Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **544,** 3710–3728 (2025) https://doi.org/10.1093/mnras/staf1810
Advance Access publication 2025 October 21

The XXL survey: LV. Galaxy cluster classification from the XXL X-ray source catalogue using a Gaussian process binary classifier trained on imperfectly labelled data

J. Cale Baguley,[1]★ M. N. Bremer,[1] Ben J. Maughan,[1] S. Bhargava,[2] C. Garrel,[2,3] E. Koulouridis,[4] M. Pierre,[2] C. Adami,[5] L. Chiappetti,[6] D. Eckert,[7] C. H. Ek,[8] L. Faccioli,[2] F. Gastaldello,[6] M. Oguri,[9,10] N. Okabe,[11] F. Pacaud,[12] S. Paltani[7] and T. Sadibekova[2]

[1]*Astrophysics Group, School of Physics, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK*
[2]*Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, F-91191 Gif-sur-Yvette, France*
[3]*Max Planck Institute for Extraterrestrial Physics (MPE), Giessenbachstrasse 1, D-85748 Garching bei München, Germany*
[4]*Institute for Astronomy & Astrophysics, Space Applications & Remote Sensing, National Observatory of Athens, GR-15236 Palaia Penteli, Greece*
[5]*Aix Marseille Université, CNRS, CNES, LAM, F-13388 Marseille CEDEX 13, France*
[6]*INAF, IASF Milano, via Corti 12, I-20133 Milano, Italy*
[7]*Department of Astronomy, University of Geneva, Ch. d'Écogia 16, CH-1290 Versoix, Switzerland*
[8]*Department of Computer Science and Technology, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK*
[9]*Center for Frontier Science, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan*
[10]*Department of Physics, Graduate School of Science, Chiba University, 1-33 Yayoi-Cho, Inage-Ku, Chiba 263-8522, Japan*
[11]*Graduate School of Advanced Science and Engineering, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan*
[12]*Argelander Institut für Astronomie, Universität Bonn, D-53121 Bonn, Germany*

Accepted 2025 October 17. Received 2025 October 10; in original form 2025 January 24

ABSTRACT

We present a Gaussian process binary classifier designed to incorporate label uncertainty in its training data, with the aim of selecting galaxy cluster candidates based on their observed X-ray properties. The classifier was trained using sources from the North and South fields of the XXL survey, with label uncertainty derived from the existing XXL galaxy cluster selection criteria. To prevent the classifier from simply replicating the existing XXL selection, we excluded the two X-ray properties originally used by XXL to identify clusters. Applying the classifier to the XXL North catalogue yielded a new sample of 623 candidate sources, recovering 225 of the 248 clusters previously identified by the standard XXL method. We validated the classifier using two independent optically selected cluster samples. Visual inspection of 530 candidates confirmed 271 cluster candidates, including 95 not previously selected by the XXL process. Accounting for 93 uninspected sources, the purity of the sample was estimated at 0.47 ± 0.02. The newly identified candidates often showed different X-ray morphologies compared to those previously selected by XXL, typically lacking a dominant X-ray component following a β-model surface brightness profile. While classifier results were robust to being trained on the North or South XXL catalogues, subtle and unresolved differences in behaviour were identified, possibly due to differences in the properties of the two fields (e.g. Galactic column and foreground differences, or time-varying instrument calibration or background characteristics). Overall, we find that the classifier is complementary to the standard XXL processing.

Key words: methods: data analysis – methods: statistical – software: machine learning – galaxies: clusters: general – X-rays: galaxies: clusters.

1 INTRODUCTION

Forming from the collapses of the largest perturbations in the initial matter distribution of the Universe, galaxy clusters provide key insights into cosmology along with the formation and evolution of their member galaxies. Driven by the need for large samples of clusters with well-defined selection functions in order to exploit clusters as cosmological and astrophysical probes, many surveys for galaxy clusters have been carried out over a wide range of wavelengths. These range from the radio regime (selecting clusters via the Sunyaev–Zeldovich, or SZ, effect) through optical (via their galaxy populations or weak gravitational lensing signals) to X-rays (via emission from their hot intracluster medium; ICM).

The first surveys for clusters of galaxies were performed in the optical, with clusters identified by eye as overdensities of galaxies in photographic images (G. O. Abell 1958). The reliability of this approach to galaxy cluster detection is limited by projection

★ E-mail: jb14389@bristol.ac.uk

effects, where galaxies that are coincident on the sky but physically unrelated are misidentified as a galaxy cluster. The contamination of galaxy cluster samples due to these projection effects can be reduced by identifying galaxies that cluster both in projection and in photometric redshift space (E. S. Rykoff et al. 2014; M. Oguri 2014).

Galaxy clusters have also been identified from optical surveys through the weak gravitational lensing of background galaxies (D. Wittman et al. 2006; S. Miyazaki et al. 2007). This approach relies on a sufficient number density of background galaxies behind the galaxy cluster such that small distortions in their shape due to gravitational lensing can be combined to produce a detectable signal in the form of a shear map.

Surveys conducted using radio wavelength observations such as those from the Planck telescope (Planck Collaboration XXIX 2014), Atacama Cosmology Telescope (M. Hasselfield et al. 2013), and South Pole Telescope (L. E. Bleem et al. 2015) detect galaxy clusters from the SZ effect where cosmic microwave background photons undergo inverse Compton scattering by high-energy electrons within the ICM. The redshift independent nature of the SZ effect makes this approach ideal for finding distant galaxy clusters, however the current sensitivity of SZ surveys means they are not able to detect clusters to such low masses as, e.g. optical surveys.

X-ray surveys for galaxy clusters have included the Einstein Observatory Extended Medium-Sensitivity Survey (I. M. Gioia et al. 1990), the *ROSAT* Brightest Cluster Sample (H. Ebeling et al. 1998, 2000), the *XMM–Newton* Cluster Survey (A. K. Romer et al. 2001), the X-Class Cluster survey (N. Clerc et al. 2012), and the XXL X-ray survey (M. Pierre et al. 2016, hereafter Paper I) used in our work. These surveys detected galaxy clusters from the X-ray emission of their ICM. Due to the typical angular resolution of X-ray telescopes it is difficult to distinguish the point source emission of active galactic nuclei (AGNs) from the extended emission of galaxy clusters, resulting in the misclassification of sources (C. H. A. Logan et al. 2018, XXL Paper XXXIII) that can additionally bias cosmological applications of clusters (S. Bhargava et al. 2023, XXL Paper LI).

Consisting of two 25 sq deg fields widely separated on the sky observed using *XMM–Newton*, the XXL X-ray survey was designed to produce the robust and reliable galaxy cluster samples (Paper I) necessary for cluster science and precision cosmology. The XXL survey has been used to produce a number of such galaxy cluster samples, the bright cluster sample (F. Pacaud et al. 2016, hereafter Paper II), and the 365 cluster sample (C. Adami et al. 2018, hereafter Paper XX). These catalogues have subsequently been used to conduct investigations into the scaling relations between galaxy cluster properties (P. A. Giles et al. 2016, XXL Paper III) and to constrain cosmological parameters (F. Pacaud et al. 2018; C. Garrel et al. 2022, XXL Papers XXV and XLVI, respectively).

In order to produce such cluster samples, XXL first identifies and measures multiple properties of each X-ray source using the XAMIN X-ray pipeline (F. Pacaud et al. 2006; L. Faccioli et al. 2018, hereafter Paper XXIV). Galaxy cluster candidates are then selected from the X-ray source catalogue by imposing simple selection criteria to the three measured source properties considered most sensitive to a source being a galaxy cluster (described in Section 2.1). By limiting the number of measured source properties and the complexity of the selection criteria, the results of this approach are relatively straightforward to interpret but ignore potentially valuable information. In particular, the simple selection inherently fails to select dimmer and less relaxed galaxy clusters that often just miss the selection criteria.

Machine learning (ML) classification techniques have led to the development of complex mathematical models designed to maximize the use of all available information when selecting a desired subset of a parent population. Within astronomy the prospect for ML classifiers to produce large, reliable source samples, combined with their greater efficiency compared to traditional methods makes them particularly appealing. The challenge of searching for sources in ever larger data sets is becoming acute due to the scale and sensitivity of imminently arriving wide-area surveys such as the *eROSITA* (P. Predehl et al. 2021) and *Euclid* (R. Laureijs et al. 2011) surveys. This has driven a significant increase in the use of ML classifiers in astrophysics. Examples include: classification of variable stars from time-series data (J. W. Richards et al. 2011); galaxy cluster detection from the Sloan Digital Sky Survey (J. Hao et al. 2010); and the use of convolutional neural networks to select galaxy clusters from XXL's sister survey X-Class (M. Kosiba et al. 2020).

The use of ML classifiers for source selection presents the astronomer with two new problems to solve. The first is that the complex nature of the statistical models (particularly neural networks and models developed from them), makes developing a physical understanding as to why a model does or does not select any given source difficult, if not impossible.

The second problem is that supervised ML techniques require larger, near perfectly labelled training data. There are several ways of creating such data sets, but each has drawbacks. The first option, given the availability of a relevant and sufficiently large source catalogue, is to have an expert or group of experts inspect and label a large number of sources. However, this requires a significant time investment for those involved. Second is the option for crowd-sourcing the labelling process of a sufficiently large source catalogue to non-experts through citizen science schemes. This approach has the disadvantage of introducing a level of uncertainty in the labels produced. The final option is the use of simulated data where a large number of sources can be generated from the simulation and the labelling process automated using the known ground truth. This approach, however, relies on the simulations accurately recreating the full population of sources that may be present within a real catalogue, including their relative densities in parameter space.

In this paper, we present an alternative approach in which sources in a training data set are labelled based on a traditional classification scheme using a subset of their measured properties. This has the advantage that the training data perfectly represent the real data, but the disadvantage is that the labels are uncertain as the traditional classifier will misclassify a subset of the training sources. We propose a novel adaption of a Gaussian process (GP) binary classifier (C. Williams & C. Rasmussen 1996) to take into account the uncertainty on the labels. Our adapted GP binary classifier is trained on the latest internally available XXL X-ray source catalogue, using the existing galaxy cluster candidate samples to provide the labels. While training on the existing cluster candidates avoids the need for a ground truth it can introduce a bias towards the types of cluster previously selected by XXL. One of the main aims in developing this approach is to supplement the existing XXL catalogue by identifying additional galaxy cluster candidates. The GP classifier achieves this by making use of the full range of source properties measured by the XAMIN pipeline but not used in the traditional classification scheme. We subsequently interpret the results using automatic relevance determination (ARD) to identify those parameters with the greatest impact on the classifiers output.

This paper is structured as follows; Section 2 describes the XXL source catalogue along with the pre-processing applied to the catalogue data and the optically selected galaxy cluster sample used

when testing the classifier. Section 3 contains a general description of a GP binary classifier and the adaptation that we introduce to take into account the imperfectly labelled training data. Section 4 describes the results of the classifier when trained on the XXL X-ray source catalogue and validation of the classifications by comparison with an optically selected cluster catalogue and through visual inspection of observations of a subsample of XXL sources. In this section, we also present our measurements of the relevance of the input source parameters to the source classification. Section 5 discusses how the model is identifying galaxy cluster candidates from the XXL source catalogue, compares the sample of cluster candidates produced by the model to those already used by XXL and the application of the model to other source catalogues. Finally, Section 6 presents our conclusions.

2 SOURCE CATALOGUES

In order to train and test our GP model, we use the latest internally available version of the XXL source catalogue covering the North and South (XXL-N, XXL-S) fields. In this catalogue, each X-ray source is characterized by multiple measured properties derived from the XAMIN pipeline analysis of the *XMM–Newton* observations, as discussed below. So that we can explore and interpret the results of the GP model, we then use the latest CAMIRA catalogue created from the HSC-SSP (Hyper Supreme-Cam-Subaru Strategic Program) S21A data set, an independent optically selected catalogue of clusters and cluster candidates, which we then cross-match with the XXL source catalogue.

2.1 XXL source catalogue

The XXL survey is described in Paper I. Past versions of the XXL source catalogue were generated from identifying sources in individual *XMM–Newton* X-ray observations and subsequently reconciling the duplications caused by the overlap between said observations (L. Chiappetti et al. 2018, XXL Paper XXVII). In the latest (unpublished) version of the catalogue − version 4.3 used here − the X-ray observations were first combined into continuous images of each field and then split into a regular mosaic of 68 arcmin × 68 arcmin tiles overlapped by 8 arcmin before source detection was carried out on each tile independently (Paper XXIV). Source detection and characterization was then carried out using version 4.3 of the XAMIN pipeline to create the resulting XXL source catalogue. The resulting North and South catalogues contain 24 412 and 18 090 sources respectively, including point sources, extended sources and sources considered spurious. Sources flagged as spurious (low detection significance) are removed from published XXL catalogues, but we retain them here to explore whether the GP is able to recover plausible galaxy cluster candidates from that population. The 8 arcmin overlap between tiles results in multiple detections of the same source. These repeated detections do not have a significant impact on this work and so are retained in the catalogue used here.

Here, we provide a brief summary of those details of version 4.3 of the XAMIN pipeline most relevant to our analysis. Sources were detected using a wavelet method. Four surface brightness models were independently fitted to the X-ray image of each detected source. Each model corresponds to a different astrophysical object or pair of objects as listed in Table 1. The four models represent the most likely scenarios for the origin of the detected X-rays in the survey while attempting to identify juxtapositions of sources that might confound a simple attribution of an extended source as emission from a massive halo. Each of these fits were performed separately in the $0.5 - 2$ keV (soft) and $2 - 12$ keV (hard) energy bands. Through experimentation, we found that including the hard band in the data provided to the ML model gave no appreciable increase in accuracy, while incurring a significant increase in the computation time. Consequently, we choose not to use the hard band results within this work.

Table 1. The four surface brightness models used by version 4.3 of the XAMIN pipeline and the astrophysical object(s) that they are intended to represent. The three letters in parentheses are the abbreviations used for each model.

Model	Astrophysical object(s)
Beta model (EXT)	Extended cluster emission
Point source model (PNT)	AGN
Double point source model (DBL)	Two AGNs with overlapping emission
Beta model with central point source (EPN)	Extended cluster emission contaminated by emission from a central AGN

A number of parameters were measured for each source by fitting each model (see Table 2 for a summary of those parameters relevant to this work). In addition, we derived the background rate values for each source from the XAMIN output. The background count rate for a given surface brightness model and detector was calculated by multiplying the background photon count by the ratio of the sources measured count rate to photon count using the same surface brightness model and detector. Noted in bold in Table 2 are the subset of parameters over which the GP model is fitted, the choice of which is discussed in Section 3.3.

2.1.1 *The XXL C1 and C2 cluster samples*

The cluster sample selected in the standard XXL analysis (Paper I) is derived from the XXL catalogues by applying cuts in the parameter space defined by the core-radius (EXT) and extension likelihood (EXT_STAT) parameters. The standard XAMIN pipeline also uses EXT_DET_STAT to select galaxy clusters (Paper XXIV), however, we found that this parameter did not significantly impact the sample selection and so do not use it in this work. The EXT parameter was determined by fitting a β-model surface brightness profile to each source (the EXT model in Table 1). The EXT_STAT parameter is determined by independently fitting the β-model and point source models to the source. EXT_STAT acts as a measure of the significance of the extended model fit over that of the point model. The results of fitting the double point-source and extended plus point source surface brightness models are not considered when selecting the $C1$ and $C2$ source samples.

In the standard XXL analysis, the cut values listed in Table 3 are chosen so that the fraction of sources within the sample that are genuine detections of X-ray emission from a galaxy cluster is 0.95 (known as the $C1$ subsample) and 0.5 (the $C2$ subsample). These cut values were determined from an analysis of simulated XXL catalogues produced for version 3.3 of the XAMIN pipeline (Paper I). With respect to version 4.3 of the XXL catalogue used in this work these values should be considered approximate (updated values are planned for the public release of version 4.3 of the catalogue).

Follow-up observations in multiple wave bands have shown that $C1$ sources are in almost all cases (> 90 *per cent*) genuine detections of clusters (Papers II and XX). Given the simulated sources were synthesized using β-models appropriate for relaxed and well-virialized clusters, the $C1$ and $C2$ subsamples of real XXL sources

Table 2. List of the source properties measured by the XAMIN pipeline that were used in our work. The first three letters of each parameter name denote the surface brightness model used when measuring that parameter (see Table 1).

Parameter	Description
EXT_STAT∗	A measure of the significance of fitting an extended model over a point model. Given by EXT_DET_STAT divided by PNT_DET_STAT
EXT∗	A measure of the physical extent of the source on the sky in arcseconds
EXT_DET_STAT	A measure of the likelihood that the fitted extended source would be detected
EXT_RATE_MOS	Count rate in mos detectors
EXT_RATE_PN	Count rate in pn detector
EXT_BG_RATE_MOS	Background count in mos detectors
EXT_BG_RATE_PN	Background count in pn detector
PNT_DET_STAT	A measure of the likelihood that the fitted point source would be detected
PNT_RATE_MOS	Count rate in mos detectors
PNT_RATE_PN	Count rate in pn detector
PNT_BG_RATE_MOS	Background count in mos detectors
PNT_BG_RATE_PN	Background count in pn detector
DBL_DET_STAT	A measure of the likelihood that the fitted double point source would be detected
DBL_RATE_MOS	Count rate in mos detectors
DBL_RATE_PN	Count rate in pn detector
DBL_BG_RATE_MOS	Background count in mos detectors
DBL_BG_RATE_PN	Background count in pn detector
DBL_SEP	Angular separation between point sources
DBL_RATIO	Flux ratio of point sources
EPN_DET_STAT	A measure of the likelihood that the fitted extended plus point sources would be detected
EPN_RATE_MOS	Count rate in mos detectors
EPN_RATE_PN	Count rate in pn detector
EPN_BG_RATE_MOS	Background count in mos detectors
EPN_BG_RATE_PN	Background count in pn detector
EPN_RATIO	Flux ratio of the point and extended source

Notes. Parameters marked with ∗ are those that are used to classify sources as cluster candidates in the standard XXL pipeline, while the parameters in bold are those over which the GP model is trained.

Table 3. Cuts used to select the $C1$ and $C2$ cluster samples from version 4.3 of the XAMIN catalogue. These cut values are taken from version 3.3 of the XAMIN catalogue (Paper I) and should be considered approximate. The standard XAMIN pipeline also applies cuts on EXT_DET_STAT (Paper XXIV) however since we find this has no significant impact on the $C1$ and $C2$ samples they are not used in this work.

Cluster sample	EXT (arcsec)	EXT_STAT	Count (N) North	South
$C1$	> 5	> 33	139	109
$C2$	> 5	15 to 33	86	81

are likely to be dominated by such systems. This was deliberate in order to make the selection well matched to the the cosmological goals of the XXL project outlined in Paper I. Less relaxed systems not dominated by a single virialized halo are likely to be less well represented in the sample.

Fig. 1 depicts the distribution of sources within the EXT and EXT_STAT parameter space and whether they fall into the $C1$, $C2$, or neither sample. Throughout this work, we refer to the (majority) set of sources not contained within the $C1$ or $C2$ subsamples as the non-$C1C2$ sample. Note that the XAMIN models are convolved with a model of the point spread function (PSF) that depends on the position of the source relative to the pointing (L. Faccioli et al. 2018), so a non-zero EXT coupled with a high EXT_STAT implies an extended source (although the $C1$ and $C2$ definitions conservatively use a minimum EXT of 5 arcsec). The $C2$ sources have similar extents to the $C1$s, but these are less-reliably measured (have a lower EXT_STAT). This can be for a variety of reasons – for example the sources could be detected with fewer counts than those typical for a $C1$ cluster or the local background/noise could be relatively high, reducing the fidelity of the fit.

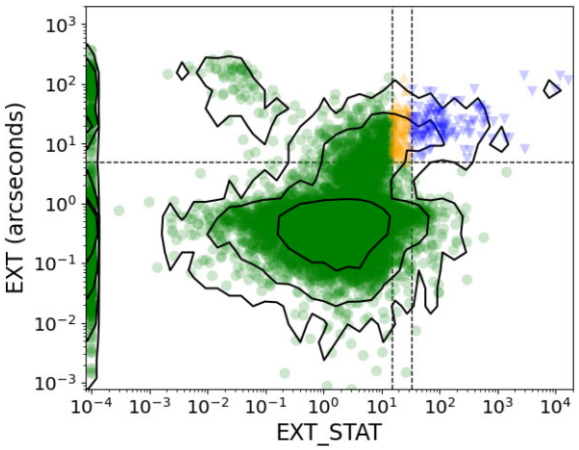


Figure 1. Distribution of XXL catalogue sources as a function of EXT_STAT against EXT labelled by whether they fall into the $C1$ (blue triangle pointing down, top right region) or $C2$ (orange triangle pointing up, top midle region) cluster samples or belong to neither (green circle, top left and bottom regions). Contours indicate a 10-fold increase in source density. Dashed lines indicate the $C1$ and $C2$ selection criteria listed in Table 3. Those sources with an EXT_STAT of zero are plotted on the left of the figure. The figure shows sources from the XXL North source catalogue. We note that the same distribution is seen for sources in XXL's southern source catalogue.

The region occupied by the $C2$ sample in Fig. 1 appears relatively small (particularly the range in EXT_STAT). Nevertheless, this is a densely populated region – the ratio of $C1$ to $C2$ sources in the North XXL catalogue catalogue being $\sim 4:3$. The narrowness in the range of EXT_STAT values indicates that relatively small perturbations in these values (for example if two otherwise identical sources fall in different regions of the mosaiced images leading to differences in backgrounds, or the PSF at the two positions) could promote a source from a classification of $C2$ to $C1$ or relegate it from being considered a cluster altogether (to a non-$C1C2$). Clearly, one hoped for advantage of a different classification scheme (such as the GP-based scheme we are exploring) is to provide an insight into these sources considered boundary cases in the $C1$ and $C2$ classification scheme.

2.1.2 *Pre-processing*

This work makes use of the XAMIN source catalogue created by fitting the four surface brightness models (Table 1) to both the combined image from *XMM–Newton*'s two MOS cameras and the image from *XMM–Newton*'s pn camera.

Before the catalogues can be used by the GP, the data must be cleaned and normalized. A total of eight X-ray observations and their associated sources were removed from the North catalogue due

to issues with their observations (e.g. increased background noise). Sources for which one or more of the parameters in Table 2 were missing were excluded from further analysis. A visual inspection of combined optical and X-ray images of these removed sources indicates that these are erroneous detections left in the catalogue. This reduces the North and South catalogues from 24 412 and 18 090 to 23 626 and 18 069 sources, respectively.

After this cleaning step, the data were normalized, reducing the variation in scale between parameters. The collection of values for each parameter were linearly rescaled by subtracting the mean and dividing by the standard deviation of the values. This removes the impact of the choice of physical units from each parameter and maps them onto a similar dynamic range. In principle this rescaling is unnecessary for a GP, but it simplifies the interpretation of which parameters are most important in influencing the outcome of the GP classification (see Sections 3.4 and 5).

2.2 CAMIRA optically selected clusters

To test the utility of the GP we applied it to a subset of the XXL sources selected by proximity to optical cluster candidates. This subset should then contain a higher proportion of real clusters than the full XXL source catalogue, and is defined independently of the sources' X-ray properties. If the GP were performing correctly it would be expected to identify more sources in this subset as potential galaxy clusters than the full XXL sample.

We used the list of optically selected cluster candidates from the latest CAMIRA catalogue produced from the HSC-SSP S21A data set (M. Oguri et al. 2018, reporting a similar catalogue produced from the S16A data set). The catalogue was produced by applying the CAMIRA red-sequence detection algorithm (E. S. Rykoff et al. 2014; M. Oguri 2014) to the SSP imaging survey (H. Aihara et al. 2022) with HSC-SSP (H. Aihara et al. 2018) in a similar manner to that of J. P. Willis et al. (2021). In the $\approx$ 22 sq deg overlap between the HSC-SSP and XXL-N, there are 572 CAMIRA selected clusters. This is larger than the 270 reported in J. P. Willis et al. (2021) as here we are using a more recent CAMIRA catalogue, with richness selection of $N > 10$ instead of $N > 15$ as used by J. P. Willis et al. (2021). In the following, we refer to these optically selected clusters as the CAMIRA sample.

The CAMIRA sample was matched with the XXL source catalogue after pre-processing (see Section 2.1.2) using a matching radius of 15 arcsec. This is a relatively conservative choice, designed to minimize chance associations between X-ray sources and CAMIRA clusters in order to yield a high-purity subset. In the case of multiple XXL sources within 15 arcsec of a CAMIRA cluster's position, we treat all XXL sources as potential X-ray counterparts. This produced a subset of 162 XXL sources with CAMIRA counterparts of which $\sim$ 29 are expected to be chance associations. This CAMIRA-matched catalogue constitutes a subset of XXL sources that should have a higher fraction of genuine clusters than would a random set of XXL sources. Since the subset is defined independently of the X-ray properties of the sources, this provides a useful tool with which to test the GP.

The fraction of XXL sources within 15 arcsec of a CAMIRA cluster that are classified as $C1$ and $C2$ is 0.21 and 0.13, respectively, compared to 4.0×10^{-3} and 2.8×10^{-3} for the full XXL source catalogue. Hence, it is clear that this CAMIRA sample contains a higher fraction of genuine X-ray detections of galaxy clusters than the full XXL North source catalogue.

We also use the GAMA spectroscopic survey (S. P. Driver et al. 2011) to produce a GAMA sample of XXL sources within 15 arcsec of a GAMA group of 10 or more members. The DR3 (I. K. Baldry et al. 2018) GAMA group catalogue[1] having been generated from the GAMA spectroscopic survey using a friends-of-friends algorithm (A. S. G. Robotham et al. 2011).

As with the CAMIRA sample, the GAMA sample showed an increased fraction of $C1$ and $C2$ sources, 0.27 and 0.10, respectively, compared to the full XXL North catalogue, 4.0×10^{-3} and 2.8×10^{-3}, indicating the sample contains a higher fraction of genuine X-ray detections of galaxy clusters than the full XXL North catalogue.

3 GAUSSIAN PROCESS MODEL

The model used in this work consists of a GP binary classifier (C. Williams & C. Rasmussen 1996) that we modify to take into account uncertainty in the labelling of the training data.

This section contains a brief description of a GP and how it is adapted for binary classification tasks (for a full description, see C. M. Bishop & N. M. Nasrabadi 2006). We also include details on the modifications made to the GP binary classifier model to accommodate uncertain training data labels (those derived from the existing XXL source classification). We subsequently outline how ARD identifies source properties as important, including how we adapted the measurement of the relevance values for our modified classifier. The section concludes with details of the measured source properties supplied to the GP (listed in bold in Table 2) and the initial probability of a source being a galaxy cluster based on the existing XXL classification.

3.1 Gaussian process

A GP is a non-parametric model designed to predict the value of an unknown continuous real function from a series of measured training points that populate some parameter space.

The core principle for a GP is that for any real function $y(\boldsymbol{x})$ of parameters $\boldsymbol{x}$ the values of the function y_1 and y_2 at two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are increasingly likely to have a small difference in value for increasingly similar points (i.e. two points may be considered increasingly similar as their separation in parameter space decreases).

The similarity of any two points is determined by the choice of symmetric kernel function, $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Within this work, we use a radial basis function (RBF) kernel encoding the premise that two objects should have a similar probability of being a galaxy cluster if they have similar measured properties.

The aim of a GP is to predict the measured value t_{N+1} at some point $\boldsymbol{x}_{N+1}$ given a set of N measured values $\boldsymbol{t}_N$ at points $\boldsymbol{x}_N$ (referred to as the training set). The joint probability distribution over the new sample point and training set is given by,

$$P(t_{N+1}, \boldsymbol{t}_N) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{t}_N \\ t_{N+1}\end{bmatrix} \middle| \begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^T & c\end{bmatrix}\right) \quad (1)$$

here the covariance matrix is explicitly separated into; (i) the covariance matrix for the N training points $\boldsymbol{C}_N$ with entries $C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \delta_{ij}\beta^{-1}$ (where δ_{ij} is the Kronecker delta and β the measurement noise on each entry of $\boldsymbol{t}$); (ii) the vector $\boldsymbol{k}$ containing the measures of the similarity between the training points and the sample point with N entries given by $k_i = k(\boldsymbol{x}_i, \boldsymbol{x}_{N+1})$; and (iii) c the kernel value for $\boldsymbol{x}_{N+1}$ plus the noise term $c = k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) + \beta^{-1}$.

[1] GAMA group catalogue DMU version G3Cv10, obtained from http://www.gama-survey.org/dr3/data/cat/GroupFinding/v10/.

In order to predict the value of t_{N+1}, it is necessary to find the conditional probability of t_{N+1} given the training points $\boldsymbol{t}_N$. For a joint multivariate normal distribution, as in equation (1), the conditional distribution over one component is given by C. M. Bishop & N. M. Nasrabadi (2006)

$$P(t_{N+1}|\boldsymbol{t}_N) = \mathcal{N}(t_{N+1}|\boldsymbol{k}^T\mathbf{C}_N^{-1}\boldsymbol{t}_N, c - \boldsymbol{k}^T\mathbf{C}_N^{-1}\boldsymbol{k}) \tag{2}$$

3.2 Binary classifier

The aim of a binary classifier is to estimate the probability of an object having a label, L, that is true ($L = 1$) or false ($L = 0$) as a function of the object's measured properties. Hence, given some training set of N labelled objects the probability of a new object $N + 1$ having a positive label is denoted as $P(L_{N+1} = 1|\boldsymbol{L}_N)$. Here, we have changed the notation from t to L to indicate that the predicted value is now a binary label instead of a continuous quantity. In the context of this work, the two labels correspond to whether an object is a galaxy cluster ($L = 1$) or not ($L = 0$). For this binary classification application, the unknown function approximated by the binary classifier GP is the probability of an object at any location in the parameter space having a positive label. The labelled training objects then act as measurements of this unknown function. The label for a new object is then likely to be the same as those of training objects with which it has a high similarity.

The general GP described in Section 3.1 predicts values for a function over the entire real axis, while for a binary classifier the probability of an object having a positive label $P(L_{N+1} = 1|\boldsymbol{L}_N)$ is restricted to the range 0 to 1, meaning that it cannot be directly predicted by a GP. This problem is solved through the use of a sigmoid function to map the output of the GP from the entire real axis to the range zero to one. The process of changing the GP output in this way results in intractable integrals that must be solved by approximation. This work makes use of expectation propagation (C. K. Williams & C. E. Rasmussen 2006), implemented in the GPY package (GPY 2012) to approximate the intractable integrals. To reduce computation time, we use the deterministic training conditional approximation, a type of sparse GP approximation (J. Quiñonero-Candela & C. E. Rasmussen 2005).

3.3 Gaussian process binary classifier adaption for data labelled by sample purity

The GP binary classifier described in the previous subsection requires a training set of correctly labelled objects covering a broad range of measured properties with examples of both positive and negative labels. By the nature of astrophysical surveys such as XXL, imperfect data and generally limited availability of followup observations mean that this type of training set is not available. Surveys instead provide source samples with estimated fraction of positive labels (i.e. the purity of the samples). For XXL, these take the form of the $C1$ and $C2$ galaxy cluster samples with an estimated fraction of galaxy cluster detections of 0.95 and 0.5, respectively. The remaining sources within the XXL catalogue then form a sample with a fraction of X-ray detections of galaxy cluster of approximately 0.

In our analysis, these purity estimates are treated as an initial estimate of the probability of a particular source having a positive label $P(L_i = 1) = u_i$ (i.e. being a galaxy cluster). For a given set of N sources, it follows that the probability of a set of labels $\boldsymbol{L}_N$ is given by multiplying the probabilities for the label of each individual source,

$$P(\boldsymbol{L}_N|\boldsymbol{u}_N) = \prod_i^N u_i^{L_i}(1 - u_i)^{1-L_i}. \tag{3}$$

Here, $\boldsymbol{u}_N$ is a vector containing the probability u_i for each source.

The probability of a new source having a positive label given N training objects labelled by source sample is then given by marginalizing over all possible combination of labels $\boldsymbol{L}_N$:

$$P(L_{N+1} = 1|\boldsymbol{u}_N) = \sum_{\boldsymbol{L}_N} P(L_{N+1} = 1|\boldsymbol{L}_N)P(\boldsymbol{L}_N|\boldsymbol{u}_N) \tag{4}$$

where $P(L_{N+1} = 1|\boldsymbol{L}_N)$ is the probability estimated by a GP binary classifier for a given set of labels $\boldsymbol{L}_N$ as described in Section 3.2. This same principle can be applied to a different binary clasification model by replacing the GP model used to calculate $P(L_{N+1}|\boldsymbol{L}_N$.

Because the number of distinct combinations of $\boldsymbol{L}_N$ scales as 2^N, the full calculation is impossible for any reasonably sized data set. We instead approximate the sum using a Monte Carlo approach, sampling $\boldsymbol{L}_N$ from $P(\boldsymbol{L}_N|\boldsymbol{u}_N)$.

A consequence of this approximation, as we will see in Section 4.4, is that the estimated probability $P(L_{N+1}|\boldsymbol{u}_N)$ output by the GP binary classifier is not a direct estimate of the probability, instead acting as a non-linear measure of the true probability of the source being a galaxy cluster. In other words, if the GP binary classifier returns a value of 0.4, this does not mean the source in question has a 40 per cent probability of being a cluster. Instead, the value relates in some non-linear way to how likely the GP binary classifier believes the source is to be a cluster. Consequently, we treat the estimated probability produced by the GP simply as a figure of merit, that we call a 'confidence value'. A source with a confidence value close to 1 is more likely to be a cluster, while a source with a confidence value close to 0 is less likely to be a galaxy cluster. Cluster samples can then be selected on the basis of their confidence value.

3.4 Hyperparameter optimization and automatic relevance determination

For a GP, the noise term β and any constants within the kernel are considered to be hyperparameters that need to be optimized for the problem being solved. The hyperparameters $\boldsymbol{\theta}$, can be optimized for a GP binary classifier described in Section 3.2 by maximizing the likelihood of the training labels $\boldsymbol{L}_N$ given $\boldsymbol{\theta}$, $P(\boldsymbol{L}_N|\boldsymbol{\theta})$ (C. M. Bishop & N. M. Nasrabadi 2006).

As described in Section 3.3, the labels are not perfectly known, and the GP binary classifier was adapted for objects labelled by sample purity $\boldsymbol{u}_N$. In order to calculate the optimal hyperparameters, it is necessary to take the average for each hyperparameter optimized to each set of labels $\boldsymbol{L}_N$ and weighted by the probability of said set of labels:

$$\bar{\boldsymbol{\theta}} = \sum_{\boldsymbol{L}_N} \boldsymbol{\theta}(\boldsymbol{L}_N)P(\boldsymbol{L}_N|\boldsymbol{u}_N) \tag{5}$$

As in Section 3.3, this sum is approximated using a Monte Carlo approach.

Having optimized the hyperparameters to the training data their values must encode some information as to the relationship between the input parameters and the model output. It is possible to determine the relevance of each input parameter from the values of the hyperparameters using ARD.

The relevance of a parameter relates to how quickly the prediction of a GP changes as a function of that parameter. More relevant

parameters are those for which a small change in value leads to a large change in the output value of the GP. The rate at which the prediction of a GP can change is determined by the form of the kernel function and the values of the kernel's hyperparameters.

This work makes use of a RBF as the kernel,

$$k(\boldsymbol{x}, \boldsymbol{x}') = V \exp\left(-\sum_{q}^{Q} \frac{(x_q - x_q')^2}{l_q^2}\right). \quad (6)$$

The quantities x_q and x_q' represent one of the Q measured properties of the two sources, indexed by q. The length-scale for each parameter, l_q, is the hyperparameter that determines the shape of the kernel along said parameter. A shorter length-scale for a parameter means the model output is more sensitive to changes in that parameter.

We express the length-scales as a fraction of the standard deviation of the corresponding parameter. These normalized length-scales are used in Section 5 to investigate the relevance of the different source properties to the GP output.

3.5 XXL implementation

The XXL catalogues contain a large number of properties measured for each source, many of which do not contain useful information for the source classification. After some initial exploration, the properties relevant to this task and subsequently provided to the GP during training and classification were reduced to the 19 parameters listed in bold in Table 2. The GP was not directly privy to any other source properties.

For the purpose of training the model, the probability of each source being labelled as a cluster ($L = 1$), was based on the estimated purity (u) of the XXL source classification as follows:

$$\begin{aligned} C1: &\ u = 0.95 \\ C2: &\ u = 0.50 \\ \text{non} - C1C2: &\ u = 0.05 \end{aligned}$$

The $C1$ and $C2$ classifications are defined in the 2D parameter space of EXT and EXT_STAT (e.g. Fig. 1), and the u values used here are chosen based on the estimated fraction of galaxy cluster detections within each sample. While the estimated purity of the set of non-$C1C2$ sources as a whole is approximately zero, in reality the distribution of u is not uniform; there is a larger probability of a source that is close to (but outside) the $C1$ and $C2$ regions being a cluster than a source further away. For this reason, we assigned a probability of $u = 0.05$ for a non-$C1C2$ source to be a cluster to better reflect sources close to the $C1C2$ boundary. This is consistent with experience, which shows that in different iterations of the XXL pipeline, sources can move across classification boundaries, so there will be real clusters near the $C1$ and $C2$ regions. In our approach, if we assigned $u = 0$, then none of the non-$C1C2$ sources would ever be labelled as a cluster in the Monte Carlo draws of the training set, reducing the probability of sources outside the $C1$ and $C2$ regions being identified as clusters by the classifier. $C1$ and $C2$ values are set based on their target sample purities.

We emphasize that, because the labels used for our training set are derived directly from the EXT and EXT_STAT parameters, we do not use those parameters in the binary classifier (the parameters used in the binary classifier being listed in bold in Table 2). Including them would lead to the model overfitting. In other words, the classifier would simply re-learn the $C1$ and $C2$ selections and would be ineffective at identifying new sources. This is done despite EXT and EXT_STAT being the most informative parameters as to a source being galaxy cluster. Any measured properties directly equivalent to EXT and EXT_STAT (e.g. EPN_EXT) are removed for the same reason.

The GP is trained separately on the two catalogues created from the North and South fields to avoid field-dependent differences (including in exposure time or Galactic column density) affecting the fit of the GP models. We subsequently use the GP trained on the North field to classify sources in the South field and vice-versa to test for overfitting and explore the effect of field-to-field differences.

The Monte Carlo approximation was run in serial for a total of 100 iterations separated in to 10 batches, each taking $\sim$ 26 h on a single node of the University of Bristol's BlueCrystal phase 4 super computer.

4 RESULTS

This section contains the results of fitting our adapted binary classifier model to the XXL North and South field catalogues separately. We first present the confidence values derived for the full set of XXL sources to which the model was fitted, along with the length-scale for each of the measured source properties. We then measure the confidence values of the subset of XXL sources that are associated with optically detected cluster candidates from the `CAMIRA` catalogue. This enables us to test the performance of the GP on a sample of X-ray sources for which a higher fraction are expected to be clusters. Finally, we conduct a visual inspection of sources selected from the North catalogue on the basis of their GP confidence, to assess the purity of samples derived from the GP.

4.1 Confidence values for XXL sources

Each source in both the North and South XXL catalogues was assigned two confidence values by sampling the adapted GP binary classifier trained on the North XXL and the South catalogues, respectively. We report the confidence values assigned to a source by the GP trained on the catalogue containing that source. These confidence values are colour-coded in a plot of EXT and EXT_STAT in Fig. 2. The high-confidence values broadly occupy the combined $C1$ and $C2$ region of the parameter space, despite the GP not having access to the EXT and EXT_STAT values. There are also sources outside the $C1$ and $C2$ regions that are assigned a high confidence. This indicates that, as intended, the GP is finding sources that are similar to clusters in parameters measured by the XXL pipeline other than EXT and EXT_STAT. These 'extra' high-confidence sources, while lying outside of the $C1C2$ region in Fig. 2, nevertheless tend to be close to it. This behaviour is expected given that large values of EXT and EXT_STAT are reliable signatures of a cluster. The probability of a source being a cluster is a continuous function of position in this region of parameter space, so a significant fraction of sources with EXT_STAT below the $C2$ threshold will still be clusters.

Fig. 2 is also diagnostic of whether the model is overfitting the data. If that were the case, the confidence values would very closely trace the initial uncertainty on the labels, showing sharp increases at the $C1$ and $C2$ boundaries. Since this behaviour is not seen, we infer that overfitting is not occurring.

To further test for overfitting by the GP, we investigated the impact on the GP output of the choice of XXL field used for training the GP. Fig. 3 compares the confidence values assigned to sources in the North and South XXL catalogues when the GP was independently trained on either catalogue. If the GP were overfitting, it would be expected that the confidence values assigned to the sources it was

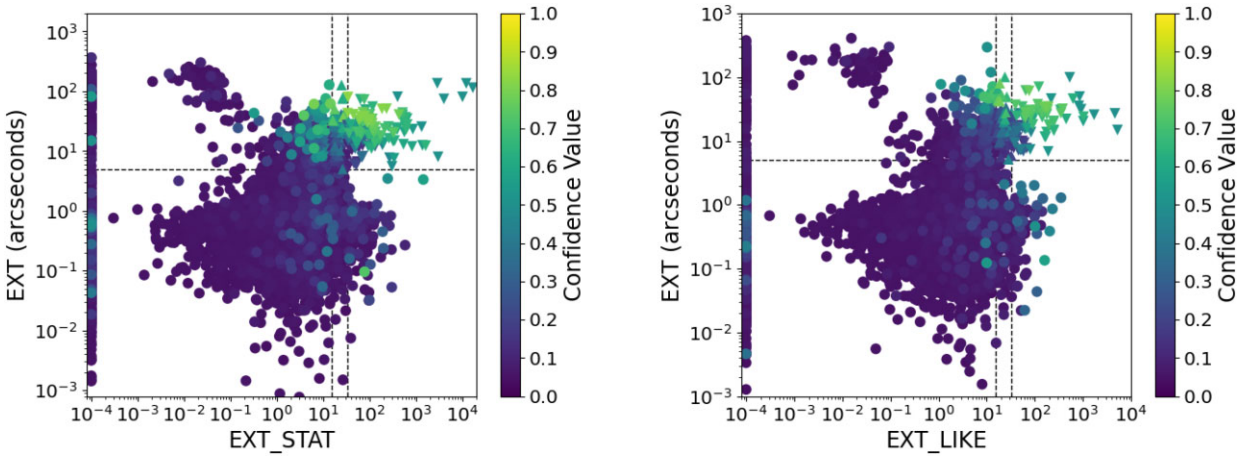


Figure 2. Confidence values as a function of EXT and EXT_STAT for the North (left) and South (right) observing fields. $C1$ and $C2$ sources are plotted as triangles pointing down and up respectively, with all remaining sources denoted by a circle. The colour denotes the assigned confidence value. The dashed lines indicate the $C1$ and $C2$ selection criteria as listed in Table 3. The sources are ordered based on their assigned confidence value such that those sources with a higher confidence value are plotted over those with a lower confidence value. This plotting order is used to make the high-confidence objects visible at the expense of obscuring some low-confidence value sources.

trained on would be close to the initial probability that a source has a positive label, while the confidence values assigned to sources that the GP was not trained on would be close to 0.5 (the default value assigned by the GP in the absence of similar data points. We do not see this behaviour in Fig. 3, indicating that overfitting is not an issue. Instead, there is a general correlation between the confidence values assigned by the GP to the same sources whether the GP was trained on North or South catalogue. For the North catalogue, the Pearson's correlation coefficient between the confidence values assigned by the GP trained on the North and those assigned when trained on the South (top panel in Fig. 3) was 0.95. For the reverse case (bottom panel of Fig. 3) it is 0.91. This implies that overfitting is not an issue, and that training the GP on one field and applying it to another produces reliable results, given the consistency in confidence values when either training set was used.

Fig. 3 shows a general tendency for the confidence values assigned to a source by the GP trained on the South catalogue to be slightly lower than those assigned to the same source when trained on the North. Since this is true of confidence values assigned to sources in both catalogues, this behaviour is due to some underlying differences between the North and South catalogues. As we will see later in Section 4.2, this is further supported by differences in the length-scales of the measured source properties between the GP trained on the North and South catalogues.

Fig. 4 shows the distribution of confidence values generated for the sources in the North and South fields. The vast majority of sources have comparatively low values, which is expected given that the XXL catalogue will be dominated by AGN rather than clusters ($\sim$ 98 per cent of the sources in the catalogue are expected to be AGN; Paper I). Although the nature of the GP means the confidence values do not map to the probabilities used to assign training labels, one might expect the confidence values of the $C1$ sources to cluster around $\sim$ 0.95, $C2$'s 0.50 and non-$C1C2$'s $\sim$ 0.05. The distributions of the confidence values assigned to the $C1$ and $C2$ sources are instead broad due to the vast majority of sources having an initial probability of 0.05 and so by weight of numbers will tend to reduce the confidence for sources that start with higher probabilities through the action of the GP. The most interesting sources are those non-$C1C2$ source whose confidence value is higher than that of the majority of non-$C1C2$ sources ($\sim$ 0.05) due to their similarity with $C1$ and $C2$ objects in the 19D parameter space. Such high-confidence sources are identified by the GP as potential galaxy clusters despite not having been selected by the standard XXL XAMIN pipeline process. If the GP has performed as intended, this population will contain a higher fraction of real clusters missed by the simple $C1$ and $C2$ classification than would a random sampling of non-$C1C2$ sources. We demonstrate that this is indeed the case in Section 4.4.

4.2 Parameter relevance

The (normalized) length-scale of the Gaussian kernel for each parameter was measured as described in Section 3.4 with the resulting values plotted in ranked order in Fig. 5. As described in Section 4.2, a smaller length-scale implies a larger importance when determining the output of the GP. Fig. 5 clearly shows the two dominant parameters driving the output confidence values when the GP is trained on the North catalogue are EXT_RATE_PN and PNT_RATE_PN. When trained on the South catalogue there are an additional two dominant parameters, EXT_RATE_MOS and PNT_RATE_MOS. The most relevant parameters in both cases are measurements of the photon count rate (RATE) from the extended (EXT) and point (PNT) source models fit to the data from the MOS or PN detectors. The remaining source parameters have length-scales significantly larger than their range of values such that they have significantly less (or no) impact on the confidence value output by the GP.

When comparing parameter importance identified using ARD it is important to note its tendency to underestimate the importance of parameters that are linearly related to the model output compared to non-linearly related parameters (T. Paananen et al. 2019). Visual inspection of those parameters deemed irrelevant by ARD showed no correlation with the confidence value output by the GP. The lack of any correlation implies that ARD is not underestimating the importance of the parameters it deems to have low relevance.

Figs 6 and 7 compare the distribution of confidence values for EXT, EXT_STAT, the four most relevant (EXT_RATE_PN, PNT_RATE_PN, EXT_RATE_MOS, PNT_RATE_MOS) and one irrelevant (EXT_BG_RATE_PN) source property as determined by

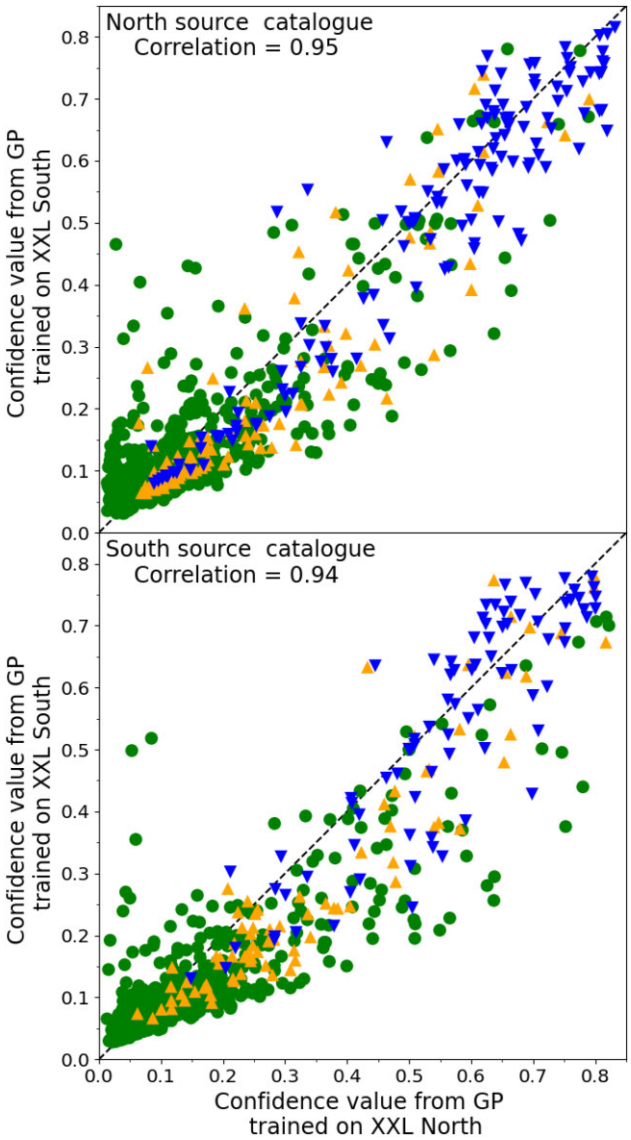


Figure 3. Comparison of confidence values assigned by the GP to sources in the North (top) and South (bottom) XXL catalogues. In both plots, the x- and y-axes show the confidence value assigned when the GP was trained on the North and South catalogues, respectively. The $C1$ and $C2$ sources are plotted as blue triangles pointing down and orange triangles pointing up respectively, with the non-$C1C2$ sources plotted as green circles.

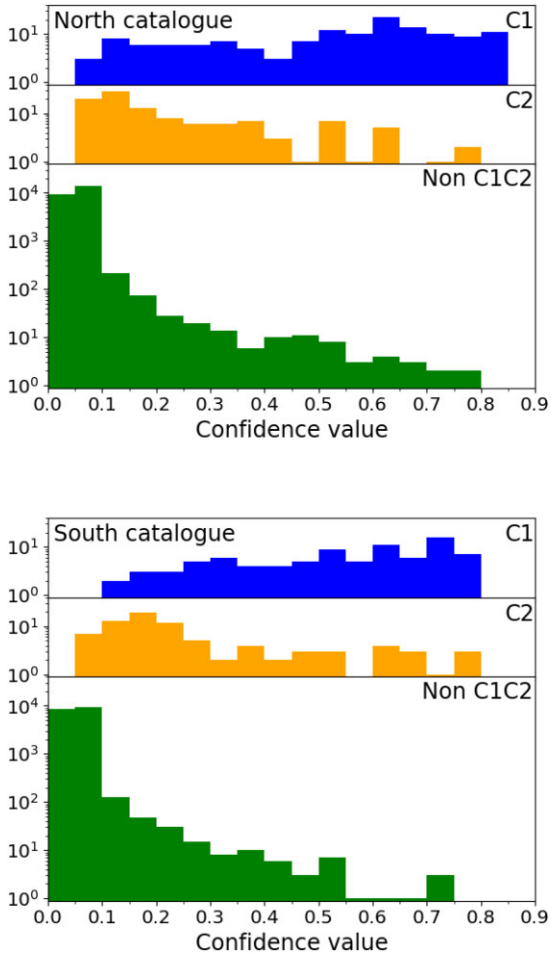


Figure 4. Distribution of confidence values for sources in the North field when the GP was trained on the North catalogue (top) and sources in the South fields when the GP was trained on the South catalogue (bottom). Within each figure the upper, middle, and lower panels show the distributions for the $C1$, $C2$, and non-$C1C2$ sources respectively.

ARD for the southern field (Fig. 5). These projections of the parameter space used by the GP illustrate that the parameters have some degree of correlation with EXT and EXT_STAT, but are not direct proxies for them. Clearly the nature of an X-ray source is encoded in some combination of its parameters other than EXT and EXT_STAT.

We note that the population of sources with a high EXT but low-confidence value in Figs 6 and 7 consists of those sources in Fig. 2 with a high EXT but low EXT_STAT value. Sources with a high EXT and low EXT_STAT are likely to be spurious detections, hence the low-confidence values assigned to them.

In order to gain a qualitative understanding as to why a source is being identified by the GP as more likely to be a galaxy cluster, let us consider the relationship between EXT_RATE_PN (the most relevant parameter in determining the output of the GP) and EXT_STAT (used to define the $C1$ and $C2$ cluster samples, Table 3). From Fig. 8, it is clear that EXT_RATE_PN is positively correlated with EXT_STAT, with a Pearson's correlation coefficient of 0.90. From the distribution of confidence values assigned by the GP, we can see that higher confidence values tend to be assigned to sources with a higher EXT_RATE_PN. This learnt association between a higher EXT_RATE_PN and an increased likelihood of a source being labelled as a galaxy cluster comes from the $C1$ and $C2$ cluster samples requiring a high measured EXT_STAT, which correlates with a higher EXT_RATE_PN. By telling the GP that a galaxy cluster looks like a $C1$ or $C2$ source we have implicitly told it that galaxy clusters tend to have a high EXT_RATE_PN.

Given the high relevance of a source's measured PNT_RATE_PN (indicated by a short length-scale in Fig. 5), it must also contain some information used by the GP when identifying potential galaxy clusters. Considering the distribution of confidence values assigned by the GP as a function of EXT_RATE_PN and PNT_RATE_PN (Fig. 9) we can see that the GP tends to assign higher confidence values to those sources with a measured EXT_RATE_PN value that is larger than their measured PNT_RATE_PN value. The difference between the two profiles means that the count rate of a truly extended

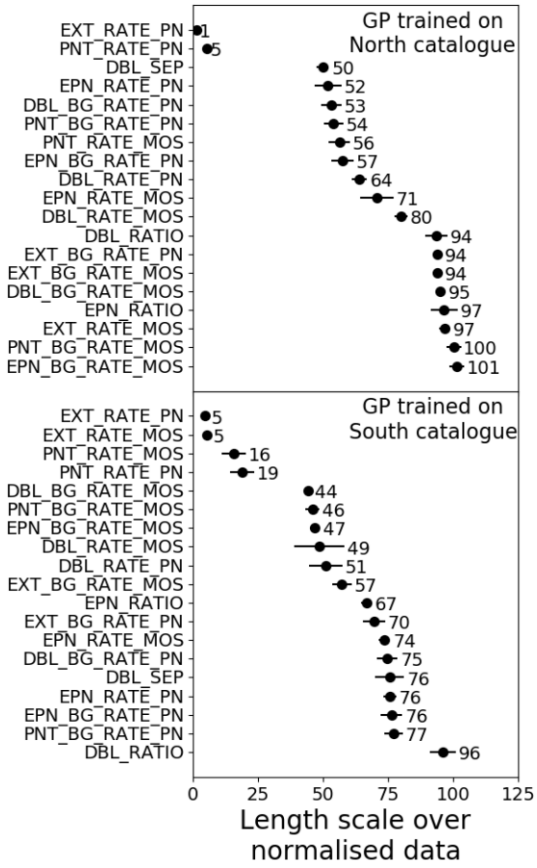


Figure 5. The length-scale of the Gaussian kernel for each parameter used by the GP. Due to the method of normalizing the data, the lengths are expressed in units of the standard deviation of each parameter in the input catalogue (see Section 2.1.2 for details). Shorter length-scales correspond to parameters which have the most influence on the confidence value output by the GP. The error bars show the 1σ uncertainty derived from the Monte Carlo process used when training the GP.

source measured using an extended source model is larger than that measured using a point source model and this is learnt by the GP.

Comparing the length-scales determined by the GP when trained on the North versus South catalogues in Fig. 5, there is a difference in the parameters considered to be relevant. When trained on the South catalogue, the GP treats EXT_RATE_MOS and PNT_RATE_MOS as important but does not do so when trained on the North catalogue. This difference in length-scales is associated with the difference in confidence values found when the GP is trained on each field (as seen in Fig. 3).

There are two possible origins for this difference occurring due to our implementation of the GP model. The first is that the difference is caused by overfitting when the training GP on the South catalogue, a series of small length-scales being symptomatic of overfitting. As discussed previously in Section 4.1, we find no evidence to suggest that this has occurred. The second possibility is that differences in the values used to normalize a measured source property when training on the different XXL catalogues is responsible for the difference in the derived length-scale of that property. We tested this by normalizing the parameter values of the South catalogue by the standard deviations of the corresponding North catalogue parameters. The length-scales output by ARD did not change significantly.

The most significant difference between the North and South field is that there exists a set of observations in the North field with significantly longer exposures than the rest of the North and South fields. Removing those sources in the North catalogue associated with these longer exposures and training the GP as before did not resolve the difference between the two fields.

Given that the extragalactic source population should have similar characteristics in both fields, the difference in length-scales most likely reflects the effect of differences in the path through the Galaxy to the two fields (i.e. absorbing column or diffuse foreground emission), or differences in the parameters of the observations carried out (depth, background, calibration changes, observing mode, etc.).

We therefore conclude that the behaviour in length-scales of parameters are likely to be due to intrinsic differences between the observations of the two fields. A more detailed investigation is needed to determine the exact origin of the difference in length-scales, but as seen in Fig. 3, the GP performs similarly when trained on either field, so the impact of this unresolved issue is small.

4.3 The GP output for the CAMIRA sample

As detailed in Section 2.2, we matched the North XXL catalogue with that of the CAMIRA cluster catalogues in order to identify a subsets of XXL sources that are spatially coincident with a CAMIRA-selected cluster candidate. A similar catalogue was not produced for the South XXL catalogue as it does not overlap the CAMIRA data. The matching was performed after the GP had been applied to the North XXL catalogue, and so did not influence the GP output.

We use the CAMIRA sample to assess how well our GP identifies true clusters, as we expect a higher fraction of the subset of XXL sources with CAMIRA counterparts to be genuine cluster detections, compared to the full XXL source catalogue. Hence, we would expect the GP to assign higher confidence values, on average, to the XXL sources with CAMIRA counterparts.

As seen in Fig. 2, the GP tends to assign a high confidence to $C1$ and $C2$ sources. The same trend can be seen in the CAMIRA subset highlighted in Fig. 10.

Focusing on the 110 non-$C1C2$ sources in the CAMIRA matched subset, Fig. 11 compares their confidence distribution to that of the full non-$C1C2$ sample. It is apparent that the distribution of confidence values for non-$C1C2$ sources within the CAMIRA sample favours higher values than that for the non-$C1C2$ sources within the full XXL North sample. Given the size of the CAMIRA non-$C1C2$ sample, if the confidence distribution were the same as the full XXL non-$C1C2$ sample, we would expect 108 sources to have a confidence value below 0.1 and 2 over 0.1. In fact, we observe 14 sources with a confidence value over 0.1. The probability of this occurring for a random sample of 110 non-$C1C2$ sources is $\sim 7.5 \times 10^{-10}$ (approximated using a binomial distribution with a probability of a source being assigned a confidence value over 0.1 equal to that for the full XXL North catalogue). Consequently, this demonstrates that the GP can select, from X-ray data alone, clusters that were missed by the X-ray-based $C1C2$ classification, but were identified as cluster candidates in the CAMIRA optical catalogue, *without access to that catalogue*.

The same analysis was repeated using the GAMA-matched sample described in Section 2.2. Of the 30 non-$C1C2$ GAMA-matched sources, we find a total of 5 with a confidence value over 0.1. The probability of randomly selecting 5 or more sources with a confidence over 0.1 in a sample of 30 being $\sim 1.4 \times 10^{-4}$. This again demonstrates that the GP classifier is identifying, from X-ray data alone, clusters that were missed by the X-ray-based $C1C2$ classification.

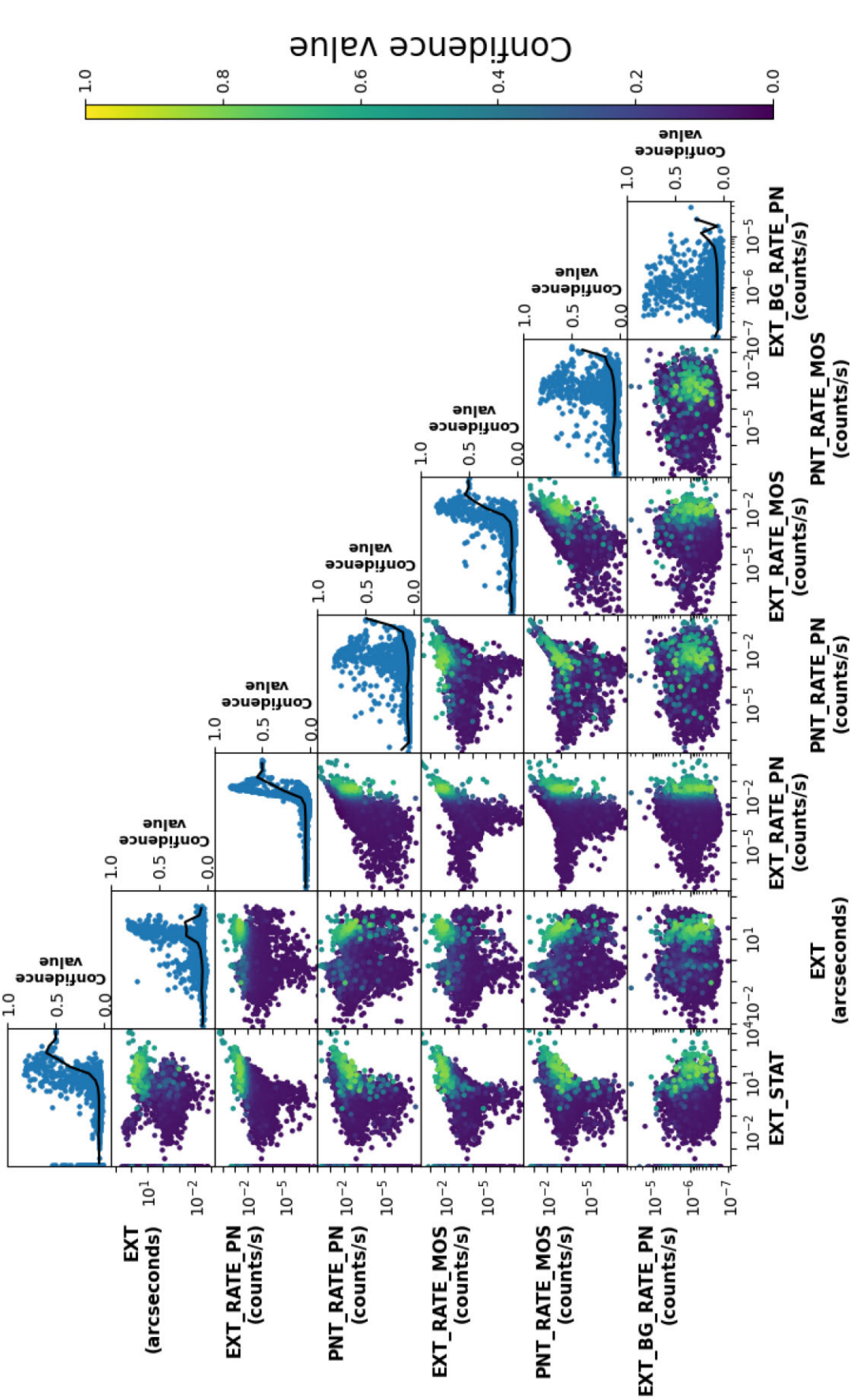


Figure 6. The distribution of the sources in the North catalogue in a subset of the full parameter space chosen to illustrate the behaviour of the GP. The parameters EXT and EXT_STAT were used to label the sources but were not input to the GP. EXT_RATE_PN, and PNT_RATE_PN were considered relevant by the GP when it was trained on either the North or South catalogues. EXT_RATE_MOS and PNT_RATE_MOS were considered relevant by the GP only when it was trained on the South catalogue. EXT_BG_RATE_PN was not considered relevant by the GP when trained on either catalogue, and is included here for comparison purposes. The off-diagonal panels show the scatter plots for each combination of parameters, colour-coded by confidence value assigned by the GP when trained on the North catalogue. Higher confidence points are plotted on top as in Fig. 2. The diagonal panels show the scatter plot of confidence against parameter value for each parameter, the black line showing the average confidence value of sources in 20 logarithmic bins evenly spaced over the parameter axis.

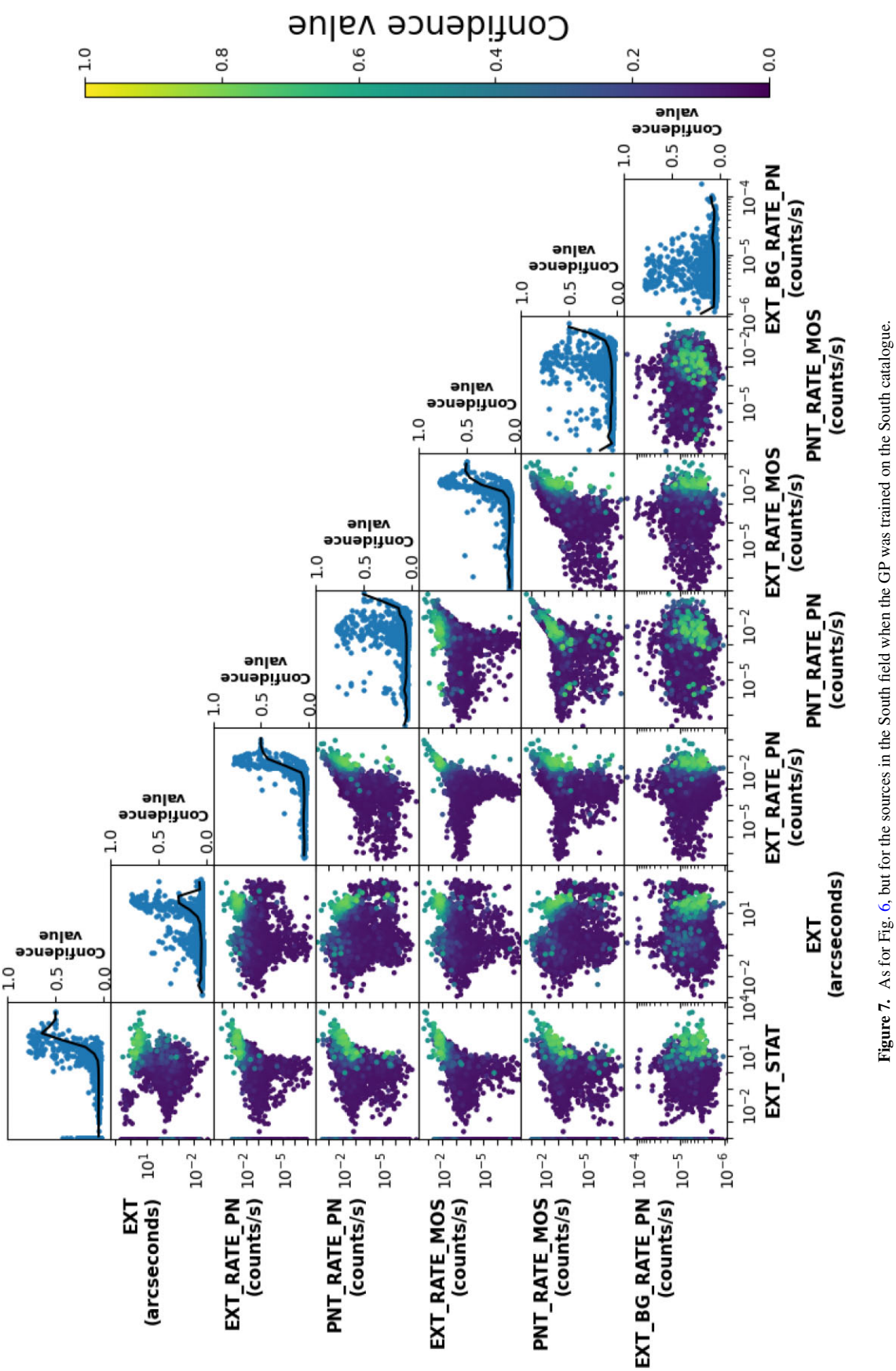


Figure 7. As for Fig. 6, but for the sources in the South field when the GP was trained on the South catalogue.

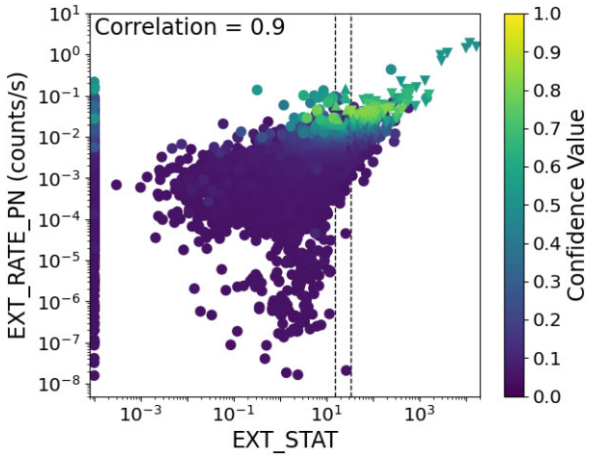


Figure 8. Confidence value as a function of EXT_RATE_PN and EXT_STAT for the North observing field. $C1$ and $C2$ sources are plotted as triangles pointing down and up, respectively, with all remaining sources denoted by a circle. Sources are colour-coded by confidence value. Higher confidence sources are plotted over lower confidence sources as in Fig. 2. Dashed lines indicate the $C1$ and $C2$ selection criteria for a sources measured EXT_STAT value (Table 3).

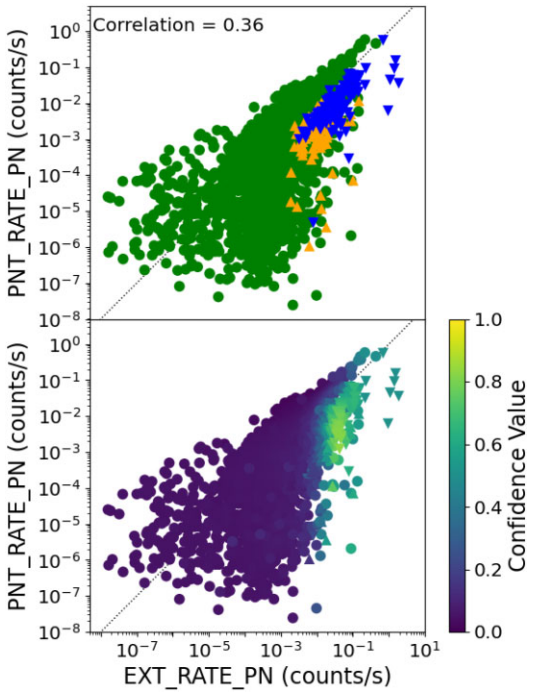


Figure 9. Distribution of sources from the XXL North field as a function of EXT_RATE_PN and PNT_RATE_PN. $C1$ and $C2$ sources are plotted as triangles pointing down and up, respectively, with all remaining sources denoted by a circle. Top, sources are colour-coded by class as $C1$ blue, $C2$ orange, and non-$C1C2$ green with plotting order top to bottom $C1$, $C2$, then non-$C1C2$. Bottom, sources are colour-coded by confidence value with high-confidence sources plotted over low-confidence sources as in Fig. 2. The dotted line indicates a one-to-one relation.

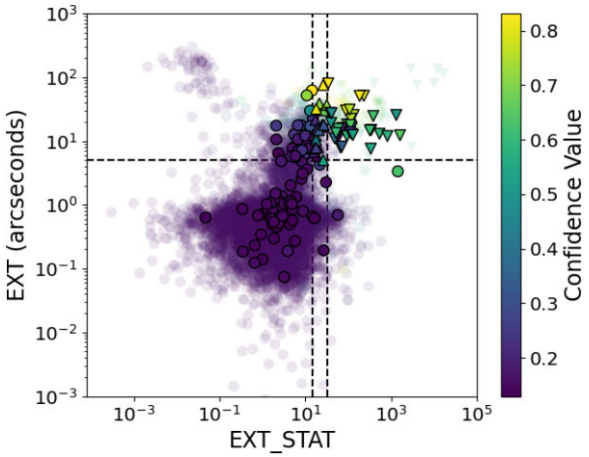


Figure 10. Confidence value as a function of EXT and EXT_STAT for the North field plotted as in Fig. 2. Sources matched to a CAMIRA optical detection are plotted on top and highlighted by a black outline.

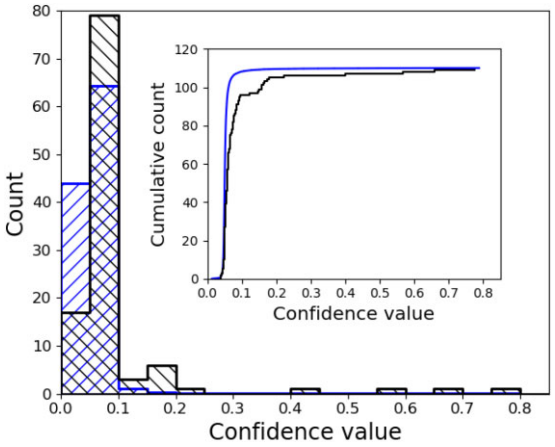


Figure 11. Distribution of confidence values assigned to the 110 non-$C1C2$ sources in the XXL North catalogue matched to a CAMIRA source (black, hash sloping down left to right) and the expected distribution of confidence values for a random sample of non-$C1C2$ sources of the same size (blue, hash sloping up left to right). The cumulative distribution of confidence values for the CAMIRA matched sources (black, lower line) and the expected distribution for a random sample of non-$C1C2$ sources (blue, higher line) is inlaid.

4.4 Visual inspection of GP-selected XAMIN detections

In order to asses the output of the GP and better understand how it identifies sources as more likely to be a cluster, we (authors JCB, MNB, and BJM) conducted a visual inspection of both the X-ray and optical images for a subset of the sources. This was conducted on sources drawn from the North catalogue using *g*-, *r*-, *i*-, and *z*-band HSC imaging (H. Aihara et al. 2018), formed into pseudo-true-colour images of the immediate fields of the X-ray sources (R. Lupton et al. 2004). X-ray images of the same fields were extracted from the XXL North X-ray mosaic in the $0.5 - 2.0$ keV band (Paper XX).

For the purposes of labelling the visually inspected sources, we define a cluster candidate to be an X-ray source with visual evidence of extended X-ray emission associated with an optical overdensity of (usually early-type) galaxies with similar colour in a given field. Any clusters in the XXL survey are typically at $z < 1.2$, and the optical data used are of sufficient depth that overdensities of galaxies associated with clusters out to this redshift are identifiable within

Table 4. Results of visual inspection of sources from various source samples created from the XXL North catalogue. For each subset of sources we report the number of sources in the sample, the number of sources from the sample that were visually inspected, the number of cluster candidates identified by visual inspection, and the estimated purity of the sample. The purity of a source sample is calculated using equation (7) to take into account the uncertainties due to the small number of sources inspected.

Source subset	Total sources in sample (N)	Sources inspected (N)	Cluster candidates (N)	Bayesian estimated purity
$C1$ sources	139	134	122	$0.91^{+0.02}_{-0.03}$
$C2$ sources	109	105	64	0.61 ± 0.04
Non-$C1C2$ CAMIRA sources	110	38	17	0.45 ± 0.08

Table 5. Results of visual inspection of sources not previously selected by XXL (non-$C1C2$) binned on source confidence value. As in Table 4, we report the number of sources in the sample, the number of sources from the sample visually inspected, the number of cluster candidates identified by visual inspection, and the estimated purity of the full sample based on the subset inspected. We note that while a cutout was produced for every source with a confidence above 0.10 a number of these were, for various reasons, not suitable for visual inspection. The reasons that a cutout was not suitable for visual inspection include, a lack of optical data and contamination by a bright optical point source.

Confidence range	Total sources in sample (N)	Sources inspected (N)	Cluster candidates (N)	Bayesian estimated purity
0.00, 0.05	9319	107	12	0.12 ± 0.03
0.05, 0.10	13661	136	39	0.29 ± 0.04
0.10, 0.15	213	176	41	0.24 ± 0.03
0.15, 0.20	74	57	25	0.44 ± 0.06
0.20, 0.25	28	25	7	0.30 ± 0.09
0.25, 0.30	20	15	6	0.41 ± 0.12
0.30, 0.35	14	9	4	0.45 ± 0.15
0.35, 0.40	6	3	2	$0.60^{+0.21}_{-0.22}$
0.40, 0.45	10	8	4	0.50 ± 0.16
0.45, 0.50	11	6	1	0.25 ± 0.15
0.50, 0.55	8	3	1	$0.40^{+0.22}_{-0.21}$
0.55, 0.60	3	2	1	0.50 ± 0.25
0.60, 0.65	4	4	1	0.33 ± 0.19
0.65, 0.60	3	2	1	0.50 ± 0.25
0.70, 0.75	2	2	1	0.50 ± 0.25
0.75, 0.80	2	2	0	$0.25^{+0.21}_{-0.19}$

the colour images. By design, this definition includes X-ray sources associated with both relaxed and unrelaxed galaxy clusters (the X-ray emission simply needs to show signs of extension rather than have a classical β-model surface density profile), and also sources associated with an X-ray halo around the dominant galaxy (or galaxies) in a relatively nearby galaxy group. All of these represent an X-ray detection of a dark matter halo on a scale larger than an individual galaxy. This definition is broader than the standard XXL selection function which is calibrated to identify clusters that resemble β-profiles.

Sources were selected for visual inspection to give a representative population for the full range of GP-assigned confidence values for non-$C1C2$ sources (Table 5), the $C1$ and $C2$ samples, and the non-$C1C2$ CAMIRA matched sources (Table 4).

To enable quantitative comparisons, we define and measure purity for the various source samples as; purity. The fraction of objects that meet our visual inspection criteria for a cluster candidate. Conventionally, the *sample* purity is calculated simply as the fraction of objects in the sample meeting the selection criteria. In this work, we use a Bayesian approach to estimate the *population* purity. This refers to the asymptotic purity of a notional infinitely large sample of objects like those in the set that were inspected. With this method, small samples for which no objects were classified as clusters (i.e. with a sample purity of zero) may produce an estimated population purity that is non-zero. Equivalently, a sample for which all members were classified as clusters (i.e. with a sample purity of one) would produce an estimated population purity that was less than one. Henceforth, unless stated otherwise, the term purity refers to the estimated population purity.

In order to estimate the purity in the manner discussed above, we take the following approach. The likelihood function for the probability of labelling n_c sources as cluster candidates from a sample of size N_s, given a purity p, follows a binomial distribution. We model the prior distribution using a beta distribution with shape parameters $\alpha = \beta = 1$ chosen such that the prior is uniform over the range zero to one. Given that a beta distribution is the conjugate prior of a binomial distribution, the posterior distribution over the purity is itself a beta distribution with shape parameters $\alpha = 1 + n_c$ and $\beta = 1 + N_s - n_c$. The mean purity is hence given by

$$\bar{p} = \frac{n_c + 1}{N_s + 2}. \tag{7}$$

This is equivalent to artificially adding two sources to a visually inspected source sample where one of the added sources is labelled a cluster candidate. The mean purity for the visually inspected samples is listed in Tables 4 and 5, with errors calculated from the binomial distribution's 16th and 84th percentiles.

A new source sample can be constructed by combining sources from different samples such as those listed in Tables 4 and 5. The purity of the new sample can be estimated as the average of each existing sample's purity weighted by the number of sources from each existing sample in the new one. For example, the purity of the

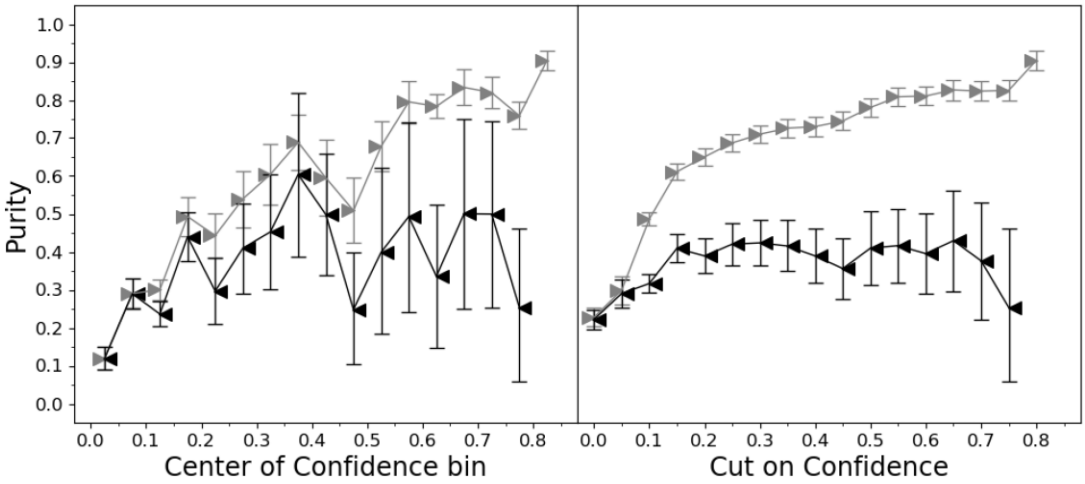


Figure 12. The purity of different source samples created from the XXL North source catalogue. See Section 4.4 for the definition of purity used in this work. Source samples are produced from the entire XXL North catalogue (grey triangle pointing to the right) and by excluding the $C1$ and $C2$ sources (black triangle pointing left). Sources are selected by binning on confidence value (left) and by selecting sources with a confidence value above some cut (right).

combined $C1C2$ sample is given by the average of the $C1$ and $C2$ sample purities (0.89 ± 0.03 and $0.64^{+0.10}_{-0.11}$, respectively) weighted by the number of $C1$ and $C2$ sources in the new $C1C2$ sample (139 and 109, respectively). The purity of the $C1C2$ sample is thus 0.78 ± 0.05. To take into account uncertainties on the measured purities, we use a Monte Carlo approximation sampling each measured purity before forming the weighted average.

We report the mean purity calculated in this way for samples produced from the full source catalogue (including $C1$ and $C2$ sources) and the non-$C1C2$ sources, binned on confidence value and by selecting sources with a confidence value above some cut (Fig. 12). The results show a clear increase in purity for samples created by selecting sources from the full XXL catalogue (including $C1$ and $C2$ sources) above a higher cut on confidence values. This is to be expected as those sources with a higher confidence value are more likely to be a $C1$ or $C2$, i.e. an unambiguous cluster detection that is very likely to be labelled as a cluster candidate by our visual inspection.

For the visually inspected samples of non-$C1C2$ sources produced by binning on confidence value there is an initial trend to higher purity with higher confidence values, but the low number count of sources in each bin at mid- to high-confidence values results in a significant increase in uncertainty for the reported purity value. Similarly, the samples of non-$C1C2$ sources selected based on a cut on confidence show an initial increase in purity with confidence cut before levelling off. The samples produced with the highest confidence cuts contain few sources and so have a significant uncertainty on their purity. The low number of non-$C1C2$ objects at high confidence is simply a consequence of the fact that any high-confidence source must be very similar to $C1$ and $C2$ sources and so is likely to be one itself.

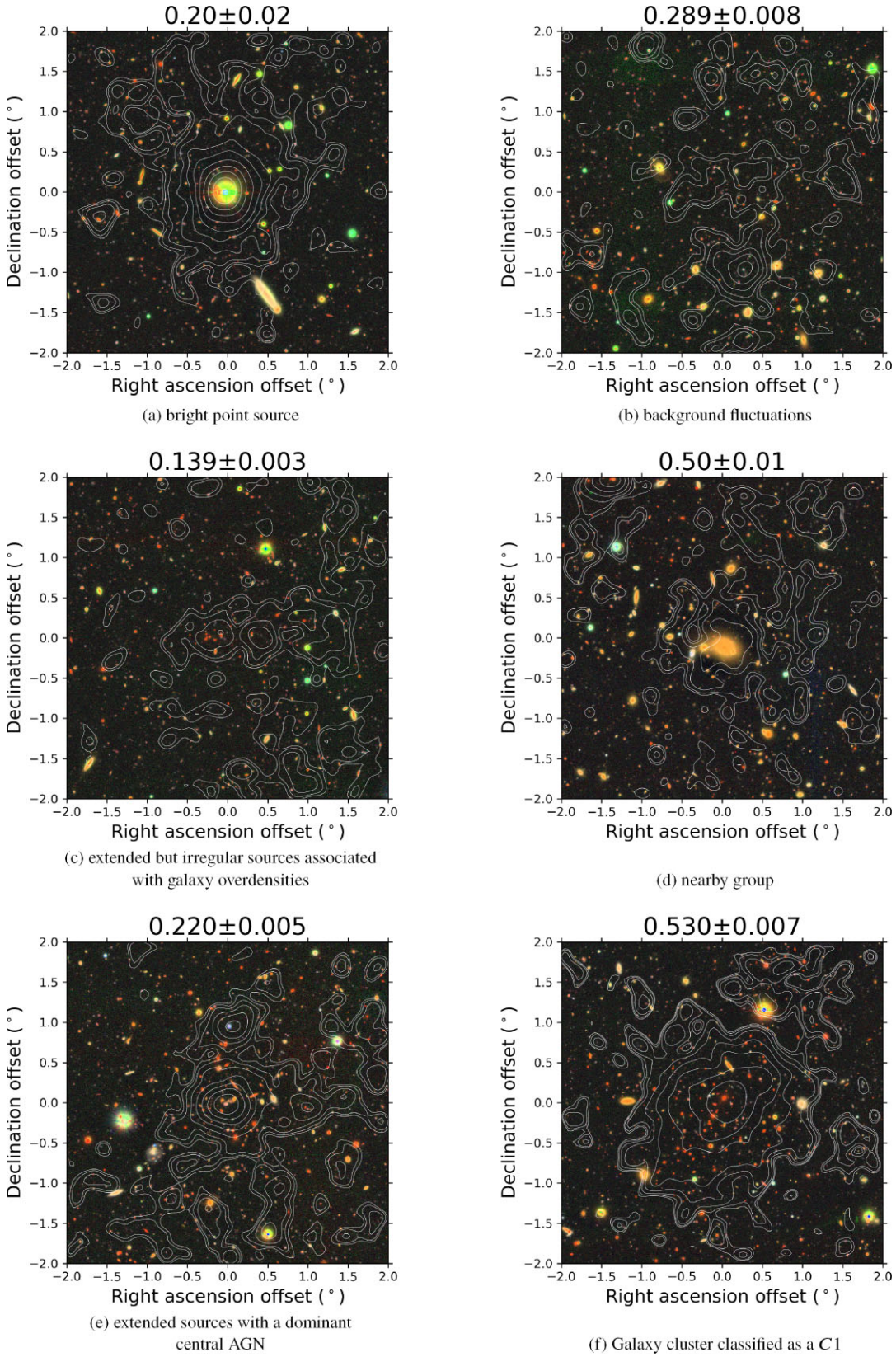
4.5 Sample selection

Given these results we define a new source sample produced by selecting sources from the full XXL North catalogue with a confidence value above 0.1. This sample contains 623 sources of which 136 and 89 were labelled as a $C1$ or $C2$ source, respectively, by XAMIN (the full $C1$ and $C2$ samples contain 139 and 109 sources, respectively, Table 6). Of the 623, 530 selected sources were visually inspected (131 $C1$'s, 85 $C2$'s, and 314 non-$C1C2$'s), the remaining 53 were excluded due to missing or bad data (5 $C1$'s, 4 $C2$'s, and 95 non-$C1C2$'s). A total of 271 of the 530 inspected sources were identified as cluster candidates, 120, 56, and 95 $C1$'s, $C2$'s, and non-$C1C2$'s, respectively. The full GP selected sample (including the 93 sources that could not be inspected) has an estimated mean purity of 0.47 ± 0.02 i.e. we estimate that 280 of the 623 sources would be classified as a cluster by our visual inspection if they were all examined. The sample selected by applying a cut on confidence of 0.1 omits 3 and 20 sources labelled as a $C1$ or $C2$, respectively. Of the 23 omitted sources all were visually inspected with 0 $C1$ and 8 $C2$ sources identified as cluster candidates.

Having visually inspected 314 non-$C1C2$ sources with confidence values above 0.2 that have optical HSC data available, we find that they can be broadly sorted into five categories: (i) bright point sources with a clear optical counterpart such as a star ($\sim$ 42 per cent, Fig. 13a); (ii) background fluctuations in the X-ray image that were combined into a broad, flat flux distribution by XAMIN's wavelet filtering ($\sim$ 32 per cent, Fig. 13b); (iii) extended but irregular sources associated with galaxy overdensities ($\sim$ 19 per cent, Fig. 13c); (iv) nearby groups where the X-ray emission appears to be dominated by the halo of the brightest galaxy ($\sim$ 6 per cent, Fig. 13d); and (v) extended sources with a dominant central AGN ($\sim$ 1 per cent,

Table 6. The number of $C1$, $C2$, and non-$C1C2$ sources in the North XXL catalogue selected by the GP (having a confidence value larger than 0.1).

XAMIN classification	GP selected (N)	Not GP selected (N)	Total (N)
$C1$	136	3	139
$C2$	89	20	109
Non-$C1C2$	398	23,228	23,626



(a) bright point source

(b) background fluctuations

(c) extended but irregular sources associated with galaxy overdensities

(d) nearby group

(e) extended sources with a dominant central AGN

(f) Galaxy cluster classified as a $C1$

Figure 13. 4×4 arcmin cutouts of XXL sources indicative of the types of sources for which the GP assigned confidence values over 0.2 (Section 4.4). The title for each example source is the confidence value assigned to it by the GP when trained on the North XXL X-ray source catalogue. The images are constructed from *g*-, *r*-, and *i*-band HSC observations (H. Aihara et al. 2018) following the method described in R. Lupton et al. (2004). X-ray contours are produced from the XXL North X-ray mosaic in the $0.5 - 2.0$ keV band (Paper XX) and smoothed by a Gaussian kernel with a standard deviation of 5 arcsec.

Fig. 13e). See Fig. 13 for examples of sources belonging to each of these categories.

After excluding point sources and background fluctuations, the dominant category of cluster candidates are the irregular extended sources. These appear to be clusters whose emission does not resemble a smooth β-model surface brightness distribution, either because of clear substructure and multimodality, or due to the limitations of the data quality. We interpret the former as dynamically younger systems compared to those systems whose X-ray surface brightness profiles are indicative of being dominated by emission from a single virialized halo.

5 DISCUSSION

5.1 GP-selected sample

Given all of the above, the first question to ask is how effective is the GP at identifying cluster candidates, particularly in comparison with the standard XXL $C1C2$ classification. We found that a sample selected on the basis of a GP-assigned confidence value above 0.1 included the vast majority, 136 and 89, of the original 139 $C1$ and 109 $C2$ sources, respectively. Fig. 3 illustrates that this success is insensitive to the choice of fields used for training and testing (i.e. the vast majority of $C1$ sources would be selected by a cut of 0.1 on either axis of either plot). The purity of a GP-selected sample is 0.45 ± 0.03, this is lower than that of the conventional $C1$ and $C2$ selection, 0.89 ± 0.03 and $0.64^{+0.10}_{-0.11}$. This implies that a GP-selected sample is less suitable for applications such as cosmological analyses, where a high purity is critical. The GP selection has the benefit of identifying a set of cluster candidates that are missed by the standard $C1$ and $C2$ selections, with an estimated purity that is sufficiently high to make feasible more detailed follow-ups, particularly if these can leverage existing (survey) data in other wavebands rather than requiring new observations.

In principle, one could obtain a larger sample at the expense of a lower purity simply by expanding the $C1C2$ selection in the EXT and EXT_STAT plane in the original XXL analysis. Based on the distribution of confidence values in the EXT and EXT_STAT plane (Fig. 2), along with our visual inspection results (Table 5) we can estimate the purity of such an extended $C1C2$ sample (following the method described in Section 4.4). Potential extended $C1C2$ source samples were identified by extending the EXT and EXT_STAT criteria to contain a number of sources within $\pm$10f the GP s-lected sample (613 to 633 sources) and their purity calculated. The highest purity is achieved by extending the EXT and EXT_STAT criteria to 4.48 arcsec and 3.52, respectively, the new extended $C1C2$ sample contains 615 sources similar to the 623 of the GP selected sample. The new extended $C1C2$ sample has an estimated purity of 0.39 ± 0.02 (calculated as described in Section 4.4) compared to 0.45 ± 0.03 for the GP-selected sample. It is simply not practical to replicate the confidence selected source sample (Fig. 2) nor the CAMIRA or GAMA matched source samples (Fig. 10) in size and purity by selecting sources based on their measured EXT and EXT_STAT properties – too many other sources would be included.

What are the X-ray characteristics of these higher confidence objects that were not selected by the $C1$ and $C2$ cuts? In general, the X-ray emission appears different to that of $C1$ clusters, which are identified by a classification scheme that was optimized to select regular β-model clusters. The implication is that $C1$s are dynamically mature clusters with the X-ray emission dominated by that from a virialized ICM. It appears that the $C1$ (and to a large extent the $C2$) sample are highly complete with respect to these bright virialized systems as we do not find them outside the $C1$ and $C2$ samples when visually inspecting sources selected by the GP (see Section 4.4). The corollary of this is that the extra non-$C1C2$ objects that we identify with a high attributed confidence value are probably less dynamically mature. Given that such objects should increasingly dominate at higher redshifts at any given mass, the GP could well be identifying systems that are useful in probing the evolution of clusters and its impact on the cluster galaxy population, particularly at higher z (greater-than ~ 0.7).

5.2 Providing additional information to the GP

One advantage of the GP (and other ML classifiers) over the simple two-parameter $C1$ and $C2$ selection is that the GP is more flexible and extensible to incorporate a wide range of additional information from beyond the XXL catalogue. In a traditional GP when the training set has binary labels, combining labels from different sources is problematic. For example, a binary label is not able to capture the information that a $C2$ source associated with a CAMIRA cluster is more likely to be real cluster than a $C2$ source without that association. Our adaptation of the GP to include uncertain labels enables us to utilize more complex training sets where there are different amounts of information about different sources, resulting in a varying degree of certainty in the labels. For instance, our approach would enable us to assign a higher initial probability than the default value if they were associated with a CAMIRA cluster. The assigned probabilities could also be changed to reflect the results of visual inspection by experts or spectroscopic followup. Such enhancing of the training labels would require a detailed investigation to avoid the introduction of unwanted biases by the new information.

It may be tempting to completely decouple the labels from the X-ray properties by assigning labels to the X-ray sources based only on their association with CAMIRA clusters. The problem is that (as seen in Fig. 10) were we to do this, $C1$ sources that are not matched to CAMIRA clusters would be labelled as not a cluster in the training set, despite the very high likelihood that they are real clusters. The GP (or any other ML binary classifier) would then be unable to separate clusters from non-clusters because it has been trained to label $C1$ clusters without CAMIRA matches as not a cluster, despite them having the X-ray characteristics associated with a robust cluster detection. The optimal approach is to combine the information from different wavelengths into initial probabilities of a positive label, precisely as enabled by our adapted GP classifier.

In addition to enhancing the prior information for labelling the training set, information from other wavelengths can be used to enhance the input data by providing additional parameters for each source. As long as the additional information can be described in the form of a vector (either an individual value or multiple values), it can be given to the GP as part of the description of a source. For example, summary statistics of the distributions of colours of optical sources within a specified distance of an X-ray source could be used to extend the description of that source beyond the X-ray properties alone.

While it is a strength of the GP and other ML models that they can use a large number of parameters from different wavebands to classify objects, this brings with it additional complexity in characterizing the selection function. To accurately model the selection function would require simulations that were sufficiently realistic so as to correctly describe the multifaceted appearance of clusters across all of the source properties that are input to the GP. Consequently,

although the GP can potentially identify a wide range of clusters, this complexity may limit its usefulness when the selection function needs to be precisely known, e.g. cosmological studies.

5.3 Applying the GP beyond XXL

In this paper, we have demonstrated the feasibility of our adapted GP classifier on the XXL source catalogue. A natural continuation of this would be to apply the technique to other, larger, X-ray survey data sets. For example, the X-CLASS survey (N. Clerc et al. 2012) uses the same detection and classification pipeline as XXL so would seem to be a straightforward application of our method. However, unlike XXL, X-CLASS is not a dedicated survey, focusing on detecting sources serendipitously in the outer parts of *XMM–Newton* observations. Given the subtle but unresolved differences we found when training the GP on the North and South XXL fields, it is possible that the greater inhomogeneity (due to different observing modes, filters, absorbing column, and Galactic foreground emission) of the X-CLASS X-ray data may cause problems when training the GP.

X-ray catalogues derived from the *eROSITA* all sky survey (P. Predehl et al. 2021) may provide more suitable data sets on which to apply our method due to the relative uniformity of the data. Even then, there is variation in the depth of the survey over the sky, for example, the data are much deeper at the poles of the satellite's orbit than at the equator. It is possible that the GP may perform better if the data were split into regions based on depth and/or absorbing column and Galactic foreground.

The true power of ML techniques, including the GP, will likely lie in the combination of *eROSITA* with all sky (or near all-sky) optical, near infrared, and SZ surveys such as those by the Vera Rubin Observatory (Ž. Ivezić et al. 2019) and *Euclid* (Euclid Collaboration 2024), amongst others. In particular, *Euclid*'s combination of imaging and spectroscopy could be of significant benefit in supplementing the labelling or measured properties of a large number of X-ray sources in order to train the GP. Alternatively, the next-generation of spectroscopic surveys, for example Wide-Area VISTA Extragalactic Survey (S. P. Driver et al. 2019), 4MOST Hemisphere Survey (E. N. Taylor et al. 2023), Multi-Object Optical and Near-infrared Spectrograph (M. Cirasuolo et al. 2014), and the Subaru Prime Focus Spectrograph Galaxy Evolution Survey (M. Takada et al. 2014) could be used to enhance the labelling in our framework.

6 CONCLUSION

We have presented a GP binary classifier adapted to take into account the uncertainties in the labels used in training data, in order to explore how well such a classifier can be used to identify galaxy clusters from their X-ray emission. Specifically, we apply the GP to pre-existing catalogues drawn from the XXL survey.

The GP was trained using a sample of clusters and cluster candidates derived from the XXL data using the standard XXL pipeline (Paper I). Applying initial probabilities of being a galaxy cluster to every X-ray source based on its classification by the XXL pipeline, the GP successfully recovered 136 and 89 of the 139 $C1$ and 109 $C2$ sources selected by XXL, respectively. We emphasize here, this is without the GP having access to those parameters that the pipeline itself used for that classification. In other words, the GP can identify cluster candidates from signatures in the original catalogue not used by the standard XXL classification.

In addition those systems already identified by the standard pipeline, the GP selected 398 sources not selected by XXL. The full GP-selected sample of 623 sources was found to have a purity of 0.45 ± 0.03 through visual inspection. Visual inspection of selected sources indicate that those non-$C1C2$ cluster candidates selected by the GP appear to have different X-ray morphologies to those selected as $C1$ by the standard pipeline. This is unsurprising as the standard XXL selection criteria were optimized to identify the most dynamically relaxed and evolved clusters. The additional candidates identified in this work tend to have multimodal X-ray emission or at least are not dominated by a single X-ray component with a typical β-model surface brightness profile.

Both the northern and southern XXL survey fields were used to explore the sensitivity of the GP's performance to the exact choice of training data. i.e. the GP was trained on and applied to different combinations of the survey fields. This analysis demonstrated that the process did not result in overfitting to the data, but it did reveal subtle, and unresolved, differences in results when different fields were used for training. While our results were robust, these differences highlight the challenges in using complex classifiers on large data sets, where small dependencies on training data may be more prevalent, but harder to detect.

ACKNOWLEDGEMENTS

JCB would like to thank the Science and Technology council for funding via the STFC Data Intensive Centre for Doctoral Training held at Cardiff, Bristol, and Swansea.

XXL is an international project based around an XMM Very Large Programme surveying two 25 deg^2 extragalactic fields at a depth of 6×10^{-15} erg cm^{-2} s^{-1} in the [0.5–2] keV band for point-like sources. The XXL website is http://irfu.cea.fr/xxl.

This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol – http://www.bris.ac.uk/acrc/.

The Saclay group acknowledges long-term support from the Centre National d'Etudes Spatiales (CNES). SB thanks CNES and CNRS for support of post-doctoral research.

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from the Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

This paper is based [in part] on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by Subaru Telescope and Astronomy Data Center (ADC) at NAOJ. Data analysis was in part carried out with the cooperation of Center for Computational Astrophysics (CfCA) at NAOJ. We are honoured and grateful for the opportunity of observing the Universe from Maunakea, which has the cultural, historical, and natural significance in Hawaii.

This paper makes use of software developed for Vera C. Rubin Observatory. We thank the Rubin Observatory for making their code available as free software at http://pipelines.lsst.io/.

The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg, and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under grant no. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation grant no. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT, and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is http://www.gama-survey.org/. This is based on observations obtained with *XMM–Newton*, an ESA science mission with instruments and contributions directly funded by ESA Member States and NASA.

DATA AVAILABILITY

The version of the XXL X-ray source catalogue used in this work is an early internal to XXL version of the catalogue and as such is not publicly available. A version of the XXL catalogue is planned for release by the XXL collaboration following updates to the XAMIN pipeline. The data will be shared on reasonable request to the XXL collaboration.

The code used within this work is available in the GitHub repository: https://github.com/CaleBaguley/The_XXL_Survey_LV.

REFERENCES

Abell G. O., 1958, ApJS, 3, 211
Adami C. et al., 2018, A&A, 620, A5 (XXL Paper XX)
Aihara H. et al., 2018, PASJ, 70, S4
Aihara H. et al., 2022, PASJ, 74, 247
Baldry I. K. et al., 2018, MNRAS, 474, 3875
Bhargava S. et al., 2023, A&A, 673, A92 (XXL Paper LI)
Bishop C. M., Nasrabadi N. M., 2006, in Jordan M., Kleinberg J., Schoelkopf B., eds, Information Science and Statistics, Vol. 4. Pattern Recognition and Machine Learning. Springer, New York, NY
Bleem L. E. et al., 2015, ApJS, 216, 27
Chiappetti L. et al., 2018, A&A, 620, A12 (xXL Paper XXVII)
Cirasuolo M. et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Proc. SPIE Conf. Ser., Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. SPIE, Bellingham, Washington, p. 91470N
Clerc N., Sadibekova T., Pierre M., Pacaud F., Le Fèvre J.-P., Adami C., Altieri B., Valtchanov I., 2012, MNRAS, 423, 3561
Driver S. P. et al., 2011, MNRAS, 413, 971
Driver S. P. et al., 2019, The Messenger, 175, 46
Ebeling H., Edge A. C., Allen S. W., Crawford C. S., Fabian A. C., Huchra J. P., 2000, MNRAS, 318, 333
Ebeling H., Edge A. C., Bohringer H., Allen S. W., Crawford C. S., Fabian A. C., Voges W., Huchra J. P., 1998, MNRAS, 301, 881
Euclid Collaboration, 2024, A&A, 697, A1
Faccioli L. et al., 2018, A&A, 620, A9 (XXL Paper XXIV)
Garrel C. et al., 2022, A&A, 663, A3 (XXL Paper XLVI)
Giles P. A. et al., 2016, A&A, 592, A3 (XXL Paper III)
Gioia I. M., Henry J. P., Maccacaro T., Morris S. L., Stocke J. T., Wolter A., 1990, ApJ, 356, L35
GPy since, 2012, GPy: A Gaussian Process Framework in Python. Available at: http://github.com/SheffieldML/GPy
Hao J. et al., 2010, ApJS, 191, 254
Hasselfield M. et al., 2013, J. Cosmol. Astropart. Phys., 2013, 008
Ivezić Ž. et al., 2019, ApJ, 873, 111
Kosiba M. et al., 2020, MNRAS, 496, 4141
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Logan C. H. A. et al., 2018, A&A, 620, A18 (XXL Paper XXXIII)
Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133
Miyazaki S., Hamana T., Ellis R. S., Kashikawa N., Massey R. J., Taylor J., Refregier A., 2007, ApJ, 669, 714
Oguri M. et al., 2018, PASJ, 70, S20
Oguri M., 2014, MNRAS, 444, 147
Paananen T., Piironen J., Andersen M. R., Vehtari A., 2019, in Chaudhuri K., Sugiyama M., eds, Proceedings of Machine Learning Research, Vol. 89, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Variable Selection for Gaussian Processes via Sensitivity Analysis of the Posterior Predictive Distribution. PMLR, p. 1743
Pacaud F. et al., 2006, MNRAS, 372, 578
Pacaud F. et al., 2016, A&A, 592, A2 (XXL Paper II)
Pacaud F. et al., 2018, A&A, 620, A10 (XXL Paper XXV)
Pierre M. et al., 2016, A&A, 592, A1 (XXL Paper I)
Planck Collaboration XXIX, 2014, A&A, 571, A29
Predehl P. et al., 2021, A&A, 647, A1
Quiñonero-Candela J., Rasmussen C. E., 2005, J. Mach. Learn. Res., 6, 1939
Richards J. W. et al., 2011, ApJ, 733, 10
Robotham A. S. G. et al., 2011, MNRAS, 416, 2640
Romer A. K., Viana P. T. P., Liddle A. R., Mann R. G., 2001, ApJ, 547, 594
Rykoff E. S. et al., 2014, ApJ, 785, 104
Takada M. et al., 2014, PASJ, 66, R1
Taylor E. N. et al., 2023, The Messenger, 190, 46
Williams C. K., Rasmussen C. E., 2006, Gaussian Processes for Machine Learning. Vol. 2. MIT Press, Cambridge, MA
Williams C., Rasmussen C., 1996, in Advances in Neural Information Processing Systems 8, Max-Planck-Gesellschaft. MIT Press, Cambridge, MA, USA, p. 514
Willis J. P. et al., 2021, MNRAS, 503, 5624
Wittman D., Dell'Antonio I. P., Hughes J. P., Margoniner V. E., Tyson J. A., Cohen J. G., Norman D., 2006, ApJ, 643, 128

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