Completeness of the Gaia-verse V: Astrometry and Radial Velocity sample selection functions in Gaia EDR3

Andrew Everall\textsuperscript{1} and Douglas Boubert\textsuperscript{2,3}

\textsuperscript{1}Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
\textsuperscript{2}Magdalen College, University of Oxford, High Street, Oxford OX1 4AU, UK
\textsuperscript{3}Rudolf Peierls Centre for Theoretical Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK

Accepted XXX. Received YYY; in original form ZZZ.

ABSTRACT

We risk reaching false scientific conclusions if we test our physical theories against subsets of the Gaia catalogue without correcting for the biased process by which stars make it into our sample. In this paper we produce selection functions for three Gaia science samples to enable the community to correct for this bias. We estimate the probability that a source in Gaia EDR3 will have i) a reported parallax and proper motion, ii) an astrometric renormalised unit weight error below 1.4, or iii) a reported radial velocity. These selection functions are estimated as a function of $G$-band apparent magnitude and position on the sky, with the latter two also being dependent on $G-G_{RP}$ colour. The inferred selection functions have a non-trivial dependence on these observable parameters, demonstrating the importance of empirically estimating selection functions. We also produce a simple estimate for the selection function of the full Gaia EDR3 source catalogue to be used with the subset selection functions. We make all of our selection functions easily accessible through the GitHub repository

\texttt{https://github.com/s.pc/e.pc/l.pc/e.pc/c.pc/t.pc/i.pc/o.pc/n.pc/f.pc/u.pc/n.pc/c.pc/t.pc/i.pc/o.pc/n.pc/s.pc}.

Key words: stars: statistics, Galaxy: kinematics and dynamics, Galaxy: stellar content, methods: data analysis, methods: statistical

1 INTRODUCTION

As observational astronomers, we use data from observatories such as the Gaia satellite (Gaia Collaboration et al. 2016) to test our physical theories about stars, galaxies and the universe. However, to reliably test these theories we must understand our data and account for regions of parameter space which couldn’t be observed. No observatory paints a complete picture of the universe and in order to fill in the gaps with our models, we need to know what could have been observed and what could not. We achieve this through a selection function. For any catalogue of observed sources the selection function tells us “Given a real or hypothetical object, what is the probability that the object would have been successfully observed and included in the catalogue?”. A detailed exposition of selection functions and their applications is provided in (Rix et al. 2021).

The Gaia DR2 selection function, modelled in Boubert & Everall (2020), has enabled us to answer the question “What is the probability that a source was observed and recorded in the Gaia DR2 source catalogue?”. However, additional information is provided for subsets of objects in the Gaia source catalogue which are vital for answering certain scientific questions. For example, the sample with measured parallax and proper motion (Lindgren et al. 2018, 2021) is enabling us to study and understand the structure and kinematics of the Milky Way, whilst the sample with radial velocities (Gaia Collaboration et al. 2018) includes the final dimension of kinematic data for a smaller, but no less impressive, set of objects.

Usually selection functions are estimated by comparing the sample to a more complete source catalogue. For the Gaia source catalogue, there is no more complete sample to compare against. Therefore we had to use our understanding of the observing strategy of Gaia to model the selection function (Boubert & Everall 2020). However, for subsets of Gaia, the source catalogue provides a sample which we know is more complete in all areas of parameter space. Previous attempts at selection functions for samples where a more complete catalogue is available involve taking the ratio of objects in the subset to source catalogue in carefully chosen on-sky position-colour-apparent magnitude bins (Bovy et al. 2012; Wojno et al. 2017; Mints & Hekker 2019, e.g.). This requires a selection function which can be defined discretely with enough objects in each bin of the subset and source catalogues such that the ratio is approximately equal to the expected completeness. This is not the case for subsets of Gaia which have complex selection criteria within the processing pipeline relating to satellite performance and the scanning law (Boubert et al. 2020, 2021b).

A strong example of this is the Gaia DR2 radial velocity catalogue (which we will address in this paper). Rybizki et al. (2021) attempted to evaluate the selection function for this sample rel-
ative to the Gaia source catalogue using number count ratios in colour-magnitude-sky bins. However, as they demonstrate, the selection function has a strong dependence on sky position over small scales due to the scanning law, crowding limitations and the Initial Gaia Source List (IGSL, Smart & Nicastro 2014). Poisson noise prevented them from going to sufficient resolution to model this structure without running out of objects in Gaia.

Everall & Das (2020) went beyond simple count ratios by modelling the source catalogue and subset CMDs as continuous Poisson count processes with Gaussian Mixture Models. Whilst this model was well-suited to multi-fibre spectrographs with well defined fields, a significant improvement was needed to fit the all-sky selection function of Gaia catalogues.

This major step forward has come from Boubert & Everall (2021) which presents a general approach to estimating selection functions for subsets of catalogues. They model the selection function as a sum of well-localised spherical needlets on the sky with weights drawn from a Gaussian Process across colour and apparent magnitude. They also use the formally correct Binomial likelihood function to fit the model.

In this work we will apply the Boubert & Everall (2021) model to three subsets of Gaia EDR3:

(i) Astrometry: the sample with published parallax and proper motion.

(ii) RUWE: Renormalised unit weight error specifies the goodness of fit of an astrometric solution and is often used to filter out poorly fit sources. We provide selection functions for the sample with RUWE < 1.4

(iii) RVS: The sample with published radial velocities from the radial velocity spectrograph (RVS).

We’ll provide a brief review of the model developed by Boubert & Everall (2021) in Section 2 and explain how binomial statistics are used to fit the likelihood function and test the results in Section 3. Readers primarily interested in the results may wish to skip this and go directly to Section 4 where the selected samples are described. We present our results in Section 5. Before concluding we also discuss other Gaia catalogues which have not been investigated here due to additional complexities but which should be a key goal for future work in this field.

2 METHODOLOGY

An in-depth description of the method for constructing sub-sample selection functions for astronomical surveys is given in Boubert & Everall (2021). Here we provide a brief summary of the key points which are used in this paper.

The aim of this work is to estimate the selection probabilities of subsets of the Gaia EDR3 relative to the source catalogue. e.g. ‘What is the probability that a source in Gaia EDR3 has parallax and proper motion?’.

These selection functions may be written down as

$$P(S_{\text{subset}} | S_{\text{source}}, \mathbf{y})$$

where \( \mathbf{y} \) are the observables (or functions of observables) over which the selection function is defined and ‘source’ refers to the more complete catalogue, in this case the Gaia source catalogue. The probability of any object being included in the subset is

$$P(S_{\text{subset}} | \mathbf{y}) = P(S_{\text{subset}} | S_{\text{source}}, \mathbf{y}) \cdot P(S_{\text{source}} | \mathbf{y})$$

where the source catalogue selection function, \( P(S_{\text{source}} | \mathbf{y}) \), is the probability that a source is included in the Gaia source catalogue. The source catalogue selection function for Gaia DR2 has previously been estimated by Boubert & Everall (2020) and a significantly improved version for the EDR3 catalogue is being developed (Fraser et al. 2021; Boubert et al. 2021a). For users who wish to apply these models in their own work sooner, we provide an initial estimate of the Gaia EDR3 source catalogue selection function in Appendix B, however, this will be superseded by Fraser et al. (2021); Boubert et al. (2021a) when published.

All selection functions estimated here will be over the variables \( \mathbf{y} = (l, b, G) \) with the RUWE and RVS selection functions additionally a function of \( G - G_{\text{BP}} \). These are the dominant variables in selecting sources to enter the samples. The scanning law produces complex selection patterns across the sky generating a heavy dependence on \((l, b)\). Whether individual observations are used in the Gaia data is dependent on the on-board estimated apparent magnitude of the source (scientific measurements of a source are only made if \( G_{\text{onboard}} < 20.7 \), Gaia Collaboration et al. (2016)), for which \( G \) is a reasonable proxy. The publication of measured radial velocity for any source in Gaia is contingent on an estimated RVS magnitude \((G_{\text{RVS}})\) calculated using the IGSL (See Section 2.1 in Sartoretti et al. 2018). As will be discussed later, \( G_{\text{RVS}} \) is more similar to \( G_{\text{BP}} \) than \( G \) which means the magnitude limit in \( G \) will be a function of \( G - G_{\text{BP}} \). Finally, the selection criteria for the RVS sample is explicitly dependent on source temperature and RUWE is implicitly dependent on colour. Making our selection function dependent on \( G - G_{\text{BP}} \) allows us to capture these dependencies. We choose to use \( G - G_{\text{BP}} \) rather than \( G_{\text{BP}} - G_{\text{RP}} \) due to calibration issues at the faint end of \( G_{\text{BP}} \) (Riello et al. 2021). The drawback of this is that \( G_{\text{BP}} \) uses larger spatial windows than \( G \) which means that our colours will be extremely red for extended sources with high excess flux (see Section 9.4 Riello et al. 2021).

Whilst we only consider magnitude, colour and position on the sky in this work, we are also conscious that Gaia is heavily limited in crowded regions of the sky such as the Milky Way bulge, LMC and globular clusters. We will discuss why we have not directly used crowding as a selection function variable and the implications of this in Section 6.

The selection function, as described in Boubert & Everall (2021), is composed of a sum of spherical needlets across the sky with coefficients subject to a Gaussian Process prior in apparent magnitude and colour. This is fit to the data using a Binomial likelihood function in HEALPix-apparent magnitude-colour bins. In the following subsections, we’ll briefly describe the maths of each of these components.

2.1 Logit probability

Our model will estimate the probability, \( q \in [0, 1] \) that an object will be included in the subset given the observables and that it’s in the source catalogue. Our model is defined in an infinite domain so we directly model \( x \in [-\infty, \infty] \), the logit-transformed probability

$$x = \logit(q) \equiv \log \left( \frac{q}{1 - q} \right),$$

$$P(S_{\text{subset}} | \mathbf{y}) = P(S_{\text{subset}} | S_{\text{source}}, \mathbf{y}) \cdot P(S_{\text{source}} | \mathbf{y})$$

where the source catalogue selection function, \( P(S_{\text{source}} | \mathbf{y}) \), is the probability that a source is included in the Gaia source catalogue. The source catalogue selection function for Gaia DR2 has previously been estimated by Boubert & Everall (2020) and a significantly improved version for the EDR3 catalogue is being developed (Fraser et al. 2021; Boubert et al. 2021a). For users who wish to apply these models in their own work sooner, we provide an initial estimate of the Gaia EDR3 source catalogue selection function in Appendix B, however, this will be superseded by Fraser et al. (2021); Boubert et al. (2021a) when published.

All selection functions estimated here will be over the variables \( \mathbf{y} = (l, b, G) \) with the RUWE and RVS selection functions additionally a function of \( G - G_{\text{BP}} \). These are the dominant variables in selecting sources to enter the samples. The scanning law produces complex selection patterns across the sky generating a heavy dependence on \((l, b)\). Whether individual observations are used in the Gaia data is dependent on the on-board estimated apparent magnitude of the source (scientific measurements of a source are only made if \( G_{\text{onboard}} < 20.7 \), Gaia Collaboration et al. (2016)), for which \( G \) is a reasonable proxy. The publication of measured radial velocity for any source in Gaia is contingent on an estimated RVS magnitude \((G_{\text{RVS}})\) calculated using the IGSL (See Section 2.1 in Sartoretti et al. 2018). As will be discussed later, \( G_{\text{RVS}} \) is more similar to \( G_{\text{BP}} \) than \( G \) which means the magnitude limit in \( G \) will be a function of \( G - G_{\text{BP}} \). Finally, the selection criteria for the RVS sample is explicitly dependent on source temperature and RUWE is implicitly dependent on colour. Making our selection function dependent on \( G - G_{\text{BP}} \) allows us to capture these dependencies. We choose to use \( G - G_{\text{BP}} \) rather than \( G_{\text{BP}} - G_{\text{RP}} \) due to calibration issues at the faint end of \( G_{\text{BP}} \) (Riello et al. 2021).
with $x$ sometimes being referred to as the log-odds ratio. This is then inverse transformed to retrieve the selection probability

$$q = \log^{-1}(x) = \frac{1}{1 + \exp(-x)}$$

(4)

de describes the Beta-Binomial expected value and p-value test which could have been selected from the source catalogue in the given bin $i$ and the selection probability $q$. This is not to suggest that our prior is "uninformative", as we will see in the results the prior does inform the model in regions of parameter space lacking data, however, we will explain in Section 6 that these regions are not of significant concern.

Some readers may be more used to our method being an example of "regularised regression". Whilst this is correct we apply our priors explicitly on the parameters of the model as opposed to having a broader regularization term. Therefore the model will always be discussed in the context of Bayesian statistics with a likelihood function on the data and priors on all parameters.

3 BINOMIAL STATISTICS

The selection function fit is based on the Binomial likelihood. For each bin, the number of sources in the sub-sample is assumed to be drawn from a Binomial distribution, given the number of sources which could have been selected from the source catalogue in the given bin $i$ and the selection probability $q$.

In this section we’ll briefly revise the likelihood function and describe the Beta-Binomial expected value and p-value test which will be used for analysing the results.

3.1 Binomial Likelihood

The source catalogue and subset are counted in HEALPix-colour-magnitude bins. The overall likelihood is given by

$$L = \prod_{i=p,m,c} \text{Binomial}(k_i | n_i, q_i)$$

(11)

where $p$, $m$, $c$ is the HEALPix-magnitude-colour bin index and $q_i$ is the model probability at the bin center (Eq. 8). Since the Binomial coefficient is independent of the selection probability, this can be dropped out as a constant and the log likelihood simplifies to

$$\log L \sim \sum_{i=p,m,c} k_i \log(q_i) + (n_i - k_i) \log(1 - q_i)$$

(12)

To optimize the likelihood function in terms of model parameters, $\beta_{jk}$, we use the L-BFGS-B algorithm (Zhu et al. 1997)

2 Referred to in those works as 'Mexican' Needlets.

Gaia Astrometry and RVS Selection Functions

2.3 Gaussian Process Prior

The free parameters $\beta_{jk}$ are modelled as a function of apparent magnitude $G$ and colour $C = G - G_{BP}$. This is performed in discrete colour-magnitude bins with a Gaussian Process prior placed on each dimension independently

$$\{\beta_{jk}\}_{mc} \sim \mathcal{GP}\left((G)_m \mid (C)_c\right)$$

(9)

where we use an independent Squared Exponential kernel for each of apparent magnitude and colour,

$$K(G, C, G', C') = \sigma^2 \exp\left(-\frac{(G - G')^2}{2l_m^2}\right) \exp\left(-\frac{(C - C')^2}{2l_c^2}\right),$$

(10)

with apparent magnitude and colour length-scales $l_m$ and $l_c$ and a variance $\sigma^2$, all of which are free parameters of that kernel.

The mean of the Gaussian Process is also a free parameter. In this work we set the mean to zero everywhere as $x = 0$ corresponds to $q = 0.5$ which is also the mean of a uniform prior reflecting the fact that we have no a priori information on the sample selection probabilities. This is not to suggest that our prior is "uninformative", as we will see in the results the prior does inform the model in regions of parameter space lacking data, however, we will explain in Section 6 that these regions are not of significant concern.

Some readers may be more used to our method being an example of "regularised regression". Whilst this is correct we apply our priors explicitly on the parameters of the model as opposed to having a broader regularization term. Therefore the model will always be discussed in the context of Bayesian statistics with a likelihood function on the data and priors on all parameters.
implemented in scpy. The boundaries are placed at $[-50, 50]$ for the unscaled parameters for which the prior distribution is a unit variance Gaussian. In other words, a parameter would have to be a 50-sigma outlier from the prior to reach the optimization boundaries.

### 3.2 Beta-Binomial Posterior

For each bin, we can independently estimate the posterior distribution of $q$ which will be useful when testing the results. This is not used to fit our model but instead to help understand the results. Consider the stars in one bin to be like identical marbles in a bag. You get handed the marbles one by one and choose whether keep the marble or give it away. Eventually, of the original bag of $n$ marbles, you’re left with $k$ in your hand. What is the expected probability of any marble being selected?

The Binomial distribution gives the likelihood of choosing $k$ marbles given the selection probability $q$.

$$P(k | n, q) = \text{Binomial}(k \mid n, q) = \binom{n}{k} q^k (1 - q)^{n-k}. \quad (13)$$

To evaluate the posterior probability distribution of $q$, we apply Bayes theorem but first we need a prior. The Beta distribution is Bayes theorem but first we need a prior. The Beta distribution is uniformly distributed between 0 and 1. This test will be used with all of our inferred selection functions to check that they have accurately captured the information in the bins.

### 3.3 p-value test

We can verify the case for a selection probability model using a p-value test (in this case we use a one-tailed p-value). This answers the question ‘Given the model, what is the probability that a measurement of an observable would not be larger than the observed value?’. For a Binomial distribution, the question is ‘What is the probability that less than $k$ sources out of $n$ are observed in this bin given the bin’s selection probability?’. The Binomial p-value is given by

$$p_{\text{value}} = U \left[ \sum_{k=0}^{k-1} \text{Binomial}(k' \mid n, q), \sum_{k=0}^{k' \mid n, q} \right]. \quad (17)$$

which we explain in more detail in Appendix A. Note that the p-value is not a deterministic value but a random variable. For example, if we have no data in a bin ($k = n = 0$) the p-value simplifies to $p_{\text{value}} = U [0, 1]$.

As in any one-tailed p-value test, if the model has successfully reproduced the data, the set of p-values for all data points will be uniformly distributed between 0 and 1. This test will be used with all of our inferred selection functions to check that they have accurately captured the information in the bins.

### 4 DATA

The selection function is estimated for three samples: Astrometry, RUWE and RVS. Here we’ll provide a brief description of each sample and how it can be accessed from the *Gaia* archive.

#### 4.1 Astrometry

The astrometry sample is the subset of *Gaia* EDR3 with published parallaxes and proper motions from the Astrometric Global Iterative Solution (AGIS, Lindegren et al. 2012). There are three criteria for a source having published parallax and proper motion (Lindegren et al. 2021):

- $N_{\text{VPU}} \geq 9$
- $G_{\text{DR2}} < 21.0$
- $\sigma_{\text{5Dmax}} < 1.2 \text{mas} \times \gamma (G)$

The VPU (VISIBILITY_PERIODS_USED) cut depends on the number of observations a source has received and their distribution in time. This produces a strong dependence of the selection function on position on the sky. The apparent magnitude cut is performed on DR2 apparent magnitudes and all magnitudes have been recomputed in EDR3. Therefore the $G_{\text{DR2}}$ cut doesn’t produce a discontinuous change in the selection function with $G$ however it will still lead to a distinct drop-off towards the faint end. Finally, the cut on $\sigma_{\text{5Dmax}}$ (ASTROMETRIC_SIGMA5D_MAX) will depend on position on the sky and apparent magnitude, because both contribute to the expected astrometric uncertainty for any source as discussed in Everall et al. (2021c). The effects of these cuts will be empirically modelled in this work without directly considering the effects of each cut on their own.

The distribution of sources as a function of colour and apparent magnitude is shown in Fig. 1 and as a function of position on the sky in Fig. 2. The SQL query for accessing the data is as follows.

```sql
SELECT * FROM gaiaedr3.gaia_source
WHERE astrometric_params_solved=3
AND phot_g_mean_mag between 1.6 and 22
```

The method requires upper and lower limits for the final bounds of the magnitude bins. There are no sources in *Gaia* EDR3 with apparent magnitude brighter than $G = 1.6$ and none with parallax or proper motion fainter than $G = 22$ so we use these as our magnitude limits. We discuss the selection function outside these limits in...
Figure 1. The number of *Gaia* EDR3 sources in each colour-magnitude bin is shown for the full sample (left) astrometry sample (middle left) astrometry with RUWE $\leq 1.4$ sample (middle right) and sample with DR2 radial velocity published in the EDR3 catalogue (right). The astrometry selection removes sources from the dim end whilst RVS is cut on $G_{\text{RVS}}$ producing an extended drop-off in $G$ as can be seen from the histogram on the right. Red dashed lines show the region of parameter space used to fit the selection functions.

Figure 2. The counts of each sample in HEALPix level 6 bins across the sky summed over the colour-apparent magnitude range used for the selection function. The RUWE sample shows slightly reduced star counts in the highest density regions compared with the Astrometry and RVS is significantly more spatially uniform due to the bright magnitude limit biasing the spectrograph towards nearby sources which are more spatially uniform and because it is challenging for RVS to assign individual windows to sources in crowded regions. The RVS star counts also show features of the scanning law and vertical stripes in the East and squares in the Equatorial South deriving from the SDSS and Schmidt photographic plate contributions to the Initial Gaia Source List (Rybizki et al. 2021).

Section 6. This results in a sample of 1,465,211,050 sources out of a total 1,806,195,366 in *Gaia* EDR3 with $1.6 < G < 22$.

4.2 RUWE

When using *Gaia* astrometry, various cuts are often placed on the sample to generate a subset with maximal information and minimal systematics. One such cut recommended by DPAC is RUWE $< 1.4$ (Gaia Collaboration et al. 2021a). Renormalised Unit Weight Error (RUWE) is the reduced chi-square of the individual position measurements of a source given the source’s astrometric solution renormalised by the 41st percentile of the catalogue as a function of colour and apparent magnitude$^3$. Sources with large RUWE have scatter between astrometric measurements which is poorly fit by the linear astrometry model. A common cause of this is binary motion (Lindegren et al. 2018; Belokurov et al. 2020; Penoyre et al. 2020) however this can also be generated by source contamination in crowded regions and possibly even astrometric microlensing.

---

$^3$ RUWE is explained in more detail in the document “Re-normalising the astrometric chi-square in *Gaia* DR2” which is available from [https://www.cosmos.esa.int/web/gaia/public-dpac-documents](https://www.cosmos.esa.int/web/gaia/public-dpac-documents)
(McGill et al. 2020). Whilst these sources can be astrophysically interesting, they introduce systematic errors in the AGIS pipeline and it is recommended to remove the more extreme cases in order to clean the sample.

There have been other recommended cuts for removing spurious astrometric solutions on ipo_gof_harmonic_amplitude (Fabricius et al. 2021) or astrometric_gof_al. (Lindgren et al. 2021) however RUWE < 1.4 is commonly used in the community so we keep to this single cut here.

Whilst an ideal telescope of Gaia’s configuration would have a reflection symmetric PSF in the perpendicular along-scan and across-scan directions, the PSF of the Gaia spacecraft is slightly asymmetric on the CCD panel. (Rowell et al. 2021), introducing chromatic behaviour to the inferred centroid location (for full discussion see Lindgren et al. 2021, Section 2.3). This propagates through to the astrometric solution where the unit weight error has a colour dependence. RUWE is the renormalised unit weight error where the renormalisation is to the 41st percentile of the unit weight error as a function of colour and apparent magnitude. As a result, the 41st percentile is achromatic however the spread of RUWE through the observed population will still have a residual colour dependence as will the fraction of sources with RUWE > 1.4. Therefore colour dependence is an important aspect of the RUWE<1.4 selection function.

The problem with modelling RUWE as a function of colour is that we require all sources to have published colour. Approximately 88% of sources in the Gaia EDR3 catalogue have published $G_{\text{RP}}$. Therefore the selection function we will actually be modelling is $P(\text{RUWE} < 1.4 | S_{\text{Gaia}}, G_{\text{RP}})$. If the event that a source has RUWE < 1.4 is entirely independent of the the event that it has published $G_{\text{RP}}$ this probability is the same as $P(\text{RUWE} < 1.4 | S_{\text{Gaia}})$.

Within the apparent magnitude range 1.6 < $G$ < 22 there are 1.81 billion sources in Gaia, 1.40 billion of which have RUWE < 1.4, therefore

\[
P(\text{RUWE} < 1.4 | S_{\text{Gaia}}) \sim 77.6\%.
\]

1.55 billion sources in Gaia with 1.6 < $G$ < 22 have published $G_{\text{RP}}, 1.29$ billion of which have RUWE < 1.4 such that

\[
P(\text{RUWE} < 1.4 | S_{\text{Gaia}}, G_{\text{RP}}) \sim 83.3\%.
\]

Therefore, whether a source has published $G_{\text{RP}}$ affects the probability that the source will have RUWE < 1.4.

We show this as a function of apparent magnitude in Fig. 4 where we give the median and 16th – 84th percentile ranges of the probability of a source having $G_{\text{RP}}$ from the full sample or given the source satisfies the RUWE cut. For 16 < $G$ < 21, a source is significantly more likely to have $G_{\text{RP}}$ if the source has published RUWE < 1.4.

Therefore we advise that the colour-dependent RUWE selection function should only be used in conjunction with the $G_{\text{RP}}$ selection function

\[
P(S_{\text{GUWE}}, S_{\text{Gaia}}, y) = P(S_{\text{GUWE}} | S_{\text{Gaia}}, y) \cdot P(S_{\text{Guia}} | S_{\text{Gaia}}, y)
\]

which is the probability of a source in EDR3 having both RUWE < 1.4 and published $G_{\text{RP}}$. Since the selection function for $G_{\text{RP}}$ is not known, and for cases where one wishes to fit a model to Gaia data without colour dependence, we also fit a magnitude-only selection function to the RUWE data including all sources with published $G$ independent of whether $G_{\text{RP}}$ was published. By evaluating this, we’re implicitly marginalising over the colour distribution of sources

\[
P(S_{\text{GUWE}} | y) = \int dC P(S_{\text{GUWE}} | S_{\text{Gaia}}, y, C) \cdot P(C | S_{\text{Gaia}}, y)
\]

where $P(C | S_{\text{Gaia}}, y)$ is the distribution of source colours in the Gaia source catalogue at the given position on the sky and apparent magnitude, $y$.

The distributions of all Gaia sources and those with RUWE < 1.4 as a function of $G$ and $y$ are shown in the first and third panels of Fig. 1 respectively. Comparing with the Gaia Catalogue of Nearby Stars (Gaia Collaboration et al. 2021b) there are sources extending out to the blue ($G - G_{\text{RP}} < -1$) and red ($G - G_{\text{RP}} > 3$) extremes of the colour distribution. The red sources will likely be extended sources or stars in crowded regions with high excess flux as mentioned in Section 2. The blue sources are very small in number and only appear at the faint end suggesting they are driven by photometric measurement errors for faint sources. We will only fit the colour-dependent selection for the range $-1 < G - G_{\text{RP}} < 7$ which contains sources with well measured colours including those with high excess flux.

\[
\text{SELECT * FROM gaiaedr3.gaia_source WHERE ruwe<1.4}
\]

\[\text{WHERE ruwe<1.4}\]
Figure 4. The probability of a source having $G_{BP}$ is evaluated from the Beta-Binomial posterior distribution in Equation 16. For $G > 15$ there are few enough sources in RVS that this isn’t significantly different to the full sample probability however for $16 \leq G \leq 21$ the RUWE $G_{BP}$ probability changes strongly from the full sample which could lead to significant biases in the RUWE colour selection function if not used with a correct $G_{BP}$ selection function.

5. AND phot_rp_n_obs>0 
6. AND phot_g_mean_mag BETWEEN 1.6 AND 22 
7. AND phot_g_mean_mag-phot_rp_mean_mag BETWEEN -1 AND 7

The RUWE sample has 1,292,152,210 sources out of a total 1,550,860,236 in the Gaia EDR3 source catalogue with colours and apparent magnitudes in the given range.

4.3 RVS

The $Gaia$ radial velocity sample is incredibly important to dynamical studies of the Milky Way (e.g. Nitschai et al. 2020). Radial velocities are measured by the $Gaia$ satellite’s on-board spectrograph using the calcium triplet. Radial velocities have not yet been released for $Gaia$ DR3, however the DR2 sample of 7 million RVS stars is still by far the largest sample of stellar radial velocities available from any single observatory. The selection criteria used to produce the DR2 sample is given in Gaia Collaboration et al. (2018). These are some of the main cuts:

- $G_\text{ext}^\text{RVS} < 12$ or $G_\text{int}^\text{RVS} < 14$
- $3550 < T_{\text{eff}} < 6900$
- No double line spectroscopic binaries or emission line stars.

$G_\text{ext}^\text{RVS}$ is the apparent magnitude estimated through photometric transformations of observations from ground-based observatories in the Initial Gaia Source List (IGSL, Smart & Nicastro 2014). $G_\text{int}^\text{RVS}$ is directly estimated from the $Gaia$ spectroscopy data in the radial velocity pipeline. In either case, $G_{BP}$ provides a better approximation to $G_{RVS}$ than $G$. However since the $Gaia$ source catalogue selection function is modelled as a function of $G$, we keep to that here in the interests of simplicity and usability.

Selection on IGSL-measured $G_{RVS}$ produces substantial structure across the sky as shown in the right-hand panel of Fig. 3 with vertical lines in the East due to SDSS (Strauss et al. 2002) IGSL sources and a grid pattern around the South equatorial pole from the Guide Star Catalogue (Lasker et al. 2008). This is demonstrated and discussed in more detail in Rybizki et al. (2021).

Motivated by Boubert et al. (2019), additional analysis and cleaning of the DR2 sample took place for the RVS sources published with EDR3 (Seabroke et al. 2021). 3,876 sources with incorrect radial velocities due to nearby neighbours were removed and 10,924 could not be successfully crossmatched with any source in
Figure 6. The results of the astrometry selection function fit (middle row) show a striking drop in high source density regions, not just in the disk and Magellanic clouds but also in the Milky Way globular clusters. There is also scanning structure at the faint end which significantly reduces the selection function probability in under-scanned regions. The model shows strong agreement with the approximate expected distribution from the data (top row). This is demonstrated more precisely with the Binomial p-value tests which are dominated by random noise. In the faint bins, some structure arises due to lack of resolution in the model, as such our model should only be trusted to ~ 2 degree scales.

Gaia EDR3. We also apply the cut \( \text{rv_nb_transits} \geq 4 \) recommended by Boubert et al. (2019) to clean out spurious radial velocity measurements.

The distribution of RVS sources as a function of apparent magnitude and \( G - G_{\text{RP}} \) colour is shown in the fourth panel of Fig. 1. The RVS sample occupies a very narrow colour range and is limited to the bright end of the Gaia magnitude range. Only the range of colour and apparent magnitude containing radial velocity sources is used for the RVS sample, namely \(-0.6 < G - G_{\text{RP}} < 2.6 \) and \(1.6 < G < 17.4 \), where we have applied the same bright-end cut as in the astrometry sample.

Once again, we are faced with the same issue as in the RUWE selection where not all RVS sources have published \( G_{\text{RP}} \). However the RVS selection function is only non-zero at brighter magnitudes where \( G_{\text{RP}} \) is much more complete. For \(1.6 < G < 17.4 \), 6.2 million out of 206 million sources, or 3.00%, have published DR2 radial velocity which rises to 3.04% of sources with published \( G_{\text{RP}} \). The RVS selection function has a much weaker dependence on \( G_{\text{RP}} \) selection than is the case for RUWE. By using the colour-dependent RVS selection function without accounting for the \( G_{\text{RP}} \) selection probability, a ~ 1% systematic uncertainty would be introduced to the results. This is also shown in Fig. 4 where a dotted blue line and shaded blue regions show that the \( G_{\text{RP}} \) probability is very high out to \( G \approx 16 \) at which point we start to run out of RVS sources so the uncertainties become large.

\[
\logit(q) = x = \log\left(\frac{q}{1-q}\right)
\]

\( x = \logit(q) \)

\[
\begin{array}{cccccc}
\text{G} = 11.5 & \text{G} = 14.5 & \text{G} = 17.5 & \text{G} = 20.5 \\
-6.9 (0.1\%) & -4.6 (1.0\%) & -2.2 (10.0\%) & 0.0 (50.0\%) & 2.2 (90.0\%) & 4.6 (99.0\%) & 6.9 (99.9\%)
\end{array}
\]

\[
\begin{array}{cccccc}
\text{Data} & \text{Model} & \text{Consistency} \\
0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 \\
\end{array}
\]

SELECT * FROM gaiaedr3.gaia_source WHERE dr2_rv_nb_transits>=4

AND phot_rp_n_obs>0
AND phot_g_mean_mag BETWEEN 1.6 AND 17.4
AND phot_g_mean_mag-phot_rp_mean_mag BETWEEN -9.6 AND 2.6

The RVS sample within the given colour-apparent magnitude range contains 6,186,950 out of a total 203,513,110 objects in the source catalogue within the same colour-magnitude range with published \( G_{\text{RP}} \).

For RVS, the distribution on the sky (shown in the right panel of Fig. 2) is no longer solely dominated by the Milky Way source distribution. Aspects of the Gaia scanning law and features of the IGSL are visible in the on-sky distribution.

4.4 Power Spectrum

The Gaussian Process prior requires a dispersion parameter \( \sigma \) which needs to be well chosen for the problem. The expected variance will depend on the Needle scale.

We start from the expected selection probability across the sky. \( E[q] = \frac{\alpha}{\alpha + \beta} = \frac{k + 1}{\ell + 2} \) (see Section 3.2). The distribution of logit \( \frac{k + 1}{\ell + 2} \) is plotted across the sky for each of the samples in Fig. 3. A spherical harmonic model can be directly estimated for the distribution from the HEALPix values using

\[
\hat{a}_{\ell m} \sim \frac{4\pi}{N_{\text{pix}}} \sum_{p=0}^{N_{\text{pix}}-1} Y_{\ell m}^*(l_p, b_p) x_p
\]  

(20)

where \( x_p \) are the pixel values and \( l_p, b_p \) are the Galactic longitude.
model which isn’t able to pick up structure on scales under bins of each sample which is caused by the lack of spatial resolution of the exceptions are slight excesses of bins with p-values of 0 and 1 in the faintest demonstrates that the model has fit the data exceptionally well. The minor magnitude bins are uniformly distributed at almost all magnitudes. This Figure 7.

harmonic coefficients. We model the power spectra with a single black lines in Fig. 5.

This is done for each sample with the power-spectra shown by the taking the square mean of mode amplitudes and latitude of the pixels. The power spectrum is then estimated by

\[
\hat{C}_\ell = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|. \tag{21}
\]

This is done for each sample with the power-spectra shown by the black lines in Fig. 5.

The power spectrum gives the expected variance of spherical harmonic coefficients. We model the power spectra with a single power-law distribution

\[
C_\ell = A(\ell + 1)^\gamma \tag{22}
\]

where \(\gamma\) describes how the power decays for smaller scales.

The sum of square spherical harmonic coefficients renormalised by their uncertainty is chi-square distributed with \(2\ell + 1\) degrees of freedom.

\[
\sum_{m=-\ell}^{\ell} \frac{|a_{\ell m}|^2}{\sigma_{\ell m}^2} \sim \chi^2(2\ell + 1) \tag{23}
\]

where \(\sigma_{\ell m}^2 = C_\ell\) is the expected coefficient variance. Therefore we can use this to derive the likelihood of the given power spectrum model

\[
P(\hat{C}_\ell | C_\ell) \propto \exp\left(\frac{-z^2}{2}\right) \tag{24}
\]

where \(z = \frac{(2\ell + 1)\hat{C}_\ell}{C_\ell}\). The 16th – 84th percentiles of this distribution provide the shaded regions in Fig. 5.

This is maximized with respect to \(A\) and \(\gamma\) in Eq. 22 to determine the best fit parameters using gradient descent with the Newton Conjugate Gradient method as implemented in scipy. We only use data with \(\ell < 135\) as this is the smallest scale spherical harmonic used in the Needlets and corresponds to a scale length on the sky \(\sim 1\) degree which is the approximate pixel size of the data we will be using with HEALPix nside = 64. The best fit power law profiles are shown by the red dashed lines in Fig. 5 for the Astrometry, RUWE (magnitude-only) and RVS samples respectively with parameter values given in Table 1.

Given the power spectrum of spherical harmonics, we want to work out what this implies for the variance of Needlet coefficients. From Appendix B of Boubert & Everall (2021), the variance of the Needlet coefficients as a function of the power spectrum is given by

\[
\langle \beta_{jk}^2 \rangle = A_{jk} \sum_{\ell=0}^{\ell_{\text{max}}} b_{jk}^2(j) C_\ell \frac{(2\ell + 1)}{4\pi} \tag{25}
\]

where \(C_\ell\) is taken from Eq. 22 using the best fit parameters and the normalisation constant is the area per pixel, \(A_{jk} = \frac{4\pi}{n_{\text{side}}^2}\).

The reader should be concerned that we have done something uncomfortably non-Bayesian. We’ve used the data to determine the appropriate prior for our model. There are two reasons why we can get away with this. Firstly, we’ve used the data aggregated over colour-apparent magnitude space and only used this to estimate two parameters of a simple power spectrum such that the vast amount of information is hidden from our prior model decision. Secondly, and more importantly, in this work we’re not attempting to infer a posterior distribution for our model. Using data to infer the prior would lead to an underestimation of posterior uncertainties as data has implicitly been double counted, however, we are only inferring the best fit selection function model.

The reason we’ve used the power spectra as described is that it makes the optimization significantly more computationally efficient as we’re using an informed start point.

5 RESULTS

When estimating each of the selection functions we used a Needlet model with \(j_{\text{max}} = 5\) and fit to data at HEALPix level 6 (nside = 64) resolution. This gives a total of 16,381 Needlets with \(\sim 2\) degree resolution fit to 49,152 pixels. Details about the samples,
binning schemes and model parameters are listed in Table 1. A detailed discussion of the motivation behind the bin sizes and model parameters is provided in Section 6.

In this Section, we’ll show the results of the fits as a function of sky position, apparent magnitude and colour. To test the veracity of the fits given the data, we’ll use the Binomial p-value test explained in Section 3.3.

### 5.1 Astrometry

The astrometry is fit in 0.2 mag bins with a magnitude scale length of 0.3 mag - long enough that neighbouring bins are correlated, but short enough that the data can produce sharp changes in the model.

The astrometry selection function across the sky is shown in the middle row of Fig. 6. The top row shows \((k+1)/(n+2)\) for the given magnitude bins, which, as discussed in Section 3.2, is the expected selection probability in the given magnitude bin independent of all others. High source density regions show much stronger selection limitations, particularly at the dimmer magnitudes. This is apparent through the white spots at \(G = 17.5\) which are centered on globular clusters such as \(\omega\) Centauri and dwarf galaxies such as the LMC and SMC.

The bottom row of Fig. 6 shows the results of the p-value test discussed in Section 3.3. At brighter magnitudes, p-values are distributed completely randomly demonstrating that the model has been successfully fit. Shifting to fainter magnitudes the selection function probability is reduced in regions of the sky with fewer scans. These ‘holes’ in the scanning law are prominent due to the selection cuts on \(N_{\text{rms}}\) and \(\sigma_{S}\) for the EDR3 astrometry sample.

Structure also appears particularly around \(\omega\) Centauri and the disk and LMC. This is where the selection function changes on scales smaller than the model can resolve. For example, the half-light radius of \(\omega\) Centauri is \(~4.8\) arc-minutes (van de Ven et al. 2006) however the Needlets can only resolve structure on 2 degree scales. Therefore the reduction in selection probability due to the globular cluster is spread out over a wider area by the Needlet. At the core, in the pixel which contains the globular cluster, the selection probability is overestimated whilst being underestimated in any neighbouring pixels. There is a further-out halo of overestimated probability due to the structure of the Needlets which go negative before returning to zero (see Fig. 1 Boubert & Everall 2021).

We provide a histogram of p-values in the top panel of Fig. 7 where we only include bins where \(n > 0\) as bins with \(n = 0\) have a uniformly distributed p-value independent of the model. The histogram for all bins is offset by a factor of 4 to make it clearer. For the vast majority of the data, we see well-behaved solutions with uniformly distributed p-values however, at the faint end, there are over-densities at \(p_{\text{value}} = 0.1\) where the resolution limitations become significant and we are under-fitting to the data.

The astrometry selection function as a function of magnitude is shown in the top panel of Fig. 8. We group level 6 HEALPix pixels by number of sources in the pixel in the Gaia source catalogue with \(1.6 < G < 22\). The selection function is the count-weighted mean of the selection functions in the given pixels. For \(G \leq 17\), the Astrometry sample is \(~99\%) complete however this drops off quickly in high source density regions. Low density regions stay close to complete out to \(G \sim 19\) before also falling rapidly. By \(G \sim 20\), less than 1% of sources from the Gaia source catalogue are included in the astrometry sample.

For \(G < 1.6\) we are not able to say anything informative about the selection function probability as there is no data in the source catalogue here to use. For \(G > 22\) the selection function should be taken as zero.

For the astrometry sample, we have shown that the selection function model reproduces the observed data down to Needlet scales of \(~2\) degrees. At high latitudes, the astrometry sample can be complete out to \(G \sim 19\) and significantly drops at \(G \sim 21\) due to the cut placed on DR2 apparent magnitude. However the selection

| Sample | Astrometry | RUWE < 1.4 (G only) | RUWE < 1.4 | RVS |
|--------|------------|---------------------|------------|-----|
| \(\sum n_i\) | 1,806,195,366 | 1,806,195,366 | 1,550,860,236 | 203,513,110 |
| \(\sum k_i\) | 1,465,211,050 | 1,400,803,102 | 1,292,152,210 | 6,186,950 |
| \(G\) range | [1.6,22] | [1.6,22] | [1.6,22] | [1.6,17.4] |
| \(G - G_{\text{HP}}\) range | - | - | [-1.7] | [-0.6,2.6] |

| Bins | HEALPix nside | 64 | 64 | 64 | 64 |
| \(G\) bins | 0.2 | 0.2 | 0.2 | 0.2 |
| \(G - G_{\text{HP}}\) bins | - | - | 2.0 | 0.4 |

| Power Spectrum | \(\log A\) | -0.2937 | -0.6206 | -0.2032 | 0.0937 |
| \(\gamma\) | -2.3985 | -2.3552 | -2.4845 | -2.2761 |

| Model | \(J_{\text{max}}\) | 5 | 5 | 5 | 5 |
| \(l_G\) | 0.3 | 0.3 | 0.3 | 0.6 |
| \(l_{G-G_{\text{HP}}}\) | - | - | 3.0 | 0.6 |
| \(B\) | 2.0 | 2.0 | 2.0 | 2.0 |

Table 1. For each Gaia sub-sample selection function, we provide the key parameters describing the data and model.
probability declines much brighter for crowded regions or where there are very few scans in EDR3.

5.2 RUWE

As discussed in Section 4, we evaluate the RUWE selection function in terms of sky-position and apparent magnitude against all sources in Gaia EDR3 and separately also as a function of $C = G - G_{\text{BP}}$ for sources where $G_{\text{BP}}$ is published.

Both RUWE selection functions are fit to $0.2\text{mag}$ bins in $G$ with a magnitude scale length of $0.3$. The colour dependent model includes four colour bins each 2mag wide with a scale length of $3\text{mag}$ which enables broad changes in the selection function probability as a function of $C$.

The results of the colour-independent model are shown in Fig. 9. At the faint end the behaviour is similar to the astrometry selection function and the binomial p-value also shows that the fits are struggling to resolve features in crowded regions at faint magnitudes. For $G = 17.5$ and brighter, however, the RUWE $< 1.4$ cut removes significant numbers of sources which have published astrometry. The selection probability is lower and for $G = 11.5$ the scanning law is highlighted as heavily scanned regions are more likely to be removed by the cut on RUWE.

The magnitude-dependent behaviour is shown more clearly in the middle panel of Fig. 8 where the RUWE $< 1.4$ selection probability actually peaks near the faint end before declining and shows the same crowding dependence as the astrometry selection. However, the bright end has significantly reduced selection probability particularly in non-crowded regions. Our interpretation is that astrometric measurements have lower variance for bright sources and therefore sources with genuine intrinsic noise (such as binary systems) would have a more significant excess noise and be removed by the RUWE $< 1.4$ cut. For $G \leq 3$ the selection probability picks up again which is likely an artefact of the prior as there is very little data in the source catalogue at these magnitudes.

The colour-dependent selection function for RUWE $< 1.4$ across the sky for four colour and apparent magnitude bins is given in Fig. 10. Because RUWE is a measure of the astrometric error above the expected uncertainty, areas with high RUWE aren’t necessarily those where Gaia performs worst but rather those where the satellite struggles unexpectedly. The selection function for $C = 0$ at the dim end shows a similar structure to the astrometry selection suggesting that the astrometry sample cuts are the limiting factor in this region of colour space. As we shift towards redder sources the picture changes significantly. For redder sources, the low RUWE sample is more complete in the galactic plane and significantly less so in un-crowded regions. This inversion appears to be more extreme at brighter magnitudes.

This structure bares some similarities to Fig. 7 of Everall et al. (2021c) where the along-scan astrometric error is typically higher in crowded regions at bright magnitudes. A possible explanation is that the Gaia attitude error model either overestimates measurement uncertainties for bright sources in Galactic plane or underestimates those at high latitudes. This would result in higher RUWE for high latitudes and explain the observed selection pattern.

Fig. 11 shows the p-values in four magnitude bins across the sky for the $C \in [1.0, 3.0]$ colour bin. As in the astrometry selection, the bright bins have very successfully modelled the data however at the faint magnitudes, spatial resolution becomes important in the Galactic plane. Once again we see a rippling effect on scales which the Needlets aren’t able to fully resolve. The second panel of Fig. 7 gives the distribution of all p-values from which we can see that the resolution issues are confined to dim magnitudes and overall the model still fits the data extremely well.

The selection function probability for the RUWE sample across the CMD is shown in Fig. 12. For the highly crowded bulge region (top panel) the selection probability peaks at $G \sim 18$ in the blue and $G \sim 11$ in the red before declining towards fainter magnitude particularly for redder sources before cutting off sharply at $G \sim 20.5$. In a low-density field (bottom), there is a stronger bimodality with selection function peaks at $G \sim 7$ and $G \sim 20$. This structure is likely connected to the different types of observations taken onboard Gaia with 1D windows for $G > 13$ sources and 2D windows $G < 13$. 

Figure 8. The selection function probability is strongly dependent on source density in all samples. By splitting pixels according to the source density in the full Gaia sample, we evaluate the selection function probability using the count-weighted average of all pixels within the given density bin. For the astrometry sample (top) the selection function drops at brighter magnitudes for higher source density regions. The RUWE $< 1.4$ colour-independent selection function (middle panel), shows a slightly different behaviour as much brighter sources are more likely to be cut out due to excess noise however the faint end shows similar behaviour to the astrometry. The RVS sample (bottom panel) at $G - G_{\text{RP}} = 0.5$ shows a more complicated pattern. At the bright end, as expected, the selection function is lower in high source density regions. However this flips at $G \sim 13$. This is because RVS is limited in $G_{\text{RVS}} - G_{\text{RP}}$ and hence is more complete for redder sources as shown by Fig. 12 which will largely be bulge fields due to dust extinction. In all samples, the model reverts to the prior mean $(x = 0)$ for $G \leq 3$ due to a lack of data in the source catalogue.
with window gating happening at still brighter magnitudes (Evans et al. 2018).

The overall trends with apparent magnitude for bluer sources are similar to the results from the colour-independent model shown in the middle panel of Fig 8. The cause of the drop-off in selection probability for faint red sources is unknown but may be related to chromaticity of the astrometric solution. An alternative explanation is that redder sources have lower $G_{RP}$ and as such are more likely to have a measured $G_{RP}$ in the source catalogue. This could push the selection function down at these colours. In this region of parameter space, one should make sure that the $G_{RP}$ selection function is used in conjunction with RUWE at the faint end of Fig. 12. For the bulge line of sight, the selection probability also drops off at much brighter magnitudes due to the faint end of Fig. 12. For the bulge line of sight, the selection probability also drops off at much brighter magnitudes due to the faint end.

5.3 RVS

The RVS sample is binned in 0.2 mag bins in colour and 0.4 mag bins in $G - G_{RP}$. We use a scale length of 0.6 in apparent magnitude and 1.2 in colour. We opt for higher colour resolution than we did for RUWE because RVS selection will be closely related to $G_{RP}$ which will generate strong $C = G - G_{RP}$ dependence in the selection function.

The radial velocity sample is dependent on CCD observations from the on-board spectrograph which occupies four out of seven CCD rows on the Gaia focal plane. Depending on the $G_{RVS}$ apparent magnitude evaluated for the source, a variety of window classes will be used as described in Section 7.1 of Cropper et al. (2018). Sources with $G_{RVS} < 7$ (Class 0) received a full 2D windows whilst fainter sources only received 1D windows (Class 1 or 2). In crowded regions, Class 1 and 2 sources with overlapping windows would have their windows truncated in the region of overlap, often leading to non-rectangular windows. These observations were not used in DR2 data processing (as explained in Sartoretto et al. 2018) leading to 40% of spectra being removed and a much higher percentage in crowded regions.

This process is manifested in the RVS selection function. The third panel in Fig. 8 shows that the RVS selection function for $G \leq 8$ (corresponding to approximately $G_{RVS} \leq 7$) is source density independent whereas sources in crowded regions are much less likely to be selected with $8 \leq G \leq 12.5$.

For $G \geq 12.5$, this behaviour changes significantly. At dimmer and redder magnitudes, the RVS population traces the distribution of dust with high completeness in regions of the sky with significant extinction. We don’t know the precise cause of this however there are some plausible explanations. The RVS catalogue filters out stars cooler than $T_{eff} = 3550$K which corresponds to $G - G_{RP} \geq 1.2$ (see Fig. 3 Andrae et al. 2018). However, in regions with high dust extinction we will have hotter stars appearing with higher $G - G_{RP}$ due to reddening. Therefore, for redder colours, the RVS cut on $T_{eff}$ will be more strict where there is less extinction and produce a higher selection probability in dusty regions of the sky.

As discussed earlier, the RVS sample is selected on proxies for the apparent magnitude in the RVS waveband, $G_{RVS}$. This is quite close to the $G_{RP}$ waveband. A sharp cut in $G_{RP}$ will lead to a correlated cut between $G$ and $G - G_{RP}$. This is exactly what we see at the faint end of Fig. 12. For the bulge line of sight, the selection probability also drops off at much brighter magnitudes due to the
crowding limits of the spectrograph. The narrow range of colour with non-zero selection probability is reflective of the RVS cuts on source effective temperature. The effect of the $G_{\text{RVS}}$ cut is most distinctive in the third row, second column of Fig. 13 where the complex structure of the IGSL is very prominent.

The p-value test, shown across the sky in Fig. 14 for $G - G_{\text{RP}} = 0.8$, demonstrates that the model is correctly representing the data across most magnitudes. The third panel of Fig. 7 also shows this with only one magnitude bin showing any poorly fit pixels with $G = 14$. This may be the model struggling to reproduce sharp changes in the behaviour of the IGSL.

As we saw for the RUWE model, the RVS selection function shoots up at $G \lesssim 3$ which can be seen in both Fig. 8 and Fig. 12 due to a lack of objects in the source catalogue. Users should be wary of this when applying the selection functions at these magnitudes.

### 5.4 RVS and Astrometry or RUWE

As well as individual subsets, various science cases will also require the intersection of subsets. For example, if one wanted full 6D phase space information for all objects in their sample, they’d want the...
Figure 12. The RUWE selection function (left) is highest at faint blue magnitudes ($G_{18}$, $G_{20}$, $G_{RP}$) and drops significantly at both the brightest and faintest ends for the crowded bulge field (top) and uncrowded Galactic pole field (bottom). There is also a strong gradient to lower selection probability for redder sources. This may be related to chromaticity of astrometric fits or relatively high completeness of $G_{RP}$ at these colours. RVS has a narrow range of colours with significant selection probability which is likely the result of the $T_{eff}$ selection used in the RVS pipeline (Sartoretti et al. 2018). At the faint end, the probability rapidly reduces but correlated with colour such that redder sources are observed to fainter magnitudes. This is due to the fact that RVS observes in the red part of the Gaia waveband and so the true RVS limit is closer to a $G_{RP}$ cut. This produces the nice diagonal cut in $G$ vs $G - G_{RP}$ corresponding to $G_{RP} - 12.5$. 

The RUWE sample only contains sources with parallax and proper motion. Therefore $P(S_{RUWE} | S_{source}, y) = 1$ and the selection function for Astrometry and RUWE is equal to the RUWE selection function as expected.

If subset 1 and subset 2 are independent

$$P(S_{subset1} \mid S_{subset2}, S_{source}, y) = P(S_{subset1} \mid S_{source}, y)$$  \hspace{1cm} (27)$$

and we can take the selection function for the intersection to be the product of the two individual selection functions.

We do this for the RVS sample where parallax and proper motion are provided and where RUWE $< 1.4$. There are 6,162,273 in Gaia EDR3 with both RVS and parallax and proper motion where 5,303,693 of these also have RUWE $< 1.4$.

We take the product of the RVS position-colour-magnitude selection function defined in the previous section with the colour-independent selection functions for Astrometry

$$P(S_{RVS}, S_{Astrometry} \mid S_{Gaia}, y) = P(S_{RVS} \mid S_{Gaia}, y) \cdot P(S_{Astrometry} \mid S_{Gaia}, y)$$  \hspace{1cm} (28)$$
Gaia Astrometry and RVS Selection Functions

Figure 13. At brighter magnitudes (top rows) and bluer colours (left columns), the RVS selection probability is significantly higher in regions of the sky with more scans and lower source density. As we move dimmer and redder, this shifts to a higher selection probability in the Milky Way disk where dust extinction causes sources to be systematically redder and therefore have be relatively brighter in $G_{\text{RP}}$ (see Fig. 12).

Figure 14. The $p$-value test for RVS bins, shown here for $G - G_{\text{RP}} \in [0.6, 1]$, is dominated by random noise with some small scale structure just visible at $G = 13.5$. This shows that the model has provided an extremely good fit to the data.

and RUWE

\[
P(S_{\text{RVS}}, S_{\text{RUWE}} \mid S_{\text{Gaia}}, y) = P(S_{\text{RVS}} \mid S_{\text{Gaia}}, y) \cdot P(S_{\text{RUWE}} \mid S_{\text{Gaia}}, y).
\]

We then compare this with the data. Fig. 15 shows the distribution of $p$-values for this sample where we only include bins with $n > 0$ as bins with $n = 0$ give a uniformly distributed $p$-value independent of the model. There is on significant sign of underfitting so we conclude that it is reasonable in this case to assume that RVS and Astrometry or RUWE samples are independent and therefore use the product of probabilities.

6 DISCUSSION

For each selection function we have made decisions about the variables to use, binning schemes to aggregate the data and the values of free model hyperparameters. In this section we justify the choices made.

6.1 Selection function variables

For our selection variables we decided to use $l, b, G$ and $G - G_{\text{RP}}$ where appropriate. As discussed in Section 2, the cuts used to generate the given $\text{Gaia}$ subsets are primarily dependent on these variables. However, we have found that all selection functions are
strongly dependent on source density in the Gaia source catalogue. Whilst this is indirectly modelled through the spatial variables, directly using a crowding variable could enable us to better model the selection function in small dense objects such as globular clusters.

We could not use crowding as a selection variable in this work because it is not a measurable. The number density of sources in the Gaia source catalogue provides some insight but this is not the true underlying density of sources as Gaia is not a complete catalogue. The detection probability of a source in Gaia is also likely to primarily depend on the number density of sources with the same apparent magnitude or brighter such that a source density variable would also need to be magnitude dependent. As a further complication, we should really be modelling the source density on the focal plane. This includes sources from both fields of view separated by a basic angle of 106.5 deg such that the source density is also a function of observation time mapped through the scanning law. Constructing selection functions as a function of crowding is a complicated problem worth solving for Gaia subsets but beyond the scope of this paper.

This means that users should be wary of applying our selection functions to synthesised catalogues in order to generate mock Gaia data. The spatial component of the selection function includes strong crowding dependence which reflect features of the Milky Way such as the bulge, LMC and globular clusters but which will be different for any other simulated galaxy.

6.2 Binning schemes

In principle we could evaluate the likelihood for every source at its exact coordinates. With 1.8 billion sources, just evaluating the spatial model at all source positions with \( j_{\text{max}} = 5 \) would involve \( \sim 10^{13} \) function evaluations. And this would need to be done for every iteration of the optimization. Instead we aggregate the data in pixels and colour-magnitude bins to make the problem computationally tractable. Our choices of bin scales for each selection function are shown in Table 1.

We have used HEALPix level 6 (inside = 64) such that the data is resolved on sub-degree scales. The selection function can change on smaller scales than this, particularly in globular clusters, however, it is computationally prohibitive to resolve the cores of globular clusters many of which have scale radii of arcminutes or smaller (Vasiliev & Baumgardt 2021).

The widths of magnitude and colour bins are decided based on our expectation of how fast we believe the selection function will vary in these parameters. The selection function is heavily magnitude dependent so we resolve in magnitude as high as we can within computational limitations. The colour dependence of RUWE is due to the rescaling of the unit weight error to account for the chromaticity of the astrometry. Therefore we only include the wide 2 mag bins in colour to encapsulate any general trends. RVS has a more significant colour dependence due to cuts placed on effective temperature so we use narrower 0.4 mag bins for this.

6.3 Model parameters

The maximum level of needlets used, \( j_{\text{max}} \) is set to one lower than the level of the data binning. We cannot use the same level or higher as this would lead to more model parameters than data points. This level of needlets enables us to fit a model which varies on 2 deg scales across the sky.

\( B \) and \( \nu \) determine the shape and width of the needlets. The relation between these parameters and the needlet shape is complicated but by setting \( B = 2, \nu = 1 \) the width of the needlets is similar to the distance between needlet centers at a given HEALPix level. This enables us to fit smooth models across the sky without significant inter-needlet fluctuations.

The magnitude and colour lengthscales of the Gaussian Process, \( l_G \) and \( l_{G-G_{\text{BP}}} \), are set to 1.5x the bin widths in almost all cases. This allows for correlations between neighbouring data points and results in a smooth model as a function of colour and apparent magnitude. The one exception is \( l_G \) for RVS which we set to 0.6 to match \( l_{G-G_{\text{BP}}} \). As discussed in Section 4, the RVS sample is selected on \( G_{\text{RVS}} \) which is a narrow band in the same magnitude range as \( G_{\text{BP}} \). Any features at fixed \( G_{\text{RVS}} \) will produce a correlated feature in \( G \) vs \( G-G_{\text{BP}} \) as we saw in Fig. 12. We set the scale lengths such that the model has equal power to resolve in magnitude and colour.

The mean and variance of the Gaussian Process are set such that the prior is only significant when there is very little data to fit. This occurs for \( G < 3 \) at which point the model reverts to the mean as is seen in Fig 8. When using the selection function, users concerned about extremely bright stars should consider truncating their models at \( G = 3 \) as the selection functions are not constrained by data brighter than this.
6.4 Magnitude and colour bounds

The selection functions are only defined within apparent magnitude and colour ranges given in Table 1 so a choice needs to be made about how to extrapolate the selection functions.

In all cases there are sources in the EDR3 source catalogue beyond the faint limit but none in the subset. The selection probability should be treated as zero fainter than this. As discussed in the previous subsection, any models using our selection function should be truncated for $G < 3$ where the prior dominates the model.

We have no a priori knowledge about the exact colour dependence of the RUWE selection function. We recommend that bluer or redder than the colour range we have modelled users should take the value of the $G - G_{\text{RP}} = [-1, 1], [5, 7]$ bins respectively.

The RVS sample drops for blue and red sources due to cuts on effective temperature so the selection function should be taken as zero for $G - G_{\text{RP}} < -0.6$ or $> 2.6$.

7 OTHER SAMPLES

In this paper we have produced selection functions for three scientifically important sub-samples of the Gaia mission. However there are many catalogues in the Gaia-verse which are used by the astronomy community for a wide range of exciting research. Here we list some which were considered but which, due to additional complexity, we were not able to evaluate with the current method.

7.1 Variables

Variable sources are incredibly important as standard candles allowing us to measure precise distances to sources and systems across the Milky Way and in external galaxies (Muraveva et al. 2018; Riess et al. 2019). This has enabled detailed maps of the distribution of stars throughout the Milky Way disk and halo (Skowron et al. 2019; Iorio & Belokurov 2019).

Holl et al. (2018) describes the Gaia DR2 variable sample following the processing detailed in Eyer et al. (2017). As discussed in Holl et al. (2018), overall completeness was not an aim of the DR2 variables sample. Considering only the lower limits placed on number of successful FoV transits required for variable processing they estimate that the average completeness is $\sim$ 80% for the FoV $\geq 12$ pipeline and as low as $\sim$ 51% for FoV $\geq 20$. This varies heavily across the sky due to the scanning law and crowding effects. Therefore a selection function for this sample would be hugely valuable for models of the Milky Way.

There are several reasons why the method we have applied here would not be directly applicable to variable samples. Firstly we need to consider what question we’re asking. We want to know e.g. “Given that a source with $l, b, G$ is an RRLyrae, what is the probability that Gaia would publish this source classified as an RRLyrae?”.

Written as a probability, this is

$$P(\text{SRRLyrae} \mid l, b, G, \text{X}_{\text{SRRLyrae}})$$

where $\text{X}_{\text{SRRLyrae}}$ is the event that the source is an RRLyrae-type star. Were we to naively use the approach we’ve used in this paper for other samples, we would actually evaluate the probability that a source is selected and it is an RRLyrae

$$P(\text{SRRLyrae}, \text{X}_{\text{SRRLyrae}} \mid l, b, G).$$

This is because we’d be comparing the population of classified RRLyraes against the full sample of Gaia sources the vast majority of which are not RRLyrae stars.

To model the selection function for variable samples, we would need to know which sources in Gaia would have been classified as a variable had they actually been one.

The second issue is in choosing the observables over which to define the selection function. The selection function for variable sources will be highly dependent on variability amplitude and period. Gaia observes stars with certain dominant frequencies due to the spin, precession and orbital dynamics of the satellite. This would make some variability periods harder to detect than others. The apparent $G$-band magnitude is also highly unreliable for variable sources (see Section 5.4 Arenou et al. 2018). Improved apparent magnitudes may be determined from the time series however this adds another layer of complexity when comparing against the full Gaia dataset.

Finally, there are also many individual variable samples, such as RRLyrae, Cepheids and Long Period Variables, which are each used for separate science aims. A full study of the selection functions of variable sources in Gaia should aim to construct selection functions for each population.

The variables selection function for the Gaia mission is a significant, complex project worthy of dedicated further study.

7.2 Parallax SNR

$1/\sigma$ is often used as a distance estimate for sources with Gaia astrometry. However, for sources with low parallax signal-to-noise (SNR) the distance uncertainty distribution can be highly asymmetric and the measured parallax is negative for many sources in Gaia EDR3 (see Bailer-Jones 2015; Luri et al. 2018, for full discussion).

A common selection to avoid this is taking only sources with $\sigma/\sigma_{\text{err}} > X$ where commonly used values for $X$ are 5 or 10 depending on the scientific aims. This choice significantly affects the selection function as a function of distance and as a result, the inferred distance to sources can be heavily biased (see Schönrich & Aumer 2017, for a detailed discussion).

One solution to this was demonstrated by Schönrich et al. (2019) who infer a distance-dependent selection function for the Gaia DR2 RVS sample. The limitations of the method used here is that they must assume an a-priori model for the distribution of sources in the Milky Way. As a result, their selection function is not model independent. Their selection function is also sky-averaged for the entire sample and doesn’t look to quantify any of the complex variation of the selection function on small spatial scales.

We could use the method from Boubert & Everall (2021) directly on the $\sigma/\sigma_{\text{err}} > 5$ sample as a subset of the Gaia catalogue and evaluate the selection function as a function of $G, G - G_{\text{RP}}, l, b$. Indeed, $\sigma_{\text{err}}$ is strong function of position on the sky due to the scanning law and a function of $G$ due to CCD photon noise (Everall et al. 2021c).

However the selection is also directly dependent on $\sigma$ so any selection function for a parallax SNR selected sample would need to feature measured $\sigma$ or observables which are directly dependent on it. For this reason, we do not attempt to model the parallax SNR selection function here.

8 ACCESS AND USE THE SELECTION FUNCTIONS

The selection functions are accessible through the GitHub repository selectionfunctions (https://github.com/gaiaverse/selectionfunctions). Here we provide a brief example of how to load the EDR3 source catalogue and RVS subsample selection
functions to estimate the probability that S5-HVS1 (Koposov et al. 2020) could have been observed in *Gaia* EDR3 with radial velocity measure from DR2. The inferred probability is \( \sim 3 \times 10^{-15} \) and indeed S5-HVS1 does not have published radial velocity.

```python
import selectionfunctions.cog_v as CoGV
import selectionfunctions.cog_ii as CoGII
from selectionfunctions.source import Source

# Load DR3 selection function
dr3_sf = CoGII.dr3_sf(version='modelAB', crowding=False)

# Load RVS selection function
rvs_sf = CoGV.subset_sf(map_fname='rvs_cogv.h5')

# Introduce source
s5_hvsl = Source('
    22h54m51.68s,
    -51d11m44.19s',
    photometry={'gaia_g':16.92,
                'gaia_bp_gaia_rp':-0.008},
    frame='icrs')

# Selection probability
print(dr3_sf(s5_hvsl) * rvs_sf(s5_hvsl))
```

There are many ways the selection functions we have produced may be used by the community. The number density of *Gaia* sources in any coordinate system may be directly corrected to infer the true underlying distribution as demonstrated in Rix et al. (2021) for white dwarfs as a function of colour and absolute magnitude. One can take a forward modelling approach applying the selection function to a model and comparing with the published *Gaia* data then updating the model parameters to find the best fit. This will be the subject of two upcoming papers Everall et al. (2021a,b).

The selection functions can also be used to generate realistic mock *Gaia* catalogues from simulations similar to projects such as Rybizki et al. (2020) and Aurigaia (Grand et al. 2018). However, as discussed in Section 6 features of the Milky Way are baked in due to crowding dependence of *Gaia*’s source sensitivity so applying these selection functions to simulated galaxies will not perfectly reproduce the samples *Gaia* would record if it were observing the simulated galaxy.

### 9 CONCLUSIONS

We have evaluated the selection functions for three subsamples of the *Gaia* EDR3 data release: (i) Sources with published parallax and proper motion, (ii) the subset with RUWE < 1.4 and (iii) sources with DR2 radial velocities published with EDR3.

All selection functions are evaluated as a function of position on the sky using a spherical needlelets model. The astrometry and RUWE selection functions are modelled as a function of *G*-band apparent magnitude and we construct selection functions for RUWE and RVS as a function of *G* and *G* − *G*$_{RP}$ colour with Gaussian Process priors.

In all cases there is a strong dependence on crowding and features of the *Gaia* scanning law. For RVS, the effects of temperature cuts on the sample and selection on *G*$_{RVS}$ using IGS magnitudes heavily impacts the selection function.

In all samples the model is able to reproduce the data down to scales ~ 2 degrees. The only areas in which the model breaks down are where the selection probability varies on smaller scales and our spatial model doesn’t have the flexibility to reproduce the data. This is an issue in the Milky Way bulge and around globular clusters.

Much of the spatial structure and magnitude and colour dependence of the selection functions is complicated, detailed and non-intuitive. This demonstrates the importance of empirically evaluating the selection functions for subsets of *Gaia* data.

We provide a first-order estimate of the *Gaia* EDR3 source catalogue selection function to be used with the sub-sample selection functions we have developed. All of these are accessible through the gaiaverse/selectionfunctions GitHub repository.5

### ACKNOWLEDGEMENTS

AE thanks the Science and Technology Facilities Council of the United Kingdom for financial support. DB thanks Magdalen College for his fellowship and the Rudolf Peierls Centre for Theoretical Physics for providing office space and travel funds. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

AE is grateful to the *Gaia* helpdesk and in particular Lennart Lindegren and Paola Sartoretti for helping to explain key features of the selection functions. The authors also thank the anonymous referee for providing detailed comments which enhanced our descriptions of the selection functions.

### DATA AVAILABILITY

The data underlying this article are publicly available from the European Space Agency’s *Gaia* archive (https://gea.esac.esa.int/archive/). Open-source implementations of the selection functions calculated in this work are available through the Python package selectionfunctions (https://github.com/gaiaverse/selectionfunctions). The data products needed to evaluate the selection functions are publicly available on the Harvard Dataverse (https://doi.org/10.7910/DVN/TGI69).

### REFERENCES

Andrae R., et al., 2018, A&A, 616, A8
Arenou F., et al., 2018, A&A, 616, A17
Bailer-Jones C. A. L., 2015, PASP, 127, 994
Baldi P., Kerkyacharian G., Marinucci D., Picard D., 2006, arXiv Mathematics e-prints, p. math/0606154
Belokurov V., et al., 2020, MNRAS, 496, 1922
Boubert D., Everall A., 2020, MNRAS, 497, 4246
Boubert D., Everall A., 2021, Accepted in MNRAS
Boubert D., et al., 2019, MNRAS, 486, 2618
Boubert D., Everall A., Holl B., 2020, MNRAS, 497, 1826
Boubert D., Fraser J., Everall A., 2021a, in prep
Boubert D., Everall A., Fraser J., Gratton A., Holl B., 2021b, MNRAS, 501, 2954
Bovy J., Rix H.-W., Liu C., Hogg D. W., Beers T. C., Lee Y. S., 2012, ApJ, 753, 148
Cropper M., et al., 2018, A&A, 616, A5
Evans D. W., et al., 2018, A&A, 616, A4
Everall A., Das P., 2020, MNRAS, 493, 2042

5 https://github.com/gaiaverse/selectionfunctions
APPENDIX A: BINOMIAL ONE-TAILED P-VALUE

The one-tailed p-value test provides the probability that the observations would be smaller (in some sense) than the actual measured data, $d$ given the hypothesised model.

For a hypothesised model with parameters $\psi$ and measured data $d$, the likelihood of the data given the parameters is $P(d \mid \psi)$. The p-value is given by the integral over this (i.e. the CDF)

$$p_{\text{value}} = \frac{\int_{d_{\text{min}}}^d d' \, P(d' \mid \psi)}{\int_{d_{\text{min}}}^{\infty} d' \, P(d' \mid \psi)}$$

where $d_{\text{min}}$ is the minimum value the data can take under the model.

Let’s say we have a sample of $k$ marbles drawn from a bag of $n$. We hypothesise that the probability of any marble being selected is Bernoulli (i.e. randomly) distributed with probability $q$. Therefore our likelihood is Binomial($k \mid n, q$). The p-value is given by

$$p_{\text{value}} = \frac{\int_{k_{\text{min}}}^k d' \, \text{Binomial}(d' \mid n, q)}{\int_{k_{\text{min}}}^{\infty} d' \, \text{Binomial}(d' \mid n, q)}.$$  

However, the CDF at $k' = k$ discontinuously jumps. This is shown by the example Binomial CDF in Fig. A1 where we have used $n = 5$, $q = 0.5$ to demonstrate. The CDF at $k = 3$ has a discontinuity and uniformly covers a range of p-values in the CDF. Therefore we can write the p-value as

$$p_{\text{value}} = U \left[ \frac{\sum_{k' = 0}^{k-1} \text{Binomial}(k' \mid n, q)}{\sum_{k' = 0}^{\infty} \text{Binomial}(k' \mid n, q)} \right].$$

This is shown by the red shaded region in Fig. A1 as the p-value for $k = 3$ given $n = 5$, $q = 0.5$ is $p_{\text{value}} \sim U[0.50, 0.81]$. Without any data ($n = 0$) the p-value becomes uniformly distributed

$$p_{\text{value}} = U \left[ \frac{\sum_{k' = 0}^{\infty} \text{Binomial}(0 \mid 0, q)}{\sum_{k' = 0}^{\infty} \text{Binomial}(0 \mid 0, q)} \right] = U[0, 1].$$
This paper has focused on evaluating selection functions for subsamples of Gaia EDR3. However, these are only useful in conjunction with the Gaia source catalogue as described in Section 2.

In Boubert & Everall (2020), the Gaia DR2 source catalogue selection function was evaluated. The method involved evaluating the observation efficiency, the probability that a source transit would be used in the astrometric solution, for every source as a function of apparent magnitude and on-sky source density. The selection function was then the probability that a source received at least five successful transits.

The method required precise predictions of the number of occasions on which the source could have been detected from the scanning law. The scanning law was calibrated using the Gaia DR2 epoch photometry (Boubert et al. 2020, 2021b). However epoch photometry for the EDR3 baseline won’t be published until the full data release in 2022 which prevents us from calibrating the EDR3 scanning law and repeating the method previously used for DR2.

Work is currently underway to produce a significantly more impressive selection function for the EDR3 source catalogue (Fraser et al. 2021; Boubert et al. 2021a). For now, we provide a simple estimate for the EDR3 selection function based on Boubert & Everall (2020) which can be used with the results of this paper but emphasise that this will be superseded when the new results are published.

The Gaia EDR3 source catalogue contains about 7 per cent more sources than Gaia DR2. Fig B1 shows the ratio of source density across the sky between EDR3 and DR2. A striking feature of this is that the increase in source density follows patterns of the scanning law. Regions of the sky which were under-scanned in DR2 but received more transits in the extra 14 months of data to EDR3 have the biggest increase in content.

Motivated by this, we propose a very simple adjustment to evaluate the EDR3 selection function. We assume that the detection efficiency Gaia EDR3 is the same as in DR2. Therefore we use the results from Fig. 7 of Boubert & Everall (2020) to evaluate the detection efficiency as a function of apparent magnitude and source density. The number of times a source was transited is evaluated from the uncalibrated Gaia EDR3 nominal scanning law. Therefore the selection function is (Eq. C1 Boubert & Everall 2020)

$$P(S_{source} | G, I, B) = 1 - \sum_{m=0}^{4} \frac{\beta(A + m, B + n - m)}{\beta(A, B)}$$

where A, B are the DR2 Beta distribution parameters of the efficiency as a function of source density and apparent magnitude. n is the number of times a source was transited in EDR3 according to the EDR3 scanning law.

The numbers of transits in DR2 and EDR3 from (Boubert et al. 2021b) and the EDR3 nominal scanning law respectively are shown in the top panels of Fig. B2. In the bottom panels we show the selection function probability at $G = 21$. The selection probability has increased across the majority of the sky but most significantly in the ecliptic plane regions where there were few transits in DR2 such as the West side of the Galactic bulge.

This result provides a first estimate of the Gaia EDR3 selection function but should be used with caution. There are three limitations to this model.

Firstly, we have assumed that the pipeline of DR2 and EDR3 will lead to the same source detection efficiency in both catalogues. With each new data release, the data processing pipeline is rerun with improved source calibration which can change whether individual source observations are used in the astrometric solution.

Secondly, in Gaia DR2, when two sources were observed to be within 0.4 arcseconds of one another, the lower priority source was removed from the source catalogue. In EDR3, this threshold was reduced to 0.18 arcseconds (Gaia Collaboration et al. 2021a; Torra et al. 2021) so EDR3 will be more complete in crowded regions independent of the scanning law.

Finally, the EDR3 nominal scanning law has not been calibrated directly against the data and will deviate from Gaia’s true scanning history by up to ~ 30 arcseconds at any point in time. Therefore the predicted observations in this model will be marginally off the true numbers over small regions of sky.

Details on how to access the Gaia EDR3 source catalogue selection function are provided in Section 8.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Figure B1. The increase of sources in the Gaia source catalogue between EDR3 and DR2 is driven by new scans. In the top panel we show the ratio of source density in HEALPix pixels between EDR3 and DR2 showing clear structure of the scanning law. The bottom panel is the approximate ratio of number of scans between the data releases using the calibrated DR2 scanning law and nominal scanning law for EDR3. Regions which received few observations in DR2 with many more in EDR3 saw large increase in number of objects in the Gaia source catalogue.

APPENDIX B: Gaia EDR3 SOURCE CATALOGUE

This paper has focused on evaluating selection functions for subsamples of Gaia EDR3. However, these are only useful in conjunction with the Gaia source catalogue as described in Section 2.

In Boubert & Everall (2020), the Gaia DR2 source catalogue selection function was evaluated. The method involved evaluating the observation efficiency, the probability that a source transit would be used in the astrometric solution, for every source as a function of apparent magnitude and on-sky source density. The selection function was then the probability that a source received at least five successful transits.

The method required precise predictions of the number of occasions on which the source could have been detected from the scanning law. The scanning law was calibrated using the Gaia DR2 epoch photometry (Boubert et al. 2020, 2021b). However epoch photometry for the EDR3 baseline won’t be published until the full data release in 2022 which prevents us from calibrating the EDR3 scanning law and repeating the method previously used for DR2.

Work is currently underway to produce a significantly more impressive selection function for the EDR3 source catalogue (Fraser et al. 2021; Boubert et al. 2021a). For now, we provide a simple estimate for the EDR3 selection function based on Boubert & Everall (2020) which can be used with the results of this paper but emphasise that this will be superseded when the new results are published.

The Gaia EDR3 source catalogue contains about 7 per cent more sources than Gaia DR2. Fig B1 shows the ratio of source density across the sky between EDR3 and DR2. A striking feature of this is that the increase in source density follows patterns of the scanning law. Regions of the sky which were under-scanned in DR2 but received more transits in the extra 14 months of data to EDR3 have the biggest increase in content.

Motivated by this, we propose a very simple adjustment to evaluate the EDR3 selection function. We assume that the detection efficiency Gaia EDR3 is the same as in DR2. Therefore we use the results from Fig. 7 of Boubert & Everall (2020) to evaluate the detection efficiency as a function of apparent magnitude and source density. The number of times a source was transited is evaluated from the uncalibrated Gaia EDR3 nominal scanning law. Therefore the selection function is (Eq. C1 Boubert & Everall 2020)

$$P(S_{source} | G, I, B) = 1 - \sum_{m=0}^{4} \frac{\beta(A + m, B + n - m)}{\beta(A, B)}$$

where A, B are the DR2 Beta distribution parameters of the efficiency as a function of source density and apparent magnitude. n is the number of times a source was transited in EDR3 according to the EDR3 scanning law.

The numbers of transits in DR2 and EDR3 from (Boubert et al. 2021b) and the EDR3 nominal scanning law respectively are shown in the top panels of Fig. B2. In the bottom panels we show the selection function probability at $G = 21$. The selection probability has increased across the majority of the sky but most significantly in the ecliptic plane regions where there were few transits in DR2 such as the West side of the Galactic bulge.

This result provides a first estimate of the Gaia EDR3 selection function but should be used with caution. There are three limitations to this model.

Firstly, we have assumed that the pipeline of DR2 and EDR3 will lead to the same source detection efficiency in both catalogues. With each new data release, the data processing pipeline is rerun with improved source calibration which can change whether individual source observations are used in the astrometric solution.

Secondly, in Gaia DR2, when two sources were observed to be within 0.4 arcseconds of one another, the lower priority source was removed from the source catalogue. In EDR3, this threshold was reduced to 0.18 arcseconds (Gaia Collaboration et al. 2021a; Torra et al. 2021) so EDR3 will be more complete in crowded regions independent of the scanning law.

Finally, the EDR3 nominal scanning law has not been calibrated directly against the data and will deviate from Gaia’s true scanning history by up to ~ 30 arcseconds at any point in time. Therefore the predicted observations in this model will be marginally off the true numbers over small regions of sky.

Details on how to access the Gaia EDR3 source catalogue selection function are provided in Section 8.

This paper has been typeset from a TEX/LATEX file prepared by the author.
Figure B2. A simple estimate of the *Gaia* EDR3 selection function is achieved by updating the DR2 selection function with the nominal scanning law for EDR3. In the top row we show the number of scans received as a function of position on the sky in Galactic coordinates for DR2 (left) and EDR3 (right). The inferred *Gaia* source catalogue selection function probability at $G = 21.0$ is shown in the bottom panel where regions which have received more scans in EDR3 have significantly higher selection probability.