|q_\Theta(v|\psi)| \implies \\ &[q_\Theta(v|\underline{y})]^{-1} \propto \mathcal{L}(Q_\Theta(v|\psi); \underline{y}), \end{aligned} \tag{12}$$

where $[q_\Theta(v|\underline{y})]^{-1}$ is the quantile form of the posterior, and the (quantile) prior density $[q_\Theta(v|\psi)]^{-1}$ is implied by QF transform $Q_\Theta(v|\psi) = \theta,\ v \in [0,1]$.

Quantile-based prior can also be used in combination with quantile-based likelihood, since, as we showed previously, regardless of the form of the likelihood used, the posterior beliefs about the parameter θ will be the same. In such a case, neither prior nor likelihood would require the PDF, and, therefore, both of them can be represented by quantile distributions.

4. Numerical inversion of quantile functions

4.1. Root-finding algorithms

The core element of the *quantile-based* likelihood method is the use of the intermediate depths $\underline{u}$, corresponding to the observables $\underline{y}$ given the parameter θ. These values can either be found analytically, as $F(y)$ for distributions with CDF, or numerically via root-finding algorithm, as $\widehat{F}_Y(y) \approxeq \widehat{Q_Y^{-1}}(y)$ e.g. in case of quantile distributions (Fig. 1).

The problem of inverting a quantile function is tantamount to finding the root of a target function

$$\Omega(u; y, \theta) = [y - Q_Y(u|\theta)], \tag{13}$$

where y is a known observation, θ is the parameter value, and u is the depth. Provided that the $Q_Y(u|\theta)$ is a non-decreasing function and y is a fixed observable value, the target function $\Omega(u; y, \theta)$ is non-increasing. The root-finding algorithm uses the target function to take an observable y and "pull in" its inverted equivalent $Q_Y(u|\theta)$ until the two values exactly meet by iteratively adjusting u.

Since the target function $\Omega(u; y, \theta)$ has a range $u \in [0,1]$, the *bracketing* root-finding algorithms, such as bisection or *regula falsi* (Atkinson, 2008; Burden and Faires, 2011) may be used, although depending on the shape of the quantile function their convergence can be slow. Modern bracketing methods, such as Chandrupatla (Chandrupatla, 1997), Ridders (Ridders, 1979), Zhang (Zhang, 2011; Stage, 2013) and TOMS748 (Alefeld et al., 1995), implemented in the Boost C++ library (Schäling, 2011), combine cubic, quadratic and linear (secant) interpolation to ensure robust and efficient convergence.

The convergence may be accelerated with the help of the *non-bracketing* root-finding algorithms e.g. Newton-Raphson, Halley and Schröder (Householder, 1970), which rely on computing the derivatives of the target function Ω. The first derivative of the target function is simply the negative QDF $\Omega'(u; y, \theta) = \frac{d[y - Q_Y(u|\theta)]}{du} = -q_Y(u|\theta)$. Unfortunately, the derivative-based algorithms do not guarantee that the root will be found and may end up in infinite loops and divergences. The bigger issue with trying to use a *non-bracketing* algorithm to find the root of the target function Ω is related to intermediate values of u falling outside of the [0,1] interval. In such a case $Q_Y(u|\theta)$ will return an error, which will cause the root-finding convergence check to fail. Modern derivative-based root-finding algorithms, such as NewtSafe (Acton, 1990; Press, 2007), perform the root search within a specified interval and fall back to bisection if the algorithm iteration leads the guess outside of the interval.

In this paper we used the Brent *bracketing* root-finding algorithm (Press, 2007) to invert the quantile functions. In R (R Core Team, 2021) the algorithm is available as the `uniroot()` function and in Stan (Gabry and Češnovar, 2022) we implemented it as a custom user-defined function. All source code is available in the Supplementary Materials.

4.2. Computational cost

Quantile-based method of inference comes at a computational cost associated with inverting a quantile function. In order to assess the cost of numerical inversion of quantile functions and to compare the integrated autocorrelation times (IAT) between the density-based and the quantile-based models, we performed simulation-based calibration (Modrák et al., 2022; Säilynoja et al., 2022; Talts et al., 2020; Cook et al., 2006) of the standard Exponential model, under the Gamma prior with $\alpha = 4$ and $\beta = 1$, which can be formulated both in the density-based and in the quantile-based form (since the exponential distribution is fully invertible). We refer to the Supplementary Materials for the details of the simulation-based calibration (SBC) algorithm.

We calculated IAT as the number of iterations of the sampler divided by the parameter's effective sample size (ESS) estimator (Betancourt, 2020). In addition to the standard rank-normalized ESS estimator, we calculated the minimum of the ESS for the 5% and 95% quantiles of the sample, known as the "tail ESS" (Vehtari et al., 2021).

We ran 200 replications of each of the models. Each SBC replication consisted of 2000 draws (of which half was used for burn-in) and 2 parallel chains (to assess the quality of chain mixing). The Stan code for running both density-based and the quantile-based Gamma-Exponential models is available in the Supplementary Materials.

We found that, on average, the numerical inversion of QF costs additional 6.43 sec/chain of 2000 samples (paired t-test 95% CI: [6.28, 6.59]). The MCMC proposals are also slightly more likely to be rejected (average increase in rejections is 0.345 based on the paired t-test with 95% CI of [0.216, 0.474]), as the quantile-based models are more dependent on the feasible initial values.

At the same time we found no significant difference in IAT for either the bulk of the samples (mean difference of 0.012 with 95% CI: [-0.0264, 0.0503], pair-wise t-test), nor the tail of the distribution (mean difference of -0.029 with 95% CI: [-0.0690, 0.0118], pair-wise t-test).

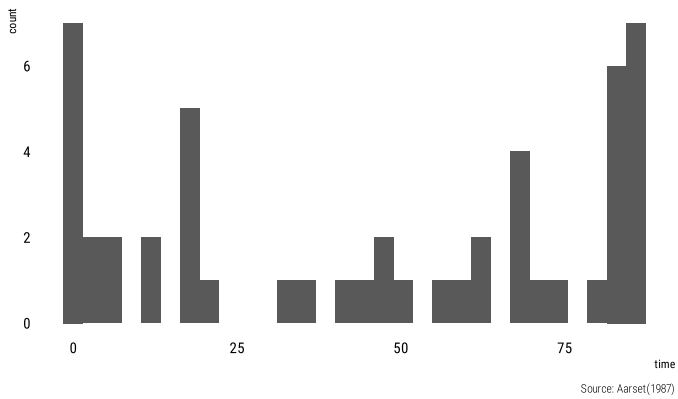


Fig. 3. Histogram of time-to-failure for 50 devices.

The cost may be more significant if the quantile function is expensive to compute, e.g. for distributions with a large number of parameters or involving computationally expensive transformations, or in the presence of covariates, as the case is for *parametric quantile regression*, discussed below.

5. Applications

In this section, we illustrate the application of the *quantile-based inference* to univariate and regression models and provide code examples for models based on the quantile sampling distributions in Stan (Gabry and Češnovar, 2022) and in R (R Core Team, 2021). For the univariate model, we update the shape parameter of a bathtub-shaped Govindarajulu distribution and for the regression model, we pick the flattened skew-logistic distribution to model the error term.

5.1. Univariate model

We take the dataset provided in Aarset (1987) on the time-to-failure of 50 devices (Fig. 3). Lifetime reliability data are often modeled using specialized distributions (Nadarajah, 2009) or 2(3)-component mixtures. Nair et al. (2020) obtained the Bayes estimator under the squared error loss function for the posterior mean of the parameter γ in the Govindarajulu distribution (Nair et al., 2012), under the generalized exponential prior (Gupta and Kundu, 2007). We reuse their example extending it to estimating the full posterior distribution by implementing the model in Stan (Gabry and Češnovar, 2022).

The QF and the QDF of the Govindarajulu distribution (Nair et al., 2012) are given by:

$$\begin{aligned} Q(u|\sigma,\gamma) &= \sigma[(\gamma+1)u^{\gamma} - \gamma u^{\gamma+1}] \\ q(u|\sigma,\gamma) &= \sigma\gamma(\gamma+1)u^{\gamma-1}(1-u), \end{aligned} \tag{14}$$

where $\sigma, \gamma > 0$. The distribution has support on $Q(u|\sigma,\gamma) \in [0,\sigma]$. Note, that the QDF in Nair et al. (2020) is slightly deviating from their original formula shown above. We refer to Nair et al. (2012) for the correct definition of the Govindarajulu distribution (including the same distribution with shifted support).

We adopt the generalized exponential prior with hyperparameters α = 0.59012 and λ = 1, used by Nair et al. (2020) for the parameter γ of the Govindarajulu distribution. The CDF and PDF of the generalized exponential distribution are given by

$$\begin{aligned} F(x|\lambda,\alpha) &= (1-e^{-\lambda x})^{\alpha} \\ f(x|\lambda,\alpha) &= \alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}, \end{aligned} \tag{15}$$

where $\alpha, \lambda > 0$. The support of the distribution is $x \in [0,\infty]$. The quantile function and the quantile density of this distribution are:

$$Q(u|\lambda,\alpha) = \frac{1}{\lambda}[-\ln(1-u^{1/\alpha})] q(u|\lambda,\alpha) = \frac{u^{(1/\alpha)-1}}{\alpha\lambda(1-u^{1/\alpha})}, \tag{16}$$

where it is visible that Generalized Exponential distribution is a p-transformed exponential distribution (Gilchrist, 2000) with the scale parameter λ, and the shape parameter α.

Nair et al. (2020) estimated the σ parameter of the Govindarajulu distribution using L-moments and assumed it to be known and equal to 93.463 for this problem, which we adopt it as a fixed parameter, as well.

Table 2
Summary of the posterior samples from the GenExp-Govindarajulu model.

parameter	mean	median	q5	q95	rhat
gamma	2.132	2.1	1.638	2.73	1.001

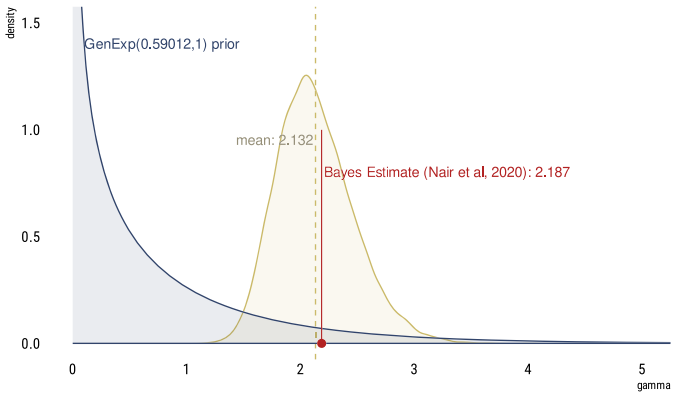


Fig. 4. Prior and posterior distributions of the parameter 'gamma' in the GenExp-Govindarajulu model.

As discussed Section 2.3, since Govindarajulu distribution does not have a closed-form CDF, it would be inappropriate to write $Y \sim \text{Govindarajulu}(\sigma, \gamma)$. Instead, our proposed notation highlights that the Govindarajulu distribution is defined via the QF, and it is, therefore, the degenerate random variate u that is inversely distributed according to this U-shaped distribution and not the observation y itself.

$$u \overset{y}{\backsim} \text{Govindarajulu}(\sigma, \gamma). \tag{17}$$

This notation also indicates the need to invert the QF to compute the random variate u corresponding to the observations y given the values of parameters σ, γ.

The resulting model takes the form

$$\begin{aligned} u &\overset{y}{\backsim} \text{Govindarajulu}(93.463, \gamma) \\ \gamma &\sim \text{GenExp}(1, 0.59012). \end{aligned} \tag{18}$$

The GenExp-Govindarajulu model has been validated using the Simulation-Based Calibration (Cook et al., 2006; Modrák et al., 2022; Talts et al., 2020). As evident from the diagnostic plots in the Supplementary Materials, the model parameters are successfully recovered for all widths of the posterior credible intervals.

We ran 2500 post-warmup iterations and 4 chains in Stan (Gabry and Češnovar, 2022). Table 2 summarizes the posterior distribution of the parameter γ of Govindarajulu distribution. We compare the prior and the posterior distribution in Fig. 4 and include the Bayes estimate by Nair et al. (2020), noting that the variation in the results could be attributed to the minor discrepancy in the quantile density formula between Nair et al. (2012) and Nair et al. (2020). The Stan code for this model can be found in the Supplementary Materials.

5.2. Parametric quantile regression

Quantile functions are useful not only for modeling the observables, but they can also be used to represent unobserved quantities of interest, such as the error term in a parametric quantile regression.

Using Gilchrist's Linear Combination rule in Table 1 any quantile function can be represented as

$$Q(u|\mu, \sigma, \theta) = \mu + \sigma Q_s(u|\theta), \tag{19}$$

where $Q_s(u)$ is a "basic" quantile function (Gilchrist, 2000), μ and σ are location and scale parameters, respectively, and θ is an optional shape parameter. Many quantile functions, such as logistic $Q(u) = \mu + \sigma \text{logit(u)}$ or normal $Q(u) = \mu + \sigma \Phi^{-1}(u)$, are already in this form. Others, such as the SLD discussed in Section 2, can have location and scale parameters added to them to enable shifted and scaled support, e.g. $Q_\ell u|\mu, \sigma, \delta) = \mu + \sigma[(1-\delta)\ln(u) - \delta \ln(1-u)]$. The basic quantile functions (i.e. $\text{logit}(u)$ for the logistic and $\Phi^{-1}(u)$ for the normal) can be useful as the building blocks for constructing more complex

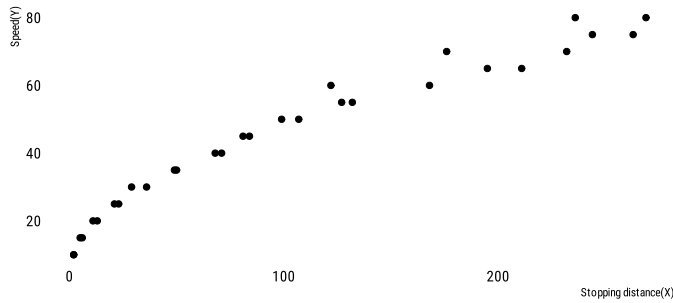


Fig. 5. Car stopping distance dataset.

distributions (Table 1). Basic quantile functions with the median centered at zero are called *standard* quantile functions (Gilchrist, 2000) denoted here as $S(u|\theta) = Q_s(u|\theta) - Q_s(0.5|\theta)$.

A simple linear regression of a random variable Y given the covariate X can be written as

$$y_i = \underbrace{\alpha + \beta x_i}_{\text{deterministic term}} + \underbrace{\varepsilon_i}_{\text{stochastic term}}, \tag{20}$$

where y_i is the i-th observation of Y, x_i is i-th observation of covariate X, α and β are unknown intercept and slope, respectively. The ε_i represents the error, which in ordinary least squares (OLS) regression is assumed (or forced through the link function) to be normally distributed with the mean of zero. An alternative way of representing the error term is through a *standard quantile function* $\varepsilon_i = S_\varepsilon(u_i|\theta)$, where u_i is the depth corresponding to the error ε_i in the regression model with the intercept α, slope β and the shape parameter θ (which are assumed to be independent).

$$y_i = \alpha + \beta x_i + S_\varepsilon(u_i|\theta), \tag{21}$$

We want to emphasize that the *traditional* quantile regression introduced by Koenker (2005) is in essence semi-parametric, because it does not require the user "to specify the distribution of the error term as it is allowed to take any form" (Yu and Moyeed, 2001). The regression Equation (21) represents the *parametric quantile regression* (PQR), because in this type of regression the error term is modeled explicitly (Gilchrist, 2008; Sharma and Chakrabarty, 2020; Su, 2015; Dean and King, 2009; Muraleedharan et al., 2016; Perri and Tarsitano, 2007, 2008).

Note that the deterministic term in (21) can be viewed as a location parameter in the quantile function

$$Q_Y(u_i|\mu_i, \theta) = \mu_i + S_\varepsilon(u_i|\theta), \tag{22}$$

where $\mu_i = \alpha + \beta x_i$. Likewise, if the stochastic component $S_\varepsilon(u_i|\theta)$ is made dependent on the covariate x_i, the resulting PQR QF can capture the heteroscedasticity of the error term.

The depth u_i can be found by inverting the quantile function $u_i \approxeq \widehat{Q_Y^{-1}}(y_i|\mu_i, \theta)$. In cases where inverting the PQR QF may be analytically difficult (e.g. when the $S_\varepsilon(u_i, \theta)$ is not invertible), the numerical approximation can be used (see Section 4 above). Once the depths $\underline{u} = \{u_1, u_2, \ldots, u_n\}$ are found, the likelihood of N observations $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ given parameter θ can be calculated using the density quantile function corresponding to the PQR QF.

Because the deterministic term μ_i in PQR QF $Q_Y(u_i|\mu_i, \theta)$ is additive and does not depend on the depth u_i it can be dropped from the derivative.

$$[q_Y(u_i|\mu_i, \theta)]^{-1} = \left[\frac{dQ_Y(u_i|\mu_i, \theta)}{du}\right]^{-1} = \left[\frac{dS_\varepsilon(u_i|\theta)}{du}\right]^{-1} = [q_\varepsilon(u_i|\theta)]^{-1}, \tag{23}$$

where $[q_\varepsilon(u_i|\theta)]^{-1}$ is the *density quantile function* of the error term.

We illustrate the application of PQR using the car stopping distance data from Gilchrist (2000), sec. 12.4. The dataset (Fig. 5) contains 30 observations of the car speed and the corresponding stopping distances. As suggested by the physics' kinetic energy equation (Lutus, 2021) the speed of the car is proportional to the square root of the braking distance. We can draw a *mean regression line* through the observations, as shown in Fig. 5 relating the car to the square root of the stopping distance using ordinary least squares (OLS). In the rest of this section we will estimate the *quantile regression lines* for the *median*, the 5th, and the 95th quantile using the PQR.

One of the simplest quantile functions which could be used to model the error in PQR is the logistic quantile function $Q(u) = \ln(u) - \ln(1-u)$. The distribution of the errors in the stopping distance model might be less "peaked" than the standard logistic distribution due to various factors (about vehicles or the drivers) not included in the model. Therefore,

adding some effect of the standard uniform quantile function u (Lampasi, 2008) might be reasonable. Flattened Logistic Distribution (FLD) described by Sharma and Chakrabarty (2019) combines the standard QFs of logistic and uniform distributions by applying positive affine transformation for scale and shape parameters (ref. Addition and Linear Combination rules in Table 1).

$$Q_\varepsilon(u|\chi,\eta,\kappa) = \chi + \eta\left[\underbrace{\ln(u) - \ln(1-u)}_{\text{logistic}} + \kappa \times \underbrace{u}_{\text{uniform}}\right], \tag{24}$$

where χ is the location parameter, $\eta, \kappa > 0$ are scale and shape parameters, respectively. For the standard quantile function $S_\varepsilon(u|\kappa) = Q_\varepsilon(u|\kappa) - Q_\varepsilon(0.5|\kappa)$, the location should be set to 0 and the scale set to 1.

The Flattened Logistic Distribution is symmetrical. This assumption might be too restrictive for modeling the residuals in the car stopping distance model (e.g. because of inertia). Sharma and Chakrabarty (2020) replaced the logistic quantile function in the FLD with the skew-logistic quantile function; the resulting QF can be referred to as the flattened skew-logistic distribution (FSLD).

The FSLD QF and the DQF are

$$\begin{aligned} Q_\varepsilon(u|\chi,\eta,\delta,\kappa) &= \chi + \eta\left(\underbrace{(1-\delta)\ln(u) - \delta\ln(1-u)}_{\text{skew-logistic}} + \kappa \times \underbrace{u}_{\text{uniform}}\right) \\ [q_\varepsilon(u|\chi,\eta,\delta,\kappa)]^{-1} &= \left[\eta\left(\frac{1-\delta}{u} + \frac{\delta}{1-u} + \kappa\right)\right]^{-1}. \end{aligned} \tag{25}$$

Since the variance in the speed Y increases with the car stopping distance X, a heteroscedastic model can be used to describe the error term in the PQR for the stopping distances. The resulting PQR QF and the corresponding DQF can be expressed as

$$\begin{aligned} Q_Y(u|\alpha,\beta,\theta;x) &= \alpha + \beta\sqrt{x} + S_\varepsilon(u;\theta)\sqrt{x} \\ [q_Y(u|\theta;x)]^{-1} &= \left[\frac{dQ_Y(u|\alpha,\beta,\theta;x)}{du}\right]^{-1} = \frac{1}{\sqrt{x}}[q_\varepsilon(u;\theta)]^{-1}, \end{aligned} \tag{26}$$

where α, β are intercept and slope, $\theta = \{\eta, \delta, \kappa\}$ represent the parameters of the standard flattened logistic distribution $S_\varepsilon(u;\theta)$ with the density quantile function $[q_\varepsilon(u|\theta)]^{-1}$, u is the *depth* corresponding to the error in the model for the speed y given the stopping distance x and the regression parameters $\{\alpha,\beta,\theta\}$. The depth u can be computed by inverting the PQR QF $u \approxeq \widehat{Q_Y^{-1}}(y|\alpha,\beta,\theta;x)$ (26).

For each of the n observations of speed in the sample $\underline{Y} = \{y_1, y_2, \ldots, y_n\}$ we can compute $\underline{Q_Y} = \{Q_Y(u_1|\alpha,\beta,\theta,x_1), \ldots, Q_Y(u_n|\alpha,\beta,\theta,x_n)\}$, such that $u_i \approxeq \widehat{Q_Y^{-1}}(y_i|\alpha,\beta,\theta,x_i)$, $i = \{1,2,\ldots n\}$.

Let's further assume that the expert's prior belief about the intercept was elicited using a set of quantile-probability pairs and the best fit was achieved using the FLD quantile function with hyperparameters $\chi = 1, \eta = 1$, and $\kappa = 10$. Similarly, the expert belief about the slope is described by the FSLD with hyperparameters $\chi = 2, \eta = 2, \delta = 0.8$, and $\kappa = 2$.

Since FLD and FSLD are quantile distributions, the prior for the parameters α and β of PQR must be defined in the quantile form. This means that the density quantile functions $f(Q_\alpha(v))$, $f(Q_\beta(w))$ and the Jacobian adjustments $|q_\alpha(v)|$, $|q_\beta(w)|$ can be dropped, as explained in Section 3.2 above.

$$\begin{aligned} f(Q_\alpha(v), Q_\beta(w), \theta|\underline{Q_Y}, x) \propto \mathcal{L}(\theta;\underline{Q_Y}, x) f(Q_\alpha(v))|q_\alpha(v)| f(Q_\beta(w))|q_\beta(w)| f(\theta) \implies \\ f(Q_\alpha(v), Q_\beta(w), \theta|\underline{Q_Y}, x) \propto \mathcal{L}(\theta;\underline{Q_Y}, x) f(\theta). \end{aligned} \tag{27}$$

Therefore, the posterior distribution of the PQR parameters $\alpha = Q_\alpha(v), \beta = Q_\beta(w)$, and θ can be expressed using the *quantile-based likelihood* (and the *quantile-based prior* for parameters α and β).

Table 3
Summary of the posterior samples from the FSLD PQR model.

parameter	mean	median	q5	q95	rhat
v	0.3452	0.3458	0.3224	0.3632	1.012
Q(v)	3.8107	3.8201	3.4817	4.0702	1.012
w	0.4599	0.4598	0.4485	0.4715	1.013
Q(w)	4.5148	4.5133	4.4254	4.6059	1.013
eta	0.2691	0.2625	0.1937	0.3655	1.014
k	0.1142	0.0804	0.0063	0.3171	1.057
dlt	0.8065	0.8329	0.5766	0.9426	1.084

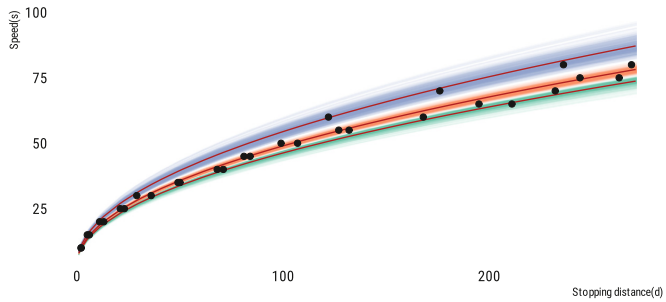


Fig. 6. Posterior predictive quantiles (0.05, 0.5, 0.95) for stopping distances. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$$
\begin{aligned}
f(\alpha,\beta,\theta|\underline{Q_Y},\underline{x}) &\propto \mathcal{L}(\theta;\underline{Q_Y},\alpha,\beta,\theta,\underline{x})f(\theta)\\
\mathcal{L}(\theta;\underline{Q_Y},\alpha,\beta,\theta,\underline{x}) &= \prod_{i=1}^{n} f(Q_Y(u_i|\alpha,\beta,\theta;x_i)) = \prod_{i=1}^{n}\left[q_\varepsilon(u_i|\theta)\sqrt{x_i}\right]^{-1}\\
Q_Y(\underline{u}|\alpha,\beta,\theta;\underline{x}) &= \alpha+\beta\sqrt{\underline{x}}+S_\varepsilon(\underline{u};\theta)\sqrt{\underline{x}}\\
u &\overset{y}{\backsim} Q_Y(\alpha,\beta,\theta;\underline{x})\\
v &\overset{\alpha}{\backsim} \mathrm{FLD}(1,1,10)\\
w &\overset{\beta}{\backsim} \mathrm{FSLD}(2,2,0.8,2)\\
\eta &\sim \mathrm{Exp}(1/10)\\
\delta &\sim \mathrm{Beta}(2,1)\\
\kappa &\sim \mathrm{Exp}(1/0.1),
\end{aligned}
\tag{28}
$$

where $f(\theta) = f(\eta)f(\delta)f(\kappa)$. Note that, as we discussed in Section 2.3, the Parametric Regression Quantile Function $Q_Y(\underline{u}|\alpha,\beta,\theta;\underline{x})$ is not invertible and therefore it would be inappropriate to write $Y \sim Q_Y(\underline{u}|\alpha,\beta,\theta;\underline{x})$. Instead, we indicate $u \overset{y}{\backsim} Q_Y(\alpha,\beta,\theta;\underline{x})$, which means that the likelihood is defined via the QF and needs to be inverted to find the random variate u corresponding to observations y. This notation also helps distinguish between the random variate used for likelihood (u) and those used by the quantile-based priors (v and w for the parameters α and β, respectively).

The PQR model has been validated using the Simulation-Based Calibration (Cook et al., 2006; Modrák et al., 2022; Talts et al., 2020) in Stan. The diagnostic plots provided in the Supplementary Materials, show that the PQR model parameters are successfully recovered for all widths of the posterior credible intervals.

We ran 2500 post-warmup iterations and 4 chains using the Robust Adaptive Metropolis algorithm by Vihola (2012) implemented in `fmcmc` package (Vega Yon and Marjoram, 2019) in R (R Core Team, 2021). The code is provided in the Supplementary Materials.

Table 3 summarizes the posterior distribution of the parameters in the parametric quantile regression model for the car stopping distances.

Posterior predictive check (Gabry et al., 2019) can be done by generating a grid of values for the car stopping distances x and using randomly sampled parameters from the posterior distribution to compute the value of the response y using the PQR QF. Since in the PQR the regression equation is expressed in terms of the depth u we can extract the coherent (non-crossing) quantile regression lines for any set of fractiles. Fig. 6 illustrates hypothetical outcome plots for the 5th, 50th, and 95th quantile regression lines. The solid red lines are the conditional mean curves, representing the respective predictive quantiles.

In order to assess the empirical goodness of fit, we calculated the proportion of data points falling below the 5th, 50th, and 95th predictive quantile. Out of $n = 30$ observations, 93% of observations fell inside the conditional 95% posterior predictive interval (shown as the outer solid red lines on the plot), while 57% of observations turned out below the predictive median curve.

6. Discussion and conclusion

In the past 20 years, many examples of using quantile distributions for the approximate Bayesian computation (ABC) appeared in the literature (Allingham et al., 2009; Drovandi and Pettitt, 2011; Dunson and Taylor, 2005; McVinish, 2012; Smithson and Shou, 2017). ABC methods normally do not require computation of the likelihood, which, in case of the quantile distributions, is convenient, as these distributions lack an explicit CDF and PDF.

Regardless of whether the distribution is defined by the CDF of the QF, the defining function sometimes needs to be inverted. If the inverse does not exist in closed form, the function has to be inverted numerically. In the case of the *density-based likelihood*, the inverse distribution function may be needed for sampling from the posterior (e.g. for the posterior predictive check). In the case of the *quantile-based likelihood*, the inverse is needed for computing the intermediate depth values, corresponding to observations (conditional on covariates) for every draw of the parameters. No numerical inversion of the quantile function is needed for defining the *quantile-based prior*. A wide selection of efficient root-finding algorithm implementations in the popular statistical software makes the inversion of custom quantile functions accessible. We provide a generic wrapper for inverting arbitrary quantile functions using Brent method in the accompanying R package (Perepolkin, 2019). Further research of custom root-finding algorithms for non-decreasing functions on unit-interval can make inverting of quantile function even more computationally efficient.

The *quantile-based inference* opens up a wide set of new distributions to serve as likelihood and/or prior in Bayesian models. Although many flexible *density-defined* distributions have been proposed in recent decades (Jones, 2015; Steel and Rubio, 2015), *quantile* distributions play an important role in certain field applications (Nair et al., 2013; Chalabi et al., 2012), as well as in expert knowledge elicitation and decision analysis (Mikkola et al., 2021; Hadlock, 2017; Powley, 2013). Besides, the flexibility offered by the distributions defined in terms of the quantile function (Gilchrist, 2007), and in particular their easily extensible nature (Table 1), allows ultimate freedom in expressing the expert-informed priors. In this paper we showed the connection of quantile parameter transformation to inverse transform sampling and used quantile distribution as a prior for regression parameters.

Multivariate versions of quantile distributions have been explored in the past (Field and Genton, 2006; Vineshkumar and Nair, 2019), but their adoption in the scientific literature remains low. One possibility of utilizing the flexibility of the quantile distributions in a multivariate setting is to employ them as marginal distributions for bivariate copulas, which can be assembled into higher-dimensional structures using vines (Czado, 2019; Kurowicka and Joe, 2011). When used as priors (Wilson, 2018), the copula structure can be elicited from the experts (Elfadaly and Garthwaite, 2017) along with the marginal quantile-probability pairs for fitting the quantile distribution (O'Hagan et al., 2006; Mikkola et al., 2021). Versatile and user-friendly multivariate quantile distributions represent an opportunity for further research.

Gilchrist (2007) provides a review of the traditional approach to quantile regression, as proposed by Koenker and Bassett (1978) and contrasts it with the *fully parametric approach* taken by PQR (Gilchrist, 2000, 2008; Su, 2015). The parametric approach to regression provides coherent (non-crossing) estimates of posterior quantiles, allowing the scientists to model the distribution of the error term explicitly (instead of making assumptions). Note that the parametric quantile regression may also be used with invertible distributions (logistic, normal, etc), as long they have computable QF and QDF ($\Phi^{-1}(u)$ and $\Phi^{-1}(u)/du$, for normal distribution).

Traditionally, the fitting of parameters in quantile distributions was performed using the matching of moments or L-moments (Gilchrist, 2008; Asquith, 2007; Karvanen and Nuutinen, 2008), matching of percentiles (Karian and Dudewicz, 2011), location and scale-free shape functionals (King and MacGillivray, 2007), distributional least squares/absolutes (Gilchrist, 2007; Sharma and Chakrabarty, 2020), and maximum likelihood (Rayner and MacGillivray, 2002; Su, 2007; Tarsitano, 2005). The various methods of obtaining parameter estimates for the quantile distributions have been extensively studied and compared, primarily in application to GλD (King and MacGillivray, 1999; Karian and Dudewicz, 2011; Fournier et al., 2007; Tarsitano, 2010), but also to some other distributions (Rayner and MacGillivray, 2002; Jeong-Soo, 2005). This paper generalizes the approach to *quantile-based likelihood* (Gilchrist, 2000; Rayner and MacGillivray, 2002; Haynes and Mengersen, 2005; Nair et al., 2020) connecting the previous research on parametric quantile regression (Gilchrist, 2008; Su, 2015; Sharma and Chakrabarty, 2020) with more recently introduced work on *quantile-based priors* (Nair et al., 2020) and implementing both of these concepts in Stan (Gabry and Češnovar, 2022) and R (Vega Yon and Marjoram, 2019).

Since the definition of quantile function is usually mathematically simpler (and more easily extendable) than the respective CDF and PDF (Gilchrist, 2000), *quantile-based priors* represent an inexpensive and flexible way of incorporating prior knowledge in Bayesian models. The unit sampling space may offer some additional computational advantage for MCMC/HMC algorithms. The quantile-based formulation of the prior may not be appropriate if the sampler constraints need to be defined on the parameter level (e.g. prior truncation). In such a case, the traditional *density-based prior* may be more useful.

Quantile-based likelihoods open a wide range of possibilities for designing flexible data generative models for special areas of application (e.g. Govindarajulu for reliability problems, Wakeby for modeling of floods, GLD in other instances) where the *density-based* equivalent is not available. The alternatives usually involve falling back to the non-Bayesian estimation

methods (Karian and Dudewicz, 2011; Asquith, 2007; Karvanen and Nuutinen, 2008) or using the approximate computation algorithms (Drovandi et al., 2011; Dunson and Taylor, 2005; McVinish, 2012), both of which are outside of the scope for this paper. The availability of efficient MCMC samplers (Gabry and Češnovar, 2022) and modern root-finding algorithms (Schäling, 2011), make quantile-based likelihood computationally feasible. Gilchrist (2000) writes: "The lack of use of maximum likelihood is surprising as it is perfectly straightforward if one uses the general-purpose maximization software available rather than look for specific formulae for estimators". We share his sentiment.

Embracing and expanding the use of quantile distributions in Bayesian analysis can enable new solutions for old problems and enrich the toolkit available to scientists for performing hard inference tasks. We hope that the *quantile-based inference* methods presented in this paper can contribute to the expanding body of knowledge about the use of quantile functions in Bayesian statistics and fuel further research in the area of quantile distributions.

Data availability

The `qpd` R package used in this paper is available on Github at https://github.com/dmi3kno/qpd. Contact corresponding author Dmytro Perepolkin (Dmytro.Perepolkin@cec.lu.se) for requests for data.

Acknowledgements

The authors have no conflict of interest to declare. U Sahlin was funded by The Crafoord Foundation (ref 20200626). B Goodrich is supported by U.S. National Science Foundation grants 2051246 and 2153019. We thank the editorial team and reviewers for their constructive feedback which helped us improve this manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107795.

References

Aarset, Magne Vollan, 1987. How to identify a bathtub hazard rate. IEEE Trans. Reliab. (ISSN 1558-1721) R-36 (1), 106–108. doi:10/ddsps5.
Acton, Forman S., 1990. Numerical Methods That Work. Mathematical Association of America, Washington, D.C. ISBN 978-0-88385-450-1.
Aldeni, Mahmoud, Lee, Carl, Famoye, Felix, 2017. Families of distributions arising from the quantile of generalized lambda distribution. J. Stat. Distrib. Appl. (ISSN 2195-5832) 4 (1), 25. doi:10/gjrbzj.
Alefeld, G.E., Potra, F.A., Shi, Yixun, 1995. Algorithm 748: enclosing zeros of continuous functions. ACM Trans. Math. Softw. 21 (3), 327–344. ISSN 0098-3500, 1557-7295, doi:10/dd9c45.
Allingham, D., King, R.A.R., Mengersen, K.L., 2009. Bayesian estimation of quantile distributions. Stat. Comput. 19 (2), 189–201. ISSN 0960-3174, 1573-1375, doi:10/dn3mfd.
Andrilli, Stephen Francis, Hecker, David, 2010. Elementary Linear Algebra, 4th ed edition. Elsevier Academic Press, Amsterdam, Boston. ISBN 978-0-12-374751-8.
Asquith, William H., 2007. L-moments and TL-moments of the generalized lambda distribution. Comput. Stat. Data Anal. (ISSN 0167-9473) 51 (9), 4484–4496. https://doi.org/10.1016/j.csda.2006.07.016.
Atkinson, Kendall E., 2008. An Introduction to Numerical Analysis. John Wiley & Sons. ISBN 81-265-1850-2.
Baey, Charlotte, Smith, Henrik G., Rundlöf, Maj, Olsson, Ola, Clough, Yann, Sahlin, Ullrika, 2022. Calibration of a bumble bee foraging model using approximate Bayesian computation.
Bernton, Espen, Jacob, Pierre E., Gerber, Mathieu, Robert, Christian P., 2019. Approximate Bayesian computation with the Wasserstein distance. J. R. Stat. Soc., Ser. B, Stat. Methodol. (ISSN 1467-9868) 81 (2), 235–269. doi:10/gjvfgp.
Betancourt, Michael, 2020. Markov chain Monte Carlo in practice. https://betanalpha.github.io/assets/case_studies/markov_chain_monte_carlo.html#422_asymptotic_variance_and_the_effective_sample_size.
Burden, Richard L., Faires, J. Douglas, 2011. Numerical Analysis, 9th ed edition. Brooks/Cole, Cengage Learning, Boston, MA. ISBN 978-0-538-73351-9.
Chalabi, Yohan, Scott, David J., Wuertz, Diethelm, 2012. Flexible distribution modeling with the generalized lambda distribution. Working paper. ETH, Zurich, Switzerland.
Chandrupatla, Tirupathi R., 1997. A new hybrid quadratic/bisection algorithm for finding the zero of a nonlinear function without using derivatives. Adv. Eng. Softw. (ISSN 0965-9978) 28 (3), 145–149. https://doi.org/10.1016/S0965-9978(96)00051-8.
Cook, Samantha R., Gelman, Andrew, Rubin, Donald B., 2006. Validation of software for Bayesian models using posterior quantiles. J. Comput. Graph. Stat. 15 (3), 675–692. ISSN 1061-8600, 1537-2715, doi:10/dgth3q.
Czado, Claudia, 2019. Analyzing Dependent Data with Vine Copulas. Springer Berlin Heidelberg, New York, NY. ISBN 978-3-030-13784-7.
Dean, Benjamin, King, A.R., 2009. Versatile regression: simple regression with a non-normal error distribution. In: Third Annual Applied Statistics Education and Research Collaboration (ASEARC) Conference, pp. 7–8.
Dedduwakumara, Dilanka S., Prendergast, Luke A., Staudte, Robert G., 2021. An efficient estimator of the parameters of the generalized lambda distribution. J. Stat. Comput. Simul. (ISSN 0094-9655) 91 (1), 197–215. doi:10/gkcqg3.
Dion, Patrice, Galbraith, Nora, Sirag, Elham, 2020. Using expert elicitation to build long-term projection assumptions. In: Mazzuco, Stefano, Keilman, Nico (Eds.), Developments in Demographic Forecasting, vol. 49. Springer International Publishing, Cham, pp. 43–62. ISBN 978-3-030-42471-8, 978-3-030-42472-5.
Drovandi, Christopher C., Pettitt, Anthony N., 2011. Likelihood-free Bayesian estimation of multivariate quantile distributions. Comput. Stat. Data Anal. (ISSN 0167-9473) 55 (9), 2541–2556. doi:10/fdwbxf.
Drovandi, Christopher C., Pettitt, Anthony N., Faddy, Malcolm J., 2011. Approximate Bayesian computation using indirect inference. J. R. Stat. Soc., Ser. C, Appl. Stat. (ISSN 1467-9876) 60 (3), 317–337. doi:10/fb92vd.
Dunson, David B., Taylor, Jack A., 2005. Approximate Bayesian inference for quantiles. J. Nonparametr. Stat. 17 (3), 385–400. ISSN 1048-5252, 1029-0311, doi:10/b2h4mp.

Elfadaly, Fadlalla G., Garthwaite, Paul H., 2017. Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. Stat. Comput. (ISSN 0960-3174) 27 (2), 449–467. doi:10/ghgk3q.
Fearnhead, Paul, Prangle, Dennis, 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc., Ser. B, Stat. Methodol. (ISSN 1467-9868) 74 (3), 419–474. doi:10/f3zg9k.
Field, Christopher, Genton, Marc G., 2006. The multivariate g-and-h distribution. Technometrics (ISSN 0040-1706) 48 (1), 104–111. doi:10/dtrr7b.
Fournier, Benjamin, Rupin, Nicolas, Bigerelle, Maxence, Najjar, Denis, Iost, Alain, Wilcox, R., 2007. Estimating the parameters of a generalized lambda distribution. Comput. Stat. Data Anal. (ISSN 0167-9473) 51 (6), 2813–2835. doi:10/d5fxtn.
Freimer, Marshall, Kollia, Georgia, Mudholkar, Govind S., Lin, C. Thomas, 1988. A study of the generalized Tukey lambda family. Commun. Stat., Theory Methods (ISSN 0361-0926) 17 (10), 3547–3567. doi:10/fpccx7.
Gabry, Jonah, Češnovar, Rok, 2022. Cmdstanr: R Interface to 'CmdStan'.
Gabry, Jonah, Simpson, Daniel, Vehtari, Aki, Betancourt, Michael, Gelman, Andrew, 2019. Visualization in Bayesian workflow. J. R. Stat. Soc., Ser. A, Stat. Soc. (ISSN 1467-985X) 182 (2), 389–402. doi:10/gftqws.
Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, Rubin, Donald B., 2013. Bayesian Data Analysis. CRC Press.
Gelman, Andrew, Hill, Jennifer, Vehtari, Aki, 2021. Regression and other stories. In: Analytical Methods for Social Research. Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore. ISBN 978-1-107-67651-0, 978-1-107-02398-7.
Gilchrist, Warren, 1997. Modelling with quantile distribution functions. J. Appl. Stat. (ISSN 0266-4763) 24 (1), 113–122. doi:10/c2bvc4.
Gilchrist, Warren, 2000. Statistical Modelling with Quantile Functions. Chapman & Hall/CRC, Boca Raton. ISBN 978-1-58488-174-2.
Gilchrist, Warren, 2008. Regression revisited. Int. Stat. Rev. (ISSN 1751-5823) 76 (3), 401–418. https://doi.org/10.1111/j.1751-5823.2008.00053.x.
Gilchrist, Warren G., 2007. Modeling and fitting quantile distributions and regressions. Am. J. Math. Manag. Sci. (ISSN 0196-6324) 27 (3–4), 401–439. doi:10/gjqt4f.
Gu, Mengyang, Bhattacharjya, Debarun, Subramanian, Dharmashankar, 2018. Nonparametric estimation of utility functions. arXiv:1807.10840 [stat].
Gupta, Rameshwar D., Kundu, Debasis, 2007. Generalized exponential distribution: existing results and some recent developments. J. Stat. Plan. Inference (ISSN 0378-3758) 137 (11), 3537–3547. doi:10/d3vvwb.
Hadlock, Christopher C., Bickel, J. Eric, 2017. Johnson quantile-parameterized distributions. Decis. Anal. (ISSN 1545-8490) 14 (1), 35–64. doi:10/f936ww.
Hadlock, Christopher C., Bickel, J. Eric, 2019. The generalized Johnson quantile-parameterized distribution system. Decis. Anal. (ISSN 1545-8490) 16 (1), 67–85. doi:10/ghhdxr.
Hadlock, Christopher Campbell, 2017. Quantile-Parameterized Methods for Quantifying Uncertainty in Decision Analysis. PhD thesis. University of Texas, Austin, TX.
Haynes, Michele, Mengersen, Kerrie, 2005. Bayesian estimation of g-and-k distributions using MCMC. Comput. Stat. (ISSN 1613-9658) 20 (1), 7–30. doi:10/dpgjv5.
Householder, A.S., 1970. The Numerical Treatment of a Single Nonlinear Equation. McGraw-Hill, London. ISBN 978-0-07-030465-9.
Jacob, Pierre, 2017. Likelihood calculation for the g-and-k distribution.
Jeong-Soo, Park, 2005. Wakeby distribution and the maximum likelihood estimation algorithm in which probability density function is not explicitly expressed. Commun. Stat. Appl. Methods (ISSN 2287-7843) 12 (2), 443–451. https://doi.org/10.5351/CKSS.2005.12.2.443.
Johnson, Alicia A., Ott, Mies Q., Dogucu, Mine, 2022. Bayes Rules! An Introduction to Bayesian Modeling with R. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton. ISBN 978-0-367-25539-8. 978-1-03-219159-1.
Johnson, Norman Lloyd, Kotz, Samuel, Balakrishnan, N., 1994. Continuous Univariate Distributions, 2nd edition. Wiley Series in Probability and Mathematical Statistics. Wiley, New York. ISBN 978-0-471-58495-7, 978-0-471-58494-0.
Jones, M.C., 2015. On families of distributions with shape parameters. Int. Stat. Rev. (Revue Internationale de Statistique) (ISSN 0306-7734) 83 (2), 175–192. doi:10/b8z6.
Karabatsos, George, Leisen, Fabrizio, 2018. An approximate likelihood perspective on ABC methods. Stat. Surv. (ISSN 1935-7516) 12 (none), 66–104. doi:10/gj47cf.
Karian, Zaven A., Dudewicz, Edward J., 2011. Handbook of Fitting Statistical Distributions with R. CRC Press, Boca Raton, FL. ISBN 978-1-58488-711-9.
Karvanen, Juha, Nuutinen, Arto, 2008. Characterizing the generalized lambda distribution by l-moments. Comput. Stat. Data Anal. (ISSN 0167-9473) 52 (4), 1971–1983. https://doi.org/10.1016/j.csda.2007.06.021.
Keelin, Thomas W., 2016. The metalog distributions. Decis. Anal. (ISSN 1545-8490) 13 (4), 243–277. doi:10/f9n7nt.
Keelin, Thomas W., Powley, Bradford W., 2011. Quantile-parameterized distributions. Decis. Anal. (ISSN 1545-8490) 8 (3), 206–219. https://doi.org/10.1287/deca.1110.0213.
King, Robert A.R., MacGillivray, H.L., 1999. A starship estimation method for the generalized lambda distributions. Aust. N. Z. J. Stat. 41 (3), 353–374. https://doi.org/10.1111/1467-842X.00089. ISSN 1369-1473, 1467-842X.
King, Robert A.R., MacGillivray, H.L., 2007. Fitting the generalized lambda distribution with location and scale-free shape functionals. Am. J. Math. Manag. Sci. (ISSN 0196-6324) 27 (3–4), 441–460. https://doi.org/10.1080/01966324.2007.10737708.
King, Robert Arthur Ravenscroft, 1999. New Distributional Fitting Methods Applied to the Generalised [Lambda] Distribution. PhD thesis. Queensland University of Technology, Australia.
Koenker, Roger, 2005. Quantile Regression. Econometric Society Monographs, vol. 38. Cambridge University Press, Cambridge, New York. ISBN 978-0-521-60827-5, 978-0-521-84573-1.
Koenker, Roger, Bassett, Gilbert, 1978. Regression quantiles. Econometrica (ISSN 0012-9682) 46 (1), 33–50. https://doi.org/10.2307/1913643.
Koller, Daphne, Friedman, Nir, 2009. Probabilistic Graphical Models: Principles and Techniques. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA. ISBN 978-0-262-01319-2.
Kurowicka, Dorota, Joe, Harry (Eds.), 2011. Dependence Modeling: Vine Copula Handbook. World Scientific, Singapore. ISBN 978-981-4299-87-9.
Lambert, Ben, 2018. A Student's Guide to Bayesian Statistics. SAGE, Los Angeles. ISBN 978-1-4739-1636-4, 978-1-4739-1635-7.
Lampasi, D.A., 2008. An alternative approach to measurement based on quantile functions. Measurement (ISSN 0263-2241) 41 (9), 994–1013. doi:10/fdkww2.
Larrain, Macarena, Van Passel, Steven, Thomassen, Gwenny, Van Gorp, Bart, Nhu, Trang T., Huysveld, Sophie, Van Geem, Kevin M., De Meester, Steven, Billen, Pieter, 2021. Techno-economic assessment of mechanical recycling of challenging post-consumer plastic packaging waste. Resour. Conserv. Recycl. (ISSN 0921-3449) 170, 105607. https://doi.org/10.1016/j.resconrec.2021.105607.
Lutus, P., 2021. The physics behind stopping a car. https://arachnoid.com/braking_physics/index.html.
Marsden, Jerrold E., Weinstein, Alan, Marsden, Jerrold E., 1985. Calculus I, 2nd edition. Undergraduate Texts in Mathematics. Springer-Verlag, New York. ISBN 978-0-387-90974-5.
McVinish, R., 2012. Improving ABC for quantile distributions. Stat. Comput. (ISSN 1573-1375) 22 (6), 1199–1207. doi:10/cgdzrt.
Mikkola, Petrus, Martin, Osvaldo A., Chandramouli, Suyog, Hartmann Oriol Abril Pla, Marcelo, Thomas, Owen, Pesonen, Henri, Corander, Jukka, Vehtari, Aki, Kaski, Samuel, Bürkner, Paul-Christian, Klami, Arto, 2021. Prior knowledge elicitation: the past, present, and future. arXiv:2112.01380 [stat].
Modrák, Martin, Moon, Angie H., Kim, Shinyoung, Bürkner, Paul, Huurre, Niko, Faltejsková, Kateřina, Gelman, Andrew, Vehtari, Aki, 2022. Simulation-based calibration checking for Bayesian computation: the choice of test quantities shapes sensitivity.

Muraleedharan, G., Lucas, C., Guedes Soares, C., 2016. Regression quantile models for estimating trends in extreme significant wave heights. Ocean Eng. (ISSN 0029-8018) 118, 204–215. https://doi.org/10.1016/j.oceaneng.2016.04.009.
Myerson, Roger B., 2005. Probability Models for Economic Decisions. Duxbury Applied Series. Thomson/Brooke/Cole, Belmont, CA. ISBN 978-0-534-42381-0, 978-0-534-42468-8.
Nadarajah, Saralees, 2009. Bathtub-shaped failure rate functions. Qual. Quant. (ISSN 1573-7845) 43 (5), 855–863. doi:10/d2rvr6.
Nair, N. Unnikrishnan, Sankaran, P.G., Vineshkumar, B., 2012. The Govindarajulu distribution: some properties and applications. Commun. Stat., Theory Methods (ISSN 0361-0926). https://doi.org/10.1080/03610926.2011.573168.
Nair, N. Unnikrishnan, Sankaran, P.G., Balakrishnan, N., 2013. Quantile-Based Reliability Analysis. Springer New York, New York, NY. ISBN 978-0-8176-8360-3, 978-0-8176-8361-0.
Nair, N. Unnikrishnan, Sankaran, P.G., Dileepkumar, M., 2020. Bayesian inference in quantile functions. Commun. Stat., Theory Methods (ISSN 0361-0926) 51 (14), 1–13. doi:10/ghkdr4.
O'Hagan, Anthony, Buck, Caitlin E., Daneshkhah, Alireza, Richard Eiser, J., Garthwaite, Paul H., Jenkinson, David J., Oakley, Jeremy E., Rakow, Tim, 2006. Uncertain Judgements: Eliciting Experts' Probabilities: O'Hagan/Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons, Ltd, Chichester, UK. ISBN 978-0-470-03331-9, 978-0-470-02999-2, 978-0-470-03330-2.
Olkin, I., Tate, R.F., 1961. Multivariate correlation models with mixed discrete and continuous variables. Ann. Math. Stat. (ISSN 0003-4851) 32 (2), 448–465. https://doi.org/10.1214/aoms/1177705052.
Parzen, Emanuel, 1979. Nonparametric statistical data modeling. J. Am. Stat. Assoc. (ISSN 0162-1459) 74 (365), 105–121. doi:10/gjh3sz.
Parzen, Emanuel, 1980. Data modeling using quantile and density-quantile functions. Technical report. Texas A & M Univ College Station Inst of Statistics.
Perepolkin, Dmytro, 2019. Qpd: tools for quantile-parameterized distributions.
Perri, Pier Francesco, Tarsitano, A., 2008. Distributional least squares based on the generalized lambda distribution. In: International Conference on Computational Statistics. COMPSTAT 2008, vol. 400. Physica-Verlag, Springer Company. ISBN 3-7908-2083-0, pp. 341–348.
Perri, Pier Francesco, Tarsitano, Agostino, 2007. Partially adaptive estimation via quantile functions. Commun. Stat., Simul. Comput. (ISSN 0361-0918) 36 (2), 277–296. https://doi.org/10.1080/03610910601158369.
Powley, Bradford W., 2013. Quantile Function Methods for Decision Analysis. Stanford University. ISBN 9798664735765.
Prangle, Dennis, 2017. Gk: an r package for the g-and-k and generalised g-and-h distributions. arXiv:1706.06889 [stat].
Press, William H. (Ed.), 2007. Numerical Recipes: The Art of Scientific Computing, 3rd ed edition. Cambridge University Press, Cambridge, UK, New York. ISBN 978-0-521-88068-8, 978-0-521-88407-5, 978-0-521-70685-8.
Price, G. Baley, 1984. The Inverse-Function Theorem. Springer New York, New York, NY, pp. 237–262. ISBN 978-1-4612-9747-5, 978-1-4612-5228-3.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rahman, Ataur, Zaman, Mohammad A., Haddad, Khaled, El Adlouni, Salaheddine, Zhang, Chi, 2015. Applicability of Wakeby distribution in flood frequency analysis: a case study for eastern Australia. Hydrol. Process. (ISSN 0885-6087) 29 (4), 602–614. doi:10/f6wzmh.
Ramberg, John S., Schmeiser, Bruce W., 1974. An approximate method for generating asymmetric random variables. Commun. ACM (ISSN 0001-0782) 17 (2), 78–82. https://doi.org/10.1145/360827.360840.
Rayner, G.D., MacGillivray, H.L., 2002. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Stat. Comput. (ISSN 1573-1375) 12 (1), 57–75. doi:10/c27574.
Reinhardt, Jason C., Chen, Xi, Liu, Wenhao, Manchev, Petar, Paté-Cornell, M. Elisabeth, 2016. Asteroid risk assessment: a probabilistic approach. Risk Anal. (ISSN 1539-6924) 36 (2), 244–261. https://doi.org/10.1111/risa.12453.
Ridders, C., 1979. A new algorithm for computing a single root of a real continuous function. IEEE Trans. Circuits Syst. (ISSN 0098-4094) 26 (11), 979–980. https://doi.org/10.1109/TCS.1979.1084580.
Säilynoja, Teemu, Bürkner, Paul-Christian, Vehtari, Aki, 2022. Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. Stat. Comput. 32 (2), 32. https://doi.org/10.1007/s11222-022-10090-6. ISSN 0960-3174, 1573-1375.
Schäling, Boris, 2011. The Boost C++ Libraries. Boris Schäling. ISBN 0-9822191-9-9.
Sharma, Dreamlee, Chakrabarty, Tapan Kumar, 2017. Some general results on quantile functions for the generalized beta family. Stat. Optim. Inf. Comput. 5 (4), 360–377. ISSN 2310-5070, 2311-004X, doi:10/gm28jc.
Sharma, Dreamlee, Chakrabarty, Tapan Kumar, 2019. The quantile-based flattened logistic distribution: some properties and applications. Commun. Stat., Theory Methods (ISSN 0361-0926) 48 (14), 3643–3662. https://doi.org/10.1080/03610926.2018.1481966.
Sharma, Dreamlee, Chakrabarty, Tapan Kumar, 2020. A quantile-based approach to supervised learning. In: Johri, Prashant, Verma, Jitendra Kumar, Paul, Sudip (Eds.), Applications of Machine Learning. Springer Singapore, Singapore, pp. 321–340. ISBN 9789811533563, 9789811533570.
Smithson, Michael, Shou, Yiyun, 2017. Cdf-quantile distributions for modelling random variables on the unit interval. Br. J. Math. Stat. Psychol. (ISSN 2044-8317) 70 (3), 412–438. doi:10/f9vhts.
Stage, Steven A., 2013. Comments on an improvement to the Brent's method. Int. J. Exp. Alg. 4 (1), 1–16.
Steel, Mark F.J., Rubio, Francisco J., 2015. On families of distributions with shape parameters: discussion. Int. Stat. Rev. (ISSN 1751-5823) 83 (2), 218–222. doi:10/gpc2db.
Su, Steve, 2007. Numerical maximum log likelihood estimation for generalized lambda distributions. Comput. Stat. Data Anal. (ISSN 0167-9473) 51 (8), 3983–3998. https://doi.org/10.1016/j.csda.2006.06.008.
Su, Steve, 2015. Flexible parametric quantile regression model. Stat. Comput. (ISSN 1573-1375) 25 (3), 635–650. https://doi.org/10.1007/s11222-014-9457-1.
Talts, Sean, Betancourt, Michael, Simpson, Daniel, Vehtari, Aki, Gelman, Andrew, 2020. Validating Bayesian inference algorithms with simulation-based calibration.
Tarsitano, Agostino, 2005. Fitting Wakeby model using maximum likelihood. In: Statistica e Ambiente, vol. 1. ISBN 88-7178-531-2, pp. 253–256.
Tarsitano, Agostino, 2010. Comparing estimation methods for the FPLD. J. Probab. Stat. 2010, 1–16. https://doi.org/10.1155/2010/295042.
Tukey, John W., 1965. Which part of the sample contains the information? Proc. Natl. Acad. Sci. 53 (1), 127–134. ISSN 0027-8424, 1091-6490, doi:10/dkvgcq.
Vega Yon, George, Marjoram, Paul, 2019. Fmcmc: a friendly MCMC framework. J. Open Sour. Softw. 4 (39). doi:10/gjvdh9.
Vehtari, Aki, Gelman, Andrew, Simpson, Daniel, Carpenter, Bob, Bürkner, Paul-Christian, 2021. Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). Bayesian Anal. (ISSN 1936-0975) 16 (2). doi:10/gh435z.
Vihola, Matti, 2012. Robust adaptive Metropolis algorithm with coerced acceptance rate. Stat. Comput. 22 (5), 997–1008. ISSN 0960-3174, 1573-1375, doi:10/bns84z.
Vineshkumar, Balakrishnapillai, Nair, Narayanan Unnikrishnan, 2019. Bivariate quantile functions and their applications to reliability modelling. Statistica (ISSN 1973-2201) 79 (1), 3–21. doi:10/gpdd87.
Wilson, Kevin James, 2018. Specification of informative prior distributions for multinomial models using vine copulas. Bayesian Anal. 13 (3), 749–766. https://doi.org/10.1214/17-BA1068. ISSN 1936-0975, 1931-6690.
Yu, Keming, Moyeed, Rana A., 2001. Bayesian quantile regression. Stat. Probab. Lett. (ISSN 0167-7152) 54 (4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9.
Zhang, Zhengqiu, 2011. An improvement to the Brent's method. Int. J. Exp. Alg. 2 (1), 21–26.

Paper III

Statistics and Computing (2024) 34:11
https://doi.org/10.1007/s11222-023-10325-0

ORIGINAL PAPER

Hybrid elicitation and quantile-parametrized likelihood

Dmytro Perepolkin[1] · Benjamin Goodrich[2] · Ullrika Sahlin[1]

Received: 4 October 2023 / Accepted: 8 October 2023

Abstract
This paper extends the application of quantile-based Bayesian inference to probability distributions defined in terms of quantiles of observable quantities. Quantile-parameterized distributions are characterized by high shape flexibility and parameter interpretability, making them useful for eliciting information about observables. To encode uncertainty in the quantiles elicited from experts, we propose a Bayesian model based on the metalog distribution and a variant of the Dirichlet prior. We discuss the resulting hybrid expert elicitation protocol, which aims to characterize uncertainty in parameters by asking questions about observable quantities. We also compare and contrast this approach with parametric and predictive elicitation methods.

Keywords Bayesian analysis · Quantile-parameterized distributions · Quantile-based distributions · Expert knowledge elicitation · Indirect inference

Mathematics Subject Classification 62C10 · 62F15 · 62G99

1 Introduction

1.1 Parametric and predictive approach to elicitation

Bayesian parametric inference is about updating prior beliefs about the model parameters in light of new observations. The underlying assumption is that an expert's prior knowledge (or lack thereof) can be translated into a subjective probability distribution of model parameters through the process of elicitation (Winkler 1967). The direct *elicitation of parameters* represents a *structural approach* to extracting an expert's knowledge (Kadane 1980). This approach requires that the expert comprehends the model and the role a specific parameter plays within it. Unfortunately, some parameters may be abstract, challenging to interpret (such as α and β parameters in the Gamma distribution), and at times not independent, as is the case with parameters in a hierarchical model.

Dmytro Perepolkin
dmytro.perepolkin@cec.lu.se

Ullrika Sahlin
ullrika.sahlin@cec.lu.se

[1] Centre for Environmental and Climate Science, Lund University, Lund, Sweden

[2] Applied Statistics Center, Columbia University, New York, USA

An alternative approach involves eliciting information about observable quantities, possibly conditioned on observable covariates (Kadane and Wolfson 1998), which may be more intuitive and relatable for experts. The *elicitation of predictions* aims to assess the expert's uncertainty regarding future observations (Gelman et al. 2013). Kadane and Wolfson (1998) advise against eliciting moments, with the exception of possibly the first moment (the arithmetic average). Instead, assessment should be carried out using quantiles or probabilities from the predictive distribution. The challenge with eliciting the predictive distribution is that it makes no distinction between the randomness explained by the model and the uncertainty about the parameters within it. Without this distinction, updating the expert predictions with the data coming from the new observations may be challenging.

While the non-Bayesian elicitation often stops at the quantiles or probabilities related to the expert's *predictive judgment* (Spetzler and Staël Von Holstein 1975; Morgan 2014; Keeney and von Winterfeldt 1991; Hanea et al. 2021), the Bayesian school of thought attempts to devise a method to infer the prior distribution, which could have led to the particular predictions expressed by the expert (Akbarov 2009; Hartmann et al. 2020; Winkler 1980; Kadane and Wolfson 1998; Bockting et al. 2023; Manderson and Goudie 2023). Kadane (1980) refer to this process of eliciting the predictive

Published online: 27 October 2023

distribution followed by inferring the prior as the *predictive approach*, as it leverages predictions to derive the distribution of parameters.

1.2 Aims of the paper

In this paper, we propose a **hybrid elicitation** approach, which combines the elicitation of observable quantities with the elicitation of the associated uncertainty. This method combines elements of both the predictive and structural approaches to elicitation and can be employed to establish the prior distribution for a model defined by a quantile-parametrized distribution.

Quantile-parameterized distributions (QPDs) (Keelin and Powley 2011; Hadlock 2017) are parameterized by a set of quantile-probability pairs describing a random variable. As a result, the parameters in a QPD are measured on the same scale as the random variable they represent. These distributions can be utilized to model either uncertainty about future observations (predictive distribution) or the distribution of an unobservable parameter (prior distribution). For a comprehensive review of quantile-parameterized distributions, we refer to Perepolkin et al. (2023b).

Until now there has been, to the best of our knowledge, no published research on how to update the quantile parameters of a QPD in light of the new observations. This paper extends the principles of *quantile-based Bayesian inference* (Perepolkin et al. 2023a) to models parameterized by quantiles and proposes a prior distribution capable of capturing uncertainty in the quantile parameters. The proposed approach enables the elicitation and Bayesian updating of a variable quantity with minimal assumptions about the underlying model structure.

1.3 Paper structure

Section 2 introduces the method of *quantile-based inference* as proposed by Rayner and MacGillivray (2002) and Nair et al. (2020), and summarized in Perepolkin et al. (2023a). This method of inference is related to using one of the quantile-based distributions (Perepolkin et al. 2023b), which lack an explicit distribution function (CDF) and probability density function (PDF), as either prior or likelihood components within a Bayesian model. Quantile-based priors and likelihoods rely on substitutions derived from the inverse distribution function, known as the quantile function (QF).

In Sect. 3, we delve into a subclass of quantile-based distributions parameterized by sets of quantile-probability pairs (Fig. 1). We provide a brief overview of the literature concerning different methods for constructing quantile-parameterized distributions (QPDs). Our particular focus is on the quantile-based quantile-parameterized metalog distribution (Keelin 2016), chosen for its parameter flexibility.

In Sect. 4 we introduce the Bayesian model in which the likelihood is expressed using the metalog distribution, parameterized by a set of quantile-probability pairs. We demonstrate how uncertainty in these parameters can be specified through a variant of the Dirichlet distribution. Section 5 describes the elicitation of the QDirichlet prior and introduces a novel *hybrid* elicitation process for obtaining quantile-probability pairs along with the uncertainty associated with them. We illustrate our approach with excerpts from a hypothetical interview. The accompanying `qpd` R package (Perepolkin 2019), implements several quantile-parameterized distributions and includes functionality for supporting the elicitation of the Dirichlet and Connor–Mosimann distributions (Elfadaly and Garthwaite 2013).

Section 6 discusses the MCMC-based algorithm used for updating parameters in a quantile-parameterized distribution. In this paper, we employ Hamiltonian Monte Carlo algorithm in Stan, interfaced by the `cmdstanr` package in R (Gabry and Češnovar 2022). An alternative implementation using the Robust Adaptive Metropolis algorithm by Vihola (2012), interfaced by the `fmcmc` package (Vega Yon and Marjoram 2019), is available in the Supplemental Materials. The models proposed in this paper have been validated using Simulation-Based Calibration (Cook et al. 2006; Modrák et al. 2022; Talts et al. 2020). The results of the simulation studies (provided in Appendix C in Supplemental Materials) demonstrate the successful recovery of the parameter values for all widths of the posterior credible intervals.

We conclude the paper by discussion and summary of the results in Sect. 7.

2 Quantile-based Bayesian inference

The use of non-invertible quantile-based distributions as either a likelihood (Rayner and MacGillivray 2002; King 1999) or a prior (Nair et al. 2020) is not a novel concept in scientific literature. Rayner and MacGillivray (2002) described a three-step process for computing the log-likelihood of a quantile-based distribution. They applied this method to estimate the parameters of the *g-and-k* and generalized *g-and-h* distributions using maximum likelihood estimation. Similarly, Nair et al. (2020) employed quantile function substitutions to express both prior and likelihood in a quantile form. They calculated the posterior Bayes estimator of the parameters in the Govindarajulu model with uniform and generalized exponential priors. Perepolkin et al. (2023a) summarized the approaches to quantile-based inference and provided several examples of applying the principles of inference with quantile functions in both univariate and regression settings.

For a random sample $\underline{x} = \{x_1, x_2, \ldots x_n\}$, the posterior distribution of θ over the parameter space Θ can be summarized as:

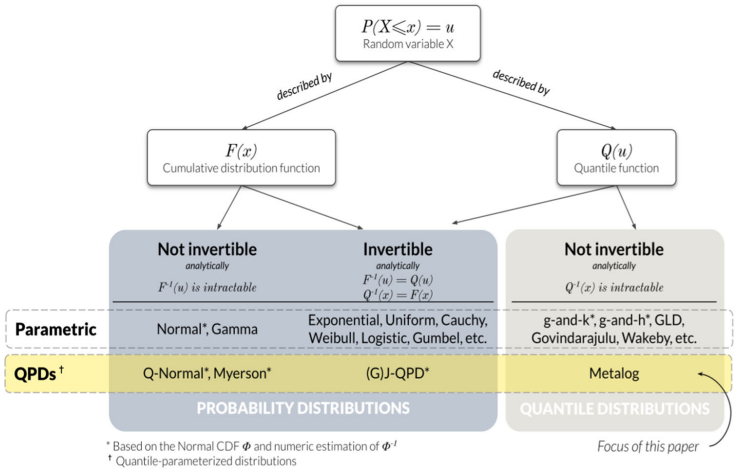


Fig. 1 Probability distributions, quantile-based distributions and parameterization by quantiles

$$f(\theta|\underline{x}) \propto \mathcal{L}(\theta; \underline{x}) f(\theta) \tag{1}$$

where $f(\theta|\underline{x})$ is the posterior distribution of θ after having observed the sample $\underline{x}$, $f(\theta)$ is the prior distribution of θ, and $\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ is the *density-based* form of the likelihood.

Consider a set of conditional probabilities $\underline{u} = \{u_i|\theta\} = \{F(x_i|\theta)\}$, $i = \{1, 2, \ldots n\}$, corresponding to the sample $\underline{x}$ of the observable x with distribution function F given the parameter θ, called *depths*. The conditional probabilities $\underline{u}$ are degenerate random variables that are entirely determined given the observations $\underline{x}$ and the value of the parameter θ. They are called *depths*, because they indicate how "deep" a particular observation is within the distribution. Using the depths $\underline{u}$, we can calculate $\underline{Q} = \{Q_1(u_1), Q_2(u_2), \ldots Q_n(u_n)|\theta\}$, where $Q(u|\theta) = F^{-1}(u|\theta)$ represents the quantile function or inverse cumulative distribution function (CDF). Since $Q(u_i|\theta) = x_i$, we can substitute $\underline{Q}$ for $\underline{x}$, and the Bayesian inference formula (1) becomes:

$$f(\theta|\underline{Q}) \propto \mathcal{L}(\theta; \underline{Q}) f(\theta) \tag{2}$$

We refer to this form of the likelihood $\mathcal{L}(\theta; \underline{Q}) = \prod_{i=1}^{n} f(Q(u_i|\theta)) = \prod_{i=1}^{n} [q(u_i|\theta)]^{-1}$ as *quantile-based* because it relies on the calculation of intermediate depths $u_i = F(x_i|\theta)$, $i = \{1, 2, \ldots n\}$. Here $[q(u_i|\theta)]^{-1}$ is reciprocal to the derivative of the quantile function $Q(u_i|\theta)$ called the *density quantile function* (Perepolkin et al. 2023a).

Both forms of the likelihood, $\mathcal{L}(\theta; \underline{Q})$ and $\mathcal{L}(\theta; \underline{x})$, are equivalent and yield the same posterior beliefs about the parameter θ (Perepolkin et al. 2023a).

3 Quantile parameterization of distributions

In this section, we consider a special class of distributions where the parameters are specified by quantile-probability pairs (Fig. 1), and see how the concept of *quantile-based inference* (Perepolkin et al. 2023a) can be applied to these distributions, as well.

A set of n quantile-probability pairs, denoted as $\mathsf{S} = \{(p_i, q_i)\}, i = \{1, 2, \ldots n\}$, can be thought of as comprising a pair of ordered vectors: a vector of probabilities p and a vector of quantiles q, with $p = \{p_1, \ldots p_n\}$, $p_i \in [0; 1]$, and $q = \{q_1, ..q_n\}$, $i = \{1, 2, \ldots n\}$. As CDF $F(x) = p$ is a non-decreasing function, the vectors p and q are considered properly ordered iif $q_i \leq q_{i+1}, \forall q_i \in q$, and $p_i \leq p_{i+1}, \forall p_i \in p$. Additionally, the quantile-probability pairs within the set S are considered distinct iff $\forall\{(p_i, q_i)\} \in \mathsf{S}, \exists!\{(p_j, q_j)\} = \{(p_i, q_i)\}$, $j \neq i$, $i = \{1, 2, \ldots n\}$, $j = \{1, 2, \ldots n\}$. In this paper, we refer to the set of n distinct, properly ordered quantile-probability pairs as a size-n *quantile-probability tuple* (QPT) denoted by $\{p, q\}_n$.

3.1 SPT-parameterization

A recent review (Perepolkin et al. 2023b) describes two methods for constructing distributions parameterized by quantile-probability pairs:

- By reparameterization of existing distributions, or
- Through an optimization step, where the distribution parameters are mapped to quantiles using least squares or similar algorithms.

Distributions falling under in the first category are typically parameterized by the *symmetric percentile triplet* (SPT), a QPT of size 3. In the SPT, the middle cumulative probability $p_2 = 0.5$ represents the median, while $p_1 = 1 - p_3 = \alpha,\ \alpha \in (0, 0.5)$ (e.g. $\{0.25, 0.50, 0.75\}$ or $\{0.10, 0.50, 0.90\}$). Examples of SPT-parameterized QPDs include the Myerson distribution (Myerson 2005), the Johnson Quantile-Parameterized distribution (J-QPD) (Hadlock and Bickel 2017), and their generalizations (Perepolkin et al. 2023b; Hadlock and Bickel 2019). A special case of SPT-parameterization also exists for the metalog distribution (Keelin 2016). For the purposes of this paper, we do not consider SPT-parameterized QPDs, including the SPT-metalog, Myerson, or J-QPD, due to their strict requirement for symmetric probability parameterization.

3.2 Parameterization using implicit functions

Keelin and Powley (2011) and Powley (2013) introduce an alternative method of parametrizing a distribution by a set of quantile-probability pairs. This method relies on the finite Taylor series expansion of parameters within a known quantile function as linear functions of the cumulative probability p.

The authors created the Simple Q-Normal (SQN) distribution by taking the quantile function of a normal distribution $x \equiv \mu + \sigma\Phi^{-1}(p)$, and making the parameters μ and σ functions of p; specifically, $\mu(p) = a_1 + a_4 p$ and $\sigma(p) = a_2 + a_3 p$. In a similar vein, Keelin (2016) proposed the metalogistic (*metalog*) distribution by making the parameters μ and s in the logistic quantile function $x \equiv \mu + s\,\text{logit}(p)$ be the functions of p, i.e. $\mu = a_1 + a_4(p-0.5) + a_5(p-0.5)^2 + \ldots$ and $s = a_2 + a_3(p-0.5) + a_6(p-0.5)^2 + \ldots$. Here, μ represents the mean, s is proportional to the standard deviation such that $\sigma = s\pi/\sqrt{3}$, $\text{logit}(p) = \ln(p/(1-p))$ is the log-odds of probability $p \in [0, 1]$ and $a_i,\ i = \{1, 2, \ldots n\}$ are real constants.

In both cases, the quantile function $Q(p)$ whose parameters also depend on p is an *implicit function*. This means that such a quantile function cannot be simply computed for arbitrary values of p. Nevertheless, with a set of n quantile-probability pairs, it is possible to determine the constants $a_i,\ i = 1, 2 \ldots n$, by solving a system of n linear equations (Keelin and Powley 2011; Powley 2013). This system can be represented as the matrix Equation (3).

$$a = \mathbb{P}^{-1} q \tag{3}$$

Keelin and Powley (2011) show the conditions under which a size-n QPT $\{p, q\}_n$ can uniquely determine the constants $a = \{a_1, \ldots a_n\}$. Additional details of the metalog distribution, including the composition of the matrix $\mathbb{P}$, can be found in Appendix A in Supplemental Materials.

The shape flexibility of the QPD increases with the number of terms added to the finite Taylor expansion of parameters within the parent distribution. To estimate the coefficients for the n-term quantile-parameterized distribution $a = \{a_1, \ldots a_n\}$, a minimum of n quantile-probability pairs is required. The order of the terms, denoted as n, is constrained by the size of the parameterizing QPT m, $(n \leq m)$, and concerns for overfitting. The QPT used for parameterizing a distribution can be obtained through expert elicitation or from the empirical CDF (ECDF), which is constructed from a sample of observations. The ECDF begins at zero and increments by $1/m$ at each of the m data points in the sample, representing the fraction of observations that are less than or equal to the specified value (Wasserman 2006).

Depending on the relationship between the size of the parameterizing QPT m and the number of terms n in the QPD QF we use the following terminology:

- When the size of the parameterizing QPT m equals the number of terms n in the QPD QF, i.e. $m = n$, we refer to the process of estimating the vector of coefficients $a = \{a_1 \ldots a_n\}$ as "fitting", and we call the resulting QPD "properly parameterized". In properly parameterized QPDs, the QF curve is guaranteed to pass through every QPT point. We label the n-term metalog parameterized by the n-size QPT $\{p, q\}_n$ as the **proper n-metalog**.
- When the size of the parameterizing QPT m exceeds the number of terms n in the QPD's QF (for example, when the QPT is derived from the sample ECDF and $m > n$), we refer to the process of estimating the vector of coefficients $a = \{a_1 \ldots a_n\}$ as "approximating", and we call the resulting QPD "approximated". Such approximation is typically achieved through optimization or regression, and the resulting QF curve is no longer guaranteed to pass through every QPT point. We designate the n-term metalog parameterized by the m-size QPT $\{p, q\}_m,\ m > n$ as the **approximate n-metalog**.

Given the matrix Equation (3), we have two alternative parameterizations for the proper n-metalog: it can either be directly parameterized by the coefficients $a = \{a_1, \ldots a_n\}$ (referred to as the A-parameterization) or indirectly parameterized by a QPT $\{p, q\}_n$ (referred to as the QPT-parameterization). Therefore, in this paper, when we mention the proper n-metalog, the notations $Q_{M_n}(u|a)$ and $Q_{M_n}(u|p, q)$ (where $u|a$ or $u|p, q$ represents the *depths* corresponding to the observation x) are used interchangeably. In the case where the metalog is approximated, only the A-parameterization is suitable because the number of metalog terms n is not determined by the m data points from the ECDF used to estimate the parameter vector a via Equation (3).

4 QDirichlet-metalog model

In this section, we introduce a Bayesian model with the likelihood defined by the proper n-term metalog. Since the metalog is a quantile-based distribution (Fig. 1), we employ the *quantile-based likelihood* following Equation (2). The quantile-based likelihood relies on the intermediate depths $\underline{u}|\theta$, which correspond to the sample of observations $\underline{x}$. Since a closed-form CDF for the metalog distribution is not available, we resort to numerical approximation, denoted as $u = \widehat{Q}_X^{-1}(x)$ (Perepolkin et al. 2023a).

4.1 Parameter uncertainty

To account for the uncertainty in the QPT parameters of a quantile-parameterized likelihood, such as the metalog, we must introduce uncertainty either in cumulative probabilities or the corresponding quantile values (or both). Coles and Tawn (1996) specified the prior for an extreme value model in terms of the quantile values for certain fixed cumulative probability values. Crowder (1992) suggested that the prior can be constructed based on the space of probabilities, with fixed quantiles. In a recent paper on predictive elicitation, Hartmann et al. (2020) divided the observable space into several exhaustive and mutually exclusive categories and asked experts to assign probabilities that the next observation falls into each of the categories, treating these probabilities as uncertain. They assigned a Dirichlet prior to these probability judgments.

We follow a similar approach by using the quantile values provided by the expert to partition the outcome space. We then characterize the uncertainty in the corresponding cumulative probabilities using the Dirichlet distribution (together referred to as the *QDirichlet* prior). Our method of constructing a prior distribution for the simplex Δ shares similarities with the approach adopted by Bürkner and Charpentier (2020) for modelling monotonic effects in ordinal regression. The parameter vector of the Dirichlet distribution, in conjunction with the vector of elicited quantiles, serves as hyper-parameters for the proposed QDirichlet prior, which captures the uncertainty in the parameters of the quantile-parameterized model.

4.2 The QDirichlet prior

Consider a size-n QPT $\{p, q\}_n$, consisting of a vector of probabilities p and a vector of quantile values q. Now, consider an extended vector of probabilities $b = \{0, p, 1\}$ of size $n + 2$, containing the vector p. Additionally, consider a forward difference $\Delta = \{\Delta_1 \dots \Delta_{n+1}\}$, where $\Delta_i = b_{i+1} - b_i$, $i = 1, 2, \dots (n + 1)$, which is a simplex of size $n + 1$. The simplex Δ is properly ordered iif it is based on the properly ordered vector b, and consequently, also p.

To transform the simplex Δ back into the vector of probabilities p, the cumulative sum $\Xi_1^n()$ can be used, so that $p = \Xi_1^n(\Delta) : \ p_j = \sum_{i=1}^{j} \Delta_i, \ j \in (1 \dots n)$, assuming the simplex Δ is properly ordered. If, for any reason, the simplex Δ can no longer be considered *properly ordered*, we can use an index vector of distinct values, denoted as $I = \{I_1 \dots I_n\} : I_j = \{1, 2, \dots n\}, \ j = \{1, 2, \dots n\}, \ \exists! I_j = I_i, \ j \neq i$. This index vector can be used to restore the proper order before accumulating the simplex Δ into the probability vector p.

To express prior uncertainty in the simplex Δ, we can use the Dirichlet distribution (Johnson et al. 1997) with a hyperparameter vector α of size $n + 1$, conditional upon the specified quantile values q. We refer to this particular variant of the Dirichlet prior as the *QDirichlet* prior, as its parameter vector α is specified in relation to the fixed quantile values q.

4.3 The metalog likelihood

We adopt the notation for quantile-based likelihoods introduced in Perepolkin et al. (2023a), where $u \overset{x}{\backsim} \dots$ should be read as "the depths u corresponding to the random variable x inversely distributed as …". Consequently, QDirichlet-Metalog model can be expressed as follows:

$$\begin{aligned} u &\overset{x}{\backsim} \text{Metalog}(p, q) \\ \Delta &\sim \text{Dirichlet}(\alpha|q); \\ p &= \Xi_1^n(\Delta); \end{aligned} \tag{4}$$

where Δ is a simplex of size $n + 1$, $\Xi_1^n()$ is the cumulative sum operator, p is a size-n vector of cumulative probabilities and q is the corresponding size-n vector of quantiles. Furthermore, u is the depth corresponding to the observable x given the parameterizing QPT $\{p, q\}$. The depths u can be computed (typically numerically) by inverting the quantile function $\widehat{Q^{-1}}(x|p, q)$. The metalog quantile function is indirectly parameterized by the QPT $\{p, q\}_n$ through the vector of metalog coefficients a, determined by the matrix Equation (3).

In Model (4), the prior is represented by the Dirichlet distribution with hyperparameter α specifying the uncertainty in the cumulative probabilities and a vector q representing the quantile values corresponding to the sampled cumulative probabilities (QDirichlet prior). The metalog (quantile-based) likelihood parameterized by the QPT $\{p, q\}_n$ relies on depths u which can be estimated using the numerical inverse of the metalog quantile function (Perepolkin et al. 2023a).

4.4 Eliciting Dirichlet distribution

Elfadaly and Garthwaite (2013) describe a method for inferring the parameter vectors of the Dirichlet and Generalized Dirichlet (Connor–Mosimann) distributions from the conditional univariate beta distributions. In this method, the expert assesses the quartiles of the probability for each category using the elicitation of the symmetric percentile triplet (SPT) as follows:

1. The expert assesses the probability quartiles for the first category p_1.
2. The expert is then asked to assume that the median value they provided in the assessment of p_1, is in fact the correct probability (true value) for the first category.
3. Next, the expert proceeds to assess the SPT for the next category, conditional upon the previous assessment, denoted as $p_2|p_1$.
4. The three quartiles of $p_2|p_1$ are divided by $1 - p_1$ to normalize them, i.e. $p_2^*|p_1$.
5. The hyperparameters of the beta distribution representing $p_2^*|p_1$ are the determined.
6. Steps 3–5 are repeated for all categories, except the last one.

Elfadaly and Garthwaite (2013) also propose an improvement to Pratt et al. (1995)'s method of fitting the beta distribution to the elicited conditional SPTs based on the normal approximation described in Patel and Read (1996). The α and β parameters of the conditional beta-distributions are then normalized and the hyperparameter vector $\alpha = \{\alpha_1 \ldots \alpha_{n+1}\}$ is estimated.

The elicitation method proposed by Elfadaly and Garthwaite (2013) can be applied to assess the uncertainty in the cumulative probabilities that parametrize the metalog likelihood. If elicitation starts with the left tail of the distribution, the first category will coincide with the first cumulative probability p_1 in the parameter vector p. Subsequent (higher) cumulative probabilities will always be conditional upon and include the median value of the lower probability. If the elicitation is performed out of order, which might be expedient to avoid anchoring effects (Spetzler and Staël Von Holstein 1975; Abbas et al. 2008), the integer index vector I of the same size can be provided along with the results of the assessment to restore the proper ordering of the simplex Δ after sampling and before its accumulation into the vector of probabilities p.

Note that the parameter vector p in the model (4) is not independent: it is paired with the vector of fixed quantile values q. There are several approaches to specifying the hyperparameter vector q.

- **Predictive distribution**. The vector q could be coming from the characterization of the prior predictive distribution. In this scenario, we could ask the expert to specify their uncertainty regarding the next observation using standard predictive elicitation techniques described in Morgan et al. (1990) or Spetzler and Staël Von Holstein (1975). Predictive elicitation results in the QPT $\{p^*, q^*\}_n$, of which vector q^* can be adopted as the true values of q, while the vector of cumulative probabilities p^* can serve as the initial values for the MCMC/HMC algorithm.
- **Hypothetical sample**. Alternatively, the vector q could be viewed as a *representative sample* from the predictive distribution. Randomly sampling the predictive distribution has the advantage that the values closer to the distribution's mode are more likely. However, if the sampled values of q are too closely spaced, fitting the Metalog to the QPT $\{p, q\}_n$ within the MCMC loop may become challenging.

The primary goal of eliciting the vector q is to *position* the prior on the data (x) scale and provide a reasonable baseline for the follow up elicitation. In fact, the hyperparameter vector q specifies the *location* of the QDirichlet prior, while the hyperparameter vector α is responsible for defining its *shape*.

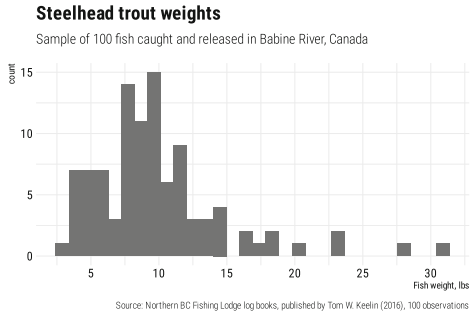
5 Applications

In this section, we provide an example of *hybrid* elicitation for parameterizing the QDirichlet prior to describe the uncertainty in the $\{p, q\}_n$ parameters of the proper metalog.

5.1 Steelhead trout weights

We take a sample of 100 observations from the records of steelhead trout weights captured and released in Babine River, Canada, spanning the years of 2006-2014 (Fig. 2). The dataset has been published by Keelin (2016) and is also included in the `rmetalog` (Faber and Jung 2021) package, which is accessible on CRAN.

Our goal is to elicit prior beliefs regarding the distribution of fish weights from a hypothetical expert and subsequently update those beliefs in light of the sampled data. Sections 5.2 and 5.3 outline the elicitation process and provide details on the required diagnostics for prior specification and posterior inference.



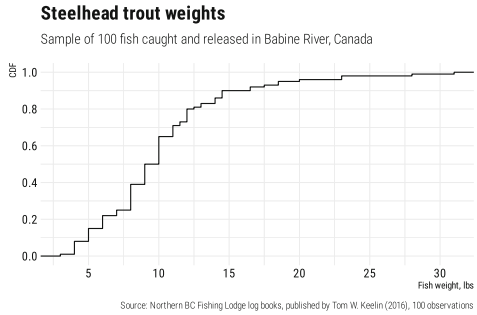


Fig. 2 Summary of the sample from the steelhead trout dataset

Table 1 Quantile-probability pairs for the predictive distribution of fish weights

Cumulative probability, p*	Weight in lbs, q*
0.1	4
0.5	9
0.9	17

5.2 Specifying location of the prior

The *hybrid* elicitation involves of two phases: the elicitation of quantile values q and the elicitation of uncertainty in the associated cumulative probabilities (i.e. potential vectors of p that could correspond to the specified q).

Imagine conducting an interview with an experienced fly-fisherman to gather information about steelhead trout weights in Canadian rivers. We begin with the elicitation of the predictive QPT using the probability-value (PV) method (Spetzler and Staël Von Holstein 1975; Abbas et al. 2008). After guiding the expert through essential preparatory steps (motivating, structuring, conditioning, encoding and verification), we elicit the following predictive QPT $\{p^*, q^*\}_3$ (Table 1).

Physical weight can be represented by non-negative values, suggesting the use of a distribution bounded on the left. To model this, we employ a semi-bounded log-metalog for the predictive QPT values (Fig. 3). Notably, the three cumulative probabilities p^* provided by the expert divide the y-axis of the CDF into four distinct bands (highlighted by different colors in Fig. 3). These bands can be seen as categories into which a weight of a randomly drawn fish could fall on the empirical CDF curve. Similar to the approach taken in Hartmann et al. (2020), we use the elicited quantile values q^* to partition the outcome space into the exhaustive and mutually exclusive categories. The widths of the bands correspond to the increments in cumulative probabilities provided by the expert, as represented by the simplex Δ in the model (Equation (4)).

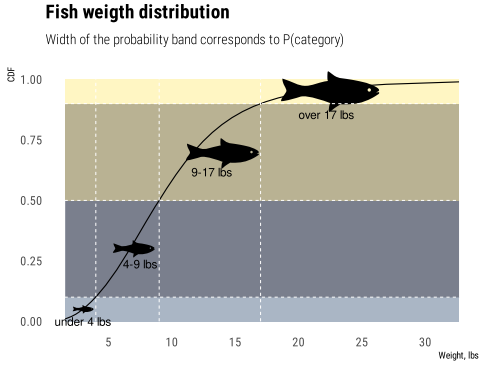


Fig. 3 Probability bands corresponding to the four fish size categories

Table 2 SPT for the count of the small fish (less than 4 lbs)

	Cumulative probability	Fish count
Small fish	0.25	70
	0.50	90
	0.75	120

All fish counts are out of the sample of 1000

In the second step of the *hybrid* elicitation we elicit uncertainty regarding these probability band widths Δ, using the values q as reference points to delineate the categories to which the random variate $p_i \in p,\ i = \{1, 2, \ldots n\}$ would be associated.

5.3 Hybrid elicitation of the QDirichlet prior

The predictive QPT $\{p^*, q^*\}_n$ we elicited earlier does not inherently incorporate uncertainty (except for potential

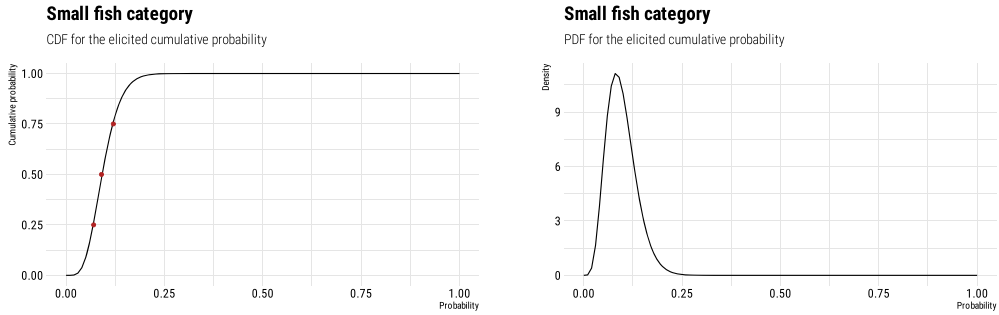


Fig. 4 Conditional beta distribution fitted to the quartiles provided by the expert

imprecision or inconsistencies in the expert's expression of belief). During this phase, we ask the expert to consider uncertainties surrounding their quantile assessments, aiming to distinguish aleatory from epistemic uncertainties (Knight 1921). A possible drawback of commencing with predictive elicitation, as we did above, is that the specified vector p^* might anchor the expert's belief to these values, potentially affecting the range of probabilities that the expert would associate with the vector q. The expert might be inclined to allow only a small and likely symmetrical variation around the initially specified p^* values, hesitating to revise their judgments. On the other hand, the specified p^* values can serve as a starting point for discussion with the expert, validating their beliefs, and deliberately challenging the expert to reassess them.

The term "aleatory" signals the use of the sampling frame, prompting us to shift from assessing the properties of population to the evaluating the properties of an imaginary sample. In this elicitation phase, we transform the cumulative probabilities into natural frequencies (Gigerenzer 2011), treating the values of $p_i \in p,\ i = \{1, 2, \ldots n\}$ as proportions within a hypothetical large sample.

Recall that the expert has supplied us with a predictive QPT $\{p^*, q^*\}_3$, where $p^* = \{0.1, 0.5, 0.9\}$ and $q^* = \{4, 9, 17\}$ (Table 1).

Interviewer:
Consider a large sample of steelhead trout caught in British Columbia over the past few years, let's say 1000 fish. Based on your assessment, it's expected that around 100 fish would weigh less than 4 lbs.

The elicited predictive QPT is interpolated with a log-metalog and presented in Fig. 3. Considering the sampling frame, the curve we've drawn through the three points provided by the expert is just one of numerous empirical CDF curves that could be constructed given the inherent sampling uncertainty.

Table 3 Conditional SPTs for the counts of the fish in the 1000 fish sample

	Category*	P25	P50	P75
1	Small fish	0.07	0.090	0.12
2	Medium fish	0.34	0.410	0.50
4	Huge fish	0.05	0.075	0.15

*Small fish is under 4 lbs, Medium fish is 4–9 lbs, Huge fish is over 17 lbs

Table 4 Dirichlet parameter vector

	Category	a
1	Small fish	3.77
2	Medium fish	12.86
4	Huge fish	2.70
3	Large fish	10.72

Table 5 Connor–Mosimann parameter vectors

	Category	a	b
1	Small fish	5.87	55.24
2	Medium fish	6.77	8.01
4	Huge fish	1.32	5.25

We proceed with the elicitation by asking the expert to contemplate the fish weight cutoff of 4 lbs.

Interviewer:
Let's delve into this hypothetical sample of 1000 fish. According to your assessment, there should be approximately 100 fish weighing less than 4 lbs. We will interpret this assessment as you believing that there's about equal chance that the actual number of "small" fish (weighing less than 4 lbs) in this sample is either above or below 100. In essence, we interpret it as the median assessment. Would you like to reconsider this value?

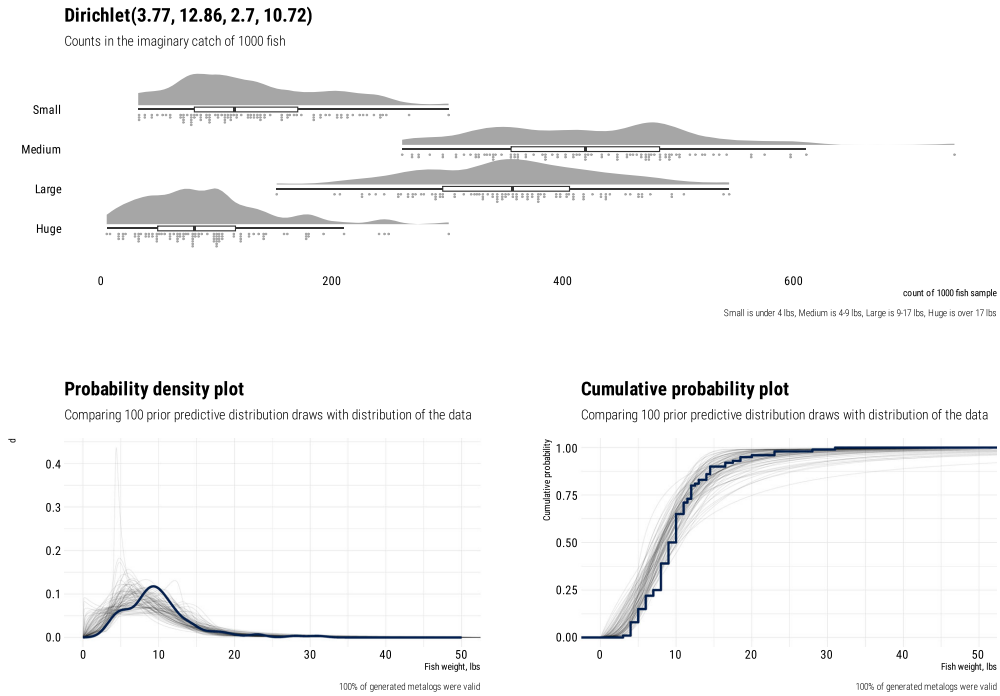


Fig. 5 Prior predictive check for Dirichlet distribution

At this stage, the expert may choose to adjust their assessment of the median. Once the median value of the first category is confirmed, we can proceed with the elicitation of the range around it. We follow the conventional *fixed probability* encoding method (Abbas et al. 2008; Spetzler and Staël Von Holstein 1975) asking the expert about the range of fish counts in the sample (which actually represent cumulative probabilities) corresponding to the quartiles or the 10th/90th percentile.

Suppose, the expert has furnished us with the revised median and the 50% Interquartile Range (IQR) around the initially assessed probability $p_1^* = 0.1$ for the count of "small" fish in the hypothetical sample of 1000, as summarized in Table 2.

From this information, we can promptly deduce the uncertainty in the "width" of our first bin. It is now characterized by a symmetric percentile triplet $\{0.07, 0.09, \text{and } 0.12\}$ with $\alpha = p_1 = 1 - p_3 = 0.25$. Employing this SPT, we can fit the beta distribution (Fig. 4) using the method proposed in Elfadaly and Garthwaite (2013).

We then proceed to with the conditional elicitation of probabilities for the remaining fish weight categories and uncertainties associated with them, conditional on the median values of the previously elicited categories. During this phase, we ask the expert:

Interviewer:
Let's assume that in the sample of 1000 fish, precisely 90 were found to be small (weighing less than 4 lbs). What would be your estimate of the number of fish that would fall within the weight range of 4 to 9 lbs in such a sample?

In this question, we are soliciting the expert's input for a conditional probability distribution. Therefore, we do not hold the expert accountable for their previous assessment, where they implied that approximately half of the population would weigh 9 lbs or less (as suggested by the cumulative probability of 0.5). We anticipate that the median count would be close to 500 fish, but not necessarily an exact match. Our aim is to elicit the count of fish weighing between 4 and 9 lbs. However, if the expert prefers to provide us with the count corresponding to the "exceedance probability" (i.e., for 2 categories combined), we should subtract the median count of the first category, which is 90.

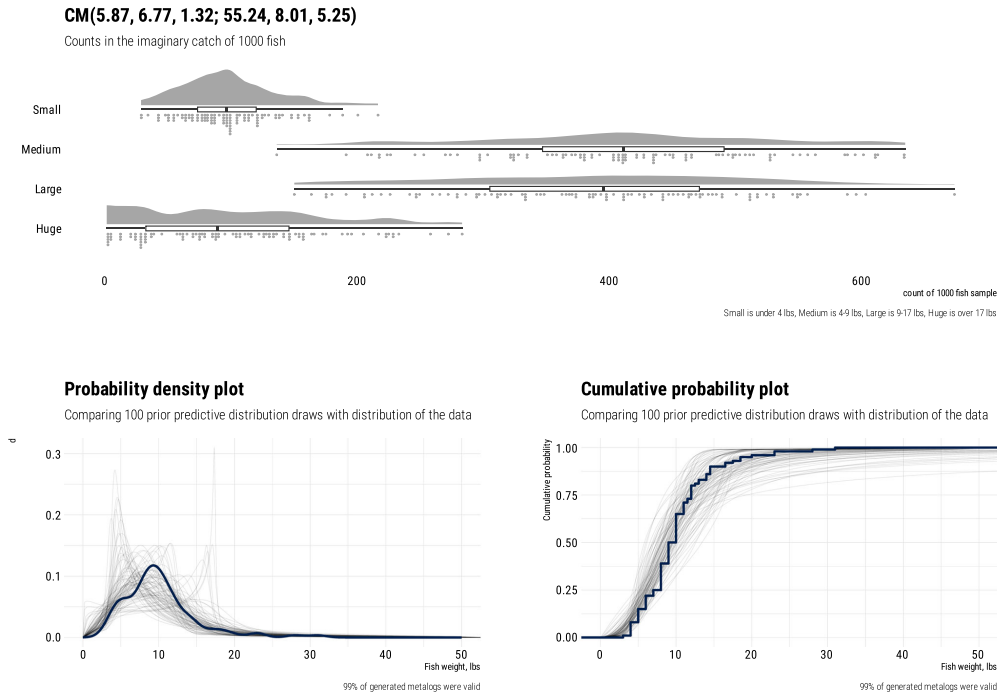


Fig. 6 Prior predictive check for Connor–Mosimann distribution

For a total of N groups, we should only need to conduct $N - 1$ elicitations (a total of 3 elicitation of triplets in our case). It might be convenient to elicit the top quantile (representing the upper tail of the CDF) as $1 - \sum_{i=1}^{3} p_i$) and leave the quantiles for the "Large fish" category (fish weighing between 9 and 17 lbs) to be calculated from the remaining information.

Let's assume that, after eliciting the three conditional SPTs and converting the hypothetical sample counts to probabilities, we obtain the following assessments (Table 3). Note, that the assessments for the category 3 Large fish are missing from the table. They are implied (and will be inferred from) the rest of the data.

5.4 Fitting Dirichlet and Connor–Mosimann distributions

We can utilize the elicited conditional SPTs to derive parameter vectors for the Dirichlet (Table 4) or Connor–Mosimann (Table 5) distributions following the process in Elfadaly and Garthwaite (2013). The algorithm for transforming the conditional SPTs into parameter vector(s) is implemented in the `qpd` R package (Perepolkin 2019).

The Dirichlet distribution defines a strong negative dependence between the elements of the simplex Δ, meaning that an increase in the probability of one element necessarily decreases the probability of every other element (Balakrishnan 2014). The Connor–Mosimann distribution relaxes this assumption of a strong negative correlation between the categories, allowing for a more flexible encoding of dependence between the quantiles (Wilson 2017).

5.5 Prior predictive check

Prior predictive checks are crucial for providing the expert with feedback on the elicited values and diagnosing potential issues (Gabry et al. 2019). Since uncertainties in the quantile probabilities were elicited as conditional probabilities, it is important to show to the expert the impact of the provided probability ranges on the overall multivariate distribution. This can be accomplished, for example, using marginal plots. We can draw samples from the Dirichlet distribution (Fig. 5) or the Connor–Mosimann distribution (Fig. 6) and present

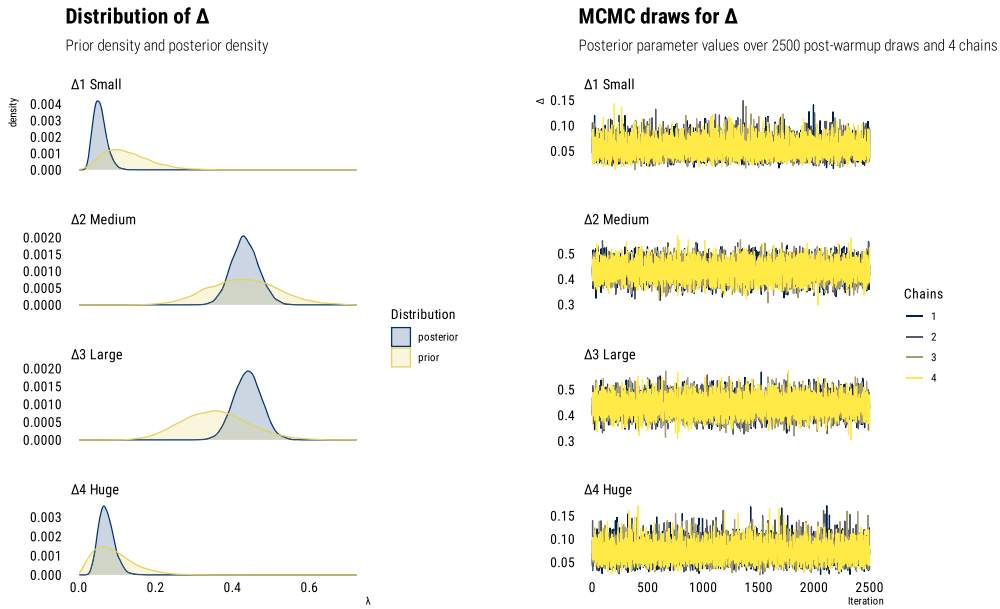


Fig. 7 Summary of the posterior draws for Δ simplex

the expert with an overview of the parameter distribution in the same format that will be used for the posterior predictive check (Fig. 8 in Sect. 6).

6 MCMC implementation

To sample from the posterior distribution of Δ,we employed the Hamiltonian Monte Carlo (HMC) algorithm in Stan, interfaced via the `cmdstanr` package (Gabry and Češnovar 2022) in R. An alternative implementation using the Robust Adaptive Metropolis algorithm by Vihola (2012), implemented in the `fmcmc` package (Vega Yon and Marjoram 2019), is provided in the Supplemental Materials

We validated the QDirichlet-Metalog model using the Simulation-Based Calibration algorithm (Cook et al. 2006; Modrák et al. 2022; Talts et al. 2020). As evident from the diagnostic plots in the Supplemental Materials Appendix C, the parameter Δ is successfully recovered for all widths of the posterior credible interval.

We have also performed Simulation-Based Calibration for the model with QCM (Generalized Dirichlet) prior. The diagnostic plots also indicate the successful recovery of the parameter vector for all widths of the posterior credible interval.

To fit the QDirichlet-Metalog model, we used 2500 post-warmup iterations and 4 chains. The posterior distribution of the parameter Δ is presented in Table 6 and Fig. 7. The results reveal a significant reduction in the uncertainty regarding the cumulative probabilities corresponding to the quantile values of 4, 9, and 17 lbs. Specifically, the lowest value 4 lbs corresponds to a cumulative probability range of 0.03−0.09, while the upper value 17 lbs corresponds to a range of 0.89−0.96, both representing 90% credible intervals.

Additionally, the posterior predictive check demonstrates a reduction in uncertainty regarding the parameter Δ. Compare the posterior predictive check in Fig. 8) with the prior predictive checks shown in Figs. 5 and 6.

7 Discussion

Over the last two decades, several probability distributions with interpretable parameters defined on the same scale as observable quantities were proposed (Myerson 2005; Keelin and Powley 2011; Hadlock and Bickel 2017). The primary goal of research into quantile-parameterized distributions is to simplify the elicitation process and make it more accessible for experts. In our proposed *hybrid* elicitation framework, tailored specifically for models with quantile-parameterized likelihoods, experts are encouraged to adopt a sampling

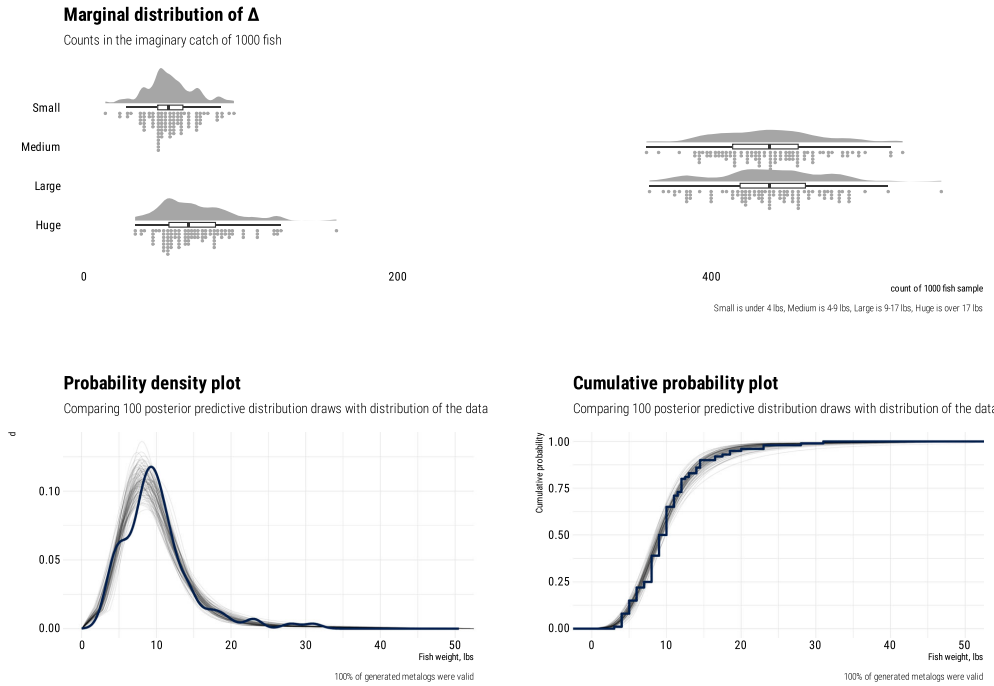


Fig. 8 Posterior predictive check for the QDirichlet-Metalog model

frame. This replaces the challenging task of expressing uncertainty about cumulative probabilities with a simpler task of expressing uncertainty about natural frequencies in a hypothetical sample (Gigerenzer 2011; Hoffrage et al. 2002, 2015).

The elicitation method for Dirichlet distribution proposed by Elfadaly and Garthwaite (2013) asks the expert to assume that the median of the previously assessed category p_{i-1} is, in fact, the true value of the probability for the category $i-1$. It then proceeds to elicit the value p_i for the next category i, conditional on this assessment. We believe that such conditioning can be made simpler if one adopts the natural frequency framework. Inputs elicited from the expert in the natural frequency frame can be easily validated through simulation, providing the expert with immediate feedback on the implications of their judgments for the model.

In Bayesian analysis, we see quantile-parameterized and parametric likelihoods as complementary. Initiating a model with a likelihood expressed by a quantile-parameterized distribution can be advantageous when only QPT judgments from experts are available, no specific choice for a parametric distribution is evident, and data is limited. As data becomes more abundant and our understanding of the data-generating process improves, a transition to a parametric likelihood can be justified.

Table 6 Posterior sample summary for the simplex Δ

Category	Mean	Median	q5	q95	rhat
Small	0.0552	0.0536	0.0306	0.0862	1.000
Medium	0.4328	0.4318	0.3763	0.4927	1.000
Huge	0.0724	0.0703	0.0427	0.1093	1.001
Large	0.4396	0.4398	0.3803	0.4986	1.000

The QDirichlet-Metalog model described in this paper can be applied in conjunction with a predictive approach to elicitation (Kadane 1980). Assuming that the only information elicited from the expert is the predictive QPT $\{p, q\}_n$, the quantiles q vector can be combined with a *uniform* Dirichlet prior, allowing the data alone to define the posterior for the simplex Δ. Given that the Dirichlet distribution is a generalization of the Beta distribution to higher dimensions, a weakly informative prior can be specified with a unit vector, i.e. Dirichlet$(1, 1, \ldots 1)$. We discuss inference using weakly informative priors in Appendix B in Supplemental Materials.

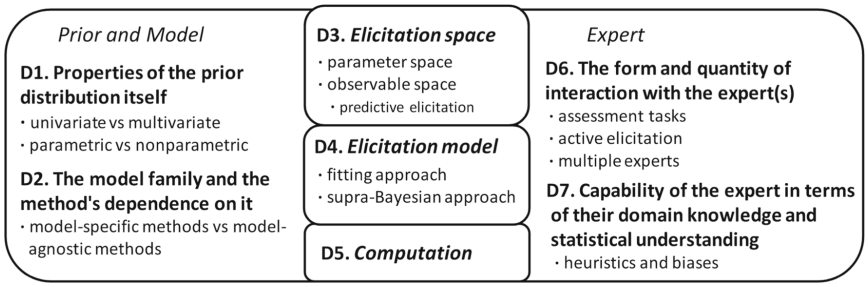


Fig. 9 Prior elicitation hypercube

Parametric elicitation aims to describe the epistemic uncertainty contained in the *parameters* of the model with the help of experts. On the other hand, predictive elicitation aims to describe the uncertainty in the *next observation* without distinguishing between the randomness in the model and the lack of knowledge about the model parameters.

Mikkola et al. (2023) proposed the prior elicitation hypercube with 7 dimensions related to the elicitation of prior distributions (Fig. 9). Following this classification, the proposed *hybrid elicitation* falls under the category of a univariate, parametric, prior-specific (D1), model-specific (D2) elicitation method, conducted in the observable space (D3). *Hybrid elicitation* leverages the approach proposed by Elfadaly and Garthwaite (2013) to derive the parameter vector(s) of the (Generalized) Dirichlet distribution (D4). This process relies on the simple arithmetic computations (D5) to transform the parameters of the conditional marginal beta distributions into the (Generalized) Dirichlet parameter vector(s). Furthermore, *hybrid elicitation* adopts an active, iterative elicitation approach (D6), requiring minal assumptions about the expert's familiarity with statistical concepts, such a detailed understanding of the underlying generative model (D7).

Hybrid elicitation begins by describing the next observation, but subsequently shifts to characterizing the uncertainty inherent in the predictive assessment itself. This is achieved by describing a hypothetical sample from the target population corresponding to the cumulative probabilities. These probabilities, in conjunction with a set of quantile values, serve as parameters within the quantile-parameterized model. Hybrid elicitation, similar to predictive elicitation, deals with observable quantities. However, like parametric (structural) elicitation, it ultimately results in characterizing the uncertainty in the model parameters. Thus, hybrid elicitation can be seen as an observation-level parametric elicitation specifically designed for for quantile-parameterized models.

Supplemental materials
Supplemental materials contain the R and Stan code for all examples used in the article. Appendix A includes the details of metalog distribution. Appendix B provides the details of the QDirichlet-metalog model with weakly informative prior. Appendix C includes the results of Simulation-Based Calibration for both models discussed in the paper.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-023-10325-0.

Acknowledgements The authors have no conflict of interest to declare. D Perepolkin and U Sahlin were funded by the strategic research environment Biodiversity and Ecosystem Services in Changing Climate (BECC). U Sahlin was also funded by the Crafoord Foundation (ref 20200626). B Goodrich is supported by U.S. National Science Foundations grants 2051246 and 2153019.

Author Contributions DP wrote the main manuscript text, BG provided substantial assistance with Stan implementation of quantile-based inference. US supervised the project and provided key input in model specification phase. All authors reviewed the manuscript.

Funding Open access funding provided by Lund University.

Data availability The `qpd` R package used in this paper is available on Github at https://github.com/dmi3kno/qpd. Contact corresponding author Dmytro Perepolkin (dmytro.perepolkin@cec.lu.se) for requests for data.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abbas, A.E., Budescu, D.V., Yu, H.T., et al.: A comparison of two probability encoding methods: fixed probability vs. fixed variable values. Decis. Anal. **5**(4), 190–202 (2008). https://doi.org/10.1287/deca.1080.0126

Akbarov, A.: Probability elicitation: Predictive approach. PhD thesis, University of Salford (2009)

Balakrishnan, N.: Continuous Multivariate Distributions. In: Wiley StatsRef: Statistics Reference Online. American Cancer Society, (2014) https://doi.org/10.1002/9781118445112.stat01249

Bockting, F., Radev, S.T., Bürkner, P.C.: Simulation-Based Prior Knowledge Elicitation for Parametric Bayesian Models. (2023) https://doi.org/10.48550/arXiv.2308.11672

Bürkner, P.C., Charpentier, E.: Modelling monotonic effects of ordinal predictors in Bayesian regression models. Br. J. Math. Stat. Psychol. **73**(3), 420–451 (2020). https://doi.org/10.1111/bmsp.12195

Coles, S.G., Tawn, J.A.: A Bayesian analysis of extreme rainfall data. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **45**(4), 463–478 (1996). https://doi.org/10.2307/2986068

Cook, S.R., Gelman, A., Rubin, D.B.: Validation of software for Bayesian models using posterior quantiles. J. Comput. Graph. Stat. **15**(3), 675–692 (2006). https://doi.org/10.1198/106186006X136976

Crowder, M.: Bayesian priors based on a parameter transformation using the distribution function. Ann. Inst. Stat. Math. **44**(3), 405–416 (1992)

Elfadaly, F.G., Garthwaite, P.H.: Eliciting Dirichlet and Connor–Mosimann prior distributions for multinomial models. TEST **22**(4), 628–646 (2013). https://doi.org/10.1007/s11749-013-0336-4

Faber, I., Jung, J.: Rmetalog: the metalog distribution (2021)

Gabry, J., Češnovar, R.: Cmdstanr: R interface to 'CmdStan' (2022)

Gabry, J., Simpson, D., Vehtari, A., et al.: Visualization in Bayesian workflow. J. R. Stat. Soc. A. Stat. Soc. **182**(2), 389–402 (2019)

Gelman, A., Carlin, J.B., Stern, H.S., et al.: Bayesian data analysis. CRC Press, Cambridge (2013)

Gigerenzer, G.: What are natural frequencies? BMJ **343**, d6386 (2011). https://doi.org/10.1136/bmj.d6386

Hadlock, C.C.: Quantile-parameterized methods for quantifying uncertainty in decision analysis. PhD thesis, University of Texas, Austin, TX, (2017) https://doi.org/10.15781/T2F18SX41

Hadlock, C.C., Bickel, J.E.: Johnson Quantile-Parameterized Distributions. Decis. Anal. **14**(1), 35–64 (2017)

Hadlock, C.C., Bickel, J.E.: The generalized Johnson quantile-parameterized distribution system. Decis. Anal. **16**(1), 67–85 (2019). https://doi.org/10.1287/deca.2018.0376

Hanea, A.M., Hemming, V., Nane, G.F.: Uncertainty quantification with experts present status and research needs. Risk Anal. (2021). https://doi.org/10.1111/risa.13718

Hartmann, M., Agiashvili, G., Bürkner, P., et al.: Flexible prior elicitation via the prior predictive distribution. (2020) arXiv:2002.09868 [stat]

Hoffrage, U., Gigerenzer, G., Krauss, S., et al.: Representation facilitates reasoning: what natural frequencies are and what they are not. Cognition **84**(3), 343–352 (2002). https://doi.org/10.1016/S0010-0277(02)00050-1

Hoffrage, U., Krauss, S., Martignon, L., et al.: Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. Front. Psychol. **6**, 1473 (2015)

Johnson, N.L., Kotz, S., Balakrishnan, N.: Discrete multivariate distributions. Wiley series in probability and statistics. Wiley, New York (1997)

Kadane, J., Wolfson, L.J.: Experiences in elicitation. J. Royal Stat. Soci.: Series D (The Stat.) **47**(1), 3–19 (1998). https://doi.org/10.1111/1467-9884.00113

Kadane, J.B.: Predictive and structural methods for eliciting prior distributions. Bayesian Anal. Econ. Stat. 18 (1980)

Keelin, T.W.: The metalog distributions. Decis. Anal. **13**(4), 243–277 (2016). https://doi.org/10.1287/deca.2016.0338

Keelin, T.W., Powley, B.W.: Quantile-parameterized distributions. Decis. Anal. **8**(3), 206–219 (2011). https://doi.org/10.1287/deca.1110.0213

Keeney, R., von Winterfeldt, D.: Eliciting probabilities from experts in complex technical problems. IEEE Trans. Eng. Manage. **38**(3), 191–201 (1991). https://doi.org/10.1109/17.83752

King, R.A.R.: New distributional fitting methods applied to the generalised [lambda] distribution. PhD thesis, Queensland University of Technology, Australia (1999)

Knight, F.H.: Risk, Uncertainty and Profit, vol 31. Houghton Mifflin (1921)

Manderson, A.A., Goudie, R.J.B.: Translating predictive distributions into informative priors. (2023). https://doi.org/10.48550/arXiv.2303.08528

Mikkola, P., Martin, O.A., Chandramouli, S., et al.: Prior knowledge elicitation: the past, present, and future. Bayesian Anal. **1**(1), 1–33 (2023). https://doi.org/10.1214/23-BA1381

Modrák, M., Moon, A.H., Kim, S., et al.: Simulation-based calibration checking for bayesian computation: the choice of test quantities shapes sensitivity, (2022). https://doi.org/10.48550/arXiv.2211.02383

Morgan, M.G.: Use (and abuse) of expert elicitation in support of decision making for public policy. Proc. Natl. Acad. Sci. **111**(20), 7176–7184 (2014)

Morgan, M.G., Henrion, M., Small, M.: Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, Cambridge (1990)

Myerson, R.B.: Probability models for economic decisions. Duxbury applied series. Thomson/Brooke/Cole, Belmont, CA (2005)

Nair, N.U., Sankaran, P.G., Dileepkumar, M.: Bayesian inference in quantile functions. Commun. Stat. - Theory Methods (2020). https://doi.org/10.1080/03610926.2020.1827430

Patel, J.K., Read, C.B.: Handbook of the normal distribution, vol. 150. CRC Press, Cambridge (1996)

Perepolkin, D.: Qpd: tools for quantile-parameterized distributions (2019)

Perepolkin, D., Goodrich, B., Sahlin, U.: The tenets of quantile-based inference in Bayesian models. Comput. Stat. Data Anal. **187**(107), 795 (2023). https://doi.org/10.1016/j.csda.2023.107795

Perepolkin, D., Lindström, E., Sahlin, U.: Quantile-parameterized distributions for expert knowledge elicitation. (2023b) https://doi.org/10.31219/osf.io/tq3an

Powley, B.W.: Quantile function methods for decision analysis. PhD thesis, Stanford University, Paolo Alto, CA (2013)

Pratt, J.W., Raiffa, H., Schlaifer, R.: Introduction to statistical decision theory. MIT press, Cambridge (1995)

Rayner, G.D., MacGillivray, H.L.: Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Stat. Comput. **12**(1), 57–75 (2002)

Spetzler, C.S., Staël Von Holstein, C.A.S.: Probability encoding in decision analysis. Manage. Sci. **22**(3), 340–358 (1975)

Talts, S., Betancourt, M., Simpson, D., et al.: Validating Bayesian inference algorithms with simulation-based calibration. (2020) https://doi.org/10.48550/arXiv.1804.06788
Vega-Yon, G., Marjoram, P.: Fmcmc: a friendly mcmc framework. J. Open Source Softw. (2019). https://doi.org/10.21105/joss.01427
Vihola, M.: Robust adaptive Metropolis algorithm with coerced acceptance rate. Stat. Comput. **22**(5), 997–1008 (2012)
Wasserman, L.: All of nonparametric statistics. Springer texts in statistics. Springer, New York (2006)
Wilson, E.: Fitting a modified Connor–Mosimann distribution to elicited quantiles of multinomial probabilities (2017)
Winkler, R.L.: The assessment of prior distributions in Bayesian analysis. J. Am. Stat. Assoc. **62**(319), 776–800 (1967). https://doi.org/10.1080/01621459.1967.10500894
Winkler, R.L.: Prior information, predictive distributions, and Bayesian model-building. Bayesian analysis in econometrics and statistics, pp. 95–109. North-Holland Publishing Company, Amsterdam (1980)

Paper IV

Observer-adjusted species distribution models using presence-only data

Dmytro Perepolkin[a,*], Johan Elmberg[b], Finn Lindgren[c], Ullrika Sahlin[a]

[a] *Lund University, Centre for Environmental and Climate Science, Lund, Sweden, 22362*

[b] *Kristianstad University, Department of environmental science, Kristianstad, Sweden, 29188*

[c] *University of Edinburgh, School of Mathematics, Edinburgh, Scotland, EH9 3FD*

Abstract

Effective waterfowl management demands precise estimates of presence and abundance. Financial constraints hinder monitoring efforts, prompting the utilization of citizen science data. However, these opportunistic datasets exhibit taxonomic preferential sampling bias, manifested in inflated implied absences. To address this, we propose an attention adjustment based on species list shortage (SLS), a measure of deviation of observation list length from local species richness. Testing its efficacy, we develop a two-stage Zero Adjusted continuous Poisson model for species distribution, applying log-transformations to mitigate overdispersion. The bias-adjusted model, evaluated against systematically observed data, emerges as a promising approach to enhance predictive performance while accounting for observer biases in opportunistic surveys.

Keywords: species distribution model, species list length, informative prior, INLA, bias

1. Introduction

1.1. Waterfowl management

The waterfowl management task entails responsible stewardship and conservation of a diverse avian family, encompassing ducks, swans, and geese, classified under the Order *Anseriformes* and the Family *Anatidae* (Williams, 1997; Roberts et al., 2018). These species hold significant ecological importance, contributing to nutritional cycles, offering recreational opportunities, and playing a pivotal role in conservation efforts (MacMillan et al., 2004).

Among the variety of species within the waterfowl family, a strong focus lies on the management of geese. In Europe, goose management is strategically organized through Adaptive Species Management Plans, primarily targeting four key species: the Greylag goose (Anser anser), the Barnacle Goose (Branta leucopsis) comprising two distinct sub-populations with separate flyways, the Pink-footed Goose (Anser brachyrhynchus), and the Taiga Bean Goose (Anser fabalis) (Madsen et al., 2017). Additionally, a sizable population of Canada Goose (Branta canadensis), recognized as an invasive species, coexists with varying degrees of regulation. Europe is also home for the Greater white-fronted goose (Anser albifrons), Lesser white-fronted goose (Anser erythropus) and Red-breasted goose (Branta ruficollis). Of particular conservation significance are the Lesser white-fronted goose and the Red-breasted goose, both globally threatened and warranting heightened conservation attention (Fox et al., 2017).

*Corresponding author

Email addresses: dmytro.perepolkin@example.com (Dmytro Perepolkin), johan.elmberg@hkr.se (Johan Elmberg), finn.lindgren@edinburgh.co.uk (Finn Lindgren), ullrika.sahlin@cec.lu.se (Ullrika Sahlin)

The regulatory landscape for these species is diverse, with some, such as the Greylag goose, being subject to relatively unrestricted hunting, while others enjoy protected status, precluding harvest (Tombre et al., 2021). All goose species face multifaceted challenges arising from environmental shifts, climate change, urban sprawl, alterations in land use, and the expansion of agricultural activities (Fox and Madsen, 2017). The overarching objective of goose management is to ensure the preservation of ecosystems essential for sustaining robust populations, concurrently addressing the adverse impacts—such as potential damage to human interests—stemming from the proliferation of these avian species. More specifically, a central and daunting task of goose management in Europe and North America is to reduce the conflict with agriculture due to growing populations and new migratory habits of geese (Bradbeer et al., 2017; Eythórsson et al., 2017; Tombre et al., 2013).

To effectively manage goose populations on a flyway scale, precise estimates of their presence and abundance are imperative. Accurate data are crucial for developing population models, determining harvest quotas, and implementing habitat enhancement strategies (Baveco et al., 2017; Jensen et al., 2008; Moriguchi et al., 2013). However, this necessitates a comprehensive understanding of the extent to which various landscapes are utilized by geese.

A species distribution model is designed to forecast the presence and abundance (count) of geese at a specific location. Given the likely disparate mechanistic processes governing the presence and abundance of geese, a model capable of handling these two aspects separately is warranted.

1.2. Presence-only data

Unfortunately, the current landscape of goose management is challenged by financial constraints, with monitoring efforts facing the brunt of shrinking budgets (Johnson et al., 2023). In response to these fiscal challenges, researchers have turned to leveraging abundant yet inherently biased opportunistically collected citizen science data found in online databases, including the Global Biodiversity Information Facility (GBIF) (GBIF, 2023) and eBird (Sullivan et al., 2009). Notably, citizen science databases present a set of biases, primarily arising from deviations in observation protocols. Observations are often recorded at the convenience of contributors, without adhering to a specified instruction or protocol (Chauvier et al., 2021; Dorazio, 2014; Phillips et al., 2009; Stolar and Nielsen, 2015; Warton et al., 2013).

1.3. Preferential taxonomic sampling

A principal challenge in opportunistic surveys lies in the oversight, neglect or misidentification of observed species. Observers, motivated by diverse interests, may selectively record favorite species, document unusual occurrences, or strive to document every species they encounter on a trip (out of habit or because of their understanding of the value of systematic approach to wildlife observation) (Di Cecco et al., 2021). Taxonomic preferential sampling, a scientific term for the phenomenon for this kind of selective attention, results in the selective registering of certain species, potentially leaving others underrecorded or unrecorded (Dorazio, 2014).

Szabo et al. (2010) used species list length as a proxy for observation effort - longer lists indicate higher commitment to registering all encountered species. In addition, the longer is the species list, the higher is the confidence in implied absences. The absences on species lists of observers with taxonomic preference bias do not signify true absence. In contrast, absences on lists of systematically observing individuals can be interpreted as true absences with higher confidence.

The primary concern in this study is related to implied absences (given the method we adopted for the pseudo-absence imputation). Even though there could be some inaccuracy in the counts reported by unexperienced observers, the reports of abundance are not subject to the same kind of bias. Therefore, we posit that presence should be modeled separately from abundance.

1.4. Proposed methodology

To address taxonomic preferential sampling, we propose a bias adjustment similar to distance sampling methodologies (Eberhardt, 1967; Farr et al., 2021; Royle et al., 2004; Yuan et al., 2017). The foundational principle in distance sampling involves identifying an *ideal state*. For example, observers are more likely to detect species in immediate proximity to roads. The probability of detection attenuates with increasing distance from the road, following a detection function with a scale parameter governing the rate of attenuation (Martino et al., 2021; Ribeiro et al., 2023; Sicacha-Parada et al., 2021).

Species list length reaches its ideal state when it equals local species richness (i.e., it includes all species expected to occur in the surrounding environment) (Boersch-Supan et al., 2019; Kelling et al., 2015). The difference between the species list length and the local species richness, which we term *species list shortage* (SLS), serves as a measure of deviation from the ideal state, quantifying the number of species overlooked and/or unreported by an observer. An SLS adjustment (which we refer to as the *attention adjustment*) is essentially correcting probability of discovery, and not abundance, which should be reflected in the model.

To summarise, the proposed methodology is comprised of these steps:

1. Build a species richness model
2. Calculate SLS for every observation from the predicted species richness and species list length
3. Build a Species Distribution Model that separates presence from abundance with SLS bias adjustment applied on the presence part of the model
4. Evaluate predictive performance against data which is assumed to be free from taxonomic sampling bias

The aim of the paper is to test whether the proposed attention adjustment using SLS can improve the predictive performance of a species distribution model when evaluated against the observations with comprehensive species lists.

To implement the proposed methodology, the following contribution are made. A Zero Adjusted Poisson (hurdle) model (Arab, 2015; Zuur, 2017) was selected to allow for bias adjustment in presence part of the species distribution model. Such isolation of the presence process is not achievable with the more commonly used Zero Inflated Poisson (Wenger and Freeman, 2008). To complement the data collected by systematic monitoring programs and augment the size of the held-out data, observations with the complete species lists were extracted from eBird (Sullivan et al., 2009).

2. Materials and Method

2.1. Data

Opportunistic data

The Global Biodiversity Information Facility (GBIF) is an international organization maintaining the online repository of wildlife observations contributed by participating institutions (GBIF, 2023). GBIF operates under the widely adopted hierarchical biodiversity data standard, Darwin Core, embraced by a majority of contributing institutions. Notably, in Sweden the foremost contributor to GBIF is the Swedish Species Gateway, also known as SLU Artdatabanken (Artportalen), developed and managed by the Swedish Species Information Centre on behalf of the Swedish Environmental Protection Agency (Bradter et al., 2018; Henckel et al., 2020; Jönsson et al., 2023; Ruete et al., 2017, 2020; Snäll et al., 2011). Artportalen serves as a platform for the collection of raw species observations from researchers, conservationists, and private individuals through its web interface, accessible on both desktop and mobile devices. Within GBIF, these observations, categorized as *occurrences*, are organized hierarchically under *events*, which, in turn, are grouped under *collections* - potentially corresponding to observations made by different organizations.

We adopt an event-centered approach (Szabo et al., 2010), where the unit of analysis is a wildlife visit by a unique observer at a specific time and location. An event may encompass observations of various species, collated into an event species list. In this paper we use the term *visit* as a more granular definition of an *event*

adopted by Darwin Core data standard. In some cases GBIF allows observations from different locations to be combined under the same `eventID`. In order to differentiate between the species lists collected in different locations, we define a *visit* as a unique combination of the observer (as recorded in `recordedBy`), GBIF's `eventID` (when available), and geographic coordinates. When `eventID` is absent, we supplement it with the combination of the data `collectionCode` (e.g., Artportalen, eBIRD, etc.), observer, verbatim `locality`, and observation date.

Systematic monitoring data

In a Swedish context, one of the most systematically collected datasets of bird observations is derived from the Swedish Bird Survey (Svensk Fågeltaxering) program (Fågeltaxering, 2023). This initiative relies mainly on the dedicated efforts of volunteers and knowledgeable ornithologists who conduct bird counts at specified times throughout the year in predefined locations, employing consistent method. The Swedish Bird Survey program encompasses various sub-programs, including standard routes, summer point routes, winter point routes, coastal bird routes, night routes, seabird routes during the breeding season, and seabirds in autumn and winter (Fågeltaxering, 2023).

A standard route consists of eight 1 km-line transects and eight five-minute points within a 2 by 2 km square. The routes are strategically distributed across a sparse and regular grid to comprehensively cover the entire country. The primary objective of the Swedish Bird Survey program is to compile statistics and report general population trends (Liljebäck et al., 2021).

The merits of the Swedish Bird Survey dataset are in its consistent temporal and spatial aspects, featuring regular observations at fixed locations covering diverse land types through a national grid system. The standardized route lengths ensure uniform observation effort, while the survey's comprehensiveness is maintained by the professionalism of participating ornitologists. This systematic approach, consistently applied over many years, yields historical records valuable for long-term analyses. Swedish Bird Survey data, integrated into GBIF through Artportalen.se, possess a unique collection code, distinguishing it from non-systematically collected observations, including those made by citizen scientists in Sweden.

Complete checklists

Despite the intrinsic value of systematically collected Swedish Bird Survey data, the comparison to more abundant presence-only (non-systematic) data prompts inquiry into the extent of their divergence. One dataset marginally inferior to systematically collected Swedish Bird Survey is eBird (Sullivan et al., 2009), an online ornithological observation database reliant on volunteer-contributed data. Distinct from Artportalen, eBird utilizes a data entry form, wherein contributors are encouraged to select a species checklist for completion. While not obligated to record every species, contributors can mark an entry as a "complete checklist" if they are submitting the records for all species observed (see Figure 1).

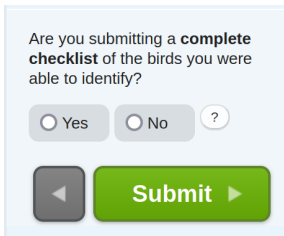
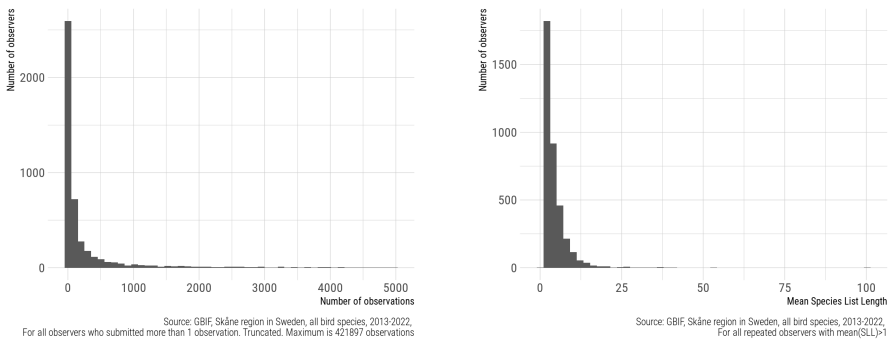


Figure 1: Confirmation menu on eBIRD.org platform

Though eBird complete checklists share some similarities with the Swedish Bird Survey, they deviate in terms of data collection, being gathered in random locations, at random times, and covering variable spatial

extents and durations. Despite these sampling challenges, they can be regarded as presence-absence data. The absence of species on eBird checklists implies true absence with higher confidence. Integration of eBird records into GBIF necessitates the identification and exclusion of these records from presence-only datasets. Unfortunately, the relevant completeness flag is not transferred to GBIF, requiring a dataset integration to discern eBird records with complete checklists and infer absences.

Artportalen also offers a checklist feature, analogous to eBird. However, this option is not a default choice and requires a separate login, which might explain lack of popularity of this data entry option. If the observation data are entered through the checklist, the observation is marked by special text in the comment field, which makes it difficult to parse these records out and impossible to see if the submitted checklist is complete, unless it is filed as part of a specific campaign or project, such as the Swedish Bird Survey (Fågeltaxering, 2023).



(a) Number of observations by observer

(b) Mean Species List Length by observer

Figure 2: Observer contributions to GBIF

Observing the observers

Citizen science contributions are notably diverse, raising concerns about the potential biases introduced by varying levels of engagement (Feldman et al., 2021; Grimmett et al., 2020; Jönsson et al., 2023). In the Swedish province Skåne, a total of 4,666,523 bird observations were submitted to the GBIF between 2013 and 2022. A total of 83% of these observations (contributed by 5,746 observers) have proper event identification, enabling their aggregation with other observations made during the same visit by the same observer.

Remarkably, the top 1% of observers (by number of observations) accounted for 47% of all avian observation records, while the top 20% contributed a staggering 96% of the total records (Figure 2a). In terms of event species lists, the top 1% of observers, defined by the number of lists, contributed 36% of all species lists, with the top 20% contributing 95% of the lists. Mean species list length (MSLL) varies from 1 to 101, with the top 1% of observers (by MSLL) beginning at an average list length of 17 (Figure 2b). This stratification shows that there is a substantial contributions made by a selected group of observers, and that there's large variation in the species list lengths among the contributors.

Environmental data

In the context of waterfowl species distribution modeling, particularly for geese, certain environmental features consistently hold relevance due to the distinct biology and behavior exhibited by these species.

For instance, Jensen et al. (2008) studied the spring-staging behavior of pink-footed geese in Mid-Norway. The authors used various environmental covariates, including the coverage proportion of roads and vertical features within the geographical grid cell, the mean coverage proportion of vertical features in the neighboring cells, distance to open water, and elevation above sea level. A study by Moriguchi et al. (2013) modeled the distribution of East-Asian greater white-fronted geese in Japan. Environmental covariates considered in this research included minimum temperature (obtained from WorldClim), average elevation, proportion of rice fields, urban and lake areas, distance to lakes, maximum snow depth, and latitude-longitude, combined by principal component analysis. Subsequently, Li et al. (2017) added human population density to the list of influential factors.

Typically, a study area is subdivided into contiguous sub-areas of equal size (cells), wherein covariate values are assumed to be constant. In our study, we utilized the 2018 Swedish National Landcover data (Nationella Marktäckedata, NMD), presented in the form of raster grids with a resolution of 10 x 10 m. To mitigate potential imprecision in geo-referencing and to smooth out spatial effects, we aggregated the raster data by a factor of 100. Additionally, we posited that the cell size of 1 x 1 km roughly corresponds to the area feasible for a single observer to inspect during an average trip. It's noteworthy that, unlike eBIRD, GBIF does not document observational efforts in terms of time or spatial extent, thus necessitating the assumption about average effort area at this resolution.

Selected covariates

We computed the proportion of the focal cell area classified as agricultural land, specifically NMD classes 3 and 42, representing arable land and other open land with more than 10% vegetation coverage, respectively. Additionally, we assessed the proportion of the focal cell area covered by water, including NMD classes 2 and 61, which encompass lakes, water courses, open sea areas, and wetlands. A spatial factor covariate indicating whether any of the aggregated cells are part of Natura-2000 protected areas (SPA, SCI) was also integrated into our model.

To quantify the distance to residence, we utilized OpenStreetMap to extract a polygon layer showing land use designated as residential. The inclusion of distance to road as either a covariate or a thinner has been a common practice in numerous studies (Adde et al., 2021; Cretois et al., 2021; Escamilla Molgora et al., 2022a,b; Geurts, 2023; Koshkina et al., 2017; Milanesi et al., 2020; Morera-Pujol et al., 2023; Ramesh et al., 2022; Sicacha-Parada et al., 2021; Stolar and Nielsen, 2015). We extracted the road network, encompassing primary, secondary, and motorway roads, along with associated link roads, regular roads, and trunks.

We implemented several data restrictions, applying spatial, temporal, and taxonomic constraints to effectively manage the model's size:

- *Spatial Subset*: The entire dataset, including GBIF/eBIRD observational records and spatial covariates (NMD), was confined to the North-East part of the Skåne province, comprised of Osby, Östra Göinge, Bromölla, Hässleholm, and Kristianstad municipalities.
- *Temporal Subset*: Observations considered in the analysis were limited to the period from 2013 to 2022, and including the months of April through August each year. The first half of this timeframe (2013-2017) was used to train the species richness model, while the subsequent data (2018-2022) was reserved for the primary species distribution model. The hold-out data for evaluation was extracted from the data for 2018-2022. This strategy ensured the estimation of species richness on entirely non-overlapping data.
- *Taxon Focus*: The species distribution model was developed to predict the distribution of the broad category of geese, comprising the genera Anser, Branta, and Chen. The amalgamation of observations of all goose species into a single category serves to provide some additional protection against possible misclassification by citizen scientist observers and alleviate individual species detection bias.

2.2. Species richness model

The species richness model aims to predict the number of unique species (defined by unique GBIF `speciesKey`) observed in each focal cell within the study area. For each raster cell, we tallied the unique species encountered across years, observers, and events. Given the assumption that bird species richness cannot be zero at any location in Sweden, cells with zero species richness counts (for which no observations in GBIF were found) were filled with missing values accordingly.

We assume that regional gamma diversity (Hunter Jr and Gibbs, 2006) varied negligibly among years and that it had no long-term trend during the study period. We also assume that even a single encounter of a unique bird species during the study period suffices to include it in the richness count. This aggregation of counts mitigates observer-specific preferential sampling bias, providing a species richness estimate somewhat independent of individual sampling preferences (though influenced by collective observer preferences).

Species richness is amenable to modeling through an inhomogeneous Poisson model. Following the notation in Fithian et al. (2015), we consider random set $\mathcal{S} = s_i \subseteq \mathcal{D}$ of locations s_i for all species varieties in a geographical domain $\mathcal{D}$ (the *species process*). The presence of a particular species at location s is akin to its participation in the local ecosystem, evidenced by at least one sighting during the observation period. The Inhomogeneous Poisson Process (IPP) characterizes the species process with an intensity function $\lambda(s)$, mapping sites s_i to non-negative values (here $\lambda(s)$ quantifies how many s_i occur in the vicinity of s). The intensity is conveniently modeled in log-linear form:

$$\log \lambda(s) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m(s)$$

Here, β_0 is the intercept, and β_m represents individual coefficients for the covariate value $x_m(s)$. In systematic sampling, data would be compiled from locations A_i in $\mathcal{D}$, registering the species richness $N_{\mathcal{S}}(A_i)$ at each location. While the target quantity - number of unique bird species observed in location s_i - represents a count, we opted for a continuous version of the Poisson distribution due to overdispersion caused by large variance in the number of visits to individual locations. To address this, we log-transformed the counts. Given the exclusion of zero counts, this transformation did not result in information loss (O'Hara and Kotze, 2010). The continuous version of the Poisson (xPoisson) distribution was used:

$$\text{Prob}(y) = \frac{\lambda^y}{y!} exp(-\lambda)$$

where $y > 0$ is the response variable, and λ is the expected value. In the continuous version, $y!$ is computed using the integer part of y.

We specified the species richness model with the following formulation:

$$\begin{aligned} \lambda_{s_i} &\sim \beta_0 + \beta_1 L_v(s_i) + \beta_2 L_w(s_i) + \beta_3 I_p(s_i) + \beta_4 D_r(s_i) + \tau(s_i) \\ \log(Y_{s_i}) &\sim \text{xPoisson}(\log(\lambda_{s_i})) \end{aligned}$$

Here, L_v and L_w represent the proportions of vegetation- and water-covered areas in the focal cell, respectively; I_p is a factor indicating whether the focal cell constitutes a protected area; D_r denotes the distance to the nearest residential area; and $\tau(s_i)$ is the spatial autocorrelation term representing the SPDE model with a Matern prior (Lindgren and Rue, 2015). The observation points are situated on a regular grid corresponding to the aggregated raster for the covariates.

We used the predicted species log-richness as a thinning adjustment to mitigate taxonomic preferential sampling bias, as explained in the subsequent section. The corresponding code is available in the Supplementary Materials.

2.3. Thinning function and prior specification

Species distribution models, especially those built on presence-only data, often incorporate adjustments to address preferential sampling bias. This bias can arise from two primary sources: preferred location and preferred taxonomy. Rarely do these biases act in isolation. Specific locations, such as bird-watching sites, may undergo more extensive sampling due to observer motivation to spot their preferred species.

Studies by Sicacha-Parada et al. (2021) and Geurts (2023) employ distance to roads as the adjustment factor to address the higher observation density in locations with road access (i.e., the preferred location bias). Although we extracted the road network from OpenStreetsMaps and computed the distance to the nearest road for each location on a grid, given the dense road network in Skåne, we observed no improvement in predictive performance. Consequently, we opted to exclude this adjustment to maintain model parsimony.

To address the preferred taxonomy bias, we suggest contextualizing the observation of target species (or genera of geese) in relation to other species observed during the same visit. Our hypothesis posits that if the species list length approximates local species richness, then omitted species from the list can be confidently treated as true absences.

We aggregated bird species observations from the study region and computed the species list length, defined as the count of unique `speciesKey` codes on the observation list for a unique visit. Records related to focused ring-recapture programs that did not include any geese were excluded.

For comparability with the log-richness predictions from the model in Section 2.2, we opted to log-transform the species list length. Since the list of species cannot be shorter than 1, no information loss occurs in the chosen transformation (O'Hara and Kotze, 2010). Therefore, the species list shortage metric, utilized for bias adjustment in the species distribution model, is computed as the non-negative difference between the predicted log-richness for location i and the log species list length from visit j (denoted as SLL_j).

$$SLS_j = \max[0, \log(Y_i) - \log(SLL_j)]$$

We employed a log-negative exponential distance function, denoted as $d(SLS, \sigma) = \exp(-\sigma SLS)$. This function scales the Species List Shortage (SLS) using the parameter σ, rendering the bias log-linear within the likelihood specification for the Zero-Adjusted xPoisson model, as discussed in the subsequent section.

2.4. Species Distribution Models

Existing species distribution models in the literature, particularly those based on presence-only data, commonly employ Poisson models (Gelfand and Shirota, 2019; Boersch-Supan et al., 2019; Cretois, 2021; Warton et al., 2013; Renner et al., 2015) and Zero-Inflated Poisson models (Martínez-Minaya et al., 2018; Wenger and Freeman, 2008; Nolan et al., 2022). Despite the increasing abundance of most goose species in recent decades, they still constitute a relatively small portion of all birds in Skåne. Biological knowledge also indicates that certain locations are unsuitable for geese due to sensitivity to noise (Simonsen et al., 2016, 2017), likely aversion of areas with tall structures (Johnson et al., 2014), optimization of locations to minimize predator encounters, lack of foraging habitat, and a general tendency to avoid humans (Fox, 2019; Humphrey et al., 2023). Hence, it is reasonable to anticipate a substantial number of locations with zero goose counts. Factors contributing to goose abundance may differ, including the availability of sufficient food, open area size, distance to roosting sites, etc (Jensen et al., 2008; Chudzińska et al., 2015).

Consequently, a model capable of handling a substantial number of inflated zeros in the data and distinguishing between the causes of presence and abundance is essential.

The Zero-Inflated Poisson (ZIP) model (Zuur, 2017), characterized by a combination of binomial and Poisson features, addresses one of these requirements, namely accommodating large number of zero counts.

$$
\begin{aligned}
\text{Prob}(y|\lambda) &= p \times 1_{y=0} + (1-p) \times \text{Poisson}(y|\lambda) \\
E(y) &= (1-p)\lambda \\
Var(y) &= (1-p)(\lambda + \lambda^2 p) \\
&= E(y)(1+\lambda p)
\end{aligned}
$$

The Zero-Adjusted Poisson (ZAP) model (Zuur, 2017), commonly known as the hurdle model, represents an advancement over ZIP by incorporating the truncated version of the Poisson distribution:

$$
\begin{aligned}
\text{Prob}(y|\lambda) &= p \times 1_{y=0} + (1-p) \times \text{Poisson}(y|y>0,\lambda) \\
E(y) &= \frac{1}{1-\exp(-\lambda)} p\lambda \\
Var(y) &= \frac{1}{1-\exp(-\lambda)} p(\lambda + p\lambda^2) - \left(\frac{1}{1-\exp(-\lambda)} p\lambda\right)^2 \\
&= E(y)\left[1 - \exp(-\lambda)E(y)\right]
\end{aligned}
$$

The ZAP model allows modeling absences ($y = 0$) separately from abundances ($y > 0$). The parameters p and λ can be modelled using distinct but possibly overlapping sets of covariates.

Due to the overdispersion of counts, we log-transformed the abundance counts of geese. Consequently, instead of using the zero-truncated Poisson in our ZAP model, we opted for the continuous (non-truncated) version, where zero log-abundance corresponds to the count of one.

In order to test the efficacy of attention adjustment we compare predictive performance of the model with SLS (implemented as a covariate and as a bias adjustment) to the predictive performance of an equivalent model without SLS.

Baseline model (no adjustment)
Zero-Adjusted xPoisson model for goose presence and abundance:

$$
\begin{aligned}
\theta_{s_i} &\sim \alpha_0 + \alpha_1 L_v(s_i) + \alpha_2 L_w(s_i) + \alpha_3 I_p(s_i) + \eta(s_i) \\
Z_{js_i} &\sim \text{Binomial}(\text{logit}(\theta_{s_i})) \\
Z_{js_i} &= \begin{cases} 0 \text{ if } Y_{js_i} = 0 \\ 1 \text{ if } Y_{js_i} > 0 \end{cases} \\
\lambda_{s_i} &\sim \beta_0 + \beta_1 L_v(s_i) + \beta_2 L_w(s_i) + \beta_3 I_p(s_i) + \tau(s_i) \\
\log(Y_{js_i}|Y_{js_i} > 0) &\sim \text{xPoisson}(\log(\lambda_{s_i}))
\end{aligned}
$$

Here $p_{s_i} = 1 - \text{logit}(\theta_{s_i})$ represent the probability of absence. Covariates in both parts of the model include the proportion of vegetation (L_v) and water (L_w)covering the target cell s_i, an indicator of whether the cell includes protected territory (I_p), and model-specific Gaussian fields $\eta(s_i)$ and $\tau(s_i)$.

Attention adjustment as covariate

The binomial part of the Zero-Adjusted xPoisson model includes the attention adjustment as a covariate as follows:

$$\theta_{js_i} \sim \alpha_0 + \alpha_1 L_v(s_i) + \alpha_2 L_w(s_i) + \alpha_3 I_p(s_i) + \alpha_4 SLS_j(s_i) + \eta(s_i)$$
$$Z_{js_i} \sim \text{Binomial}(\text{logit}(\theta_{js_i}))$$

Attention adjustment as thinner

The binomial part of the Zero-Adjusted xPoisson model with the attention adjustment included as a thinner:

$$\theta_{js_i} \sim \alpha_0 + \alpha_1 L_v(s_i) + \alpha_2 L_w(s_i) + \alpha_3 I_p(s_i) + d(SLS_j, \sigma) + \eta(s_i)$$
$$Z_{js_i} \sim \text{Binomial}(\text{logit}(\theta_{js_i}))$$
$$\sigma \sim \text{logitMyerson}(1, 3, 10, 0.1)$$

where $d(SLS, \sigma) = \exp(-\sigma SLS)$ is log-negative-exponential distance function.

A custom prior for parameter σ was derived from expert judgment, using elicited quantiles as parameters in the logit-Myerson distribution (Perepolkin et al., 2023) (Figure 3). The implied lower bound of the parameter value σ is $Q(0, 1, 3, 10, 0.1) = 0.2$.

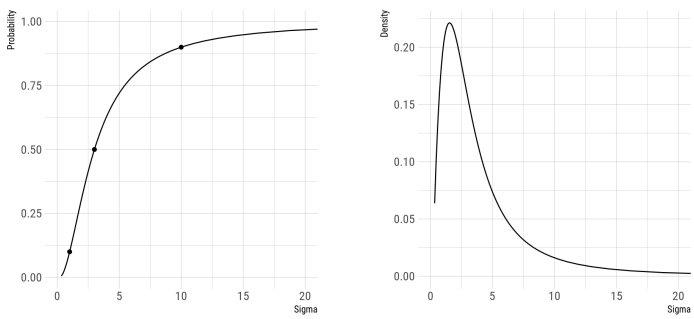


Figure 3: The CDF and the PDF of the logit-Myerson distribution used as a prior for the scale parameter in the thinning function

We implemented the models using the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009; Lindgren and Rue, 2015), interfaced by the inlabru package (Bachl et al., 2019) in R. The code is provided in the Supplementary materials.

2.5. Prediction scores

For assessing model performance, the following prediction scores were used (lower scores indicate better performance):

- Absolute error (AE): Measures the absolute deviation of observed counts from the predicted median, calculated as $\text{AE}_i = y_i - \text{median}_i$. AE serves as a proper scoring rule for the median.

- Squared error (SE): Calculated as the squared difference between counts and the predicted mean, expressed as $\text{SE}_i = [y_i - E(Y_i|\text{data})]^2$. SE functions as a proper scoring rule for the expectation.
- Dawid-Sebastiani (DS) score. Given by $\text{DS}_i = \frac{[y_i - E(Y_i|\text{data})]^2}{V(Y_i|\text{data})} + \log[V(Y_i|\text{data})]$. DS serves as a proper scoring rule for the predictive mean ($E(Y_i)$) and variance ($V(Y_i)$).

The posterior predictive variance ($Var(Y_i)$) for the predicted count Y_i is based on the count expectations (μ_i) and variances (ξ_i^2) of the model predictions for each grid cell, conditioned on the model predictor x_i.

The posterior predictive variance of the count Y_i is:

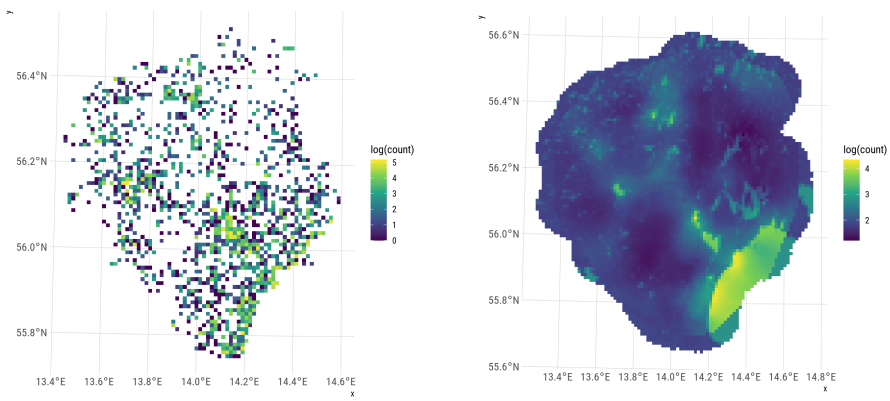
$$\begin{aligned} Var(Y_i) &= E(V(Y_i|x_i)) + Var(E(Y_i|x_i)) \\ &= E(\xi_i^2) + V(\mu_i) \end{aligned}$$

- (Negated) Log score (LG): Represents the logarithm of the observation probability, defined as $\text{LG}_i = -\log[P(Y_i = y_i|\text{data})]$. LG is a strictly proper score (Gneiting and Raftery, 2007).

3. Results

3.1. Species richness predictions

We assume that the species richness, corresponding to the regional gamma diversity (Hunter Jr and Gibbs, 2006), remained relatively stable from 2013 to 2022. Accordingly, we utilize the location-specific predictions derived from the model trained on observations spanning the initial five years (Figure 4a) as an estimate of the species richness in corresponding locations for the subsequent 5-year period (Figure 4b). The model predicts higher log-richness values along the coastal regions and the lakes.



(a) Log-richness data on 1x1 km grid

(b) Predicted log-richness

Figure 4: Species log-richness and predictions

3.2. Parameter estimates

All covariates contribute to predictions, with the exception of the proportion of vegetation in the binomial part of the baseline model (Table 1 provides).

The coefficients in the two bias adjusted models are very similar (Table 2, Table 3). The negative coefficient for SLS as a covariate implies that for a given species richness the probability of presence increases with the length of the species list. The influence of the thinner is also negative (indicated as `-sigma*sls` in Table 3).

Table 1: Fixed effect estimates for the baseline SDM

	mean	sd	0.025quant	0.5quant	0.975quant
L_w_present	2.061	0.308	1.460	2.061	2.668
L_v_present	0.531	0.279	-0.016	0.530	1.079
Intercept_present	-4.450	0.175	-4.797	-4.448	-4.111
L_w	0.585	0.200	0.193	0.585	0.976
L_v	0.939	0.188	0.570	0.939	1.308
Intercept	-13.533	0.117	-13.764	-13.532	-13.306

Table 2: Fixed effect estimates for the SDM with attention adjustment as covariate

	mean	sd	0.025quant	0.5quant	0.975quant
L_w_present	3.618	0.299	3.034	3.617	4.207
L_v_present	1.223	0.266	0.701	1.222	1.746
sls	-1.558	0.048	-1.651	-1.558	-1.464
Intercept_present	-2.725	0.165	-3.051	-2.723	-2.406
L_w	0.579	0.205	0.177	0.580	0.979
L_v	0.933	0.191	0.556	0.933	1.307
Intercept	-13.531	0.118	-13.765	-13.530	-13.300

Table 3: Fixed effect estimates for the SDM with attention adjustment as thinner

	mean	sd	0.025quant	0.5quant	0.975quant
L_w_present	3.621	0.299	3.037	3.620	4.211
L_v_present	1.224	0.266	0.702	1.223	1.747
-sigma*sls	-0.772	0.037	-0.845	-0.772	-0.698
Intercept_present	-2.725	0.165	-3.052	-2.724	-2.406
L_w	0.579	0.204	0.179	0.580	0.978
L_v	0.933	0.191	0.558	0.934	1.306
Intercept	-13.531	0.118	-13.764	-13.530	-13.301

3.3. Predictive performance

The model incorporating SLS as a covariate exhibits superior performance (smaller scores) across the means of all metrics (Table 4), but not when we compare the quantiles for AE and SE (Figure 5).

Both models with SLS contribute to a reduced variance in predictive scores seen over observations in the held-out dataset (Figure 5). The scores that rely on computed variance and observation probability (DS and LG) are more concentrated.

Table 4: Mean predictive scores by model

Model	MAE	MSE	MDS	MLG
SLS_covariate	1.111	1.727	2.165	2.095
SLS_thinner	1.293	1.923	2.580	2.120
Baseline	1.210	1.970	8.621	3.416

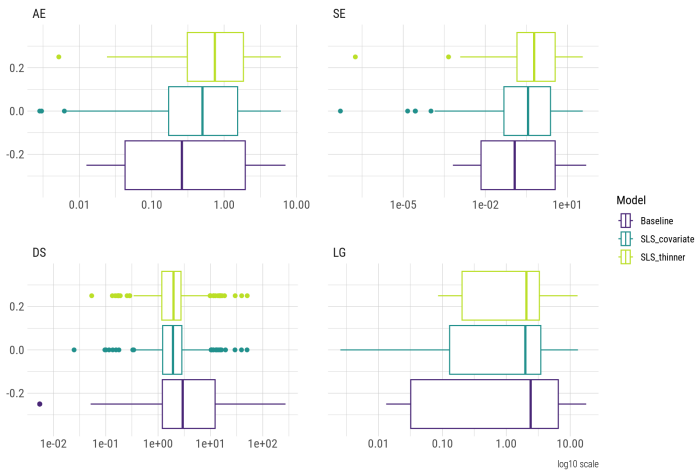


Figure 5: Comparison of Absolute Error (AE), Squared Error (SE), Dawid-Sebastiani Score (DS), and (Negated) Log Score (LG) across the three species distribution models. Smaller is better.

4. Discussion

The model comparison shows that the species list shortage contributes to improved predictive performance, primarily by concentrating the distribution of scores across observations. The method of implementing SLS made little difference.

Model limitations

Predicting waterfowl distribution poses challenges due to the intricate nature of bird biology, habitat preferences, communal behavior, and their considerable mobility. The model in this study overlooks the daily local commutes of birds between foraging grounds and water roosts. While our environmental covariates capture general habitat preferences, introducing a time-aware model could enhance predictions, considering the interaction between the time of day and preferred habitat.

Our observations are confined to the period of April through August each year, missing variations in bird behavior. Specifically, during the nesting period in April-May, birds tend to stay closer to water, while by late summer, increased mobility is observed as offspring mature (Cramp and Simmons, 1977).

Migratory patterns of some goose species were not considered in our model. While the majority of geese, mainly Greylag Goose and Canada Goose, are considered residents in Skåne, some species within the broad

goose genera migrate through Sweden in fall and spring, or come to Skåne to spend the winter (Ekberg and Nilsson, 1994). Models focusing on flyway scales are more suitable for estimating the migration patterns of waterfowl.

The model we presented lacks consideration for the biological distinctions among various goose species in Southern Sweden. For instance, the nesting preference of Canada Goose for lakes of boreal type differs from wetlands in open areas favored by Greylag Goose. Our model benefits from mitigating preferential sampling bias through simultaneous modeling of multiple species, as indicated by Fithian et al. (2015). However, separate modeling of individual species within the broader goose genera could offer a more nuanced understanding of their biology and preferences.

Our environmental covariates are rudimentary, lacking data on crop types in Sweden's agricultural landscapes. Considering the varying nutritional preferences of geese over their annual cycle (Cramp and Simmons, 1977) could likely enhance predictive performance. However, the temporal window (breeding time) of the data set in the present study makes the lack of crop data less problematic, as this is when geese stay mainly in and close to wetlands, and spend less time foraging in agricultural areas. Modeling local-scale movements of geese may be better accomplished with an Agent-Based Model approach, explicitly accounting for daily energy balance and the tradeoff between flying to remote, energy-rich foraging grounds and staying close to roosting locations.

Human-geese interactions were omitted from our model. While distance to human residence positively contributed to the richness model, we refrained from including it in the species distribution model due to its potential impact on both the probability of discovery (birds closer to human residences are more likely to be spotted and reported) and the probability of presence (wild geese tend to avoid human settlements). The same applies to another popular bias-correction covariate, distance to road. As explained by Fithian et al. (2015), addressing this confounding factor requires specialized approaches, likely involving hierarchical and simultaneous modeling of multiple species.

Our species richness model relies on a strong assumption about the probability of discovery, related to the challenge of collating counts across locations with disparate sampling efforts, as evident in Figure 4. The uneven distribution of visits, particularly to popular locations, skews the probability of discovery towards complete account for all resident species, while locations with infrequent visits may not have had a chance to adequately explore local species richness. The preferential sampling we aim to address results in an uneven distribution of the probability of detection across locations s_i, rendering our pooling of species across time, visits, and observers only partially successful.

To appropriately capture the probability of discovery for each species k in the richness model, explicit modeling of the probability of detection is necessary. Such modeling should consider factors such as the probability of detection in a single visit, the number of visits, and specific environmental covariates influencing the probability of discovery.

Future research

Despite these limitations, it is evident that Species List Length is an important feature that warrants explicit consideration when informing models using citizen science data. The results confirm that longer species lists align with higher observation effort (Roberts et al., 2007; Szabo et al., 2010). The proposed methodology of combining the species list length and richness can be developed further. The model would benefit from more accurate richness estimates.

Citizen science data remains a valuable but underexplored resource. We encourage scientists to scrutinize observation circumstances, fostering the adoption of scientifically grounded sampling practices, including checklist adherence and formal observation protocols. The data collection process is as informative as the data itself, and meticulous sampling protocols are paramount to yield meaningful insights. The bias-adjusted model, evaluated against systematically observed data, emerges as a promising approach to enhance predictive performance while accounting for observer biases in opportunistic surveys.

5. Acknowledgments

This work would not have be possible without excellent computational environment created by the `{targets}` package. We want to thank Will Landau for making *simple models easy and complex models possible*.

6. Supplementary Materials

The data and the code are made available online at https://osf.io/m5t4r/.

References

Adde, A., Casabona i Amat, C., Mazerolle, M.J., Darveau, M., Cumming, S.G., O'Hara, R.B., 2021. Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data. Ecosphere 12, e03790. doi:10.1002/ecs2.3790.

Arab, A., 2015. Spatial and Spatio-Temporal Models for Modeling Epidemiological Data with Excess Zeros. International Journal of Environmental Research and Public Health 12, 10536–10548. doi:10.3390/ijerph120910536.

Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., 2019. Inlabru: An R package for Bayesian spatial modelling from ecological survey data. Methods in Ecology and Evolution 10, 760–766. doi:10.1111/2041-210X.13168.

Baveco, J.M., Bergjord, A.K., Bjerke, J.W., Chudzińska, M.E., Pellissier, L., Simonsen, C.E., Madsen, J., Tombre, I.M., Nolet, B.A., 2017. Combining modelling tools to evaluate a goose management scheme. Ambio 46, 210–223. doi:10.1007/s13280-017-0899-5.

Boersch-Supan, P.H., Trask, A.E., Baillie, S.R., 2019. Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. Biological Conservation 240, 108286. doi:10.1016/j.biocon.2019.108286.

Bradbeer, D.R., Rosenquist, C., Christensen, T.K., Fox, A.D., 2017. Crowded skies: Conflicts between expanding goose populations and aviation safety. Ambio 46, 290–300. doi:10.1007/s13280-017-0901-2.

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A., Snäll, T., 2018. Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models of less common species? Methods in Ecology and Evolution 9, 1667–1678. doi:10.1111/2041-210X.13012.

Chauvier, Y., Zimmermann, N.E., Poggiato, G., Bystrova, D., Brun, P., Thuiller, W., 2021. Novel methods to correct for observer and sampling bias in presence-only species distribution models. Global Ecology and Biogeography 30, 2312–2325. doi:10.1111/geb.13383.

Chudzińska, M.E., van Beest, F.M., Madsen, J., Nabe-Nielsen, J., 2015. Using habitat selection theories to predict the spatiotemporal distribution of migratory birds during stopover – a case study of pink-footed geese Anser brachyrhynchus. Oikos 124, 851–860. doi:10.1111/oik.01881.

Cramp, S., Simmons, K. (Eds.), 1977. Handbook of the Birds of Europe, the Middle East, and North Africa: The Birds of the Western Palearctic. Repr ed., Oxford Univ. Pr, Oxford.

Cretois, B., 2021. Transforming the Use of Citizen Science Data for Biodiversity Conservation at Different Scales. Ph.D. thesis. NTNU. Trondheim, Norway.

Cretois, B., Simmonds, E.G., Linnell, J.D.C., van Moorter, B., Rolandsen, C.M., Solberg, E.J., Strand, O., Gundersen, V., Roer, O., Rød, J.K., 2021. Identifying and correcting spatial bias in opportunistic citizen science data for wild ungulates in Norway. Ecology and Evolution 11, 15191–15204. doi:10.1002/ece3.8200.

Di Cecco, G.J., Barve, V., Belitz, M.W., Stucky, B.J., Guralnick, R.P., Hurlbert, A.H., 2021. Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. BioScience 71, 1179–1188. doi:10.1093/biosci/biab093.

Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23, 1472–1484. doi:10.1111/geb.12216.

Eberhardt, L.L., 1967. Some Developments in 'Distance Sampling'. Biometrics 23, 207–216. doi:10.2307/2528156, arXiv:2528156.

Ekberg, B., Nilsson, L., 1994. Skånes fåglar. Signum, Lund, Sweden.

Escamilla Molgora, J.M., Sedda, L., Diggle, P., Atkinson, P.M., 2022a. A joint distribution framework to improve presence-only species distribution models by exploiting opportunistic surveys. Journal of Biogeography 49, 1176–1192. doi:10.1111/jbi.14365.

Escamilla Molgora, J.M., Sedda, L., Diggle, P.J., Atkinson, P.M., 2022b. A taxonomic-based joint species distribution model for presence-only data. Journal of The Royal Society Interface 19, 20210681. doi:10.1098/rsif.2021.0681.

Eythórsson, E., Tombre, I.M., Madsen, J., 2017. Goose management schemes to resolve conflicts with agriculture: Theory, practice and effects. Ambio 46, 231–240. doi:10.1007/s13280-016-0884-4.

Fågeltaxering, 2023. Metoder | Svensk Fågeltaxering. http://www.fageltaxering.lu.se/inventera/metoder.

Farr, M.T., Green, D.S., Holekamp, K.E., Zipkin, E.F., 2021. Integrating distance sampling and presence-only data to estimate species abundance. Ecology 102, e03204. doi:10.1002/ecy.3204.

Feldman, M.J., Imbeau, L., Marchand, P., Mazerolle, M.J., Darveau, M., Fenton, N.J., 2021. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. PLOS ONE 16, e0234587. doi:10.1371/journal.pone.0234587.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6, 424–438. doi:10.1111/2041-210X.12242.
Fox, A.D., 2019. Urban Geese – looking to North America for experiences to guide management in Europe? Wildfowl 69, 3–27–27.
Fox, A.D., Elmberg, J., Tombre, I.M., Hessel, R., 2017. Agriculture and herbivorous waterfowl: A review of the scientific basis for improved management. Biological Reviews 92, 854–877. doi:10.1111/brv.12258.
Fox, A.D., Madsen, J., 2017. Threatened species to super-abundance: The unexpected international implications of successful goose conservation. Ambio 46, 179–187. doi:10.1007/s13280-016-0878-2.
GBIF, 2023. What is GBIF? https://www.gbif.org/what-is-gbif.
Gelfand, A.E., Shirota, S., 2019. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs 89, e01372. doi:10.1002/ecm.1372.
Geurts, E., 2023. Examining Spatial Biases in the Community Science Platform, iNaturalist, Using British Columbia, Canada, as a Case Study. Thesis. University of Victoria. Victoria BC, Canada.
Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437.
Grimmett, L., Whitsed, R., Horta, A., 2020. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. Ecological Modelling 431, 109194. doi:10.1016/j.ecolmodel.2020.109194.
Henckel, L., Bradter, U., Jönsson, M., Isaac, N.J.B., Snäll, T., 2020. Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. Diversity and Distributions 26, 1276–1290. doi:10.1111/ddi.13128.
Humphrey, J.E., Haslem, A., Bennett, A.F., 2023. Housing or habitat: What drives patterns of avian species richness in urbanized landscapes? Landscape Ecology 38, 1919–1937. doi:10.1007/s10980-023-01666-2.
Hunter Jr, M.L., Gibbs, J.P., 2006. Fundamentals of Conservation Biology. John Wiley & Sons.
Jensen, R.A., Wisz, M.S., Madsen, J., 2008. Prioritizing refuge sites for migratory geese to alleviate conflicts with agriculture. Biological Conservation 141, 1806–1818. doi:10.1016/j.biocon.2008.04.027.
Johnson, F.A., Madsen, J., Clausen, K.K., Frederiksen, M., Jensen, G.H., 2023. Assessing the value of monitoring to biological inference and expected management performance for a European goose population. Journal of Applied Ecology 60, 132–145. doi:10.1111/1365-2664.14313.
Johnson, W., Schmidt, P., Taylor, D., 2014. Foraging flight distances of wintering ducks and geese: A review. Avian Conservation and Ecology 9. doi:10.5751/ace-00683-090202.
Jönsson, M., Kasperowski, D., Coulson, S.J., Nilsson, J., Bína, P., Kullenberg, C., Hagen, N., van der Wal, R., Peterson, J., 2023. Inequality persists in a large citizen science programme despite increased participation through ICT innovations. Ambio doi:10.1007/s13280-023-01917-1.
Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., Sorte, F.A.L., Moore, T., Wiggins, A., Wong, W.K., Wood, C., Yu, J., 2015. Can Observation Skills of Citizen Scientists Be Estimated Using Species Accumulation Curves? PLOS ONE 10, e0139600. doi:10.1371/journal.pone.0139600.
Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M., Stone, L., 2017. Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. Methods in Ecology and Evolution 8, 420–430. doi:10.1111/2041-210X.12738.
Li, X., Si, Y., Ji, L., Gong, P., 2017. Dynamic response of East Asian Greater White-fronted Geese to changes of environment during migration: Use of multi-temporal species distribution model. Ecological Modelling 360, 70–79. doi:10.1016/j.ecolmodel.2017.06.004.
Liljebäck, N., Bergqvist, G., Elmberg, J., Haas, F., Nilsson, L., Lindström, Å., Månsson, J., 2021. Learning from long time series of harvest and population data: Swedish lessons for European goose management. Wildlife Biology 2021. doi:10.2981/wlb.00733.
Lindgren, F., Rue, H., 2015. Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software 63, 1–25. doi:10.18637/jss.v063.i19.
MacMillan, D., Hanley, N., Daw, M., 2004. Costs and benefits of wild goose conservation in Scotland. Biological Conservation 119, 475–485. doi:10.1016/j.biocon.2004.01.008.
Madsen, J., Williams, J.H., Johnson, F.A., Tombre, I.M., Dereliev, S., Kuijken, E., 2017. Implementation of the first adaptive management plan for a European migratory waterbird population: The case of the Svalbard pink-footed goose Anser brachyrhynchus. Ambio 46, 275–289. doi:10.1007/s13280-016-0888-0.
Martínez-Minaya, J., Cameletti, M., Conesa, D., Pennino, M.G., 2018. Species distribution modeling: A statistical review with focus in spatio-temporal issues. Stochastic Environmental Research and Risk Assessment 32, 3227–3244. doi:10.1007/s00477-018-1548-7.
Martino, S., Pace, D.S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., Jona Lasinio, G., 2021. Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. Ecography 44, 1533–1543. doi:10.1111/ecog.05843.
Milanesi, P., Mori, E., Menchetti, M., 2020. Observer-oriented approach improves species distribution models from citizen science data. Ecology and Evolution 10, 12104–12114. doi:10.1002/ece3.6832.
Morera-Pujol, V., Mostert, P.S., Murphy, K.J., Burkitt, T., Coad, B., McMahon, B.J., Nieuwenhuis, M., Morelle, K., Ward, A.I., Ciuti, S., 2023. Bayesian species distribution models integrate presence-only and presence–absence data to predict deer distribution and relative abundance. Ecography 2023, e06451. doi:10.1111/ecog.06451.
Moriguchi, S., Amano, T., Ushiyama, K., 2013. Creating a potential distribution map for Greater White-fronted Geese wintering

in Japan. Ornithological Science 12, 117–125. doi:10.2326/osj.12.117.
Nolan, V., Gilbert, F., Reader, T., 2022. Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models. Journal of Biogeography 49, 215–232. doi:10.1111/jbi.14268.
O'Hara, R.B., Kotze, D.J., 2010. Do not log-transform count data. Methods in Ecology and Evolution 1, 118–122. doi:10.1111/j.2041-210X.2010.00021.x.
Perepolkin, D., Lindström, E., Sahlin, U., 2023. Quantile-parameterized distributions for expert knowledge elicitation. doi:10.31219/osf.io/tq3an.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecological Applications 19, 181–197. doi:10.1890/07-2153.1.
Ramesh, V., Gupte, P.R., Tingley, M.W., Robin, V.V., DeFries, R., 2022. Using citizen science to parse climatic and land cover influences on bird occupancy in a tropical biodiversity hotspot. Ecography 2022, e06075. doi:10.1111/ecog.06075.
Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods in Ecology and Evolution 6, 366–379. doi:10.1111/2041-210X.12352.
Ribeiro, R., Matthiopoulos, J., Lindgren, F., Tello, C., Zariquiey, C.M., Valderrama, W., Rocke, T.E., Streicker, D.G., 2023. Incorporating Environmental Heterogeneity and Observation Effort to Predict Host Distribution and Viral Spillover from a Bat Reservoir. Preprint. Ecology. doi:10.1101/2023.04.04.535562.
Roberts, A., Eadie, J.M., Howerter, D.W., Johnson, F.A., Nichols, J.D., Runge, M.C., Vrtiska, M.P., Williams, B.K., 2018. Strengthening links between waterfowl research and management. The Journal of Wildlife Management 82, 260–265. doi:10.1002/jwmg.21333.
Roberts, R.L., Donald, P.F., Green, R.E., 2007. Using simple species lists to monitor trends in animal populations: New methods and a comparison with independent data. Animal Conservation 10, 332–339. doi:10.1111/j.1469-1795.2007.00117.x.
Royle, J.A., Dawson, D.K., Bates, S., 2004. Modeling Abundance Effects In Distance Sampling. Ecology 85, 1591–1597. doi:10.1890/03-3127.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology 71, 319–392. doi:10.1111/j.1467-9868.2008.00700.x.
Ruete, A., Arlt, D., Berg, Å., Knape, J., Żmihorski, M., Pärt, T., 2020. Cannot see the diversity for all the species: Evaluating inclusion criteria for local species lists when using abundant citizen science data. Ecology and Evolution 10, 10057–10065. doi:10.1002/ece3.6665.
Ruete, A., Pärt, T., Berg, Å., Knape, J., 2017. Exploiting opportunistic observations to estimate changes in seasonal site use: An example with wetland birds. Ecology and Evolution 7, 5632–5644. doi:10.1002/ece3.3100.
Sicacha-Parada, J., Steinsland, I., Cretois, B., Borgelt, J., 2021. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. Spatial Statistics 42, 100446. doi:10.1016/j.spasta.2020.100446.
Simonsen, C.E., Madsen, J., Tombre, I.M., Nabe-Nielsen, J., 2016. Is it worthwhile scaring geese to alleviate damage to crops? – An experimental study. Journal of Applied Ecology 53, 916–924. doi:10.1111/1365-2664.12604.
Simonsen, C.E., Tombre, I.M., Madsen, J., 2017. Scaring as a tool to alleviate crop damage by geese: Revealing differences between farmers' perceptions and the scale of the problem. Ambio 46, 319–327. doi:10.1007/s13280-016-0891-5.
Snäll, T., Kindvall, O., Nilsson, J., Pärt, T., 2011. Evaluating citizen-based presence data for bird monitoring. Biological Conservation 144, 804–810. doi:10.1016/j.biocon.2010.11.010.
Stolar, J., Nielsen, S.E., 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Diversity and Distributions 21, 595–608. doi:10.1111/ddi.12279.
Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142, 2282–2292. doi:10.1016/j.biocon.2009.05.006.
Szabo, J.K., Vesk, P.A., Baxter, P.W.J., Possingham, H.P., 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. Ecological Applications 20, 2157–2169. doi:10.1890/09-0877.1, arXiv:29779611.
Tombre, I.M., Eythórsson, E., Madsen, J., 2013. Towards a Solution to the Goose-Agriculture Conflict in North Norway, 1988–2012: The Interplay between Policy, Stakeholder Influence and Goose Population Dynamics. PLOS ONE 8, e71912. doi:10.1371/journal.pone.0071912.
Tombre, I.M., Fredriksen, F., Jerpstad, O., Østnes, J.E., Eythórsson, E., 2021. Population control by means of organised hunting effort: Experiences from a voluntary goose hunting arrangement. Ambio doi:10.1007/s13280-021-01590-2.
Warton, D.I., Renner, I.W., Ramp, D., 2013. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. PLOS ONE 8, e79168. doi:10.1371/journal.pone.0079168.
Wenger, S.J., Freeman, M.C., 2008. Estimating Species Occurrence, Abundance, and Detection Probability Using Zero-Inflated Distributions. Ecology 89, 2953–2959. doi:10.1890/07-1127.1.
Williams, B.K., 1997. Approaches to the Management of Waterfowl under Uncertainty. Wildlife Society Bulletin (1973-2006) 25, 714–720.
Yuan, Y., Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., Buckland, S.T., Rue, H., Gerrodette, T., 2017. Point Process Models for Spatio-Temporal Distance Sampling Data from a Large-Scale Survey of Blue Whales. The Annals of Applied Statistics 11, 2270–2297. doi:10.1214/17-AOAS1078, arXiv:26362186.
Zuur, AF., 2017. Beginner's Guide to Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-Inla: Using Glm and Glmm Volume I. Hightland Statistics Ltd., Sl OCLC 973745327, 61.

Doctoral theses published in Environmental Science, Lund University

1. Georg K.S. Andersson (2012) Effects of farming practice on pollination across space and time. Department of Biology/Centre for environmental and climate research

2. Anja M. Ödman (2012) Disturbance regimes in dry sandy grasslands – past, present and future. Department of Biology/Centre for environmental and climate research

3. Johan Genberg (2013) Source apportionment of carbonaceous aerosol. Department of Physics/Centre for environmental and climate research

4. Petra Bragée (2013) A palaeolimnological study of the anthropogenic impact on dissolved organic carbon in South Swedish lakes. Department of Geology/Centre for environmental and climate research

5. Estelle Larsson (2013) Sorption and transformation of anti-inflammatory drugs during wastewater treatment. Department of Chemistry/Centre for environmental and climate research

6. Magnus Ellström (2014) Effects of nitrogen deposition on the growth, metabolism and activity of ectomycorrhizal fungi. Department of Biology/Centre for environmental and climate research

7. Therese Irminger Street (2015) Small biotopes in agricultural landscapes: importance for vascular plants and effects on management. Department of physical geography and ecosystem science/ Department of Biology/Centre for environmental and climate research

8. Helena I. Hanson (2015) Natural enemies: Functional aspects of local management in agricultural landscapes. Department of Biology/Centre for environmental and climate research

9. Lina Nikoleris (2016) The estrogen receptor in fish and effects of estrogenic substances in the environment: ecological and evolutionary perspectives and societal awareness Department of Biology/Centre for environmental and climate research

10. Cecilia Hultin (2016) Estrogen receptor and multixenobiotic resistance genes in freshwater fish and snails: identification and expression analysis after pharmaceutical exposure. Centre for environmental and climate research

11. Annika M. E. Söderman (2016) Small biotopes: Landscape and management effects on pollinators. Department of Biology/Centre for environmental and climate research

12. Wenxin Ning (2016) Tracking environmental changes of the Baltic Sea coastal zone since the mid-Holocene. Department of Geology/Centre for environmental and climate research

13. Karin Mattsson (2016) Nanoparticles in the aquatic environment, Particle characterization and effects on organisms. Department of Chemistry/Centre for environmental and climate research

14. Ola Svahn (2016) Tillämpad miljöanalytisk kemi för monitorering och åtgärder av antibiotika- och läkemedelsrester I Vattenriket. School of Education and Environment, Kristianstad University/Centre for environmental and climate research

15. Pablo Urrutia Cordero (2016) Putting food web theory into action: Local adaptation of freshwaters to global environmental change. Department of Biology/Centre for environmental and climate research

16. Lin Yu (2016) Dynamic modelling of the forest ecosystem: Incorporation of the phosphorous cycle. Centre for environmental and climate research

17. Behnaz Pirzamanbein (2016) Recontruction of past European land cover based on fossil pollen data: Gaussian Markov random field models for compositional data. Centre for Mathematical Sciences/Centre for environmental and climate research

18. Arvid Bolin (2017) Ecological interactions in human modified landscapes –Landscape dependent remedies for the maintenance of biodiversity and ecosystem services. Department of Biology/Centre for environmental and climate research

19. Johan Martinsson (2017) Development and Evaluation of Methods in Source Apportionment of the Carbonaceous Aerosol. Department of Physics/Centre for environmental and climate research

20. Emilie Öström (2017) Modelling of new particle formation and growth in the atmospheric boundary layer. Department of Physics/Centre for environmental and climate research

21. Lina Herbertsson (2017) Pollinators and Insect Pollination in Changing Agricultural Landscapes. Centre for environmental and climate research

22. Sofia Hydbom (2017) Tillage practices and their impact on soil organic carbon and the microbial community. Department of Biology/Centre for environmental and climate research

23. Erik Ahlberg (2017) Speeding up the Atmosphere: Experimental oxidation studies of ambient and laboratory aerosols using a flow reactor. Department of Physics/Centre for environmental and climate research

24. Laurie M. Charrieau (2017) DISCO: Drivers and Impacts of Coastal Ocean Acidification. Department of Geology/Centre for environmental and climate research

25. Kristin Rath (2018) Soil salinity as a driver of microbial community structure and functioning. Department of Biology/Centre for environmental and climate research

26. Lelde Krumina (2018) Adsorption, desorption, and redox reactions at iron oxide nanoparticle surfaces. Department of Biology/Centre for environmental and climate research

27. Ana Soares (2018) Riverine sources of bioreactive macroelements and their impact on bacterioplankton metabolism in a recipient boreal estuary. Department of physical geography and ecosystem science/Centre for environmental and climate research

28. Jasmine Livingston (2018) Climate Science for Policy? The knowledge politics of the IPCC after Copenhagen. Centre for environmental and climate research

29. Simon David Herzog (2019) Fate of riverine iron over estuarine salinity gradients. Department of Biology/Centre for Environmental and Climate Research

30. Terese Thoni (2019) Making Blue Carbon: Coastal Ecosystems at the Science-Policy Interface. Centre for Environmental and Climate Research

31. Lovisa Nilsson (2019) Exploring synergies – management of multifunctional agricultural landscapes. Centre for Environmental and Climate Research

32. Zhaomo Tian (2019) Properties and fungal decomposition of iron oxide-associated organic matter. Centre for Environmental and Climate Research

33. Sha Ni (2020) Tracing marine hypoxic conditions during warm periods using a microanalytical approach. Department of Geology/Centre for Environmental and Climate Research

34. Julia Kelly (2021) Carbon exchange in boreal ecosystems: upscaling and the impacts of natural disturbances. Centre for Environmental and Climate Science

35. William Sidemo Holm (2021) Effective conservation of biodiversity and ecosystem services in agricultural landscapes. Centre for Environmental and Climate Science

36. John Falk (2021) Ice out of Fire: Ice and cloud condensation nucleation of aerosol particles emitted from controlled soot generation and combustion of renewable fuels. Department of Physics/Centre for Environmental and Climate Science

37. Maria Blasi i Romero (2021) Wild bees in agricultural landscapes: Modelling land use and climate effects across space and time. Centre for Environmental and Climate Science

38. Ivette Raices Cruz (2021) Robust analysis of uncertainty in scientific assessments. Department of Biology/Centre for Environmental and Climate Science

39. Adrian Gustafson (2022) On the role of terrestrial ecosystems in a changing Arctic. Department of Physical Geography and Ecosystem Science/Centre for Environmental and Climate Science

40. Klas Lucander (2022) Direct and indirect pressures of climate change on nutrient and carbon cycling in northern forest ecosystems – Dynamic modelling for policy support. Department of Physical Geography and Ecosystem Science/Centre for Environmental and Climate Science

41. Johan Kjellberg Jensen (2023) Understanding the urban ecosystem – interactions between plants, animals, and people. Department of Biology/Centre for Environmental and Climate Science

42. Paoloa Michaela Mafla-Endara (2023) Encounters at the microscale: Unravelling soil microbial interactions with nanoplastics. Department of Biology/Centre for Environment and Climate Science

43. Dmytro Perepolkin (2023) Scientific methods for integrating expert knowledge in Bayesian models. Centre for Environment and Climate Science