Author:
Baguley, Cale

Title:
On the selection of galaxy clusters using an adapted Gaussian Process binary classifier

On the selection of galaxy clusters using an adapted Gaussian Process binary classifier.

Jake. Cale. Baguley

A thesis submitted to the University of Bristol
in accordance with the requirements of the degree of
Doctor of Philosophy
in the Faculty of Science

School of Physics,
H.H. Wills Physics Laboratory,
Tyndall Avenue,
Bristol,
BS8 1TL

23^{rd} of June 2023

$\sim 40,000$ words

Abstract

The aim of this work is to select galaxy cluster candidates from the XXL X-ray source catalogue by applying a supervised machine learning based selection method. The biggest hurdle when applying supervised machine learning selection methods to astrophysical catalogues is the need for a sufficiently large, perfectly labelled set of training data that accurately reflects real data. The creation of such training sets for astrophysics is a highly involved complex problem. This work presents an alternative approach. By adapting the machine learning model to account for uncertainty on the training labels we remove the need for a perfectly labelled training set, instead requiring one that can be created by labelling a source catalogue based on the purity of existing source samples. We describe in chapter 3 the adaption of a Gaussian process binary classifier to account for uncertainties on the training labels.

The adapted classifier was separately trained on the North and South XXL X-ray source catalogues labelled based on the existing XXL cluster selection samples (chapter 4). To avoid the model simply re-learning the existing selection criteria those measured source properties used by XXL to select galaxy clusters were not provided to the model. The capability of the model with respect to cluster selection was assessed using three methods. We first made use of a simulated XXL catalogue with labelled galaxy cluster detections, but it was found to insufficiently recreate the real XXL catalogues to be of use. A set of XXL sources with evidence of an increased likelihood of being a galaxy cluster detection based on their association with an optically-selected cluster, showed the model is able to distinguish such sources from the general population. Finally we visually inspect a subset of sources within the North catalogue to determine a reliable cluster selection criteria based on the output of the ML model. The cluster sample produced contains 623 sources from the North catalogue. Of the 248 sources previously selected by XXL 225 were recovered by this sample. The sample was found to have a purity of $0.45^{+0.03}_{-0.03}$ and contain an expected 280 cluster candidates, 101 of which were not previously selected by XXL. The new candidates were often found to differ in their X-ray morphologies from those previously selected by XXL, tending not to be dominated by a single X-ray component that follows a β-model surface brightness profile.

Interpretation of the model's selection criteria (chapter 5) showed it learnt to identify clusters based on a sources count rates measured by separately fitting an extended and point source emission model. We note

that while the output of the binary classifier was robust to being trained on either the North or South XXL source catalogues, our investigation into the selection criteria showed a subtle and unresolved difference in behaviour, possibly due to differences in the properties of the two fields (e.g. differences in Galactic column and foreground, or time-varying instrument calibration or background characteristics). Overall, we find that the classifier is complementary to the standard XXL processing. However, the advantage of the Gaussian process is that it allows for additional information (e.g. from other wavebands) to be incorporated into the uncertainties on the labels used for training, or in the classification process (chapter 6).

Acknowledgments

I would like to thank my supervisors, Malcolm Bremer and Ben Maughan for their support and guidance throughout this PhD. I am particularly grateful for their patience as I learnt how to write an academic paper. I would also like to thank Carl Henric Ek for providing his expertise on machine learning and Gaussian processes.

To the University of Bristol's astrophysics group, thank you for providing a supportive and friendly work environment. You have all made coming into work a genuinely enjoyable experience.

Finally I would like to thank my parents, Paul Baguley and Jane Fielder, and brother Torin Baguley, without their love and support I would not have been able to achieve what I have. I am particularly thank full for their accommodating me during COVID when I was working from home. I would also like to thank my father for checking my spelling and grammar throughout this work.

COVID-19 statement

The COVID pandemic resulted in progress towards the thesis being slowed and made more difficult, both during and for a period after the lock-downs. In particular, the remote working reduced access to my supervisors and stopped me from benefiting from the expertise of other students slowing the rate at which I was able solve problems. The impact of COVID on other members of the XXL collaboration also delayed the production of data products used within this work. In the absence of COVID restrictions had time allowed one or several of the topics below would have likely been included in this work;

- Training the GP model on both hard and soft band X-ray data and analysis of the results.
- Testing the use of additional information to improve the training labels.
- Post processing of the GP selected catalogue to further improve the samples purity.
- The application of the GP model to select galaxy clusters from the X-Class survey catalogue.

Declaration

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. This work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree. Any views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol. The dissertation has not been presented to any other university for examination either in the United Kingdom or overseas.

I note here any contributions to the work from collaborators.

- Both the XXL and simulated XXL catalogues used within this work were produced by the XXL collaboration and provided prior to public release.
- The Hyper Suprime-Cam collaboration provided the most recent `CAMIRA` optically selected galaxy cluster catalogue for use within this work.

Contents

List of Figures

List of Tables

1

Introduction

In the moments after the big bang the distribution of matter within the Universe was not uniform, instead random quantum fluctuations led to over and under dense regions of space. The largest overdensities within this density field seeded the formation of galaxy clusters through gravitational collapse. By its very nature the process of galaxy cluster formation and evolution is intrinsically linked to the initial state and evolution of the universe as a whole, making the study of galaxy cluster populations integral in developing our understanding of cosmology. Galaxy clusters are also important for the study of galaxies, owing to the internal environment of a cluster significantly impacting the process of galaxy formation and evolution. To conduct the studies of galaxy cluster populations necessary for cosmological research, along with their affect on galaxy formation and evolution, astrophysicists require comprehensive and reliable galaxy cluster catalogues with well understood selection functions. Driven by this need, surveys dedicated to the detection and cataloguing of galaxy clusters have been carried out over a wide range of wavelengths. These range from the radio regime (selecting clusters via the Sunyaev-Zeldovich effect) through optical (via their galaxy populations or weak gravitational lensing signals) to X-rays (via emission from their hot Intra Cluster Medium; ICM).

The first surveys for galaxy clusters were performed using optical observations, with clusters identified by eye as overdensities of galaxies in photographic images (Abell, 1958). With the advent of digital computing galaxy cluster surveys progressed to an automated algorithmic approach to detection and characterisation (Shectman, 1985; Lumsden et al., 1992). While such automated approaches are tailored to

the telescope and wavelength at which the galaxy cluster survey is conducted, they follow the same general structure. First sources are identified from observations using some source detection algorithm, such as source extractor (Bertin & Arnouts, 1996). Detected sources are promptly characterised by measuring some set of their physical properties. Sources are subsequently selected as galaxy cluster candidates, if their measured source properties satisfy some selection criteria. Not only did the use of automated approaches increase the rate at which observational data could be processed and the number galaxy clusters identified, but the use of selection criteria made the selection process highly interpretable. This interpretability simplifies the galaxy cluster selection function, a key requirement for galaxy cluster based cosmological studies.

While efficient enough for past surveys, existing detection and characterisation algorithms are prohibitively computationally expensive when applied to large data sets. This limitation of existing algorithms is becoming increasingly apparent when considering the volumes of data expected from recent and near future cluster surveys such as eROSITA (Predehl et al., 2021) and Euclid (Laureijs et al., 2011). The prospect for Machine Learning (ML) classifiers to produce large, reliable source samples with greater efficiency than past selection methods makes ML classifiers increasingly appealing to astronomers. Examples include: classification of variable stars from time series data (Richards et al., 2011); galaxy cluster detection from the Sloan Digital Sky Survey (Hao et al., 2010) and the use of convolutional neural networks to select galaxy clusters from the X-Class serendipitous X-ray cluster survey (Kosiba et al., 2020).

While the main driver behind the adoption of ML is the efficient processing of the large data sets produced by modern surveys, ML can also be beneficial in extracting new information from older surveys. ML models are designed to identify and exploit high dimensional trends in data that, due to their high dimensionality, are not easily identified by the scientist. Combining this ability of ML with those measured source properties not previously used when selecting interesting objects from a catalogue, has the potential to identify novel interesting sources. In the case of XXL this means that ML could be used to identify novel galaxy cluster candidates based on source properties beyond the three currently used.

The main aim of this work is to apply a ML model to to identify novel galaxy cluster candidates from the XXL X-ray source catalogue that were not selected by the standard processing pipeline, XAMIN (Faccioli et al., 2018). This is achieved first by adapting a Gaussian Process (GP) binary classifier such that it can account for uncertainties on the training labels (chapter 3). The model is trained on the XXL X-ray catalogue before being tested. The testing process makes use of simulated data, a sample of XXL sources with X-ray independent evidence of being a galaxy cluster, and visual inspection of XXL sources (chapter 4). A detailed investigation into the selection criteria used by the ML model to select cluster candidates is conducted to identify the limitations of the model when applied to XXL (chapter 5).

The remainder of this chapter gives an overview of galaxy clusters with specific emphasis on the detection methods used by galaxy cluster surveys. This is followed by a discussion of ML models with emphasis on object classification in the context of astrophysical surveys.

1.1 Galaxy Clusters

The end goal of any galaxy cluster survey is to produce a reliable catalogue of galaxy clusters with which astronomers can conduct research. Such research includes studies of: the evolution of galaxy clusters; how the environment within a cluster affects its constituent galaxies; and cosmology. To produce a galaxy cluster survey capable of achieving this goal, it is important to understand not only the properties of clusters and how they appear in observations, but also the requirements on a catalogue imposed by cluster research.

1.1.1 Galaxy Cluster Formation and Evolution

Developing an accurate model of galaxy cluster formation is key to not only understanding their physical properties, but the intricate relations between those properties and the cosmological evolution of the universe. The creation of such a model first requires a description of the initial distribution of matter in the universe. Given this distribution, analytical and computational models of gravitational collapse can subsequently be used to describe the formation of galaxy clusters and their evolution from unrelaxed dynamic systems to a more relaxed steady state. Further additions to the model are needed to account for the affect of the environment around clusters and the and non-gravitational processes within them, such as radiative cooling and AGN feedback.

In order for galaxy clusters to form under gravitational collapse, the initial distribution of matter in the universe must be non-uniform at a local scale. A perfectly uniform distribution of matter having no overdense regions from which to form a galaxy cluster. This non-uniform distribution is produced by random quantum fluctuations and is generally modelled by describing the matter density field as a random Gaussian field. One benefit of modelling the initial density distribution as a random Gaussian is that the probability of a peak in the density field, smoothed over a given scale, having a given mass is described by a Gaussian (Kravtsov & Borgani, 2012).

The nature of the random density field is such, that the exact evolution of an overdensity into a galaxy cluster can not be described analytically. Despite this it is instructive to consider the evolution of an isolated, perfectly spherical overdensity with some uniformly distributed mass content. Initially the expansion rate of the universe dominates the evolution of the overdensity, growing it from a quantum fluctuation to the macroscopic scale. During this process of expansion the inward force of gravity due to the excess

mass relative to the mean density of the universe, causes the overdensity to grow slightly slower than the universe around it. Over time the reduced growth rate increases the contrast in density between the universe and the overdensity. This in turn further reduces the growth rate of the overdensity with respect to the universe. Eventually the expansion rate of the overdensity reaches zero before inverting as it begins to collapse under gravity. The time taken to reach this point is the turn around time, and for a universe with $\Omega_\Lambda = 0$, is equal to the free fall time of the spherical overdensity at its maximum size (Kravtsov & Borgani, 2012).

Having detached from the expansion of the universe the overdensity collapses under self gravitation, forming a galaxy cluster. As the galaxy cluster collapses it converts gravitational potential energy to kinetic energy. The increased kinetic energy produces an outward gas pressure slowing the rate of collapse. After a period roughly equal to the turn around time, the rate of collapse reaches zero, the outward gas pressure induced by the clusters internal kinetic energy being equal but opposite to the inward pressure from gravity. In this steady state the system satisfies the virial theorem and is considered virialised.

Modelling the formation of a galaxy cluster in this simplified way, leads to the self similar model of galaxy cluster evolution. The self similar model stipulates that the properties of a cluster are only dependent on its mass and redshift. This directly informs the creation of X-ray scaling relations among others, that link a clusters X-ray observable properties to its mass (for a recent review see Lovisari & Maughan, 2022). By measuring the observed properties of a virialised cluster it is hence possible to not only infer its mass, but given a sufficiently large population of clusters, probe the cosmological evolution of the universe (Allen et al., 2011; Pratt et al., 2019).

While informative, the spherical collapse model is an imperfect approximation, unable to account for a number of complicating factors during cluster formation and evolution. While modifications can be made to the spherical model to account for more complex ellipsoidal systems, modelling the evolution of realistic overdensities is not analytically possible, requiring a computational approach instead. Simulations of galaxy cluster formation and evolution include the Millennium (Springel et al., 2005b) and EAGLE (Schaye et al., 2015) simulations. The exact details of how such cosmological simulations work are beyond the scope of this work (See Angulo & Hahn, 2022, for a recent review). A description of the main points of cluster formation found through simulations and how they differ from the spherical collapse model is included here.

The spherical collapse model treats the initial overdensity as a perfectly spherical region of space with uniform mass distribution, in reality not only is the overdensity non-spherical but the internal distribution of matter is non-uniform. The affect of this non-uniform density, is that there is no longer a singular time at which the overdensity stops growing and begins to collapse. The collapse of the system follows

a hierarchical process instead. The highest density peaks within the overdensity evolving quickest and beginning to collapse first. Initially these collapsing density peaks are isolated, but as they grow and more regions within the overdensity begin to collapse, they merge under gravity, eventually coalescing into a singular collapsed cluster. Over time the inhomogeneity due to the initial matter distribution will abate, the galaxy cluster moving from a dynamical young state towards a relaxed virialized one.

In further contrast to the spherical collapse model galaxy clusters are not isolated objects, instead surrounded by the large scale structure that makes up the cosmic web. In addition to galaxy clusters the cosmic web can be broken down into a number of components. Voids are regions of space with reduced matter content evolving from initial underdensities in the matter density field. Filaments and walls are bridges of matter (including galaxies) connecting neighbouring galaxy clusters. The combined gravitational force of neighbouring clusters, acting to draw matter onto the line or plane drawn by joining the clusters to form a filament or wall respectively. As such structures gain mass, their own gravitational force increases, further enhancing this effect. The material that makes up a filament or wall flows along it and onto the clusters due to their gravitational force. This irregular flow of material onto a galaxy cluster producing bow shocks as the infalling material interacts with the ICM (Markevitch & Vikhlinin, 2007). The effect of this process disturbs the cluster away from a virialised state towards a dynamically younger, more turbulent one.

In rare instances galaxy clusters pass sufficiently close to one another to tidally interact, or in the most extreme cases merge, such as the bullet cluster (Tucker et al., 1995). The gravitational interaction between the two clusters imparts significant amounts of kinetic energy into the system, leading to a highly disturbed dynamically younger state. The process of merging between two galaxy clusters also acts to separate a clusters galaxy and dark matter content from its ICM. This process occurs due to the negligible scattering cross section of galaxies and dark matter compared to the more strongly self interacting ICM (Tulin & Yu, 2018). Studies of the dark matter distribution during and shortly after merger events provide key insights in to dark matter self interaction.

This description of galaxy clusters has yet to discuss the non-gravitational processes present within a cluster. The baryonic component of the cluster is subject to heating and cooling mechanisms impacting the form of the ICM. It is expected that the hot ICM should cool through thermal bremsstrahlung radiation over time, leading to a cool core at the centre of a cluster (Lovisari & Maughan, 2022). This cooling is however opposed by heating mechanisms such as the energy imparted into to the ICM via AGN jets. At smaller masses supernova and galactic winds can impart sufficient energy into the ICM to also oppose cooling processes (Menci & Cavaliere, 2000). These non-gravitational processes distort the distribution and temperature of the ICM such that the cluster appears less virialised in X-ray observations.

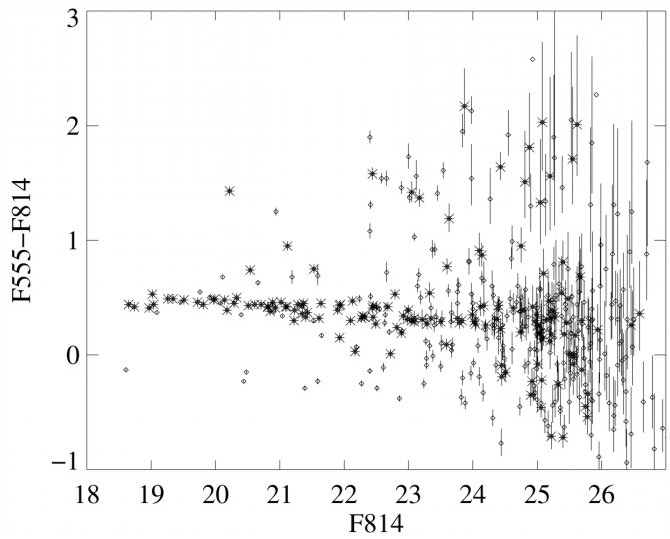


Figure 1.1: Colour-magnitude diagram of the redshift 0.231 galaxy cluster Abell 2390 (figure from Gladders & Yee, 2000) produced using two filter Huble space telescope observations of the cluster core. The galaxies magnitude measured in one of the two filters is plotted on the x-axis. The galaxies colour, measured by subtracting the galaxies magnitude in the higher frequency filter from that in the lower one such that a redder galaxy has a higher value, is plotted on the y-axis. Galaxies morphological identified as early types are plotted as asterisks, the remaining galaxies are plotted as diamonds. The galaxy clusters red sequence can be clearly seen as a strong linear trend of early type galaxies in colour-magnitude space.

1.1.2 Galaxy Cluster Properties

From the description of the formation process of a galaxy cluster it is clear that the matter content of a cluster can be broadly split into three components; galaxies, dark matter and the ICM. Each of these components can be directly or indirectly studied through observations, providing important insights into the nature of a cluster and the underlying physics that govern them.

Clusters of galaxies generally contain hundreds to thousands of galaxies with a number density significantly higher than that of the general galaxy population. A significantly higher fraction of the galaxy population within clusters are red elliptical compared to that of the field population. Such early type galaxies have a strong linear correlation between their colour and magnitude. For a galaxy cluster this presents as its red sequence, a linear correlation between the colour and apparent magnitude of galaxies contained by the cluster (figure 1.1, with a dependence on the distance to the cluster (Bower et al., 1992). The increased fraction of early elliptical galaxies within a cluster is due to a number of complex processes stemming from the cluster's internal environment. Such processes include: an increased rate

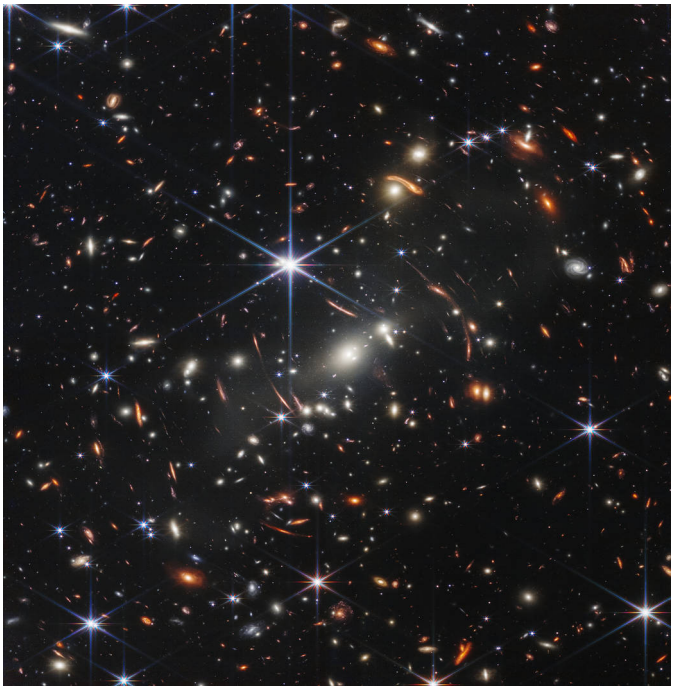


Figure 1.2: Image of the gravitational lensing of background galaxies by the redshift 0.39 galaxy cluster SMACS 0723 taken by NASA's James Webb Space Telescope. The 2.4 arcminuet across image shows a clear central brightest cluster galaxy (BCG) surrounded by diffuse emission. Those galaxies belonging to the cluster tending to appear as early types with similar colour to the BCG. The redder background galaxies show clear distortions in shape due to gravitational lensing by the cluster. Within this image the appearance of foreground stars is dominated by a set of six large and two small diffraction spikes due to the hexagonal shape of JWST's mirrors. Image credit: NASA, ESA, CSA, and STScI

of gravitational interactions and mergers between galaxies due to their higher number density (Springel et al., 2005a); ram pressure stripping of material from a galaxy as it moves through the ICM (Gunn & Gott, 1972); and the suppression of the infall rate of material onto a galaxy (van den Bosch et al., 2008). These processes most strongly influencing galaxies closer to the centre of a cluster.

The brightest cluster galaxy (BCG) is generally found close to the centre of a galaxy cluster's gravitational potential well and is surrounded by a halo of diffuse stellar emission. The BCG forms at the centre of the galaxy cluster and grows hierarchically through mergers with other galaxies (De Lucia & Blaizot, 2007). Deviations of the BCG from a cluster's centre of mass indicate an unrelaxed system and can be identified by comparison with other tracers of a cluster's centre of mass, such as the peak in the ICM's X-ray emission.

The velocity dispersion of galaxies within a cluster provide an indirect observation of the dark matter content of a cluster (Zwicky, 1933). The density of the dark matter halo being found to follow a Navarro-Frenk-White (NFW) profile (Navarro et al., 1996) through N-body gravity simulations. In addition to the the velocity dispersion of a galaxy cluster's constituent galaxies, the gravitational effect of the dark matter halo can be observed due to the gravitational lensing of coincident background sources. The strength of the gravitational lensing increases both with the mass of the cluster and how closely the unlensed line of sight to the background source passes the cluster's centre of mass. The projected image of a source when strongly lensed is distorted to form a large arc or Einstein ring. Figure 1.2 shows an example of strong lensing by galaxy cluster SMACS 0723. The presence of such strongly lensed objects is rare, the majority of coincident background galaxies being weakly lensed by a cluster. The weak lensing of a galaxy acts to distort its image, to appear slightly elongated orthogonal to the line joining said galaxy and the clusters centre of mass. Assuming that the average unlensed galaxy is spherical, stacking the images of background galaxies makes it possible to measure the average distortion due to weak lensing for a given point on the sky. Combining weak lensing measurements for multiple points on the sky produces both a gravitational shear, and convergence map, indicating the distribution of and total mass within a cluster (see Bartelmann & Schneider, 2001a, for a review). The use of weak lensing as a method to study galaxy clusters' dark matter halos is subject to uncertainties due to the presence of unlensed foreground galaxies and random deviations in the shape of galaxies from spherical.

The final component of a galaxy cluster, and the most relevant to this work, is the optically thin plasma that makes up the intra cluster medium (ICM). Having been sufficiently heated by the gravitational collapse of the cluster to form a plasma, the dominant cooling mechanism with the ICM is thermal bremsstrahlung emission as X-rays. Thermal bremsstrahlung emission is the production of a photon by a charged particle when decelerated by another charged particle (Rybicki & Lightman, 1986). Within the ionised plasma that makes up the ICM this takes the form of free electrons being decelerated when interacting with protons. The result is an X-ray spectrum such as that in figure 1.3 with a shape dependent on the temperature and electron density of the ICM. The spectral energy distribution for thermal bremsstrahlung radiation is, (Böhringer & Werner, 2010)

$$\epsilon(\nu) = \frac{16 q_e^6}{3 m_e c^2} \left(\frac{2\pi}{3 m_e k_B T} \right)^{\frac{1}{2}} n_e n_i q_i^2 g_{ff}(q_i, T, \nu) e^{\frac{-h\nu}{k_B T}}. \tag{1.1}$$

Here q_e, m_e, and n_e are the charge, mass and number density of electrons, q_i and n_i are the charge and number densities of the protons, c is the speed of light, k_B is Boltzmann's constant, T is the temperature of the plasma, and ν is the photon frequency. The function $g_{ff}(q_i, T, \nu)$ is the gaunt factor.

By assuming that bremsstrahlung emission is the dominant emission from a galaxy clusters ICM, to the exclusion of all other emission, it is possinble to derive the luminosity mass relation as described in

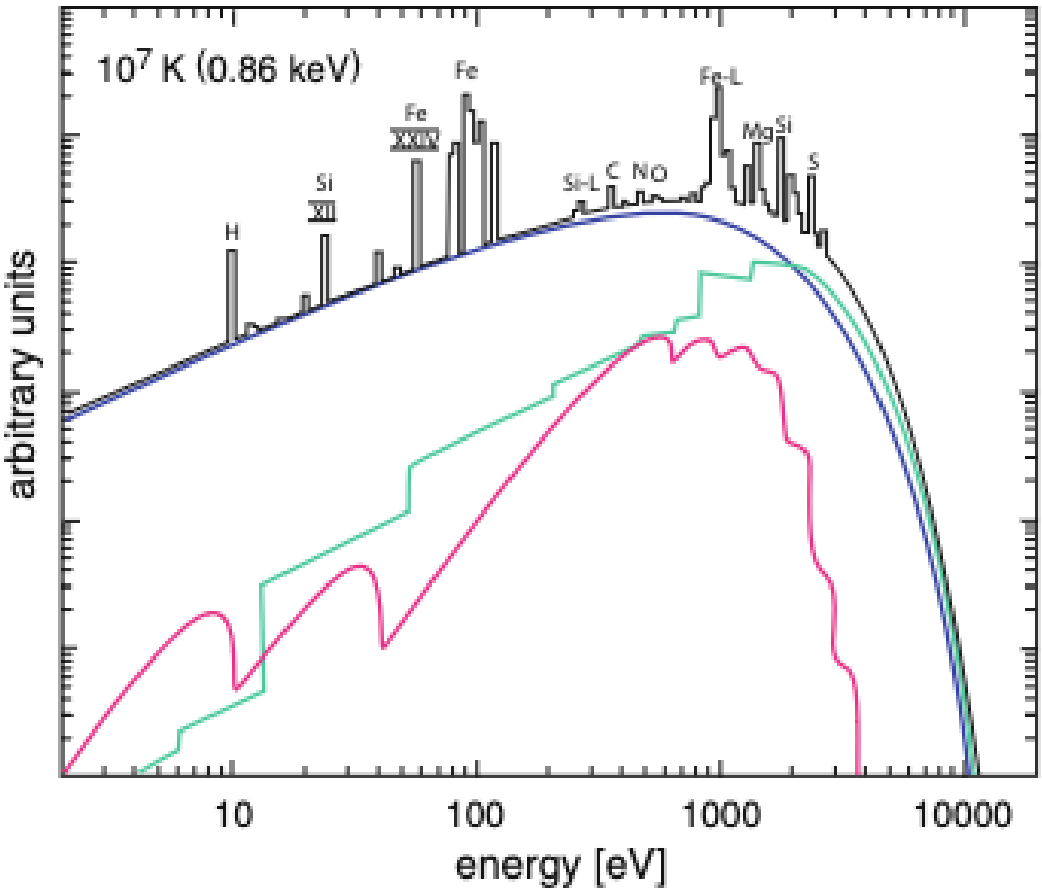


Figure 1.3: X-ray spectrum for a solar abundance plasma at a temperature of $10^7 K$ by Böhringer & Werner (2010). The X-ray spectrum is dominated by bremsstrahlung emission (blue), with aditional components from recombination radiation (green) and 2-photon radiation (red). Additionally the spectrum exhibits emission lines due to the presence of non-ionised elements. The prominent emission lines are labelled based by element (the -L label indicates transitions in to L-shell ions and the roman numerals indicating a specific atomic ion).

Lovisari & Maughan (2022). Given the equation for bremsstrahlung emission the total bolometric X-ray luminosity of a galaxy clusters ICM is proportional to the following integral over the volume of the cluster,

$$L_{X,Bol} \propto \int \epsilon dV \propto \int \rho_{gas}^2 T_X^{1/2} dV \tag{1.2}$$

Where ϵ is that for equation 1.1 and ρ_{gas} and T_X are the density and temperature of the ICM respectively (ρ_{gas} being proportional to the electron and proton densities, n_e and n_i in equation 1.1). By assuming that all galaxy clusters follow self similarity, i.e. their properties depend only on their mass and redshift, and a uniform temperature profile the integral 1.2 simplifies to,

$$L_{X,Bol} \propto \rho_{gas}^2 T_X^{1/2} R^3 \tag{1.3}$$

where R is the radius out to which the clusters luminosity is measured. For the self similar model of galaxy cluster evolution a clusters gas density can be described as $\rho_{gas} \propto M/R^3$, allowing for the removal of ρ_{gas} from the equation.

$$L_{X,Bol} \propto \frac{M}{R^3} M T_X^{1/2} \tag{1.4}$$

The ratio of the cluster mass and the cube of its radius is proportional to its density $M/R^3 \propto \rho_c(z)$. As described previously the density of a galaxy cluster evolves with the expansion of the universe, first reducing as the cluster expands before increasing when the cluster begins to collapse under gravity. This change in density can be described as proportional to the square of $E(z) = \sqrt{\Omega_M (1+z)^3 + \Omega_k (1+z)^2 + \Omega_\Lambda}$,,

$$L_{X,Bol} \propto E(z)^2 M T_X^{1/2} \tag{1.5}$$

Substituting the mas temperature relation for galaxy clusters $T_x \propto E(z)^{2/3} M^{2/3}$ (Lovisari & Maughan, 2022) gives the relation between a galaxy clusters bolometric X-ray luminosity and mass assuming self similarity

$$L_X \propto E(z)^{7/3} M^{4/3} \tag{1.6}$$

While this derivation applies to the bolometric luminosity under the assumption of self similarity the equation can be generalised by determining the powers empirically.

$$L_X = A_{LM} E(z)^{\gamma_{LM}} M^{B_{LM}} \tag{1.7}$$

Here A_{LM} is some constant and γ_{LM} and B_{LM} are the power components previously set to $7/3$ and $4/3$.

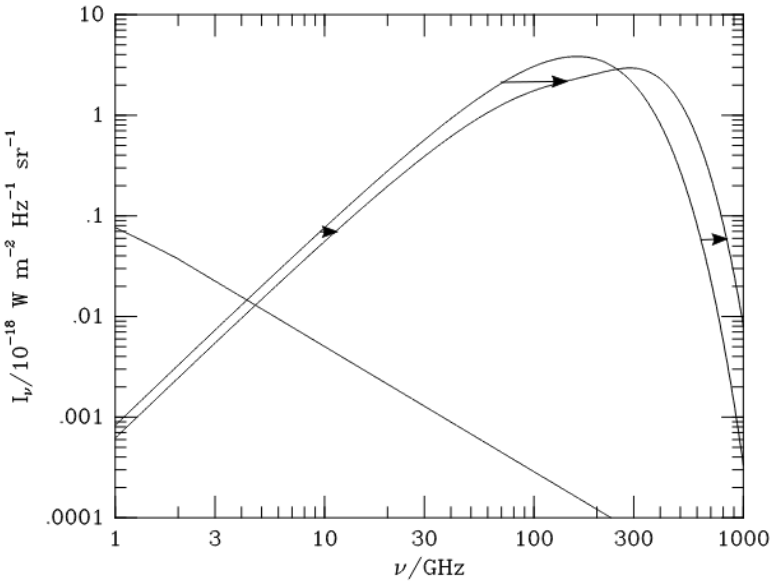


Figure 1.4: Depiction of the shift of the cosmic microwave background spectrum to higher frequencies by an exaggerated Sunyaev-Zeldovich (SZ) effect by Birkinshaw (1999). It is clear from the figure that the SZ effect for a given frequency can be vied as a fractional increase (decrease) in intensity at higher (lower) frequencies. Included for comparison is the integrated emission from the bright radio source Cygnus A observed by a telescope with a solid angle of one square degree.

X-ray observations of cluster have shown that the surface brightness profile of virialised clusters are well fit by a β profile (Cavaliere & Fusco-Femiano, 1976). The description of a cluster's surface brightness profile by a β model is purely phenomenological in nature with the profile deviating from that predicted by the model near the centre of a cluster (Ettori et al., 2013). For non-virialised galaxy clusters X-ray observations provide an insight into the structure and disturbances within the ICM. The temperature and density dependence of thermal Bremsstrahlung radiation making structures such as shocks and cooling flows differentiable from the ICM (Ryu et al., 2003).

Radio obsservations of galaxy clusters are able to map the contents of the ICM through the Sunyaev-Zeldovich (SZ) effect. The SZ effect is the inverse Compton scattering of cosmic microwave background photons by free electrons within the ICM (Sunyaev & Zeldovich, 1972). This up scattering process acts to shift the spectrum of the CMB to higher wavelengths that manifests as a fractional increase (decrease) in the apparent power of the CMB at higher (lower) wavelengths (Figure 1.4). For an isothermal cluster the apparent distortion of the CMB can be expressed as a change in temperature,

$$\frac{\Delta T_{SZE}}{T_{CMB}} = f(x) \int n_e \frac{k_B T_e}{m_e c^2} \sigma_T dl \tag{1.8}$$

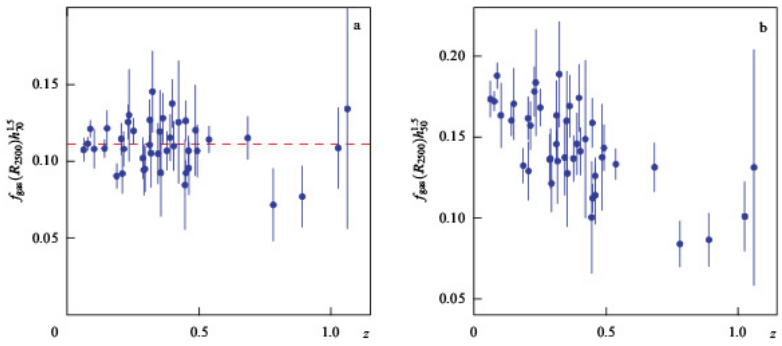


Figure 1.5: Example of galaxy cluster f_{gas} fraction as a function of redshift calculated from the Chandra X-ray data (Allen et al., 2008) using the standard ΛCDM cosmology (left) and the same calculation assuming an incorrect cosmology (right) from Vikhlinin et al. (2014). Where the f_{gas} fraction has been calculated using the standard ΛCDM cosmology f_{gas} is constant with redshift while the same f_{gas} calculated using an incorrect cosmology shows a strong redshift dependence.

(Carlstrom et al., 2002). Here $f(x)$ is the frequency dependence of the CMB spectrum for a unitless frequency x, n_e and T_e are the electron density and temperature of the cluster respectively, k_B is the Boltzmann constant, $m_e c^2$ is the electron rest mass energy and σ_T is the Thomson cross-section. The fractional change in power means that the strength of the SZ effect is independent of redshift, making it ideal for studying clusters at higher redshifts (Birkinshaw, 1999). The presence of other redshift dependent effects, such as the angular size of the cluster, do however limit the detectability of the SZ signal at increasing redshifts.

As with the thermal bremsstrahlung emission the strength of the SZ signal is dependent on the temperature and density of the ICM (T_e and n_e in equation 1.8), but does so in a physically and mathematically different way. Combining X-ray and SZ observations of a cluster makes it is possible to exploit these differences to extract additional information on the structure of the ICM. Further, due to the different dependence on redshift between the X-ray and SZ signal, combined observations can be used to measure both the physical distance to clusters and H_0 (Bonamente et al., 2006).

1.1.3 Galaxy Clusters for Cosmology

It is clear that both the properties of the galaxy cluster population and the formation of individual clusters are closely related to the cosmological formation and evolution of the universe. This close relationship makes galaxy clusters a key component in cosmological studies, with galaxy clusters having provided the first evidence of dark matter through measurements of the velocity dispersion of galaxies within the Coma cluster (Zwicky, 1933). Methods of cosmological analysis include the use of galaxy cluster redshifts determined using the redshift dependence of a clusters X-ray and SZ properties to accurately

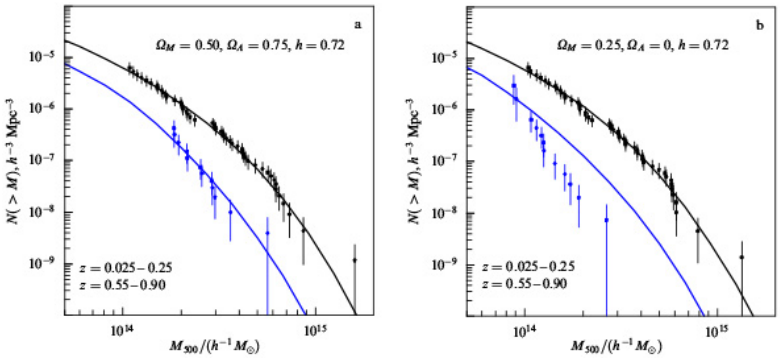


Figure 1.6: Example of the sensitivity of the cluster mass function to differing cosmological models by Vikhlinin et al. (2014). The figure depicts the predicted cluster mass function (solid line) in two redshift bins ($z = 0.025 - 0.25$, black and $z = 0.55 - 0.90$ blue), compared to observations for both redshift ranges. It is clear that the cluster mass function predicted for a cosmology close to the current accepted values (left) matches the observed distribution compared to that predicted for the cosmology (right)

measure Hubble's constant. Alternatively cosmological models can be tested based on the requirement that the f_{gas} fraction of galaxy clusters (the ratio of baryonic mass to total cluster mass) be constant with redshift (Figure 1.5). In instances where f_{gas} is calculated based on an incorrect cosmological model a strong trend with redshift can be seen. Alternatively one can make use of the cluster mass function (the number density of galaxy clusters as a function of mass) to asses different cosmological models (Fig 1.6) due to its strong dependency on cluster evolution and initial mass distribution. A full review of cluster cosmology is beyond this work (for recent reviews see Vikhlinin et al., 2014; Allen et al., 2011; Pratt et al., 2019). Here we instead describe the key requirements such research imposes on cluster surveys and catalogues.

X-ray studies of galaxy clusters have been key in providing significant constraints on cosmological models (Clerc & Finoguenov, 2022). The benefit of using X-ray observations to conduct cosmological studies is the relationship between the properties of the ICM when observed in the X-ray band and a clusters mass and redshift (Lovisari & Maughan, 2022). Such scaling relations inherently rely on the self similar model of galaxy cluster properties. The self similar model and hence X-ray scaling relations reliably describe virialised clusters. The more disturbed a cluster is, the less reliably the self similar model describes its properties, reducing the accuracy of X-ray scaling relations. The reliability of X-ray scaling relations being tied to a clusters' dynamic state, results in a need for large catalogues of approximately virialised clusters for cosmological studies. This is an important consideration when assessing the cluster candidates selected by the GP binary classifier used in this work.

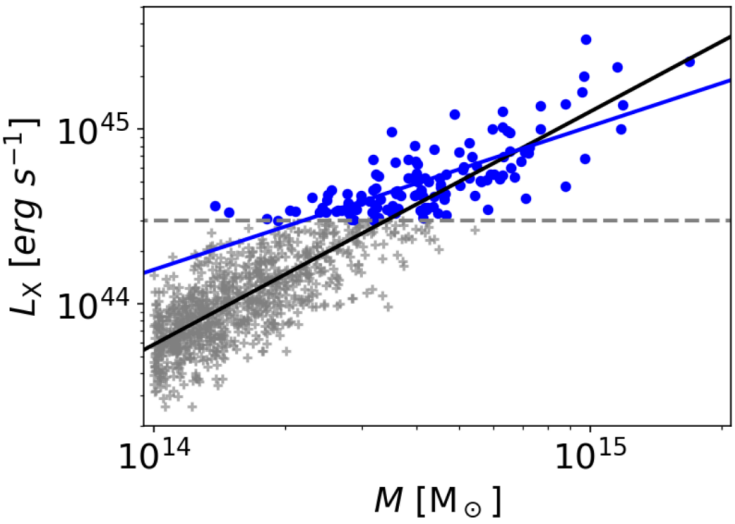


Figure 1.7: Illustration of the impact of the selection function when measuring the galaxy cluster luminosity mass relation by Lovisari & Maughan (2022). The measured luminosity mass relation (blue) differs from the true relation (black) as the luminosity selected clusters (blue dots) do not accurately replicate the true cluster population (all scatter points). In order to accurately recover the true relation it is necessary to account for the impact of selecting clusters on luminosity.

The accuracy of any cosmological studies that make use of galaxy cluster catalogues depends on knowledge of the catalogue's selection function. The selection function of a galaxy cluster catalogue describes the probability that a galaxy cluster with given properties, (i.e. mass and redshift) is included within the catalogue. Inaccurate modelling of the selection function can result in the under or over estimation of the true population of clusters with a given set of properties, systematically shifting cosmological measurements from their true values. An example of such an effect is shown in Figure 1.7 from Lovisari & Maughan (2022). For samples produced using simpler cluster detection algorithms (e.g. flux limited samples) the description of the selection function is relatively straight forward. Simpler detection algorithms however are unable to make use of all available information when identifying clusters, reducing the samples' completeness. In contrast, increasing the amount of information used by the detection algorithm improves its ability to detect clusters, but at the expense of complicating the selection function. This is of particular concern when using ML for cluster detection due to the inherent internal complexities of ML models, making determining the selection function difficult.

1.1.4 Galaxy Cluster Surveys

Galaxy cluster surveys have historically been conducted over a wide range of wavelengths with each tailored to the observational properties of a cluster at the chosen wavelength. The different observational

properties of galaxy clusters introduce a number of advantages and disadvantages for cluster survey's depending on the surveys chosen wavelength. Further surveys at each wavelength probe different components of the galaxy cluster population, as characterised by their selection function. The remainder of this subsection covers galaxy cluster surveys conducted using optical, radio and X-ray observations.

1.1.4.1 Galaxy cluster detection in optical observations

The first large scale galaxy cluster survey by Abell (1958) identified clusters from overdensities of galaxies on photographic plates. The majority of false positives when identifying clusters from overdensities of galaxies, is due to unrelated coincident galaxies projected onto the same region of the sky. Projection effects limit number density based detection methods to high richness clusters, the probability of an apparent overdensity occurring due to projection effects reducing as richness increases.

To solve the issue imposed by projection effects it is necessary to introduce additional information when identifying clusters. The conceptually simplest approach is to include a galaxies redshift, such that clusters can be identified from overdensities of galaxies within three dimensional space instead of two. While spectroscopic redshifts provide accurate measurements of a galaxy's redshift, the process of taking such measurements is observationally intensive, making it impractical for use in large scale cluster surveys. Photometric redshifts are significantly less observationally intensive at the expense of precision and the need for accurate modelling of galaxies stellar populations (Bolzonella et al., 2000). The reduced precision of photometric redshifts limits their usefulness in distinguishing clusters from projections.

Without a method for efficiently measuring precise redshifts for a large number of galaxies, it is necessary to identify an observable difference between those galaxies within a cluster and the field population. As described earlier (section 1.1.2), galaxy clusters contain a higher fraction of red elliptical galaxies compared to that of the field. The higher fraction of red elliptical galaxies within a cluster presents as a clusters red sequence (Figure 1.1). Since an apparent overdenstiy of unrelated field galaxies will not contain a red sequence, galaxy clusters can be distinguished from projections of field galaxies by the presence of a red sequence (Gladders & Yee, 2000; Rykoff et al., 2014). By exploiting a clusters red sequence to improve detection, optical cluster surveys such as `CAMIRA` (Oguri et al., 2018) are able to detect clusters with a lower richness.

While red sequence detection provides an improvement over using galaxy number density to identify galaxy clusters alone, there are still a number of limiting factors. The detectability of a galaxy clusters red sequence reduces with its richness, impacting the ability of optical surveys to detect and identify the smallest clusters and groups. Reliable red sequence identification requires multi band observations that span a cluster's 4000 angstrom break to maximise the colour difference between the early type galaxies contained by a cluster and the late type galaxies that populate the field (Gladders & Yee, 2000). In

addition to this an accurate and calibrated model of a cluster's red sequence over the survey's targeted redshift range.

Alternatively galaxy clusters can be identified from the weak gravitational lensing (Bartelmann & Schneider, 2001b) of coincident background galaxies due to their mass (Wittman et al., 2002). The location of a galaxy cluster is identified from peaks in the convergence map (Wittman et al., 2006). The benefit to this approach over other cluster detection methods, is it depends only on the mass content of the cluster. Other detection methods, such as red sequence detection, are dependent on our understanding of the internal processes within a cluster and their effect on its observable properties (Wittman et al., 2006). When searching for galaxy clusters using gravitational lensing, there are a number of limitations to consider. The presence of unlensed foreground galaxies within an observation act to reduce the strength of the lensing signal and need to be reliably identified and removed. Random deviations in the true shape of lensed background galaxies from spherical, introduces noise into the lensing measurement. Projection effects due to large scale structure within the universe, introduces false positives that require followup observations to identify (White et al., 2002; Hennawi & Spergel, 2005).

1.1.4.2 Galaxy cluster detection in Microwave observations

Galaxy cluster surveys conducted at radio wavelengths, such as those by the Plank telescope (Planck Collaboration et al., 2014), Atacama Cosmology Telescope (Hasselfield et al., 2013) and South Pole Telescope (Bleem et al., 2015), are able to identify galaxy clusters due to the SZ effect. Due to the SZ signal appearing as a relative shift in photon energy, the strength of the observed signal is effectively redshift independent (Birkinshaw, 1999). The strength of the SZ signal hence is only dependent on the electron temperature and density of the ICM (equation 1.8), which in turn depends on cluster mass with some noise. Any SZ selected galaxy cluster catalogue is hence mass limited, unlike flux limited X-ray and optical detection methods. This mass limit makes SZ surveys ideal for identifying higher redshift galaxy clusters at the expense of identifying the low mass, low redshift clusters included in other surveys.

1.1.4.3 Galaxy cluster detection in X-ray observations

X-ray based galaxy cluster surveys identify clusters via thermal bremstralung emission from the ICM (Rybicki & Lightman, 1986). Examples of galaxy cluster surveys conducted usng X-ray observations include, the Einstein Observatory Extended Medium-Sensitivity Survey (Gioia et al., 1990), the Wide Angle ROSAT Pointed Survey (WARPS; Scharf et al., 1997), the ROSAT Brightest Cluster Sample (BCS; Ebeling et al., 1998, 2000), the bright and southern SHARC survey's (Romer et al., 2000; Burke et al., 2003), the 400 square degree ROSAT survey (400d; Burenin et al., 2007a), the XMM Cluster Survey (XCS; Romer et al., 2001) and the XXL survey (Pierre et al., 2016). The large scales over which a cluster's ICM is distributed, makes a cluster's X-ray emission appear extended on the sky. There exist

a number of different methods to efficiently detect emission from extended sources such as a cluster. For example the WARPS (Scharf et al., 1997) survey makes use of Voronoi tessellation to identify and distinguish regions of increased flux due to a source from general background emission. Alternatively wavelet filtering can be used to extract extended emission from a set of observations, this is used by both XXL (Pacaud et al., 2016) and 400d (Vikhlinin et al., 1998; Burenin et al., 2007b). When applied, a wavelet filter preferentially highlights structures within the observation of comparable size to the filter. By applying a range of differently sized wavelet filters to an obssevation, sources of differing extensions can be detected. This method is of particular use in X-ray cluster detection due to its reliability at low photon count rates (Starck & Pierre, 1998).

The main complicating factor in the identification of galaxy clusters from X-ray observations, is the presence of point like emission from the significantly larger AGN population. Due to the often complex point spread function of X-ray telescopes, emission from AGN appears somewhat extended making it difficult to distinguish AGN from clusters. 400d (Vikhlinin et al., 1998; Burenin et al., 2007b) solves this problem by identifying and masking point sources prior to searching for genuinely extended objects. In contrast, XXL (Pacaud et al., 2016) and XClass (Clerc et al., 2012) detect both extended and point like sources, AGN and clusters are then separated based on the results of fitting a set of surface brightness profiles to each source (see section 2.1 for a more detailed explanation).

1.2 Machine Learning

Machine learning (ML) is a subset of artificial intelligence that focuses on the development of automated methods to identify and exploit multi dimensional trends and relationships within data sets to perform tasks. The applications and subsequent use of ML approaches have increased over recent decades due to the volume of data produced in the modern world. Examples of such applications include medical diagnosis (Deo, 2015), self driving vehicles, spam email filtering (Dada et al., 2019), image generation and language translation. Within astronomy the use of ML has increased rapidly in response to ever larger and more complex observational and simulated data sets. The percentage of refereed astrophysics papers submitted to NASA's ADS service (Kurtz et al., 2000) mentioning ML rising from $\sim 0.3\%$ in 2012 to $\sim 2.9\%$ in 2022, a near ten fold increase in as many years.

At the heart of any ML process is a mathematical model designed to describe a wide range of potential relationships between the input data and some output needed to solve a given task. The approach through which a model encodes such relationships can be broadly separated into two types, parametric and non-parametric. Parametric models, such as polynomial curve fitting, neural networks and models based on them, explicitly describe the functional form that maps the input to the desired output. Parametric models have a fixed number of parameters independent of the data provided to them. A parametric model is

subsequently trained by optimising the values of its internal parameters to achieve the desired output. While less adaptable, parametric models are generally more efficient to train due to self imposed limits on the functional forms of the learnt solutions.

In contrast, non-parametric models do not explicitly describe the potential mapping functions between the input and output, relying instead on the training data to infer the form of said functions. Non-parametric models can contain individual parameters that are optimised to the training set, but the number of parameters is inferred during the training process. Examples of non-parametric models include; K nearest neighbours where the output of the model is the average of the K nearest training points in some high dimensional space, decision trees where the number of branching decisions is determined during the training process and the Gaussian Processes (GP) used in this work (see chapter 3 for a detailed description of a GP). By not explicitly outlining the form of potential mapping functions, non-parametric models are highly adaptable, able to model more complex relationships and identify trends not considered when designing the model. The adaptability of non-parametric models does however lead to a general need for larger training sets and a tendency for such models over fitting to the training data. Over fitting occurring when a model precisely replicates the training input at the expense of accurately recovering the underlying truth. It is worth noting that the benefits and limitations of any specific model are more heavily dependent on said model than it being parametric or non-parametric. For example the GP used in this work was chosen due to its interpretability over its non-parametric nature.

In order for a ML model to solve a task it needs to be supplied with a set of training data from which to infer a solution. Due to the wide range of ML models, there exist a number of different approaches to training. Despite the large number of approaches to training, they can be broadly separated based on what and how information on the desired output is provided to the model during the training process. Of the types of training process, three are of particular relevance to astronomy, supervised, unsupervised and semi-supervised learning.

Supervised learning utilises labelled training data. Each example input in the training data set is paired with the corresponding desired value for the output of the model. The model is subsequently trained to replicate the desired output for each given input. Supervised learning is generally used to solve regression problems and those classification problems where the desired classes are already known. Examples of models that make use of supervised learning include neural networks, decision trees and support vector machines. The need for a large accurately labelled training set can be prohibitive in instances where the process of labelling each component of the training set is resource and time intensive.

Unsupervised learning differs from supervised in that there is no explicit desired output value for any given input. There exists instead some measure of how well the output of the model for the training set

as a whole solves the given task. An unsupervised model is subsequently trained by maximising this measure. For example unsupervised clustering algorithms are designed to efficiently describe the training set as originating from some number of classes, based on the assumption that objects from the same class are locally correlated in parameter space. Which of the classes any one training object belongs to is not predetermined, but inferred by the unsupervised classifier. The benefit of this approach compared to supervised learning is the freedom in the model's output. The model is free to determine from the data the best solution to the task at hand. The freedom provided to the model does limit the scientist's ability to control the output though. Uses of unsupervised learning include dimensionality reduction, outlier detection and generative tasks.

Semi-supervised learning (Chapelle et al., 2010) is a combination of supervised and unsupervised learning where an often small labelled training set is supplemented by a large unlabelled training set. By requiring a relatively small labelled training set, semi-supervised learning reduces the time and resource cost needed when creating a sufficiently large labelled training set. The reduction in labelled training data does however reduce the accuracy of the output of the trained model. Within this work we make use of a supervised training approach as it allows us to explicitly train the model to identify galaxy clusters.

There are both a number of advantages and disadvantages to ML models when compared to bespoke solutions created using specific knowledge of the problem being solved. With respect to this work, a bespoke solution designed using domain specific knowledge would be existing cluster detection algorithms designed using our understanding of galaxy cluster properties. The generalisability of ML models make them highly applicable to a range of problems, removing the need to create a bespoke solution, freeing the resources that would have been used for other research. Since a ML model does not need to explicitly replicate the underlying mechanics of the problem being solved there is significantly more freedom in their design, allowing for an emphasis on speed and efficiency. The resulting models are able to quickly and efficiently process large data sets (post training) at the expense of the accuracy provided by a bespoke solution that makes use of domain specific knowledge.

While the generalised nature of ML models means they are able to identify and exploit novel information to solve a given problem, the often abstract nature of components within a model obfuscate the learnt solution. This is of particular issue for astronomers looking to physically interpret and learn from the solution identified by the model. While there exist methods for interpreting ML models it is not always clear which approach is best for any given situation.

When applying a ML model to solve a given problem the main limiting factor is the available training data. While a bespoke method can be constructed to make use of our physical understanding of a problem to minimise the impact of a lack of, or inaccuracies in, the available data, ML methods are unable to do so.

Further the reliability of a ML model drastically reduces when presented with data unlike that used to train it, despite which the ML model will provide an answer. While it is possible to extract a measurement of the error on the answer provided by some models, they can be confidently incorrect when presented data that differs from the training set. It is necessary to create a set of training data that accurately recreates and fully spans all possible inputs to the model. Accurate replication of real data is a particular concern when applying a model trained on simulated data to real data.

1.2.1 Machine Learning in Astronomy

With the growth in the volume and complexity of modern astrophysical data sets, astronomers are increasingly looking to apply ML solutions to both solve and automate complex tasks. Here we provide a brief discussion on the applications and use of ML models within astronomy. For a more complete discussion on the use of ML we direct the reader to reviews by Baron (2019) and Djorgovski et al. (2022). For reviews with a greater focus on applications of ML in cosmology see those by Moriwaki et al. (2023), Lahav (2023), and the white paper by Ntampaka et al. (2019a).

One common and conceptually simple application of ML within astronomy, is the measurement of an object's properties directly from observational data, such as the mass of a galaxy cluster (Haider Abbas, 2019; Cohn & Battaglia, 2020), or redshift (Hatfield et al., 2020). Such measurement tasks are generally approached as a supervised regression problem. The benefit of using ML solutions over more traditional measurement methods are two fold, the lower computational requirement of the trained model allows for the efficient processing of large data sets and the ability of a ML model to identify and exploit additional information within the observations can lead to reductions in the scatter on the measurement (Ntampaka et al., 2019b). The difficulty in using ML to solve such a problem is the creation of a reliably labelled training set that accurately reflects real sources. Any training set that makes use of real observations that have been labelled using existing methods, inherently introduces uncertainties on those labels, limiting the accuracy of the ML models output. Alternatively one could construct the training set from simulated data where the label is known, for example the majority of ML solutions for measuring galaxy cluster mass are trained on simulated mock X-ray (Green et al., 2019; Ferragamo et al., 2023), radio (Krippendorf et al., 2023), or velocity data from a clusters constituent galaxies (Ntampaka et al., 2015; Ho et al., 2019; Ntampaka et al., 2019c). When using simulations in this way it is important to minimise any artefacts or inaccuracies within the simulation that can bias the results of the ML model when applied to real observations.

The identification of novel object classifications is an important part of astronomy, and unsupervised clustering techniques present an opportunity to make use of all available information when doing so. Unsupervised classifiers have been applied to galaxy spectra (Sánchez Almeida et al., 2010) and glitches

within the LIGO/Virgo detectors (George et al., 2018). An additional benefit of taking an unsupervised approach to classification is that the results are not biased by prior knowledge, the final classes depending only on the information within the training data set. The downside is a lack of control over the output and the need for an expert to visually inspect and interpret the classes produced. It is also not guaranteed that the classes produced by the unsupervised classifier are due to the physical properties of the observed objects, ML models being unable to distinguish between features, due to the nature of the observed object and those due to observational affects.

Within any sufficiently large survey there will exist erroneous and interesting detections that differ from the main population. Manual identification of such detections is not practical given the scale and complexity of modern surveys. Unsupervised anomaly detection can be used to somewhat automate this task, identifying a significantly reduced subset of detections that are sufficiently distinct to warrant visual inspection. For example Reis et al. (2018) makes use of an unsupervised random forest to detect anomalies in the Apache Point Observatory Galactic Evolution Experiment. The detection of outliers using such approaches is limited by the accuracy of the input data, sufficient noise acting to disguise outliers as the result of measurement error. As with unsupervised classification, anomaly detection models are unable to distinguish between anomalies due to unique astrophysics and those due to observational effects. Visual inspection by an expert is needed to make such distinctions. This approach is particularly useful when searching for rare types of sources with too few existing examples for training a classifier.

1.2.2 Supervised Machine Learning Classifiers

The aim of any astrophysical survey is the identification and cataloguing of a specific type, or types of source from observational data. With the volume and complexity of data expected from both current and near future surveys such as eROSITA (Predehl et al., 2021) and Euclid (Laureijs et al., 2011), astronomers face a number of new challenges. A particularly appealing solution to these challenges are ML classifiers given their ability to both efficiently process large amounts of data and exploit all available information when distinguish between source types. Over the last decade this has led to a number examples of ML being used for source selection from survey data including: classification of variable stars from time series data (Richards et al., 2011); clasification of X-ray point sources from Chandra data (Kumaran et al., 2023); the selection of stars from the eROSITA final equatorial depth survey (Schneider et al., 2022); identification of galaxy cluster member galaxies (Angora et al., 2020); galaxy cluster selection from the Sloan Digital Sky Survey (Hao et al., 2010; Grishin et al., 2023); and the use of convolutional neural networks to select galaxy clusters from XXL's sister survey X-Class (Kosiba et al., 2020).

The use of supervised ML classifiers for source selection presents the astronomer with two new problems to solve. The first is the requirement for large sets of perfectly labelled training data. There are

several ways of creating such data sets, but each has drawbacks. One option, assuming there exists a relevant and sufficiently large source catalogue, is to have an expert or group of experts inspect and label a sufficient number of sources to create a training set. This however requires a significant time investment for those involved and relies on the ability of experts to accurately label sources. Alternatively the labelling process of a sufficiently large source catalogue can be done by non-experts through citizen science schemes (Dieleman et al., 2015). This approach has the disadvantage of introducing a level of uncertainty in the labels produced along with the resources needed to setup and run such a project. An additional option is the use of simulated data, where a large number of sources can be generated and automatically labelled based on the full information from the simulation. This approach however, relies on the simulations accurately recreating all sources that may be present within a real catalogue.

The second problem is the complex nature of the statistical models (particularly neural networks and those models developed from them) and makes developing a physical understanding as to why a model does or does not select any given source difficult, if not impossible. Since the interpretability of a model tends to decrease with increasing complexity, this is an important factor to consider when choosing the best ML classifier. This is of particular importance for galaxy cluster surveys, where an accurate model of the selection function is needed to cosmological studies.

1.3 Summary

Forming from the gravitational collapse of the largest perturbations in the initial matter distribution of the universe, galaxy clusters provide a key insight into cosmology and the evolution of galaxies. The study of galaxy clusters for these purposes, requires large reliable and well understood cluster catalogues. As galaxy cluster surveys grow in both size and detail, a new more efficient approach is needed to identify clusters from observational data.

One promising prospect is the use of ML classifiers for source selection. The use of ML classifiers however presents new challenges for the astronomer, including the creation of large reliably labelled training data that accurately recreates real survey data and the interoperability of the source selection criteria learnt by the model.

The aim of this work is to use a Gaussian Process (GP) binary classifier adapted for imperfectly labelled training data to select galaxy clusters from the XXL X-ray source catalogue Pierre et al. (2016). The XXL X-ray source catalogue is outlined in chapter 2 along with the CAMIRA optically selected galaxy cluster catalogue (Oguri et al., 2018) used to test the results of GP classifier once applied to the XXL catalogue. The GP binary classifier model and its adaption to account for uncertainties on the training labels is outlined in chapter 3. The model is applied to and assessed on its ability to identify galaxy cluster

candidates from the XXL X-ray source catalogue in chapter 4. The selection criteria learnt by the model are identified and interpreted in Chapter 5. Chapter 6 containing a summary and conclusion of the work.

2

Galaxy Cluster Surveys and Source Catalogues Relevant to this Work

In order to achieve its aim of applying a ML model to select galaxy cluster candidates from the XXL X-ray source catalogues, this work makes use of source catalogues and data products from both the XXL X-ray survey (Pierre et al., 2016) and the Hyper Suprime-Cam Subaru Strategic Program (HSC, Aihara et al., 2018b). A simulated XXL X-ray source catalogue and optically selected cluster catalogue produced by applying the `CAMIRA` cluster detection algorithm (Oguri et al., 2018) to the HSC survey, are used to asses the output of the model. In addition, observational data from the XXL and HSC surveys are used to produce source cutouts for visual inspection. This chapter outlines the XXL and HSC surveys along with those catalogues used in this work.

2.1 The XXL X-ray Survey

The XXL X-ray survey was designed with the goal of producing a well-defined X-ray selected catalogue of galaxy clusters out to a redshift of one to be used for precision cosmological studies (Pierre et al., 2016). The XXL X-ray survey additionally provides a large catalogue of AGN detections and observational data for the purpose of studying the X-ray background. To achieve its main goal the XXL survey consists of two 25 square degree observing fields widely separated on the sky (Figure 2.1). The two XXL observing fields are named the North and South fields. Both fields are constructed from a series of approximately 10ks observations by the XMM Newton X-ray telescope, though there exists some

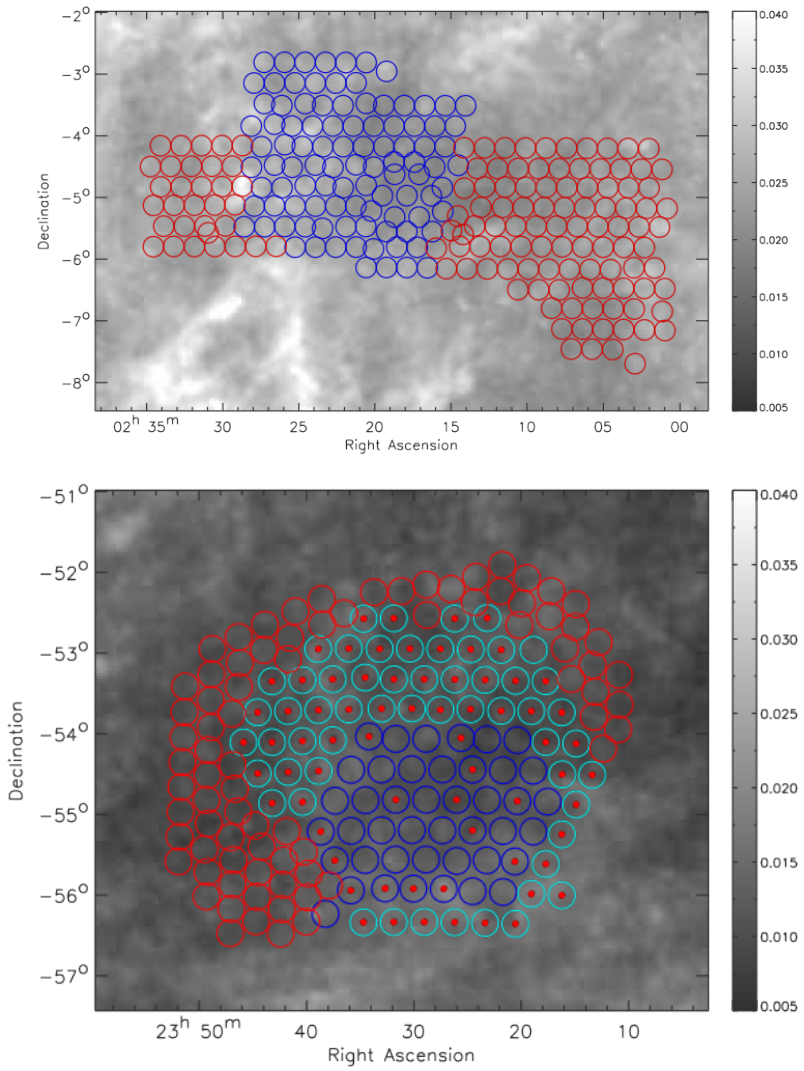


Figure 2.1: XXL observational layout for the North (top) and South (bottom) XXL fields by Pierre et al. (2016). The observational layout is overlaid onto maps of the dust column density calibrated to E(B-V) reddening in magnitude (Schlegel et al., 1998)

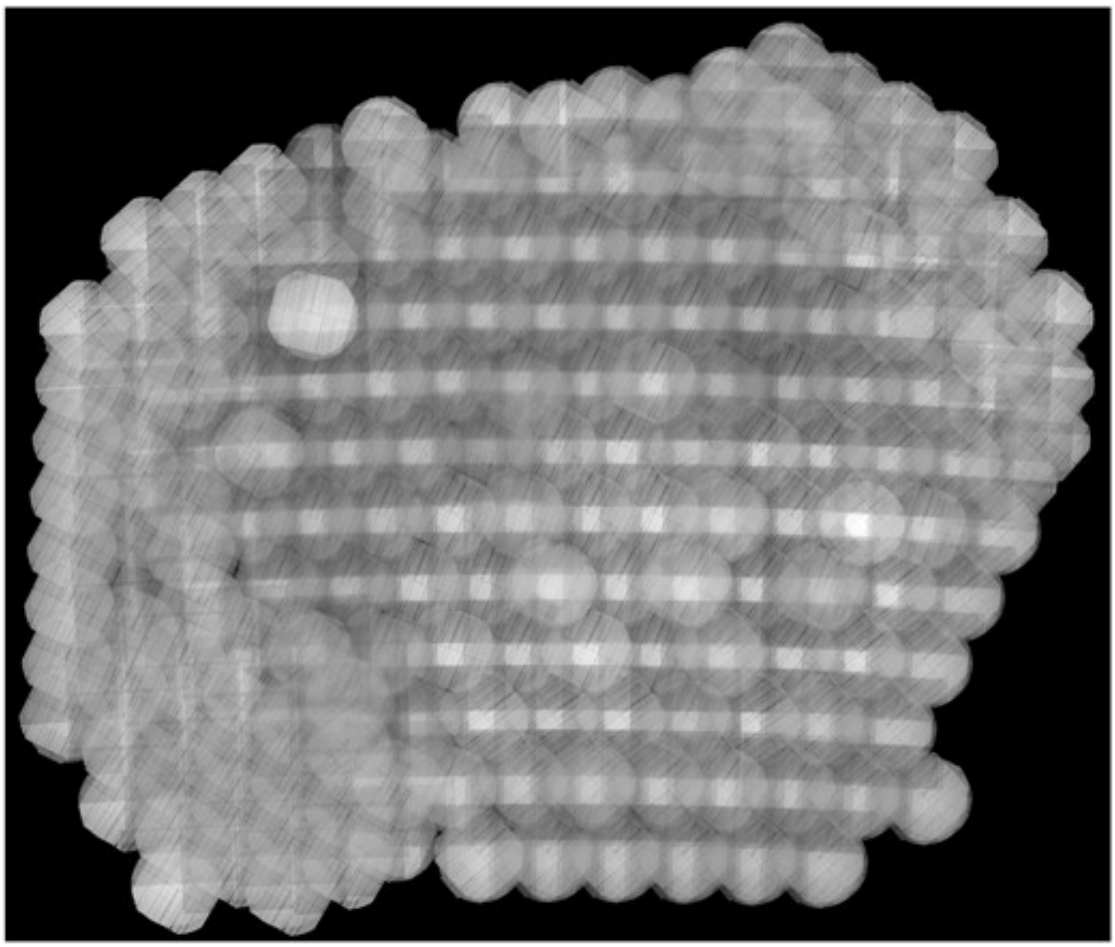


Figure 2.2: Composite mosaic of XMM Newton observations in the southern field showing the variation in exposure (Upsdell et al., 2023). The lighter areas showing regions with increased exposure time. The apparent straight dark lines on each pointing correspond to the gaps between individual ccd's within XMM Newton.

deviations in exposure time (figure 2.2) that must be accounted for when using the observations. A set of deeper observations are present in the North field covering an approximate 5 square degree region.

X-ray sources were identified and characterised from the observational data using the XAMIN pipeline (Pacaud et al., 2006; Faccioli et al., 2018) to produce the North and South source catalogues. This work makes use of the latest XXL X-ray source catalogues produced by version 4.3 of the XAMIN pipeline (Faccioli et al., 2018) a general description of which is given in subsubsection 2.1.1. Two galaxy cluster samples were produced from the XXL X-ray source catalogue by selecting potential galaxy clusters based on their properties as measured by XAMIN. Named the C1 and C2 cluster samples they were designed to achieve purities of 1 and 0.5 respectively. The details of the C1 and C2 samples including their selection criteria are detailed later in subsubsection 2.1.2. The XXL bright cluster sample (Pacaud et al., 2016) containing the 100 brightest galaxy clusters detected by the XXL survey was produced as part of the first data release. The larger 365 galaxy cluster catalogue (Adami et al., 2018) produced as part of the second data release, contains those clusters included in the C1 and C2 samples with spectroscopic confirmed redshifts. An additional set of optically selected galaxy clusters with associated X-ray emission too week to be characterised by XAMIN is included in the 365 catalogue, labelled as the C3 sample.

Those cluster catalogues produced by the XXL survey have been used for a range of cosmological and galaxy cluster studies. Such studies include, measurements of the cluster luminosity temperature relation (Giles et al., 2016), constraining the values of cosmological parameters (Pacaud et al., 2018; Garrel et al., 2022), measurements of the cluster mass temperature relation (Lieu et al., 2016) and investigations into clusters' baryonic content (Eckert et al., 2016).

2.1.1 Xamin detection and characterisation pipeline

Here we provide a general description of the latest XAMIN processing pipeline (version 4.3) used by XXL to detect and characterise X-ray sources. For a full description we point the interested reader to the work by Faccioli et al. (2018) and the references therein.

In previous versions of the pipeline, XAMIN independently identified sources from each XMM pointing (Pacaud et al., 2006). In doing so the XAMIN pipeline was unable to make use of all available information in regions where two or more XMM pointings overlapped. The most recent version of the pipeline instead splits the observing field into a set of mosaicked square tiles, before identifying and characterising sources from each tile separately. Each tile is 68 by 68 arcminutes and they are arranged as a grid with spacing of 60 arcminutes to cover the entire field. The chosen tile size and spacing produces a 4 arcminute overlap between each tile designed to avoid issues during processing due to the edge of a tile.

Each tile is constructed by co-adding data collected from multiple observations by the MOS1, MOS2 and PN detectors. The combined image is subsequently smoothed using a wavelet approach. Sources are detected from the smoothed image using the mixed detection method outlined by Valtchanov et al. (2001). This mixed approach independently applies two filtering methods to detect sources from the observational data. The first filtering method convolves of the observational data with Gaussian kernels. The second filtering method is the wavelet function histogram method (Starck & Pierre, 1998). Sources are independently identified from the two filtered images using Source extractor (Bertin & Arnouts, 1996) to produce a source catalogue.

Having created a catalogue of sources for a tile, XAMIN moves to characterise each source by independently fitting a set of four surface brightness profile models to data from relevant XMM Newton observations. The surface brightness models each correspond to a different physical configuration for a source (Table 4.2). i) The EXT profile models the extended emission from a galaxy cluster's ICM as a β profile,

$$S_X(r) \propto \left[1 + \left(\frac{r}{r_c}\right)^2\right]^{-3\beta+0.5}. \tag{2.1}$$

Here S_X is the surface brightness at a projected radius r from the cluster centre, r_c is the core radius of the profile and β is fixed to $2/3$ as the generally low number of counts from XXL clusters is insufficient to robustly constrain β (Faccioli et al., 2018). ii) The unresolved emission from AGN is described by a point source model. iii) The double point source model describes the appearance of two point like AGN with overlapping emission. iv) The extended plus point source model describes the extended emission from a cluster contaminated by the point like emission from an AGN. The presence of the contaminating AGN may be due to projection or the presence of an AGN within one of the clusters galaxies, often the brightest central galaxy (Logan et al., 2018; Bhargava et al., 2023). All surface brightness profiles are convolved with XMM Newtons point spread function (PSF) at the location of the detected source during the fitting process. The source properties measured by fitting each surface brightness profile are recorded within the source catalogue. A subset of these measured source properties are what XAMIN and the binary classifier model use when selecting galaxy cluster candidates. A list of those measured source properties relevant to this work can be found in Table 4.3

The fitting process is conducted separately for the data collected by the two MOS detectors, the PN detector, and the combined MOS and PN detectors. This work makes use of the properties measured by the fit to the combination of the MOS and PN detectors to maximise the information available to the ML binary classifier. We may find that the ML model learns to only makes use of results from one of the MOS or PN detectors, but at this point it is not possible to predict this.

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

Table 2.1: 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.

At this point there exists separate catalogues for each tile that makes up the North and South fields. We create a full catalogue for the North field by simply appending the catalogue from each of the tiles that make up the North field, this process is repeated using the Southern tiles to create the South catalogue. The North and South catalogues contain a total of $24,412$ and $18,090$ X-ray source respectively. Due to the 4 arcminute overlap between tiles it is possible for the same source to be included in the catalogue two or four times. The fraction of sources within the catalogue that are duplicates will be relatively low given the small fraction of the full 25 square degrees covered by multiple tiles. Given the low fraction of sources that are duplicates, they should have minimal impact on the results of the binary classifier.

2.1.2 The C1 and C2 galaxy cluster candidate samples

The C1 and C2 galaxy cluster samples are selected from the North and South source catalogues by applying cuts in the parameter space defined by the core-radius (EXT) and extension likelihood (EXT_LIKE) parameters. The standard XAMIN pipeline also uses EXT_DET_STAT to select galaxy clusters (Faccioli et al., 2018), however, we found that this parameter did not significantly impact our sample selection so it is not used in this work. EXT is a measure of the physical extent of the source on the sky and is measured when fitting the Extended Beta Model. The EXT_LIKE parameter is determined by independently fitting the extended and point source models to the source. The value of the EXT_LIKE parameter being the ratio of the likelihood of detecting the extended source and that for the point source. EXT_LIKE 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.

As described previously the aim is for the C1 and C2 cluster samples to have purities of 1 and 0.5 respectively. The cut values listed in Table 2.3 are chosen to achieve the targeted purities for the C1 and C2 samples. These cut values were determined from an analysis of simulated XXL catalogues produced for version 3.3 of the XAMIN pipeline, where the ground truth of an extended X-ray source is known at

Parameter	Description
EXT_LIKE∗	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 how likely 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 how likely 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 how likely 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 how likely 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

Table 2.2: 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 4.2). 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 Gaussian process model is trained.

Cluster Sample	EXT (arcseconds)	EXT_LIKE
$C1$	> 5	> 33
$C2$	> 5	15 to 33

Table 2.3: 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 (Pierre et al., 2016) and should be considered approximate. The standard XAMIN pipeline also applies cuts on EXT_DET_STAT (Faccioli et al., 2018) however since we find this has no significant impact on the C1 and C2 samples they are not used in this work.

that location (Pierre et al., 2016). 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).

The subsequent follow-up of such sources in multiple wavebands broadly confirms this level of reliability for the identification of sources labelled as C1 or C2 (e.g Pacaud et al., 2016; Adami et al., 2018). In other words, follow-up observations in multiple different wave bands have shown that C1 sources are in almost all cases ($> 90\%$) genuine detections of clusters. Given the simulated sources were synthesised using β-models appropriate for relaxed and well-virialised clusters, the C1 and C2 subsamples of real XXL sources 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 Pierre et al. (2016). However, less relaxed systems not dominated by a single virialised halo are likely to be less well-represented in the sample, something that the ML approach used here may help mitigate against.

Further analysis of the galaxy clusters selected by XXL as a C1 or C2 allows for the characterisation of XAMIN's selection function (Pacaud et al., 2016; Adami et al., 2018). The mass redshift distribution of the bright XXL cluster sample (Pacaud et al., 2016) in figure 2.3 shows that XXL is able to detect clusters out to a redshift of approximately unity, with the majority of clusters found around a redshift of 0.3. The masses of clusters at high redshifts is found to be relatively consistent with low mass clusters ($< 10^{14} M_{\odot}$) only being detected at lower redshifts. One hoped for benefit of applying a ML algorithm to select galaxy clusters from the XXL catalogue is selection of low mass galaxy clusters at higher redshifts.

Figure 2.4 depicts the distribution of sources within the EXT and EXT_LIKE parameter space and whether they fall into the C1, C2 or neither sample. Through out this work we refer to the (majority) set of sources not contained within the C1 or C2 sub-samples as the non-C1C2sample. It is immediately apparent that the C1 cluster sub-sample consists of X-ray sources for which the EXT model measures a significant extension. Note that the XAMIN models are convolved with an appropriate model of the

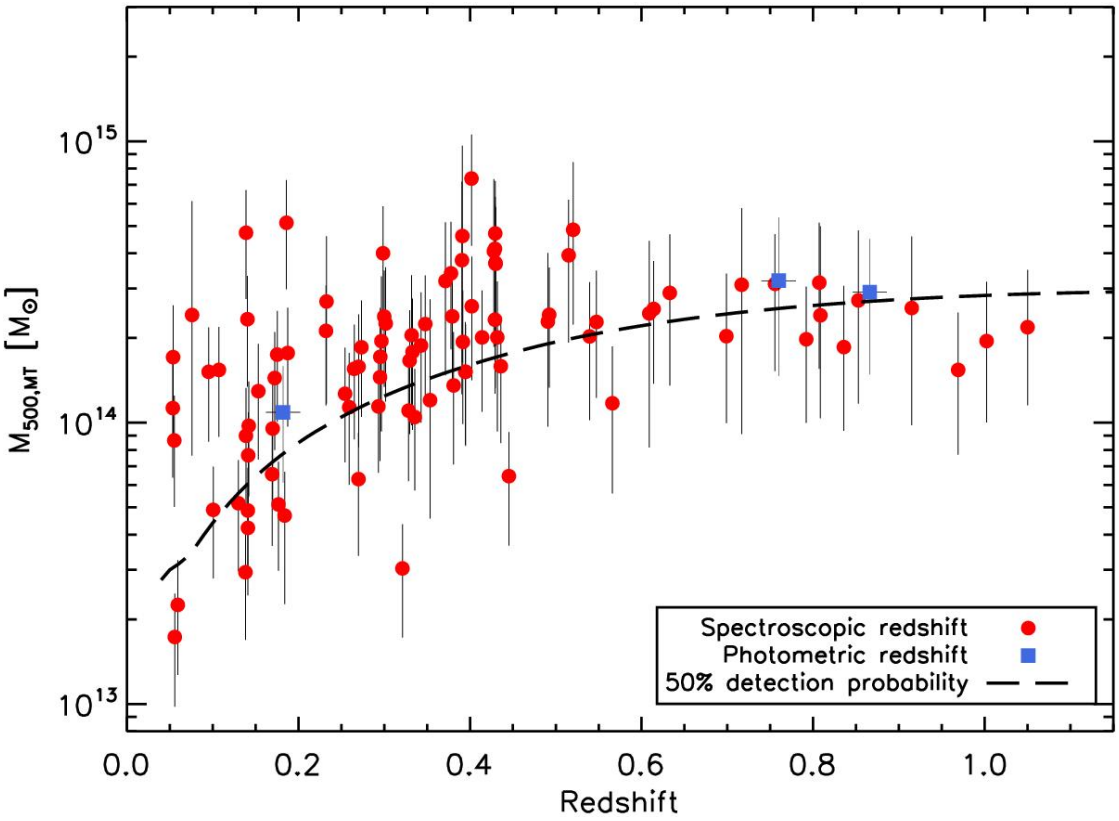


Figure 2.3: Redshift mass distribution of XXL's bright cluster sample produced by Pacaud et al. (2016). At higher redshifts the mass of clusters is fairly consistent with low mass clusters ($< 10^{14} M_\odot$) only being found at lower redshifts. The dashed line shows the 50% completeness limit calculated for a WMAP9 cosmology using the method described in section 6.1 of Pacaud et al. (2016).

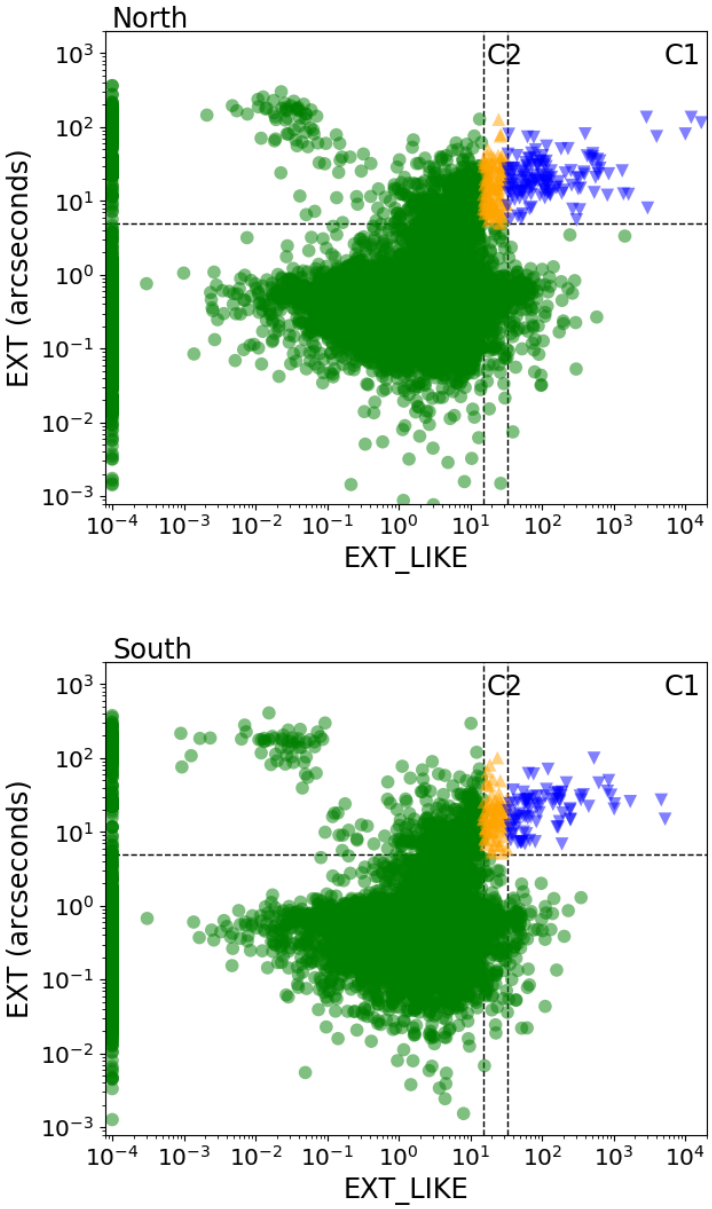


Figure 2.4: Distribution of sources for the North (top) and South (bottom) catalogues as a function of EXT_LIKE against EXT. Sources are labelled by whether they fall into the C1 (blue triangle pointing down) or C2 (orange triangle pointing up) cluster samples or belong to neither (green circle). Dashed lines indicate the C1 and C2 selection criteria listed in table 2.3. Those sources with an EXT_LIKE of zero are plotted on the left hand side of the figure.

point spread function (PSF), so a non-zero EXT coupled with a high EXT_LIKE implies an extended source (although the C1 and C2 definitions conservatively use a minimum EXT of 5 arcseconds). The C2 sources have similar extents to the C1s, but these are less-reliably measured (and hence have a lower EXT_LIKE). This can be for a variety of reasons - for example the sources could be detected with fewer counts than those typical for a C1 selected source or the local background/noise could be relatively high, reducing the fidelity of the fit. The lower EXT_LIKE threshold means the C2 sample can include higher-z galaxy clusters than the C1 sample.

The region occupied by the C2 sample in Figure 2.4 appears relatively small (particularly the range in EXT_LIKE). 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_LIKE 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 even variation in the final 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 in this classification scheme (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. This may be accomplished by utilising other parameters returned by the analysis pipeline (i.e. those listed in Table 4.3). Ideally such an approach would recover all the C1s and a high purity subset of C2s, plus additional clusters that fall outside the C1 and C2 definition and which could not be reliably identified as clusters using the standard selection illustrated in Fig. 2.4.

In our scheme we use the C1 and C2 classifications defined in table 2.3 to label those sources that have a high (C1) or a moderate (C2) likelihood of being true clusters for the purpose of training our GP. Essentially we are using the GP to explore whether combinations of values of source parameters (other than EXT and EXT_LIK) can identify clusters comparably to, or better than, the simple two parameter cut used in the original classifications. We will explore whether such combinations can identify clusters and cluster candidates missed by the C1 and C2 classification, for example the less relaxed systems which might become more important at higher redshifts, particularly when exploring the astrophysics of the growth of clusters and their galaxy populations.

2.1.3 The simulated XAMIN version 4.3 source catalogue

Simulated source catalogues are a useful tool for both training ML models and assessing the reliability of source selection processes (be they ML based or otherwise). What makes simulated catalogues useful is the ability to accurately label each source in the simulated catalogue based on its physical origin. For example, sources within a simulated XXL catalogue can be labelled as genuine detections of a cluster,

AGN or as a spurious detection. Given a fully labelled simulated catalogue it is possible to asses any source selection process based on which simulated sources are selected and their labels.

Within this work we make use of a simulated XXL source catalogue (produced by the XXL collaboration) to asses the output of the ML classifier and potential source samples (section 4.3). The first step in creating the simulated catalogue is the creation of a mock 25 square degree observing field. A realistic galaxy cluster population containing mass and redshift was created from the Euclid flagship dark matter only simulation. Scaling relations were used to calculate each clusters luminosity and temperature. Given the luminosity and temperature the extended emission for each simulated cluster was subsequently modelled by a spherical β profile (equation 2.1) where the value of β is set to $2/3$. Following this a set of simulated AGN from the Cosmo OWLS simulation (Le Brun et al., 2014) were added to the field following the method described in (Koulouridis et al., 2018).

Having constructed the survey field a set of XMM Newton observations were simulated mimicking the spacing between the pointings that make up the real XXL survey and covering the full 25 square degree field. Each simulated observation included background from both AGN emission and particle interactions with the telescope.

The field was split into 25 tiles of size 68" by 68" before processing by the XAMIN pipeline. We combine the individual tile catalogues to produce a single catalogue containing $28,256$ sources in the same way as that for the real XXL data. Those XAMIN detections within 12 arcseconds of a simulated overdensity halo are labelled as detections of a galaxy cluster. We note that this does allow for multiple detections to be labelled as detections of the same simulated cluster, this is done to account for duplicate detections including those produced by the overlap between tiles. This simulated labelled catalogue is subsequently used in section 4.3 to test the output of the adapted GP binary classifier.

2.2 The Hyper Suprime-Cam Subaru Strategic Program optical survey

The Hyper Suprime-Cam Subaru Strategic Program (HSC Aihara et al., 2018b) is a $\approx 1,400$ square degree, wide band, optical survey conducted using the Subaru telescopes Hyper Suprime-Cam instrument (Miyazaki et al., 2018). The main aims of the survey are to provide sufficient observational data to conduct studies of dark matter and dark energy, the formation and evolution of galaxies, and the re-ionisation of the inter galactic medium.

The HSC survey is of particular use within this work due to the approximately 22 square degree overlap between it and XXL's North field. In section 4.5 we make use of optical source cutouts from HSC's third

public data release (Aihara et al., 2022) when conducting visual inspection of XXL sources to asses the output of the ML binary classifier. In addition we make use of the latest CAMIRA cluster catalogue constructed from HSC's galaxy catalogue to select a subset of XXL sources with an increased probability of being a galaxy cluster. The CAMIRA cluster catalogue and the process by which XXL sources are matched to CAMIRA clusters is outlined in more detail in the following subsection.

2.2.1 The CAMIRA galaxy cluster catalogue

In order to test the utility of the GP, it is instructive to examine its results for a subset of an X-ray catalogue for which there is evidence (independent of X-rays) of a galaxy cluster being associated with a given X-ray source, as a proxy for a ground-truth. For any subset of an X-ray catalogue with a higher fraction of galaxy clusters (compared to the catalogue as a whole) the GP should, if working correctly, assign higher confidence values to those source in the subset compared to the whole catalogue. Hence a subset of X-ray sources identified as containing a higher fraction of galaxy clusters can be used to test if the GP is performing correctly, as it should return higher confidence values compared to the whole catalogue. By creating the subset using evidence that is independent of the sources' X-ray properties it avoids any bias that would be introduced given that the initial training labels are assigned based on each source's X-ray properties, even if the GP is not trained on the exact X-ray properties used for labelling (EXT and EXT_LIKE).

We are able to make use of a catalogue of optically selected cluster candidates in the form of the latest CAMIRA catalogue produced from the HSC-SSP S21A data set (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 (Rykoff et al., 2014; Oguri, 2014) to the Subaru Strategic Program imaging survey (Aihara et al., 2022) with Hyper Supreme-Cam (HSC-SSP) (Aihara et al., 2018a) in a similar manner to that of Willis et al. (2021) in order to explore how well the GP performs. In the ≈ 22 square degree overlap between the HSC-SSP and XXL-N, there are 572 CAMIRA selected clusters within 13 arcminutes of the centre of an XMM Newton pointing. The number of clusters within 13 arcminutes differs from the 270 reported in Willis et al. (2021) as here we are using a more recent CAMIRA catalogue, with a richness selection of $N > 10$ instead of $N > 15$ as used by 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 4.1) using a matching radius of 15 arcseconds. This is a relatively conservative choice, designed to minimise 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 arcseconds 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. This CAMIRA-matched catalogue constitutes a subset of XXL sources that should have a higher fraction of genuine clusters than a random subset because we expect a significant fraction

of the CAMIRA-identified clusters to have sufficiently strong extended X-ray emission that they will be detected as X-ray sources in the XXL catalogue. 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.

Of the 572 CAMIRA clusters, 440 did not match to an X-ray source within 15 arcseconds. The most likely reasons for these 440 CAMIRA clusters not being matched are: i) the conservative nature of our 15 arcsecond matching radius and ii) the X-ray emission associated with these clusters not being detected by the XAMIN pipeline, as discussed in Willis et al. (2021). For our purposes, this is not a concern, as the goal of our matching is to define a high purity sample of optically-selected clusters with which to assess the performance of the GP.

Prior to examining the output of the GP for this subsample we confirmed that it does indeed contain an increased fraction of genuine cluster detections compared to a purely random sub sample, on the basis of the standard XXL source classification. The fraction of these CAMIRA matched sources 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. The probability of a set of 141 random sources from the XXL North catalogue having a total fraction of $C1$ and $C2$ sources greater than, or equal to, 0.34 is effectively zero. 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 might therefore expect that X-ray sources associated with CAMIRA detections should be identified by the GP as potential clusters at a higher rate than X-ray sources drawn at random, even though the GP is blind to the association with any CAMIRA source. We perform this test in section 4.4

Having outlined those catalogues and data products from the XXL and HAC surveys used in this work, the next chapter describes the theory behind a Gaussian process and the method by which we adapt it to account for imperfectly labelled training data.

3

Gaussian Processes for Binary Object Classification With Imperfectly Labelled Training Data

A Gaussian Process (GP) is a specific form of Stochastic Process defined by the following; a (potentially infinite) collection of random variables is considered to be a Gaussian Process if the joint probability distribution of every finite combination of said random variables can be described by a multivariate normal distribution. By defining a GP in this way it inherits the ability of a stochastic process to act as a probability distribution over functional space (a function being treated as an infinite set of variables indexed by some continuous vector space) without explicitly parameterising any functions. Further, since any normally distributed random variable exists over the entire real axis, any function randomly drawn from a GP can itself exist over the entire real axis. These properties of a GP make it an ideal starting point when building non-parametric models of unknown functions.

Within astronomy Gaussian processes have been used for a range of tasks including the modelling of time series data (Foreman-Mackey et al., 2017; Boone, 2019) and classification tasks (de Beurs et al., 2022; Morales-Álvarez et al., 2018). We do not find the use of Gaussian processes for classification with imperfectly labelled training data.

The aim of this chapter is to first introduce a GP as a probability distribution over functional space and how it can be used to solve regression problems. Following this we describe how a GP can be used for binary classification and the adaption we developed to allow such a model to account for imperfectly labelled training data. The content of this chapter follows and expands on the derivations outlined by Bishop & Nasrabadi (2006). The adapted GP binary classifier model introduced here is then used in later chapters to select galaxy cluster candidates from the XXL X-ray source catalogue.

3.1 Gaussian Processes as a probability distribution over functions

Before moving onto applying a GP to solve a regression problem let us first consider in detail how a GP acts as a probability distribution over functional space. To best understand how the GP encodes the properties of a function in a probability distribution let us consider the values of some unknown arbitrary function y_1 and y_2 at two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. By the definition of a GP we can write the joint probability distribution on the values of y_1 and y_2 as a multivariate normal distribution,

$$P(y_1, y_2) = \mathcal{N}\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \middle| \, \boldsymbol{\mu} \, , \, \bar{\boldsymbol{k}} \right). \tag{3.1}$$

Here the vector $\boldsymbol{\mu}$ is the mean and $\bar{\boldsymbol{k}}$ is the covariance matrix. The normal distribution restricts the values of y_1 and y_2 to the whole real axis. This limits a GP to modelling real functions.

The mean vector $\boldsymbol{\mu}$ simply contains a prior prediction of the most likely values for y_1 and y_2. We can write the components of $\boldsymbol{\mu}$ more explicit as functions of the positions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\boldsymbol{x}_1) \\ \mu(\boldsymbol{x}_2) \end{bmatrix}. \tag{3.2}$$

Here $\mu(\boldsymbol{x})$ is the prior prediction of the unknown function $f(\boldsymbol{x})$. Since a normal distribution covers the entirety of the real axis independent of its mean, the choice of $\mu(\boldsymbol{x})$ does not limit the values of $y(\boldsymbol{x})$. For simplicity the mean is often set to zero for all $\boldsymbol{x}$.

The contents of the covariance matrix $\bar{\boldsymbol{k}}$ encode the relationship between the values y_1 and y_2. For any function the relationship between the values of the function at two points is dependent on said points. The relation between y_1 and y_2 depends on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The exact parameterisation of this relation depends on the type of function being modelled, but it always takes the form of a kernel function, $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. A kernel function being a specific form of symmetric positive semidefinite function (Bishop & Nasrabadi, 2006). The covariance matrix $\bar{\boldsymbol{k}}$ is then explicitly written as,

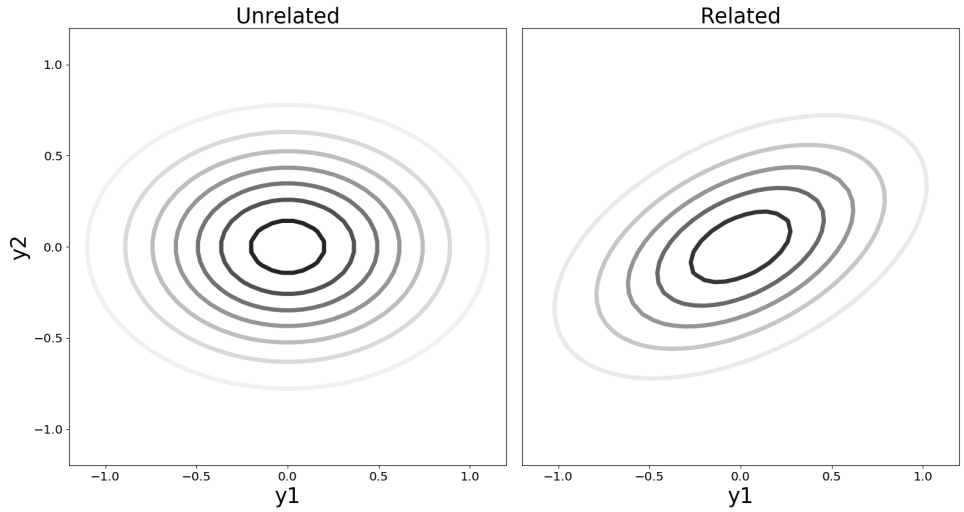


Figure 3.1: Example of the joint multivariate normal distribution over the values of the function y_1 and y_2. The left hand panel depicts the distribution when the values of y_1 and y_2 are unrelated. The right hand panel depicts the probability distribution when the values are related. In both examples the expected values for y_1 and y_2 are zero.

$$\bar{\boldsymbol{k}} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) \end{bmatrix} \tag{3.3}$$

and is itself positive semidefinite.

To understand how the output of the kernel affects the joint distribution over y_1 and y_2 consider first the situation where the kernel encodes no relation between y_1 and y_2. The absence of any relation is encoded by returning zero when given both x_1 and x_2, $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\boldsymbol{x}_2, \boldsymbol{x}_1) = 0$. In this instance the covariance matrix is diagonal and the probability distributions over y_1 and y_2 are wholly independent. The diagonal components of $\bar{\boldsymbol{k}}$ simply determins the standard deviations over y_1 and y_2. The impact of this on the joint probability distribution can be seen in the left hand panel of figure 3.1.

In contrast when there is some relation between y_1 and y_2, the kernel returns a non-zero value for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The result is some covariance between the values of y_1 and y_2 as can be seen in the right hand panel of figure 3.1. Note that this doesn't affect the expected values of y_1 and y_2, these are purely dependent on the priors $\boldsymbol{\mu}$.

In the case where $\boldsymbol{x}_2$ tends to $\boldsymbol{x}_1$ the the components of the covariance matrix, as determined by the kernel, all tend towards the same value, $k(\boldsymbol{x}_1, \boldsymbol{x}_1)$. The result is that as $\boldsymbol{x}_2$ tends towards $\boldsymbol{x}_1$ the GP

requires that y_2 is increasingly likely to be the same value as y_1. At the point where $\boldsymbol{x}_2$ equals $\boldsymbol{x}_1$ the same must be true of y_2 and y_1. This requirement restricts the GP to modelling continuous functions.

The joint distribution can be extended to include the value of the function at N points by simply increasing the size of each component. The resulting joint distribution can be written explicitly as,

$$P(y_1, \ldots, y_N) = \mathcal{N}\left(\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \middle| \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{bmatrix}\right).$$

For the sake of brevity this is written as,

$$P(\boldsymbol{y}_N) = \mathcal{N}(\boldsymbol{y}_N | \boldsymbol{\mu}_N, \bar{\boldsymbol{k}}_{N \times N}). \tag{3.4}$$

While increasing the number of points included in the joint probability distribution results in greater complexity with regards to the mathematics, the core principles of a GP still apply. The kernel encodes the relationship between the values of the random unknown function at any two points. The random unknown function exists over the entirety of the real axis and is continuous.

Throughout the discussion above, the exact form of the kernel function, $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and its effects on the GP have intentionally been left out because such effects are dependent on the choice of kernel. There are a myriad of possible kernel function to choose from, each of which restrict the types of functions that can be modelled by the GP. Here we discuss a Gaussian kernel in the context of a GP and the form of the functions that it models. Each kernel is presented in the context of an example function with differing properties modelled by the choice of kernel.

3.1.1 Gaussian Kernel

To understand the use of a Gaussian kernel, consider measuring the height of the ground at two locations. The smaller the physical distance between the two locations, the smaller the difference in height one could expect to measure. In contrast, as the distance between the two locations increases so does the possible difference in height. This doesn't mean that the height at the two widely separated locations is necessarily different, simply that they are not required to be similar as is the case for two physically close locations. Essentially the relation to be encoded, is that the heights of the ground at two locations that are relatively close, are themselves likely to be similar. From this description it is also evident that the correlation in height is purely dependent on the distance between the two locations and hence is unaffected if both locations are subject to some translation. To model this problem using a GP we need a kernel that is able to encode the condition, the height at two points are increasingly likely to be the same, the smaller the physical separation between them, the Gaussian kernel is a natural choice in this instance.

The Gaussian kernel for a function with a A dimensional input space is given by

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\sum_{\alpha}^{A} \frac{(x_i^\alpha - x_j^\alpha)^2}{{l^\alpha}^2}\right) , \tag{3.5}$$

where x_i^α and x_j^α are the α component of the position vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ respectively, and l^α is a hyper parameter used to scale the α dimension. From this equation we can see that the value of the kernel only depends on the separation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ making it independent of translations in parameter space. Such kernels that only depend on the difference between the two points being referred to as stationary kernels.

With respect to the covariance matrix (3.3) it is immediately apparent that the diagonal components always equal one. It must then be the off diagonal components that govern the relation between y_1 and y_2. When $\boldsymbol{x}_1$ significantly differed from $\boldsymbol{x}_2$, the off axis components reduce towards zero, the covariance matrix tending towards the identity matrix. The values of y_1 and y_2 becoming uncorrelated. In contrast the as $\boldsymbol{x}_2$ tends towards $\boldsymbol{x}_1$ the off diagonal components tend towards one. The entries within the covariance matrix all equalling one, means that not only are y_1 and y_2 correlated but equal. Note: this does not say anything about the exact values of y_1 or y_2, just their relation to one another.

Returning to the problem of modelling the height of the ground at two locations, it is clear that the Gaussian kernel imposes the required condition. As the separation between the locations on the ground decreases, the correlation between the height of the ground measured by the Gaussian kernel increases as needed.

As we shall see later in section 3.3, this form of kernel is useful when classifying objects. A Gaussian kernel can be used when classifying objects to encode the principle that, the smaller the difference between two objects properties (described by some vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$) the greater the chance that they are the same type of object (i.e should be labelled the same).

3.2 Gaussian Processes for regression

Having discussed how a GP acts as a probability distribution over function space, here we apply a GP to solve a regression problem. The aim is to predict the value of some unknown function $f(\boldsymbol{x})$ given a set of N noisy measurements of the function at a series of points. These are referred to individually as training points and the collection as the training set. The values associated with the i^{th} training point are denoted as t_i the noisy measurement of the function, y_i the true value of the function and $\boldsymbol{x}_i$ the location of the training point. For the sake of brevity we define the vectors $\boldsymbol{t}_N$ and $\boldsymbol{y}_N$ as containing the noisy measurements and true value of the function for the training points respectively. The i^{th} entry of $\boldsymbol{t}_N$ and $\boldsymbol{y}_N$ corresponding to t_i and y_i respectively.

The noise on the measurements of the function that make up the training set are assumed to be Gaussian in nature and constant as a function of $\boldsymbol{x}$. The resulting joint distribution on the measurements $\boldsymbol{t}_N$, given the values of the function $\boldsymbol{y}_N$, is described by a multivariate normal distribution,

$$P(\boldsymbol{t}_N|\boldsymbol{y}_N) = \mathcal{N}(\boldsymbol{t}_N|\boldsymbol{y}_N, \beta^{-1}\bar{\boldsymbol{I}}_{N\times N}). \tag{3.6}$$

Here β is a hyper parameter that models the error on the measurements. The exact value of β depends on the system being modelled and can be determined as described in section 3.5. $\bar{\boldsymbol{I}}_{N\times N}$ is an identity matrix of size N.

As described in section 3.1 the joint probability distribution over the true values of the function $\boldsymbol{y}_N$ at the training points is given by the multivariate normal distribution in equation 3.4. The joint probability distribution for the measured values $\boldsymbol{t}_N$ given the locations at which the function is measured, is hence given by convolving equations 3.6 and 3.4 over $\boldsymbol{y}_N$,

$$P(\boldsymbol{t}_N) = \mathcal{N}(\boldsymbol{t}_N|\boldsymbol{\mu}_N, \bar{\boldsymbol{C}}_{N\times N}). \tag{3.7}$$

As previously the components of $\boldsymbol{\mu}_N$ correspond to a prior prediction of the value of the unknown function at each training point. The matrix $\bar{\boldsymbol{C}}_{N\times N}$ is the sum of the covariance matrices in 3.4 and 3.6,

$$\bar{\boldsymbol{C}}_{N\times N} = \beta^{-1}\bar{\boldsymbol{I}}_{N\times N} + \bar{\boldsymbol{k}}_{N\times N}. \tag{3.8}$$

The aim of regression is to predict the measured value of the function t_{N+1} at some new sample point $\boldsymbol{x}_{N+1}$ given the training data and a prior prediction μ_{N+1}. The sample point can be included in the GP model by extending the multivariate normal distribution in equation 3.7,

$$P(t_{N+1}, \boldsymbol{t}_N) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{t}_N \\ t_{N+1}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{\mu}_N \\ \mu_{N+1}\end{bmatrix}, \begin{bmatrix}\bar{\boldsymbol{C}}_{N\times N} & \boldsymbol{k}_N \\ \boldsymbol{k}_N^T & c_{N+1}\end{bmatrix}\right). \tag{3.9}$$

The covariance matrix here being explicitly written as a block matrix separated based on the training and sample points as in 3.8. $\bar{\boldsymbol{C}}_{N\times N}$ is the covariance matrix for the training points. $\boldsymbol{k}_N$ is a vector containing the relations between each training point and the sample point determined by the kernel. The i^{th} entry of $\boldsymbol{k}$ being given by $k_i = k(\boldsymbol{x}_i, \boldsymbol{x}_{N+1})$. The value c is the sum,

$$c_{N+1} = k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) + \beta^{-1}. \tag{3.10}$$

To predict the value of the unknown function at the sample point, the joint distribution in equation 3.9 needs to be rearranged to give the conditional distribution over the measured value at the sample point given that measured at the training points. For a joint multivariate normal distribution, as in Equation 3.9, the conditional distribution over one component is a normal distribution with mean and standard deviation (Bishop & Nasrabadi, 2006),

$$m(\boldsymbol{x}_{N+1}) = \mu_{N+1} + \boldsymbol{k}_N^T \boldsymbol{C}_N^{-1}(\boldsymbol{t}_N - \boldsymbol{\mu}_N) \tag{3.11}$$

$$\sigma^2(\boldsymbol{x}_{N+1}) = c_{N+1} - \boldsymbol{k}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{k}_N \tag{3.12}$$

For the purposes of simplicity the initial prior on the value of the function $\mu(\boldsymbol{x})$ is set to zero for all $\boldsymbol{x}$ reducing equation 3.11 to,

$$m(\boldsymbol{x}_{N+1}) = \boldsymbol{k}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{t}_N. \tag{3.13}$$

Setting the initial prior on the value of the function to zero for all $\boldsymbol{x}$ does not limit the possible values the function can take. This approach is generally taken as there is no prior information as to the value of the unknown function and it reduces the number of calculations needed, particularly for large training sets. Throughout the rest of this work when discussing regression with a GP, the prior on the initial value of the function is set to zero.

Having derived the mean (equation 3.13) and standard deviation (equation 3.12) needed to conduct regression using a GP, the remainder of this section focuses on understanding how the relation between the training data and the sample point affect the prediction of the GP model. First consider a sample point that is increasingly unrelated to all training points. The values of the training points provide decreasing amounts of information as to the value of the sample point. The value of the kernel comparing said unrelated sample point to any training point reflects this lack of information by tending towards zero. Since the vector $\boldsymbol{k}_N$ contains the relation between the sample point and each training point its entries tend towards zero. From equations 3.13 and 3.12 the mean and standard deviation reduce to zero and c respectively. Essentially when there is no relation between the sample point and the training points, there is no information about the value of the function at that point. The result is that, in the absence of any information, the probability distribution over the function defaults to a prior with mean zero and standard deviation c.

Interpreting the output of the GP model for a sample point related to multiple training points is not necessarily straightforward. To help in interpreting the output of the GP model first consider regression with a single training point. For a single training point the mean and standard deviation (equations 3.13 and 3.12 respectively) can be explicitly written,

$$m(\boldsymbol{x}_{1+1}) = \frac{k(\boldsymbol{x}_1, \boldsymbol{x}_{1+1})}{(k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1})}\, t_1, \tag{3.14}$$

$$\sigma^2(\boldsymbol{x}_{1+1}) = (k(\boldsymbol{x}_{1+1}, \boldsymbol{x}_{1+1}) + \beta^{-1}) \; - \; \frac{k(\boldsymbol{x}_1, \boldsymbol{x}_{1+1})}{(k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1})}\, k(\boldsymbol{x}_{1+1}, \boldsymbol{x}_1). \tag{3.15}$$

Here the indices 1 and $1+1$ denote the training and sample point respectively. $m(\boldsymbol{x}_{1+1}$ can be interpreted as a weighted average of the single training point and the prior (in this instance the prior is zero). The weight assigned to the training point is the fraction $k(\boldsymbol{x}_1, \boldsymbol{x}_{1+1})/(k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1})$ and that assigned to the prior is one minus this fraction.

Consider the situation where the position of the sample point tends to that of the training point, $\boldsymbol{x}_{1+1} \rightarrow \boldsymbol{x}_1$

$$\lim_{\boldsymbol{x}_{1+1} \rightarrow \boldsymbol{x}_1} m(\boldsymbol{x}_{1+1}) = \frac{k(\boldsymbol{x}_1, \boldsymbol{x}_1)}{k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1}} \, t_1, \tag{3.16}$$

$$\lim_{\boldsymbol{x}_{1+1} \rightarrow \boldsymbol{x}_1} \sigma^2(\boldsymbol{x}_{1+1}) = \beta^{-1} + (1 - \frac{k(\boldsymbol{x}_1, \boldsymbol{x}_1)}{k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1}}) k(\boldsymbol{x}_1, \boldsymbol{x}_1). \tag{3.17}$$

Since all kernel functions are positive semi-definite, the value output by a kernel when the two inputs are the same must be greater than or equal to zero. The fraction present in both equations then exists over the range zero to one (the noise term also being positive),

$$\frac{k(\boldsymbol{x}_1, \boldsymbol{x}_1)}{k(\boldsymbol{x}_1, \boldsymbol{x}_1) + \beta^{-1}} \in [0, 1]. \tag{3.18}$$

As this fraction approaches one, the mean tends towards the value of the training point and the standard deviation tends towards a minimum, β^{-1}. β^{-1} being the error when measuring the function at any point. In contrast, as the fraction reduces towards zero the mean tends to zero, the prior on the value of the function. The standard deviation increases to a maximum value of $\beta^{-1} + k(\boldsymbol{x}_1, \boldsymbol{x}_1)$. Essentially the fraction is acting as a measure of how reliable the training point is. The more reliable the training point, the closer the fraction is to one. The GP regression model then predicts that the value of the function is most likely the same as the training point. The error on the prediction is then equal to that when measuring the value of the function. The less reliable the training point the closer the fraction to zero, the model reverting to predicting the prior on the value of the function, zero.

For multiple training points the mean in equation 3.11 acts as a weighted average of the value of y_{N+1} predicted by each training point and the prior. Essentially the more informative and reliable a training point, the larger the weighting assigned to said training point when making a prediction. Furthermore, the more informative the relation between a set of training points, the larger the weighting assigned to those training points over other training points and the prior. This reflects there being more information available as to the true value of the function at these strongly related training points.

The standard deviation in equation 3.12 reflects the number and reliability of the training points weighted by their relation to the sample point. By increasing the number and reliability of the training points related to the sample point, equation 3.12 reduces towards β^{-1}. This reduction in the standard deviation reflects the increase in information about the value of the function at the sample point.

As in section 3.1 the description above has not been explicit as to the form of the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This is intentional in order for the description of a GP and its response to training data to be applicable to all choices of kernel. The following subsection gives an explicit example of using a GP to predict an unknown function using Gaussian kernel.

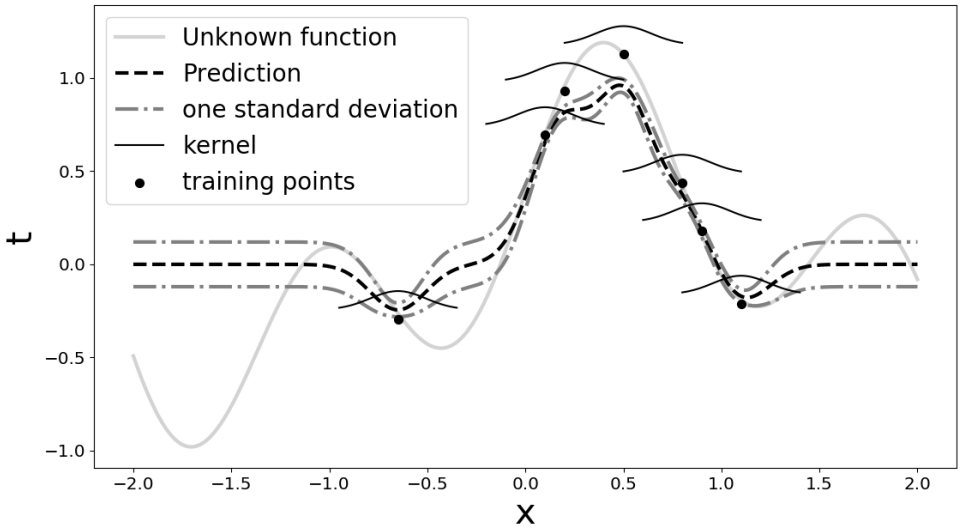


Figure 3.2: Plot illustrating the use of a Gaussian process with a Gaussian kernel function to approximate an unknown function (solid grey line). The dashed line shows the predicted value of t (i.e. the value of the function that would be measured at some point x) given a set of seven training points (black circles). The dot-dashed lines indicate the 1σ confidence interval on the prediction of t_{N+1}. The solid lines illustrate the value of the kernel for each training point (plotted at an arbitrary height). This illustrates that the uncertainty on the predicted value of t is smallest at locations that are similar to the training points.

3.2.1 Regression example with a Gaussian kernel

Consider the unknown function in Figure 3.2 (indicated by the light grey line). The training set consists of seven training points each with some noisy measurement of the function (black circles). Sampling the GP model at a range of x values gives a prediction of the form of the function (dashed black line) along with the one standard deviation error on that prediction (the dot dashed grey line).

It is clear to see that when the sample point has a large separation from (is unrelated to) the training points, the mean output by the GP reverts to the prior of zero and the uncertainty increases to a maximum value. In contrast, when the sample point is close (related) to the isolated training point at $x \approx -0.75$ the mean becomes a weighted average of the training point and the prior. The weight assigned to the training point is given by the fraction in 3.18 and that assigned to the mean (in this case zero) is one minus this value. As the difference in x between the sample and isolated training point decreases (i.e. they become more closely related), the value output by the kernel relating the two points increases to one. The change

in the kernels output increases the weighting of the training point towards to a maximum of

$$\frac{1}{1+\beta^{-1}} \tag{3.19}$$

shifting the predicted value closer to that of the training point. Further, as the difference in x decreases, the uncertainty on the prediction decreases, reflecting an increase in available information.

Where the sample point is close (related) to a number of training points, the mean output by the GP is an average of said training points weighted by the strength of the kernel between the sample point and the training points. The uncertainty on the prediction can also be seen to reduce when sampling in this region as there is a larger amount of information available.

Having discussed the use of a GP for regression the next step is to apply this to binary classification. This will involve using a GP to model the probability of an object having a positive label as some unknown function of its relation to other labelled training objects.

3.3 Gaussian Processes for supervised binary classification

The aim of a binary classifier is to predict if an object does or does not belong to a given class of objects based on its properties. This involves assigning a positive label ($L = 1$) or a negative label ($L = 0$) to the object if it does or does not belong to the given class of objects respectively. Rarely can the label on an object be perfectly determined from its properties. Instead there exists some probability that the object belongs to the class. This probability of the object belonging to the class can be expressed as a function of its properties, $P(L = 1|\boldsymbol{x})$. The aim of a GP binary classifier is to model the probability of some new object having a positive label given a set of N labelled training objects, $P(L_{N+1} = 1|\boldsymbol{L}_N)$. $\boldsymbol{L}_N$ being a vector containing the label assigned to each training object. The dependence of the probability on the the properties of the new and training objects is not explicitly included here for reasons of brevity.

Calculating $P(L_{N+1} = 1|\boldsymbol{L}_N)$ using the GP model described in section 3.2 presents two problems; i) a GP model predicts functions over the entire real axis, while the probability $P(L_{N+1} = 1|\boldsymbol{L}_N)$, exists over the range zero to one and ii) the labels assigned to the training data are not measurements of the unknown probability, $P(L_i = 1|\boldsymbol{x}_i)$, instead they are a random draw from said probability. To solve the first problem, a new value, a_i, is defined for each object such that the probability for a new object object having a positive label can be written as,

$$P(L = 1|a_i) = \sigma(a_i). \tag{3.20}$$

Here $\sigma(a_i)$ is a logistic sigmoid function that maps the value of a_i from a space spanning the entire real axis to a probability space (one continuous over the range zero to one). Since the value a_i exists in a

space spanning the entire real axis, the associated variable for a new object, a_{N+1}, can be predicted using a GP,

$$P(a_{N+1}|\boldsymbol{a}_N) = \mathcal{N}(a_{N+1} \mid \boldsymbol{k}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{a}_N,\ c_{N+1} - \boldsymbol{k}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{k}_N). \tag{3.21}$$

Here $\boldsymbol{a}_N$ contains the values of a_i associated with the training objects. The remaining variables being the same as those defined in section 3.2.

The probability of the new object having a positive label given the values $\boldsymbol{a}_N$ associated with the training objects can be calculated by marginalising over a_{N+1},

$$P(L_{N+1} = 1|\boldsymbol{a}_N) = \int P(L_{N+1} = 1|a_{N+1})\ P(a_{N+1}|\boldsymbol{a}_N)\ da_{N+1}\ . \tag{3.22}$$

Where $P(L_{N+1} = 1|a_{N+1})$ is the logistic sigmoid of a_{N+1} (equation 3.20) and $P(a_{N+1}|\boldsymbol{a}_N)$ is GP prediction over the value of a_{N+1} given the values associated with the training objects, $\boldsymbol{a}_N$ (equation 3.21). Since the values of $\boldsymbol{a}_N$ are not explicitly known it is necessary to marginalise over these values as well giving,

$$P(L_{N+1} = 1|\boldsymbol{L}_N) = \int\int P(L_{N+1} = 1|a_{N+1})\ P(a_{N+1}|\boldsymbol{a}_N)\ P(\boldsymbol{a}_N|\boldsymbol{L}_N)\ da_{N+1}\ d\boldsymbol{a}_N\ . \tag{3.23}$$

Here the probability $P(\boldsymbol{a}_N|\boldsymbol{L}_N)$ is is given by,

$$P(\boldsymbol{a}_N|\boldsymbol{L}_N) = \Pi_{i=1}^N\ \sigma^{-1}(L_i) \tag{3.24}$$

While using a sigmoid function to map to a new space over the entire real axis has solved the first of the two problems outlined previously (a GP only modelling functions over the entire real axis), it has not solved the second. The second problem being that the training values, L_N, are not measurements of the true probability of the training objects having a positive label given their properties, but samples from it. This results in a problem when calculating $P(\boldsymbol{a}_N|\boldsymbol{L}_N)$, as the inverse sigmoid function is not defined for the values of one and zero used to label the training data. The use of a sigmoid function presents an additional issue when marginalising over a_{N+1} as the integral in equation 3.22 is intractable.

To calculate $P(L_{N+1} = 1|\boldsymbol{L}_N)$ from equation 3.23 it is hence necessary to use an approximate method. There exist three techniques to approximating $P(L_{N+1} = 1|\boldsymbol{L}_N)$ (Bishop & Nasrabadi, 2006). The first uses a form of variation inference to approximate the sigmoid function as a product of Gaussians (Gibbs & Mackay, 2000). The second uses a Laplace approximation to model the distribution over a_{N+1} as a Gaussian (Bishop & Nasrabadi, 2006). The final approach, and that used within this work, makes use of expectation propagation to calculate $P(L_{N+1} = 1|\boldsymbol{L}_N)$ (Opper & Winther, 2000; Minka, 2001; Williams & Rasmussen, 2006).

While this complicates the calculations, the underlying theory of a GP still applies. The labelled training objects act as the measurements of the function, in this case the probability of an object having a positive label. The GP then uses the kernel to predict the probability for the new object.

3.4 Adaption for imperfectly labelled training data

When creating a labelled training set for the purpose of classification there is often uncertainty on the labels assigned to each object. As it stands the GP binary classifier described in the previous section assumes perfect labelling of the training data. There exist a number of classification problems where ML would be beneficial but its application is limited by the limited availability of perfectly labelled training data for example medical diagnostic tasks where existing diagnosis test have a known uncertainty, the classification of astrophysical sources where limitations within the data make the labelling of all sources with 100% certainty impossible. The XXL X-ray source catalogue described in section 2.1 is one such astrophysical example where there exists no definitive classification of sources on which to train. In order to apply a GP binary classifier to select galaxy cluster candidates from the XXL X-ray source catalogue we propose a method to adapt the binary classifier to account for imperfectly labelled training data.

Given the uncertainty on each of the training labels, the probability for a single training object being assigned a positive label is written $P(L_i = 1|u_i) = u_i$. The joint probability distribution for the entire training set can then be written as the multiplication of the individual probabilities,

$$P(\boldsymbol{L}_N|\boldsymbol{u}_N) = \Pi_i^N \; u_i^{L_i} \; (1-u_i)^{1-L_i} \; . \tag{3.25}$$

$\boldsymbol{u}_N$ being a vector containing the probability that each object has a positive label, u_i.

The probability of a new object having a positive label given N training objects with some uncertainty on their labels, is then given by marginalising over all possible combination of labels $\boldsymbol{L}_N$.

$$P(L_{N+1} = 1|\boldsymbol{u}_N) = \sum_{\boldsymbol{L}_N} P(L_{N+1} = 1|\boldsymbol{L}_N) P(\boldsymbol{L}_N|\boldsymbol{u}_N) \tag{3.26}$$

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 3.3.

The sum in equation 3.26 scales as the number of possible combinations of $\boldsymbol{L}_N$. Since each entry in $\boldsymbol{L}_N$ has two possible values (one or zero) this results in 2^N combinations, each of which is summed over in equation 3.26. For any reasonably sized training set the sum becomes impractical to calculate analytically. Instead it is necessary to approximate the sum using a Monte Carlo approach where the entirety of the contents of the vector $\boldsymbol{L}_N$ is sampled from $P(\boldsymbol{L}_N|\boldsymbol{u}_N)$.

A consequence of this approximation, as we will see in subsection 4.2, is that the estimated probability $P(L_{N+1}|\boldsymbol{u}_N)$ output by the GP binary classifier is no longer a direct estimate of the probability. The value output by the GP instead acts as a nonlinear measure of the true probability of the object having a positive label. In other words, if the GP binary classifier returns a value of 0.4, this doesn't mean the new object in question has a 40% probability of having a positive label. Instead, the value relates in some non-linear way to how likely the GP binary classifier believes that the object should have a positive label. Consequently, the estimated probability produced by the GP should be treated as a figure of merit. To make this distinction clear the output of this adapted GP binary classifier is referred to as a "confidence value" throughout the rest of this work. The key takeaway here is that an object with a confidence value to close to 1 is more likely to have a positive label, while an object with a confidence value close to 0 is less likely to have a positive label. Objects can hence be classified on the basis of their confidence value.

3.5 Hyper parameter optimisation

When describing the use of a GP model for regression in section 3.2 the noise term β was introduced as a hyper parameter. The exact value of β was said to be dependent on the system being modelled. Further when describing the Gaussian kernels in subsection 3.1.1, the various constants were also referred to as hyper parameters. Hyper parameters are the parameters of a machine learning model that control the output of the model, but are not updated as part of the training process. Examples of hyper parameters include the number and size of each layer in a neural network, the number of decision trees in a random forest and the length scales, l^{α}, in the Gaussian kernel (equation 3.5). Since these values have a significant impact on the training and hence results of a machine learning model they are often optimised to the best results for the problem being solved.

In the case of a GP for regression, the hyper parameters denoted by $\boldsymbol{\theta}$ can be optimised by maximising the joint probability of the training data. The joint distribution over the training data from equation 3.7 can be written to explicitly show its dependence on the the hyper parameters as,

$$P(\boldsymbol{t}_N|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{t}_N|\boldsymbol{0}, \bar{\boldsymbol{C}}_{N\times N}(\boldsymbol{\theta})). \tag{3.27}$$

Here the mean of the distribution has been set to zero for all training points.

The same approach can be taken when applying a GP to solve a binary classification problem as described in section 3.3. The joint probability of the training labels, $\boldsymbol{L}_N$, given the hyper parameters, $\boldsymbol{\theta}$, takes the form,

$$P(\boldsymbol{L}_N|\boldsymbol{\theta}) = \int P(\boldsymbol{L}_N|\boldsymbol{a}_N)P(\boldsymbol{a}_N|\boldsymbol{\theta})d\boldsymbol{a}_N. \tag{3.28}$$

Here $P(\boldsymbol{L}_N|\boldsymbol{a}_N)$ is the multiplication of the Bernoulli distribution of each $\sigma(a_i)$ (this is necessary to account for both possibilities, $L_i = 1$ and $L_i = 0$)

$$P(\boldsymbol{L}_N|\boldsymbol{a}_N) = \Pi_i^N \sigma(a_i)^{L_i}(1-\sigma(a_i))^{1-L_i} \tag{3.29}$$

and $P(\boldsymbol{a}_N|\boldsymbol{\theta})$ is the joint normal distribution for N points given by writing Equation 3.7 as an explicit function of the hyper parameters $\boldsymbol{\theta}$,

$$P(\boldsymbol{a}_N|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{a}_N|\boldsymbol{0}, \boldsymbol{C}_N(\boldsymbol{\theta})). \tag{3.30}$$

The solution to maximising Equation 3.28 is dependent on the approximation used for $P(L_{N+1}|\boldsymbol{L}_N)$. The work in later chapters uses expectation propagation (Williams & Rasmussen, 2006) to approximate the integral in Equation 3.28, implemented within the GPy (GPy, 2012) python package.

3.5.1 Hyper parameter optimisation for imperfectly labelled training data

To optimise the hyper parameters for binary classification where there is some uncertainty on the training labels as described in section 3.4, there is no clear joint probability that can be maximised for the training data. Instead we chose to take the average of the hyper parameters optimised to each set of labels $\boldsymbol{L}_N$ weighted by the probability of said set of labels given the uncertainty u_i,

$$\bar{\boldsymbol{\theta}} = \sum_{\boldsymbol{L}_N} \boldsymbol{\theta}(\boldsymbol{L}_N) P(\boldsymbol{L}_N|\boldsymbol{u}_N) \tag{3.31}$$

The function $\boldsymbol{\theta}(\boldsymbol{L}_N)$ denotes the values of the hyper parameters optimised to a given set of labels $\boldsymbol{L}_N$. As when predicting the label of an object given imperfectly labelled training objects (section 3.4) it is necessary to approximate this sum using a Monte Carlo approach.

3.5.2 Automatic Relevance Determination

Having optimised the hyper parameters to a training set, the values for each hyper parameter must inherently contain some information about the training data. The simplest example of this is the noise term β, included in a GP when solving a regression problem. Since this is used in equation 3.6 to encode the measurement noise on the training data, it stands to reason that the value of β determined by optimising the hyper parameters acts as a measurement of said noise. This shows that by optimising the hyper parameters to the training data, a GP is able to extract information about said training data.

By optimising those hyper parameters within the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that govern the response to changes in individual dimensions of $\boldsymbol{x}$, it is possible to infer the relevance of each dimension with respect to the output of GP. A dimension is considered to be increasingly relevant to the prediction of the GP the larger the potential change in the prediction given a small change in said dimension. In contrast, if a

large change in a given dimension produces little to no change in the prediction of the GP it is considered irrelevant. An alternative description being, the larger the maximum gradient in the prediction of the GP for a given dimension, the greater the relevance of said dimension in the prediction by the GP. Since a kernel's hyper parameters govern how the output of the kernel and in turn the GP's response to changes in each parameter, these must relate in some way to the relevance of the parameter.

The exact relation between a kernel's hyper parameter and the relevance of a given dimension depends on the kernel function. Here we give an example of how the hyper parameters in a Gaussian kernel relate to the relevance of each dimension. A Gaussian kernel takes the form described in equation 3.5,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\sum_{\alpha}^{A} \frac{(x_i^\alpha - x_j^\alpha)^2}{{l^\alpha}^2}\right) . \tag{3.32}$$

Where there are A dimensions, the variance of the distribution is given by V and each dimension has some characteristic length scale l^α.

The length scales for each dimension in a Gaussian kernel are what determine the response of the kernel (and hence GP) to a change in a given dimension. A short length scale for a particular dimension means the Gaussian kernel is more tightly peaked in that dimension, resulting in a more rapid change in the GP's prediction in response to variations in that dimension. The opposite is true for a large length scale. Hence the shorter the length scale l^α for a given dimension, the more relevant that dimension is in determining the output of the GP.

The key point to take away here, is that by optimising the hyper parameters that govern the response of the kernel and hence GP to changes in each dimension. We will see later that this can be used to determine which of an objects properties are most informative as to its classification. This is a useful tool when interpreting the results of the GP, as it allows direct comparison of the relative importance of each. In order to allow direct comparisons between dimensions with different units and dynamic ranges, it is necessary to normalise each. By normalising a dimension it removes any scaling due to the choice of units or dynamic range of the data. The relevance values for each dimension post normalisation are then purely in terms of the distribution of data over said dimension, such that they can be directly compared.

3.6 Summary

This chapter has outlined the core principles of a GP as a probability distribution over functions and how the choice of kernel encodes additional information about the form of these functions. It described how a GP is used to model an unknown function to solve a regression problem, providing an example using a Gaussian kernel. Using the GP model for regression as a starting point this chapter showed

how a GP can be used to solve a binary classification problem where there is a set of perfectly labelled training data. Following this, a method for adapting a GP binary classifier has been outlined to account for uncertainties on the labels of the training data. In addition to outlining how a GP can be used to model both regression and classification, this chapter has described how hyper parameter optimisation can be used to not only optimise the hyper parameters to the training data, but provide information as to the relevance of each dimension in determining the output of a GP model. This is particularly useful when physically interpreting the behaviour of a GP model.

In later chapters, the GP model adapted to account for uncertainties on the labels of the training data will be applied to identify galaxy cluster candidates from the XXL X-ray source catalogue. Further analysis of the output of the model applied to this problem will also make use of automatic relevance determination to identify relevant parameters and aid interpretation.

4

Adapted Gaussian Process Binary Classifier Applied to Select Galaxy Clusters From the XXL X-ray Source Catalogue

Within astronomy, novel astrophysical objects are detected and catalogued by conducting surveys of all or part of the sky. The exact design of a survey is heavily dependent on the type of object being searched for, but all aim to maximise the total number of target objects identified whilst minimising the number of objects or spurious detections miss identified as a target object. Those surveys searching for galaxy clusters in particular needing to produce large galaxy cluster samples while minimising the number of non-galaxy cluster objects accidentally selected. An additional requirement for a successful galaxy cluster survey is the need for a well understood selection function in order for the resulting cluster catalogue to be used for cosmological studies (Nord et al., 2008).

Object selection as part of astrophysical surveys can generally be described in the following steps;

1. The identification of potential sources from observations using software such as source extractor (Bertin & Arnouts, 1996).

2. The characterisation of each source into a set of measured properties, such as flux, colour and surface brightness. For XXL this characterisation is conducted by fitting four surface brightness profiles to each source (Faccioli et al., 2018).
3. The selection of a subset of source based on some of their measured properties, the aim being to select a specific type of source. Typically this involves simple selection criteria such as requiring a source to have a value for a specific measured property within some range, such as the cuts on EXT and EXT_LIKE used to define the C1 and C2 samples in XXL (Pierre et al., 2016).
4. Follow up investigation of the selected objects to confirm they are genuine detections of the specific type of object being looked for.

With the scale of astrophysical surveys increasing, we are reaching a point where the volume of data produced by observations introduces significant computational difficulty when conducting the first three of these steps, a problem that existing methods struggle to solve. One solution is the use of machine learning (ML) methods due to their ability to quickly and efficiently process large complex data sets. The different types and adaptability of ML methods mean that there exist a large number of methods applicable to each of the first three steps when selecting objects from astrophysical surveys. A single ML model could in fact be used to achieve the three steps at once, however doing so greatly increases the models complexity, increasing computation time and the difficulty in interpreting the selection process learnt by the model. Such an increase in complexity generally outweighs any benefits to using a single model for all three steps when three separate models can achieve the same or better results (each being individually tuned to the step they are applied to).

The final step, source follow up, is also affected by the increase in data. One hoped-for benefit of using ML methods in the earlier steps is to reduce the number of unwanted objects reaching this stage and hence the time spent conducting follow up investigations on such objects.

The aim of the work detailed in this chapter is to use the adapted GP binary classifier outlined in section 3.4 to produce a sample of galaxy cluster candidates from the XXL cluster survey (Pierre et al., 2016) (described in section 2.1). The existing selection criteria being used to label sources with some known uncertainty of being a galaxy cluster. The hope being that a GP selected sample is able to accurately recover those galaxy cluster candidates previously identified by the XAMIN pipeline along with a number of new cluster candidates. Clusters are to be identified from the latest version of the north and south XXL catalogues produced by version 4.3 of the XAMIN pipeline.

To achieve its aim this chapter first outlines the prepossessing of the XXL source catalogue and how the GP model was trained on the North and South catalogues to identify galaxy cluster candidates. The performance of the GP is subsequently analysed by investigating its output on a number of X-ray source

Cluster Sample	EXT (arcseconds)	EXT_LIKE
C1	> 5	> 33
C2	> 5	15 to 33

Table 4.1: 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 (Pierre et al., 2016) and should be considered approximate when applied to version 4.3 as done here. The standard XAMIN pipeline also applies cuts on EXT_DET_STAT (Faccioli et al., 2018) however since we find this has no significant impact on the C1 and C2 samples they are not used in this work.

catalogues, the North and South XXL catalogues, the simulated XXL catalogue and the catalogue of XXL source matched to a CAMIRA source. A sub-sample of the XXL North source catalogue was visually inspected producing a set of sources labelled as a likely detection of a galaxy cluster or not. The labelled subset was then used to asses source samples selected based on the output of the GP.

4.1 Training the adapted Gaussian Process binary classifier model on the XXL X-ray source catalogue

Before training the adapted GP binary classifier model (described in subsection 3.4) to select galaxy clusters from the North and South XXL catalogues(see section 2.1), it is necessary to prepare the data contained within said catalogues. The catalogues were prepared by; i) producing a set of training labels for all sources based on the existing C1 and C2 samples, ii) selecting the measured source properties over which the GP model is to be trained and iii) cleaning the catalogues by removing those sources with missing information or non-physical measured properties.

In order to produce the training labels for the GP model, first consider the existing method used to select galaxy cluster candidates from the XXL catalogues. Sources are classified by applying cuts on three measured properties, EXT, EXT_LIKE and EXT_DET_STAT. We find that EXT_DET_STAT has little to no impact on the classification and so do not use it within this work. The remaining selection criteria are listed in table 4.1 and are designed to produce the C1 and C2 galaxy cluster samples with purities of ~ 1 and 0.50 respectively (Pierre et al., 2016). The values reported in table 4.1 were determined for version 3.3 of the XAMIN using simulated data (Adami et al., 2018) and should be considered approximate when applied to version 4.3 of the XXL catalogues used here. Of the remaining non-C1C2 sources (those not included in the C1 or C2 samples), galaxy clusters make up a negligible fraction, the vast majority of sources being either AGN or spurious detections. The purity of the non-C1C2 source sample with respect to galaxy clusters is hence approximately zero.

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

Table 4.2: 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.

The purity of each source sample directly relates to the probability of a source contained by said sample being a galaxy cluster. This information can be fed into the GP binary classifier adapted for imperfectly labelled training data as the uncertainty on the labels. The simplest approach would be to set the initial probability of a source being a cluster (u_i in equation 3.25) equal to the purity of the source sample it is included in, i.e 1 for a C1 0.5 for a C2 and 0 for a non-C1C2 source. While this approach to labelling sources would be sufficient, we elect to use the following values for the initial probability of a source being a galaxy cluster;

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

While the estimated purity of the non-C1C2 is effectively zero, in reality the distribution of u_i is not uniform, there being a non-zero probability of a source that is close to (but outside) the C1 and C2 regions being a cluster. For this reason we assigned a probability of $u_i = 0.05$ for non-C1C2sources to be a cluster. This decision allows the adapted GP classifier more flexibility to identify such sources as clusters if their other source parameters are sufficiently similar to those of sources labelled as a C1 or C2. This is consistent with experience that in different iterations of the XXL pipeline, sources can move across classification boundaries, such that genuine detections of clusters must exist outside of the C1 and C2 regions. In our approach, if we assigned $u_i = 0$, then none of the non-C1C2sources 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 cluster candidates by the GP classifier. Similar logic is applied when setting $u = 0.95$ for C1 objects despite the C1 sample having an effective purity of one.

Having labelled the sources, the next step is to determine which of the source properties measured by XAMIN are to be provided to the classifier model. As described in section 2.1.1, version 4.3 of the XAMIN pipeline characterises each source by fitting four surface brightness profiles, (table 4.2) each

Parameter	Description
EXT_LIKE∗	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 how likely 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 how likely 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 how likely 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 how likely 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

Table 4.3: List of the source properties measured by the 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 4.2). 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 Gaussian process model is trained.

corresponding to a different astrophysical object or pair of objects. The fitting process produces a set of measured source properties for each of the four surface brightness profiles measured in both the $0.5 - 2$ keV (soft) and $2 - 12$keV (hard) energy bands. In order to reduce the complexity of the data (and hence computation time) it is necessary to reduce the number of measured properties used to characterise each source. The first reduction in the number of source properties is to consider only those properties measured using data from the soft energy band. The number of source properties is further reduced by removing those that are not informative as to the nature of the source (i.e the position on the sky). This leaves the source properties listed in table 4.3. The background rate values for each source were determined straightforwardly from the XAMIN output.

Because the EXT and EXT_LIKE properties were used to define the training labels these properties should not be provided to the GP. Were the GP to be provided with the EXT and EXT_LIKE values for each source, it would simply identify that the source label correlates with EXT and EXT_LIKE and return a smoothed version of the uncertainty on the training labels. Further the GP would only consider EXT and EXT_LIKE when assigning confidence values to the sources (based on the results of automatic relevance determination 3.5.2). While this means that the GP is able to recover the initial C1 and C2 samples, it is not providing any improvement or new information with respect to selecting previously missed galaxy clusters.

The _DET_STAT source properties are also not provided to the GP for a similar reason. These correspond to the likelihood that the given type of or pare of objects with physical properties as determined by fitting the appropriate model would be detected by XAMIN. Since the EXT_LIKE value is determined from EXT_DET_STAT and PNT_DET_STAT the GP would again learn to directly replicate the selection criteria that depend on EXT_LIKE. This leaves those measured source properties listed in bold in table 4.3 as the ones provided to the GP.

The final stage prior to training the GP model is to clean and normalise the data. This process involves first removing any source that for whatever reason, is missing one or more of the measured properties provided to the GP, EXT and EXT_LIKE. EXT and EXT_LIKE are included during this specific part of the processing step as they are needed to label each source. EXT and EXT_LIKE are not considered in the remaining steps.

Those sources with one or more measured source properties with a value less than or equal to zero are removed, such values being nonphysical. An inspection of the images for these sources indicates that these are (low significance or artefact-generated) 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. The distribution of the remaining sources among the C1 C2 and non-C1C2 samples are listed in table 4.4.

XXL Catalogue	C1 sources	C2 sources	non-C1C2 sources	total sources
North	139	109	23,378	23,626
South	86	81	17,902	18,069

Table 4.4: The number of sources within $V4.3$ of the XXL catalogues after the removal of those sources with one or more source properties that are either missing or nonphysical. Included are the number of sources within each of the sub-samples.

In order to make use of automatic relevance determination (ARD, section 3.5.2) to directly compare the relative importance of the different measured source properties, it is necessary to normalise the data. Each source property was normalised by dividing the measured value for each source by the standard deviation over the measured property for all sources. While a number of different normalisation methods were tested (such as dividing by the maximum value for a given source property), it was found that dividing by the standard deviation is needed for this specific data set in order to directly compare the relevance of the measured source properties.

Having prepared the data as outlined above, the adapted GP binary classifier was trained separately on the North and South catalogues to avoid any field-dependent differences affecting the results. For both catalogues the training process consisted of ten batches of ten Monte Carlo iteration for a total of 100 iterations. Each batch taking ~ 26 hours on a single node of the University of Bristol's BlueCrystal phase 4 supercomputer.

4.2 Classifier results for the XXL X-ray catalogue

Having trained two versions of the GP model, one on the North and the other on the South XXL catalogues, each source is assigned two confidence values, one by each of the models. The resulting confidence values for both XXL catalogues are colour-coded in a plot of EXT and EXT_LIKE in figure 4.1. It is clear that the distributions of confidence values within this space follow a smoothed version of the initial probabilities of a source being a cluster, assigned based on the C1 and C2 classifications. The high-confidence values broadly occupy the combined C1 and C2 regions. Since truly extended sources are expected to have a higher EXT and EXT_LIKE this indicates that the GP is performing as expected, despite not having access to sources EXT and EXT_LIKE values.

There also exist a number of non-C1C2 sources assigned a high confidence value despite being significantly different in EXT and EXT_LIKE from the combined C1 and C2 region, including some sources with an EXT_LIKE value of zero. If the GP has worked as expected it implies that such sources have an increased likelihood of being a galaxy cluster compared to others labelled as a non-C1C2 source. It is

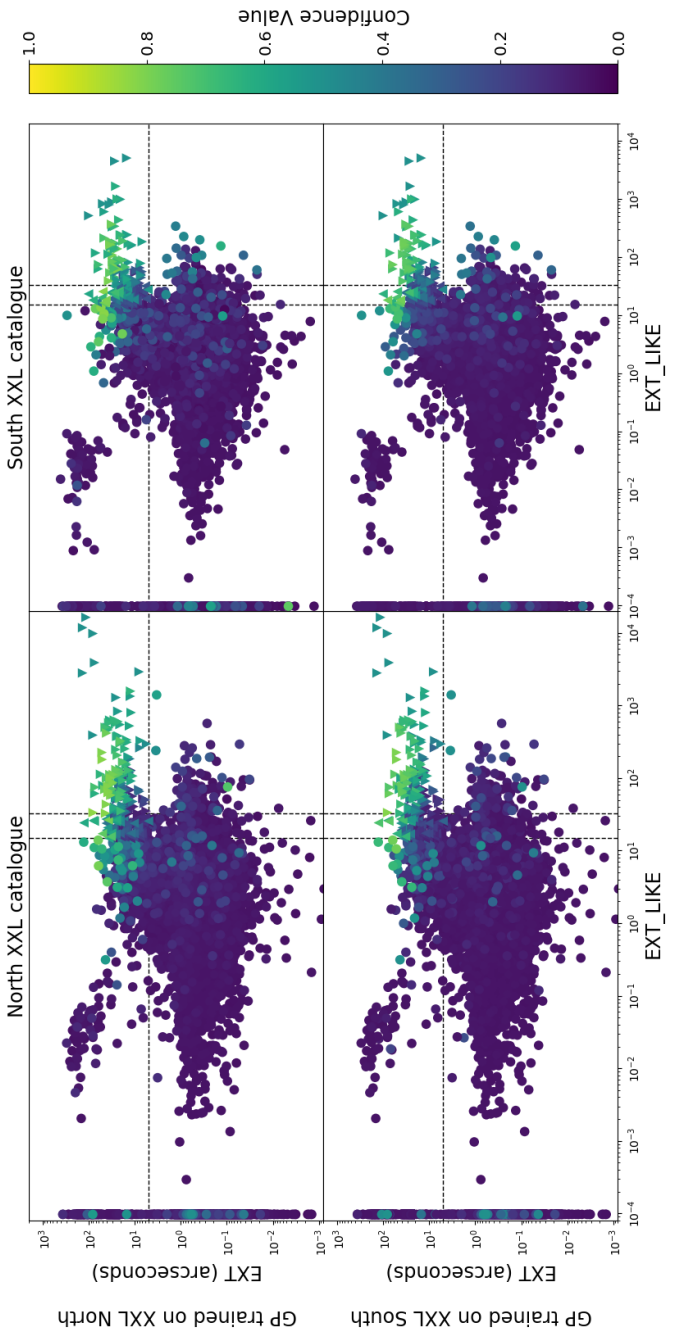


Figure 4.1: Confidence values assigned by the GP trained on the North (top row) and South (bottom row) XXL catalogues as a function of EXT and EXT_LIKE for the North (left column) and South (right column) catalogues. C1 and C2 sources are plotted as triangles pointing down and up respectively, with all 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 4.1. 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. It is clear that the GP is assigning higher confidence values to sources with grater measured EXT and EXT_LIKE values despite not being privy to this information. The result is that the GP is able to highlight sources with an increased probability of being a galaxy cluster that were not previously selected by XAMIN as a C1 or C2 source.

important to remember here that while an increase in confidence value indicates an increased likelihood for a source to be a galaxy cluster, there is not a direct linear relation between the two. Such sources could not be selected by simply expanding the C1 or C2 regions in EXT, EXT_LIKE space without significantly reducing the fraction of real galaxy clusters in the resulting sample. This ability of the GP to identify sources significantly far from the C1 and C2 regions is a significant benefit of using a GP, should (as will be shown later) such sources contain a high fraction of galaxy cluster detections.

Figure 4.1 can also be used to test if the GP model has over-fit to the training data. If the GP model had over-fit to either of the North or South catalogues when trained on them it would be apparent from figure 4.1 in two ways. The first being that the confidence values assigned to the sources used to train the GP would closely trace the initial uncertainty on the labels, showing sharp increases at the C1 and C2 boundaries. This is not seen in the top left and bottom right plots in figure 4.1 (those depicting the distribution of confidence values assigned to sources used to train the GP), indicating that the GP has not over-fit when trained on eather XXL catalogue.

The second indication of overfitting would be that the confidence value assigned to those sources not used to train the GP would be close to 0.5 (the default value assigned by the GP in the absence of similar training points, as described in subsection 3.3). Instead, and as will be shown in more detail later (subsection 4.2.1), the confidence values assigned to a source by both the GP trained on the catalogue containing and trained on the catalogue not containing the source are of similar value. This further indicates that the GP when trained on either catalogue has not over-fit.

Figure 4.2 depicts the distribution of sources as a function of the confidence value assigned by the GP when trained on the North catalogue and when trained on the South catalogue. The vast majority of sources have comparatively low confidence values. This is to be expected as the XXL catalogue is dominated by detections of AGN ($\sim 98\%$ of the sources in the catalogue are expected to be AGN, Pierre et al., 2016). Although the nature of the GP means the confidence values don't map to the probabilities used to assign training labels, one might expect the confidence values of the C1 sources to cluster around ~ 0.95, C2's 0.50 and non-C1C2's ~ 0.05. The distributions of the confidence values assigned to the C1 and C2 sources are instead broad. This broad distribution is 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 value for sources with higher initial probabilities through the action of the GP.

The most interesting sources are those non-C1C2 sources whose confidence value is higher than that of the majority of non-C1C2 sources (~ 0.05), the increase in confidence value predominantly occurring due to a sources similarity withC1 and C2 objects in the 19-dimensional parameter space. Such high confidence non-C1C2 sources are identified by the GP as potential galaxy clusters despite not being selected

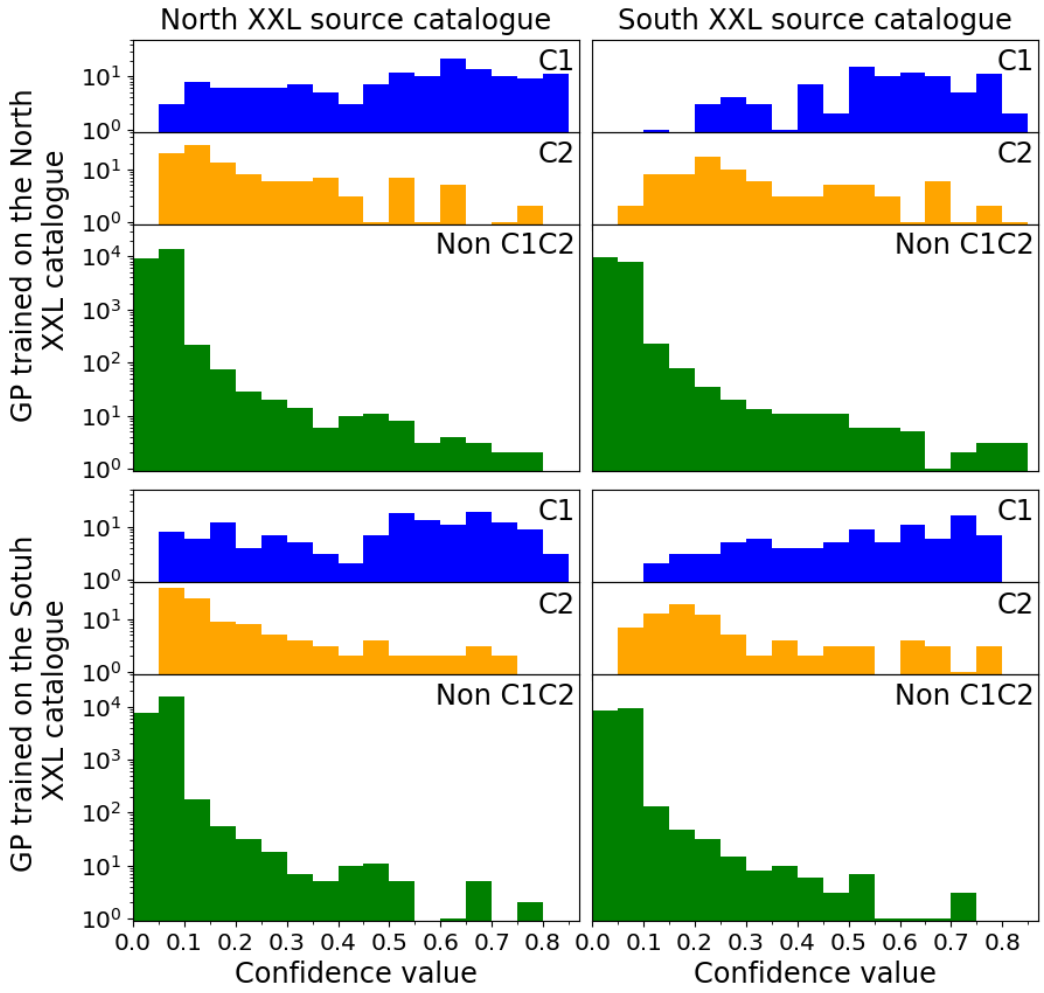


Figure 4.2: Distribution of confidence values for the North XXL catalogue (left column) and South catalogue (right column) assigned by the GP when trained on the North catalogue (top row) and South catalogue (bottom row). Within each Figure the upper, middle and lower panels show the distribution of confidence values for the C1, C2 and non-C1C2 sources respectively.

by the standard XAMIN pipeline. These sources can be seen as the portion of the green distribution in Figure 4.2, which extends significantly above a confidence value of 0.05. If the GP has performed as intended, this population will contain a higher fraction of real clusters missed by the simpleC1 and C2 classification than would a random sampling of non-C1C2 sources (this is indeed the case, as will be demonstrated in subsection 4.5).

The precision with which the GP determines the confidence value is an important factor when assessing the reliability of using said confidence values as a method to select galaxy cluster candidates. In the instance where the GP assigns confidence values with low precision, samples selected on confidence value are highly variable with respect to the sources they contain, reducing their reliability and reproducibility. In order for the GP model to produce reliable samples of galaxy cluster candidates it needs to be able to assign confidence values with high precision. To asses the precision of the confidence values assigned by the GP, figure 4.3 shows the standard error of the confidence value as a function of confidence value. To calculate the standard error on the confidence value after 100 Monte Carlo iterations the standard deviation is calculated for the confidence value after ten Monte Carlo iterations and then divided by the square root of ten.

Figure 4.3 shows that the standard error on the confidence value is, for the vast majority of sources, significantly smaller than the confidence value. By calculating the average of the standard error for sources binned on confidence value (the black solid line in figure 4.3) it is clear that the the average standard error increase to a maximum value around 0.5 before decreasing. The average standard error varies as a function of confidence value in the way seen in figure 4.3 due to the nonlinear mapping imposed by the sigmoid function when converting the GP output from an infinite space to the range zero to one. Essentially the standard error is uniform in the infinite space, then due to the non-linear mapping the error is more compressed as the confidence value moves away from 0.5.

In figure 4.3, there exist an interesting population of predominantly C1 objects with a confidence value of ~ 0.5 and a lower standard error than the main population of sources with similar confidence values. The most obvious explanation for this population of objects is that they are dissimilar to those objects used to train the GP, hence the GP defaults to the prior estimate for the confidence value of 0.5. This fails to explain why such a population includes sources that were included in the training data (top left and bottom right subplots in figure 4.3). When assigning a confidence value to a source that is included in the training set, the source must be related to at least one training source, itself. In such a situation one could reasonably assume that the confidence value assigned to a source on which the GP was trained would return the initial probability of said source being a cluster (0.95, 0.5 and 0.05 for a C1, C2 or non-C1C2 source respectively). This is clearly not the case. From the existence of this population of source with confidence values of ~ 0.5 and low standard errors we must conclude that, for an isolated source

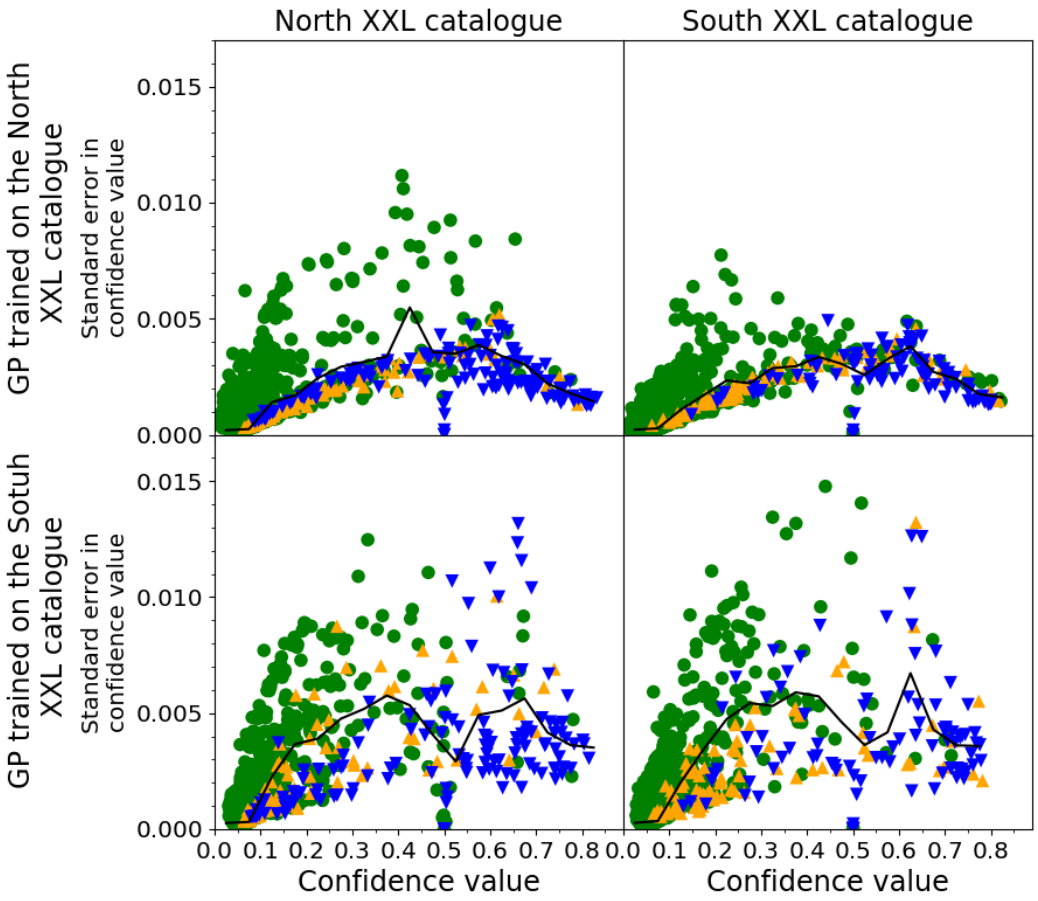


Figure 4.3: Standard deviation in confidence value as a function of the confidence values for the North XXL catalogue (left column) and South catalogue (right column) assigned by the GP when trained on the North catalogue (top row) and South catalogue (bottom row). Within each figure C1 and C2 sources are plotted as blue triangles pointing down and green triangles pointing up respectively, the non-C1C2 sources being plotted as green circles. The solid black line is the mean standard deviation in confidence value for all sources binned on confidence with a bin size of 0.05. Comparing the standard error in confidence value between that assigned by the GP when trained on the North or South catalogue shows a clear increase for the South trained GP. This indicates there exists some difference between the North and South source catalogues that is impacting the output of the GP (see subsubsection 4.2.1).

included in the training set, the GP model adapted for imperfectly labelled data favours the prior value of 0.5 inherent to a GP binary classifier over the initial probability of a source having a positive label.

To summarise the GP binary classifier is able, despite being blind to EXT and EXT_LIKE, to recover the existing C1 and C2 sources along with identifying a number of other sources as potential galaxy cluster candidates. This is achieved without over fitting to the training data (the model does not simply replicated the C1and C2 samples and works when applied to the XXL catalogue it was not trained on) meaning there must exist some information within the measured source properties provided to the GP pertinent to a source being a galaxy cluster.

Up to this point when describing the results of the GP we have intentionally focused on the general properties of the confidence values assigned to sources. In the following subsection we compare the differences between the confidence values assigned by the GP when trained on the North or South XXL source catalogues.

4.2.1 Comparing the results of the classifier when trained on the North and South XXL catalogues

Figure 4.4 compares the confidence values assigned to sources in the North (top) and South (bottom) catalogues by the GP when trained separately on the North and South catalogues. It is immediately apparent that there exists a general correlation between the confidence values assigned to a source by the GP trained on the North or South catalogues. For the North catalogue, the Pearson's correlation coefficient between the confidence values assigned by the GP trained on the North and trained on the South (top panel in Figure 4.4) was 0.95. For the south catalogue (bottom panel of Figure 4.4) it is 0.91. Given the sizes of both catalogues (post cleaning) these coefficients correspond to Student T-values of 425 and 266. The probability of achieving a larger T-value by chance from uncorrelated data is effectively zero. This is further evidence that over-fitting is not an issue, and that training the GP on one field and applying it to another produces reliable results.

Despite the correlation found above, figure 4.4 shows a general tendency for the GP trained on the South catalogue to assign slightly lower confidence values to a source than the GP trained on the North. Since this is true of confidence values assigned to sources in both catalogues, this behaviour is due to underlying differences between the North and South catalogues and is not due to the GP sampling sources that make up its training set (if that were the case, the two plots would be mirrored in the $y = x$ line in Figure 4.4). This trend implies that there exists some inherent difference in the measured properties of the X-ray sources contained in the North and South catalogues, affecting the output of the GP.

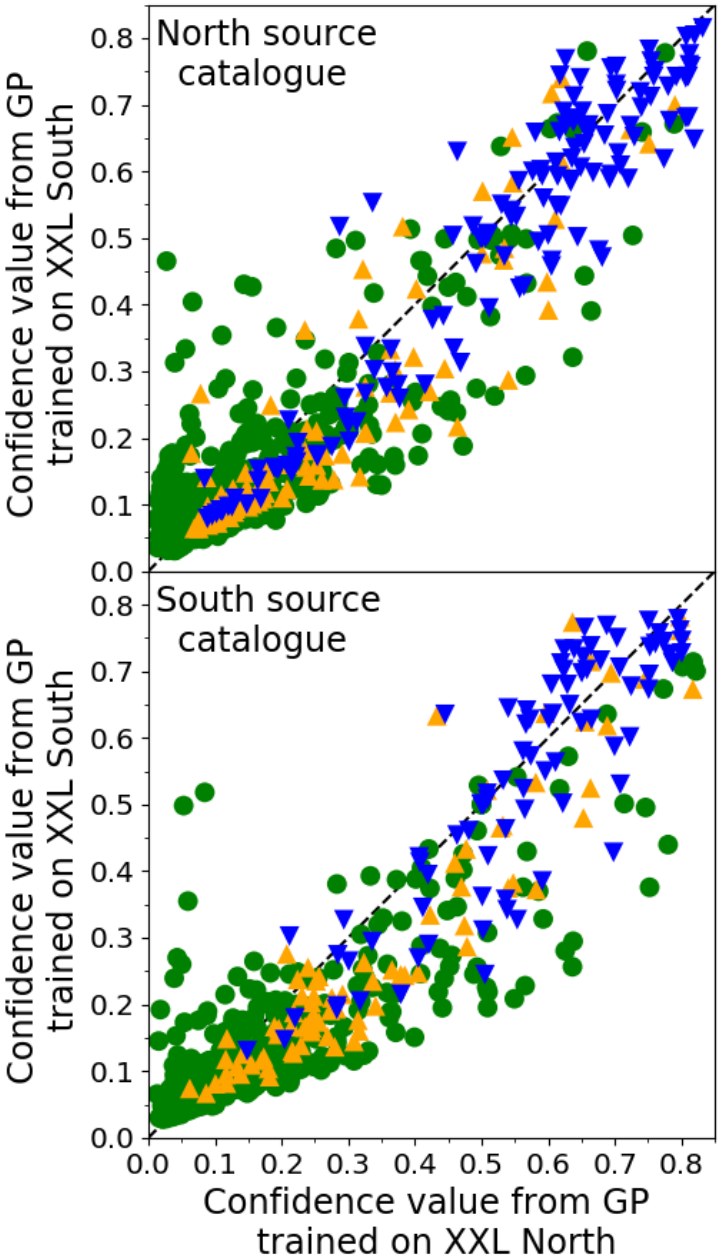


Figure 4.4: Comparison of confidence values assigned by the GP to sources in the North (top) and South (bottom) XXL catalogues. In both plots, the x- and y-axes shows the confidence value assigned when the GP was trained on the North and South catalogues respectively. The C1 and C2 sources are plotted as blue triangles pointing down and orange triangles pointing up respectively, with the non-C1C2 sources plotted as green circles.

Further differences between the output of the GP when trained on the North or South source catalogues can be seen in the standard error on the confidence value (figure 4.3). It is immediately apparent that the standard error on confidence values assigned by the GP when trained on the South XXL catalogue (bottom row of figure 4.3) is generally higher than that for the GP trained on the North catalogue (top row). The difference in error being most prominent around a confidence value of 0.5 due to the sigmoid function suppressing the error for values closer to zero or one. This difference in the error on the confidence value is additional evidence for a difference between the North and South catalogues affecting the output of the GP. As we will see later in section 5.3, this is further supported by differences in the relative importance of the measured source properties for the GP trained on the North or the South catalogue. A discussion as to the potential sources of this difference is left for 5.3, where it can be conducted in the context of the relative importance of each measured source property.

The following section uses a simulated version of the XXL source catalogue to test the output of the GP. The benefit of using a simulated catalogue, being that each detection within the catalogue can be labelled as a genuine detection or not of a galaxy cluster. The labelled catalogue can then be used to asses the confidence value output by the GP and hence any sample selected based on confidence value.

4.3 Classifier results for the simulated XXL X-ray source catalogue

In order to make use of the simulated XXL catalogue described in section 2.1.3, each simulated source was assigned two confidence values, one by the GP trained on the North XXL source catalogue and the other by the GP trained on the south catalogue. Both GP's were trained as in the previous section over 100 Monte Carlo samples separated into ten batches of ten samples. The reliability and accuracy of the GP can then be assessed by considering the difference in confidence values assigned to those simulated sources labelled as either a detection or not a detection of a galaxy cluster.

Figure 4.5 shows the distribution of confidence values assigned by the two GP's as a function of EXT and EXT_LIKE, separated into sources labelled as a detection or not a detection of a galaxy cluster. From figure 4.5 it is immediately apparent that both GP's assign high confidence values to simulated sources with a high EXT and EXT_LIKE value, despite not being provided this information. This matches the behaviour seen in the previous section when applying the GP's to real data (figure 4.1) and is expected given the definitions of the C1 and C2 samples used when training the two GP models.

When considering only those simulated sources labelled as not a detection of a galaxy cluster, there is a clear difference in the distribution of confidence values compared to that for real sources. For the simulated sources there exists a clear anomalous population of non-cluster detections with an EXT close

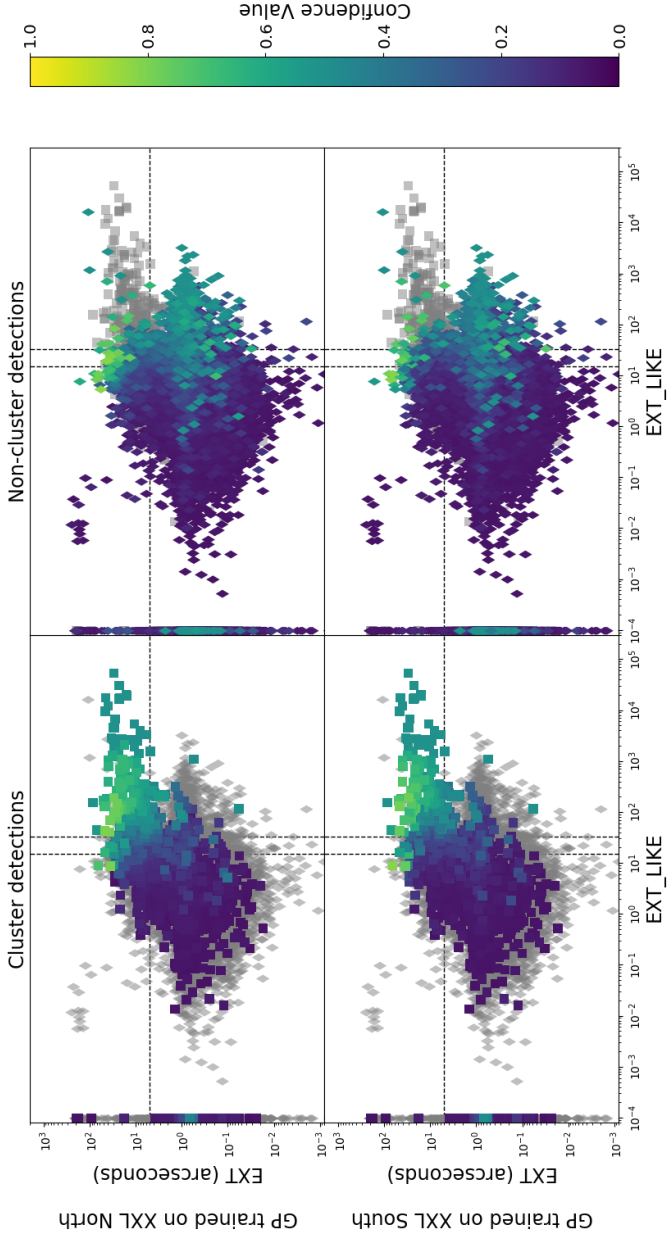


Figure 4.5: The distribution of simulated sources labelled as a detection of a galaxy cluster (square) or not a detection of a galaxy cluster (diamond) as a function of EXT and EXT_LIKE. Those sources labelled as a galaxy cluster detection and not a galaxy cluster detection in the left and right hand columns respectively are colour coded based on the confidence value assigned to them by the GP trained on the North XXL catalogue (top row) and South XXL catalogue (bottom row). The dashed lines indicate the C1 and C2selection criteria as listed in table 4.1. The coloured 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. There exists a clear population of non-cluster simulated sources with an extent of order unity that the GP assigns a high confidence that is not present for real data (figure 4.1) This highlights that the GP (and other ML classifiers) is both sensitive to and can enable identification of discrepancies between real and simulated data.

to one arcsecond and a high EXT_LIKE assigned a high confidence value by both GP's. Such a population is not present in the north or south catalogues (figure 4.1). There exist two possible causes for the existence of this anomalous population, i) the GP's are assigning inaccurate confidence values to sources not present in their training set or ii) within the simulated catalogue there exists a population of non-cluster detections whose measured source properties (specifically those provided to the GP's) don't accurately reflect the distribution of real sources. The first possible cause can be ruled out by considering the confidence values assigned by the GPs to the real XXL catalogue they were not trained on. Since we do not see a population of high confidence sources with an EXT of around one and a high EXT_LIKE in the top right and bottom left panels of figure 4.1 the presence of this anomalous population in figure 4.5 is not due to GPs. Having ruled out the GP as the potential source, the only remaining possibility is that the anomalous population represents a set of simulated non-cluster detections that are not present in the north and south XXL catalogues. This highlights that the GP (and other ML classifiers) is both sensitive to and can enable identification of discrepancies between real and simulated data.

Due to the presence of the anomalous population of non-cluster detections assigned a high confidence by the adapted GP binary classifiers, the simulated catalogue does not sufficiently replicate a real XXL catalogue for the purposes of assessing the GP models. To asses the GP models, the next section instead uses a sample of sources from XXL north that have X-ray independent evidence of being a galaxy cluster. Specifically those sources in the north XXL catalogue that are matched to an optically-detected CAMIRA cluster.

4.4 Classifier results for XXL sources matched to CAMIRA detections

As detailed in section 2.2.1, a set of sources were selected from the north XXL source catalogue if they were within 15 arcseconds of a CAMIRAselected cluster candidate (the specific CAMIRA catalogue used in this work being a more recent version of that detailed in Oguri et al., 2018). The matching process was performed independently of the GP such that it does not influence the output of the model. While we know that these X-ray sources are aligned with optical cluster candidates, the GP was blind to this, attributing confidence values purely on the basis of the parameters determined from the X-ray observations. This produces a set of XXL sources in the Northern catalogue that have X-ray independent evidence as to being a galaxy cluster that the GP is not made privy to.

The resulting CAMIRA/XXL sample is a combination of (i) chance superpositions of optically selected cluster candidates and non-associated X-ray sources (whether the optical cluster candidates are real or

not); and (ii) true clusters with detected optical and X-ray emission (whether from the ICM or an embedded AGN). Given the number of CAMIRA and XXL sources in the North field there are expected to be ~ 31 chance superpositions in the CAMIRA/XXL sample. Since the CAMIRA/XXL sample contains a total of 162 XXL sources chance superpositions are expected to make up $\sim 20\%$ of the sample. The CAMIRA/XXL sample is used here to assess how well the GP identifies true clusters, as we expect a higher fraction of the sample 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 CAMIRA/XXL sample, if it is able to correctly identify sources with an increased probability of being a galaxy cluster.

As seen in Figure 4.1, the GP when trained on the North or the South XXL catalogue tends to assign a high confidence to C1 and C2 sources. Highlighting the CAMIRA/XXL sample in figure 4.6 shows that the tendency for the GP's to assign higher confidence values to C1 and C2 is consistent within the CAMIRA/XXL source sample. The comparison with the CAMIRA sample enables us to test how well the GP is able to recover clusters from the large non-C1C2 population. For this reason we exclude C1 and C2 sources and focus on those non-C1C2 sources in the CAMIRA/XXL sample. Furthermore, the difference in the fraction of genuine cluster detections between the CAMIRA/XXL sample and full XXL North sample should be greatest for non-C1C2 sources leading to a clearer difference between the distribution of confidence values of each.

Figure 4.7 shows the distribution of non-C1C2 sources within CAMIRA/XXL sample as a function of the confidence value assigned by the two GP's, one trained on the north and the other the south XXL catalogue. This is directly compared to the expected distribution of confidence values for a purely random subsample of non-C1C2 sources of the same size as the non-C1C2 CAMIRA/XXL subsample. It is apparent that the distribution of confidence values assigned by both GP's to non-C1C2 sources within the CAMIRA sample favours higher values than that for the non-C1C2 sources within the full XXL North sample.

For a random subsample of non-C1C2 sources drawn from the north XXL catalogue and equal in size to that of the non-C1C2 CAMIRA/XXL sample, the expected number of sources with a confidence value greater than 0.1 is ~ 2 for both the GP trained on the North catalogue and trained on the South catalogue. In fact we observe 14 and 11 sources with a confidence value over 0.1 (for the GP trained on the north and south catalogues respectively). The probability of either of these occurring for a random sample of 110 non-C1C2 sources is effectively zero. This implies two things that must both be true. Firstly (and as expected), the subset of non-C1C2 XXL sources matched to a CAMIRA cluster candidate contains a higher fraction of genuine X-ray cluster detections compared to all non-C1C2 sources. Secondly, for those XXL sources not within the C1 and C2 samples, a higher confidence value assigned by the GP is correlated with a greater likelihood of the source being a genuine X-ray cluster detection. If one or both of

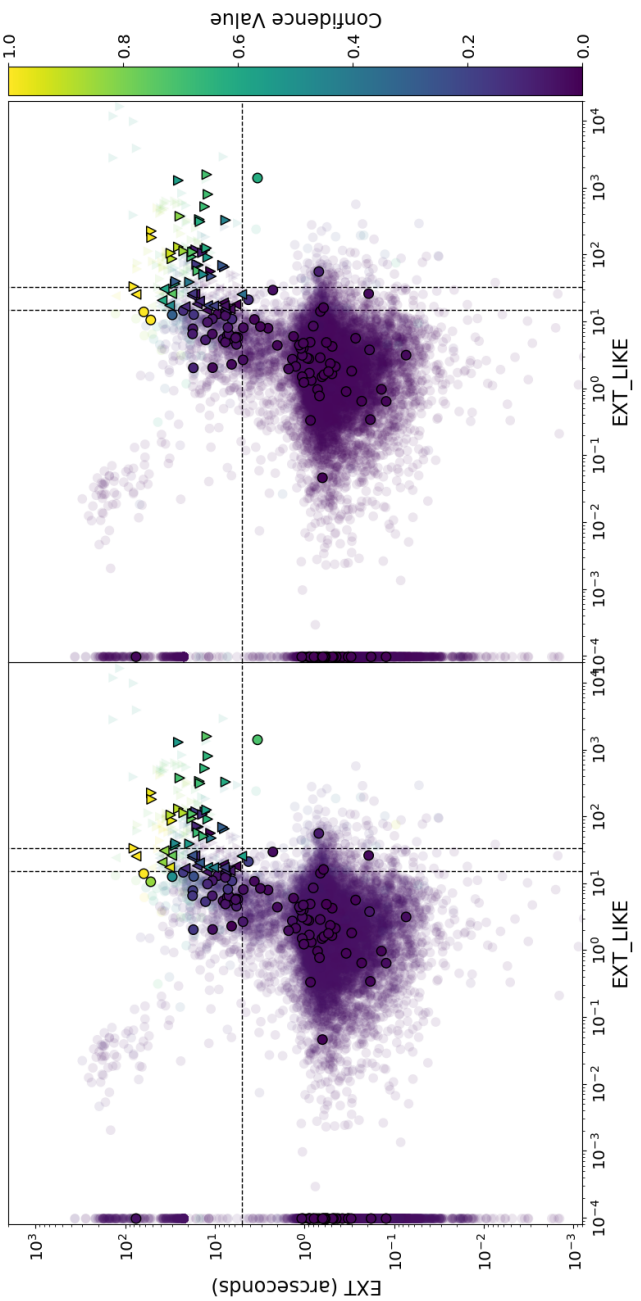


Figure 4.6: Confidence values assigned to sources in the North XXL catalogue by the GP trained on the North (left) and South (right) XXL fields. Confidence values are plotted as a function of EXT and EXT_LIKE as in Figure 4.1. Sources matched to a CAMIRA optical detection are plotted on top and highlighted by a black outline.

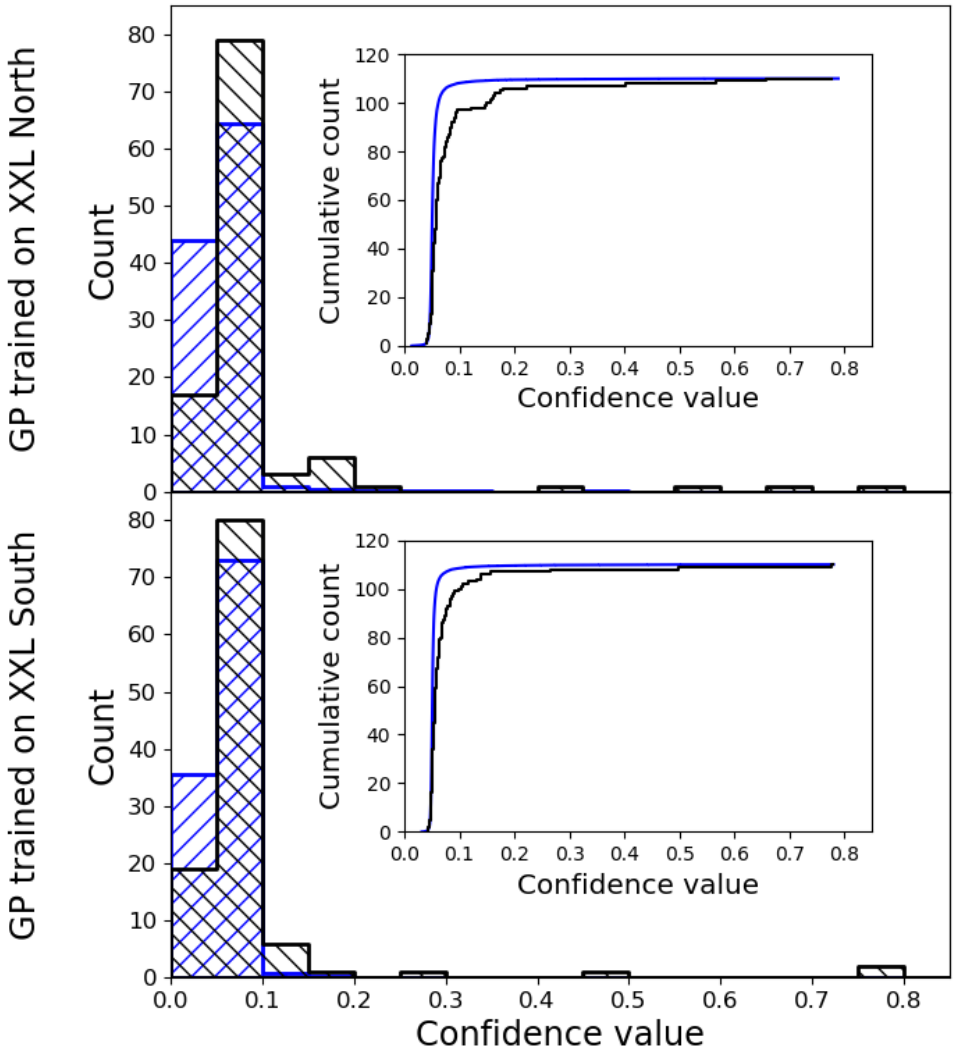


Figure 4.7: Distribution of non-C1C2sources within 15 arcseconds of a CAMIRA/ cluster candidate (black) compared to the expected distribution of a random sample of non-C1C2XXL sources (blue) as a function of confidence value. The confidence value for each source being assigned by the GP when trained on the north (top) or south (bottom) XXL catalogues. Inlaid is the cumulative distribution of sources as a function of confidence value

these statements were not true, the subset of non-C1C2 sources matched to a CAMIRA cluster candidate would have a confidence distribution that matches that for the full non-C1C2 sample. 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 tendency for CAMIRA-selected clusters to be assigned higher confidence values implies that the high-confidence sources that are non-C1C2s and lack CAMIRA counterparts are also potential galaxy cluster detections. This could arise for example, because CAMIRA may be sensitive to a more restricted redshift and richness range than the XXL X-ray data. Non-C1C2 sources that lack a CAMIRA counterpart and have a comparatively high confidence value are unlikely to be X-ray point sources (as these sources should retain their initial confidence value of 0.05). Their higher confidence value must reflect the X-ray emission having some spatially-extended component even if this is not strong enough to classify the source as a C1 or C2.

The CAMIRA/XXL sample has shown that the GP is able to identify and assign an increased confidence to potential galaxy cluster detections not previously identified by XXL. The sample does not however, provide sufficient information for us to accurately evaluate source samples produced by selecting on the assigned confidence value. The next section aims to provide the information necessary to evaluate source samples selected on confidence value by conducting a visual inspection and classification of sources from the XXL north catalogue. Source samples selected on confidence value are then evaluated using this information in section 4.6.

4.5 Visual Inspection of Sources

The visual inspection of XXL sources was limited to the XXL north catalogue due to its overlap with optical observations conducted by the Hyper Suprime-Cam Subaru Strategic Program (Aihara et al., 2022). Despite limiting the visual inspection to only the north catalogue, it is not practical to inspect the entire catalogue ($\sim 24,000$ sources), instead a subset of sources were selected for visual inspection. The north XXL catalogue is dominated by non-C1C2 sources assigned a low confidence value by the GPs, hence a randomly selected subset of sources would not contain a sufficient number of C1, C2 or high confidence non-C1C2 sources for meaningful analysis. Instead of randomly selecting sources from the whole catalogue, the catalogue is split into a series of subsets designed to characterise the catalogues specific components. The north catalogue is hence separated into the following subsets: C1 sources, C2 sources and the non-C1C2 sources separated into bins on confidence value of size 0.05. Sources were subsequently randomly selected from each subset and combined together in a random order to avoid bias when labelling. The number of sources selected from each subset was chosen to accurately characterise

the subsets contents, whilst keeping the number of sources to be inspected of a practical size. The number of sources visually inspected from the C1 and C2 source samples are listed in table 4.5, with the same information for the non-C1C2 samples listed in table 4.6.

Also included as part of the visual inspection are those non-C1C2 sources that are within 15 arcseconds of a CAMIRA cluster candidate. The purpose of including the non-C1C2 CAMIRA/XXL sources is as an additional test of the assumption that said source sample contains an increased fraction of galaxy cluster detections. The number of sources visually inspected from the CAMIRA/XXL sample is listed in table 4.5.

Having selected the sources for visual inspection, it is necessary to create a set of cutouts of each source. Each set of cutouts must convey sufficient information for each source to be reliably labelled as ether a genuine X-ray detection of a galaxy cluster, or not. Figures 4.8, 4.9 and 4.10 show an example cutout created for a C1 source, AGN and spurious X-ray detection respectively.

Two pseudo true-colour optical images of the immediate field (five by five arcminuets) of each source were constructed from Hyper Supreme-Cam imaging data (Aihara et al., 2018a) following the method described in Lupton et al. (2004). The first image being a combination of the G, R and I optical bands and the second the R, I and Z bands. X-ray images of the same fields were extracted from the XXL North X-ray mosaic in the $0.5 - 2.0$ keV band (Adami et al., 2018). X-ray contours were generated from the X-ray image smoothed using a Gaussian kernel with standard deviation of five arcseconds. Due to variation in background and source flux the value of the contours were independently calculated for each source, being given by the median photon count of all pixels in the smoothed image plus $(2^{-5}, 2^{-4}, ..., 2^{4}, 2^{5})$. The distribution of X-ray counts in the smoothed image was plotted for the full five by five arcminuet cutouts along with the two by two and one by one arcminuet central regions. Plotting this distribution separated into different regions of the image helps those conducting the visual inspection in identifying how strongly the sources X-ray emission differs from the background. The code used to generate these cutouts is available from the following Github repository; `https://github.com/CaleBaguley/XXL-galaxy-cluster-lotto-code`.

Following the creation of the cutouts, each source was visually inspected by J.C. Baguley, M.N. Bremer and B.J. Maughan and labelled as ether a galaxy cluster candidate or not. The inspection order for the sources was randomised and at no point were those conducting the inspection privy to the confidence value or the classification of the source by XAMIN. This information was withheld in order to avoid biasing the results.

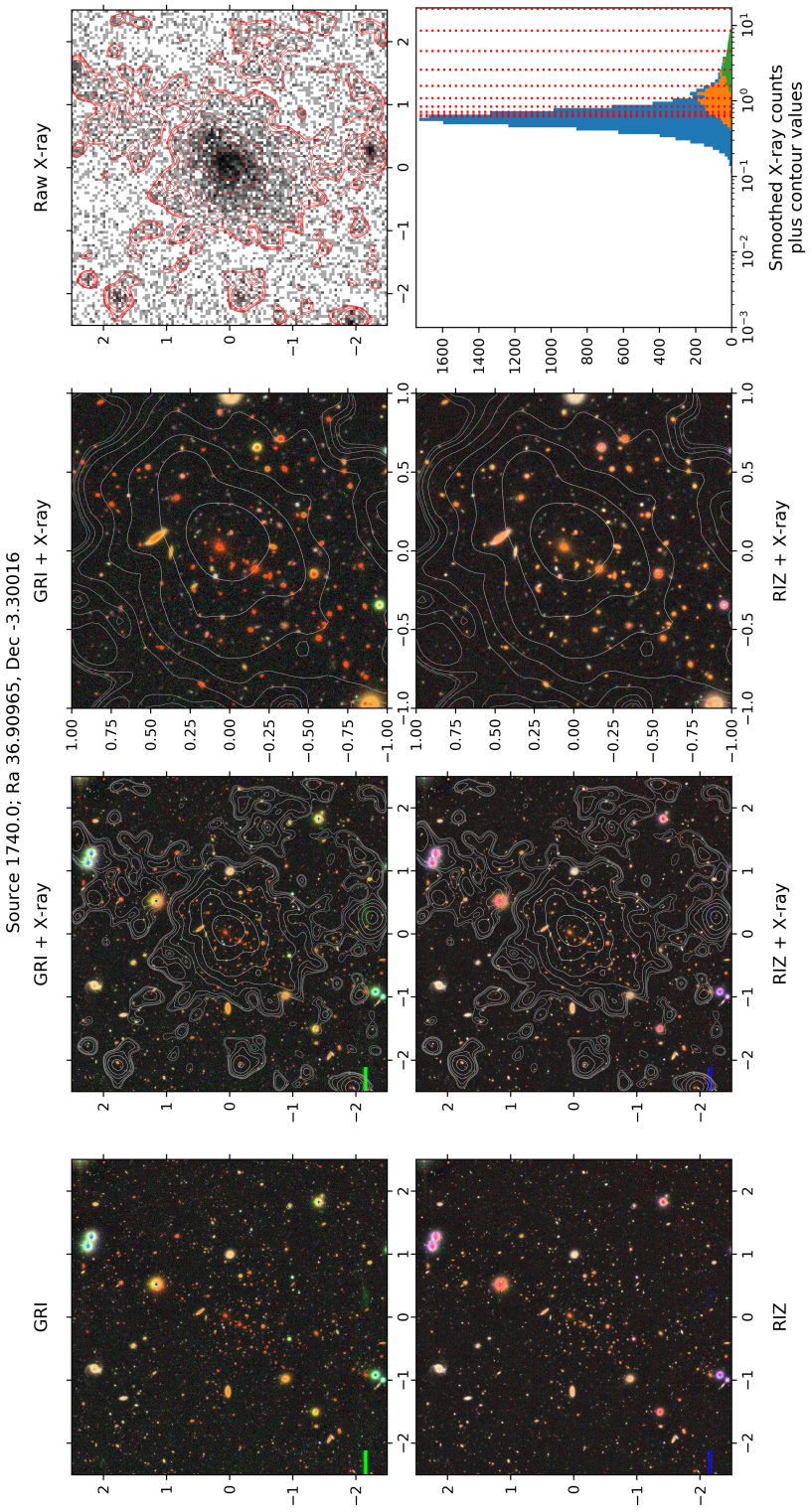


Figure 4.8: Example of the cutouts created for visual inspection for a C1 source.

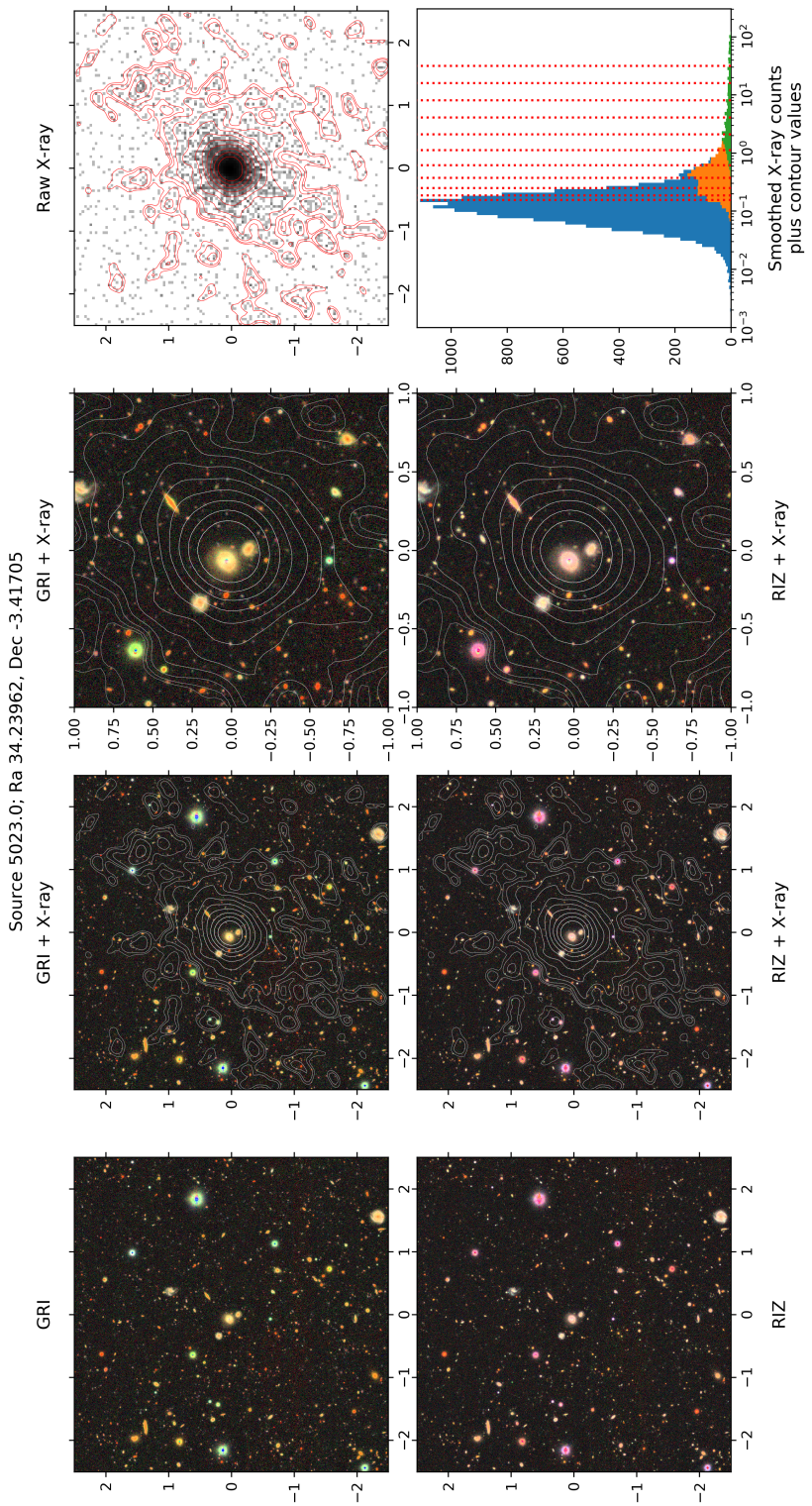


Figure 4.9: Example of the cutouts created for visual inspection for an AGN.

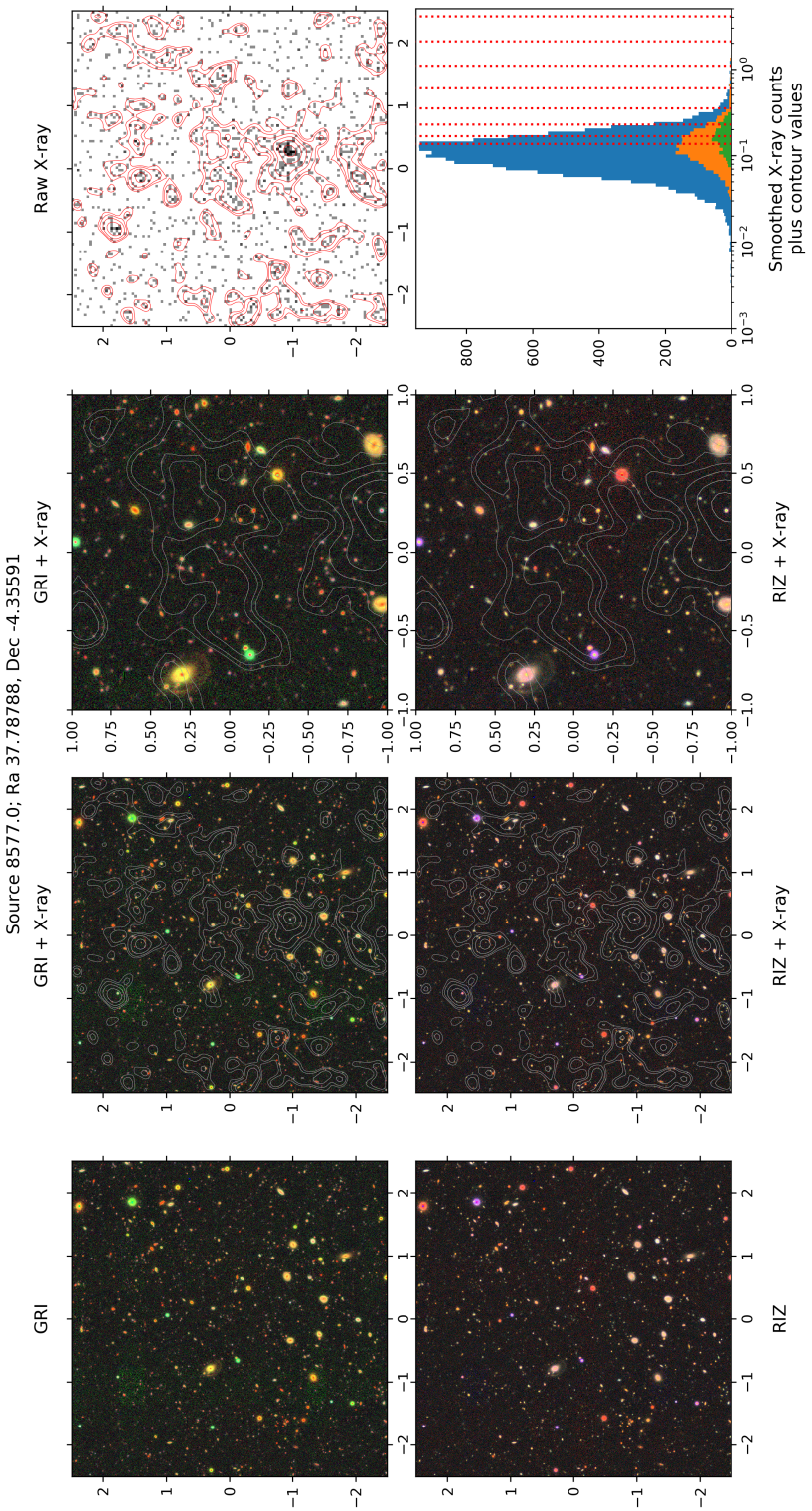


Figure 4.10: Example of the cutouts created for visual inspection for a spurious detection.

Source subset	Total sources in sample	Sources inspected	Cluster candidates	Bayesian estimated purity
C1 sources	139	121	109	$0.90^{+0.02}_{-0.03}$
C2 sources	109	20	13	$0.64^{+0.10}_{-0.11}$
non-C1C2 CAMIRA sources	110	38	17	$0.45^{+0.08}_{-0.08}$

Table 4.5: 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 4.1. This is done to account for uncertainties due to the small number of sources inspected.

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 is of sufficient depth that overdensities of galaxies associated with clusters out to this redshift are identifiable within 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 intentionally broader than the standard XXL selection function which is calibrated to identify clusters that resemble β-profiles, as part of the aim of applying ML to identify clusters from the XXL catalogue is to find novel galaxy clusters with a larger range of morphologies than those selected by XXL. The number of sources that met this criteria and hence were labelled as galaxy cluster candidates are listed for each subset of the XXL catalogue in tables 4.5 and 4.6.

The next section takes the information collected through the visual inspection process and uses it to asses the efficiency of different selection criteria with respect to identifying galaxy cluster candidates from the XXL catalogue.

4.6 Cluster candidate sample results

In order to determine the best selection criteria for identifying galaxy cluster candidates it is necessary to define two metrics with which to asses the performance of each possible selection criteria. The first

Confidence range	Total sources in sample	Sources inspected	Cluster candidates	Bayesian estimated purity
0.00, 0.05	9319	96	4	$0.05^{+0.02}_{-0.03}$
0.05, 0.10	13661	115	17	$0.15^{+0.03}_{-0.03}$
0.10, 0.15	213	32	8	$0.26^{+0.08}_{-0.08}$
0.15, 0.20	74	26	4	$0.18^{+0.07}_{-0.07}$
0.20, 0.25	28	24	4	$0.19^{+0.08}_{-0.08}$
0.25, 0.30	20	10	3	$0.33^{+0.14}_{-0.13}$
0.30, 0.35	14	11	2	$0.23^{+0.11}_{-0.11}$
0.35, 0.40	6	5	1	$0.29^{+0.17}_{-0.16}$
0.40, 0.45	10	9	3	$0.36^{+0.14}_{-0.14}$
0.45, 0.50	11	5	1	$0.29^{+0.17}_{-0.16}$
0.50, 0.55	8	5	0	$0.14^{+0.12}_{-0.11}$
0.55, 0.60	3	2	1	$0.50^{+0.25}_{-0.25}$
0.60, 0.65	4	2	0	$0.25^{+0.21}_{-0.19}$
0.65, 0.60	3	3	1	$0.40^{+0.22}_{-0.21}$
0.70, 0.75	2	2	0	$0.25^{+0.21}_{-0.19}$
0.75, 0.80	2	1	0	$0.33^{+0.27}_{-0.25}$

Table 4.6: Results of visual inspection of sources not previously selected by XXL (non-C1C2) binned on source confidence value. As in table 4.5 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 0.20 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.

metric we use to assess the contents of a source sample is the purity:

- **Purity**. This is the fraction of objects that meet the selection criteria that have the characteristic we are looking for (i.e. meeting the visual inspection criteria needed to be labelled a cluster candidate). This directly relates to the probability of a source having this characteristic given that it meets the selection criteria. 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.

For those selection criteria that correspond directly to visually inspected samples, as described in the previous section, the probability density function for the purity, p, is modelled using the following Bayesian 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. For the purpose of simplicity we select a beta distribution as the prior over p as it is the conjugate prior of a binomial distribution. The shape parameters of the beta distribution $\alpha = \beta = 1$ are chosen such that the prior is uniform over the range zero to one. The resulting 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

$$\text{Purity} = \frac{n_c + 1}{N_s + 2}. \tag{4.1}$$

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 along with the the 16^{th} and 84^{th} percentiles for the visually inspected samples are listed in tables 4.5 and 4.6.

We do not wish to simply consider those samples that were visually inspected, hence we need a way to model the purity for new samples. For any new sample, constructed by combining visually inspected samples, the purity can be estimated by taking the average purity of the visually inspected samples weighted by the number of sources from each inspected sample in the new sample. This does not however account for the uncertainty on the purity of the visually inspected samples. To do so we use a Monte Carlo approximation, randomly sampling the purity for each of the inspected samples from the beta distribution described above. The resulting probability density distribution over the purity of the new sample is used to calculate the mean along with the 16^{th} and 84^{th} percentiles.

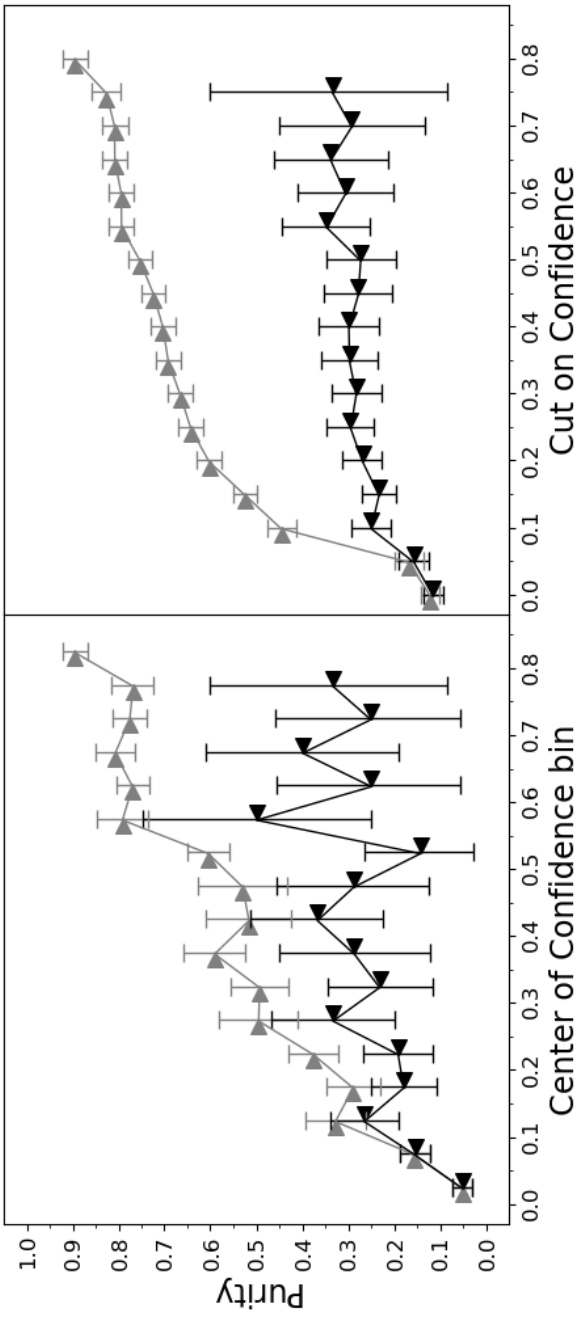


Figure 4.11: The Figure shows the purity of different source samples created from the XXL North source catalogue. See subsection 4.6 for the definition of purity used in this work. Source samples are produced from the entire XXL North catalogue (grey triangle pointing to the right) and by excluding the 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).

The mean purity is calculated as above for samples produced from the full north XXL catalogue (including C1 and C2 sources) and only the non-C1C2 sources. The first set of samples were constructed by binning sources based on the confidence value assigned to them by the GP trained on the north catalogue. The second set of samples being created by selecting sources with a confidence value (again assigned by the GP trained on the north catalogue) above some cut. The purities calculated for the different samples being reported in figure 4.11. The results show a clear increase in purity for samples containing sources with higher 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 samples of non-C1C2 sources produced by binning on the confidence value, there is an initial trend to higher purity with increasing confidence value. The low number count of sources in each bin at mid to high confidence values however results in a significant increase in uncertainty for the reported sample purity. Similarly, the samples of non-C1C2 sources selected based on a cut on confidence value show an initial increase in purity with cut on confidence value before levelling off. The samples produced with the highest confidence cuts contain few sources and so again 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 most of the high-confidence sources are C1s or C2s.

Having described the results of the first performance metric, purity, let us now define the second metric, completeness.

- **Completeness**. This is the fraction of those sources with the characteristic we are looking for that would be detected by XXL that meet the selection criteria (i.e. the fraction of XXL sources that would be labelled as a galaxy cluster candidate following visual inspection of the entire north catalogue that meet the selection criteria). This directly relates to the probability that a source with this characteristic that would be detected by XXL is included in the sample. This deviates from the standard definition of completeness used in source detection in that, it is not with respect to the whole population of galaxy clusters, just those detected by XXL. This does not impact the analysis of the GP as it can only ever select sources detected by XXL.

The completeness for any source sample is calculated using a Monte Carlo approximation. For each iteration, the purity of each of the visually inspected source samples is randomly drawn from the beta distributions calculated previously. The completeness is then calculated for the given set of purities by taking the ratio of the expected number of cluster candidates within the sample and the expected number for the whole catalogue,

$$\text{Completeness} = \frac{\sum_i^{\text{samples}} \text{Purity}_i n_i}{\sum_i^{\text{samples}} \text{Purity}_i N_i} \tag{4.2}$$

Here both sums are over all visually inspected samples, purity$_i$ is the purity of the i^{th} inspected sample, n_i is the number of sources from the i^{th} inspected sample in the new source sample and N_i the total number of sources in the i^{th} inspected sample. The expected completeness calculated in this way along with the 16^{th} and 84^{th} percentiles are shown in figure 4.12.

For source samples produced by binning sources on confidence values it is clear that the completeness is maximum for the 0.05 to 0.1 bin on confidence value. This is expected, as despite the purity of the sample being relatively low it contains the vast majority of sources (figure 4.2), hence by weight of numbers it would contain the vast majority of sources labelled as a galaxy cluster candidate if we were to visually inspect all source within the north catalogue. The same reasoning explains why, for those samples selected by a cut on confidence value, only those with a low cut on confidence value have a high completeness. This also explains why there is only a small reduction in completeness when the C1 and C2 sources are removed, as they contain a small fraction of all galaxy cluster candidates despite their higher purity.

Having visually inspected 79 non-C1C2 sources with confidence values above 0.2 and available optical HSC data, we can infer the types of sources that a GP-selected sample with a confidence value cut of 0.2 would find. The non-C1C2 sources that were inspected can be broadly sorted into five categories: (i) bright point sources with a clear optical counterpart ($\sim 42\%$); (ii) background fluctuations in the X-ray image that were combined into a broad, flat flux distribution by Xamin's wavelet filtering ($\sim 32\%$); (iii) extended but irregular sources associated with galaxy overdensities ($\sim 19\%$); (iv) nearby groups where the X-ray emission appears to be dominated by the halo of the brightest galaxy ($\sim 6\%$); and (v) extended sources with a dominant central AGN ($\sim 1\%$). See Figure 4.13 for examples of sources belonging to each of these categories. A detailed discussion of why the GP identifies such sources as extended is left for chapter 5, where it is conducted in the context of the relative importance of the measured source properties in determaning the confidence value output by the GP.

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 virialsed halo. These sources differ from those contained by the C1 and to a lesser extent the C2 samples. The majority of those galaxy cluster candidates within the C1 sample haveing a smoother β-model surface brightness distribution, indicating dynamically older systems.

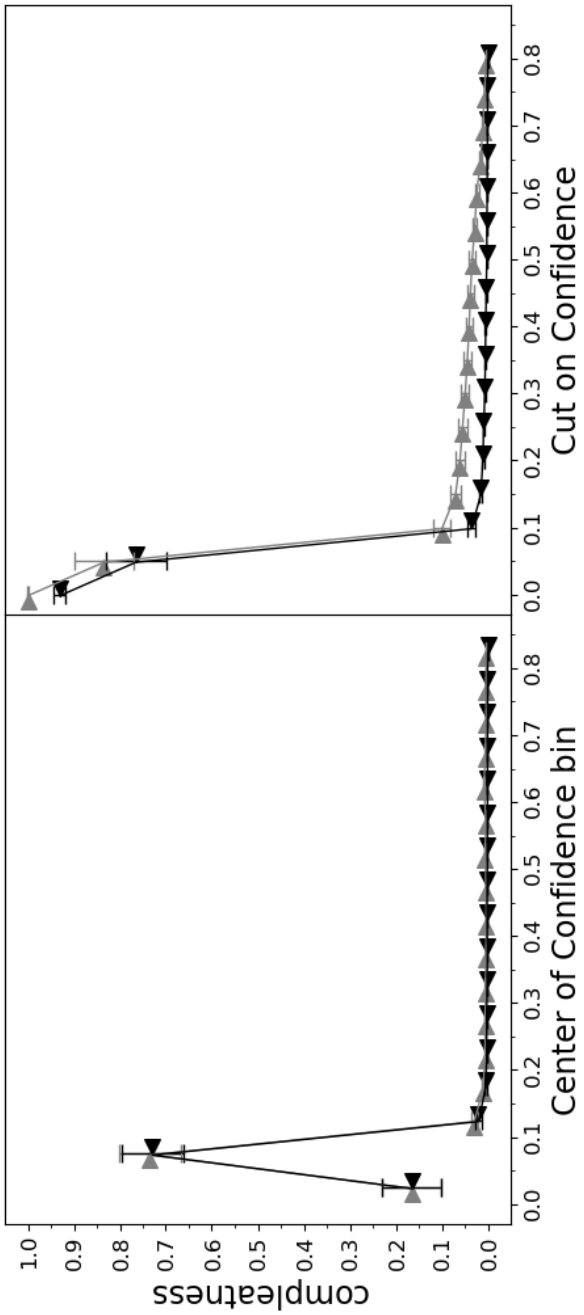
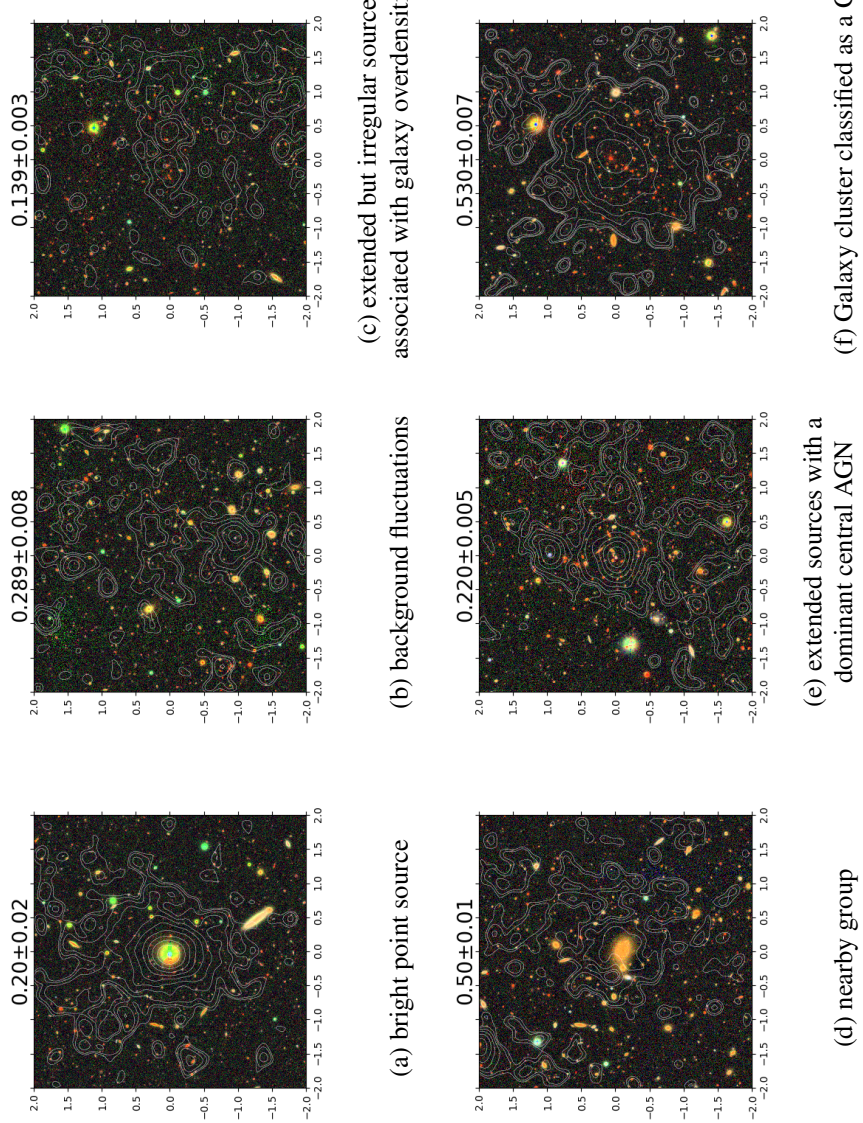


Figure 4.12: The Figure shows the completeness of different source samples created from the XXL North source catalogue. See subsection 4.6 for the definition of completeness used in this work. Source samples are produced from the entire XXL North catalogue (grey triangle pointing to the right) and by 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).



(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 4.13: Four by four arcminute cutouts of XXL sources indicative of the types of sources for which the north trained GP assigned confidence values over 0.2. 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 (Aihara et al., 2018a) following the method described in Lupton et al. (2004). X-ray contours are produced from the XXL North X-ray mosaic in the $0.5 - 2.0$ keV band (Adami et al., 2018) and smoothed by a Gaussian kernel with a standard deviation of five arcseconds.

Given the results outlined above we define the GP-selected source sample. The GP-selected sample is 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 are labelled as a C1 or C2 source respectively (the C1 and C2 samples containing 139 and 109 sources respectively). The sample has an estimated mean purity of $0.45^{+0.03}_{-0.03}$ i.e. we estimate that 280 of the 623 sources would be classified as a galaxy cluster candidate by our visual inspection if we were to visually inspect them all. Of the estimated 280 galaxy cluster candidates within the GP-selected sample 122 and 57 are expected to also be labelled as a C1 or C2 source respectively. This leaves 101 expected cluster candidates not found in the C1 or C2 samples. The sample selected by this cut on confidence omits 3 and 20 sources labelled as a C1 or C2 respectively. Given the fractions of C1s and C2s that we visually classified as cluster candidates, we estimate the loss of 3 and 13 visually-classified cluster candidates from the C1 and C2 source samples respectively. This assumes a constant purity for the C1 and C2 samples as a function of confidence. In fact it is reasonable to expect those C1 and C2 sources with a lower confidence value to be less likely to be visually classified as a cluster candidate than those with a higher confidence value, so the actual number of galaxy cluster candidates labelled as a C1 or C2 missed by the GP-selected sample could be lower.

4.7 Discussion

Given the results above, the key question to ask is how effective is the GP at identifying cluster candidates. In particular, how well does the content of the GP-selected source sample compare to that of the existing C1 and C2 samples. We found that a sample selected on the basis of sources having a confidence value assigned by the GP trained on the north XXL catalogue above 0.1, includes the vast majority of the original C1 and C2 sources. For example, making such a selection for sources in the South field when the GP was trained on the North, selects all C1 sources and all but one C2. Figure 4.4 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). However, the purity of a GP-selected sample is lower than that of the conventional C1 and C2 samples. This implies that a GP-selected sample is less suitable for applications such as cosmological analysis, 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 followup investigations, particularly if these can leverage existing (survey) data in other wavebands rather than requiring new observations.

The existing C2 sample can, in principle, be extended by reducing the EXT and EXT_LIKE selection criteria to include a higher number of sources, in doing so the purity of the C2 sample would be reduced. Based on the distribution of confidence values in the EXT and EXT_LIKE plane (Figure 4.1), along with our visual inspection results (Table 4.6) we can estimate the purity of an extended C2 sample. The

EXT and EXT_LIKE criteria used to select the combined C1C2 sample (Table 4.1) can be lowered such that the new sample includes roughly the same number of sources as the GP selected sample. Selecting those sources in the north catalogue with an EXT greater than 2.37 arcseconds and an EXT_LIKE greater than 4.10 gives a sample of 614 sources with an estimated purity of $0.37^{+0.02}_{-0.02}$. The values for the EXT and EXT_LIKE cut used here were identified by stepping through the values for different sources and selecting the highest purity sample that contained a number of sources within the range $[608, 683]$. This in comparison to the GP-selected sample of XXL north which contains 623 sources with an estimated purity of $0.45^{+0.03}_{-0.04}$. It is simply not practical to identify clusters assigned a high confidence value (figure 4.1) or a CAMIRA detection (figure 4.6) simply by selecting sources on the basis of EXT and EXT_LIKE alone. Doing so would require the selection of the majority of the XXL catalogue, making visual inspection and observational followup impractical.

As noted previously the X-ray characteristics of those non-C1C2 XXL sources assigned a higher confidence value by the GP, differ from those sources selected by the C1 and C2 criteria. In general the X-ray emission appears different to that of C1 clusters, which are identified by a classification scheme that was optimised to select regular β-model clusters. The implication is that C1s are dynamically mature clusters with the X-ray emission dominated by that from a virialised ICM. It appears that the C1 (and to a large extent the C2) sample are highly complete with respect to these bright virialised systems, as we do not find them outside the C1 and C2 samples when visually inspecting sources selected by the GP (see subsection 4.5). If there existed a set of virialised systems outside the C1 and C2 samples then we would reasonably expect at least some of them to have similar measured properties to the C1 and C2 sample, resulting in them being assigned a high confidence by the GP and hence identified during the visual inspection. 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).

The critical advantage of the GP 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 about the sources. Incorporating additional information into the labelling of a traditional GP or other binary classifiers is not possible due to the need for binary training labels. 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. Adapting a GP binary classifier to include uncertainty on the labels enables us to utilise more complex training sets where there are varying amounts of information about different sources. The more information as to the nature of a source the greater the certainty with which it can be labelled. For instance, our approach would enable us to assign a higher initial probability of

being labelled a galaxy cluster during training to XXL sources that were associated with a CAMIRA cluster. A C1 or C2 source could be assigned a higher initial probability than the default value if they were associated with a CAMIRA cluster, while a non-C1C2 source that was associated with a CAMIRA cluster could be assigned an initial probability based on the estimated purity of the CAMIRA sample. The assigned probabilities could also be changed to reflect the results of visual inspection by experts or spectroscopic follow-up. The resulting training set should then enable the GP to be more effective at defining high-purity samples with less dependence on the similarity of sources to the purely X-ray selected C1 and C2 subsets used for labelling in our main analysis.

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 4.6) 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 given their C1 classification. The GP (or any other machine learning binary classifier) would then be unable to separate clusters from non-clusters because it would be explicitly trained to label C1 clusters without a matched CAMIRA source 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 measured properties 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 its X-ray properties alone.

While it is a strength of the GP (and other ML classifiers) that it can use a large number of measured source properties from different wavebands to classify objects, this brings with it additional complexity in both training the model and characterising the selection function. Increasing the number of measured source properties increases the number of calculations conducted to determine the value of the kernel function for two sources as well as the time taken to optimise hyper parameters that define the kernel functions response to inputs. Further accurately modelling the selection function of the GP requires simulations that are sufficiently realistic so as to correctly describe the multi-faceted 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 in studies where the selection function needs to be precisely known, such as using well-defined samples of clusters for cosmological studies.

A simulated XXL source catalogue was used to asses the confidence values output by the GP in section 4.3. The simulation used is limited, both in terms of the range of source parameters it covers and the relative numbers of AGN and X-ray emitting massive halos included in the catalogue. Consequently, it did not sufficiently recreate the distribution of measured X-ray source properties when compared to the North and South catalogues to produce reliable results for this study. Particularly, the simulation did not sufficiently recreate the distribution of properties of real sources needed for accurate easement and characterisation of the galaxy cluster selection function.

The content of this chapter has focused on the application of the adapted GP binary classifier to XXL. Due to the adaptable nature of the GP classifier (and other ML techniques) the technique used here is not limited to XXL. The technique is applicable to a range of not only different galaxy cluster surveys but astrophysical surveys designed to identify other types of sources. A discussion of potential applications of the GP in the context of other galaxy cluster surveys is presented in section 6.2.

4.8 Summary

This chapter presented the results of applying the adapted GP binary classifier described in section 3.4 to select galaxy cluster candidates from the north and south XXL source catalogues. The GP was trained using a sample of clusters and cluster candidates derived from the XXL data using the standard XAMIN pipeline (Pierre et al., 2016). Applying initial probabilities of being a galaxy cluster to every X-ray source based on its classification by the XAMIN pipeline, the GP successfully recovered the vast majority of the sources previously identified as clusters or cluster candidates (those sources in the C1 and C2 samples). We emphasise here, this is without the GP having access to those parameters that the pipeline itself used for that classification (EXT and EXT_LIKE). In other words, the GP can identify clusters or cluster candidates from more subtle signatures in the original catalogue than used by a standard photometry-based pipeline.

In addition those systems already identified by the standard pipeline, the GP identified further systems that are plausible cluster candidates, resulting in a sample with a reasonable degree of purity. These extra cluster candidates often appear to have different X-ray morphologies to those identified as the most secure cluster detections by the standard pipeline. This is unsurprising as the standard XXL selection criteria were optimised to identify the most dynamically relaxed and evolved clusters. The additional candidates identified in this work tend to have multi-modal 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 over-fitting to the data, but it did reveal subtle differences in results when different fields were used for training. In order to best identify potential sources of the subtle differences between the GP when trained on the north catalogue or the south catalogue a more detailed understanding of how the GP selects sources is needed. The following chapter uses the results of automatic relevance determination (described in section 3.5.2) to identify and investigate those measured source properties considered most informative as to a source being a galaxy cluster by the GP.

5

Interpreting the Binary Classifiers Selection Criteria

A key component of applying a machine learning (ML) model to solve complex problems, such as object classification, is to develop an understanding as to how the model is solving the problem. By developing an understanding of the solution identified by the model, the scientist is able to not only determine the limitations of the model's solution, but gain an insight into the underlying physical mechanisms and properties used to solve the problem. Due to the often complex and differing nature of ML models, the best approach to interpreting a models output is not only heavily dependent on the model itself, but is an area of active research.

For a Gaussian Process (GP), interpreting the trained model is made significantly easier compared to other models (such as a neural network) by the use of automatic relevance determination (ARD). ARD (as described in section 3.5.2) identifies those variables input to the GP model that are most informative when determining the output of the model. With respect to this work specifically, ARD is used to identify those measured source properties most informative as to a source being a galaxy cluster. Given this information, it significantly reduces the number of measured source properties that need to be inspected to understand the criteria by which the model is identifying sources as potential galaxy clusters.

This chapter follows this method of interpreting a trained GP model. It first reports the results of ARD for the model when trained separately on the North and South XXL source catalogues. Having identified

those measured source properties relevant when determining the output of the GP, the distribution of confidence values over those properties is investigated and compared to the distributions over EXT and EXT_LIKE. In addition this chapter looks at differences in the relevance of different source properties when the GP model is trained on the North or South XXL catalogues.

5.1 Automatic relevance determination results

The (normalised) length scale of the Gaussian kernel for each measured source property provided to the GP was measured as described in subsection 3.5.2, with the resulting values plotted in Figure 5.1 ranked by shortest to longest length scale. Recalling that, as described in section 3.5.2, the smaller the length scale for a parameter, the more relevant that parameter is in determining the confidence value assigned to a source. A shorter length scale allows the assigned confidence value to change by a larger amount for a given change in the corresponding measured source property. Figure 5.1 clearly shows that the two dominant measured properties 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 source properties, EXT_RATE_MOS and PNT_RATE_MOS. The most relevant source properties in both cases are measurements of the photon count rate (RATE) from the extended (EXT) and point (PNT) source models, when the models having been 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. This is comparable to the $C1C2$ selection criteria which makes use of values derived from fitting both the extended and point models. Specifically EXT_LIKE is the ratio of the detection likelihood of the extended fit over that of the point fit.

Having used ARD to identify those measured source properties considered most relevant when determining the confidence value output by the GP, the next section investigates the distribution of confidence values as a function of these source properties. It further compares the distribution of confidence values with that over the EXT and EXT_LIKE source properties used to generate the initial C1C2 source classifications.

5.2 Interpreting the distribution of confidence values

Figures 5.2, 5.3, 5.4 and 5.5, compare the distribution of confidence values for EXT, EXT_LIKE, 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 the GP trained on the southern field (bottom panel of Figure 5.1). The measured source properties, EXT_RATE_PN and PNT_RATE_PN, are identified as the most relevant and EXT_BG_RATE_PN as irrelevant when the GP is trained on the North XXL

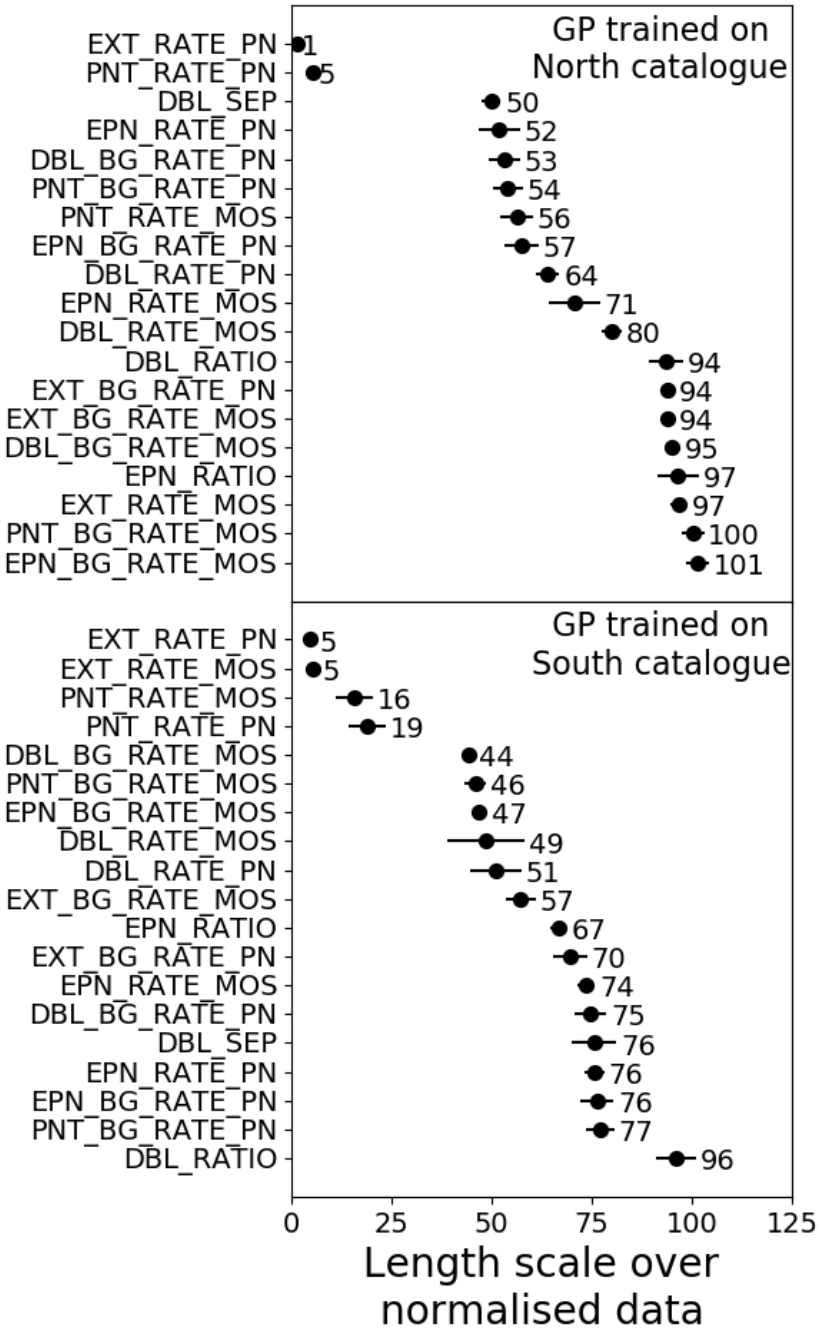


Figure 5.1: The length scale of the Gaussian kernel for each parameter used by the GP. Due to the method of normalising the data, the lengths are expressed in units of the standard deviation of each parameter in the input catalogue (see Section 4.1 for details). Shorter length scales correspond to parameters 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.

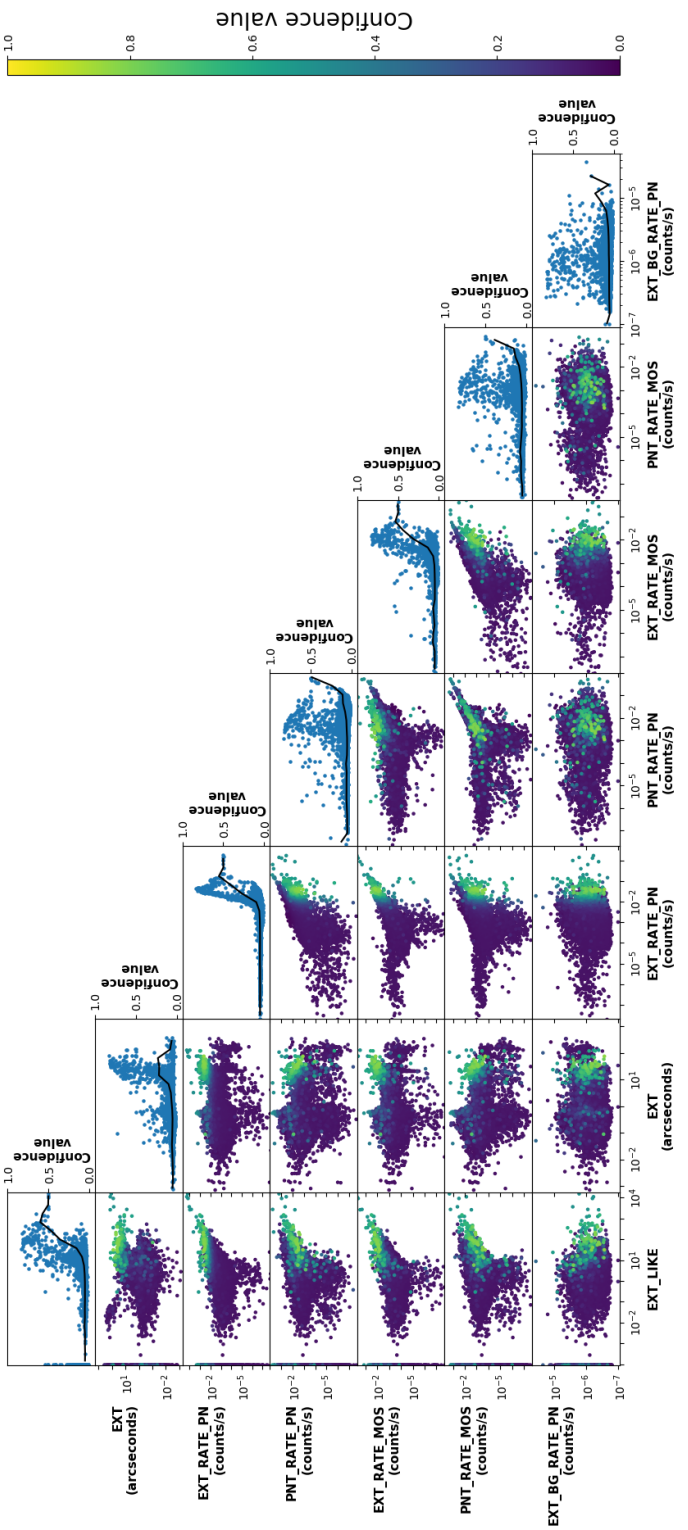


Figure 5.2: The distribution of the sources in the North catalogue in a subset of the full parameter space chosen to illustrate the behaviour of the GP. The parameters EXT and EXT_LIKE were used to label the sources but were not input to the GP. EXT_RATE_PN and PNT_RATE_PN were considered relevant by 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 Figure 4.1. 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.

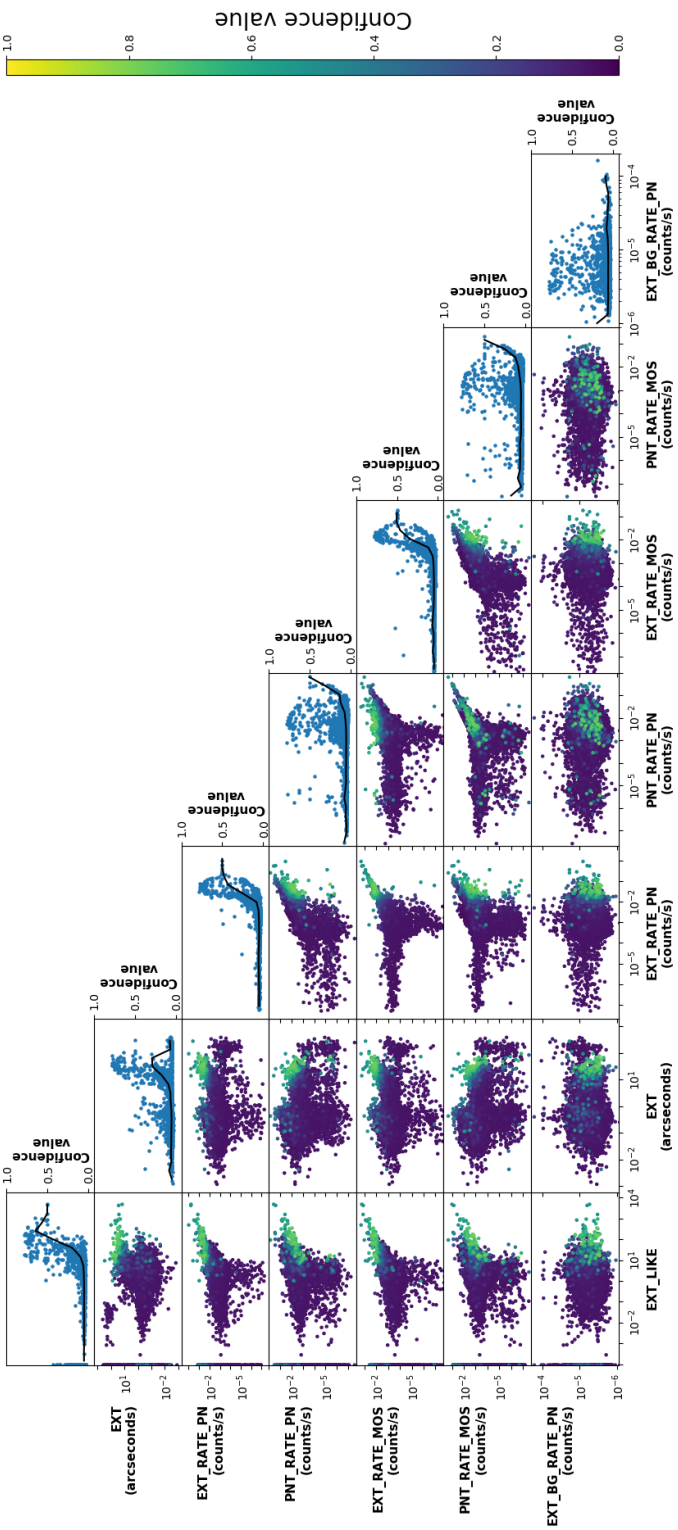


Figure 5.3: As for Figure 5.2, but for the sources in the South field when the GP was trained on the South catalogue.

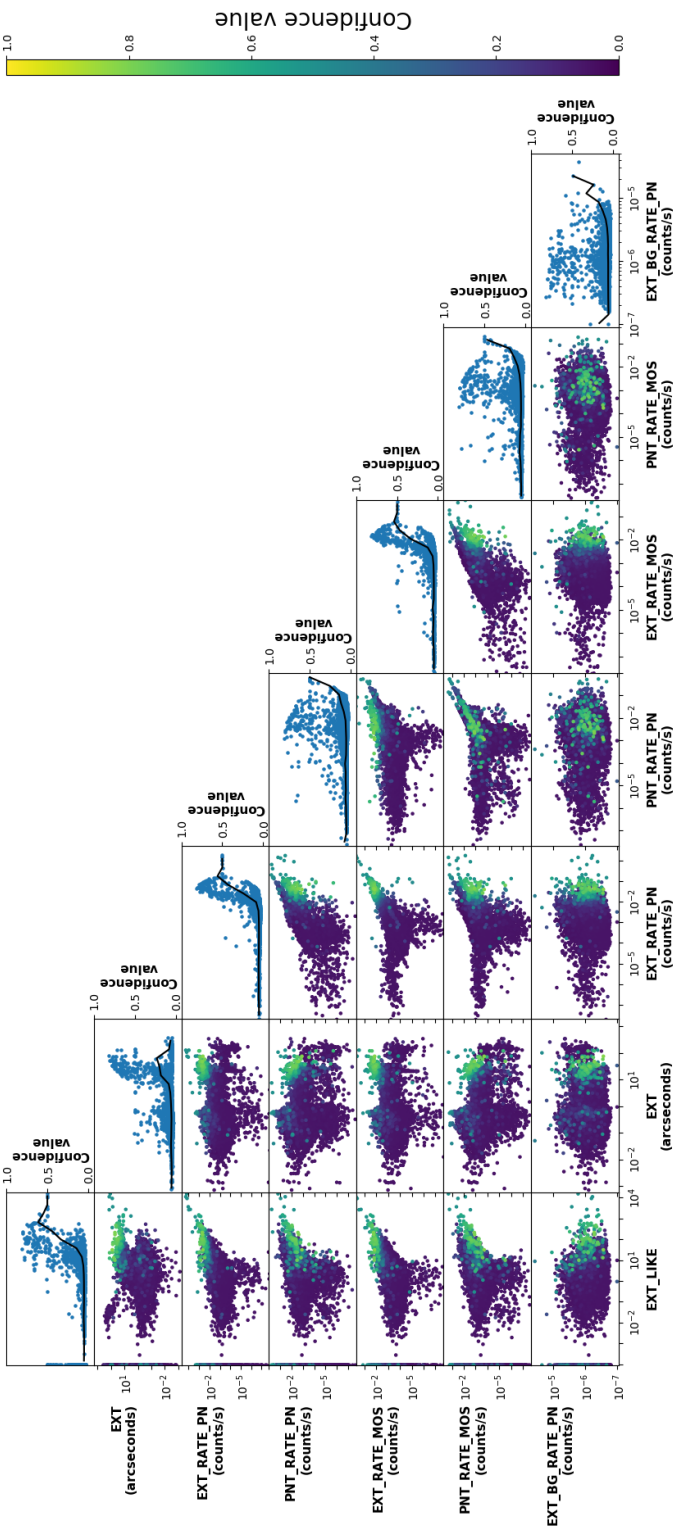


Figure 5.4: As for Figure 5.2, but for the sources in the North field when the GP was trained on the South catalogue.

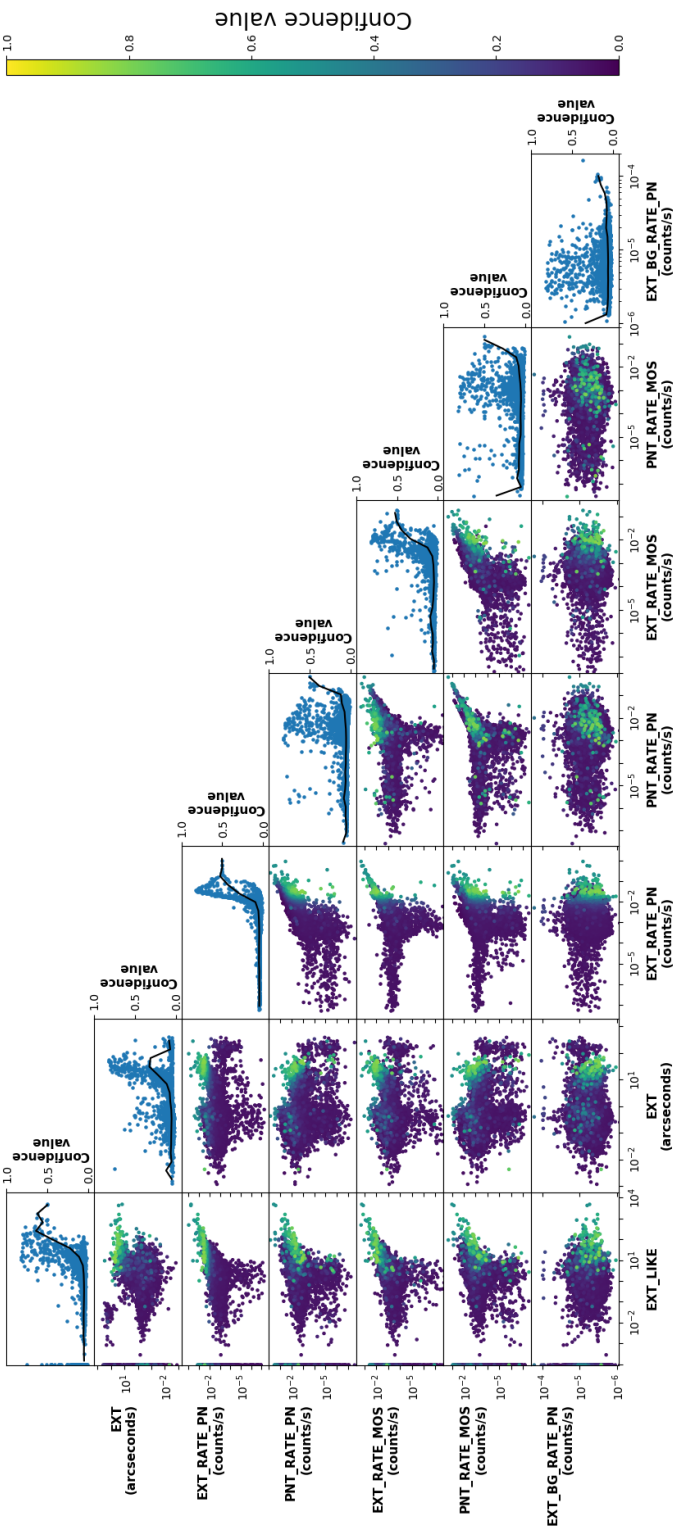


Figure 5.5: As for Figure 5.2, but for the sources in the South field when the GP was trained on the North catalogue.

source catalogue. Each plot shows the position in the two-dimensional parameter space formed by pairs of measured properties for each source, colour coded by the confidence value returned by the GP for that source. The scatter plots at the top of each column show confidence value against the measured source property, the black line indicating the mean confidence value in logarithmic bins.

It is immediately apparent from Figures 5.2, 5.3, 5.4 and 5.5, that there is no correlation between EXT_BG_RATE_PN and the confidence value assigned to a source by the GP trained on either XXL catalogue. This supports the previous interpretation of EXT_BG_RATE_PN having little to no impact on the confidence value, based on its length scale reported in Figure 5.1. Similar inspection of the other source properties determined irrelevant based on their length scales, shows the same absence of any correlation with the confidence value assigned to a source.

Of those four measured source properties considered relevant, based on the length scales determined by ARD when trained on the southern catalogue, there is clearly some relationship with the confidence value assigned by the GP trained on either the North or South catalogue. It is clear that sources with higher confidence values tend to have higher values for the four measured source properties, particularly EXT_RATE_PN and EXT_RATE_MOS. To determine why this relation exists and how the GP has identified it, we first investigate the relation between EXT_RATE_PN and EXT_LIKE. Subsequently we investigate the relation between a sources position in the two dimensional space defined by EXT_RATE_PN and PNT_RATE_PN, and its measured EXT. We leave analysis as to why this signal is present for both the GP trained on the North and that trained on the South catalogue for the following section, despite ARD deeming that EXT_RATE_MOS and PNT_RATE_MOS are only relevant when the GP is trained on the South catalogue.

Figure 5.6 depicts the distribution of sources as a function of their EXT_LIKE and EXT_RATE_PN values, with sources colour coded by their confidence values assigned by the GP when trained on the North or South catalogues. It is apparent that there exists a correlation between EXT_LIKE and EXT_RATE_PN. This correlation is to be expected as EXT_LIKE is calculated by taking the ratio of the likelihood of the fitted extended source being detected by the XAMIN pipeline and that for the fitted point source. Since a brighter source is more likely to be detected by XAMIN, the brighter the extended source (larger EXT_RATE_PN) the larger the value of EXT_LIKE measured by XAMIN. While the GP is not privy to EXT_LIKE, it is given the C1C2 classifications in the form of the training labels which are themselves partially based on EXT_LIKE. This means that the training labels inherently correlate with EXT_RATE_PN, C1 and C2 sources tending to have higher values of EXT_RATE_PN. The GP has identified this trend and exploited it when assigning confidence values.

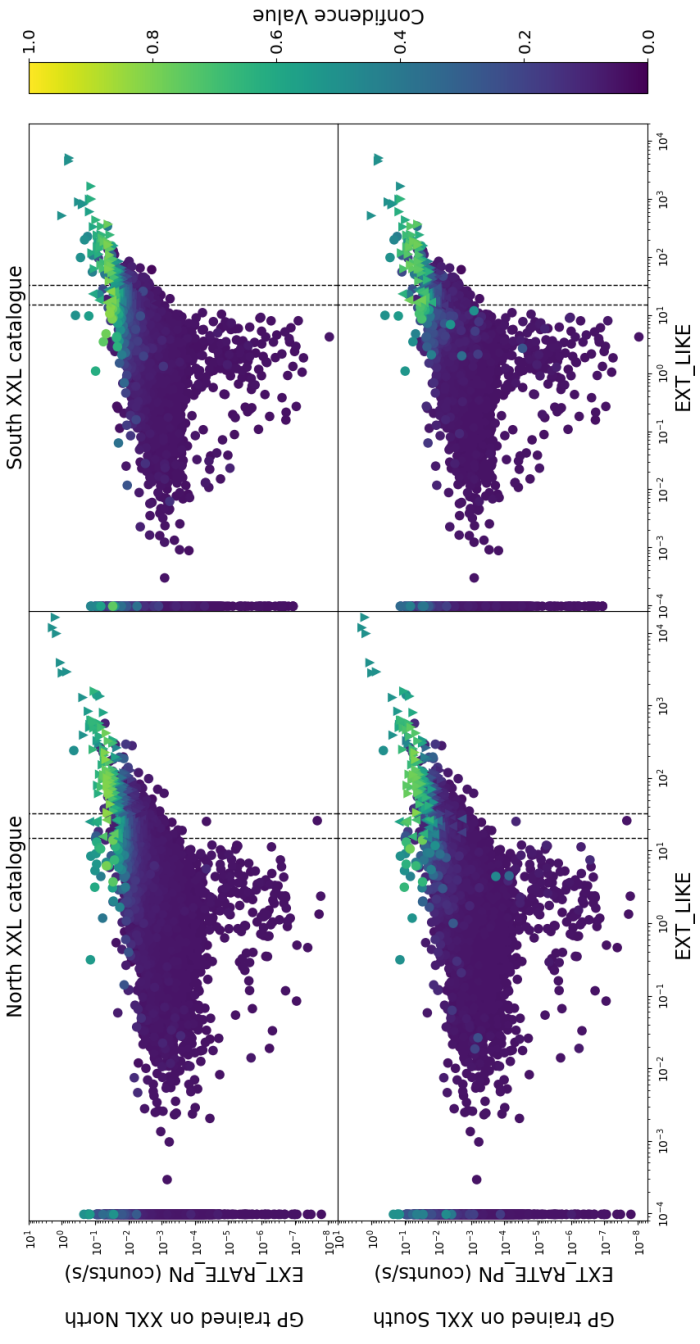


Figure 5.6: Scatter plot of sources EXT_RATE_PN and EXT_LIKE for the North XXL catalogue (left column) and South catalogue (right column), colour coded by confidence value assigned by the GP when trained on the North catalogue (top row) and South catalogue (bottom row). Within each plot, $C1$ and $C2$ sources are plotted as triangles pointing down and up respectively with all remaining sources denoted by a circle. Higher confidence sources are plotted over lower confidence sources as in Figure 4.1. Dashed lines indicate the C1 and C2 selection criteria for a sources measured EXT_LIKE value (table 4.1).

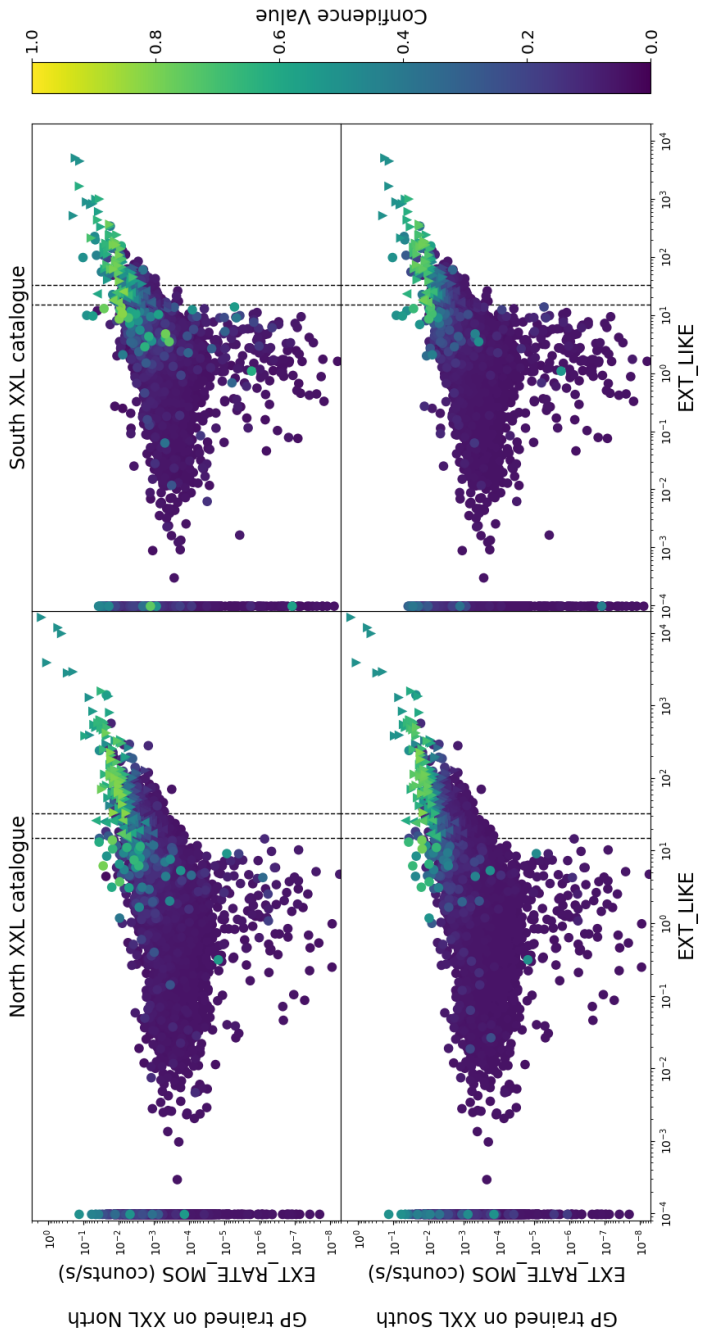


Figure 5.7: The same as figure 5.6 but for the distribution over EXT_RATE_MOS and EXT_LIKE.

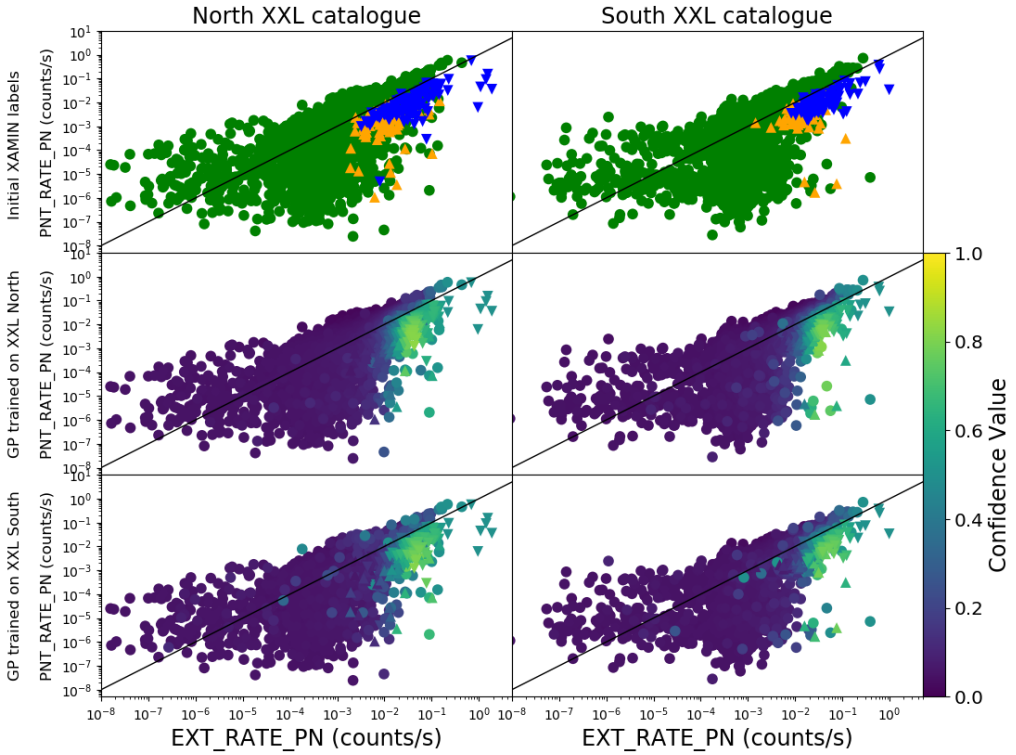


Figure 5.8: Distribution of sources from the XXL North (left column) and South (right column) fields as a function of EXT_RATE_PN and PNT_RATE_PN. From top to bottom each row is colour coded by, the initial labels (C1 blue, C2 orange, and non-C1C2 green), the confidence value assigned by the GP trained on the North catalogue, and the confidence value assigned by the GP trained on the South catalogue. Within each plot C1 and C2 sources are plotted as triangles pointing down and up respectively, with all remaining sources denoted by a circle. For sources colour coded by confidence value high confidence value sources are plotted over low confidence sources as in Figure 4.1. The dotted line indicates a one to one relation.

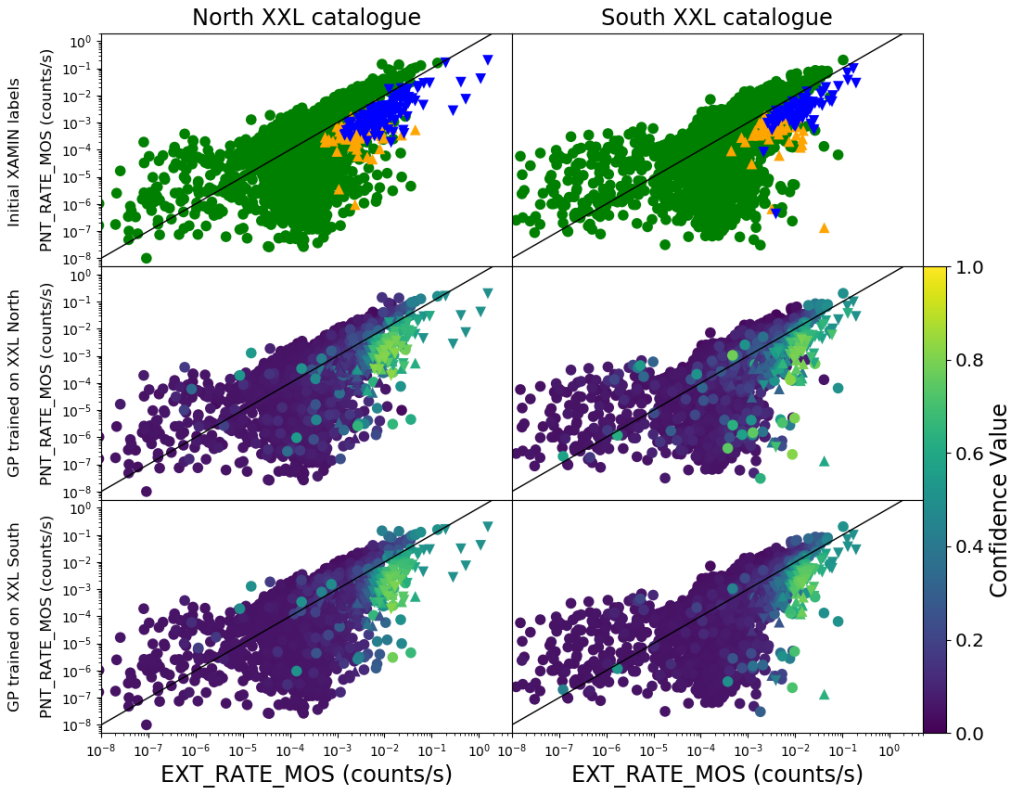


Figure 5.9: The same as Figure 5.8, but for the distribution over EXT_RATE_MOS and PNT_RATE_MOS.

With respect to the distribution of confidence values over the two dimensional space defined by EXT_-RATE_PN and PNT_RATE_PN (Figure 5.8), there is a distinct population of higher confidence sources. The population not only exists at large values of EXT_RATE_PN and PNT_RATE_PN as noted earlier, but favours those sources with a higher EXT_RATE_PN than PNT_RATE_PN (the high confidence sources sitting below the dotted one to one line). This same behaviour can be seen in the distribution of C1 and C2 sources compared to that of the whole source population. While the GP has simply identified this trend based on the initial C1C2 labels, there must exist some relationship between this trend and whether a source is a detection of the extended X-ray emission from a galaxy cluster.

To understand the trend seen in Figure 5.8, consider the results of fitting a point source model to an extended object. When fitting a point model to an extended source, the model is unable to treat emission from the outer regions of the source as coming from the source, the point model only allowing for emission from a singular point on the sky. The emission from the outer regions is instead treated as background. This has the effect of producing a lower measurement of the count rate of said source, compared to that measured when fitting an extended model. The result of this quirk of the fitting process is that genuinely extended sources can be identified from their position in EXT_RATE_PN and PNT_RATE_PN space, as seen in Figure 5.8.

While the investigation above has focused on the source properties measured for the PN camera, the same trends are seen when investigating the measured properties for the combined MOS cameras. Specifically, the trends for sources assigned a high confidence value to have a high EXT_RATE_MOS (Figure 5.7) that is also larger than their PNT_RATE_MOS (Figure 5.9) is seen for both the GP trained on the North catalogue and trained the South catalogue. This is despite ARD determining that EXT_RATE_MOS and PNT_RATE_MOS are only relevant, when the GP is trained on the South XXL catalogue and not when trained on the north. The next section considers the details of this difference and discusses potential causes.

5.3 Differences between the North and South Catalogues

With respect to the relevance of individual measured source properties, as determined by ARD, the main difference between the GP when trained on the North or the South catalogue is the relative importance of the EXT_RATE_MOS and PNT_RATE_MOS source properties. EXT_RATE_MOS and PNT_RATE_MOS are considered irrelevant when the GP is trained on the North catalogue (they have a large length scale, Figure 5.1). In contrast when the GP is trained on the South catalogue they are a considered relevant, with length scales comparable to EXT_RATE_PN and PNT_RATE_PN. There is also an increase in the length scales determined for EXT_RATE_PN (from one to five) and PNT_RATE_PN (from five to nineteen). While the length scales of the remaining measured source properties do change, depending on

which field is used for training, the values are sufficiently large that they are irrelevant in determining the confidence value and hence their relative value does not contain any useful information.

Before considering potential differences between the North and South fields, there are two possible origins for the difference in length scales relating to the implementation of the GP model. The first is that the difference is caused by over-fitting or some other systematic effect that occurs when training the GP. As discussed previously in section 4.2 we find no evidence to suggest that this has occurred. The second possibility is that differences in the values used to normalise a measured source property when training on the different XXL catalogues are responsible for the difference in the derived length scale of that property. As described in 4.1, each measured source property was normalised by the standard deviation of its values in the training data. This normalisation is what allows us to directly compare the length scales, hence any significant changes in the standard deviation between the two XXL catalogues could affect the relative importance of a source property. An investigation found that the changes in standard deviation for all source properties was insufficient to explain the differences seen in the derived length scales and the inferred importance of each source property. Having investigated the possible causes of the difference in length scale when the GP is trained on the North or South catalogues associated with the GP, we rule out the GP as a possible cause of the difference.

Any differences between the North and South catalogues that affect the output of the GP must exist within the four count rates identified as relevant when the GP is trained on either the North or South catalogue. To investigate potential differences Figure 5.10 depicts the cumulative distribution of sources from both the North (blue) and South (orange) catalogues as a function of each of the four count rates. It is immediately apparent from Figure 5.10 that for all four count rates the distribution of sources within the North catalogue is shifted to lower measured count rates than that for the south. The main difference between the North and South observations is the presence of a region of deeper observations within the North field (Pierre et al., 2016), the remainder of the North field is observed to the same depth as the whole of the South field. The presence of these deeper observations allows for the detection of fainter sources and hence shifts the distribution of count rates seen in Figure 5.10 to lower values. Removing those sources from the North catalogue that are detected within one of these deeper observations, accounts for the shift in measured rate values (Figure 5.11).

The simplest way to test if the inclusion of those sources detected within a deeper observation, causes the difference in length scales, is to train the GP on the North catalogue with said sources removed. The test was conducted over a total of 30 Monte Carlo iterations split into three batches of ten iterations, with the resulting length scales plotted in the top panel of Figure 5.12. It is apparent when comparing Figures 5.1 and 5.12 that the removal of those sources detected within the deeper observations of the North catalogue do not significantly affect the length scales. This effectively rules out the presence of the

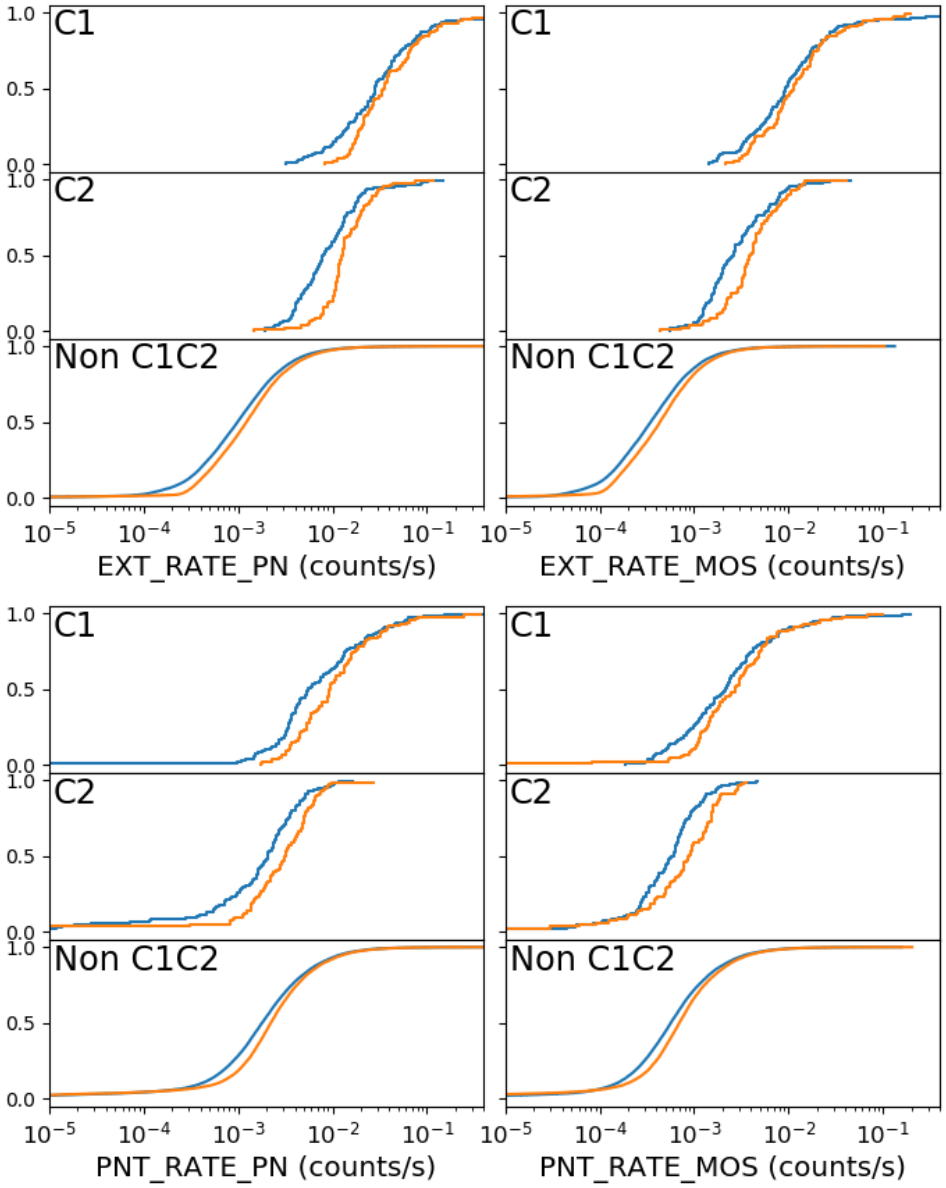


Figure 5.10: Normalised cumulative distribution of sources within the North (blue) and South (orange) XXL catalogues as a function of EXT_RATE_PN (top left), EXT_RATE_MOS (top right), PNT_RATE_PN (bottom left) and PNT_RATE_MOS (bottom right). Within each panel the source catalogues are split into the C1, C2 and non-C1C2 source samples.

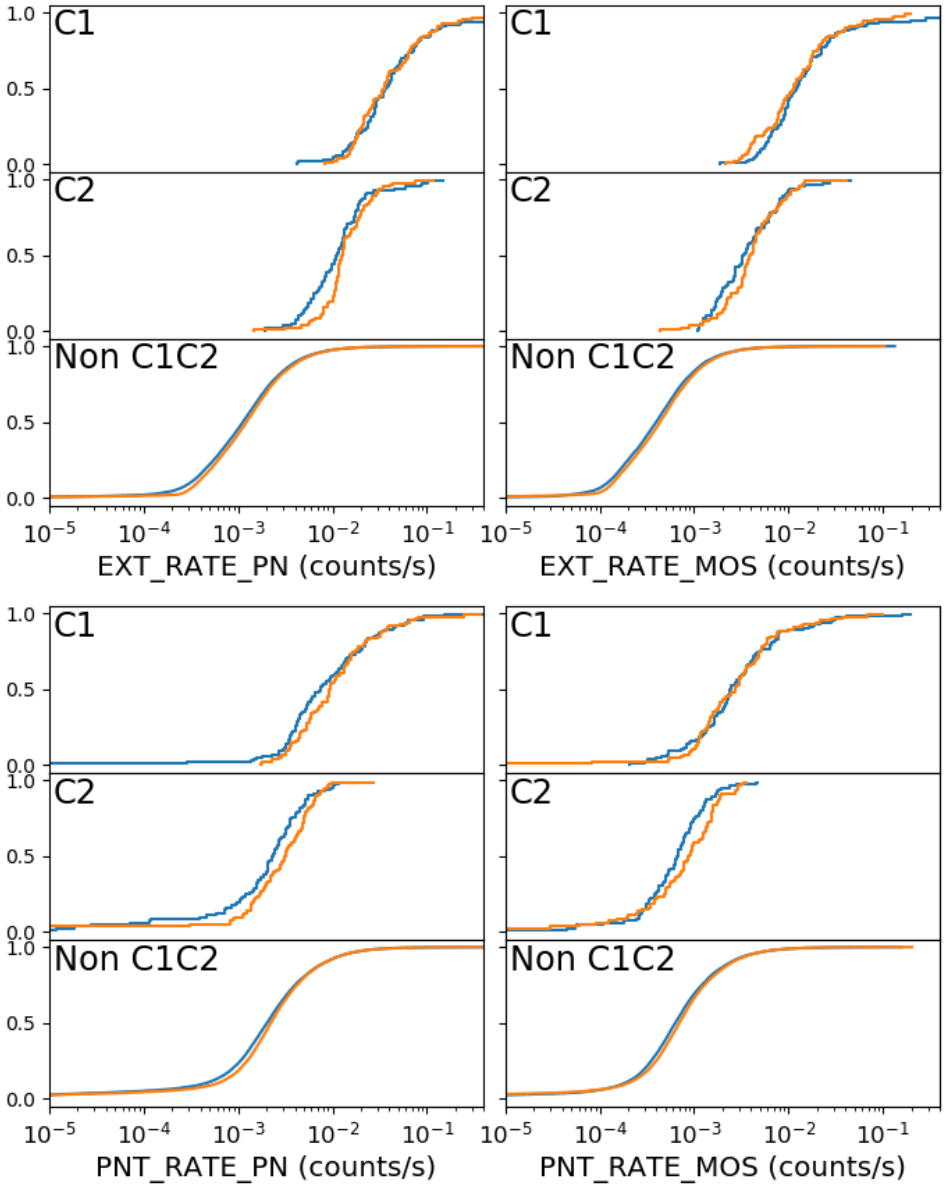


Figure 5.11: As in Figure 5.10 but with those sources detected in the region of the North field with deeper observations removed from the North catalogue.

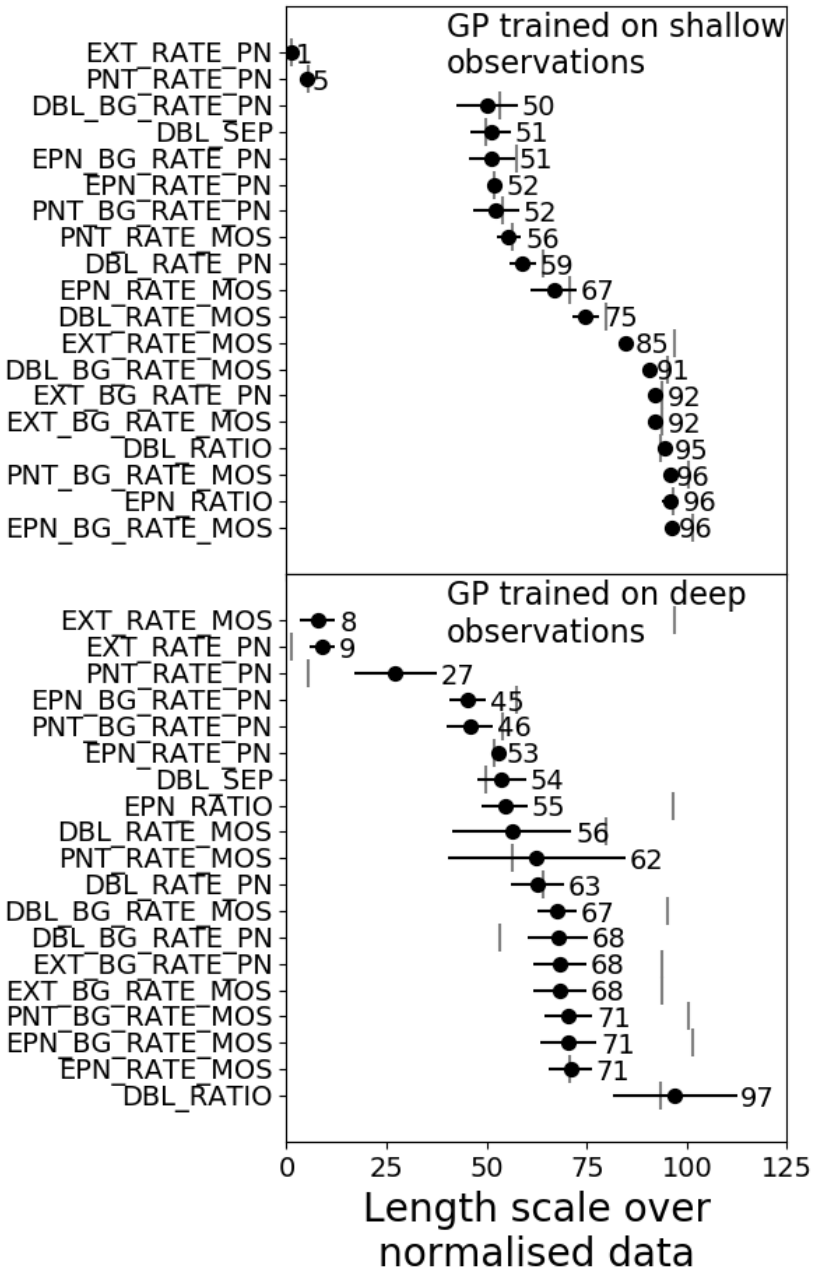


Figure 5.12: The same as Figure 5.1 except the length scales are determined for the GP when trained on the North catalogue with those sources detected within the deeper observations removed (top) and trained on only those sources detected within the deeper observations (bottom). The grey horizontal lines show the corresponding length scales determined by the GP when trained on the full North catalogue.

deeper observations within the North field as the source of the discrepancy in length scales and confidence values between the North and South fields.

For comparative purposes, the GP was also trained on only those sources within the North catalogue detected within deeper observations (bottom panel of Figure 5.12). When the GP is trained only on sources in the North catalogue detected in a deep observation, the length scales are significantly different from those determined when the GP is trained on the whole North catalogue. This further emphasises that those sources detected within the deeper observations are not the source of the discrepancy between the GP when trained on the full North or South catalogues.

Both the GP and the presence of a set of deeper observations in the North field have been ruled out as potential causes of the difference in length scales in Figure 5.1. Further, the XAMIN pipeline is ruled out as a possible cause as it was not changed between being applied to the North and South fields. While the potential exists for differences between the extragalacitc source populations that make up the North and South fields (Migkas et al., 2021) to be the cause of the behaviour seen in Figure 5.1, any affect due to these differences would most likely be masked by other, stronger factors assuming the GP were able to identify it at all. The behaviour seen in Figure 5.1 is more likely the effect of differences in the path through the Galaxy to the two fields (i.e. absorbing column as shown in figure 2.1 or diffuse foreground emission), or differences in the parameters of the observations carried out (depth, background, calibration changes, observing mode etc). A more detailed investigation is needed to determine the exact origin of the difference in length scales. Any such investigation benefiting from the inclusion of data from the XClass serendipitous X-ray galaxy cluster survey (Clerc et al., 2012) due to it's wider distribution over the sky.

While the reason for this behaviour is not clear, there exists a significant correlation between the confidence values assigned to sources by the GP when trained on either field (figure 4.4). The existence of the correlation implies the confidence values output by the GP are robust to the differences between the two fields. For example, both instances of the GP identify mostly the same non-C1C2 sources as being more likely to be a cluster. It is not possible to say which confidence value is more appropriate (that assigned by the GP trained on the North or South catalogue), but selecting a subsample based on the north-trained confidence values would be more effective at separating out C1 and C2 clusters and identifying non-C1C2 sources with comparable confidence values.

5.4 Discussion

In this chapter is has been established that the GP is identifying sources as galaxy cluster candidates based on two criteria; their measured EXT_RATE_PN is high and the ratio of their measured values for

PNT_RATE_PN and EXT_RATE_PN is approximately less than one (section 5.2). Given this information it is possible to investigate why, when selecting galaxy cluster candidates from the non-C1C2 sample (section 4.6), the sample is found to include both bright point sources and spurious detections of background fluctuations, as identified during visual inspection (section 4.5). Further, it is possible to explore the limitations of the GP selected sample with respect to the three types of genuinely extended X-ray sources also identified during visual inspection. The three types of genuinely extended X-ray sources found being; (i) extended but irregular sources associated with galaxy overdensities; (ii) nearby groups with X-ray emission dominated by the halo of the brightest galaxy; and (iii) extended sources polluted by a dominant central AGN. See Figure 4.13 for examples of each type of source.

Despite not being physically extended, bright point sources have a sufficiently high number of detected photons that XAMIN, even when fitting an extended model, recovers a high count rate. This high number of photons alone does not however explain how a bright point source achieves the second criteria, having values of EXT_RATE_PN and PNT_RATE_PN such that their ratio is approximately less than one. There are two explanations for how a point source is able to satisfy this criteria. The first is that a source with a sufficiently high EXT_RATE_PN (> 0.1) will be assigned a confidence value of 0.5 regardless of the ratio between EXT_RATE_PN and PNT_RATE_PN (Figure 5.8). This occurs due to the isolated nature of such sources with a high EXT_RATE_PN causing the GP to default to its prior estimate of the confidence value of 0.5. While such sources could be easily removed from the sample by applying a maximum value for EXT_RATE_PN or PNT_RATE_PN there also exist a number of sources with similarly high EXT_RATE_PN and PNT_RATE_PN values, identified as galaxy cluster candidates via visual inspection (section 4.5).

Alternatively, the appearance of a point source can be distorted to appear physically extended. The main factor in the distortion of a sources appearance, being XMM-Newtons point spread function (Read et al., 2011), though other factors including coincident sources, gradients in the background flux and the edges of overlapping pointings also contribute. The result of a point source being distorted to appear more physically extended is that, despite XAMIN accounting for the point spread function, the extended model is better able to model the surface brightness profile of the source. This has the result of increasing the count rate measured by the extended model as photons are distributed over an extended region of sky. The opposite affect occurs for the point source model, reducing the count rate it measures. The increase and decrease in count rate for the extended and point source models respectively, reduces the ratio of PNT_RATE_PN and EXT_RATE_PN, making a point source appear more like an extended source to the GP. The result of the distortion and isolation of high count rate point sources, is for the GP to assign them higher confidence values, leading to their inclusion in the sample of cluster candidates (as described in section 4.6).

The other type of unwanted sources found in the GP selected sample are spurious detections of background fluctuations. Such detections occur when the wavelet filtering applied by XAMIN (Pacaud et al., 2006) convolves random background fluctuations into an apparent broad, flat "source". Due to the apparent size of the "source" and mislabelling of background photons, the count rate measured when fitting is sufficiently high to satisfy the first condition identified by the GP (a high EXT_RATE_PN). We note that the typical EXT_RATE_PN measured for spurious background detections is generally less than that measured for the bright point sources discussed previously.

With respect to the second condition imposed by the GP, the ratio of PNT_RATE_PN and EXT_RATE_PN being approximately less than one, spurious background detections by their nature have little to no photons detected in their centre resulting in a low measurement of PNT_RATE_PN. The ratio of PNT_RATE_PN and EXT_RATE_PN hence is much lower than one. Combining this with the high EXT_RATE_PN of the spurious detection explained above, means that GP assigns the source a high enough confidence value to be selected. Such sources could be removed from the sample by only selecting sources with a PNT_RATE_PN above some cut, however this would also remove genuinely extended sources.

The remaining types of sources not previously selected by XAMIN but included in the GP selected sample, are all genuine detections of sources with extended X-ray emission and so should be included in the GP selected sample. As noted previously the extended sources fall into three types; (i) extended but irregular sources associated with galaxy overdensities; (ii) nearby groups with X-ray emission dominated by the halo of the brightest galaxy; and (iii) AGN contaminated extended sources (AC). Having identified the criteria with which the GP assigns confidence values it is possible to conduct a qualitative investigation of the limitations of the GP when selecting each type of genuinely extended sources. The main limiting factor for all three types of object is the need for a measured EXT_RATE_PN that is $\gtrsim 0.02$ counts^{-1}. This factor reduces the ability of the GP to select genuinely extended sources that are fainter and/or more distant. Visual inspection of sources with confidence below the selection cut (section 4.5), has shown the existence of sources considered galaxy cluster candidates with measured values of EXT_RATE_PN below this cut of ~ 0.02 counts per-second.

Among the three types of extended source identified, AC objects are less susceptible to having a low EXT_RATE_PN due to the point source contributing to the measured rate. The bright point source however will also increase the measured value for the PNT_RATE_PN, this increase being larger than that for EXT_RATE_PN. The result is an increase in the ratio of PNT_RATE_PN to EXT_RATE_PN, such that the source appears to the GP to be less extended, reducing the confidence value assigned by the GP. This limits the GP to identifying AC objects where the extended emission is dominant.

XAMIN attempts to solve the problem of identifying AC sources by explicitly looking for them in addition to looking for uncontaminated extended sources (Logan et al., 2018; Bhargava et al., 2023). The selection of AC sources using the results of fitting an extended plus point source (EPN) model as part of the XAMIN pipeline. While the outputs of the EPN model are provided to the GP, Figure 5.1 shows that such parameters are inconsequential when the GP assigns a confidence value to a source. This ignoring of the EPN model by the GP is not because they are not informative in identifying AC sources, but because the GP was not explicitly told to look for such sources. The GP being told to look for extended sources similar to those that satisfy the C1 or C2 selection criteria. An AC source is less likely to satisfy the C1 or C2 selection criteria compared to a purely extended source, the presence of the AGN distorting the appearance of the AC source to be less extended and more point like. To better select extended AC sources, the GP would need to be explicitly trained to look for such sources, be it by including them as an extended object or by training the GP to identify them as a separate class entirely.

5.5 Summary

The purpose of this chapter was to investigate and identify the selection criteria used by the GP to classify X-ray sources as potentially extended. To achieve this goal, ARD was used to identify those source parameters that most strongly influence the confidence value output by the GP. Specifically when trained on the North XXL source catalogue, the output of the GP was found to strongly depend on EXT_RATE_PN and PNT_RATE_PN, with all other parameters having little to no effect on the output. Analysis of the distribution of confidence values as a function of EXT_RATE_PN and PNT_RATE_PN showed that the GP uses two criteria when assigning a source a high confidence value; i) a source must have a high EXT_RATE_PN and ii) the ratio of PNT_RATE_PN to EXT_RATE_PN must be less than approximately one.

In explaining the reason for the GP to learn such criteria, it was shown that the first criteria (the need for a high EXT_RATE_PN) is learnt by the GP because the cut on EXT_LIKE used to select C1 and C2 sources is only satisfied by bright sources. Hence when the GP is told to look for objects similar to and including those labelled as a C1 or C2 source, it learns to look for bright sources. The second criteria; the ratio of the measured values of PNT_RATE_PN and EXT_RATE_PN being less than approximately one, relates to how extended a source is. The more extended a source, the less well the point source model is able to fully capture the source's count rate in comparison to the extended model. The result being that the ratio of PNT_RATE_PN to EXT_RATE_PN decreases as a source becomes more extended. The GP is able to identify and exploit this trend since the C1 and C2 sources it is told to look for, exhibit a low value for this ratio due to being extended.

When conducting the same analysis of the GP trained on the southern XXL catalogue it was found that, not only was the confidence value dependent on EXT_RATE_PN and PNT_RATE_PN, but also EXT_RATE_MOS and PNT_RATE_MOS. While the difference between the GP when trained on the two catalogues was not found to have a significant impact on the confidence value assigned to a source (see section 4.2.1), the discrepancy in parameter relevance is, indicating some underlying difference between the North and South XXL catalogues. A number of possible sources of the discrepancies were ruled out, including the GP being applied differently, the presence of a set of pointings in the North field conducted with longer exposure times, and the XAMIN pipeline. The remaining conclusion is that there exists some inherent difference in the path through the Galaxy to the two fields or the parameters of the observations used to construct the two fields.

The presence of point sources and spurious background detections within the GP selected sample (section 4.5), was discussed in the context of the criteria used by the GP. It was shown that, despite not being extended, XAMIN will measure a high EXT_RATE_PN for a sufficiently bright point source. Further distortions in the shape of a point source, (including those due to the point spread function of XXM Newton) make it appear more extended, reducing the ratio of PNT_RATE_PN to EXT_RATE_PN. In contrast, spurious detections of background fluctuations tend to produce broad flat profiles, such that they appear extended to the GP, whilst also encompassing a large enough number of photons to produce a high measurement of EXT_RATE_PN.

6

Conclusions and Future Work

The aim of this work was to develop and apply a machine learning (ML) solution to select galaxy clusters from the XXL X-ray source catalogue (Pierre et al., 2016). The Gaussian Process (GP) adapted to achieve this aim was described in chapter 3. The GP was applied to select galaxy clusters from the XXL X-ray source catalogue in chapter 4, where its ability to select cluster candidates was also assessed. Following this, chapter 5 detailed the identification of the criteria by which the model selects galaxy cluster candidates. This final chapter provides a summary of this work before discussing potential ways to build on it, including ways to supply the GP model with additional information to better select clusters and the applicability of the model beyond searches for galaxy clusters.

6.1 Summary

When looking to apply a ML binary classifier to select sources of a given type from survey data, the need for a sufficiently large and perfectly labelled training set that accurately replicates real data poses a significant issue. There currently exist three approaches to solving this problem. The first approach involves the accurate labelling of real sources by experts which, in order to produce a sufficiently large training set, requires a significant and often prohibitive time and resource investment on the part of the astronomer. The second approach makes use of citizen science, where a large number of non-experts label the data with the hope that, on average, the correct label is applied to each source. This crowd sourcing approach however produces less accurate labels than those by experts and can still require a significant investment to setup. The final approach is the use of simulated data labelled using knowledge

of the simulated sources properties. This approach relies on the simulation accurately recreating real data to avoid the model learning a solution that does not transfer to real data, despite accurate results when tested on the simulated data.

This work presented an alternative approach that makes use of existing source samples with known purities as the labels on the training set. We outlined in chapter 3 the core principles of a GP and how a GP binary classifier was adapted to account for uncertainties on the training labels. The benefit of adapting a GP to account for these uncertainties is that it is able to take as training input a source catalogue labelled by how likely each source is to be the targeted object type. An initial prediction of probability that any one source is the desired type can be estimated by the purity of the sample containing the source with respect to that type. For example in the case of XXL we set the initial probability of a source being a galaxy cluster to 0.95, 0.5, or 0.05 if it is contained by the C1, C2 or neither cluster sample respectively. While the results of this approach will be less accurate than a perfectly labelled real training set, the use of existing source samples to create labels significantly reduces the resource cost when labelling the training set whilst also avoiding the issue of simulated data not accurately recreating real data.

The adapted GP binary classifier is applied to select galaxy cluster candidates from the XXL X-ray source catalogue (Pierre et al., 2016) in chapter 4. The model was separately trained on the North and South catalogues labelled using the existing C1 and C2 cluster samples as outlined above. Each source was subsequently assigned two different confidence values, one by the GP trained on the North catalogue and the other by the GP trained on the South catalogue. Comparisons of the assigned confidence values showed no signs of over fitting and a slight tendency for the South trained GP to assigned lower confidence values than the North trained GP.

The ability of the GP to identify sources with an increased probability of being a galaxy cluster was tested in three ways in chapter 4. First, each source in a simulated XXL catalogue covering 25 square degrees (described in section 2.1.3) was assigned a confidence value by the GP trained on the North and by that trained on the South. It was found that the simulated XXL catalogue did not sufficiently replicate the distribution of source properties found in the North or South catalogues to be of use when testing the output of the GP (section 4.3).

A set of optically selected galaxy clusters from the `CAMIRA` cluster catalogue were used to identify (in a way that is independent of the GP) a set of XXL X-ray sources that have an increased likelihood of being a galaxy cluster relative to the general population of XXL sources (see section 2.2.1 for a description of how said sources were identified). It was shown that this set of sources were assigned an increased confidence value relative to the general population of XXL sources (section 4.4). This relative increase in the confidence values of sources that were independently identified as having an increased likelihood of

being a cluster detection indicates that the GP, when trained on the North or South catalogue, is identifying sources with an increased likelihood of being a cluster.

The final test involved the visual inspection and classification of a subset of sources within the North catalogue. The visual inspection process making use of a combination of XXL X-ray observations and Hyper Suprime-Cam (HSC) observations to produce multi wavelength cutouts of each source. The set of visually inspected sources was used to assess cluster samples selected on confidence value (section 4.5). It was identified that selecting those sources with a confidence value greater than 0.1 from the North field produced a galaxy cluster sample of reasonable size and purity. The galaxy cluster sample produced in this way contained a total of 623 sources with a purity of $0.45^{+0.03}_{-0.03}$. Given this, we expect the sample to contain a total of 280 sources that would be labelled as a galaxy cluster via visual inspection, 101 of which are expected to not previously have been selected as part of the C1 or C2 samples.

One issue with applying a ML solution to the selection of astrophysical sources is the difficulty in identifying the criteria by which the model has learnt to identify sources of interest. The benefit of using a GP based approach and one reason it was selected for this work, is the ability to use automatic relevance determination (ARD; see section 3.5.2) to optimise and identify the most important inputs. This work made use of ARD in chapter 5 to identify those source properties measured by XAMIN that are most relevant when calculating the confidence value output by the GP (figure 5.1). We found that the count rates measured when separately fitting the extended and point source models are the most relevant measured property when determining the confidence value assigned by the GP to a source. Further investigation showed that the GP assigns increased confidence values to a source if they have a high measured count rate when fitting an extended source that is also approximately larger than the measured count rate when fitting a point source model.

Comparing the relative importance of different measured source properties for the GP when trained on the North and South catalogues showed significant discrepancies in the relevance of the count rates measured when fitting a point source or extended source to data from the MOS cameras (Figure 5.1). A detailed investigation of potential causes of this discrepancy effectively ruled out both the implementation of the GP and the presence of deeper observations in the North XXL field. With XAMIN, the observing methodology and the properties of the cluster populations being consistent across both fields, the only remaining potential cause is the existence of some intrinsic difference along the line of sight to the clusters. A more detailed follow up investigation is needed to establish the exact cause.

6.2 Future work

Of the galaxy cluster sample created by selecting those sources in the North catalogue assigned a confidence value higher than 0.1, by the GP trained on the North catalogue, only a subset have been visually inspected to identify galaxy cluster candidates. To produce a finalised sample of galaxy cluster candidates it will be necessary to visually inspect the remaining sources within the sample, filtering out non-clusters. Those galaxy cluster candidates identified by visual inspection will subsequently need some form of spectroscopic confirmation before being categorically labelled a galaxy cluster. The same process is necessary to produce a catalogue of galaxy clusters from the South catalogue. The process of visually inspecting sources in the South catalogue is complicated however by the absence of any overlapping HSC observations. It will be necessary to identify a different set of optical observations in order to visually inspect sources in the South field.

If the galaxy cluster catalogues produced using the GP are to be of use for cosmological analysis it is necessary to determine the catalogues' selection function. The complex selection criteria learnt by the GP make a purely analytic description of the selection function impossible, instead it is necessary to make use of simulated data. As described by Pacaud et al. (2016), when determining the cluster selection function for the bright cluster sample a combined analytic and Monte Carlo approach was used. A similar approach should be applicable when calculating the selection function for the GP selected cluster catalogue. As with any approach that makes use of simulated data to determine a catalogue's selection function it is necessary for the simulation to accurately recreate all source properties used by the GP when selecting galaxy cluster candidates. This presents a potential issue for any ML based selection method because they make use of a larger number of source properties, compared to non-ML methods, all of which must be accurately replicated by the simulated data. A solution to this problem beyond focusing on maximising the accuracy of the simulation is not immediately apparent and needs investigation. One potential benefit of using a GP here is the ability to identify those source properties relevant to the output confidence value. Given this information it is possible to focus on accurately simulating those properties over the irrelevant ones.

Looking beyond XXL, the GP can be applied to select galaxy cluster candidates from other larger X-ray surveys. For example the X-CLASS survey (Clerc et al., 2012), which makes use of the same detection and classification pipeline as XXL, presents an apparent straightforward application of the model. X-CLASS however differs from XXL in that it does not make use of dedicated observations instead serendipitously identifying clusters from the outer regions of targeted XMM Newton observations. Despite all observational data used by X-CLASS being truncated to similar depths (10ks and 20ks) to reduce in homogeneity in the data there will still exist large variations between observations compared

to the closely related XXL observations. Given the issues presented by the unresolved differences between the North and South fields (described in section 5.3) differences along the line of sight between widely separated pointings present a potential barrier to applying the GP to X-CLASS. The X-ray source catalogue derived from the eROSITA all sky survey (Predehl et al., 2021) may provide a more suitable data set 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 is much deeper at the poles of the satellite's orbit compared to that at the equator. It is possible that the GP may perform better if the data were split into regions based on the depth, absorbing column and/or Galactic foreground.

While the sample of galaxy cluster candidates produced from the XXL X-ray source catalogue is of reasonable size for visual inspection the size of the sample produced for larger surveys, such as X-CLASS and eROSITA may be prohibitive. One could simply choose to inspect those sources with the highest confidence values, but at the highest confidence values the majority of those sources selected by the GP are also selected by the existing C1 and C2 samples (4.11). Those sources assigned a high confidence value but not selected by the C1 or C2 samples are predominantly not clusters. The result is that at the highest confidence values the existing selection criteria is preferable. The solution to the issue of the large number of sources needing inspection is to better filter out non-galaxy clusters prior to visual inspection. One approach is to apply a form of post processing to filter non-clusters from the GP selected sample. The choice of filter can be based on the understanding of the selected sample gained when investigating the GP's selection criteria. For example, when visually inspecting the GP selected sample of cluster candidates it was clear that AGN tend to have a high central count rate compared to clusters, the inverse being true for spurious detections of background emission when compared to clusters. A fraction of the unwanted AGN could hence be automatically filtered by removing those selected sources with a central X-ray count rate above some threshold. Repeating this for sources with a central count rate below some threshold could also be used to remove unwanted spurious detections. The low number of visually inspected and hence labelled sources does however limit our ability to reliably filter only non-cluster sources in this way.

Alternatively, the GP selection could be improved by providing it with additional information about each source. Due to the adaptability of the GP (and ML models in general) it is relatively trivial to provide the model with additional information in the form of measured source properties. The only criteria for providing the GP with additional information is imposed by the use of a Gaussian kernel which requires as input a vector of constant size for all sources. With this in mind one could supply the model with additional information measured by XAMIN from hard band observations. The hope in providing the GP with hard band information is that it learns to distinguish AGN from clusters due to differences in their hardness ratios.

When providing additional information to the GP we are not limited to information from X-ray observations. For example, in addition to a source's measured X-ray properties one could supply the GP model with, the number density of optically detected galaxies within different radii of the X-ray detection, measurements of the source's Sunyaev-Zeldovich properties and/or the measurements of the weak lensing signal. This approach does however require overlapping multi-wavelength observations for all sources. All-sky (or near all-sky) surveys such as those carried out by eROSITA (Predehl et al., 2021), the Vera Rubin Observatory (Ivezić et al., 2019) and Euclid (Laureijs et al., 2011), are ideal for this purpose. The limitation in providing additional information to the GP is the increased computation time, both when training and sampling the GP. In particular, the use of a Gaussian kernel in this work requires the calculation of the difference in properties between all sources, an increasingly computationally expensive calculation the more source properties used. While it may be necessary to only provide those measured source properties deemed most informative to the model to reduce computation, doing so may remove useful but unknown information and trends that the GP could have identified and used for selection.

Providing the GP with additional information is not limited to adding more measured source properties, one can instead choose to provide additional information through the uncertainties on the training labels. For example, XXL X-ray sources within 15 arcseconds of a `CAMIRA` cluster could be assigned an increased initial probability of being a galaxy cluster reflecting the additional information from the `CAMIRA` catalogue, i.e a C2 source within 15 arcseconds of a `CAMIRA` cluster could be assigned an initial probability of 0.8 instead of the current 0.5 (we note that the value of 0.8 is used here as an example not a suggested value). The inclusion of information from the `CAMIRA` catalogue during the training process does however mean it can no longer be treated as an independent test of the output of the GP. In addition, those XXL sources with existing spectroscopic confirmation as to being a galaxy cluster could be assigned an initial probability of one.

In addition to not being limited to X-ray surveys the GP is applicable to the selection of sources beyond galaxy clusters. For the approach used within this work to be applicable to the selection of a given type of source from a catalogue there are two requirements. First, there must exist some method of assigning each source an initial probability of being the targeted type, such as source samples of known purity. Second, each source needs to be described by a consistent set of measured source properties that can be presented as a vector. Where it is not possible or inefficient to provide information to the GP in the form of a vector it is necessary to make use of a different kernel, one that is able to take as input the available type of information. This is of course no guarantee that the GP, or any other ML method, will find a better solution than existing methods. It is worth noting that, despite the discussion around ML focusing on its applicability to current and near future surveys, the GP based approach is also applicable to existing surveys providing astronomers the opportunity to exploit previously unused information to identify and select sources that, despite being detected were not correctly identified.

6.3 Conclusion

This work has presented a supervised Gaussian Process binary classifier based approach to the selection of sources from a catalogue that is able to account for uncertainty on the training labels. The model has been applied to the selection of galaxy clusters from the XXL X-ray source catalogue (Pierre et al., 2016) producing a sample of 623 sources selected from XXL's North observing field that the GP considers more likely to be galaxy clusters. The sample recovers the vast majority of the previous C1 and C2 samples in addition to previously unselected sources. Following visual inspection the sample was found to have a purity of $0.45^{+0.03}_{-0.03}$ meaning we expect 280 of the selected sources would be considered a galaxy cluster candidate following visual inspection. Of the expected 280 galaxy cluster candidates, 179 are expected to be selected by the C1 and C2 samples, leaving an expected 101 novel cluster candidates. Looking forward it is clear that this method is applicable to a wide range of surveys for different types of sources, and is well placed to exploit the expected availability of large area multi wavelength data from current and near future surveys for source selection.

Bibliography

Abell G. O., 1958, ApJS, 3, 211

Adami C., et al., 2018, A&A, 620, A5

Aihara H., et al., 2018a, PASJ, 70, S4

Aihara H., et al., 2018b, PASJ, 70, S4

Aihara H., et al., 2022, Publications of the Astronomical Society of Japan, 74, 247

Allen S. W., Rapetti D. A., Schmidt R. W., Ebeling H., Morris R. G., Fabian A. C., 2008, MNRAS, 383, 879

Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409

Angora G., et al., 2020, A&A, 643, A177

Angulo R. E., Hahn O., 2022, Living Reviews in Computational Astrophysics, 8, 1

Baron D., 2019, arXiv e-prints, p. arXiv:1904.07248

Bartelmann M., Schneider P., 2001a, Phys. Rep., 340, 291

Bartelmann M., Schneider P., 2001b, Phys. Rep., 340, 291

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Bhargava S., et al., 2023, arXiv e-prints, p. arXiv:2301.11196

Birkinshaw M., 1999, Phys. Rep., 310, 97

Bishop C. M., Nasrabadi N. M., 2006, Pattern recognition and machine learning. Vol. 4, Springer

Bleem L. E., et al., 2015, ApJS, 216, 27

Böhringer H., Werner N., 2010, A&A Rev., 18, 127

Bolzonella M., Miralles J. M., Pelló R., 2000, A&A, 363, 476

Bonamente M., Joy M. K., LaRoque S. J., Carlstrom J. E., Reese E. D., Dawson K. S., 2006, ApJ, 647, 25

Boone K., 2019, AJ, 158, 257

Bower R. G., Lucey J. R., Ellis R. S., 1992, MNRAS, 254, 601

Burenin R. A., Vikhlinin A., Hornstrup A., Ebeling H., Quintana H., Mescheryakov A., 2007a, ApJS, 172, 561

Burenin R. A., Vikhlinin A., Hornstrup A., Ebeling H., Quintana H., Mescheryakov A., 2007b, ApJS, 172, 561

Burke D. J., Collins C. A., Sharples R. M., Romer A. K., Nichol R. C., 2003, Monthly Notices of the Royal Astronomical Society, 341, 1093

Carlstrom J. E., Holder G. P., Reese E. D., 2002, ARA&A, 40, 643

Cavaliere A., Fusco-Femiano R., 1976, A&A, 49, 137

Chapelle O., Scholkopf B., Zien A., 2010, Semi-Supervised Learning. Adaptive Computation and Machine Learning series, MIT Press, `https://books.google.co.uk/books?id=A3ISEAAAQBAJ`

Clerc N., Finoguenov A., 2022, arXiv e-prints, p. arXiv:2203.11906

Clerc N., Sadibekova T., Pierre M., Pacaud F., Le Fèvre J.-P., Adami C., Altieri B., Valtchanov I., 2012, Monthly Notices of the Royal Astronomical Society, 423, 3561

Cohn J. D., Battaglia N., 2020, MNRAS, 491, 1575

Dada E. G., Bassi J. S., Chiroma H., Abdulhamid S. M., Adetunmbi A. O., Ajibuwa O. E., 2019, Heliyon, 5, e01802

De Lucia G., Blaizot J., 2007, MNRAS, 375, 2

Deo R. C., 2015, Circulation, 132, 1920

Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441

Djorgovski S. G., Mahabal A. A., Graham M. J., Polsterer K., Krone-Martins A., 2022, arXiv e-prints, p. arXiv:2212.01493

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

Ebeling H., Edge A. C., Allen S. W., Crawford C. S., Fabian A. C., Huchra J. P., 2000, MNRAS, 318, 333

Eckert D., et al., 2016, A&A, 592, A12

Ettori S., Donnarumma A., Pointecouteau E., Reiprich T. H., Giodini S., Lovisari L., Schmidt R. W., 2013, Space Sci. Rev., 177, 119

Faccioli L., et al., 2018, A&A, 620, A9

Ferragamo A., et al., 2023, MNRAS, 520, 4000

Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, AJ, 154, 220

GPy since 2012, GPy: A Gaussian process framework in python, `http://github.com/SheffieldML/GPy`

Garrel C., et al., 2022, A&A, 663, A3

George D., Shen H., Huerta E. A., 2018, Phys. Rev. D, 97, 101501

Gibbs M., Mackay D., 2000, IEEE Transactions on Neural Networks, 11, 1458

Giles P. A., et al., 2016, A&A, 592, A3

Gioia I. M., Henry J. P., Maccacaro T., Morris S. L., Stocke J. T., Wolter A., 1990, ApJ, 356, L35

Gladders M. D., Yee H. K. C., 2000, AJ, 120, 2148

Green S. B., Ntampaka M., Nagai D., Lovisari L., Dolag K., Eckert D., ZuHone J. A., 2019, ApJ, 884, 33

Grishin K., Mei S., Ilić S., 2023, A&A, 677, A101

Gunn J. E., Gott J. Richard I., 1972, ApJ, 176, 1

Haider Abbas M., 2019, arXiv e-prints, p. arXiv:1912.05316

Hao J., et al., 2010, ApJS, 191, 254

Hasselfield M., et al., 2013, J. Cosmology Astropart. Phys., 2013, 008

Hatfield P. W., Almosallam I. A., Jarvis M. J., Adams N., Bowler R. A. A., Gomes Z., Roberts S. J., Schreiber C., 2020, MNRAS, 498, 5498

Hennawi J. F., Spergel D. N., 2005, ApJ, 624, 59

Ho M., Rau M. M., Ntampaka M., Farahi A., Trac H., Póczos B., 2019, ApJ, 887, 25

Ivezić Ž., et al., 2019, ApJ, 873, 111

Kosiba M., et al., 2020, MNRAS, 496, 4141

Koulouridis E., et al., 2018, A&A, 620, A4

Kravtsov A. V., Borgani S., 2012, ARA&A, 50, 353

Krippendorf S., et al., 2023, arXiv e-prints, p. arXiv:2305.00016

Kumaran S., Mandal S., Bhattacharyya S., Mishra D., 2023, MNRAS, 520, 5065

Kurtz M. J., Eichhorn G., Accomazzi A., Grant C. S., Murray S. S., Watson J. M., 2000, A&AS, 143, 41

Lahav O., 2023, arXiv e-prints, p. arXiv:2302.04324

Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193

Le Brun A. M. C., McCarthy I. G., Schaye J., Ponman T. J., 2014, MNRAS, 441, 1270

Lieu M., et al., 2016, A&A, 592, A4

Logan C. H. A., et al., 2018, A&A, 620, A18

Lovisari L., Maughan B. J., 2022, in , Handbook of X-ray and Gamma-ray Astrophysics. Edited by Cosimo Bambi and Andrea Santangelo. p. 65, doi:10.1007/978-981-16-4544-0˙118-1

Lumsden S. L., Nichol R. C., Collins C. A., Guzzo L., 1992, MNRAS, 258, 1

Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133

Markevitch M., Vikhlinin A., 2007, Phys. Rep., 443, 1

Menci N., Cavaliere A., 2000, MNRAS, 311, 50

Migkas K., Pacaud F., Schellenberger G., Erler J., Nguyen-Dang N. T., Reiprich T. H., Ramos-Ceja M. E., Lovisari L., 2021, A&A, 649, A151

Minka T. P., 2001, PhD thesis, Massachusetts Institute of Technology

Miyazaki S., et al., 2018, PASJ, 70, S1

Morales-Álvarez P., Pérez-Suay A., Molina R., Camps-Valls G., 2018, IEEE Transactions on Geoscience and Remote Sensing, 56, 1103

Moriwaki K., Nishimichi T., Yoshida N., 2023, arXiv e-prints, p. arXiv:2303.15794

Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563

Nord B., Stanek R., Rasia E., Evrard A. E., 2008, MNRAS, 383, L10

Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Póczos B., Schneider J., 2015, ApJ, 803, 50

Ntampaka M., et al., 2019a, BAAS, 51, 14

Ntampaka M., et al., 2019b, ApJ, 876, 82

Ntampaka M., et al., 2019c, ApJ, 876, 82

Oguri M., 2014, MNRAS, 444, 147

Oguri M., et al., 2018, PASJ, 70, S20

Opper M., Winther O., 2000, Neural Computation, 12, 2655

Pacaud F., et al., 2006, MNRAS, 372, 578

Pacaud F., et al., 2016, A&A, 592, A2

Pacaud F., et al., 2018, A&A, 620, A10

Pierre M., et al., 2016, A&A, 592, A1

Planck Collaboration et al., 2014, A&A, 571, A29

Pratt G. W., Arnaud M., Biviano A., Eckert D., Ettori S., Nagai D., Okabe N., Reiprich T. H., 2019, Space Sci. Rev., 215, 25

Predehl P., et al., 2021, A&A, 647, A1

Read A. M., Rosen S. R., Saxton R. D., Ramirez J., 2011, A&A, 534, A34

Reis I., Poznanski D., Baron D., Zasowski G., Shahaf S., 2018, MNRAS, 476, 2117

Richards J. W., et al., 2011, ApJ, 733, 10

Romer A. K., et al., 2000, ApJS, 126, 209

Romer A. K., Viana P. T. P., Liddle A. R., Mann R. G., 2001, ApJ, 547, 594

Rybicki G. B., Lightman A. P., 1986, Radiative Processes in Astrophysics

Rykoff E. S., et al., 2014, ApJ, 785, 104

Ryu D., Kang H., Hallman E., Jones T. W., 2003, ApJ, 593, 599

Sánchez Almeida J., Aguerri J. A. L., Muñoz-Tuñón C., de Vicente A., 2010, ApJ, 714, 487

Scharf C. A., Ebeling H., Perlman E., Malkan M., Wegner G., 1997, ApJ, 477, 79

Schaye J., et al., 2015, MNRAS, 446, 521

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Schneider P. C., Freund S., Czesla S., Robrade J., Salvato M., Schmitt J. H. M. M., 2022, A&A, 661, A6

Shectman S. A., 1985, ApJS, 57, 77

Springel V., Di Matteo T., Hernquist L., 2005a, MNRAS, 361, 776

Springel V., et al., 2005b, Nature, 435, 629

Starck J. L., Pierre M., 1998, A&AS, 128, 397

Sunyaev R. A., Zeldovich Y. B., 1972, Comments on Astrophysics and Space Physics, 4, 173

Tucker W. H., Tananbaum H., Remillard R. A., 1995, ApJ, 444, 532

Tulin S., Yu H.-B., 2018, Phys. Rep., 730, 1

Upsdell E. W., et al., 2023, MNRAS, 522, 5267

Valtchanov I., Pierre M., Gastaud R., 2001, A&A, 370, 689

Vikhlinin A., McNamara B. R., Forman W., Jones C., Quintana H., Hornstrup A., 1998, ApJ, 502, 558

Vikhlinin A. A., Kravtsov A. V., Markevich M. L., Sunyaev R. A., Churazov E. M., 2014, Physics Uspekhi, 57, 317

White M., van Waerbeke L., Mackey J., 2002, ApJ, 575, 640

Williams C. K., Rasmussen C. E., 2006, Gaussian processes for machine learning. Vol. 2, MIT press Cambridge, MA

Willis J. P., et al., 2021, MNRAS, 503, 5624

Wittman D. M., et al., 2002, in Tyson J. A., Wolff S., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4836, Survey and Other Telescope Technologies and Discoveries. pp 73–82 (arXiv:astro-ph/0210118), doi:10.1117/12.457348

Wittman D., Dell'Antonio I. P., Hughes J. P., Margoniner V. E., Tyson J. A., Cohen J. G., Norman D., 2006, ApJ, 643, 128

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

de Beurs Z. L., Islam N., Gopalan G., Vrtilek S. D., 2022, ApJ, 933, 116

van den Bosch F. C., Aquino D., Yang X., Mo H. J., Pasquali A., McIntosh D. H., Weinmann S. M., Kang X., 2008, MNRAS, 387, 79