Giant galactic outflows and shocks in the cosmic web

Oei, M.S.S.L.

The history of astronomy is a history of receding horizons.
Edwin P. Hubble, American astronomer, *The Realm of the Nebulae (1936)*

8

Constraining the giant radio galaxy population with machine learning–accelerated detection and Bayesian inference

R. I. J. Mostert, **M. S. S. L. Oei**, B. Barkus, L. Alegre, M. J. Hardcastle, K. J. Duncan, H. J. A. Röttgering, R. J. van Weeren, M. Horton — *Astronomy & Astrophysics*, submitted

Abstract

Context Large-scale sky surveys at low frequencies, like the LOFAR Two-metre Sky Survey (LoTSS), allow for the detection and characterisation of unprecedented numbers of giant radio galaxies (GRGs, or 'giants'). This, in turn, enables us to study giants in a cosmological context. A tantalising prospect of such studies is a measurement of the contribution of giants to cosmic magnetogenesis. However, finding large GRG samples requires the creation of radio–optical catalogues for well-resolved radio sources and a suitable statistical framework to infer intrinsic GRG population properties.

Aims By automating the creation of radio–optical catalogues, we aim to expand significantly the census of known giants. With the resulting sample and a forward model mindful of selection effects, we aim to constrain their intrinsic length distri-

bution, number density, and lobe volume-filling fraction (VFF) in the Cosmic Web.

Methods We combine five existing codes into a single machine learning–driven (ML) pipeline that automates radio source component association and optical host identification for well-resolved radio sources. We create a radio–optical catalogue for the entire LoTSS Data Release 2 (DR2) footprint and subsequently select all sources that qualify as possible giants. We combine the list of ML pipeline GRG candidates with an existing list of LoTSS DR2 crowd-sourced GRG candidates and visually confirm or reject all members of the merged sample. To infer intrinsic GRG properties from GRG observations, we develop further a population-based forward model and constrain its parameters using Bayesian inference.

Results Roughly half of all radio sources that our ML pipeline identifies as giants (of at least $l_{\mathrm{p,GRG}} := 0.7$ Mpc long) indeed turn out to be upon visual inspection, whereas the success rate is one in eleven for the previous best giant-finding ML technique in the literature. We confirm $5,596$ previously unknown giants from the crowd-sourced LoTSS DR2 catalogue and $2,592$ previously unknown giants from the ML pipeline. Our confirmations and discoveries bring the total number of known giants to at least $11,524$. Our forward model for the intrinsic GRG population is able to provide a good fit to the data. Our posterior indicates that the projected lengths of giants are consistent with a curved power law probability density function whose initial tail index $\xi(l_{\mathrm{p,GRG}}) = -2.8 \pm 0.2$ changes by $\Delta\xi = -2.4 \pm 0.3$ over the interval up to $l_{\mathrm{p}} = 5$ Mpc. We predict a comoving GRG number density $n_{\mathrm{GRG}} = 13 \pm 10\ (100\ \mathrm{Mpc})^{-3}$, close to a current estimate of the number density of luminous non-giant radio galaxies. With the projected length distribution, number density, and additional assumptions, we derive a current-day GRG lobe VFF $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0) = 1.1 \pm 0.9 \cdot 10^{-5}$ in clusters and filaments of the Cosmic Web.

Conclusions We have created a state-of-the-art ML-accelerated pipeline for finding giants, whose complex morphologies, arcminute extents, and radio-emitting surroundings pose challenges. Our data analysis suggests that giants are more common than previously thought. More work is needed to make estimates of the GRG lobe VFF reliable, but the first results indicate that it is possible that magnetic fields originating from giants permeate significant ($\sim$10%) fractions of today's Cosmic Web.

Key words: Surveys – Methods: data analysis – Catalogues – Galaxies: active – Radio continuum: galaxies

8.1 Introduction

Recent radio Stokes-I imaging and rotation measure observations show that filaments of the Cosmic Web are magnetised (e.g. Govoni et al., 2019; de Jong et al., 2022; Carretti et al., 2023) with $B \sim 10^0$–10^2 nG (e.g. Vazza et al., 2021a). However, the origin of these magnetic fields remains highly uncertain. In a primordial magnetogenesis scenario (e.g. Subramanian, 2016), the seeds of intergalactic magnetic fields can be traced to the Early Universe. This scenario is not problem-free: primordial magnetic fields that arise before the end of inflation are typically too weak to match observations, while fields that arise after inflation (but before recombination) typically have coherence lengths that are too small. Alternatively, in an astrophysical magnetogenesis scenario, the seeds of intergalactic magnetic fields are predominantly spread by energetic astrophysical phenomena in the more recent Universe, such as radio galaxies (RGs) and supernova explosion driven winds (e.g. Vazza et al., 2017). In this latter scenario, giant radio galaxies (GRGs, or 'giants') may play a significant role in the magnetisation of the intergalactic medium (IGM), as their associated jets can carry magnetic fields of strength $B \sim 10^2$ nG from host galaxies to cosmological, megaparsec-scale distances (e.g. Oei et al., 2022a).

Efforts to measure the contribution of giants to astrophysical magnetogenesis in filaments of the Cosmic Web have only recently begun, with the advent of systematically processed, sensitive, low-frequency sky surveys such as the Low Frequency Array (LOFAR; van Haarlem et al., 2013) Two-metre Sky Survey (LoTSS; Shimwell et al., 2017). By carrying out a manual search for giants in LoTSS DR2 (Shimwell et al., 2022) pipeline products and a subsequent statistical analysis, Oei et al. (2023a) inferred a key statistic: the volume-filling fraction (VFF) of GRG lobes within clusters and filaments of the Local Universe, $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$. However, considerable uncertainty remains as to its precise value, which requires inference of both the intrinsic GRG length distribution and the intrinsic GRG number density, as well as information about the typical shape of GRG lobes.

As the number of observed radio galaxies rapidly increases with decreasing angular length, the time taken, as part of the manual process of associating radio source components and identifying optical host galaxies, logically increases. Machine learning (ML)–based techniques have the potential to massively accelerate the detection of specific radio sources, to complement or eventually replace manual searches (e.g. Proctor, 2016; Gheller et al., 2018; Lochner & Bassett, 2021; Mostert et al., 2023). The potential for detecting GRGs was demonstrated by Dabhade et al. (2020a), who visually inspected the 1,600 ML-predicted GRG candidates of Proctor (2016) and thereby discovered 151 giants. By combining multiple ML-based and rule-based algo-

rithms that automate both the radio component association process and the optical host identification process into a single pipeline, we aim to improve upon the 9% precision achieved by the ML-predictions of Proctor (2016).

In the current work, we construct a LoTSS DR2 GRG sample of unparalleled size, by combining results from a manual visual search (Oei et al., 2023a), a citizen science–based visual search (Hardcastle et al., 2023), and a machine learning–accelerated search (this article; Sect. 8.4). With a definitive LoTSS DR2 GRG sample in hand, we refine the Bayesian forward model presented in Oei et al. (2023a), and finally constrain several key geometric quantities pertaining to giants.

In Sect. 8.2, we briefly recap, generalise, and enrich the statistical GRG geometry theory of Oei et al. (2023a). In Sect. 8.3, we introduce the LoTSS DR2 data in which we search for giants. In Sect. 8.4, we describe the methods that we use to build our definite LoTSS DR2 GRG sample, and explain how we use the theory of Sect. 8.2 in practice to infer GRG quantities of interest. In Sect. 8.5, we present our findings regarding the projected proper length distribution for giants, their comoving number density, and their instantaneous lobe volume-filling fraction (VFF) in clusters and filaments of the Cosmic Web. In Sect. 8.6, we discuss caveats of the current work, compare our results with previous results, and propose promising directions for future work, before we conclude in Sect. 8.7.

We assume a flat, inflationary ΛCDM model with parameters adopted from Planck Collaboration et al. (2020); i.e. $h = 0.6766$, $\Omega_{\mathrm{BM},0} = 0.0490$, $\Omega_{\mathrm{M},0} = 0.3111$, $\Omega_{\Lambda,0} = 0.6889$, where $H_0 := h \cdot 100\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$. We define giants as radio galaxies with a projected proper[1] length $l_{\mathrm{p}} \geq l_{\mathrm{p,GRG}} := 0.7$ Mpc. We define the spectral index α such that it relates to flux density F_ν at frequency ν as $F_\nu \propto \nu^\alpha$; under this convention, most radio spectral indices are negative.

8.2 Theory

To infer the intrinsic length distribution, number density, and lobe volume-filling fraction of giants, we use a Bayesian forward modelling approach that incorporates selection effects. We adopt the framework described in Oei et al. (2023a), but generalise a few key formulae. Furthermore, in a change that allows for the extraction of tighter parameter constraints from the data, we now predict joint projected proper length–redshift histograms rather than projected proper length distributions.

[1] In Cosmic Web filament environments, where giants appear most common (Oei et al., in prep.), lobes may expand along the Hubble flow, rendering their proper and comoving extents different. To avoid ambiguity, we stress that our projected lengths are proper instead of comoving. A less precise synonym for 'projected proper length' often found in the literature is 'largest linear size' (LLS).

8.2.1 RG total and projected proper lengths

The central geometric quantity predicted by models of radio galaxy evolution (e.g. Turner & Shabala, 2015; Hardcastle, 2018) is, simply, the RG's intrinsic proper length l. Once the probability distribution of the intrinsic proper length random variable (RV) L is known, one can estimate other geometric quantities of interest, such as the VFF of RG lobes in the Cosmic Web. However, for the vast majority of observed RGs only a projected proper length l_p is available, as accurate measurements of jet inclination angles θ are currently challenging. In order to fit statistical models to data from surveys such as LoTSS DR2, models should therefore predict the distribution of the projected proper length RV L_p.

8.2.2 GRG projected proper length: general

We now show, first without adopting a specific parametric form for the distribution of L, how the cumulative density function (CDF) and probability density function (PDF) of the GRG projected proper length RV $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ can be calculated. In particular, suppose that L has support from some length $l_\mathrm{min} \geq 0$ onwards. It holds that $L_\mathrm{p} = L \sin \Theta$, where Θ is the inclination angle RV. Assuming that — at least on cosmological scales — all RG orientations in three dimensions are equally likely, the CDF of L_p relates to the PDF of L via

$$F_{L_\mathrm{p}}(l_\mathrm{p}) = \begin{cases} 0 & \text{if } l_\mathrm{p} \leq 0; \\ 1 - \int_{l_\mathrm{min}}^{\infty} \sqrt{1 - \left(\frac{l_\mathrm{p}}{l}\right)^2} f_L(l)\,\mathrm{d}l & \text{if } 0 < l_\mathrm{p} \leq l_\mathrm{min}; \\ 1 - \int_{l_\mathrm{p}}^{\infty} \sqrt{1 - \left(\frac{l_\mathrm{p}}{l}\right)^2} f_L(l)\,\mathrm{d}l & \text{if } l_\mathrm{p} > l_\mathrm{min}. \end{cases} \tag{8.1}$$

We note that, in the usual scenario $l_\mathrm{min} = 0$, the second case disappears. Equation 8.1 generalises Eq. A.8 from Oei et al. (2023a); its derivation closely follows the one presented there.

The CDF of the GRG projected proper length RV $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ is

$$F_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p}) = \begin{cases} 0 & \text{if } l_\mathrm{p} < l_\mathrm{p,GRG}; \\ 1 - \dfrac{\int_{l_\mathrm{p}}^{\infty} \sqrt{1 - \left(\frac{l_\mathrm{p}}{l}\right)^2} f_L(l)\,\mathrm{d}l}{\int_{l_\mathrm{p,GRG}}^{\infty} \sqrt{1 - \left(\frac{l_\mathrm{p,GRG}}{l}\right)^2} f_L(l)\,\mathrm{d}l} & \text{if } l_\mathrm{p} \geq l_\mathrm{p,GRG}. \end{cases} \tag{8.2}$$

This result follows from combining Eq. 8.1 and Eq. A.12 from Oei et al. (2023a).[2] As PDFs follow from CDFs by differentiation, we find that the PDFs of L_p and $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ relate as

$$f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p}) = \begin{cases} 0 & \text{if } l_\mathrm{p} < l_\mathrm{p,GRG}; \\ \dfrac{f_{L_\mathrm{p}}(l_\mathrm{p})}{\int_{l_\mathrm{p,GRG}}^{\infty} \sqrt{1-\left(\frac{l_\mathrm{p,GRG}}{l}\right)^2} f_L(l)\,\mathrm{d}l} & \text{if } l_\mathrm{p} \geq l_\mathrm{p,GRG}. \end{cases} \tag{8.3}$$

We note that, throughout the support of $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$, $f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$ and $f_{L_\mathrm{p}}(l_\mathrm{p})$ are directly proportional — the quantity in the denominator of Eq. 8.3 is merely a normalisation constant.[3]

To find $f_{L_\mathrm{p}}(l_\mathrm{p})$ if $l_\mathrm{p} > l_\mathrm{min}$, it can be helpful to perform a change of variables. By defining $\eta := \frac{l}{l_\mathrm{p}}$, we rewrite

$$F_{L_\mathrm{p}}(l_\mathrm{p}) = 1 - l_\mathrm{p} \int_1^\infty \sqrt{1-\frac{1}{\eta^2}} f_L(l_\mathrm{p}\eta)\,\mathrm{d}\eta \quad \text{if } l_\mathrm{p} > l_\mathrm{min}. \tag{8.5}$$

This form has the advantage that — within the integral — l_p occurs only in the integrand, whereas the form of Eq. 8.1 features l_p in both the integrand and in the lower integration limit. By differentation,

$$\begin{aligned} f_{L_\mathrm{p}}(l_\mathrm{p}) = &- \int_1^\infty \sqrt{1-\frac{1}{\eta^2}} f_L(l_\mathrm{p}\eta)\,\mathrm{d}\eta \\ &- l_\mathrm{p} \int_1^\infty \sqrt{1-\frac{1}{\eta^2}} \frac{\mathrm{d}f_L(l_\mathrm{p}\eta)}{\mathrm{d}l_\mathrm{p}}\,\mathrm{d}\eta \quad \text{if } l_\mathrm{p} > l_\mathrm{min}. \end{aligned} \tag{8.6}$$

To arrive at concrete expressions for the GRG projected proper length PDF of Eq. 8.3, we must choose a specific parametric form for the distribution of L or L_p.

8.2.3 GRG projected proper length: curved power law

Oei et al. (2023a) have shown that models that assume a Paretian tail for the RG in-

[2] We have also assumed that $l_\mathrm{p,GRG} > l_\mathrm{min}$, which is the obvious case to consider.

[3] This is an example of a more general rule: for any RV X,

$$f_{X \mid X \geq y}(x) = \begin{cases} 0 & \text{if } x < y; \\ \frac{f_X(x)}{1-F_X(y)} & \text{if } x \geq y. \end{cases} \tag{8.4}$$

trinsic proper length distribution, and that include angular and surface brightness selection effects, can tightly reproduce the observed GRG projected proper length distribution. The PDF of a Pareto-distributed RV is a simple power law, which is fully specified by a lower cut-off l_{min} and a tail index ξ. However, there is a good reason to believe that the true GRG projected proper length PDF deviates from simple power law behaviour. The true RG projected proper length PDF f_{L_p} will peak around a value set by the typical jet power, environment, lifetime, and inclination angle (amongst other properties). Below this value, f_{L_p} will necessarily be an increasing function of l_p; above this value, f_{L_p} will be a decreasing function.[4] As giants embody the large-length tail of the distribution of L_p, it is likely that the slope of $f_{L_\mathrm{p}\,|\,L_\mathrm{p}\geq l_{\mathrm{p,GRG}}}(l_\mathrm{p})$ at least somewhat decreases (i.e. steepens) as l_p increases — even in log–log space.

To remain close to the seemingly effective Pareto assumption of Oei et al. (2023a), we assume in this work that, at least for $l_\mathrm{p} \geq l_{\mathrm{p,GRG}}$, the RG projected proper length PDF is a curved power law:

$$f_{L_\mathrm{p}}(l_\mathrm{p}) \propto \left(\frac{l_\mathrm{p}}{l_{\mathrm{p,GRG}}}\right)^{\xi(l_\mathrm{p})} \quad \text{if } l_\mathrm{p} \geq l_{\mathrm{p,GRG}}, \tag{8.7}$$

where the exponent

$$\xi(l_\mathrm{p}) \coloneqq \xi(l_{\mathrm{p},1}) + \frac{l_\mathrm{p} - l_{\mathrm{p},1}}{l_{\mathrm{p},2} - l_{\mathrm{p},1}} \left(\xi(l_{\mathrm{p},2}) - \xi(l_{\mathrm{p},1})\right) \tag{8.8}$$

is a linear function of l_p. As long as $l_{\mathrm{p},1} \neq l_{\mathrm{p},2}$, both projected proper length constants can be chosen arbitrarily; however, $l_{\mathrm{p},1} \coloneqq l_{\mathrm{p,GRG}}$ seems to be a natural choice. Adopting this choice, and defining $\Delta\xi \coloneqq \xi(l_{\mathrm{p},2}) - \xi(l_{\mathrm{p},1})$, leads to the final exponent formula

$$\xi(l_\mathrm{p}) = \xi(l_{\mathrm{p,GRG}}) + \frac{l_\mathrm{p} - l_{\mathrm{p,GRG}}}{l_{\mathrm{p},2} - l_{\mathrm{p,GRG}}} \Delta\xi. \tag{8.9}$$

We adopt $\xi(l_{\mathrm{p,GRG}})$ and $\Delta\xi$ as two parameters of our model. We furthermore choose $l_{\mathrm{p},2} \coloneqq 5$ Mpc, which is close to the largest currently known radio galaxy projected proper length (Oei et al., 2022a, 2023a). Being the first-order Taylor polynomial of an arbitrary function $\xi(l_\mathrm{p})$ at $l_{\mathrm{p,GRG}}$, Eq. 8.8 represents a natural generalisation of the constant tail index assumption of Oei et al. (2023a). In particular, if model parameter $\Delta\xi = 0$, we recover the earlier Paretian model.

By the same reasoning as before, we find that if the RG projected proper length

[4] This line of reasoning implicitly assumes that the distribution of L_p is unimodal.

PDF is a curved power law for $l_\mathrm{p} \geq l_\mathrm{p,GRG}$, then the GRG projected proper length PDF is also a curved power law over this range:

$$f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p}) \propto \left(\frac{l_\mathrm{p}}{l_\mathrm{p,GRG}}\right)^{\xi(l_\mathrm{p})} \quad \text{if } l_\mathrm{p} \geq l_\mathrm{p,GRG}. \tag{8.10}$$

The factors required to normalise $f_{L_\mathrm{p}}(l_\mathrm{p})$ and $f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$ can be obtained numerically.

Whereas Oei et al. (2023a) parametrised $f_L(l)$ and derived $f_{L_\mathrm{p}}(l_\mathrm{p})$ and $f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$, we now parametrise $f_{L_\mathrm{p}}(l_\mathrm{p})$ and derive only $f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$. It is possible to start modelling at the level of $f_L(l)$, also in the context of curved power law PDFs, but the resulting expressions for $f_{L_\mathrm{p}}(l_\mathrm{p})$ and $f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$ become tedious and rather uninsightful. For simplicity, we therefore choose to parametrise $f_{L_\mathrm{p}}(l_\mathrm{p})$; we explore the alternative set-up in Appendix 8.A1.

8.2.4 GRG observed projected proper length

Equation 8.10 describes a distribution of GRG projected proper lengths in the absence of observational selection effects. Unfortunately, this distribution cannot be directly tested against GRG samples obtained from surveys, which are always affected by selection. For a thorough description and derivation of selection effect modelling in the context of our framework, we refer the reader to Sect. 2.8 and Appendix A.8 of Oei et al. (2023a); here, we shall only briefly introduce the expressions that we require.

A key result, adopted from Eq. 21 of Oei et al. (2023a), is that the GRG observed projected proper length RV $L_\mathrm{p,obs} \mid L_\mathrm{p,obs} \geq l_\mathrm{p,GRG}$ can be expressed as

$$f_{L_\mathrm{p,obs} \mid L_\mathrm{p,obs} \geq l_\mathrm{p,GRG}}(l_\mathrm{p}) = \begin{cases} 0 & \text{if } l_\mathrm{p} < l_\mathrm{p,GRG}; \\ \dfrac{C(l_\mathrm{p}) f_{L_\mathrm{p}}(l_\mathrm{p})}{\int_{l_\mathrm{p,GRG}}^{\infty} C(l''_\mathrm{p}) f_{L_\mathrm{p}}(l''_\mathrm{p})\, \mathrm{d}l''_\mathrm{p}} & \text{if } l_\mathrm{p} \geq l_\mathrm{p,GRG}, \end{cases} \tag{8.11}$$

where $C(l_\mathrm{p}) = C(l_\mathrm{p}, z_\mathrm{max})$ is the completeness function. More precisely, $C(l_\mathrm{p}, z_\mathrm{max})$ denotes the fraction of all RGs with projected proper length l_p in the volume up to cosmological redshift z_max that is detected and identified through the survey considered — in this work, this will be LoTSS DR2. The repeated factors in numerator and denominator reveal that, in order to compute $f_{L_\mathrm{p,obs} \mid L_\mathrm{p,obs} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$, we need to know $f_{L_\mathrm{p}}(l_\mathrm{p})$ for $l_\mathrm{p} \geq l_\mathrm{p,GRG}$ only — and within this range up to a constant only. More concerningly, we also see that selection effects that reduce the completeness by the same factor for all $l_\mathrm{p} \geq l_\mathrm{p,GRG}$ leave no imprint on $f_{L_\mathrm{p,obs} \mid L_\mathrm{p,obs} \geq l_\mathrm{p,GRG}}(l_\mathrm{p})$. Therefore, such selection effects cannot be constrained by a GRG observed projected proper length

analysis alone.

Under the assumption that the RG projected proper length PDF $f_{L_\mathrm{p}}(l_\mathrm{p})$ does not evolve between redshifts $z = z_\mathrm{max}$ and $z = 0$, the completeness function becomes

$$C\left(l_\mathrm{p}, z_\mathrm{max}\right) = \frac{\int_0^{z_\mathrm{max}} p_\mathrm{obs}\left(l_\mathrm{p}, z\right) r^2\left(z\right) E^{-1}\left(z\right)\, \mathrm{d}z}{\int_0^{z_\mathrm{max}} r^2\left(z\right) E^{-1}\left(z\right)\, \mathrm{d}z}, \tag{8.12}$$

where the observing probability $p_\mathrm{obs}(l_\mathrm{p}, z)$ is the probability that an RG of projected proper length l_p at redshift z is detected by a survey and its subsequent analysis steps (such as the machine learning pipeline considered in this work), r denotes comoving radial distance, and $E(z)$ is the dimensionless Hubble parameter[5]. The appropriate form of $p_\mathrm{obs}(l_\mathrm{p}, z)$ is determined by the selection effects relevant to the survey of interest and its analysis.

In this work, we will consider GRG lobe surface brightness selection, which at present renders part of the GRG population inherently undetectable (on an individual level, at least), and selection by limitations of our analysis steps, which causes in principle detectable giants to evade sample inclusion. We describe the former effect parametrically, and determine the latter effect empirically. The effects yield functions $p_\mathrm{obs,SB}(l_\mathrm{p}, z)$ and $p_\mathrm{obs,ID}(l_\mathrm{p}, z)$, respectively, which then combine to form a single observing probability function through

$$p_\mathrm{obs}(l_\mathrm{p}, z) = p_\mathrm{obs,SB}(l_\mathrm{p}, z) \cdot p_\mathrm{obs,ID}(l_\mathrm{p}, z). \tag{8.14}$$

Selection effects: surface brightness limit

RG lobes whose surface brightnesses are lower than some threshold value $b_{\nu,\mathrm{th}}$, which typically equals the survey noise level σ times a low factor of order unity, cannot be detected. Following Sect. 2.8.3 of Oei et al. (2023a), we model surface brightness (SB) selection by assuming that the lobe surface brightnesses $B_\nu(\nu, l, z)$ at $\nu = \nu_\mathrm{obs}$ of radio galaxies of intrinsic proper length $l = l_\mathrm{ref}$ that reside at redshift $z = 0$ are lognormally distributed. More precisely, we parametrise $B_\nu(\nu_\mathrm{obs}, l_\mathrm{ref}, 0) = b_{\nu,\mathrm{ref}}\, S$, where $b_{\nu,\mathrm{ref}}$ is the median SB, and S is a lognormally distributed RV with median 1 and dispersion

[5] In a *flat* FLRW universe, the dimensionless Hubble parameter E is

$$E\left(z\right) := \frac{H(z)}{H_0} = \sqrt{\Omega_{\mathrm{R},0}\left(1+z\right)^4 + \Omega_{\mathrm{M},0}\left(1+z\right)^3 + \Omega_{\Lambda,0}}. \tag{8.13}$$

parameter σ_{ref}. The observing probability due to SB selection then is

$$p_{\mathrm{obs,SB}}\left(l_{\mathrm{p}}, z\right) = \int_{s_{\mathrm{min}}}^{\infty} \sqrt{1 - \left(\frac{s_{\mathrm{min}}}{s}\right)^{-\frac{2}{\zeta}}} f_S\left(s\right) \mathrm{d}s; \tag{8.15}$$

$$s_{\mathrm{min}} = \frac{b_{\nu,\mathrm{th}}}{b_{\nu,\mathrm{ref}}} \left(\frac{l_{\mathrm{p}}}{l_{\mathrm{ref}}}\right)^{-\zeta} (1+z)^{3-\alpha}; \tag{8.16}$$

$$f_S\left(s\right) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{ref}} s} \exp\left(-\frac{\ln^2 s}{2\sigma_{\mathrm{ref}}^2}\right). \tag{8.17}$$

Here, α is the typical RG lobe spectral index, which we will assume fixed at $\alpha = -1$. The exponent ζ determines how the SB distribution scales with projected proper length l_{p}.

In contrast to the choice made in Oei et al. (2023a), we do not fix $\zeta = -2$, but rather leave ζ a free parameter which we fit to the data. Deviations from $\zeta = -2$ occur in at least two cases: when giant growth is not shape-preserving, and if the radio luminosity distributions of giants of different l_{p} are distinct. Dynamical models of radio galaxies in general predict that both cocoons (e.g. Fig. 4 of Turner & Shabala, 2015) and lobes (e.g. Fig. 9 of Hardcastle, 2018) change shape over time, with a dependence on jet power. There remains considerable uncertainty as to how shapes change throughout the giant phase: axial ratio–like measures generally show that RG lobes become more elongated during growth, but this trend could possibly reverse for giants, whose lobes might protrude from the clusters and filaments in which they are born. Simulations suggest that, for such protrusions, the usual constant power law profile assumptions for the ambient baryon density and temperature break down (e.g. Fig. 8 of Gheller & Vazza, 2019). If lobes of giants widen over time, then ζ would decrease. The second case occurs if the end-of-life lengths of radio galaxies increase with jet power, so that the subpopulation that survives up to some l_{p} has its jet power distribution — and thus its radio luminosity distribution — shifted upwards with respect to subpopulations at smaller l_{p}. This effect, which appears plausible given models (e.g. Fig. 8 of Hardcastle, 2018), would increase ζ. At present, it seems hard to predict the net result on ζ of these counteracting effects.

Selection effects: nonidentification

Every current-day survey search method (such as visual inspection by scientists, visual inspection by citizen scientists, and machine learning–based approaches) will fail to identify some giants that are in principle identifiable (in the sense that they lie above the detection threshold set by the noise). For automated approaches, such

as the machine learning–based approach presented in this work, identification can become more challenging for larger angular lengths φ: one reason being the increased number of unrelated, interloping radio sources that cover the solid angle occupied by the RG. We call the probability that an identifiable RG is indeed identified — and thus makes it into the final sample — $p_{\mathrm{obs,ID}}(l_{\mathrm{p}}, z)$.

Say we have M methods to search for giants in the same survey. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ be the set of all identified giants (so that $|\mathcal{G}| = N$), and let $\mathcal{G}_i \subseteq \mathcal{G}$ be the subset identified by method i. Figure 8.1 provides an overview of the set-up. We take $l_{\mathrm{p}}(g)$ and $z(g)$ to mean the projected proper length and cosmological redshift of giant g. To determine the identification probability $p_{\mathrm{obs,ID},i}(l_{\mathrm{p}}, z)$ for method i, we first assume it to be of logistic form

$$p_{\mathrm{obs,ID},i}(l_{\mathrm{p}}, z) = \frac{1}{1 + \exp\left(-(\beta_{0,i} + \beta_{l_{\mathrm{p}},i} \cdot l_{\mathrm{p}} + \beta_{z,i} \cdot z)\right)}. \tag{8.18}$$

We obtain best-fit parameters $\hat{\beta}_{0,i}$, $\hat{\beta}_{l_{\mathrm{p}},i}$, and $\hat{\beta}_{z,i}$ by performing binary logistic regression with two explanatory variables on the set of pairs $\mathcal{D}_i$, where

$$\mathcal{D}_i := \left\{ \left(\left[l_{\mathrm{p}}(g), z(g) \right], \mathbb{I}(g \in \mathcal{G}_i) \right) \mid g \in \bigcup_{j=1, j\neq i}^{M} \mathcal{G}_j \right\}. \tag{8.19}$$

The first element of each pair is a point in projected length–redshift space, whilst the second element is 0 or 1: $\mathbb{I}$ denotes the indicator function. Qualitatively, $\mathcal{D}_i$ stores for each giant in the union of all GRG subsets except $\mathcal{G}_i$ its projected length–redshift coordinates, together with the success or failure of its identification by method i.

The implicit assumption here is that all $g \in \bigcup_{j=1, j\neq i}^{M} \mathcal{G}_j$ are typical examples of identifiable giants at the relevant projected proper length and redshift. We caution that this might not be true: giants with a peculiar morphology, or those lying in parts of the sky where optical identification is hard (e.g. towards the Galactic Plane or crowded regions of large-scale structure), may be identifiable in a radio surface brightness sense, but will nonetheless evade sample inclusion more often than other giants. As a result, giants that *do* end up in a sample — such as $\bigcup_{j=1, j\neq i}^{M} \mathcal{G}_j$ — will have more regular morphologies than giants in general and will lie in regions of the sky where optical identification is easier than for giants in general. Typically, such giants are also more likely to be found by method i, and as a result our approach will probably render $p_{\mathrm{obs,ID},i}$ biased high.

Given a set of M functions $\{p_{\mathrm{obs,ID},i}(l_{\mathrm{p}}, z) \mid i \in \{1, 2, \ldots, M\}\}$, several possibilities exist to combine them into a single $p_{\mathrm{obs,ID}}(l_{\mathrm{p}}, z)$. At the minimum, $p_{\mathrm{obs,ID}}(l_{\mathrm{p}}, z)$ is

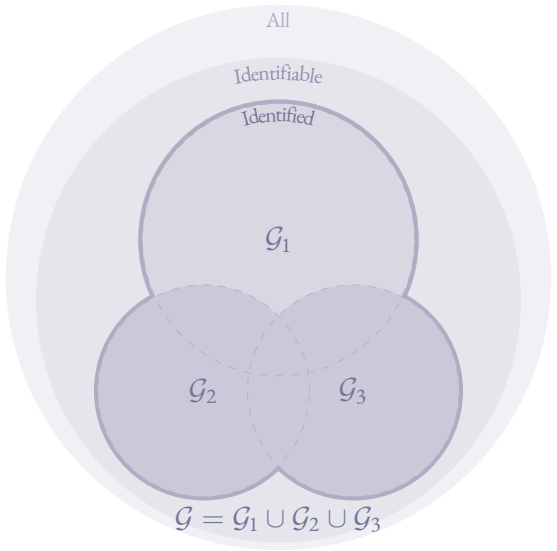

Figure 8.1: Schematic of a three-method search for giants. Of all giants in the survey footprint up to $z = z_{\rm max}$, only those for which the lobe surface brightness at the observing frequency $\nu_{\rm obs}$ is above detection threshold $b_{\nu,\rm th}$ are identifiable. $\mathcal{G}$ denotes the actually identified set of giants. $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ are the subsets identified by each method individually. As an example, we shade $\mathcal{G}_2 \cup \mathcal{G}_3$, which has overlap with $\mathcal{G}_1$, and which can be used to measure $p_{\rm obs,ID,1}(l_{\rm p}, z)$.

given by a point-wise maximum:

$$p_{\rm obs,ID}(l_{\rm p}, z) = \max_{i\in\{1,2,\ldots,M\}} p_{\rm obs,ID,i}(l_{\rm p}, z), \tag{8.20}$$

which is appropriate if methods tend to find the same identifiable giants — as in our case.[6]

8.2.5 GRG number density

The preceding theory allows us to find the intrinsic, comoving number density of giants, $n_{\rm GRG}$, if we know the observed number of giants within a solid angle of extent Ω and in the volume up to $z_{\rm max}$, $N_{\rm GRG,obs}(\Omega, z_{\rm max})$. We assume that, up to this redshift, $n_{\rm GRG}$ remains constant. We note that we cannot calculate $n_{\rm GRG}$ using Eq. 30 from

[6] In case methods tend to find independent subsets of identifiable giants,

$$p_{\rm obs,ID}(l_{\rm p}, z) = 1 - \prod_{i=1}^{M}(1 - p_{\rm obs,ID,i}(l_{\rm p}, z)). \tag{8.21}$$

We note that it is possible to design methods that find subsets of identifiable giants that have even less overlap than independent subsets have.

Oei et al. (2023a): this equation assumes $\xi(l_{\mathrm{p},1}) = \xi(l_{\mathrm{p},2})$. We derive a more general expression by first noting that the number of giants observed within a solid angle of extent Ω in the volume up to z_{max} and with projected proper lengths between l_{p} and $l_{\mathrm{p}} + \mathrm{d}l_{\mathrm{p}}$ is

$$\mathrm{d}N_{\mathrm{GRG,obs}}(l_{\mathrm{p}}, \Omega, z_{\mathrm{max}}) = \frac{\Omega}{4\pi}\, n_{\mathrm{GRG}} f_{L_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}}(l_{\mathrm{p}})\, \mathrm{d}l_{\mathrm{p}} \cdot \int_0^{z_{\mathrm{max}}} p_{\mathrm{obs}}(l_{\mathrm{p}}, z)\, 4\pi r^2(z)\, \frac{\mathrm{d}r}{\mathrm{d}z}\, \mathrm{d}z. \tag{8.22}$$

Because

$$N_{\mathrm{GRG,obs}}(\Omega, z_{\mathrm{max}}) = \int_{l_{\mathrm{p,GRG}}}^{\infty} \mathrm{d}N_{\mathrm{GRG,obs}}(l_{\mathrm{p}}, \Omega, z_{\mathrm{max}}), \tag{8.23}$$

we find — by isolating n_{GRG} — that

$$n_{\mathrm{GRG}}(l_{\mathrm{p,GRG}}, z_{\mathrm{max}}) = \frac{H_0}{c}\, \frac{4\pi}{\Omega}\, N_{\mathrm{GRG,obs}}(\Omega, z_{\mathrm{max}}) \cdot \left(\int_{l_{\mathrm{p,GRG}}}^{\infty} f_{L_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}}(l_{\mathrm{p}}) \int_0^{z_{\mathrm{max}}} p_{\mathrm{obs}}(l_{\mathrm{p}}, z)\, 4\pi r^2(z)\, E^{-1}(z)\, \mathrm{d}z\, \mathrm{d}l_{\mathrm{p}} \right)^{-1}. \tag{8.24}$$

This expression is valid also beyond the context of power law or curved power law PDFs $f_{L_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}}(l_{\mathrm{p}})$. We remark that n_{GRG} can depend sensitively on the projected proper length used to define giants, $l_{\mathrm{p,GRG}}$.

In contrast to the approach of Oei et al. (2023a), in this work we do not calculate n_{GRG} in a step *following* inference of the framework's parameters, but rather include it as a parameter to be constrained *during* inference.

8.2.6 GRG lobe volume-filling fraction

To constrain the contribution of giants to astrophysical magnetogenesis, we wish to know the volume-filling fraction of their lobes in clusters and filaments of the Cosmic Web. Under the approximation that GRG shapes are independent of their volumes, we have

$$\mathcal{V}_{\mathrm{GRG-CW}}(z = 0) = \mathbb{E}[\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}] \cdot \mathbb{E}[L_{\mathrm{p}}^3 \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}] \cdot n_{\mathrm{GRG}} \cdot (\mathcal{V}_{\mathrm{CW}}(z = 0))^{-1}, \tag{8.25}$$

where $\Upsilon_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ is a random variable denoting the ratio between a giant's combined lobe volume and its cubed projected proper length: $\Upsilon_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG} \coloneqq \frac{V}{L_\mathrm{p}^3} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$. The distribution of $\Upsilon_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ can be inferred by fitting geometric models to GRG images, as has been explored in Oei et al. (2022a, 2023b). $\mathcal{V}_\mathrm{CW}(z = 0)$ denotes the volume-filling fraction of clusters and filaments in the Local Universe.

8.2.7 GRG angular lengths

An object's angular length φ, projected proper length l_p, and cosmological redshift z are related through

$$\varphi(l_\mathrm{p}, z) = \frac{l_\mathrm{p}(1+z)}{r(z)}. \tag{8.26}$$

Due to the expansion of the Universe, there exists a minimum angular length for objects of a given projected proper length. If one defines giants as radio galaxies with projected proper lengths $l_\mathrm{p} \geq l_\mathrm{p,GRG} \coloneqq 0.7$ Mpc, as in this work, then all giants have an angular length $\varphi \geq 1.3'$ (Oei et al., 2023a). This fact has important consequences for GRG search campaigns. At the LoTSS resolution of $\theta_\mathrm{FWHM} = 6''$, it implies that giants are always resolved sources, spanning at least 13 resolution elements. Thus, to model the detectability of giants at this resolution, one must consider their surface brightness (profiles), rather than their flux densities.

8.2.8 Inference

Finally, we detail how the framework's six free parameters $\theta \coloneqq [\xi(l_\mathrm{p,GRG}), \Delta\xi, b_{\nu,\mathrm{ref}}, \sigma_\mathrm{ref}, \zeta, n_\mathrm{GRG}]$ can be inferred from a dataset containing a projected length and redshift for each observed giant. In particular, we consider a rectangle in projected proper length–cosmological redshift parameter space, within which our model assumptions are expected to hold. We partition this rectangle into N_b equiareal bins of width Δl_p and height Δz. We denote the coordinates of bin i's centre as $(l_{\mathrm{p},i}, z_i)$.

On the data side, we construct a two-dimensional histogram using these bins. The number of giants found in bin i, N_i, is a random variable with a Poisson distribution: $N_i \sim \mathrm{Poisson}(\lambda_i)$. Its expectation λ_i depends on the model parameters θ. Assuming that the $\{N_i\}$ are independent, the log-likelihood becomes

$$\ln \mathcal{L}(\{N_i\} \mid \theta) = \sum_{i=1}^{N_\mathrm{b}} N_i \ln \lambda_i(\theta) - \lambda_i(\theta) - \ln(N_i!). \tag{8.27}$$

The last term on the right-hand side of Eq. 8.27 is the same for all θ, and need not be calculated if one is interested in $\mathcal{L}$ up to a global constant only.[7] Following Eq. 8.22, but avoiding integration over z and assuming narrow bins in both dimensions, we approximate

$$\lambda_i \approx n_{\mathrm{GRG}} V_i \cdot f_{L_\mathrm{p} \mid L_\mathrm{p} \geq l_{\mathrm{p,GRG}}}(l_{\mathrm{p},i}) \Delta l_\mathrm{p} \cdot p_\mathrm{obs}(l_{\mathrm{p},i}, z_i), \tag{8.29}$$

where the volume in which the giants of bin i fall, V_i, is

$$V_i = \Omega r^2(z_i) \Delta r_i, \text{ with } \Delta r_i = \frac{c}{H_0} \frac{\Delta z}{E(z_i)}. \tag{8.30}$$

Appendix 8.A2 details a particularly efficient trick to compute the likelihood for a range of n_{GRG}, whilst leaving the other parameters fixed. By multiplying the likelihood function with a prior distribution, for which we shall choose a uniform distribution, we obtain a posterior distribution over θ — up to a constant.

8.3 Data

We applied our automated radio–optical catalogue creation methods to all total intensity (Stokes-I) maps from LoTSS DR2 (Shimwell et al., 2022).[8] The observations in LoTSS DR2 cover the 120–168 MHz frequency range, have a 6″ resolution, a median RMS sensitivity of 83 μJy beam^{-1}, and a flux density scale uncertainty of approximately 10%. The observations are split into a region centred at 12h45m +44°30′ and a region centred at 1hoom +28°00′; both avoid the Galactic Plane. These regions span 4,178 and 1,457 square degrees respectively, and together cover 27% of the Northern Sky. The observations consist of 841 partly overlapping pointings with diameters of 4.0°. The vast majority of the pointings were observed for 8h, all within the 2014-05-23 to 2020-02-05 time frame.

Apart from the LoTSS DR2 Stokes-I maps, the ML radio catalogue pipeline that we describe in this manuscript (Sect. 8.4), relies on an infrared-optical source catalogue. This catalogue combines the positions, magnitudes and colour information of

[7] If one includes the term, it only needs to be calculated once. For numerical stability, it is helpful to note that

$$-\sum_{i=1}^{N_\mathrm{b}} \ln\left(N_i!\right) = -\sum_{i=1}^{N_\mathrm{b}} \sum_{j=2}^{N_i} \ln j. \tag{8.28}$$

[8] LoTSS DR2 is publicly available at https://lofar-surveys.org/dr2_release.html.

the infrared sources of the unWISE data release (Schlafly et al., 2019) from the Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010), and the optical sources in the DESI Legacy Imaging Surveys DR9 (Dey et al., 2019).

We tried to maximise the identification of giants $p_{\mathrm{obs,ID}}$ within the LoTSS DR2 data. To do so, we complemented the sample of giants detected by our ML pipeline with all giants in the value-added LoTSS DR2 radio catalogue (Hardcastle et al., 2023). For large ($\varphi > 15''$) radio components (and thus most giants), the radio source component association and most of the optical host identifications of the value-added LoTSS DR2 radio catalogue, described by Hardcastle et al. (2023), were performed via a public project named 'LOFAR Galaxy Zoo' on Zooniverse. We will thus refer to the value-added LoTSS DR2 radio catalogue as the 'LGZ catalogue' and to the giants in that catalogue as the 'LGZ giants'. Zooniverse is an online citizen science platform for crowd-sourced visual inspection.[9] The detailed source component information provided by the LGZ catalogue allowed us to homogenise the angular length estimates of our ML pipeline giants and the LGZ giants (see Sect. 8.4.6). We further complemented our GRG sample with other GRG samples in the literature, see Sect. 8.4.8.

8.4 Methods

To derive the projected length distribution, number density, and lobe VFF for the intrinsic population of giants, we followed a two-stage approach. In the first stage, we gathered all giants that we detected in the LoTSS DR2 Stokes-I images using our automatic ML pipeline and added all other giants that we found in the LGZ catalogue. We re-evaluated and homogenised the source size estimates over the combined GRG sample, and manually inspected the plausibility of the associated radio source components and optical/infrared host galaxy. Finally, we merged this GRG sample with the other GRG samples from the literature. In the second stage, we search for the most likely parameters for the forward model presented in Sect. 8.2 that describe the GRG observed projected proper length distribution and the selection effects of the merged GRG sample. Figure 8.2 shows an overview of our approach.

8.4.1 Radio emission detection

We started out with the publicly available calibrated LoTSS DR2 Stokes-I images (Shimwell et al., 2022). For each of the 841 pointings, we ran the PyBDSF radio blob detection software (Mohan & Rafferty, 2015) using the same parameters as used in

[9] The Zooniverse website is `https://zooniverse.org`.

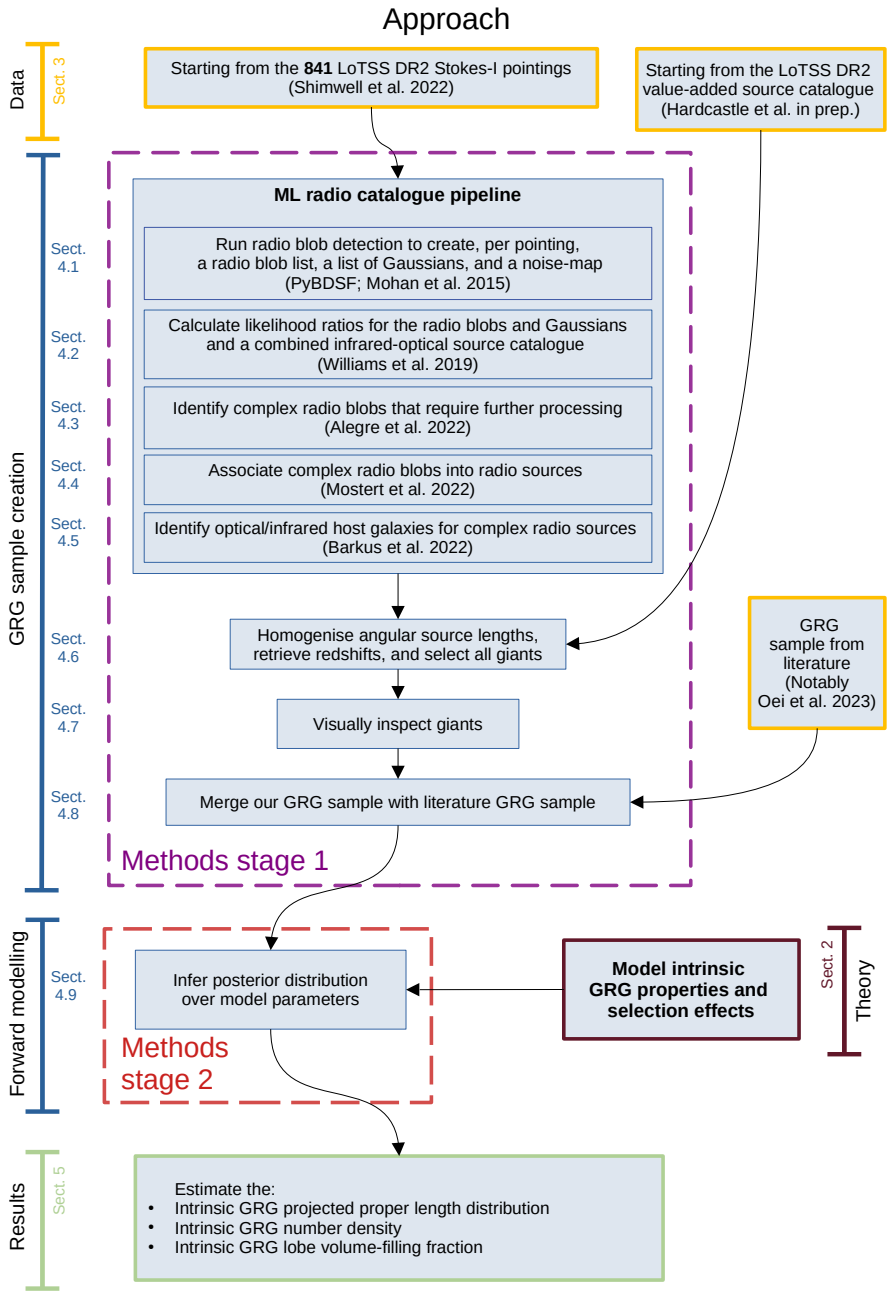

Figure 8.2: Overview of our approach, which consists of two stages. In the first stage we built a GRG sample, and in the second stage we inferred the properties of the intrinsic GRG population using a forward model. The brackets indicate the different parts of our approach and mention the sections containing the corresponding details.

LoTSS DR2 (Shimwell et al., 2022) — notably, this means that we used a 5σ detection threshold. Appendix 8.A3 provides the full list of PyBDSF parameters and their values.

The output we generated consists of a list of radio blobs with their location and properties. PyBDSF can decompose each radio blob it detects, into one or more 2D Gaussians. For each radio blob, we also saved the corresponding list of Gaussians. These Gaussians function as a source model for each radio blob and will be used in later steps in the ML pipeline.[10]

8.4.2 Calculating radio to optical/infrared likelihood ratios

For radio sources, the location of the host galaxy on the sky is close to the flux-weighted centre of the radio source.[11] The likelihood ratio method, which exploits this idea, quantifies the likelihood that a source in one spectral window is the correct counterpart to a source in another spectral window (e.g. Richter, 1975; de Ruiter et al., 1977; Sutherland & Saunders, 1992). Williams et al. (2019) used this method to cross-match the unresolved — and some resolved — radio sources of LoTSS DR1 to a combined catalogue of infrared and optical sources. More specifically, the infrared sources came from AllWISE (Cutri et al., 2021), whilst the optical sources came from the Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1; Chambers et al., 2016) DR1 3π steradian survey. The likelihood ratio function that Williams et al. (2019) used is a function of the angular distance between the flux-weighted centre of the radio source and the flux-weighted centre of the optical or infrared source, the magnitude of the optical or infrared source, and the colour of the optical or infrared source. The likelihood ratio function also takes into account uncertainties in each of these three dependencies.

We adopted the same procedure as detailed by Williams et al. (2019) to cross-match our simple radio sources (where 'simple' is to be understood as in Sect. 8.4.3) to a combined catalogue of infrared and optical sources. The infrared sources came from unWISE (Schlafly et al., 2019), and the optical sources were now taken from the DESI Legacy Imaging Surveys DR9 (Dey et al., 2019), which boasts deeper imagery than Pan-STARRS1 DR1 used for LoTSS DR1. The unWISE (Schlafly et al., 2019) and DESI Legacy Imaging Surveys DR9 source catalogues will be used for LoTSS DR2 cross-matching more generally (Hardcastle et al., 2023). Per pointing, we applied the likelihood ratio method to the full list of radio blobs and to the full list of Gaussians.

[10] However, as we discuss in Sect. 8.4.6 these source models are not always adequate for extended, well-resolved radio sources.

[11] In this context, 'close' refers to angular distances comparable to the survey's resolution.

For both the blobs and the Gaussians, we stored the identifier of the optical or infrared source that produced the highest likelihood ratio, alongside this highest likelihood ratio itself.

8.4.3 Sorting radio emission with a gradient boosting classifier

Most radio sources that consist of a single radio blob (mostly unresolved or barely resolved radio sources) can be cross-matched using the likelihood ratio method. However, some resolved radio sources, and certainly most resolved giant radio galaxies, consist of multiple radio blobs, and thus require radio blob association and cannot be cross-matched using the likelihood ratio alone. To separate the simple from the complex radio blobs in LoTSS DR1, a considerable amount of visual inspection was applied (Williams et al., 2019). For LoTSS DR2, Alegre et al. (2022) trained a gradient boosting classifier (GBC; Breiman, 1997; Friedman, 2001) to classify radio blobs as either 'simple' or 'complex' based on the properties of the radio blobs, the properties of the Gaussians fitted to these blobs, the likelihood ratios for each, and the distance to and properties of the nearest neighbours.

We adopted the procedure of Alegre et al. (2022) and use their trained GBC to separate the simple radio blobs from those that require radio component association beyond PyBDSF's capabilities and/or optical host identification beyond the scope of the likelihood ratio method as described by Sutherland & Saunders (1992). We expect most giant radio galaxies to fall in the latter case.

8.4.4 Associating radio emission into radio sources

We proceeded with automatic radio source component association for the complex radio blobs. Following the procedure laid out by Mostert et al. (2022), for each of these radio blobs, we created a $300'' \times 300''$ LoTSS DR2 image cutout centred on the radio blob. Next, a Fast region-based convolutional neural network (Fast R-CNN; Girshick, 2015), adapted and trained for this purpose by Mostert et al. (2022), was applied to these cutouts to predict which (if any) other radio blobs — whether they be complex or simple — should be associated to the centred radio blob for it to form a single physical radio source. For example, the two lobes of an RG, each represented by a radio blob, might be associated together to form a single physical radio source. Due to the fixed $300'' \times 300''$ image size for which the Fast R-CNN was trained, we expect most radio sources that are associated in our pipeline to have an angular length $\varphi < 424''$.[12]

[12] If predicted associations from neighbouring cutouts have an overlapping radio blob, the associations will be merged. For example: in cutout 1 lobe A and core B are associated and in cutout 2

The result is a radio source catalogue in which some of the radio blobs have been merged, and a component catalogue that lists for each radio blob to which radio source it belongs. The radio and the component catalogue were completed by appending to them the remaining list of simple radio blobs.

8.4.5 Optical or infrared host galaxy identification

Barkus et al. (2022) created a method to identify an extended radio source's optical or infrared host. The method described by Barkus et al. (2022) takes the radio morphology into account by drawing a ridgeline along the regions of high flux density. The method continues with the application of the likelihood ratio method to quantify which pairs of host galaxy candidates and radio sources are a plausible match. The likelihood ratio LR used in this context follows Eq. 1 of Sutherland & Saunders (1992), with the slight simplification of having the latter's dependence on two angular offsets replaced by a dependence on a *radial* angular offset only:

$$LR = \frac{q(m, c)f(r)}{n(m, c)}, \tag{8.31}$$

where $q(m, c)$ is a prior on the magnitude m and colour c of the optical host, $f(r)$ is a function of the angular offset between the optical centroid and the radio centroid, and $n(m, c)$ normalises for the number density of optical sources with a certain magnitude m and colour c in the catalogue used for the cross-matching.

To adapt the likelihood ratio for use in the case of extended radio sources, Barkus et al. (2022) implemented the different components of the ratio as follows. For $n(m, c)$, Barkus et al. (2022) estimated the probability density over m and c for a distribution of 50, 000 randomly sampled sources from a combined Pan-STARRS–AllWISE catalogue in the region of the sky that overlapped with LoTSS DR1. For $q(m, c)$, they estimated the probability over m and c for sources from the combined Pan-STARRS–AllWISE catalogue that were manually selected to be the most likely optical/near-infrared host for a sample of 950 radio sources with angular length $\varphi > 15''$. For both $n(m, c)$ and $q(m, c)$, the AllWISE W1 magnitudes were used for m, the Pan-STARRS i-band magnitudes minus the AllWISE W1 magnitudes were used for colour c, and the PDF was formed using a 2D kernel density estimator (KDE; e.g. Pedregosa et al., 2011) with a Gaussian kernel and a bandwidth of 0.2. For extended asymmetric or bent radio galaxies, the optical host is not likely to be found at the radio centroid.

core B and lobe C are associated, then the set (lobe A, core B, and lobe C) will enter the catalogue as a single radio source. Thereby creating the possibility of detecting radio sources with angular length $\varphi > 424''$.

Thus, Barkus et al. (2022) proposed $f(r)$ to be a function of both the distance between the radio centroid and the optical source $r_{\text{opt,centroid}}$ and the smallest distance between the optical source and a ridgeline fitted to the radio source $r_{\text{opt,ridge}}$. Specifically,

$$f(r) = f_{\text{ridge}}(r_{\text{opt,ridge}}) \cdot f_{\text{centroid}}(r_{\text{opt,centroid}}), \tag{8.32}$$

with

$$f_{\text{ridge}}(r_{\text{opt,ridge}}) = \frac{1}{2\pi\sigma_r^2} e^{\frac{-r_{\text{opt,ridge}}^2}{2\sigma_r^2}}, \tag{8.33}$$

and

$$f_{\text{centroid}}(r_{\text{opt,centroid}}) = \frac{1}{2\pi\sigma_c^2} e^{\frac{-r_{\text{opt,centroid}}^2}{2\sigma_c^2}}, \tag{8.34}$$

where $\sigma_r^2 = \sigma_{\text{opt}}^2 + \sigma_{\text{radio}}^2 + \sigma_{\text{astr}}^2$ and the chosen value for the astrometric uncertainty σ_{astr} is $0.2''$, the optical position uncertainties σ_{opt} are taken from the optical catalogue (generally $\sim 0.1''$), the radio position uncertainty σ_{radio} is fixed to $3''$, and the uncertainty in the centroid position σ_c is empirically estimated at 0.2 times the length of the considered radio source. The parameters in f_{ridge} are expressed in arcsec and those in f_{centroid} as a fraction of the radio source length. For the 30 optical sources closest to the radio ridgeline Barkus et al. (2022) calculate the value of LR, whereby the optical source with the highest LR value is considered to be the most likely host galaxy.

We use the method by Barkus et al. (2022) but propose three minor adaptations. First, we introduce explicit regularisation for $q(m, c)$ and $n(m, c)$. As the PDF estimates for $q(m, c)$ and $n(m, c)$ are 2D KDEs over sampled (m, c)-distributions, the parts of the (m, c)–parameter space that are sparsely sampled can lead to probabilities that are effectively zero when the realistic theoretical probability should be small but non-zero. Through the q/n-fraction in Eq. 8.31, the resulting values of LR in the sparsely sampled parts of the (m, c)–parameter space blow up to unrealistic large values or collapse to almost 0 (see Fig. 8.12). In practice, these unsampled parts of parameter space are almost never visited by new sources for which we calculate LR. Even so, we add a constant factor to the KDE estimate of q and n to get more robust LR values (see Fig. 8.13) and to express the model uncertainties in our functions of q and n. Using 10-fold cross-validation, we empirically select the bandwidths for the KDEs leading to q and n to be 0.4. Second, we propose an alternate form of $f(r)$. For giants, $f(r)$ is rarely dominated by errors in the position of the optical source or that of the radio source. As $r_{\text{opt,centroid}}$ and $r_{\text{opt,ridge}}$ are slightly correlated, multiplication of $f_{\text{ridge}}(r_{\text{opt,ridge}})$ and $f_{\text{centroid}}(r_{\text{opt,centroid}})$ underestimates the chance of low values of $r_{\text{opt,centroid}}$ or $r_{\text{opt,ridge}}$. Therefore, we combine $r_{\text{opt,centroid}}$ and $r_{\text{opt,ridge}}$ into a single parameter r_{mean} that is the mean of the two distance parameters. Furthermore, we ob-

serve that the empirical distributions of $r_{\text{opt,centroid}}$, $r_{\text{opt,ridge}}$ and r_{mean} for a sample of radio sources with angular length $> 1'$ $A_{\text{radio,opt}}$ for which optical counterparts were determined via visual inspection do not follow a normal distribution as assumed by Barkus et al. (2022) but rather a lognormal distribution (see Fig. 8.14). Instead of estimating the values of the different error components (astrometric error, error in optical position, error in radio position) we use the empirical values of the distribution of $f(r)$ for the sources in $A_{\text{radio,opt}}$; see Appendix 8.A4 for details. Third, we replaced the Pan-STARRS1 DR1 catalogue (from which colour c was derived) with the DESI Legacy Imaging Surveys DR9 catalogue, as the latter goes up to an i-band magnitude of 24.

We applied the modified ridgeline method to all radio sources in our pipeline catalogue with angular lengths larger than $1'$ and brighter than 10 mJy. We limit the ridgeline procedure to these sources to save time, as the procedure takes multiple seconds per radio source.

After detecting the optical host galaxies for our radio sources, we checked for corresponding spectroscopic redshift estimates from SDSS (VizieR catalogue `V/147/sdss12`, Ahn et al., 2012), or if not available, for photometric redshift estimates from DESI (VizieR catalogue `VII/292`, Duncan, 2022). The SDSS catalogue also provides us with velocity dispersions and a quasar flag. The DESI VizieR catalogue includes a flag (FCLEAN $= 1$) that indicates that the optical source used for the photometric redshift prediction is free from blending or image artefacts. The catalogue also includes a column (PSTAR) that estimates how likely it is that the optical source is a star based on its colours. We discard all sources in both the pipeline catalogue and the LGZ catalogue for which either FCLEAN $\neq 1$ or PSTAR > 0.2.

8.4.6 Re-assessing angular source lengths

Next, we proceeded to re-assess the angular source lengths, for both the radio catalogue created using the ML pipeline and those reported by the LGZ catalogue described by Hardcastle et al. (2023). Up to this point, the angular source lengths in both catalogues are the full width at half maximum (FWHM) of the combined Gaussian components that make up a source, if the source is only composed of a single radio blob. If the radio source is composed of multiple radio blobs, the reported size is the distance between the two furthest removed points on a convex hull that encloses the FWHMs of the blobs that make up the radio source. However, in the literature, the length of a giant is often reported to be the maximum distance between the signal of a radio source that exceeds three times the image noise σ.

To get the 3σ angular lengths, we applied five steps to all sources in both catalogues

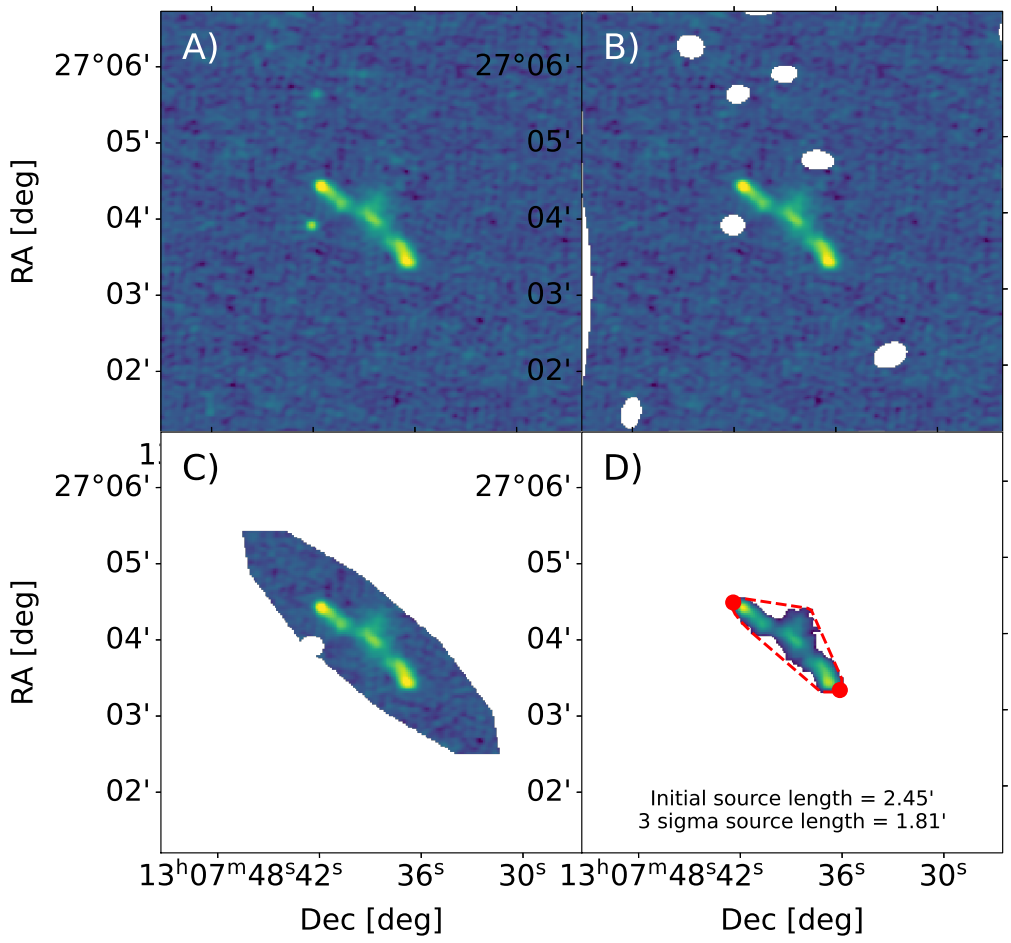

Figure 8.3: Four panels, depicting the re-evaluation of the angular length of radio source ILTJ130738.79+270355.1. Panel A) shows the initial cutout, B) shows the removal of neighbouring sources, C) the masking of emission outside a convex hull based on the old source length, D) shows the emission that is left after masking all emission below 3 times the local noise σ. The red dashed line shows the convex hull around the left-over emission, and the red points indicate the furthest removed points on this convex hull. The distance between these points yields the 3σ angular length.

with a reported angular length $\varphi > 1'$. First, we created a square image cutout with a width and height equal to 1.5 times the old angular source length. Second, we mask all neighbouring radio emission. Third, we mask all emission outside an ellipse with a major axis that is the old source length, and a minor axis that is 1.1 times the old source width or a quarter of the old source length if that value is bigger. These numbers are a result of the observation that, with respect to the 3σ angular lengths, the old lengths were almost always significantly overestimated, while the source width tended to be underestimated. Fourth, we mask all remaining emission that is below 3 times

the local noise. Fifth, we fitted a convex hull around the remaining emission and determined the distance between the points on this convex hull that were farthest apart. See Fig. 8.3 for an illustrative example.

The entire process from source detection (Sect. 8.4.1) to source list with optical identifications and updated angular lengths (this subsection) took roughly half an hour to one hour per LoTSS DR2 pointing, depending on the detected number of sources. Each pointing can be processed independently, which allowed us to spread the processing of all 841 LoTSS DR2 pointings over 5 nodes of a heterogeneous computer cluster with 80 physical CPU cores in total for three to four days.

Finally, for both the ML pipeline and LGZ catalogues, we calculate the projected proper lengths using the 3σ angular lengths and the redshift estimates corresponding to each source, and discard all sources that do not meet the $l_\mathrm{p} \geq l_\mathrm{p,GRG} = 0.7\,\mathrm{Mpc}$ GRG criterion. For the ML pipeline catalogue, we discarded all internally duplicate GRG candidates using a $1'$ cone search. The LGZ catalogue did not contain any internal duplicates. That left us with $7,001$ GRG candidates in the pipeline catalogue and $7,044$ GRG candidates in the LGZ catalogue.

8.4.7 Manual verification of obtained GRG sample

The following step we took in the creation of our GRG sample, was a manual visual inspection of all giant candidates. For the LGZ giant candidates, as described by Hardcastle et al. (2023), at least five different volunteers already inspected the radio and corresponding optical emission. The purpose of our manual visual inspection was therefore to exclude only those sources where either the radio component association or the host identification was obviously incorrect. For each giant, a single expert looked at a panel showing the giant with its neighbouring sources masked and most neighbouring emission masked (akin to panel C in Fig. 8.3) and a panel showing the giant in its wider context (akin to panel A in Fig. 8.3); additionally, the location of the optical host was indicated. We sorted the GRG candidates into three categories: candidates that looked reasonable, candidates that clearly missed (or included too many) significant radio components, and candidates that showed a very unlikely host galaxy location. For the ML pipeline giant candidates, we initially followed the same procedure as for the LGZ giant candidates. To make visual inspection feasible, we skipped the $4,272$ ML pipeline giant candidates that overlapped with the verified LGZ giants. After inspecting the ML pipeline giant candidates once, we subjected all giants that were not rejected to a second round of visual inspection. The second round was aided by inspecting LoTSS DR2 radio contours over a Legacy Survey DR9 (g, r, z) image cube, where sources from the combined optical–infrared catalogue within the

field of view were highlighted.

For the LGZ catalogue, we judged 6,550 (93%) GRG candidates to be without issues, 389 (6%) to have radio component issues, and 105 (1%) to have been assigned an unlikely host galaxy. For the 5,864 (unique) ML pipeline giant candidates, we judged 2,722 (47%) candidates to be without issues, 1,963 (33%) to have radio component issues, and 1,179 (20%) to have been assigned an unlikely host galaxy. From the 6,550 LGZ giants, we find 5,596 of those to be newly discovered (not appearing in previous literature), and for the 2,722 ML pipeline giants, we find 2,592 to be newly discovered.

Qualitatively, from the visual inspection, we noticed that the verified ML pipeline GRG sample contained more symmetric giants with colinear jets, while the verified LGZ GRG sample contained more giants with complex, bent structures indicative of interaction with the IGM. The ML pipeline did also detect giants with complex structures, but was often unable to fully separate them from all neighbouring unrelated emission. An in-depth comparison between the ML pipeline and LGZ GRG samples is beyond the scope of this work. Figures 8.4 and 8.5 each show six examples of previously unknown giants found through our ML-based approach. Through cutouts covering $3' \times 3'$, Fig. 8.4 shows angularly compact giants; through cutouts covering $6' \times 6'$, Fig. 8.5 shows more angularly extended specimen.

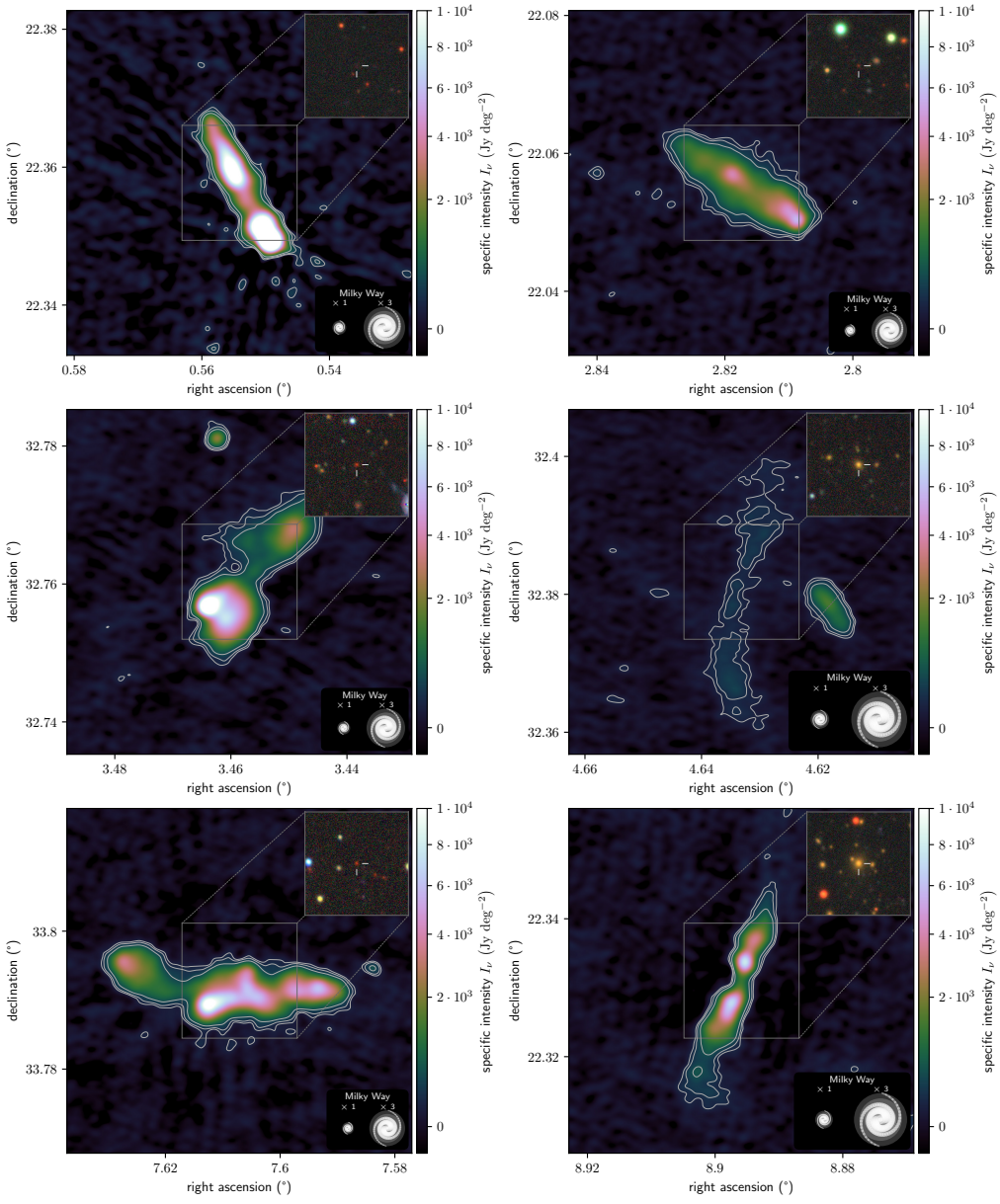

Figure 8.4: LoTSS DR2 cutouts at central observing frequency $\nu_{\text{obs}} = 144$ MHz and resolution $\theta_{\text{FWHM}} = 6''$, centred around the hosts of newly discovered giants. Each cutout covers a solid angle of $3' \times 3'$. Contours signify 3, 5, and 10 sigma-clipped standard deviations above the sigma-clipped median. For scale, we show the stellar Milky Way disk (with a diameter of 50 kpc) generated using the Ringermacher & Mead (2009) formula, alongside a 3 times inflated version. Each DESI Legacy Imaging Surveys DR9 (g, r, z) inset shows the central $1' \times 1'$ square region. As all giants obey $\varphi \geq 1.3'$, they must — if not oriented along one of the square's diagonals — necessarily protrude from this region. Rowwise from left to right, from top to bottom, these giants are ILTJ000212.45+222116.2, ILTJ001115.77+220316.6, ILTJ001350.25+324530.8, ILTJ001831.84+322247.7, ILTJ003025.90+334729.2, and ILTJ003534.45+221937.8.

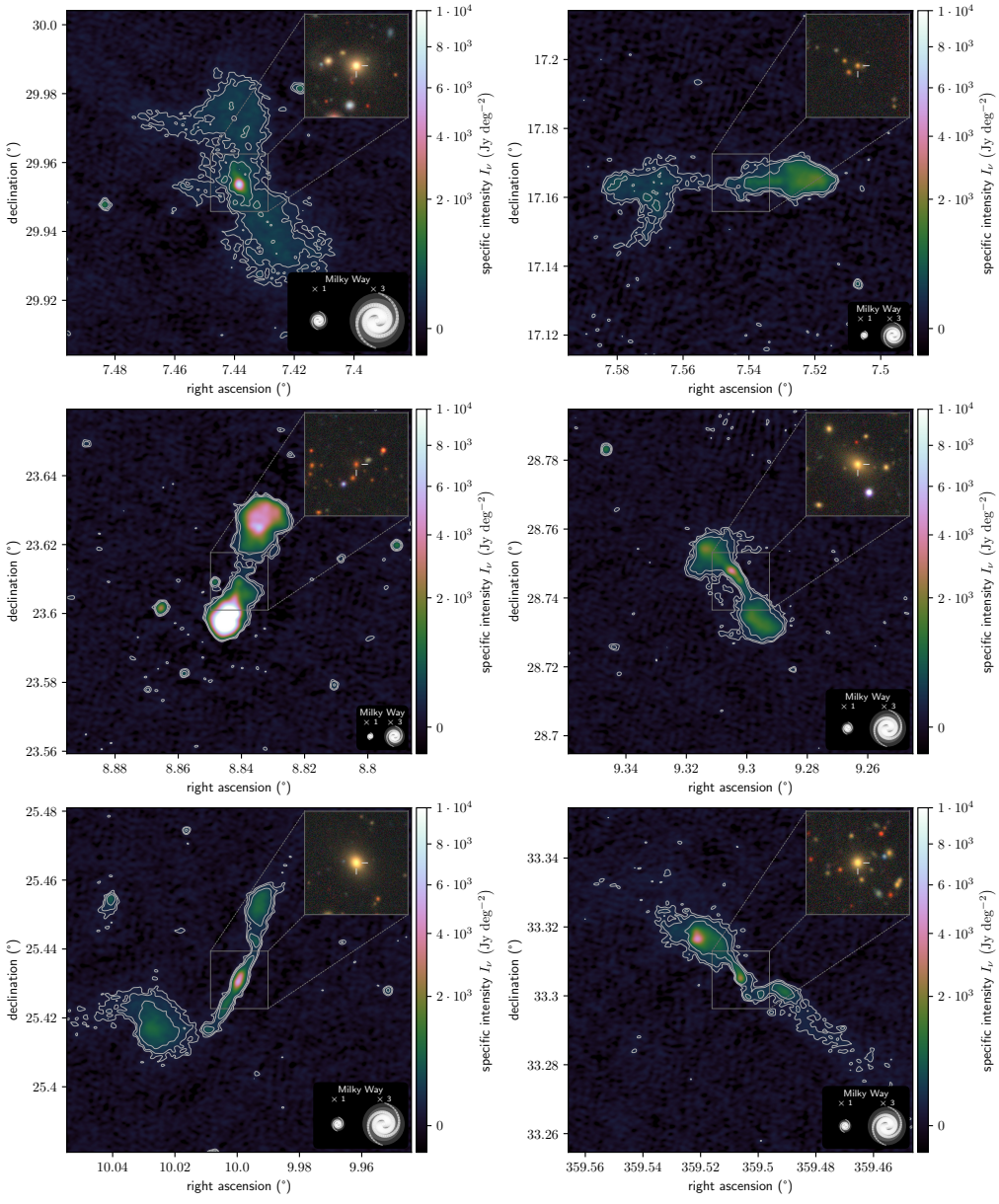

Figure 8.5: LoTSS DR2 cutouts at central observing frequency $\nu_{\text{obs}} = 144$ MHz and resolution $\theta_{\text{FWHM}} = 6''$, centred around the hosts of newly discovered giants. Each cutout covers a solid angle of $6' \times 6'$. Contours signify 3, 5, and 10 sigma-clipped standard deviations above the sigma-clipped median. For scale, we show the stellar Milky Way disk (with a diameter of 50 kpc) generated using the Ringermacher & Mead (2009) formula, alongside a 3 times inflated version. Each DESI Legacy Imaging Surveys DR9 (g, r, z) inset shows the central $1' \times 1'$ square region. As all giants obey $\varphi \geq 1.3'$, they must — if not oriented along one of the square's diagonals — necessarily protrude from this region. Rowwise from left to right, from top to bottom, these giants are ILTJ002943.72+295700.3, ILTJ003010.58+170948.6, ILTJ003521.87+233625.9, ILTJ003712.91+284436.8, ILTJ004002.30+252550.9, and ILTJ235802.49+331838.5.

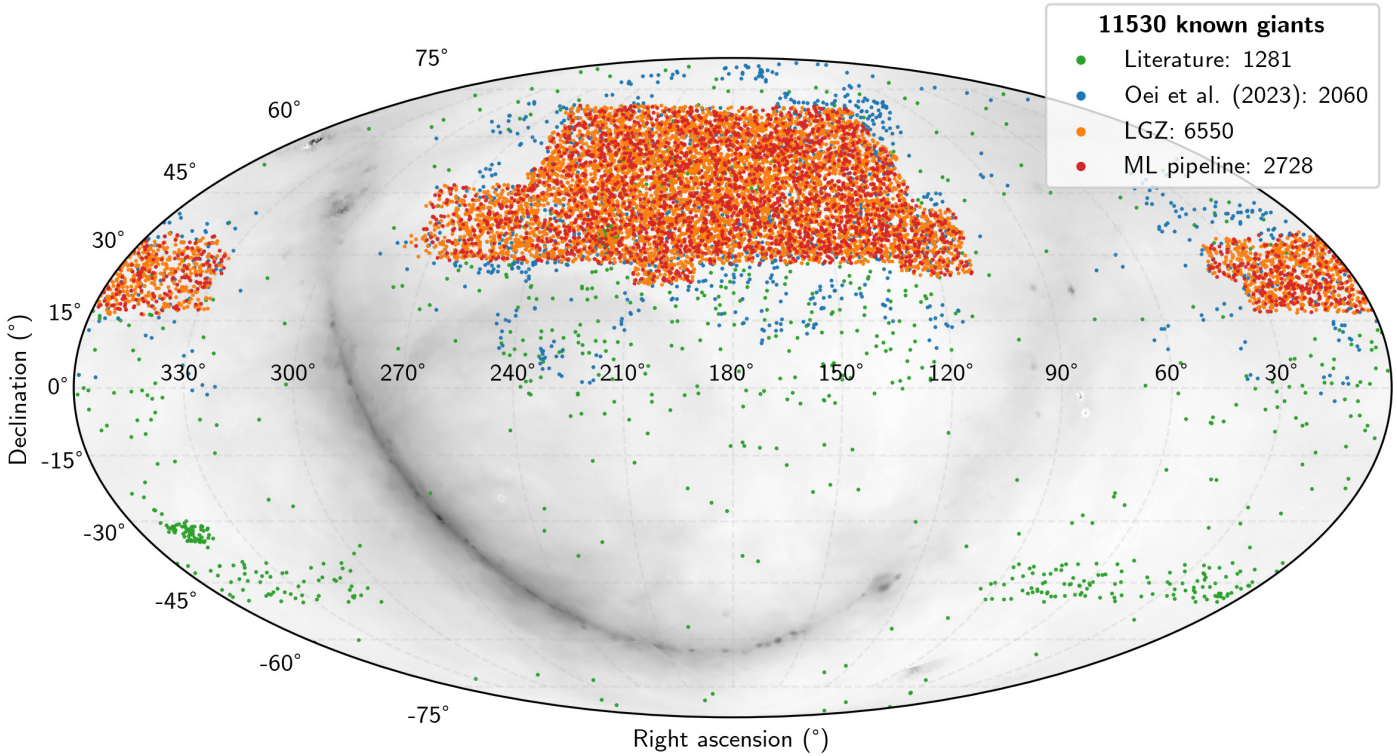

Figure 8.6: With 11, 524 unique sources, we present the largest collection of giants discovered as of yet. The ML pipeline and LGZ samples (red and orange markers) are strictly confined to the LoTSS DR2 area, while the sample by Oei et al. (2023a) extends to yet-to-be-released LoTSS pointings processed with the DR2 pipeline.

8.4.8 Merging our sample with the GRG sample in the literature

To complete our GRG sample, we iteratively added giants from the literature, going from the newest to the oldest publication. This approach follows from the assumption that newer publications are generally based on more sensitive and higher-resolution observations, leading to more accurate angular length estimates. In an effort to avoid having duplicate giants in the final sample, we only added giants when their host galaxies were more than $1'$ away from all host galaxies of the already aggregated giants.

The LGZ–ML pipeline GRG sample contains 9, 272 giants. We added 1, 432 out of the 2, 193 giants presented by Oei et al. (2023a), 41 out of the 69 giants presented by Simonte et al. (2022), 62 out of the 62 giants presented by Gürkan et al. (2022), 163 out of the 263 giants presented by Mahato et al. (2022), 178 out of the 178 giants presented by Andernach et al. (2021), 0 out of the 1 giants presented by Masini et al. (2021), 2 out of the 2 giants presented by Delhaize et al. (2021), 1 out of the 2 giants presented by Bassani et al. (2021), 1 out of the 4 giants presented by Tang et al. (2020), 372 out of the 694 giants presented by Dabhade et al. (2020a), and 0 out of the 6 giants presented by Ishwara-Chandra et al. (2020). These additions result in the final catalogue containing 11, 524 unique giants. This is the first catalogue of giants to contain more than 10^4 specimen.

Figure 8.6 shows a Mollweide view of the sky with the locations of both the newly confirmed giants and the giants from the literature. Almost all discovered giants stay clear of the Galactic Plane, where radio emission from the Milky Way — of which we show the specific intensity function at $\nu_{\rm obs} = 150$ MHz in greyscale (Zheng et al., 2017) — makes calibration and imaging harder. In addition, optical host identification is much harder near the Galactic Plane. The default field of view set-up of both our ML pipeline (Sect. 8.4.4) and of LGZ favours the discovery of giants with angular lengths of a few arcminutes at most. By contrast, the GRG campaign of Oei et al. (2023a) featured a 'fuzzy' $\sim 5'$ lower threshold to allow for an exhaustive manual search with an interactive and dynamic field of view (using Aladin; Bonnarel et al., 2000). In Fig. 8.7, we demonstrate that these design choices lead to GRG samples with markedly different angular length distributions.

As a result, the samples complement each other: the sample of Oei et al. (2023a) is more complete at lower redshifts and higher projected lengths, while the LGZ and ML pipeline samples are more complete at higher redshifts and lower projected lengths. Figure 8.8 demonstrates this point, while Table 8.1 presents the corresponding statistics of the GRG samples.

For comparison of the 3σ lengths of the ML pipeline and LGZ giants to those in

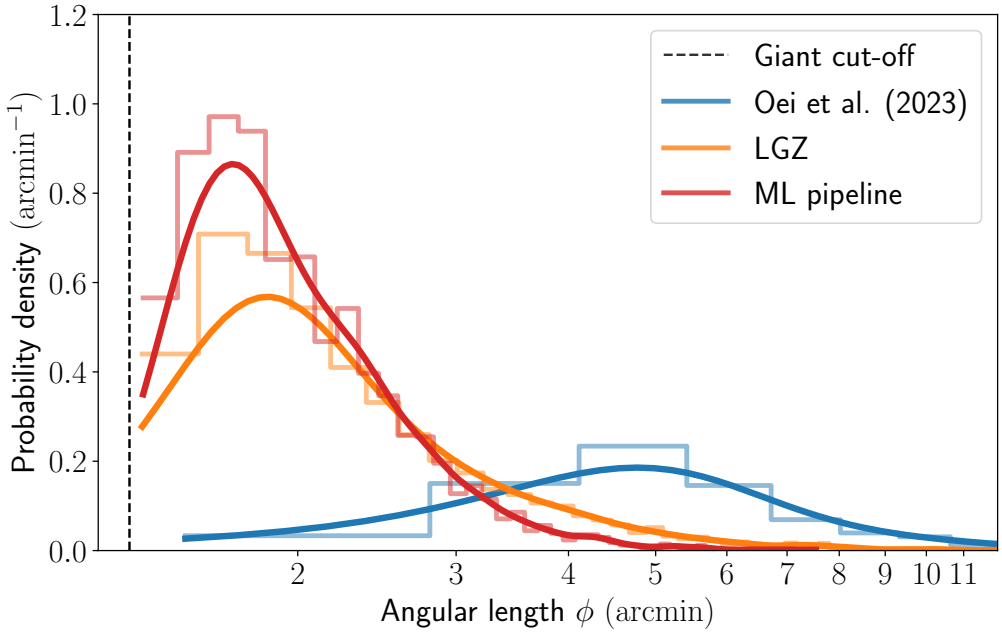

Figure 8.7: Observed distributions for angular length φ, showing that our three LoTSS DR2 search methods target different ranges of φ. The largest angular lengths detected by Oei et al. (2023a), LGZ, and the ML pipeline are 132′, 43′, and 8′ respectively, but we limit the horizontal axis to 12′ for interpretability. The vertical line marks the minimum angular length that giants can attain: $\varphi_{\mathrm{GRG}}(l_{\mathrm{p,GRG}} = 0.7\,\mathrm{Mpc}) = 1.3'$.

Table 8.1: Statistics of the GRG (sub)samples that we discovered, confirmed, or aggregated. From left to right, the columns provide the number of giants in each sample, N, and the 10th, the median, and the 90th percentile of the angular length φ, redshift z, and projected proper length l_{p}.

Sample	N	$\varphi_{10\mathrm{th}}$ (′)	$\varphi_{\mathrm{median}}$ (′)	$\varphi_{90\mathrm{th}}$ (′)	$z_{10\mathrm{th}}$	z_{median}	$z_{90\mathrm{th}}$	$l_{\mathrm{p},10\mathrm{th}}$ (Mpc)	$l_{\mathrm{p,median}}$ (Mpc)	$l_{\mathrm{p},90\mathrm{th}}$ (Mpc)
ML pipeline	2,722	1.50	1.97	3.09	0.44	0.87	1.28	0.72	0.87	1.29
LGZ	6,550	1.56	2.19	4.40	0.31	0.75	1.19	0.73	0.91	1.57
Known giants	11,524	1.57	2.33	5.70	0.23	0.72	1.19	0.73	0.94	1.68

other surveys, we inform the reader that the central frequency and the average surface brightness threshold of the observations that we use are $\nu_{\mathrm{obs}} = 144$ MHz and $b_{\nu,\mathrm{th}} = 25\,\mathrm{Jy\,deg^{-2}}$ respectively.

8.4.9 Bayesian parameter estimation

After having refined our statistical GRG framework (Sect. 8.2), and after having assembled the largest sample of giants yet (Sects. 8.4.1–8.4.8), we combined both advances to perform inference of the length distribution, number density, and lobe volume-filling fraction of giants.

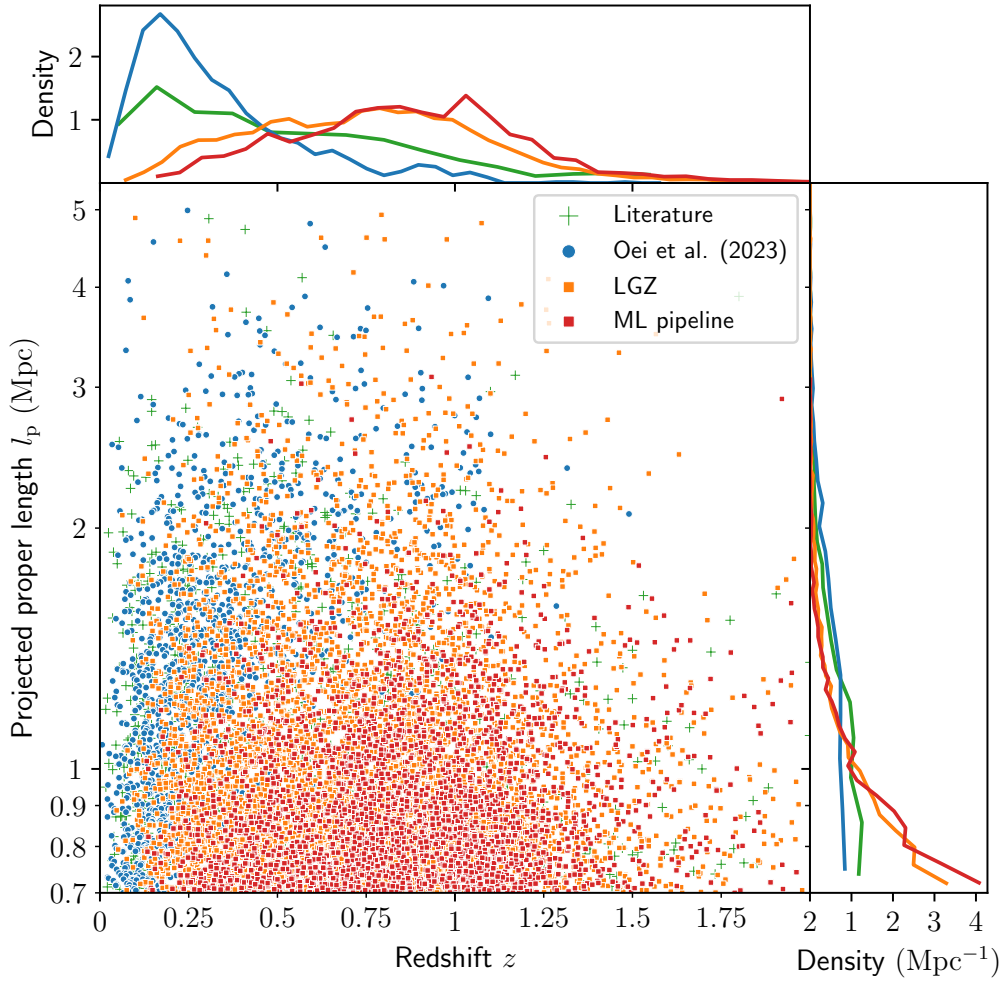

Figure 8.8: Our sample of LGZ giants (orange squares) and ML pipeline giants (red squares) effectively complements the sample of giants with large angular lengths (blue dots) from the manual search of Oei et al. (2023a). The remaining giants (green pluses) are from earlier literature, as specified in Sect. 8.4.8.

Given that our goal has been to infer properties of the full population of giants, rather than just of those currently observed, we included two main selection effects in our forward modelling. As detailed in Sect. 8.2.4, we parametrised surface brightness selection with three parameters, which are free parameters of the model. As detailed in Sect. 8.2.4, a second cause of selection is the imperfect operation of our three LoTSS DR2 search methods, all of which fail to identify a significant fraction of giants with lobe surface brightnesses *above* the survey noise level. We modelled this identification selection $p_{\mathrm{obs,ID}}$ with a set of logistic functions, regressed to GRG data.

We now provide details of this process.

Identification probability functions

To estimate $p_{\mathrm{obs,ID}}(l_{\mathrm{p}}, z)$ from data, we first selected all giants detected by the joint efforts of our machine learning pipeline, LGZ, and the manual, visual search of Oei et al. (2023a). Next, we retained only those giants that are located in regions of the sky that have been scanned by all three searches. This overlap region in principle corresponds to the full LoTSS DR2 coverage — were it not for the fact that the search of Oei et al. (2023a) skipped over the LoTSS DR1, which had already been scanned by Dabhade et al. (2020b). Thus, the actual overlap region amounts to the LoTSS DR2 coverage with a spherical quadrangle removed, whose minimum and maximum right ascensions are $\alpha_{\mathrm{min}} = 160^\circ$ and $\alpha_{\mathrm{max}} = 230^\circ$ and whose minimum and maximum declinations are $\delta_{\mathrm{min}} = 45^\circ$ and $\delta_{\mathrm{max}} = 56^\circ$. In Appendix 8.A5, we provide an explicit decomposition of our assumed LoTSS DR2 coverage — and thus implicitly of the overlap region — in terms of disjoint spherical quadrangles.

Some of the retained giants have been detected only in the combined ML–LGZ search, others have been detected only in the Oei et al. (2023a) search, and yet others have been detected in both. We note that, had it operated flawlessly, the combined ML–LGZ search would have detected all sources claimed by Oei et al. (2023a) (or at least those that are genuine giants — which should be the vast majority). Thus, by mapping the (in)ability of the ML–LGZ search to detect the giants of Oei et al. (2023a) as a function of l_{p} and z, one can estimate the ML–LGZ search's identification probability function, $p_{\mathrm{obs,ID,1}}(l_{\mathrm{p}}, z)$. More precisely, for each giant detected by Oei et al. (2023a), we evaluated whether it was also detected in the ML–LGZ search, and stored a corresponding Boolean (that is to say, either 1 or 0). We show these Booleans, at the (l_{p}, z) coordinates of the giants they belong to, as yellow (representing 1) and blue (representing 0) dots in the top-left panel of Fig. 8.9. Viewing the Boolean at (l_{p}, z) as a realisation of a Bernoulli RV with success probability $p = p_{\mathrm{obs,ID,1}}(l_{\mathrm{p}}, z)$, we recognise the inference of the identification probability function as a binary logistic regression problem with two explanatory variables. The background of Fig. 8.9's top-left panel shows the corresponding best fit.

By symmetry, this approach can be reversed to estimate the Oei et al. (2023a) search's identification probability function, $p_{\mathrm{obs,ID,2}}(l_{\mathrm{p}}, z)$. Thus, for each giant detected in the ML–LGZ search, we evaluated whether it was also detected by Oei et al. (2023a), and stored a corresponding Boolean. In the same way as before, we show these Booleans in the middle-left panel of Fig. 8.9. The panel's background shows the best logistic fit.

We combine the two identification probability functions, $p_{\mathrm{obs,ID},1}(l_\mathrm{p}, z)$ and $p_{\mathrm{obs,ID},2}(l_\mathrm{p}, z)$, in point-wise fashion as to obtain a single function $p_{\mathrm{obs,ID}}(l_\mathrm{p}, z)$. To do so, we follow the minimal combination rule of Eq. 8.20.

We remark that, by giving each Boolean in these logistic regressions an equal weight, the resulting functions are tuned to fit crowded regions of projected length–redshift parameter space best — at the expense of accuracy in sparser regions. To increase the accuracy of the functions for the parameter space at large, we performed a simple rebalancing step. First, we calculated the mean number density in the parameter space given by $l_\mathrm{p} \in [0.7, 5\ \mathrm{Mpc}] \times [0, 0.5] \ni z$. We then selectively subsampled the data in crowded regions, following the rule that the number density in each bin of width 0.5 Mpc and height 0.05 should not exceed twice the mean number density of the entire parameter space. We show the rebalanced data, alongside refitted logistic models, in the right column of Fig. 8.9. We report the rebalanced model coefficients in Table 8.2, and treat them as constants during the Bayesian inference.

Inference in practice

In this work, we constrain the parameters of Sect. 8.2's GRG population model via a projected length–redshift histogram. From our most extensive sample of giants, we select those with $0.7\ \mathrm{Mpc} =: l_\mathrm{p,GRG} < l_\mathrm{p} < 5.1\ \mathrm{Mpc}$ and $0 < z < z_\mathrm{max} := 0.5$ that lie in the LoTSS DR2 coverage as specified in Appendix 8.A5. We do not include the giants from Oei et al. (2023a) for which only a lower bound to the redshift is known. This selection retains 2,685 out of 11,524 giants. We use these giants to fill a histogram with bins of width $\Delta l_\mathrm{p} = 0.1\ \mathrm{Mpc}$ and $\Delta z = 0.02$. We have not systematically explored the effect of these bin size parameters on the resulting inference. However, the smaller one chooses the bins, the higher the numerical cost will be. On the other hand, if the bins are chosen much larger than the typical scales over which the underlying observed projected length–redshift distribution[13] varies, then some ability to extract parameter constraints will be lost.

To compute the posterior distribution over the six parameters $\theta = [\xi(l_\mathrm{p,GRG}), \Delta\xi, b_{\nu,\mathrm{ref}}, \sigma_\mathrm{ref}, \zeta, n_\mathrm{GRG}]$, we assumed a uniform prior and brute-force evaluated the likelihood function over a regular grid that covers a total of $2.1 \cdot 10^9$ parameter combinations.[14] In doing so, we applied the Poissonian likelihood trick described in Ap-

[13] With the 'underlying' observed projected length–redshift distribution, we mean the observed projected length–redshift distribution one would obtain in the limit of an infinite number of observed giants.

[14] This approach is feasible by virtue of the low numerical cost of each likelihood function evaluation. Its main advantage is its simplicity: there are no parameters to tune that govern the method's convergence behaviour. Once the model is expanded to include more parameters, or when selection ef-

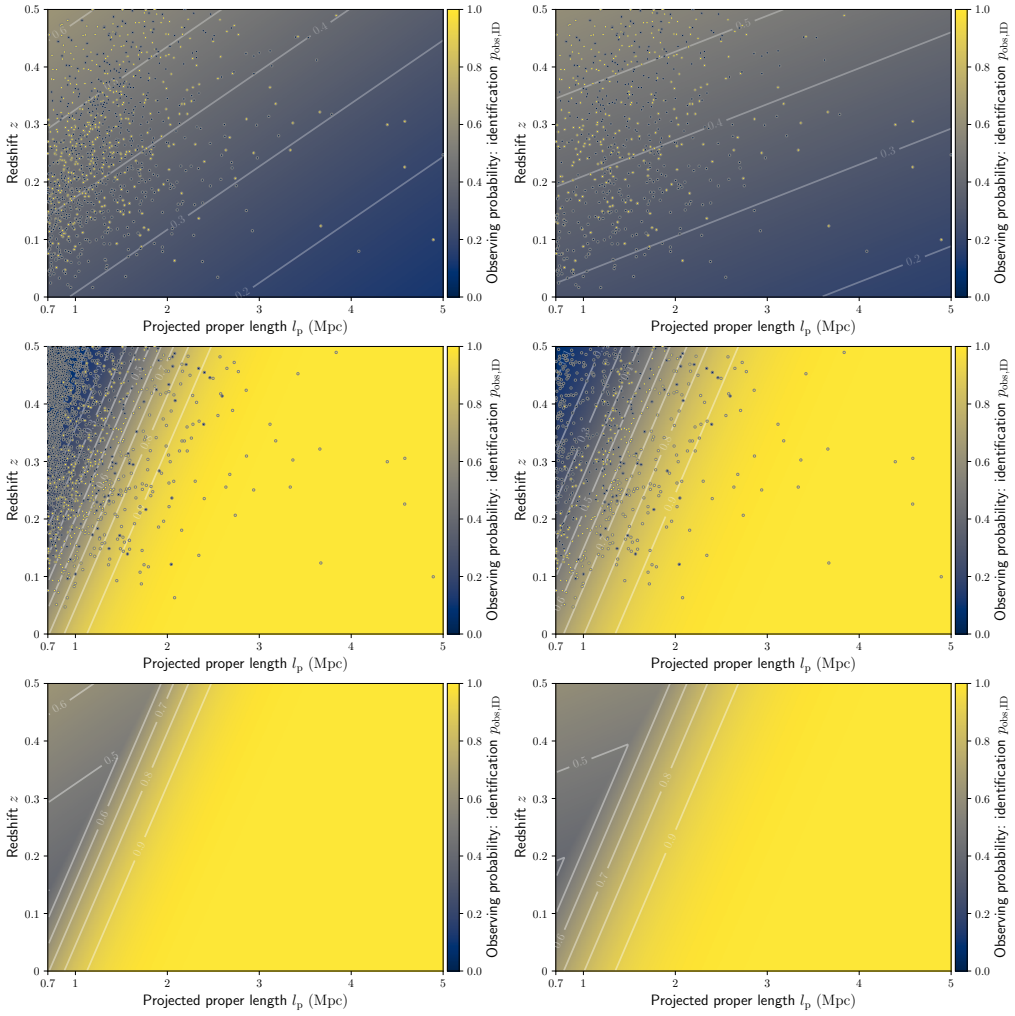

Figure 8.9: Overview of our determination of the probability to identify giants in the LoTSS DR2 with above-noise surface brightnesses, as a function of projected length and redshift — through machine learning or LOFAR Galaxy Zoo (top row), through the search of Oei et al. (2023a) (middle row), and through these methods in unison (bottom row). Each of the upper four panels shows a binary logistic regression following the theory of Sect. 8.2.4 and the practical considerations of Sect. 8.4.9. The left column shows results from all available data, whilst the right column shows results from rebalanced data. In our Bayesian inference, we use the latter results.

pendix 8.A2, which sped up our computations by one to two orders of magnitude. Table 8.2 provides the parameter ranges for which we evaluated the likelihood (which coincide with their prior distribution ranges), alongside all of the model's constants

fects with higher numerical cost are incorporated, more efficient (though more complicated) methods such as Markov chain Monte Carlo or nested sampling will become necessary.

and their assumed values. Because each likelihood function evaluation can be computed independently of the others, the problem is fully parallelisable. In practice, we distributed the $\sim 10^4$ core-hours Python calculation over $\sim$1500 virtual cores, which were spread across $\sim$20 nodes of a computer cluster. Next, we generated samples from the posterior distribution using rejection sampling (e.g. Rice, 2006). We subsequently used these samples to calculate probability distributions for derived quantities.[15]

8.5 Results

By combining an unparalleled sample of giant radio galaxies with a rigorous forward model, we have produced a posterior distribution over parameters that characterise the intrinsic population of giants. Figure 8.10 summarises the posterior over parameter hexads $\theta = [\xi(l_{\mathrm{p,GRG}}), \Delta\xi, b_{\nu,\mathrm{ref}}, \sigma_{\mathrm{ref}}, \zeta, n_{\mathrm{GRG}}]$ by means of its one- and two-dimensional marginal distributions. In this section, we analyse our newfound parameter constraints.

8.5.1 GRG length distribution

Radio galaxies enrich the IGM with magnetic fields, but giants — given their megaparsec-scale reach — appear uniquely capable of seeding the more remote regions of the Cosmic Web. Consequently, scientific interest in quantifying the length distribution of giants has arisen from the possibility that giants contribute significantly to cosmic magnetogenesis. The question at hand is deceivingly simple: how common are giants of various lengths?

As pointed out by Oei et al. (2023a), due to selection effects, the observed projected length distribution is not a reliable estimate of the *true* projected length distribution. Worse still, the relevant selection effects might not be quantitatively known a priori, requiring joint inference of the length distribution, and the selection effect parameters. Oei et al. (2023a) performed such a joint inference, and found that their data were consistent with an underlying population of giants with Pareto-distributed lengths, characterised by tail index $\xi = -3.4 \pm 0.5$. In the current work, we have relaxed the assumption of perfect Paretianity, and explore whether the data are consis-

[15] To calculate probability distributions over quantities that are a function of the parameters, such as the Local Universe GRG lobe VFF, $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$, or the joint search completeness function C, we could in principle evaluate these quantities for each parameter combination of the aforementioned grid and weigh each grid point's result by the associated likelihood (or, equivalently, posterior probability). However, some derived quantities are costly to compute, so that excessive evaluations should be avoided.

Table 8.2: Parameters and constants of Sect. 8.2's GRG population forward model alongside their prior ranges and values, as used in the Bayesian inference presented in Sect. 8.5. The first six constants serve to define the quantitative meaning of the parameters and set the scope of the analysis. The other eleven constants are not arbitrary: they affect the likelihood function and posterior distribution for a given set of parameter definitions and scope.

Parameter	Uniform prior range	Explanation
$\xi(l_{\mathrm{p},1} = l_{\mathrm{p,GRG}})$	$[-3.5, -2]$	Sect. 8.2.3
$\Delta\xi$	$[-3.5, -1.5]$	Sect. 8.2.3
$b_{\nu,\mathrm{ref}}$	$[1, 100] \cdot \mathrm{Jy\,deg}^{-2}$	Sect. 8.2.4
σ_{ref}	$[0.5, 2]$	Sect. 8.2.4
ζ	$[-0.5, 0]$	Sect. 8.2.4
n_{GRG}	$[0, 50] \cdot (100\,\mathrm{Mpc})^{-3}$	Sect. 8.2.5

Constant	Value	Explanation
$l_{\mathrm{p,GRG}}$	0.7 Mpc	Sect. 8.1
$l_{\mathrm{p},1}$	0.7 Mpc	Sect. 8.2.3
$l_{\mathrm{p},2}$	5 Mpc	Sect. 8.2.3
l_{ref}	0.7 Mpc	Sect. 8.2.4
ν_{obs}	144 MHz	Sect. 8.2.4
z_{max}	0.5	Sect. 8.2.4
α	-1	Sect. 8.2.4
$b_{\nu,\mathrm{th}}$	$25\,\mathrm{Jy\,deg}^{-2}$	Sect. 8.2.4
$\beta_{0,1}$	-1.0	Sect. 8.2.4
$\beta_{0,2}$	-1.0	Sect. 8.2.4
$\beta_{l_{\mathrm{p}},1}$	$-0.1\,\mathrm{Mpc}^{-1}$	Sect. 8.2.4
$\beta_{l_{\mathrm{p}},2}$	$2.4\,\mathrm{Mpc}^{-1}$	Sect. 8.2.4
$\beta_{z,1}$	2.8	Sect. 8.2.4
$\beta_{z,2}$	-6.4	Sect. 8.2.4
Δl_{p}	0.1 Mpc	Sect. 8.4.9
Δz	0.02	Sect. 8.4.9
Ω	1.62 sr	Appendix 8.A5

tent with a curved power law PDF for the GRG projected proper length RV $L_{\mathrm{p}} \,|\, L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}$. The marginals of Fig. 8.10 suggest that they are — in fact, the data strongly favour curvature, with a tail index at $l_{\mathrm{p},1} \coloneqq l_{\mathrm{p,GRG}} \coloneqq 0.7$ Mpc of $\xi(l_{\mathrm{p,GRG}}) = -2.8 \pm 0.2$ and a total increase in tail index up to $l_{\mathrm{p},2} \coloneqq 5$ Mpc of $\Delta\xi = -2.4 \pm 0.3$. Given the small relative uncertainty on the latter value, our data appear inconsistent with perfect Paretianity ($\Delta\xi = 0$). We note that our notion of 'data' is different from

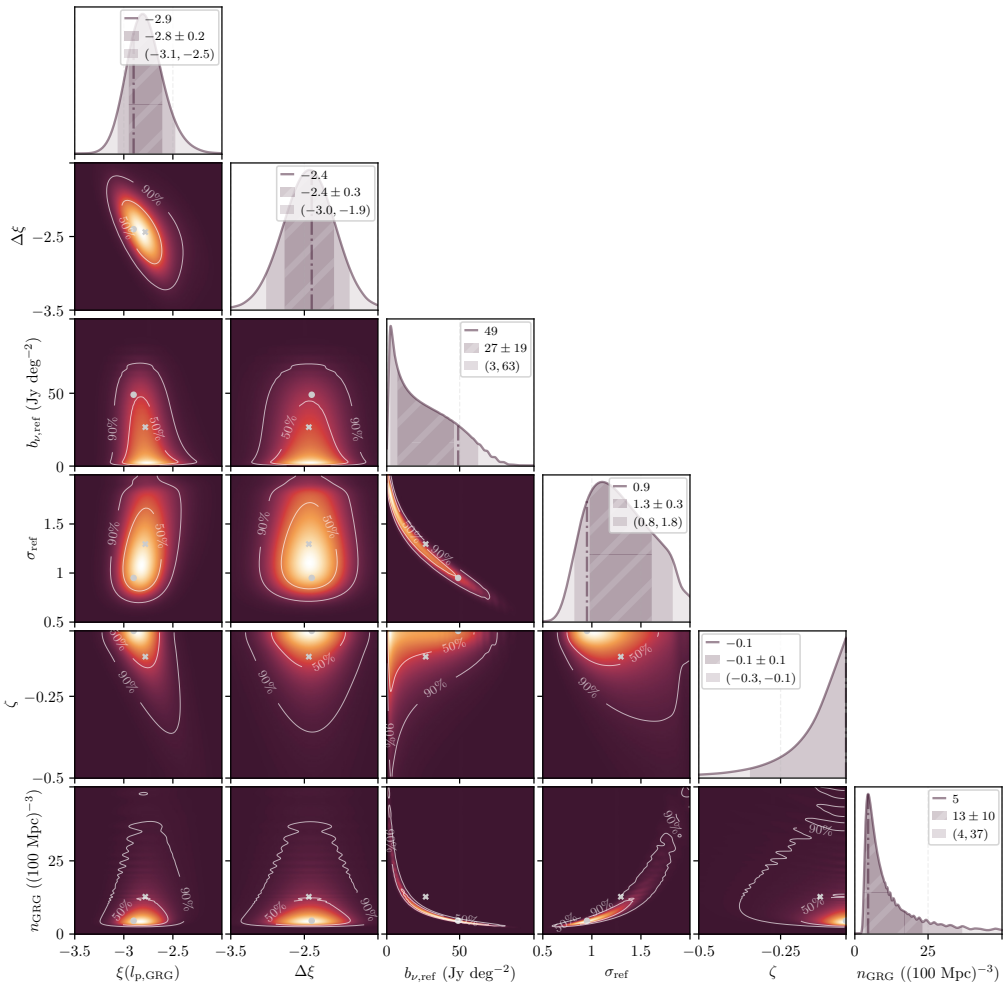

Figure 8.10: Likelihood function over $\theta = [\xi(l_{\mathrm{p,GRG}}), \Delta\xi, b_{\nu,\mathrm{ref}}, \sigma_{\mathrm{ref}}, \zeta, n_{\mathrm{GRG}}]$, based on 2685 projected lengths and redshifts of giants up to $z_{\mathrm{max}} = 0.5$. We show all two-parameter marginals of the likelihood function, with contours enclosing 50% and 90% of total probability. We mark the maximum likelihood estimate (MLE) values (grey dot) and the likelihood mean values (grey cross). The one-parameter marginals again show the MLE (dash-dotted line), a mean-centred interval of standard deviation–sized half-width (hashed region), and a median-centred 90% credible interval (shaded region).

that in Oei et al. (2023a): not only do we use more than a thousand additional giants, we also make more effective use of their redshift information. For further discussion, see Sect. 8.6.2.

It remains an open question whether giants can be understood as part of the ordinary radio galaxy population, or whether they evolve through qualitatively differ-

ent physical processes. As pointed out in Sect. 4.1.5 of Oei et al. (2023a), a curved power law PDF for $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ is consistent with a scenario in which giants share a broader length continuum with smaller radio galaxies. More specifically, if the broader radio galaxy length distribution is approximately lognormal, as appears justifiable on statistical grounds, then ξ should decrease throughout the distribution's right tail — that is, throughout the GRG range. Future research should determine whether such a unified non-giant RG–GRG scenario is also *quantitatively* consistent with the decrease in ξ we have inferred here. In addition, our inferences of $\xi(l_\mathrm{p,GRG})$ and $\Delta\xi$ are important in constraining Sect. 8.5.3's GRG lobe volume-filling fraction.

8.5.2 GRG number density

The extent to which giants have contributed to cosmic magnetogenesis depends on their intrinsic number density — which need not necessarily be a constant, but could have evolved over time. Observationally, giants are considered rare in comparison to smaller radio galaxies. However, because giants are presumably strongly affected by surface brightness selection, this current-day observed rarity might not translate to an intrinsic rarity. Excitingly, by forward modelling selection effects — and in particular surface brightness selection — we can constrain the intrinsic comoving GRG number density between $z = 0$ and $z = z_\mathrm{max}$, which we denote simply by n_GRG.

The bottom-right one-dimensional marginal of Fig. 8.10 shows a strongly skewed distribution for n_GRG, with a marginal mean $\mathbb{E}[n_\mathrm{GRG}] = 13 \pm 10\ (100\ \mathrm{Mpc})^{-3}$ and a 95% probability that $n_\mathrm{GRG} > 4\ (100\ \mathrm{Mpc})^{-3}$. These number densities are a factor of order unity higher than those of Oei et al. (2023a), who inferred a marginal mean $\mathbb{E}[n_\mathrm{GRG}] = 4.6 \pm 2.4\ (100\ \mathrm{Mpc})^{-3}$ and a 90% probability that $n_\mathrm{GRG} < 6.7\ (100\ \mathrm{Mpc})^{-3}$.

The joint marginal distribution of n_GRG and $b_{\nu,\mathrm{ref}}$ reveals a strong inverse relationship, whose origin is easy to grasp. Models in which giants are relatively rare (i.e. with low n_GRG) but with relatively mild surface brightness selection (i.e. with high $b_{\nu,\mathrm{ref}}$) are about as successful in reproducing the data-derived projected length–redshift histogram as models in which giants are relatively common (i.e. with high n_GRG) but with relatively severe surface brightness selection (i.e. with low $b_{\nu,\mathrm{ref}}$). The narrowness of the joint distribution also suggests that, if estimates of $b_{\nu,\mathrm{ref}}$ would reveal it to be $\gtrsim 10\ \mathrm{Jy\ deg^{-2}}$, it should be possible to break the degeneracy and accurately determine n_GRG.

Recent work (Oei et al., in prep.) suggests that the comoving number density of luminous, non-giant radio galaxies (LNGRGs), understood to have radio luminosities at 150 MHz of $l_\nu \geq 10^{24}\ \mathrm{W\ Hz^{-1}}$ and projected lengths $l_\mathrm{p} < l_\mathrm{p,GRG} := 0.7$ Mpc, is

$n_{\mathrm{LNGRG}} = 12 \pm 1\,(100\,\mathrm{Mpc})^{-3}$. Our work suggests that giants might be comparably common. If this is indeed the case, then the widespread belief that giants form a rare population of radio galaxies must be revised.

8.5.3 GRG lobe volume-filling fraction

The present-day volume-filling fraction of the lobes of giants in clusters and filaments of the Cosmic Web, $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$, is not a parameter of our model, but rather a derived quantity. As briefly discussed in Sect. 8.4.9, we compute its probability distribution using the parameter hexads that we have obtained by rejection sampling from the posterior.

For each sampled hexad, we compute $\xi(l_{\mathrm{p}})$ using $\xi(l_{\mathrm{p,GRG}})$, $\Delta\xi$, and Eq. 8.9, then $f_{L_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}}(l_{\mathrm{p}})$ using Eq. 8.10, and finally $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$ using n_{GRG} and Eq. 8.25. This last step also requires an estimate of $\mathbb{E}[\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}]$, the expectation value of the ratio between the combined lobe volumes and cubed projected lengths of giants. Problematically, only few accurate data currently exist to estimate this quantity. In Oei et al. (2022a), the authors estimated that Alcyoneus, which measures $l_{\mathrm{p}} = 4.99 \pm 0.04$ Mpc, boasts a combined lobe volume $V = 2.5 \pm 0.3$ Mpc^3. Similarly, Oei et al. (2023b) estimated that the giant generated by NGC 6185, which measures $l_{\mathrm{p}} = 2.45 \pm 0.01$ Mpc, has a combined lobe volume $V = 0.35 \pm 0.03$ Mpc^3. These cases give $\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}} = 2.0 \pm 0.2\%$ and $\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}} = 2.4 \pm 0.2\%$, respectively. We note that these cases concern giants generated by an elliptical and a spiral galaxy, respectively. The highly preliminary sample mean thus is $\langle \Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}} \rangle = 2.2 \pm 0.2\%$, which we treat as an approximation of $\mathbb{E}[\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}]$. As in Oei et al. (2023a), we assume that clusters and filaments comprise 5% of the Local Universe's volume (Forero-Romero et al., 2009): $\mathcal{V}_{\mathrm{CW}}(z = 0) = 5\%$. Propagating all uncertainties, we obtain the posterior distribution over $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$ shown in Fig. 8.11.

This probability distribution inherits its skewness from the skewed marginal of n_{GRG}. We find a posterior mean and standard deviation (SD) of $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0) = 1.1 \pm 0.9 \cdot 10^{-5}$. This result appears statistically consistent with that of Oei et al. (2023a), who found $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0) = 5^{+8}_{-2} \cdot 10^{-6}$. While this appears low at first sight, we speculate that these numbers are consistent with a scenario in which giants contribute significantly to cosmic magnetogenesis. To see why, we first note that the number of giants that have ever existed might exceed those that exist now by an order of magnitude. Second, a large fraction of the giants that existed throughout cosmic history might have lived at $z \gtrsim 1$, when the Universe's volume was an order of magnitude smaller. As a consequence, the instantaneous volume-filling fraction

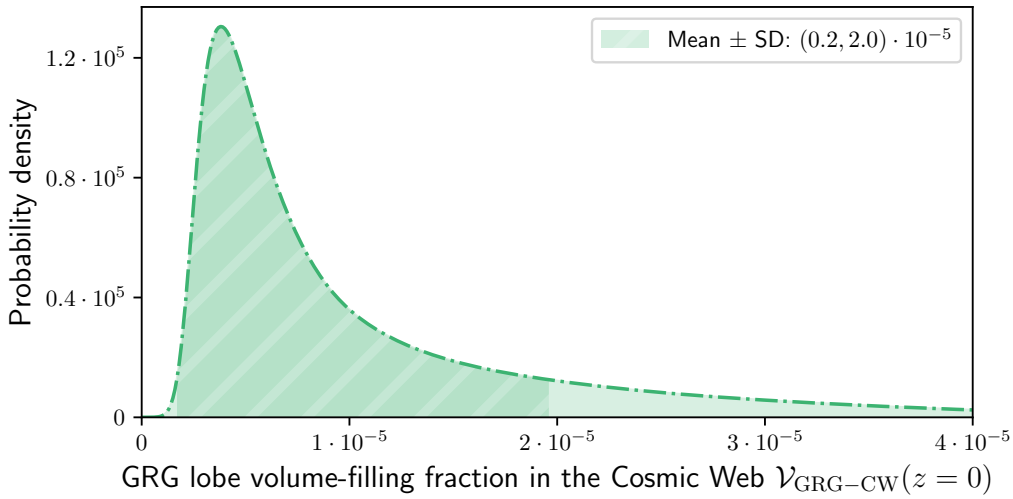

Figure 8.11: Posterior distribution for the instantaneous, current-day GRG lobe volume-filling fraction in clusters and filaments of the Cosmic Web, $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$.

$\mathcal{V}_{\mathrm{GRG-CW}}(z)$ would have been an order of magnitude larger — at least, if the instantaneous comoving number density and distributions over proper length and shape remain roughly constant over time. Third, buoyant lobes might deposit magnetic fields in their wake, while diffusion might have spread the magnetic fields of GRG lobes further through the IGM. Taken together, these three effects could render the current-day VFF of magnetic fields that were once contained in the lobes of giants higher than $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$ by several (i.e. three or more) orders of magnitude. This, in turn, suggests a significant astrophysical seeding potential. For instance, assuming four orders of magnitude, $\sim$10% of the volume of today's Cosmic Web could have been magnetised by giants.

We finally point out that giant-induced IGM magnetic fields could have strengths consistent with observational constraints. At the moment, the lowest magnetic field strengths measured in giant radio galaxy lobes, as inferred from images of Alcyoneus and the giant generated by NGC 6185 assuming the equipartition or minimum energy condition, are 400–500 nG (Oei et al., 2022a, 2023b). If such field strengths would be typical, and buoyancy and diffusion lowers the density of field lines by an order of magnitude, then the typical giant-induced IGM field strength would be $\sim$10 nG. This is in agreement with recent radio estimates and limits (e.g. Table 1 of Vazza et al., 2021a). We note that this argument ignores possibly significant amplification and decay mechanisms, such as turbulent amplification and decay.

8.6 Discussion

Below, we discuss how our ML pipeline and GRG population inference compare to earlier work.

8.6.1 Comparison with previous ML GRG search techniques

Proctor (2016) applied an ML approach to search for GRG candidates by looking for likely pairs of (unresolved) radio lobes with the required angular length in the NRAO VLA Sky Survey (NVSS; Condon et al., 1998). For this radio source component association problem, Proctor (2016) trained an oblique classifier (a type of decision tree ensemble; Murthy et al., 1993), using six source finder–derived features on 51,195 pairs of radio components, 48 of which were verified giants. This method proved to be useful under the assumption that giants generally appear as an isolated pair of unresolved radio blobs, which is the case for NVSS with its 45″ resolution and 450 μJy beam^{-1} sensitivity. Dabhade et al. (2020a) visually inspected the 1,600 GRG candidates presented by Proctor (2016) and confirmed 151 giants, which implies a 9% precision for the GRG candidate predictions. However, Proctor (2016) expect that giants with resolved lobes — which rule-based source finders often incorrectly break down into multiple separate sources — require a different approach, and we would like the reader to note that virtually all GRG lobes in LoTSS are resolved.[16] It works in our favour that the convolution neural network in our ML pipeline (Sect. 8.4.4) was specifically designed to use the morphology of the resolved, extended emission as a cue for the radio source component association. Furthermore, as the source suggestions from our ML pipeline include optical host identifications, the candidates that we inspected not only have the required angular length but also have a host galaxy and corresponding redshift estimate assigned. This allows us to visually inspect only those radio sources that fulfil the projected proper length $l_{\mathrm{p,GRG}} := 0.7\,\mathrm{Mpc}$ requirement. Overall, our ML pipeline has a precision of 47% for the GRG candidates that it suggests.

8.6.2 Comparison with previous inference strategies

Compared to the approach of Oei et al. (2023a), our approach makes better use of the redshift information available for each giant. More specifically, we use the redshifts to make a 'redshift-resolved' observed projected length histogram, while Oei et al. (2023a) only compared a 'redshift-collapsed' distribution of observed projected

[16] As all giants have angular lengths $\varphi \geq 1.3'$, they cover at least 13 LoTSS 6″ beams. This suggests that a single GRG lobe will cover multiple beams, too.

lengths to forward model predictions of $L_{\mathrm{p,obs}} \mid L_{\mathrm{p,obs}} \geq l_{\mathrm{p,GRG}}$. Effectively, Oei et al. (2023a) thus used for each giant only *Boolean* redshift information, $\mathbb{I}(z_i < z_{\mathrm{max}})$: that is, a truth value indicating whether or not the giant with index i resides at a redshift below the maximum considered value.

In addition, our work changed the comoving number density of giants, n_{GRG}, from a derived quantity to a model parameter. This approach acknowledges the fact that the observed number of giants, either for a specific projected length–redshift bin or for the parameter space in its entirety, scales linearly with n_{GRG} (if the selection effects remain the same). Thus, there is intrinsic population information contained in the observed *number* of giants. However, by comparing predicted and observed *probability distributions* only, Oei et al. (2023a) did not exploit this fact.

8.6.3 Future work

With the advent of large-scale, sensitive, low-frequency sky surveys such as the LoTSS, the Evolutionary Map of the Universe survey (EMU; Norris et al., 2011), and the arrival of next-generation instruments such as the SKA (Dewdney et al., 2009) and the DSA-2000 (e.g. Hallinan et al., 2019; Connor et al., 2022) later this decade, opportunities shall arise to detect many more giants than have been found hitherto. It is therefore likely that automated approaches to giant finding and host association will become only more relevant in the future.

Regarding our own machine learning–based pipeline, there is significant room to improve both the radio component association and the host association. Visual inspection indicated a precision of 47% and the empirically determined $p_{\mathrm{obs,ID}}$ in Fig. 8.9 showed that even in combination with the LGZ sample, the ML pipeline recall does not surpass 70%. Sensible paths to improve the radio component association within the ML pipeline architecture include switching from rectangular bounding box–based object detection (the Fast R-CNN used in this article) to pixel-based instance segmentation and using a larger convolutional backbone (e.g. Liu et al., 2022; Wright et al., 2010) or a transformer-based backbone (e.g. Liu et al., 2021; Zhang et al., 2022; Li et al., 2022). Mostert et al. (2022) conclude that a larger convolutional neural network is not effective unless one also significantly increases the quantity of high-quality training data, and in general, transformers require even more training data than convolutional neural networks (e.g. Wang et al., 2022). To that extent, adding a filtered version[17] of the available LoTSS DR2 LGZ annotations (Hardcastle et al., 2023) to the training data can be considered. Furthermore, assembling a joined

[17] For example, by identifying a handful of very active and expert volunteers and increasing the weight of their votes.

dataset encompassing the (labelled) survey data of other low frequency radio telescopes can be considered. Pre-training on this dataset can benefit radio source component association, host identification and morphological classification tasks across the board.

Finally, there appear to be clear opportunities to make the population-based forward model presented in Sect. 8.2 more accurate. For example, at present, we have neglected photometric redshift uncertainties; however, the consequences of these uncertainties appear perfectly possible to forward model. One such currently ignored consequence is Eddington bias: as RGs with projected lengths $l_\mathrm{p} = 0.6$ Mpc are intrinsically more common than those with projected lengths $l_\mathrm{p} = 0.8$ Mpc, redshift error–induced projected length errors have the net effect of falsely raising the number of supposed giants with projected lengths near $l_\mathrm{p,GRG} \coloneqq 0.7$ Mpc. This effect could contaminate the inference of $\xi(l_\mathrm{p,GRG})$. Somewhat more challenging, but plausibly of greater value, would be a further exploration of how surface brightness selection is effectively modelled. A major focus of such an exploration would be to analyse the surface brightness characteristics of hitherto discovered giants. As the masked cutouts of Fig. 8.3 suggest, the machine learning–based pipeline described in this work offers the exciting potential to amass — fully automatically — surface brightness properties for thousands of giants. The availability of such properties for a large fraction of observed giants also allows one to compare the forward model's predictions with an observed projected length–redshift–surface brightness histogram, rather than with an observed projected length–redshift histogram only. It is highly likely that adding another dimension to the data yields tighter parameter constraints. To make the identification probability functions of Fig. 8.9 more accurate, it appears promising to have an expert visually (and exhaustively, i.e. without imposing angular length thresholds) comb through a small representative region of LoTSS DR2 in search of giants. The resulting dataset would provide a better basis for determining the identification probability functions than the ML–LGZ or Oei et al. (2023a) datasets used in this work. We note that the Boötes LOFAR Deep Field search of Simonte et al. (2022) does not appear suited for this purpose, as the increased depth of this field renders it unrepresentative of LoTSS DR2 as a whole. Finally, the model could be expanded in an attempt to measure cosmological evolution of, for example, n_GRG. However, we note that adding additional parameters to the model necessitates adopting more efficient inference techniques, such as Markov chain Monte Carlo or nested sampling. The associated numerical gain would, in part, be negated by losing the speed-up associated to the likelihood trick of Appendix 8.A2.

Currently, a major uncertain factor in the determination of $\mathcal{V}_\mathrm{GRG-CW}(z = 0)$ is the value of $\mathbb{E}[\Upsilon_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}]$. To improve this situation, we recommend ex-

panding the capabilities and automating the parametric Bayesian lobe volume estimation method introduced by Oei et al. (2022a, 2023b). This method could then be applied to thousands of our ML pipeline's masked cutouts, such as the one in Fig. 8.3. This effort would increase the number of giants on which our estimate of $\mathbb{E}[\Upsilon_{\mathrm{p}} \mid L_{\mathrm{p}} \geq l_{\mathrm{p,GRG}}]$ is based by several (i.e. two or three) orders of magnitude.

8.7 Conclusions

In this work, we concatenated an existing crowd-sourced radio–optical catalogue, a new ML pipeline to automate radio–optical catalogue creation, and a Bayesian forward model to build a next-generation giant radio galaxy discovery and characterisation machine. Applying this setup to the LOFAR Two-metre Sky Survey, we uncovered thousands of previously unknown giants, confirmed thousands of GRG candidates, and constrained the properties of the underlying population.

1. The LoTSS is an on-going sensitive, high-resolution, low-frequency radio survey whose second data release (DR2) covers 27% of the Northern Sky. As the number of detected sources already ranges in the millions, it has become unfeasible (at least for small scientific teams) to conduct manual, visual searches for giants — in particular for those with angular lengths close to the lower limit of 1.3′.

2. To address this challenge, we scanned all 841 LoTSS DR2 pointings — which together cover more than five thousand square degrees of Northern Sky — with an ML pipeline that crucially includes the convolutional neural network of Mostert et al. (2022), designed for the association of radio components for highly resolved radio galaxies, and an adapted version of the automated optical host galaxy identification heuristic developed by Barkus et al. (2022). Used as a GRG detection system, our ML pipeline has a precision of 47%, a significant improvement over the 9% precision obtained using the previous state-of-the-art ML GRG detection model (Proctor, 2016; Dabhade et al., 2020a). We merged the resulting giant candidate sample with that of the LGZ citizen science campaign (Hardcastle et al., 2023), homogenised the angular lengths, and subjected the candidates to a visual, expert quality check. The result is a sample of more than eight thousand newly confirmed giants, of which a large fraction is considered genuine beyond reasonable doubt. More than 10^4 unique giants are now known to the literature.

3. We expand the population-based statistical forward model of Oei et al. (2023a) aimed at constraining the geometric properties of giants. In particular, by

modelling the PDF of the radio galaxy projected length RV L_p as a curved power law, we automatically also model the PDF of the *giant* radio galaxy projected length RV $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ as a curved power law. We assume that these projected length distributions do not undergo intrinsic evolution between cosmological redshift $z = z_\mathrm{max}$ and $z = 0$, and likewise assume an intrinsically constant comoving GRG number density throughout this redshift range. We model surface brightness selection by assuming a lognormal lobe surface brightness distribution at the survey's central frequency ν_obs, valid for radio galaxies of intrinsic proper length l_ref at redshift $z = 0$. We relate lobe surface brightness distributions for radio galaxies of other lengths and at other redshifts to this reference distribution. In addition, we model selection caused by the imperfect ability of search methods to identify all in principle identifiable giants. For this purpose, we use logistic functions of projected length l_p and redshift z.

4. We then sought to identify all model parameter hexads that can reproduce the projected length–redshift histogram of the joint ML–LGZ–Oei et al. (2023a) LoTSS DR2 GRG sample. Through a simple Poissonian likelihood and a uniform prior distribution, we constructed a posterior distribution over the model parameters. By confronting the model with an observed projected length–redshift histogram, rather than with an observed projected length distribution only (as has been done in Oei et al. (2023a)), we obtain tighter parameter constraints.

5. We find evidence for the claim that the projected lengths of giant radio galaxies follow a curved power law PDF, whose tail index equals $\xi(l_\mathrm{p,GRG}) = -2.8 \pm 0.2$ at $l_\mathrm{p,1} = l_\mathrm{p,GRG} := 0.7$ Mpc and increases by $\Delta\xi = -2.4 \pm 0.3$ (i.e. decreases by 2.4 ∓ 0.3) in the projected length interval leading up to $l_\mathrm{p,2} = 5$ Mpc. The predicted median lobe surface brightness at $\nu_\mathrm{obs} = 150$ MHz, $l_\mathrm{ref} = 0.7$ Mpc, and $z = 0$ equals $b_{\nu,\mathrm{ref}} = 30 \pm 20$ Jy deg^{-2}. This surface brightness level is lower than previously thought. Tight degeneracies resembling inverse relations exist between $b_{\nu,\mathrm{ref}}$ and the reference surface brightness dispersion measure σ_ref, and between $b_{\nu,\mathrm{ref}}$ and the GRG number density n_GRG. The latter relation suggests that giant radio galaxies might be more common than previously thought. At $n_\mathrm{GRG} = 13 \pm 10\,(100\,\mathrm{Mpc})^{-3}$, giant radio galaxies appear to be of an abundance comparable to that of luminous *non*-giant radio galaxies. Strikingly, we conclude that, at any moment in time, a significant fraction of the radio galaxy population is in a GRG phase. As an immediate consequence, the fraction of radio galaxies that end their lives as giants must

be even higher.

6. Finally, we generate a posterior distribution for the instantaneous volume-filling fraction of GRG lobes in clusters and filaments of the Cosmic Web, $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0)$ — a key statistic required for determining the cosmic magnetogenesis potential of giants. We find $\mathcal{V}_{\mathrm{GRG-CW}}(z = 0) = 1.1 \pm 0.9 \cdot 10^{-5}$. If a giant population similar to that in the Local Universe has existed for most of the Universe's lifetime, and IGM mixing processes are significant, then it appears possible that magnetic fields originating from giants permeate significant ($\sim$10%) fractions of today's Cosmic Web.

Using modern automation and inference techniques — that still leave much room for future improvements — we have conducted the most detailed study yet of the abundance and geometry of giant radio galaxies. These cosmic colossi may provide a previously underappreciated contribution to astrophysical magnetogenesis.

The full GRG catalogue with host identifications and the Stokes-I cutouts containing the segmented giants will soon be available on Zenodo. M.S.S.L. Oei and R.J. van Weeren acknowledge support from the VIDI research programme with project number 639.042.729, which is financed by the Dutch Research Council (NWO). B. Barkus is grateful for support from the UK STFC. L. Alegre is grateful for support from the UK STFC via CDT studentship grant ST/P006809/1. M.J. Hardcastle acknowledges support from the UK STFC [ST/V000624/1]. We like to thank Huib Intema for enabling the cross-institute collaboration on the Leiden Observatory computer infrastructure. We like to thank Frits Sweijen for coding the very useful `https://github.com/tikk3r/legacystamps`. This research has made use of the Python `astropy` package (The Astropy Collaboration et al., 2018); the VizieR catalogue access tool (Ochsenbein et al., 2000), CDS, Strasbourg, France (DOI: 10.26093/cds/vizier); and the 'Aladin Sky Atlas' developed at CDS, Strasbourg Observatory, France (Bonnarel et al., 2000; Boch & Fernique, 2014). LOFAR data products were provided by the LOFAR Surveys Key Science project (LSKSP; `https://lofar-surveys.org/`) and were derived from observations with the International LOFAR Telescope (ILT). LOFAR (van Haarlem et al., 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and which are collectively operated by the ILT foundation under a joint scientific policy. The efforts of the LSKSP have benefited from funding from the European Research Council, NOVA, NWO, CNRS-INSU, the SURF Co-operative, the UK Science and Technology Funding Council and the Jülich Supercomputing Centre. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

Author contributions

Rafaël and Martijn together came up with the idea of the study. Rafaël built and ran the machine learning–accelerated pipeline, merged the results with those of the LGZ pipeline, and visually judged all GRG candidates. Rafaël also generated the final GRG catalogue. Martijn developed the forward model, ran it on the data, and inferred the GRG lobe VFF. Rafaël and Martijn jointly wrote the article.

8.A1 Curved power law PDF for L

In Sect. 8.2.3, we have started modelling the geometry of radio galaxies at the level of the projected proper length RV L_p. While algebraically easier — when curved power laws are considered, at least — this approach is more limited than starting the forward model at the level of the intrinsic proper length RV L. In this appendix, we calculate the distribution of L_p upon modelling L with a curved power law.

Let us assume that, for $l \geq l_\mathrm{min}$,

$$f_L(l) \propto \left(\frac{l}{l_\mathrm{min}}\right)^{\xi(l)}, \tag{8.35}$$

where $\xi(l) = al + b$. We now use the identity that for $f(x) = \left(\frac{x}{c}\right)^{ax+b}$, one finds

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \left(\frac{x}{c}\right)^{ax+b}\left(a\ln\frac{x}{c} + a + \frac{b}{x}\right) = f(x)\left(a\ln\frac{x}{c} + a + \frac{b}{x}\right). \tag{8.36}$$

Therefore,

$$\frac{\mathrm{d}f_L(l)}{\mathrm{d}l} = f_L(l)\left(a\ln\frac{l}{l_\mathrm{min}} + a + \frac{b}{l}\right), \tag{8.37}$$

and

$$\frac{\mathrm{d}f_L(l_\mathrm{p}\eta)}{\mathrm{d}l_\mathrm{p}} = f_L(l_\mathrm{p}\eta)\left(a\ln\frac{l_\mathrm{p}\eta}{l_\mathrm{min}} + a + \frac{b}{l_\mathrm{p}\eta}\right)\eta. \tag{8.38}$$

Thus, finding the PDF of L_p requires calculating three different integrals over η:

$$\begin{aligned} f_{L_\mathrm{p}}(l_\mathrm{p}) = &-(1+b)\int_1^\infty \sqrt{1-\frac{1}{\eta^2}}\, f_L(l_\mathrm{p}\eta)\,\mathrm{d}\eta \\ &- l_\mathrm{p}a\left(1+\ln\frac{l_\mathrm{p}}{l_\mathrm{min}}\right)\int_1^\infty \sqrt{\eta^2-1} f_L(l_\mathrm{p}\eta)\,\mathrm{d}\eta \\ &- l_\mathrm{p}a\int_1^\infty \sqrt{\eta^2-1} f_L(l_\mathrm{p}\eta)\ln\eta\,\mathrm{d}\eta \quad \text{for } l_\mathrm{p} > l_\mathrm{min}. \end{aligned} \tag{8.39}$$

The PDF of $L_\mathrm{p} \mid L_\mathrm{p} \geq l_\mathrm{p,GRG}$ follows through Eq. 8.3.

8.A2 Likelihood trick

Thanks to its Poissonian form, there exists a particularly numerically efficient way of computing the likelihood presented in Sect. 8.2.8 as a function of n_{GRG}, for fixed values of the other parameters. Defining

$$A(\theta) \coloneqq \sum_{i=1}^{N_{\mathrm{b}}} N_i \ln \lambda_i(\theta) \quad \text{and} \quad B(\theta) \coloneqq \sum_{i=1}^{N_{\mathrm{b}}} \lambda_i(\theta), \tag{8.40}$$

one interested in the log-likelihood up to a constant only needs to compute

$$\ell(\theta) \coloneqq \ln \mathcal{L}(\{N_i\} \mid \theta) + \sum_{i=1}^{N_{\mathrm{b}}} \ln (N_i!) = A(\theta) - B(\theta). \tag{8.41}$$

The quantity $B(\theta)$ has a simple interpretation: it is the total number of giants expected to be observed under θ within the entire projected length–redshift parameter space considered.

How does ℓ change upon changing n_{GRG}? When $n_{\mathrm{GRG}} \mapsto a \cdot n_{\mathrm{GRG}}$, $\lambda_i \mapsto a \cdot \lambda_i$, so that

$$\begin{aligned} \ell(n_{\mathrm{GRG}}) &\mapsto \sum_{i=1}^{N_{\mathrm{b}}} N_i \ln (a \cdot \lambda_i) - a \cdot \lambda_i \\ &= A(n_{\mathrm{GRG}}) - a \cdot B(n_{\mathrm{GRG}}) + \ln a \cdot \sum_{i=1}^{N_{\mathrm{bins}}} N_i. \end{aligned} \tag{8.42}$$

(In the notation $\ell(n_{\mathrm{GRG}})$, $A(n_{\mathrm{GRG}})$, and $B(n_{\mathrm{GRG}})$, we suppress the dependence on the other five parameters.) We conclude that, when n_{GRG} increases by a factor a, the A-term in ℓ remains the same, the B-term in ℓ becomes a factor a bigger, and an extra factor emerges: namely, the product of $\ln a$ and the total number of giants in the dataset.

The significance of this result is that, once A and B are known at some reference number density $n_{\mathrm{GRG,ref}}$, we can rapidly evaluate ℓ for any other number density. In this work, we implement this 'likelihood trick' by evaluating ℓ for two different values of n_{GRG} (and for many different values of the other parameters). We then solve for

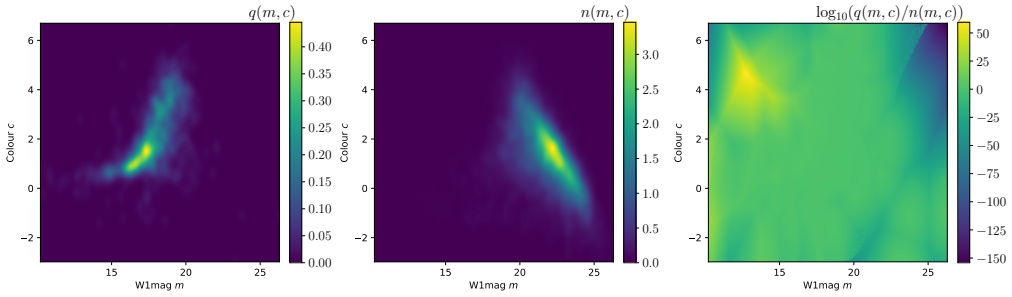

Figure 8.12: Unregularised KDE estimates for q in the left panel, n in the second panel, and q/n with logarithmic colour bar in the third panel. The KDE bandwidth of 0.2 stems from Barkus et al. (2022).

$A(n_{\mathrm{GRG,ref}})$ and $B(n_{\mathrm{GRG,ref}})$, and use

$$\ell(n_{\mathrm{GRG}}) = A(n_{\mathrm{GRG,ref}}) - \frac{n_{\mathrm{GRG}}}{n_{\mathrm{GRG,ref}}} \cdot B(n_{\mathrm{GRG,ref}}) + \ln \frac{n_{\mathrm{GRG}}}{n_{\mathrm{GRG,ref}}} \cdot \sum_{i=1}^{N_{\mathrm{b}}} N_i. \tag{8.43}$$

8.A3 PyBDSF parameters

As described in Sect. 8.4.1, the GRG detection pipeline uses PyBDSF for the initial radio blob detection. For reproducibility, we provide the specific parameters used, which we adopted from Shimwell et al. (2022):

```
bdsf.process_image(<filename>,thresh_isl=4.0,
thresh_pix=5.0, rms_box=(150,15), rms_map=True,
mean_map='zero', ini_method='intensity',
adaptive_rms_box=True, adaptive_thresh=150,
rms_box_bright=(60,15), group_by_isl=False,
group_tol=10.0, output_opts=True, atrous_do=True,
atrous_jmax=4, flagging_opts=True,
flag_maxsize_fwhm=0.5, advanced_opts=True,
blank_limit=None, frequency=143.65e6)
```

8.A4 Adaptations of the radio ridgeline based host galaxy identification

Here we elaborate on two small adaptations of the radio-optical crossmatch method introduced by Barkus et al. (2022). First, we explicitly regularised $q(m,c)$ and $n(m,c)$.

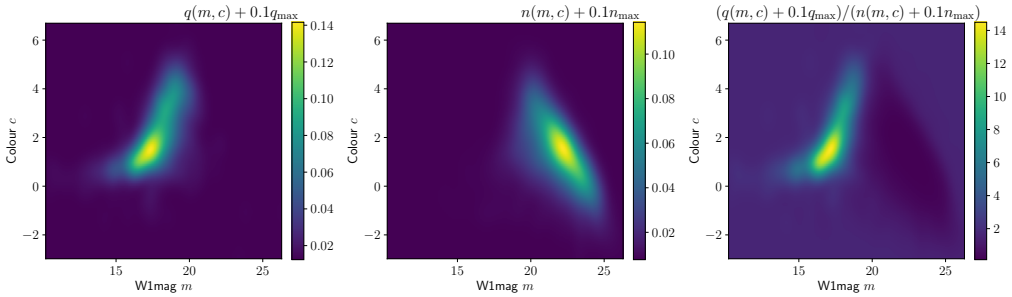

Figure 8.13: Regularised KDE estimates for q in the left panel, n in the second panel, and q/n with logarithmic colour bar in the third panel. The KDE bandwidth of 0.4 stems from 10-fold cross-validation.

Figure 8.12 shows that the unregularised forms of q and n can take on extreme values in the LR (eq. 8.31) in sparsely sampled regions of the (m, c)–parameter space. The 2D KDE that models $q(m, c)$ was fitted on the m and c values of all 905 sources with an angular length $\varphi > 1'$ from 40 randomly picked LoTSS DR2 pointings. The 2D KDE that models $n(m, c)$ was fitted on the m and c values of 10, 000 sources that were randomly sampled from the entire combined infrared–optical catalogue. By simply adding a small constant factor to $q(m, c)$ and $n(m, c)$ we get more robust LR values, see Fig. 8.13. We added a constant factor $0.1 \cdot q_{\max}$ and $0.1 \cdot n_{\max}$ for q and n respectively, where $q_{\max}$ is the maximum of the KDE for q and $n_{\max}$ is the maximum of the KDE for n. We set the KDE bandwidths to 0.4 following a 10-fold cross-validation.

Second, we changed the form of $f(r)$. Theoretically, we might expect both the distance between the 'true' optical counterpart and the radio ridgeline $r_{\text{opt,ridge}}$ and the distance between the 'true' optical counterpart and the radio centroid $r_{\text{opt,centroid}}$ to be Rayleigh distributed.[18] However, as Fig. 8.14 demonstrates, the lognormal distribution clearly provides the best empirical fit to the distances. The figure shows a histogram of the distance measures for radio sources to their optical counterpart as manually identified through LGZ. Specifically, we plot the distances for the same 905 radio sources, with an angular length $\varphi > 1'$, from 40 randomly selected paintings as above. Thus we update $f(r)$ to be:

$$f(r_{\text{mean}}) = \frac{1}{r_{\text{mean}}\sigma\sqrt{2\pi}} e^{\left(-\frac{(\ln r_{\text{mean}} - \mu)^2}{2\sigma^2}\right)}, \tag{8.44}$$

[18] In two dimensions, the Euclidean distance between the origin and a point whose Cartesian coordinates are independent, zero-mean, and equal-variance normal random variables, is Rayleigh distributed. This motivates modelling the angular distance between the optical counterpart and the radio centroid with a Rayleigh distribution. The appropriate value of the distribution's parameter likely depends (positively) on the angular length of the radio source considered; as such, one would not expect a single Rayleigh distribution to work for all radio sources.

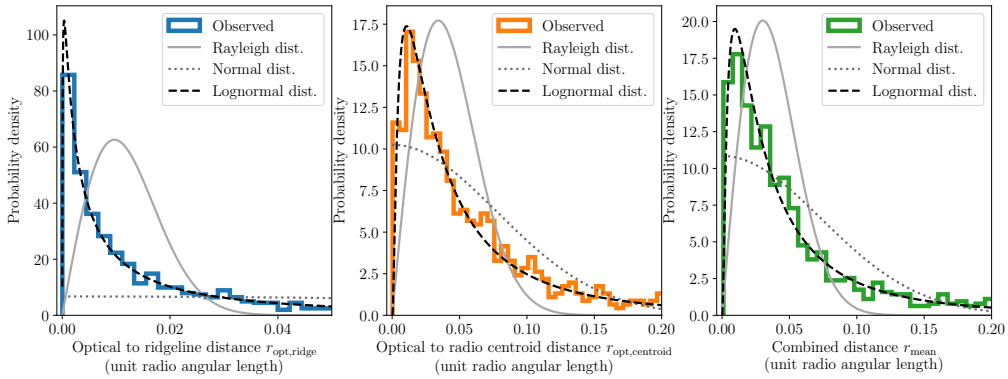

Figure 8.14: Each panel shows the histogram of a different distance measure between 905 radio sources with $\varphi > 1'$ and their optical host. The grey, dark grey, and black lines show empirical fits to these histograms for Rayleigh, normal, and lognormal distributions respectively. The tails of the histograms are long, for visualisation purposes we only plot the x-axis up to 0.05 and 0.20.

where we empirically determine σ and μ using our sample of 905 radio-sources,

$$\mu = \frac{\sum_i \ln r_{\mathrm{mean},i}}{n} = -3.37 \tag{8.45}$$

and

$$\sigma^2 = \frac{\sum_i (\ln r_{\mathrm{mean},i} - \mu)^2}{n} = 1.28, \tag{8.46}$$

with $n = 905$ the size of our sample.

8.A5 Sky coverages

As an extension of Sect. 8.4.9, this appendix details the sky coverages of our analyses. In particular, Table 8.3 provides a decomposition — in terms of disjoint spherical quadrangles — of the sky coverage common between the ML pipeline, LGZ, and the combined manual search of Dabhade et al. (2020b) and Oei et al. (2023a). For simplicity, and as an acknowledgement of the wiggle room inherent to defining this joint sky coverage, we chose integer coordinates. Together, these spherical quadrangles cover $\Omega = 5327.9\,\mathrm{deg}^2 = 1.62\,\mathrm{sr}$ (25.8%) of the Northern Sky. We shall refer to this coverage simply as the 'LoTSS DR2 coverage'.

The ML–LGZ–Oei et al. (2023a) overlap region amounts to the LoTSS DR2 coverage with the LoTSS DR1 spherical quadrangle removed. The minimum and maximum right ascensions of this quadrangle are $\alpha_{\mathrm{min}} = 160°$ and $\alpha_{\mathrm{max}} = 230°$, while its

minimum and maximum declinations are $\delta_{\rm min} = 45^\circ$ and $\delta_{\rm max} = 56^\circ$. This smaller overlap region covers 4838.9 deg^2 (23.5%) of the Northern Sky. It is the sky coverage relevant to estimating the identification probability functions of Sect. 8.4.9 and Fig. 8.9: $p_{\rm obs,ID,1}(l_{\rm p}, z)$, $p_{\rm obs,ID,2}(l_{\rm p}, z)$, and $p_{\rm obs,ID}(l_{\rm p}, z)$.

Table 8.3: Sky coordinates and solid angles of disjoint spherical quadrangles whose union forms the LoTSS DR2 sky coverage — over which we have performed our inference. For each spherical quadrangle, we provide the minimum and maximum right ascension, $\alpha_{\rm min}$ and $\alpha_{\rm max}$, the minimum and maximum declination, $\delta_{\rm min}$ and $\delta_{\rm max}$, and its solid angle, Ω. We list the largest quadrangles first. The second and third object touch along the 360°–0° right ascension coordinate discontinuity, and could be viewed as a single quadrangle.

$\alpha_{\rm min}$ (°)	$\alpha_{\rm max}$ (°)	$\delta_{\rm min}$ (°)	$\delta_{\rm max}$ (°)	Ω (deg^2)
120	253	28	69	3536.7
0	35	16	35	597.5
338	360	16	35	375.6
253	269	28	47	240.1
109	120	25	41	147.1
269	277	31	47	99.2
330	338	17	30	95.2
191	210	23	28	85.7
35	41	24	32	42.3
253	260	58	69	34.3
120	131	25	28	29.5
277	281	41	47	17.3
327	330	17	20	8.5
277	280	32	35	7.5
117	120	53	57	6.9
260	264	66	69	4.6

A

Gaussian random field ionosphere model extension: the curved Earth

In Chapter 2, we considered, at one instant of time, a thick, single-layered ionosphere whose free electron density (FED) n_e is a Gaussian random field. For algebraic and numerical simplicity, we took the layer to be parallel to a *flat* Earth. For this configuration, we calculated the differential total electron content (DTEC) $\Delta\tau$ covariance function, and showed its superiority over ad-hoc covariance functions in DTEC Gaussian process regression (GPR).

For radio interferometers with long baselines, this 'flat Earth' approximation becomes coarse. To avoid associated errors in DTEC GPR, this appendix generalises Chapter 2's model to take proper account of the (approximate) sphericity of the Earth.[1]

We now assume the ionosphere to be a *spherical shell* with thickness b, centered around some point $\mathbf{x}_c \in \mathbb{R}^3$. We require the planet's rotational axis to also pass through this point. Although we specify the ionosphere's geometry, in general we need not impose a requirement on Terra's shape itself — it might be perfectly spherical, slightly ellipsoidal, or spherical with plate tectonics–induced surface height variations, for ex-

[1] The model thus becomes incompatible with flat Earth beliefs — whose modern adherents constitute a *global* movement depicting, in my opinion, the human intellect in a rather *unflattering* light.

ample. More relevant to the model are the locations of the antennae — or, in the LOFAR context, *stations* — $\mathbf{x}_i \in \mathbb{R}^3$.[2] The height of the ionosphere is defined such that the middle of the layer lies a distance a above the reference station at $\mathbf{x}_0 \in \mathbb{R}^3$. Let A denote the height of the middle of the ionospheric shell above $\mathbf{x}_c$; thus

$$A := ||\mathbf{x}_0 - \mathbf{x}_c||_2 + a. \tag{A.1}$$

A point $\mathbf{y} \in \mathbb{R}^3$ lies in the shell as long as

$$A - \frac{b}{2} < ||\mathbf{y} - \mathbf{x}_c||_2 < A + \frac{b}{2}. \tag{A.2}$$

Let us suppose that $\mathbf{y}$ lies on a line through $\mathbf{x}_i$ heading in skybound direction $\hat{\mathbf{k}} \in \mathbb{S}^2$. A parametrisation of this line is $\mathbf{y}(s) = \mathbf{x}_i + \hat{\mathbf{k}}s$, with $s \in \mathbb{R}$. The line segment that lies fully within the shell is demarcated by the two values $s_i^{\pm} \in \mathbb{R}_{>0}$ such that

$$\begin{aligned} ||\mathbf{y}(s_i^{\pm}) - \mathbf{x}_c||_2 &= A \pm \frac{b}{2}, \text{ or} \\ ||\mathbf{x}_i + \hat{\mathbf{k}}s_i^{\pm} - \mathbf{x}_c||_2 &= A \pm \frac{b}{2}. \end{aligned} \tag{A.3}$$

Squaring both sides, and writing the result as a second-degree polynomial equation in $s_i^{\pm}$, we find

$$\left(s_i^{\pm}\right)^2 + 2\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\, s_i^{\pm} + ||\mathbf{x}_i - \mathbf{x}_c||_2^2 - \left(A \pm \frac{b}{2}\right)^2 = 0. \tag{A.4}$$

By the quadratic formula, we find

$$\begin{aligned} s_i^{\pm} &= -\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c) \boxed{\pm} \sqrt{\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right)^2 - ||\mathbf{x}_i - \mathbf{x}_c||_2^2 + \left(A \pm \frac{b}{2}\right)^2} \\ &= \left|\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right| \left(-\operatorname{sgn}\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right) \boxed{\pm} \sqrt{1 + \frac{\left(A \pm \frac{b}{2}\right)^2 - ||\mathbf{x}_i - \mathbf{x}_c||_2^2}{\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right)^2}} \right). \end{aligned} \tag{A.5}$$

A prima facie, the second line seems more complicated than the first. However, the expression is insightful.

[2] Naturally, we assume the stations to lie underneath the ionosphere — in other words, *enclosed* by the ionospheric shell.

First, we remark that $\mathcal{A} \pm \frac{b}{2} > ||\mathbf{x}_i - \mathbf{x}_c||_2$ for all $\mathbf{x}_i$, as we have assumed the stations to lie underneath the ionosphere.

Furthermore, we remark that for all practical situations, $\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c) > 0$: the vector pointing from the centre of the shell to the station with index i lies in the same *hemisphere* as the observing direction. The reason is that Terra is approximately spherical, and that observers always make sure that sources stay above the local horizon with sufficient elevation during observing runs. In case of a perfectly spherical Earth with $\mathbf{x}_c$ as its centre, and stations put on top of its smooth surface, this has a neat geometrical interpretation. Under this idealisation, $\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c) > 0$ whenever the source is above the local horizon — and thus visible! As a consequence, $-\mathrm{sgn}(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)) = -1$ in practice.

Third, because $\hat{\mathbf{k}}$ points from Terra towards the *Great Unknown*, both $s_i^{\pm}$ are positive.

In order for the second line of Eq. A.5 to yield positive values, we need the $\boxed{+}$ of $\boxed{\pm}$. The equation reduces to

$$s_i^{\pm} = \left|\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right| \left(\sqrt{1 + \frac{\left(\mathcal{A} \pm \frac{b}{2}\right)^2 - ||\mathbf{x}_i - \mathbf{x}_c||_2^2}{\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right)^2}} - 1 \right). \tag{A.6}$$

Using the *first* line of Eq. A.5, we find the path length through the shell — for a station at $\mathbf{x}_i$ observing in direction $\hat{\mathbf{k}}$ — to be

$$\begin{aligned} \Delta s_i &:= s_i^+ - s_i^- && \text{(A.7)} \\ &= \sqrt{\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right)^2 - ||\mathbf{x}_i - \mathbf{x}_c||_2^2 + \left(\mathcal{A} + \frac{b}{2}\right)^2} \\ &\quad - \sqrt{\left(\hat{\mathbf{k}} \cdot (\mathbf{x}_i - \mathbf{x}_c)\right)^2 - ||\mathbf{x}_i - \mathbf{x}_c||_2^2 + \left(\mathcal{A} - \frac{b}{2}\right)^2}. && \text{(A.8)} \end{aligned}$$

Now we introduce time-dependence. While stations track a source of fixed $\hat{\mathbf{k}}$ (which we could identify with a tuple containing a right ascension and a declination), Δs_i will change (for each station indexed by i). This is because in reference frames which are *not* moving and spinning along with Terra, both the planet's core and the station positions are time-dependent: $\mathbf{x}_c = \mathbf{x}_c(t)$ and $\mathbf{x}_i = \mathbf{x}_i(t)$. Especially for sources in Sol's planetary system and in Via Lactea, the motion of Terra around Sol induces a parallax effect that causes no source to truly retain a fixed $\hat{\mathbf{k}}$; however, here we neglect this fact — most radio sources are extragalactic anyways. In the same spirit, we also

neglect the effect of *aberration of light*.

We end up with

$$\Delta s_i(\hat{\mathbf{k}},t) = \sqrt{\left(\hat{\mathbf{k}}\cdot(\mathbf{x}_i(t)-\mathbf{x}_c(t))\right)^2 - ||\mathbf{x}_i-\mathbf{x}_c||_2^2 + \left(A+\frac{b}{2}\right)^2} - \sqrt{\left(\hat{\mathbf{k}}\cdot(\mathbf{x}_i(t)-\mathbf{x}_c(t))\right)^2 - ||\mathbf{x}_i-\mathbf{x}_c||_2^2 + \left(A-\frac{b}{2}\right)^2}. \tag{A.9}$$

Note that within each square root, the second and third term are time-independent, as they are left unshaken by Terra's rotation around her axis.

In a shell-centred reference frame, Eq. A.9 simplifies. In such frames, we have $\mathbf{x}_c(t) = 0$, so that, upon invoking Eq. A.1, we find

$$\Delta s_i(\hat{\mathbf{k}},t) = \sqrt{\left(\hat{\mathbf{k}}\cdot\mathbf{x}_i(t)\right)^2 - ||\mathbf{x}_i||_2^2 + \left(||\mathbf{x}_0||_2+a+\frac{b}{2}\right)^2} - \sqrt{\left(\hat{\mathbf{k}}\cdot\mathbf{x}_i(t)\right)^2 - ||\mathbf{x}_i||_2^2 + \left(||\mathbf{x}_0||_2+a-\frac{b}{2}\right)^2}. \tag{A.10}$$

If Terra has a spherical surface cocentric with the ionospheric shell, and all stations lie at a distance R away from the centre, then Eq. A.10 further simplifies into

$$\Delta s_i(\hat{\mathbf{k}},t) = \sqrt{\left(R\cos\varphi_i(\hat{\mathbf{k}},t)\right)^2 - R^2 + \left(R+a+\frac{b}{2}\right)^2} - \sqrt{\left(R\cos\varphi_i(\hat{\mathbf{k}},t)\right)^2 - R^2 + \left(R+a-\frac{b}{2}\right)^2}. \tag{A.11}$$

where we acknowledge $\varphi_i = \varphi_i(\hat{\mathbf{k}},t)$ is the local zenith angle for direction $\hat{\mathbf{k}}$ at $\mathbf{x}_i\,(t)$. By taking out R, we make more explicit that Δs_i has dimensions of *length*, and find

$$\Delta s_i(\hat{\mathbf{k}},t) = R\left(\sqrt{\cos^2\varphi_i(\hat{\mathbf{k}},t) - 1 + \left(1+\frac{a}{R}+\frac{1}{2}\frac{b}{R}\right)^2} - \sqrt{\cos^2\varphi_i(\hat{\mathbf{k}},t) - 1 + \left(1+\frac{a}{R}-\frac{1}{2}\frac{b}{R}\right)^2}\right). \tag{A.12}$$

If this expression is correct, then for fixed a and b we should regain Chapter 2's flat

geometry when $R \to \infty$. To test this, consider the Taylor polynomial P of degree 1 of $\Delta s_i(\hat{\mathbf{k}}, t)$ around $\xi := \frac{b}{R} = 0$:

$$P_i\left(\hat{\mathbf{k}}, t, \xi\right) := \Delta s_i(\hat{\mathbf{k}}, t, \xi = 0) + \frac{\partial \Delta s_i\left(\hat{\mathbf{k}}, t, \xi\right)}{\partial \xi}|_{\xi=0} \cdot \xi. \tag{A.13}$$

Now $\Delta s_i(\hat{\mathbf{k}}, t, \xi = 0) = 0$, and

$$\begin{aligned}\frac{\partial \Delta s_i(\hat{\mathbf{k}}, t, \xi)}{\partial \xi} = \frac{R}{2}\Big(&\frac{1 + \frac{a}{R} + \frac{1}{2}\xi}{\sqrt{\cos^2 \varphi_i - 1 + \left(1 + \frac{a}{R} + \frac{1}{2}\xi\right)^2}} + \\ &\frac{1 + \frac{a}{R} - \frac{1}{2}\xi}{\sqrt{\cos^2 \varphi_i - 1 + \left(1 + \frac{a}{R} - \frac{1}{2}\xi\right)^2}}\Big).\end{aligned} \tag{A.14}$$

Thus,

$$\begin{aligned}P_i\left(\hat{\mathbf{k}}, t, \xi\right) &= \frac{\partial \Delta s_i\left(\hat{\mathbf{k}}, t, \xi\right)}{\partial \xi}|_{\xi=0} \cdot \xi \\ &= \frac{R\left(1 + \frac{a}{R}\right)}{\sqrt{\cos^2 \varphi_i - 1 + \left(1 + \frac{a}{R}\right)^2}} \cdot \frac{b}{R} \\ &= \frac{b\left(1 + \frac{a}{R}\right)}{\sqrt{\cos^2 \varphi_i - 1 + \left(1 + \frac{a}{R}\right)^2}}.\end{aligned} \tag{A.15}$$

Using now also that $a \ll R$, we recover

$$P_i\left(\hat{\mathbf{k}}, t\right) \to \frac{b}{\sqrt{\cos^2 \varphi_i}} = \frac{b}{|\cos \varphi_i|} = b \sec \varphi_i \tag{A.16}$$

for $0 \leq \varphi_i \leq 90°$.

What are typical values of $\frac{a}{R}$ and $\frac{b}{R}$ for Terra? Because $R \approx 6400$ km, and $a \approx 320$ km, we have $\frac{a}{R} \approx \frac{1}{20}$. Also $\frac{b}{R}$ is on the same scale, because $b \approx 320$ km is possible. So in practical scenarios, the regimes $\frac{a}{R} \ll 1$ and $\frac{b}{R} \ll 1$ do not fully hold.

The DTEC expectation value for stations with indices i and j under an ionosphere

with a constant FED mean μ, is

$$\mathbb{E}\left[\tau_{ij}\left(\hat{\mathbf{k}}, t\right)\right] = \mu\left(\Delta s_i(\hat{\mathbf{k}}, t) - \Delta s_j(\hat{\mathbf{k}}, t)\right). \tag{A.17}$$

In the special case that Terra's radius is very large compared to the ionosphere's height above the surface and thickness, we recover

$$\mathbb{E}\left[\tau_{ij}\left(\hat{\mathbf{k}}, t\right)\right] = \mu b\left(\sec\varphi_i(\hat{\mathbf{k}}, t) - \sec\varphi_j(\hat{\mathbf{k}}, t)\right), \tag{A.18}$$

as in Chapter 2.

B

Flux scale–induced spectral index uncertainties for high-SNR sources

Abstract

Context A standard problem in radio astronomy is the estimation of the spectral index α of a source with a power-law spectrum, based on images at two different frequencies. For sources imaged at high signal-to-noise ratio (SNR), the dominant flux density uncertainty — or specific intensity uncertainty — is typically the image's flux *scale* uncertainty. The LOFAR Two-metre Sky Survey (LoTSS) DR2, for example, suffers from a $\sim$10% flux scale uncertainty.

Aims We analytically and numerically characterise the effect of flux scale uncertainties on spectral index measurements $\hat{\alpha}$ for high-SNR sources.

Methods We derive an expression for the spectral index error $\varepsilon := \hat{\alpha} - \alpha$, whose distribution is, under a plausible assumption, exactly Gaussian. We derive expressions for its mean μ_ε and standard deviation σ_ε. The mean is non-zero: flux scale errors make the standard spectral index formula a *biased* estimator of the true spectral index. The magnitude of the relative bias $\frac{\mu_\varepsilon}{\sigma_\varepsilon}$ is independent of the images' frequencies.

Results For a range of realistic flux scale uncertainties, we numerically calculate the spectral index bias μ_ε, uncertainty σ_ε, and relative bias $\frac{\mu_\varepsilon}{\sigma_\varepsilon}$. The bias is insignificant compared to the uncertainty. The uncertainty itself, however, *is* significant.

Conclusions For high-SNR sources, flux scale uncertainties are the main obstacle to accurately measuring spectral indices. Conveniently, under reasonable conditions, the distribution of the measured spectral index $\hat{\alpha}$ approaches perfect Gaussianity as the source SNR increases. Generally speaking, $\hat{\alpha}$ is a biased estimator of α — remarkably even when the flux scale corrections applied to the imagery are on average correct. However, the bias appears small under realistic conditions: $\mu_\varepsilon \sim 10^{-3}$–$10^{-2}$ for LoLSS–LoTSS spectral index measurements. We offer a simple formula to remove this bias, actionable once flux scale uncertainties are quantified.

B.1 Introduction

Modern radio astronomical surveys, such as the LOFAR (van Haarlem et al., 2013) Two-metre Sky Survey (LoTSS) DR2 (Shimwell et al., 2022), go sufficiently deep and cover large enough areas of sky to detect millions of sources with high significance (i.e. $> 5\sigma$, where σ is the image background noise). Flux density uncertainties of such *high-σ* or *high-SNR* sources, as we will call them here, are no longer dominated by additive background noise of thermal nature. Instead, *multiplicative* noise becomes dominant. As an example, for the LoTSS DR2, flux scale uncertainty limits the accuracy of flux densities and specific intensities to 10% — regardless of a source's brightness.

Similar surveys at other frequencies are planned or currently ongoing, which will enable spectral index measurements of millions of high-σ sources in the foreseeable future. Such measurements are scientifically interesting for a myriad of reasons. One such reason is the search for ultra-steep-spectrum (USS) radio galaxies (RGs), which have been found to reside preferentially at high redshifts. Such RGs could help to probe the physics of the Epoch of Reionisation (e.g. Miley & De Breuck, 2008; Saxena et al., 2018). Another reason to amass spectral index measurements is to boost the search for MHz-peaked spectrum (MPS) and GHz-peaked spectrum (GPS) sources, which likely represent youthful RGs, RGs in dense environments, or transients. As such, they serve as important observational probes to test models of RG evolution (e.g. O'Dea et al., 1991; O'Dea & Saikia, 2021).

In this work, we analytically and numerically characterise the spectral index error introduced by multiplicative noise sources in the high-SNR limit.

B.2 Analytical results

The simplest synchrotron-emitting sources have a power-law spectrum: $F_\nu \propto \nu^\alpha$. Let us now consider such a source. In particular, let $F_\nu(\nu_1)$ and $F_\nu(\nu_2)$ denote the ground-

truth flux densities of the source at observing frequencies ν_1 and ν_2 respectively, and let α be the ground-truth spectral index. Then

$$\frac{F_\nu(\nu_1)}{F_\nu(\nu_2)} = \left(\frac{\nu_1}{\nu_2}\right)^\alpha, \text{ or } \alpha = \frac{\ln \frac{F_\nu(\nu_1)}{F_\nu(\nu_2)}}{\ln \frac{\nu_1}{\nu_2}} \text{ provided that } \nu_1 \neq \nu_2. \tag{B.1}$$

We usually approximate α through the estimator $\hat{\alpha}$, computed with the above formula, but with the ground-truth flux densities replaced by *observed* flux densities $\hat{F}_\nu(\nu_1)$ and $\hat{F}_\nu(\nu_2)$. For high-σ sources, the observed flux densities differ from the ground-truth flux densities by a *multiplicative* factor, because *additive* image noise is negligible (by definition): $\hat{F}_\nu(\nu_1) \approx F_\nu(\nu_1) \cdot C_1$ and $\hat{F}_\nu(\nu_2) \approx F_\nu(\nu_1) \cdot C_2$. Then

$$\hat{\alpha} := \frac{\ln \frac{\hat{F}_\nu(\nu_1)}{\hat{F}_\nu(\nu_2)}}{\ln \frac{\nu_1}{\nu_2}} \approx \frac{\ln \frac{F_\nu(\nu_1)\cdot C_1}{F_\nu(\nu_2)\cdot C_2}}{\ln \frac{\nu_1}{\nu_2}} = \frac{\ln \frac{F_\nu(\nu_1)}{F_\nu(\nu_2)} + \ln \frac{C_1}{C_2}}{\ln \frac{\nu_1}{\nu_2}} =: \alpha + \varepsilon. \tag{B.2}$$

We read off that the spectral index error is

$$\varepsilon = \frac{\ln C_1 - \ln C_2}{\ln \frac{\nu_1}{\nu_2}}. \tag{B.3}$$

What is the distribution of ε?

The multiplicative factors C_1 and C_2 might deviate from 1 because of a myriad of effects, such as errors in beam and ionospheric calibration. Just as the central limit theorem (CLT) predicts that the normalised sum (i.e. the average) of many independent random variables (RVs) follows a normal distribution, the *multiplicative* CLT predicts that the normalised *product* of many independent RVs follows a *lognormal* distribution. This justifies the assumption that $C_i \sim \text{Lognormal}(\mu_i, \sigma_i^2)$ for $i \in \{1, 2\}$. This in turn implies $\ln C_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, so that — elegantly — Eq. B.3's spectral index error has a normal distribution! The parameters μ_i and σ_i^2 are the mean and variance of the corresponding normal distribution: $\mathbb{E}[\ln C_i] = \mu_i$ and $\mathbb{V}[\ln C_i] = \sigma_i^2$. They relate to the mean and variance of C_i through

$$\mu_i = \ln \frac{\mu_{C_i}^2}{\sqrt{\mu_{C_i}^2 + \sigma_{C_i}^2}}, \quad \sigma_i^2 = \ln\left(1 + \frac{\sigma_{C_i}^2}{\mu_{C_i}^2}\right). \tag{B.4}$$

The most optimistic case occurs when $\mu_{C_i} = 1$, so that $\mathbb{E}\left[\hat{F}_\nu(\nu_i)\right] \approx F_\nu(\nu_i)\,\mathbb{E}[C_i] = F_\nu(\nu_i)\,\mu_{C_i} = F_\nu(\nu_i)$: the observed flux density is an unbiased estimator of the ground-

truth flux density. Although C_i might vary per pointing, teams will generally try to ensure $\mu_{C_i} = 1$, and we will consider this case from here onwards:

$$\mu_i\left(\mu_{C_i}=1\right) = -\frac{1}{2}\ln\left(1+\sigma_{C_i}^2\right),\ \ \sigma_i^2\left(\mu_{C_i}=1\right) = \ln\left(1+\sigma_{C_i}^2\right), \tag{B.5}$$

so that $\mu_i = -\frac{1}{2}\sigma_i^2$.

The mean of the spectral index error, which we will call the *(spectral index) bias*, is

$$\mathbb{E}\left[\varepsilon\right] = \frac{\mathbb{E}\left[\ln C_1\right] - \mathbb{E}\left[\ln C_2\right]}{\ln\frac{\nu_1}{\nu_2}} = \frac{\ln\left(1+\sigma_{C_1}^2\right) - \ln\left(1+\sigma_{C_2}^2\right)}{2\ln\frac{\nu_2}{\nu_1}}, \tag{B.6}$$

while the variance of the spectral index error, assuming that C_1 and C_2 are independent, is

$$\mathbb{V}\left[\varepsilon\right] = \frac{\mathbb{V}\left[\ln C_1\right] + \mathbb{V}\left[\ln C_2\right]}{\ln^2\frac{\nu_1}{\nu_2}} = \frac{\ln\left(1+\sigma_{C_1}^2\right) + \ln\left(1+\sigma_{C_2}^2\right)}{\ln^2\frac{\nu_2}{\nu_1}}. \tag{B.7}$$

Because ε is Gaussian, $\mathbb{E}\left[\varepsilon\right]$ and $\mathbb{V}\left[\varepsilon\right]$ fully characterise its distribution.

We recall that $\sqrt{x^2} = |x| = x\cdot\operatorname{sgn}(x)$ for all $x\in\mathbb{R}$ and $\operatorname{sgn}(x) = \operatorname{sgn}^{-1}(x)$ for all $x\in\mathbb{R}_{\neq 0}$, where sgn is the signum function. The standard deviation (SD) of the spectral index error, or more succinctly the *spectral index uncertainty*,[1] is

$$\sigma_\varepsilon \coloneqq \sqrt{\mathbb{V}\left[\varepsilon\right]} = \frac{\sqrt{\ln\left(1+\sigma_{C_1}^2\right) + \ln\left(1+\sigma_{C_2}^2\right)}}{\left|\ln\frac{\nu_2}{\nu_1}\right|}. \tag{B.8}$$

By combining Eqs. B.6 and B.8, we find the *relative bias* to be

$$\frac{\mu_\varepsilon}{\sigma_\varepsilon} = \frac{\ln\left(1+\sigma_{C_1}^2\right) - \ln\left(1+\sigma_{C_2}^2\right)}{2\sqrt{\ln\left(1+\sigma_{C_1}^2\right) + \ln\left(1+\sigma_{C_2}^2\right)}}\cdot\operatorname{sgn}\left(\nu_2-\nu_1\right). \tag{B.9}$$

We remark that the absolute value of the relative bias, $\left|\frac{\mu_\varepsilon}{\sigma_\varepsilon}\right|$, is independent of the measurement frequencies.

[1] Because $\varepsilon \coloneqq \hat{\alpha} - \alpha$ and α is a constant, $\mathbb{V}[\varepsilon] = \mathbb{V}[\hat{\alpha}]$. Consequently, $\sigma_\varepsilon \coloneqq \sqrt{\mathbb{V}[\varepsilon]} = \sqrt{\mathbb{V}[\hat{\alpha}]} =: \sigma_{\hat{\alpha}}$. As it is apt to call $\sigma_{\hat{\alpha}}$ the *spectral index uncertainty*, this name also applies to σ_ε.

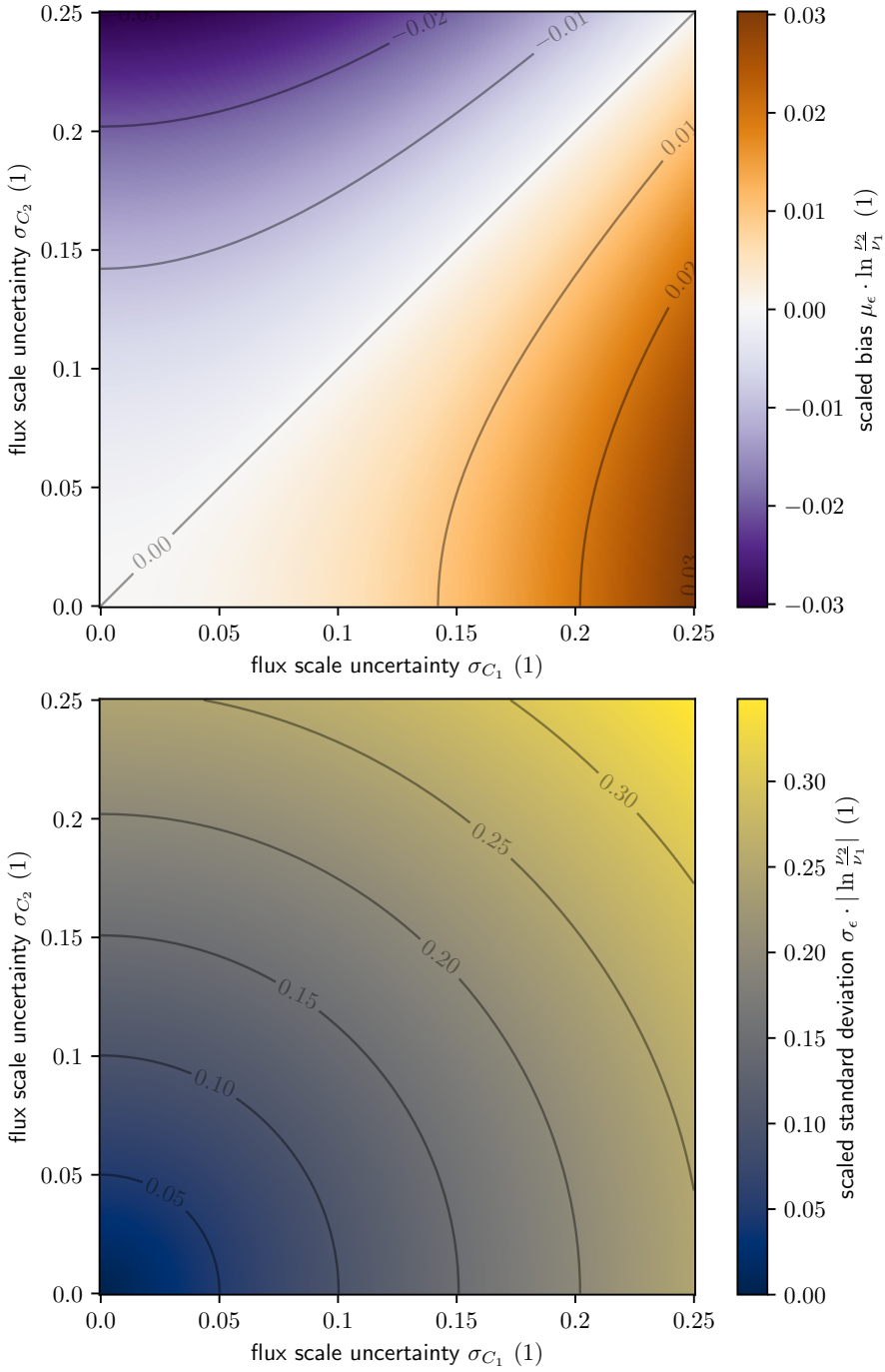

Figure B.1: Mean (top panel) and standard deviation (bottom panel) of the spectral index error ε for high-SNR sources, as a function of the flux scale uncertainties σ_{C_1} and σ_{C_2}. We scale both mean and standard deviation by a frequency-dependent factor. (This factor drops out if ν_2 is a factor e larger than ν_1. This condition is approximately satisfied for LoLSS–LoTSS spectral index measurements, where $\nu_1 = 54$ MHz and $\nu_2 = 144$ MHz.)

B.3 Numerical results

Having derived the statistical properties of the spectral index error ε that arises in spectral index measurements of bright (i.e. high-SNR) sources, we proceed to evaluating μ_ε, σ_ε, and $\frac{\mu_\varepsilon}{\sigma_\varepsilon}$ numerically.

B.3.1 Spectral index bias

From Eq. B.6, we see that the inferred spectral index $\hat{\alpha}$ is biased *positive* if the flux scale uncertainty for the lower-frequency image is *larger* than for the higher-frequency image. Similarly, $\hat{\alpha}$ is biased *negative* if the flux scale uncertainty for the lower-frequency image is *smaller* than for the higher-frequency image. We explicitly calculate the spectral index bias for the top panel of Fig. B.1. More precisely, we show $\mu_\varepsilon \cdot \ln \frac{\nu_2}{\nu_1}$, with the intent of reporting results that are independent of the particular pair of frequencies (ν_1, ν_2) used. We find that the bias is small: it is $\sim 10^{-2}$ if the σ_{C_i} differ by $\sim 10^{-1}$, and it is $\sim 10^{-3}$ if the σ_{C_i} differ by $\sim 10^{-2}$.

B.3.2 Spectral index uncertainty

From Eq. B.8, we see that the spectral index uncertainty σ_ε is invariant under exchange of σ_{C_1} and σ_{C_2}: the values of both flux scale uncertainties matter, but not to which image they belong. We explicitly calculate the spectral index uncertainty for the bottom panel of Fig. B.1. We multiply the results by the absolute value of the same frequency-dependent factor as in Sect. B.3.1. For LoTSS-like flux scale uncertainties, spectral index uncertainties for high-SNR sources are significant: $\sigma_\varepsilon \sim 10^{-1}$ for $\sigma_{C_i} \sim 10^{-1}$. Such uncertainties can interfere with drawing scientific conclusions from the estimated spectral indices.

B.3.3 Spectral index relative bias

We explicitly calculate the relative bias in Fig. B.2. For typical conditions, the bias appears to be a sub-dominant problem, being just a few percent of the spectral index uncertainty.

B.4 Conclusion

As radio surveys become more sensitive and ionospheric calibration methods improve, more source flux densities and specific intensities will be measured at high significance with respect to the image noise. For this growing population of high-SNR

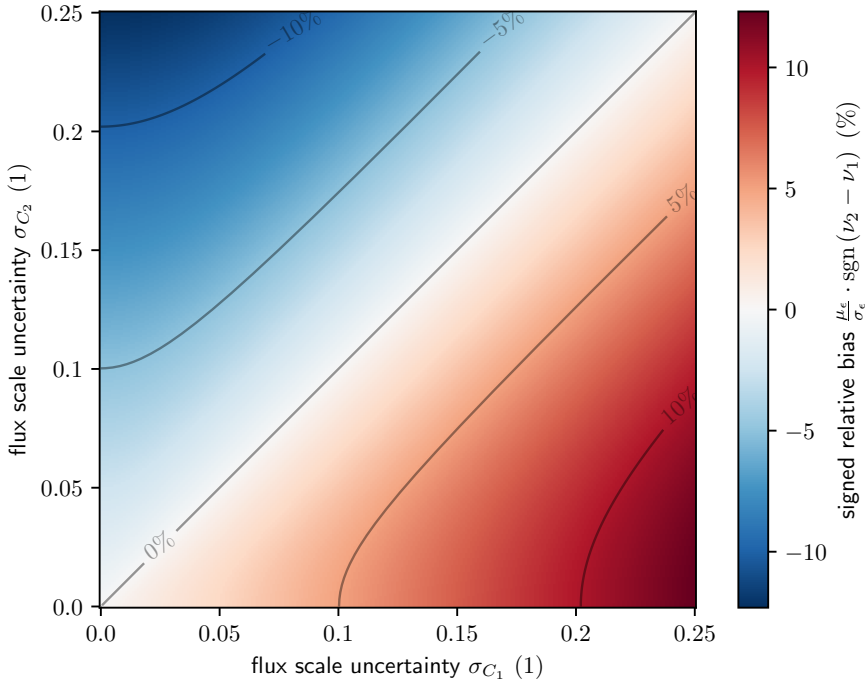

Figure B.2: Spectral index bias as a fraction of the spectral index uncertainty for high-SNR sources, as a function of the flux scale uncertainties σ_{C_1} and σ_{C_2}. The bigger the difference in flux scale uncertainty, the larger the relative bias.

sources, flux *scale* uncertainties typically dominate the flux density and specific intensity uncertainty budget. In this short work, we have considered — for such high-SNR sources — the statistical distribution of the measured spectral index $\hat{\alpha}$ and its error $\varepsilon := \hat{\alpha} - \alpha$, where α is the ground truth spectral index. We assume the standard definition for $\hat{\alpha}$: i.e. the logarithm of the ratio between the observed flux densities divided by the logarithm of the ratio between the observing frequencies. Under the assumptions that the image background noise is of negligible importance, and that the flux scale factors C_i are lognormally distributed — as suggested by the multiplicative CLT — $\hat{\alpha}$ is *exactly* Gaussian. We show that $\hat{\alpha}$ is a biased estimator of α, even when the flux scale corrections applied to the imagery are on average correct (i.e. $\mu_{C_i} = 1$). Fortunately, the spectral index bias appears small under realistic conditions. More precisely, if $\sigma_{C_2} - \sigma_{C_1} \sim 10^{-2}$–$10^{-1}$, as expected for LoLSS–LoTSS spectral index measurements, the bias $\mu_\varepsilon \sim 10^{-3}$–$10^{-2}$. More problematically, the spectral index uncertainty does not appear small: $\sigma_\varepsilon \sim 10^{-1}$. Thus, flux scale uncertainties $\sigma_{C_i} \sim 10^{-1}$ can impede the physical interpretation even of sources that are detected at very high significance with respect to the image noise (e.g. with *infinite* SNR). When the σ_{C_i} are known, Eq. B.6 offers a simple formula to remove spectral index bias.