Characterising the optical properties of galaxy clusters with GMPhoRCC

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **469,** 3851–3871 (2017)
Advance Access publication 2017 April 21

doi:10.1093/mnras/stx940

Characterizing the optical properties of galaxy clusters with GMPhoRCC

R. J. Hood★ and R. G. Mann★

Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

Accepted 2017 April 18. Received 2017 April 14; in original form 2016 November 23

ABSTRACT

We introduce the Gaussian Mixture full Photometric Red sequence Cluster Characteriser (GMPhoRCC), an algorithm for determining the redshift and richness of a galaxy cluster candidate. By using data from a multiband sky survey with photometric redshifts, a red sequence colour–magnitude relation (CMR) is isolated and modelled and used to characterize the optical properties of the candidate. GMPhoRCC provides significant advantages over existing methods, including treatment of multimodal distributions, variable width full CMR red sequence, richness extrapolation and quality control in order to algorithmically identify catastrophic failures. We present redshift comparisons for clusters from the GMBCG, NORAS, REFLEX and XMM Cluster Survey catalogues, where the GMPhoRCC estimates are in excellent agreement with spectra, showing accurate, unbiased results with low scatter ($\sigma_{\delta z/(1+z)} \sim 0.017$). We conclude with the evaluation of GMPhoRCC performance using empirical Sloan Digital Sky Survey (SDSS) like mock galaxy clusters. GMPhoRCC is shown to produce highly pure characterizations with very low probabilities (<1 per cent) of spurious, clean characterizations. In addition, GMPhoRCC is shown to demonstrate high rates of completeness with respect to recovering redshift, richness and correctly identifying the brightest cluster galaxy (BCG).

Key words: galaxies: clusters: general – galaxies: distances and redshifts.

1 INTRODUCTION

Galaxy clusters are excellent probes of cosmology; as the largest observable objects, these are great indicators of the large-scale structure and evolution of mass distribution in the universe. As this is highly sensitive to the form of the expansion of the universe, their study gives valuable constraints on cosmological models (see Peebles 1980; Jenkins et al. 2001; Sheth, Mo & Tormen 2001; Rozo et al. 2010; Allen, Evrard & Mantz 2011; Tinker et al. 2012, etc.). In addition, clusters provide an excellent opportunity for studying galaxies themselves, particularly formation, evolution and the impact of the environment (see Gladders et al. 1998; Voit 2005, etc.). With the recent surge in cluster detections, from the Sunyaev–Zel'dovich signal in the cosmic microwave background (Reichardt et al. 2013; Planck Collaboration XXIX 2014, etc.), X-ray emission of the intracluster medium (Lloyd-Davies et al. 2011; Clerc et al. 2012, etc.), and spatial and optical cluster finding (Hao et al. 2010; Murphy et al. 2012; Rykoff et al. 2014, etc.), galaxy clusters are proving to be an ever more valuable area of research.

While the most useful cosmological analysis of galaxy clusters involves the study of mass and redshift, these are difficult and time-consuming to determine directly, requiring gravitational lensing and spectroscopy. Optical characterization offers quick estimates of cluster properties using multiband optical photometry alone, and with the abundance of the such data from the Sloan Digital Sky Survey (SDSS; Ahn et al. 2014), Canada–France–Hawaii Telescope Lensing Survey (Heymans et al. 2012), VLT Survey Telescope ATLAS (Shanks & Metcalfe 2012; Shanks et al. 2015) and the Panoramic Survey Telescope and Rapid Response System 3π survey (Magnier et al. 2013), there is significant scope for such analysis.

The main focus of this research is the development of a new characterization algorithm, the Gaussian Mixture full Photometric Red sequence Cluster Characteriser (GMPhoRCC), which aims to provide optical characterization of potential clusters previously detected by other observations such as X-ray emission. While the specific motivation for this lies with the determination of cluster redshifts for forthcoming XMM Cluster Survey (XCS; Romer et al. 2001; Mehrtens et al. 2012, etc.) data releases, GMPhoRCC is designed for general use, providing characterizations for any list of positions of cluster candidates and any multicolour galaxy catalogue with photometric redshifts.

This paper is structured as follows: Section 2 discusses existing characterization methods focusing on the motivation and key features desired in a new robust algorithm. Section 3 explores the details of the GMPhoRCC with evaluation using comparisons to known and mock clusters, followed by following in Section 4, with a detailed investigation of the purity, completeness and effectiveness of the quality control system in Section 4. Finally, this paper concludes with a summary and discussion in Section 5. This paper assumes a flat Λ cold dark matter cosmology with $\Omega_m = 0.27$, $\Omega_\lambda = 0.73$ and $h = 0.71$.

★ E-mail: rjh@roe.ac.uk (RJH); rgm@roe.ac.uk (RGM)

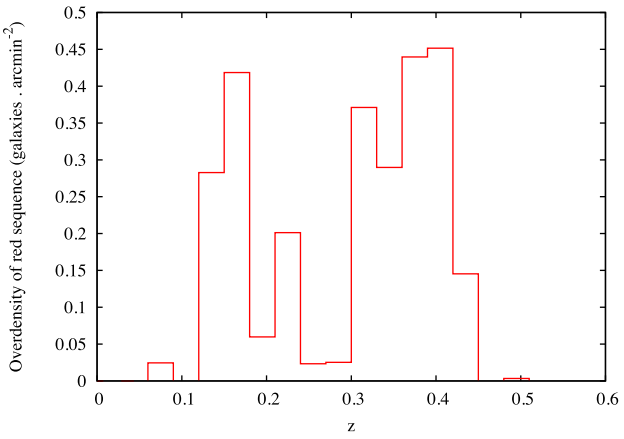

Figure 1. A histogram of the red-sequence overdensity as a function of redshift showing a multimodal distribution resulting from projection effects. These peaks correspond to spatially overlapping clusters at different redshifts, and without any additional information, it is difficult to determine which represents the target cluster.

2 CHARACTERIZING THE OPTICAL PROPERTIES OF GALAXY CLUSTERS

Many cluster detection/classification algorithms have been developed in recent years: C4 (Miller et al. 2005), maxBCG (Koester et al. 2007), GMBCG (Hao et al. 2010) and redMaPPer (Rykoff et al. 2014), to name a few. While the exact details vary, the basis of these methods is to isolate the red-sequence, early-type galaxies with similar metallicities and colours that dominate the members, and use these to infer the bulk properties of the cluster. While optical cluster finders search for additional spatial clustering, the simplest form of red-sequence modelling is to find clustering in colour space.

The simplest case of characterization relates to a well-defined easily observed red sequence as an overdensity in colour or redshift space; however, many clusters do not conform to this due to projection effects and background fluctuations. Fig. 1 demonstrates the redshift clustering of a field with two overlapping clusters. Due to the projection effect, it is unclear which peak represents the target cluster, and methods looking for maximum overdensities such as that from High et al. (2010) may fail to adequately describe the situation. While the overlapping of significant overdensities presented in Fig. 1 is rare, multimodal distributions are noted to be common, found to be present in ∼40 per cent of the GMBCG catalogue (Hood 2014); it is clear that any standalone characterization algorithm must account for these potentially ambiguous cases.

While a simple colour overdensity is a good approximation, the red sequence itself is described by a colour–magnitude relation (CMR). The CMR relates to the physical properties of the cluster, where slope encodes the mass–metallicity relation and scatter, the age variation, etc. (Gladders et al. 1998), and hence it is desirable to model the full CMR rather than simple overdensities, as is the case with GMBCG, XCS, etc.

With the red sequence isolated, it remains to determine cluster properties. Redshift can be determined using red-sequence colour–redshift models, as in the cases of Koester et al. (2007) and Mehrtens et al. (2012); however, this introduces model dependence and additional complexity in the analysis. In the case of GMBCG, redshift is obtained from the photometric estimate of the brightest cluster galaxy (BCG). Although relying on correct identification, as a bright galaxy, the photometric estimate is more easily obtained.

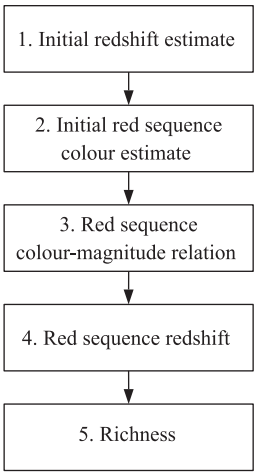

Figure 2. A flowchart summarizing the basic steps of GMPhoRCC to determine the red-sequence CMR, cluster redshift and richness.

The following lists several initial features drawn from existing algorithms that have driven the development of GMPhoRCC:

(i) Red-sequence detection: GMPhoRCC will isolate the red sequence and use these galaxies to infer the optical properties of the cluster.

(ii) Photometric redshifts: There is no assumed colour–redshift model.

(iii) Full red-sequence CMR: The red sequence will be described by a full CMR determined by the GMPhoRCC.

(iv) Multiple red-sequence bands: To maximize the efficiency of the GMPhoRCC and to cover a large range of cluster redshifts, multiple redshift-dependent colour bands will be necessary, as demonstrated in Hao et al. (2010).

(v) Multimodal distributions: Without resorting to a full finder approach where overlapping clusters can be separately identified and analysed, GMPhoRCC must deal with multimodal distributions as several potential clusters.

(vi) Quality control: Extending beyond simple error analysis, it is necessary to assess the probability of catastrophic failure. By introducing quality control, several subsets can be produced where problem clusters and possible outliers can be identified and removed to produce a clean subset.

3 GMPHORCC

GMPhoRCC is designed as an optical follow-up tool to confirm and characterize galaxy cluster candidates. The main feature of GMPhoRCC is to identify the red sequence and use the properties of these galaxies to analyse the cluster. A basic outline of the procedure used to isolate the red sequence and to ultimately determine redshift and richness is shown in Fig. 2. While each of these steps is explored in detail in subsequent sections, it is first noted that many of these require the modelling of cluster distributions such as colour and photometric redshifts, which, as shown by Hao et al. (2009), can be approximated by a Gaussian mixture.

3.1 Modelling cluster distributions with error-corrected Gaussian mixtures

Galaxy distributions, whether colour or redshift, are well modelled by Gaussian mixtures that use a sum of several Gaussian components to describe any features or substructures. Fitting the mixture

model proceeds using the error-corrected expectation maximization procedure from Hao et al. (2009), which accounts for associated measurement errors. Considering the distribution of galaxy colours as an example, the probability of galaxy n having true colour $\tilde{c}_n$, given the parameters θ, is defined as

$$p(\tilde{c}_n \mid \theta) = \sum_k N(\tilde{c}_n \mid \mu_k, \sigma_k) P(k), \quad (1)$$

$$N(\tilde{c}_n \mid \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp - \frac{(\tilde{c}_n - \mu_k)^2}{2\sigma_k^2}, \quad (2)$$

where θ represents the collective parameters of the model, namely $\mu_k, \sigma_k, P(k)$, the component means, standard deviations and weights, respectively. Combining these with Gaussian measurement errors for all the galaxies, Hao et al. (2009) have shown that this leads to the following form for the likelihood of the parameters, given the data:

$$p(\theta \mid c_n) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \frac{P(k)}{\sqrt{2\pi(\sigma_k^2 + \delta_n^2)}} \exp \left[- \frac{(c_n - \mu_k)^2}{2(\sigma_k^2 + \delta_n^2)} \right] \right), \quad (3)$$

where c_n is the observed galaxy colour with Gaussian error δ_n, K is the total number of components and N is the number of galaxies. Maximizing this likelihood using the expectation maximization procedure of Hao et al. (2009) gives the optimized parameters.

In order to apply this modelling to the distributions of cluster members, it is necessary to account for the effect of the background when analysing a candidate field. Rather than modelling the whole field and selecting components to separately describe the cluster and background as explored by the GMBCG algorithm (Hao et al. 2010), a subtraction approach is used:

$$\mathrm{GM}_{\mathrm{cluster}} = \mathrm{GM}_{\mathrm{field}} - \mathrm{GM}_{\mathrm{background}}, \quad (4)$$

where GM represents a Gaussian mixture density model. Modelling clusters in this manner with background subtraction helps to remove the ambiguity in component selection seen in GMBCG. Values of interest, such as an approximate red-sequence colour from the colour distribution, are determined with an associated uncertainty from the width of the distributions around the peaks. While this section describes in detail the Gaussian mixture fitting procedure for a given area, the selection of the cluster area, background and field is explored in Section 3.2.

Applying this to the colour distribution of cluster GMBCG J197.87292−01.34109 indeed shows the Gaussian mixture to be a good representation of the cluster around a $g - r$ colour of ∼1.2 mag as shown in Fig. 3.

3.2 Red-sequence CMR and redshift

With the framework in place to model the various cluster distributions from optical data, the first main goal of GMPhoRCC is to identify the red sequence, modelling a full CMR with intrinsic scatter. As the previous section described, GMPhoRCC models cluster distributions with background subtraction; hence, the first step is the extraction of the cluster region and background from the candidate field. The cluster region is taken as the cone with a 30-arcmin radius centred on the detection observation, for example, the peak of the X-ray emission. The local background is taken as the annulus around this cone up to a radius of 60 arcmin.

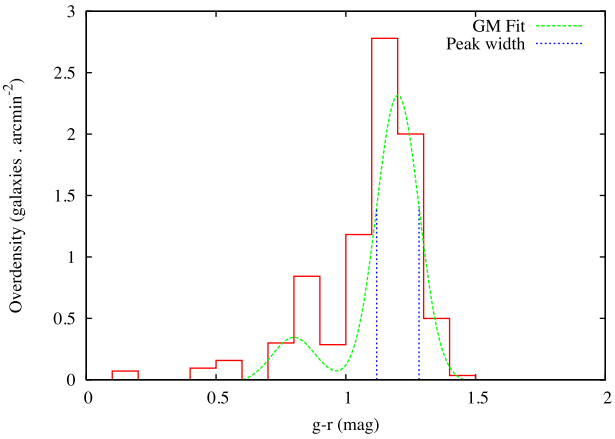

Figure 3. A histogram of the $g - r$ colour overdensity of GMBCG J197.87292−01.34109 showing the best- fitting Gaussian mixture highlighting the associated width used as an asymmetric error bar around the peak.

3.2.1 *Initial redshift estimate*

To aid the isolation of the red sequence, initial redshift and colour estimates for the cluster are determined, which allows the selection of an appropriate colour band and provides broad filtering to remove field galaxies. While this is perhaps best achieved by modelling the joint colour–redshift density distribution, extending the fitting procedure of Section 3.1 to higher dimensions is not trivial. Even without the error correction, the fitting procedure often fails to converge, fails to recover fine structure and is highly sensitive to the initial estimate of the parameters. Hence, separate, error-corrected modelling of redshift and colour proceeds.

Starting with redshift, Fig. 4 expands procedure 1 of Fig. 2 showing in detail the procedure used to arrive at an initial estimate for an inner cluster region. The redshift distribution of the inner cluster is modelled by taking the mixture model from a series of cones across a range of radii, 1–4 arcmin, and subtracting the background model. The inner cluster radius is then selected to produce the largest sum of the amplitudes of the peaks in the distribution. In addition to ensuring that peaks can still be found in the case of miscentring, this gives preference to regions producing multimodal distributions where each peak can be subsequently analysed and used to assist with the characterization. While this could be found with an investigation of the radial profile, this method is less sensitive to issues with overlapping clusters that indeed can be common (∼20 per cent of the GMBCG catalogue has a neighbour within 3 arcmin).

For a rigorous treatment of multimodal redshift distributions, a secondary peak is investigated as a potential cluster, provided the amplitude is at least 20 per cent of the primary. This threshold allows the analysis of potential structure without exploring noise or low-level fluctuations in the distribution. Considering multiple peaks in this way occurs throughout GMPhoRCC, resulting in a potentially large number of possible candidates from which the cluster is selected.

3.2.2 *Initial colour estimate*

With the initial redshift estimate, an initial colour estimate for the inner cluster region proceeds, as shown in Fig. 5. First, an appropriate colour band is selected based on the initial redshift in line with Hao et al. (2010), as shown in Table 1. These values ensure that the main spectral feature of red-sequence galaxies, the 4000 Å

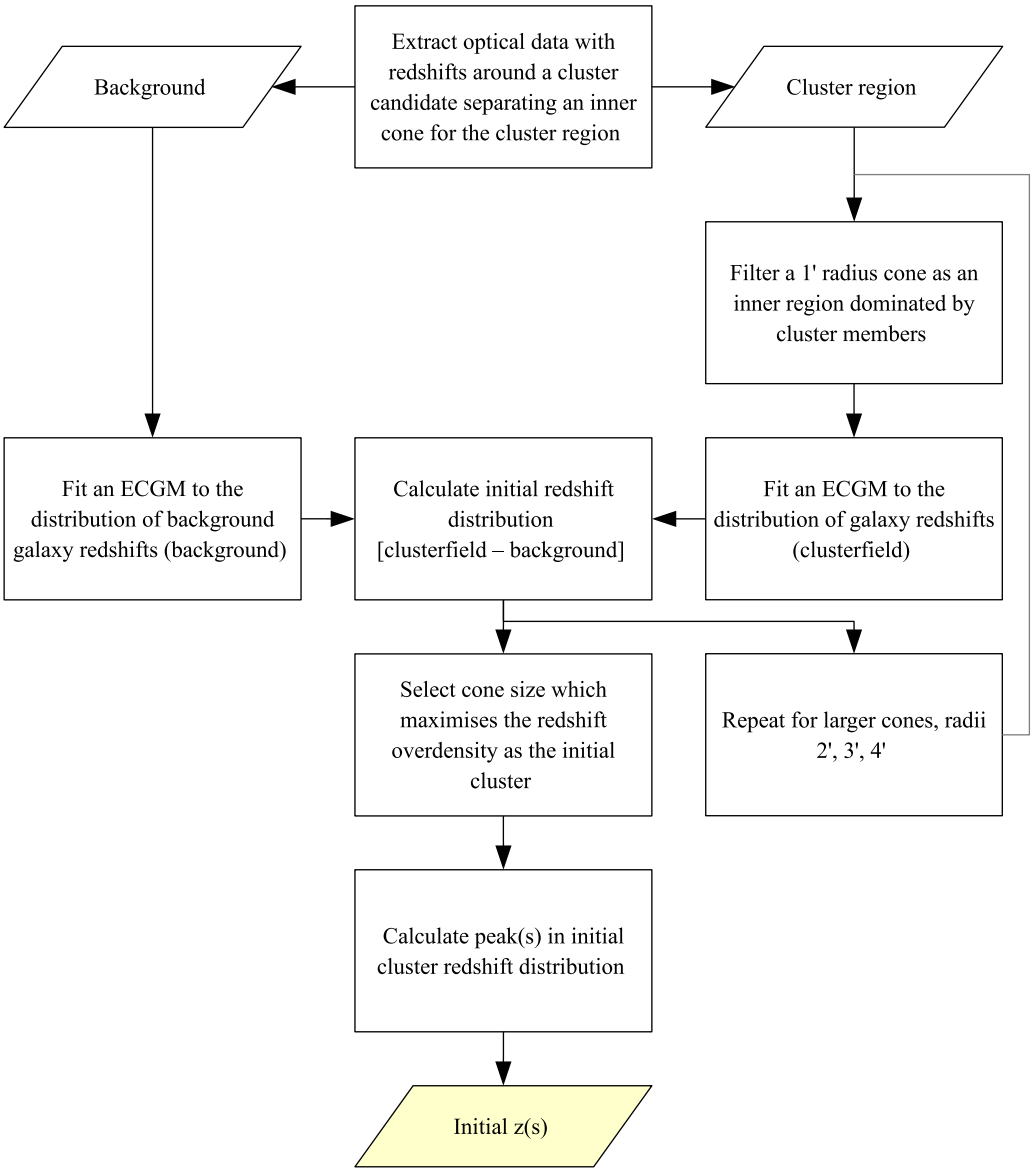

Figure 4. A detailed flowchart showing the procedures of GMPhoRCC to determine an initial cluster redshift estimate highlighted as procedure 1 in Fig. 2.

break, remains in the band at a given redshift. This is important as this ensures the strongest colour clustering and contrast against the background. Additionally, the redshift values overlap to account for possible failures of the initial estimate and help clusters where the 4000 Å break sits between bands.

Before estimating an initial colour, the field is cleaned by removing potential spurious galaxies from the background and cluster region that do not conform to the initial redshift. This is achieved by removing those where the galaxy photometric redshift lies outside a redshift window, centred on the initial estimate and those fainter than would be considered in such a cluster. Specifically for the SDSS, galaxies were removed where their redshift was more than 0.25 from the initial estimate. This large window relative to the typical photometric errors ($\sim$ 10 times greater) helps to ensure that only outliers are removed. The faint-end cut is taken as $m_*(z) + 2$, where the redshift-dependent m_* is taken from Hao et al. (2010), which was derived from the luminosity function of field galaxies. Initial red-sequence colour estimation proceeds in a similar manner to redshift, where the colour estimate is taken as the peak in the background-subtracted colour distribution of an inner cluster region. The inner cluster region is determined again by considering cones with a range of radii and selecting the cone that maximized the sum of the peak amplitudes. This region is used as the inner cluster region in subsequent analysis.

3.2.3 *Red-sequence CMR*

The red-sequence CMR is determined using the initial estimates of redshift, colour and inner cluster radius, as shown in Fig. 6. First, galaxies from the inner region are filtered in a manner similar to that used to identify red-sequence galaxies; all galaxies within 2σ of the initial colour estimate are kept for further analysis, where

$$\sigma^2 = \sigma^2_{\rm colour} + \sigma^2_{\rm RS}. \quad (5)$$

A broad initial red-sequence width is used with $\sigma_{\rm RS} = 0.1$ to ensure that only field contamination is removed.

With the red sequence dominating the remaining galaxies from the inner region, fitting a CMR proceeds using the bivariate

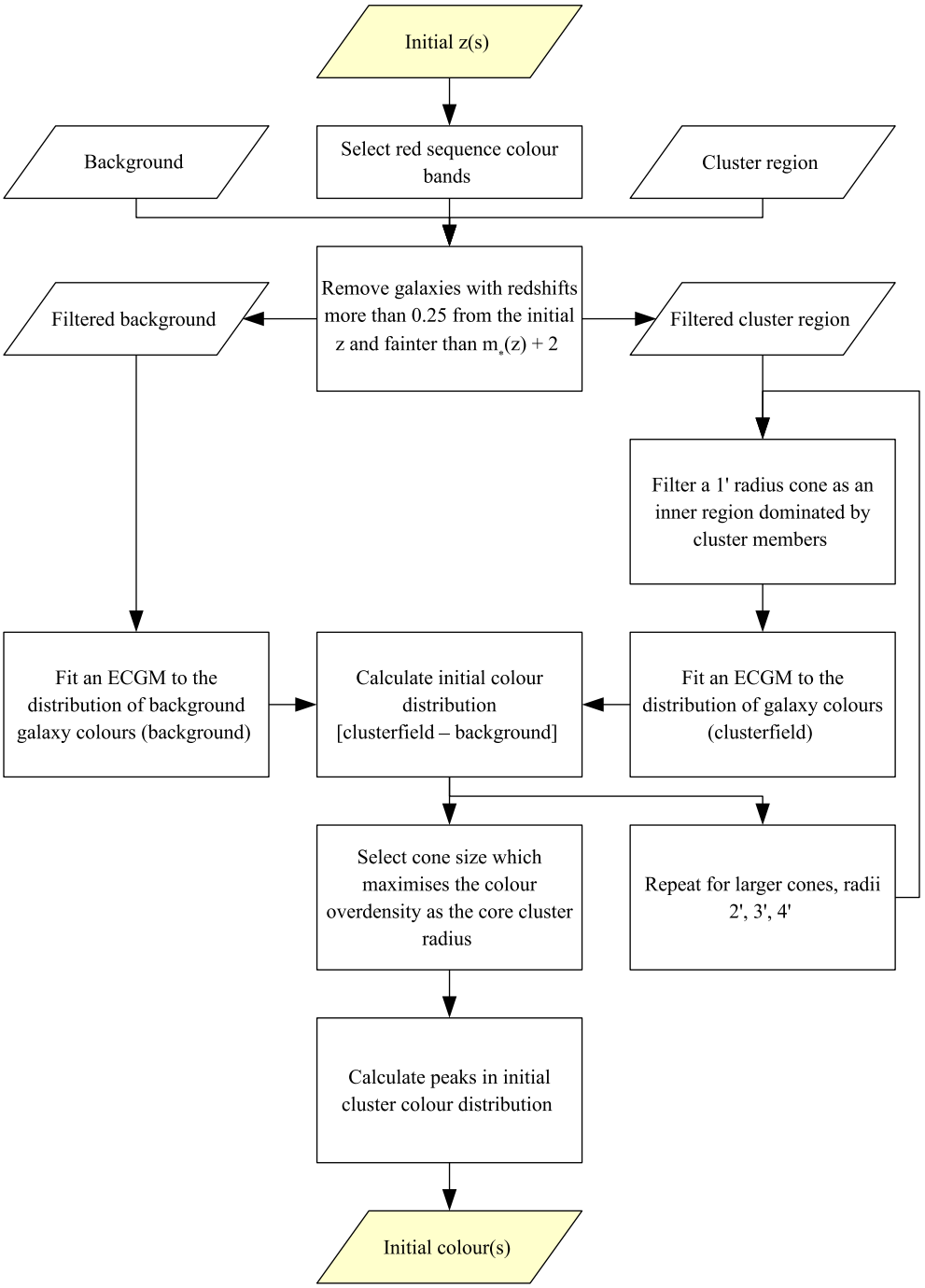

Figure 5. A detailed flowchart showing the procedures of GMPhoRCC to determine an initial red-sequence colour estimate, highlighted as procedure 2 in Fig. 2.

correlated errors and intrinsic scatter (BCES) method (Akritas & Bershady 1996). This extends the standard least-squares method to account for intrinsic scatter and potentially correlated errors in both the dependent and independent variables. Additionally, the intrinsic width of the red sequence is determined from the distribution of colours around the CMR.

In practice, the distribution of galaxies around the CMR found using BCES takes the form of a Gaussian. By correcting for the slope and using a single component, the width of the error-corrected Gaussian mixture gives a good estimate of the intrinsic scatter of the red sequence.

3.2.4 Red-sequence redshift

Isolating the red sequence provides a sample dominated by cluster members from which redshift can be derived. At this stage, several estimates for the red-sequence redshift are determined as shown below and can be subsequently used in later stages of the algorithm:

(i) Colour: derived from the properties of the CMR;
(ii) Photometric: derived from the distribution of photometric redshifts;
(iii) Spectroscopic: derived from spectroscopic red-sequence members;

Table 1. The most suitable red-sequence colour bands for the initial redshift estimate. These values overlap to account for the uncertainty in the initial redshift and those close to a transition.

Red-sequence band	Redshift range
$g-r$	$0.0 \leq z < 0.5$
$r-i$	$0.3 \leq z < 0.8$
$i-z$	$0.6 \leq z$

Table 2. A tighter redshift–band relation for the red sequence than presented in Table 1. The reduced overlap helps to remove the same candidates analysed in multiple bands.

Red-sequence band	Redshift range
$g-r$	$0.00 \leq z < 0.45$
$r-i$	$0.35 \leq z < 0.75$
$i-z$	$0.65 \leq z$

(iv) BCG: derived from the redshift of the brightest red-sequence member alone.

Selection of which estimate to use will depend primarily on the quality of the data, and while spectroscopic provides the most desirable estimate, it is not always available *en masse*.

For the remainder of this paper, unless otherwise stated, the cluster redshift estimate is determined using the photometric redshift distribution of the red-sequence galaxies. Although clusters also have blue galaxies that are not considered, using the red sequence ensures estimation on a sample dominated by cluster members where spurious contaminations have been minimized. Having isolated the red sequence from the background and inner cluster region using the determined CMR with measured intrinsic width with the 2σ approach, the peaks in the background-subtracted photometric redshift distribution provide the potential cluster redshift.

3.2.5 Candidate selection

With the possibility of multiple candidates and the frequent ambiguity of cluster selection, the final step, shown in Fig. 8, filters the results to produce a primary as the most likely cluster candidate and a secondary as the next likely possibility. To help this process, the potential clusters are first filtered to ensure that the initial redshift, red-sequence redshift and BCG redshift are all appropriate for the colour band used as shown in Table 1. Reducing the redshift overlap as depicted in Table 12 helps to remove the same candidates analysed in multiple bands. If this removes all the potential clusters, the filter is not applied and the selection process continues.

The remaining candidates are then ranked, based first on the consistency of the three main redshift estimators, initial, red sequence and BCG. Four cleanness bands are introduced, as shown in Table 3, where the most desirable candidates have the highest value.

These consistency checks help to remove candidates that may not represent clusters but rather random enhancements in the background or foreground. Finally, to further rank clusters and break degeneracy, those where the red-sequence redshift best matches the initial estimate provide the best selection, reducing the chance that the best candidates are spurious. The primary clusters are simply the cleanest candidates that best match the primary initial redshift estimate. A secondary cluster is assigned as the cleanest candidate associated with the earliest secondary peak (i.e. initial secondary redshifts considered first, etc.) that best matched the initial estimate. This selection procedure is shown in detail in Fig. 7.

While comparing the GMPhoRCC estimates to previously characterized spectroscopic clusters, it is found that, on average, 35 per cent of targets characterized have an associated secondary cluster, and of these, only 13 per cent (5 per cent of the total) better match spectra than the primary. Again, it has not only been shown that dealing with multimodal distributions is necessary, but also been shown that the selection process of GMPhoRCC is able to reliably select the most appropriate characterization from the potential candidates for the cluster.

3.3 Richness

GMPhoRCC measures richness as the number of red-sequence galaxies, defined by the 2σ filtering, within a given radius fainter than the BCG and brighter than some redshift-dependent cut-off, $m_*(z) + 1$. This takes the form of the maxBCG (Koester et al. 2007) and GMBCG (Hao et al. 2010) richness, which ensures that the red-sequence i-band magnitude range is consistent as a function of redshift. In agreement with maxBCG, GMBCG and redMaPPer (Rykoff et al. 2014), m_* is taken from the luminosity function of field galaxies determined by Blanton et al. (2003).

For consistency across a range of cluster sizes, GMPhoRCC considers n_{200}, the richness inside the characteristic radius r_{200}. As measuring r_{200} directly is possible only with gravitational lensing, an intermediate 0.5 h^{-1} Mpc fixed aperture richness, $n_{\rm gals}$, is used following the analysis of maxBCG and GMBCG, from which r_{200} is found. Using the maxBCG clusters and the weak-lensing-derived r_{200}–$n_{200}^{\rm maxBCG}$ scaling relation from Hansen et al. (2009), the following scaling relation was found by binning clusters by $n_{\rm gals}$ and fitting r_{200}. Direct derivation of this relation using weak-lensing r_{200} will reduce the scatter in this relation and is left for future work.

$$r_{200} = 0.237(n_{\rm gals})^{0.4}. \tag{6}$$

In addition to counting galaxies, richness is also estimated using the luminosity function method of High et al. (2010). This involves fitting a luminosity function within a magnitude range where the photometry is believed to be complete, and then integrating up to $m_* + 1$. Finding this range is done simply by inspecting a magnitude histogram where a limit can be assigned after which the density drops with increasing magnitude. Rather than using the binning approach to fit a Schechter function, a new probabilistic approach has been developedthat gives more reliable fits across a range of cluster richnesses. With appropriate normalization, the luminosity function (equation 7) gives the number of galaxies within the magnitude range $m \rightarrow m + {\rm d}m$, and hence the probability that a galaxy has a particular magnitude, given that the parameters of the Schechter function can be approximated by equation (8):

$$\begin{aligned}\phi(m,\theta){\rm d}m = {} & 0.4[\ln(10)]\phi_* 10^{-0.4(m-m_*)(\alpha+1)} \\ & \times \exp[-10^{-0.4(m-m_*)}]{\rm d}m,\end{aligned} \tag{7}$$

$$P(m|\theta) = \frac{\phi(m,\theta){\rm d}m}{\text{total number of galaxies}}, \tag{8}$$

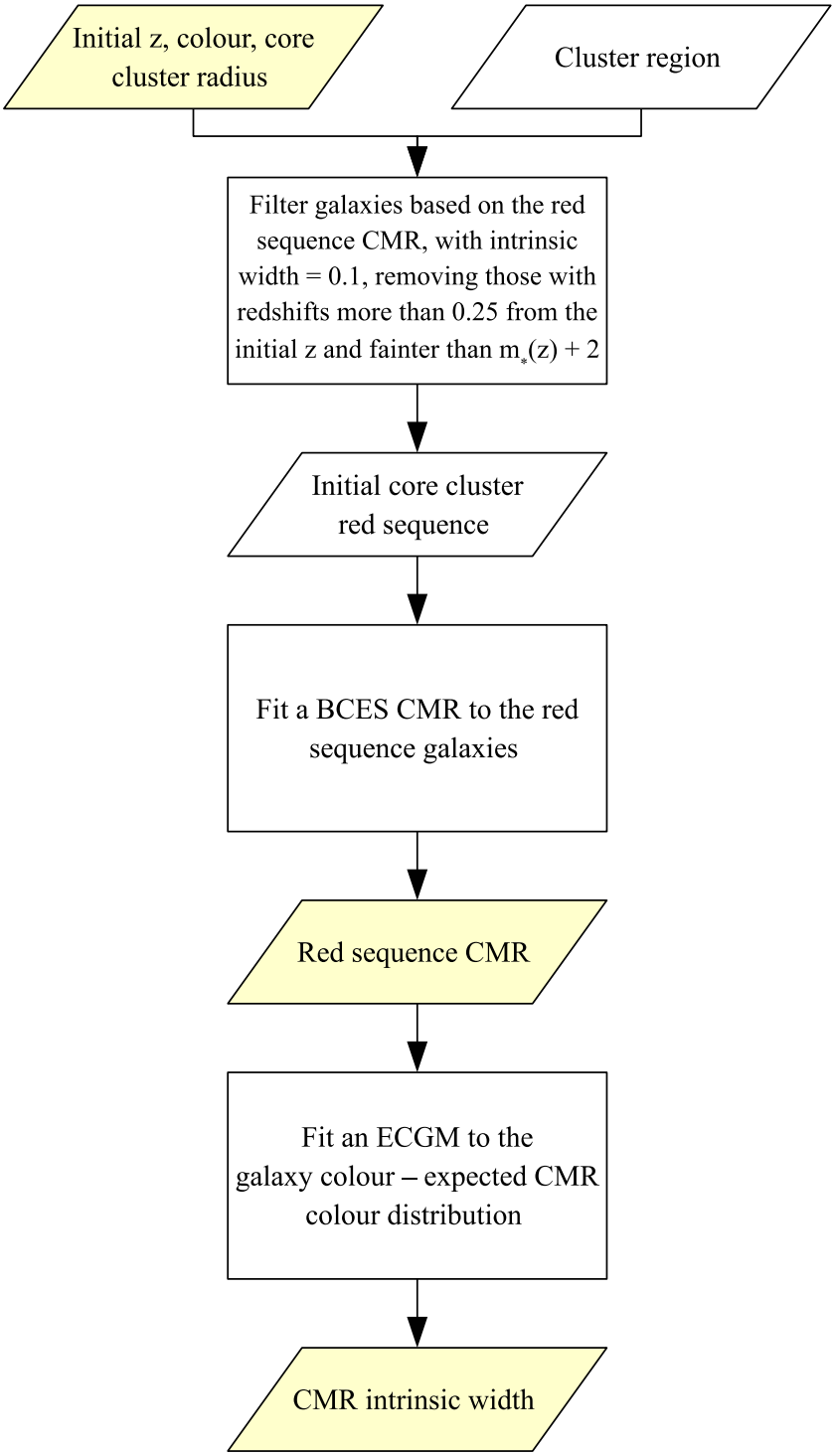

Figure 6. A detailed flowchart showing the procedures of GMPhoRCC to filter remaining contamination and model the red-sequence CMR, highlighted as procedure 3 in Fig. 2.

where θ represents the parameters of the Schechter function, namely ϕ_*, m_* and α. Using Bayes' theorem, the likelihood of the parameters, $\mathcal{L}$, is given by combining the probabilities from all galaxies:

$$\mathcal{L} = \prod_{k=1} P(\theta|m_k) = \prod_{k=1} \frac{P(m_k|\theta)P(\theta)}{P(m_k)}. \quad (9)$$

The cluster luminosity function is then defined by the parameters that maximize this likelihood. By assuming flat priors, this is equivalent to minimizing the log-likelihood shown in equation (10):

$$\ln(\mathcal{L}) \propto \sum_{k=1} \ln \phi(m_k, \theta). \quad (10)$$

Table 3. A list of the cleanness bands where the most desirable candidates have the highest band value. Redshifts are considered consistent with the colour band, as shown in Table 2.

Cleanness band	Description
1	One or more of the main redshift estimators have not been found
2	The red sequence and BCG redshift disagree by more than 0.1
3	All three redshift estimators are not consistent with the colour band
4	All remaining candidates

In addition to using the High et al. (2010) fixed faint end ($\alpha = -1$), constraining the parameters to satisfy equation (11) greatly increases the reliability of the fit:

$$\int \phi(m, \theta) \mathrm{d}m = \text{total number of galaxies.} \tag{11}$$

This ensures that a reasonable luminosity function is recovered, which, when integrated across the previously determined magnitude range, returns the total number of galaxies observed.

Minimizing equation (10) subject to the constraint shown in equation (11) proceeds using the standard sequential least-squares method described by Kraft (1988).

Although an improvement, this method still produces unreliable results for very low numbers of galaxies; hence, when fitting five or fewer data points, m_* is fixed based on the cluster redshift and the luminosity function of field galaxies determined by Blanton et al. (2003).

Combining these methods, Fig. 9 expands procedure 5 from Fig. 2 in more detail, showing the steps taken to estimate cluster richness. By using an input radius of 0.5 h^{-1} Mpc, the intermediate richness, n_{gals}, is determined, with n_{200} found by using r_{200}.

3.4 Optical data

Although designed for use with any optical data, initial calibration and development of GMPhoRCC has been driven with the use of optical data from the SDSS. The 10th data release presented in Ahn et al. (2014) provides coverage of 14 555 deg^2 in the Northern hemisphere, giving 95 per cent completeness down to 21.3 mag in the i band, giving ∼90 million suitable galaxies, ∼1.9 million with spectra.

The input optical data were selected from the Galaxy view of the PhotoObjAll table using the following query to ensure cleanness and completeness in photometry:

```
SELECT * from GALAXY
WHERE

(dered_i) < 21.0 AND

(modelMagErr_g / dered_g) < 0.1 AND
(modelMagErr_r / dered_r) < 0.1 AND
(modelMagErr_i / dered_i) < 0.1 AND

(insidemask)=0 AND
(clean)=1 AND
(extinction_i) < 0.5

(dered_g-dered_r) BETWEEN -1.5 and 5 AND
(dered_r-dered_i) BETWEEN -1.5 and 4 AND
(dered_i-dered_z) BETWEEN -1.5 and 4 AND.
```

The 10 per cent cut on colour errors with extinction and masking constraints, as used by Hao et al. (2010), ensures that the optical data are clean, which greatly improves the Gaussian mixture fitting procedure. In addition, the colour cuts and i-band constraint help to remove extreme objects with likely erroneous photometry that adversely bias the Gaussian mixture models of the cluster candidates. While the i-band cut is specific to the

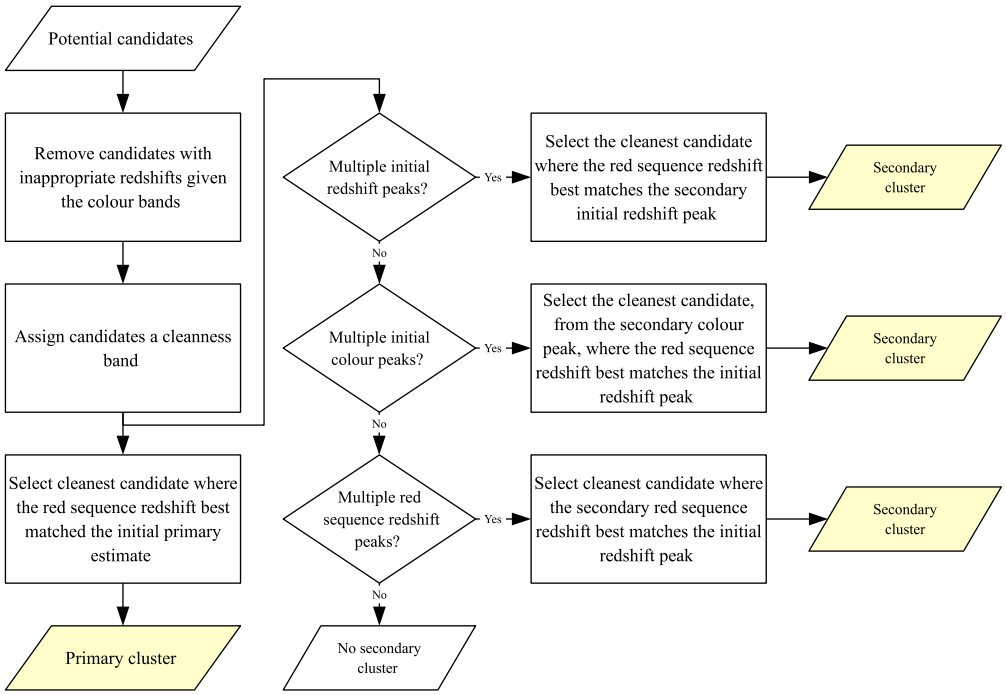

Figure 7. A detailed flowchart showing the procedures of GMPhoRCC to select a primary and a secondary cluster from a list of potential candidates.

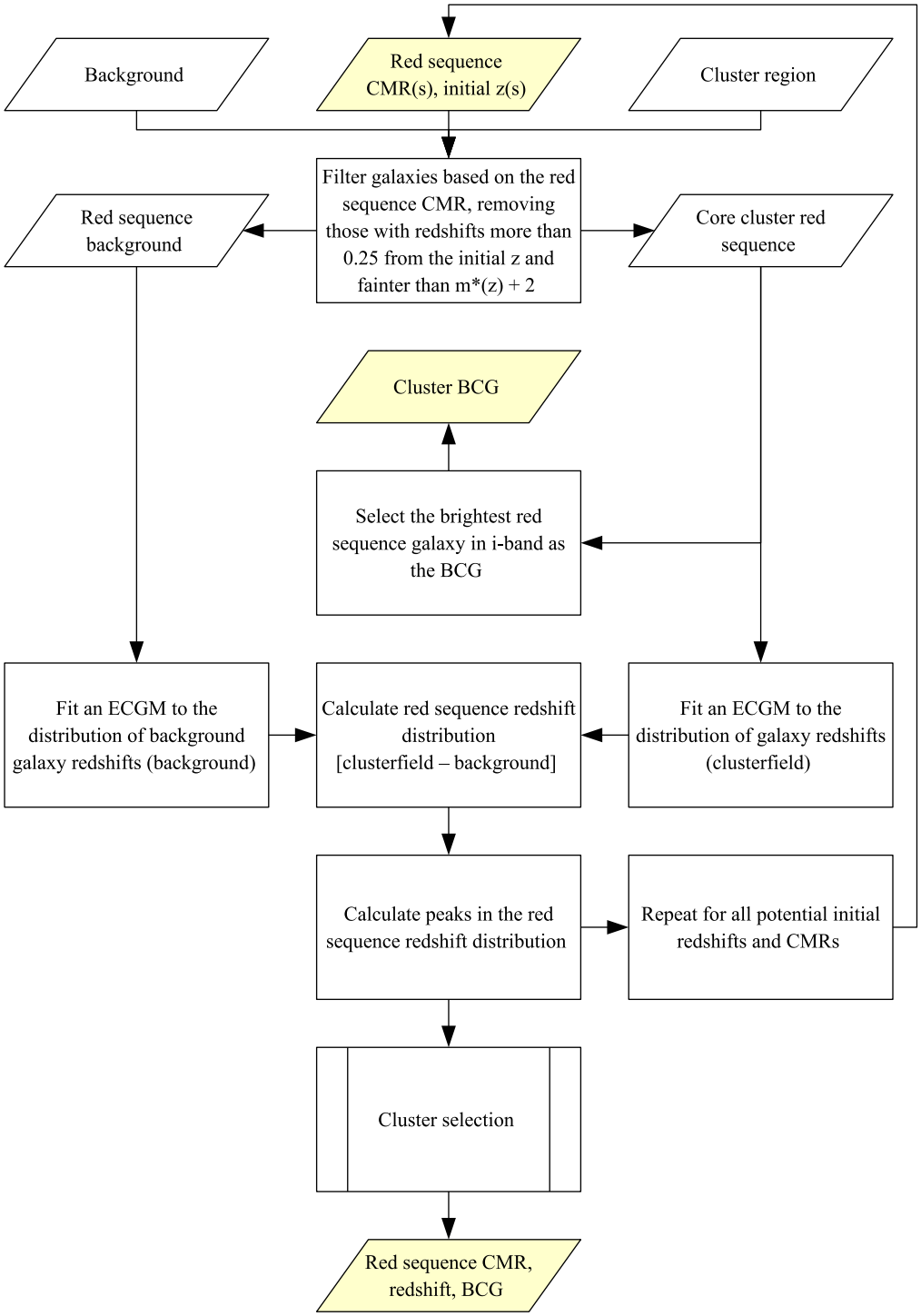

Figure 8. A detailed flowchart showing the procedures of GMPhoRCC to estimate the cluster redshift from the red-sequence CMR, highlighted as procedure 4 in Fig. 2. The cluster-selection process selects a primary and a secondary cluster, as outlined in Fig. 7.

SDSS, the colour and error constraints are recommended for GMPhoRCC, regardless of the source of the optical data, to ensure clean photometry.

In addition to multiband photometry, GMPhoRCC makes use of photometric redshifts rather than using assumed colour–redshift relations. Within the SDSS DR10, the PhotozRF table provides the most suitable redshifts, calculated using the random forest regression technique of Carliles et al. (2010). While not essential, these provide well-understood Gaussian errors that are ideal for use with the error-corrected Gaussian mixture models of Section 3.1.

3.5 Quality control

One of the main goals of GMPhoRCC is to provide a means of quality control to help identify possible catastrophic failures. As part of this, many flags have been introduced to trace how clusters

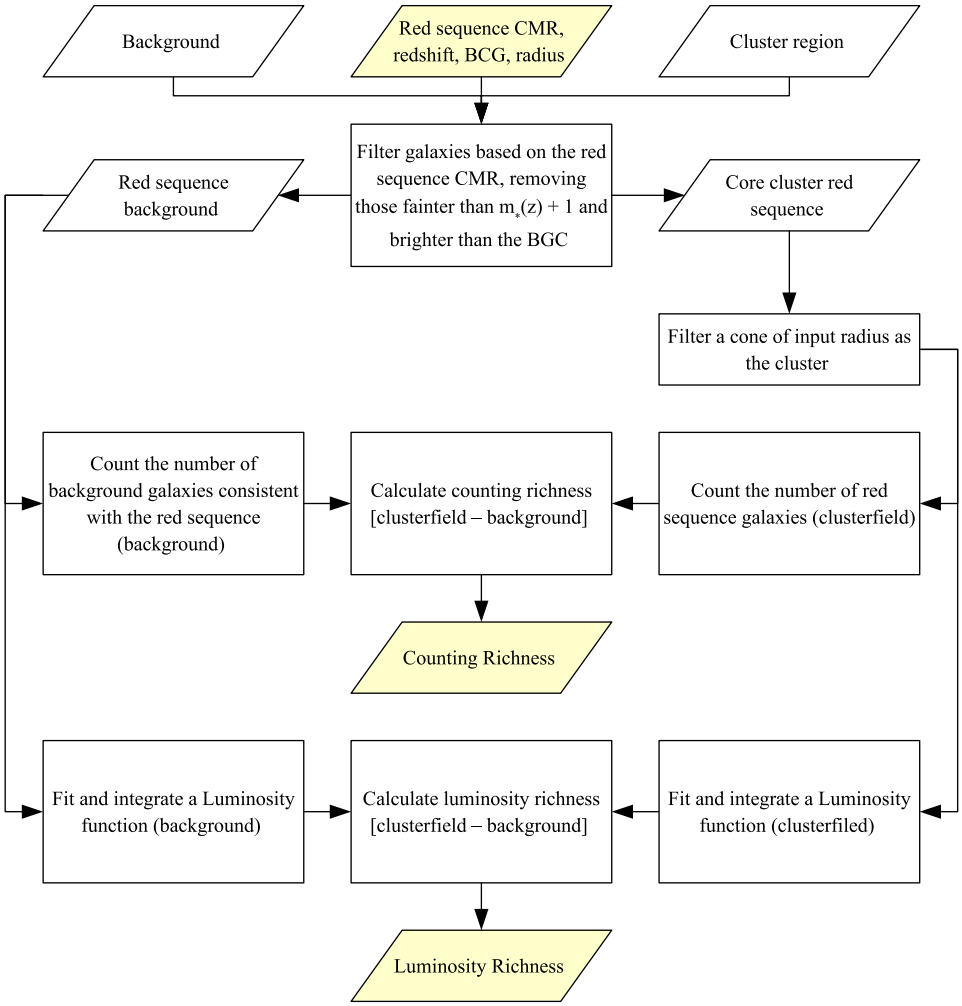

Figure 9. A detailed flowchart showing the procedures of GMPhoRCC to estimate the cluster richness, highlighted as procedure 5 in Fig. 2. An input radius of 0.5 h^{-1} Mpc leads to the intermediate richness, $n_{\rm gals}$, with n_{200} given by r_{200}.

propagate through the algorithm. These flags trace potential issues with fits, multimodal distributions and inconsistent redshifts, with a full list given in Appendix A.

Although a large source of ambiguity and potential failure results from the presence of multimodal distributions, with the prevalence of these (seen in ∼70 per cent of clusters) and the success of the candidate selection shown in Fig. 7, these are not sufficient to identify catastrophic failures alone. The strongest indicators of failure, however, are the presence of inconsistent redshifts or low richnesses.

Considering low richness, first, this indicates that the distribution modelling may be unreliable, fitting many parameters to only a few data points. More importantly, this could indicate an issue with the red sequence: Either the candidate cannot be optically confirmed as a cluster or the red sequence has been missed altogether. In either case, this is the strongest indication of catastrophic failure.

Large discrepancies between the red sequence and the BCG are also a strong indicator of catastrophic failure, particularly with regard to redshift. Large discrepancies in redshift (larger than expected considering the measurement error) can indicate a problem either with red-sequence modelling, BCG selection or cluster redshift.

By combining these flags, quality markers are assigned to clusters as an indicator of the reliability of the optical characterization, shown in Table 4. While these quality markers are universal, it is noted that the redshift inconsistency flags, `INCONSISTENT_Z_PHOT` and `INCONSISTENT_Z_SPEC`, assigned for cases where $|z_{\rm RS} - z_{\rm BCG-phot}| > \Delta z_{\rm cp}$ and $|z_{\rm RS} - z_{\rm BCG-spec}| > \Delta z_{\rm cs}$ (see Table A4), must be calibrated for specific sources of the photometric redshifts. Here $\Delta z_{\rm cp}$ and $\Delta z_{\rm cs}$ are, respectively, photometric and spectroscopic redshift consistency bounds where, for the SDSS DR10, $\Delta z_{\rm cp} = 0.035$ and $\Delta z_{\rm cs} = 0.025$. With less reliable photometric redshifts, these should be relaxed to larger bounds.

Table 4. A list of the quality markers assigned to clusters based on the GMPhoRCC flags.

Quality	Description
−1	No optical coverage
0	No characterization found
1	$n_{200} < 1$, large redshift inconsistencies, masking issues
2	$n_{200} < 3$, small redshift inconsistencies
3	Clean

With these quality markers, the characterizations can be separated into various quality subsets, as shown in Table A5: 'clean' with *quality* ≥ 3, the cleanest set with most problem clusters removed; 'mid' with *quality* ≥ 2, a middle subset with only the worst clusters

Table 5. A list of cluster subsets based on the GMPhoRCC quality marker used to remove potentially erroneous characterizations.

Subset	*quality*	Description
Detection	≥ 1	All clusters considered to have been detected, i.e. estimates were found for both redshift and richness
Mid	≥ 2	An intermediate subset removing the worst outliers, i.e. removing clusters with very low richness or large discrepancies between redshift estimates
Clean	≥ 3	The cleanest subset removing the majority of outliers, i.e. removing clusters with low richness and discrepancies between redshift estimates

removed; and 'detection' with *quality* ≥ 1, the full list of clusters considered to have been detected.

3.6 Computational performance

GMPhoRCC is aimed primarily for use with standard desktop computers and, as such, does not require substantial computational resources. Development has proceeded using PYTHON v2.7.3 with the SCIPY[1] module providing many of the mathematics routines, particularly the sequential least-squares method used to fit luminosity functions. GMPhoRCC experiences two main bottlenecks: first, from the retrieval of the optical data either from a data base or local files; and, secondly, from fitting Gaussian mixtures. While little can be done with the data retrieval, the Gaussian mixture fitting is developed using FORTRAN 90, which provides a factor of 10 speed improvement over native PYTHON and is twice as fast as the C++ version employed by Hao et al. (2010). The final performance improvement comes from the utilization of multiple threads available in even the most basic computers. While GMPhoRCC does not implement full parallelization at the FORTRAN level, the PARALLEL PYTHON[2] module allows for several cluster candidates to be analysed simultaneously. Although more were available, little improvement was found beyond six threads due to restrictions in the retrieval of the optical data.

As an example of typical performance, six threads from an Intel 3770k 4.2 −GHz processor with 16 GB of PC3-19200 RAM, accessing the optical data locally from a hard disc, have a characterization time of 42 s per cluster per thread, allowing the full characterization of the XCS catalogue, i.e. 503 clusters, within 59 min.

4 EVALUATION

Evaluation of GMPhoRCC proceeds with a two-prong approach: first, by comparing characterizations with other algorithms using spectroscopic clusters; and, secondly, by investigating mock galaxy clusters. In addition to driving the development process, particularly the calibration of the quality control system, these comparisons allow for detailed understanding of the GMPhoRCC optical selection function.

[1] http://www.scipy.org/
[2] http://www.parallelpython.com/

4.1 Comparison with existing catalogues

Comparisons with existing catalogues proceeded using spectroscopic clusters selected from the GMBCG (Hao et al. 2010), NORAS (Böhringer et al. 2000), REFLEX (Böhringer et al. 2004) and XCS (Mehrtens et al. 2012) catalogues. As richness measures are specific to the exact form of the algorithm and optical data, evaluation of the GMPhoRCC richness is thus deferred to analysis with mock clusters where comparisons with 'true' cluster values are possible.

With a total of 706 X-ray and 3795 optically detected clusters with spectra, direct evaluation of the GMPhoRCC redshift estimate is possible. Of the 4501 clusters, redshift estimates based on the photometric distribution of the red sequence were found for 97.3 per cent and compared to spectra, as shown in Fig. 10. Although some discrepancies are present, the quality markers are shown to identify and remove the worst outliers. Additionally, while the majority of all estimates are within $|z_{\rm RS} - z_{\rm spec}|/(1 + z_{\rm spec}) < 0.01$, the clean subset attains a larger fraction within this bound and less contamination with outliers. It is noted, however, that at low redshifts, $z < 0.1$, many cluster estimates are erroneous, where limitations in field area and poor contrast against the background result in cases where field galaxies dominate the cluster distributions, making it difficult to isolate the red sequence. Incompleteness and increasing measurement errors in the photometry at high redshift again cause issues with the red-sequence detection. In addition to these redshift limitations, it is expected that low-richness clusters produce the most outliers, where it is more difficult to isolate and model the red sequence with a sparse number of galaxies.

In addition to comparisons with spectra, a subset of 131 XCS clusters with both spectroscopic and photometric redshifts provides an excellent resource to compare the performance of GMPhoRCC and XCS. Fig. 11 shows the substantial improvement offered by GMPhoRCC, providing more accurate estimates with lower scatter around the spectroscopic redshifts. In addition to providing more accurate redshifts, the estimates are independent of any colour–redshift model as employed by XCS.

4.2 Richness scaling

To assess the validity of the GMPhoRCC richness as a mass proxy, an initial investigation of richness scaling is explored for the X-ray clusters from the XCS catalogue. Of particular interest is the determination of X-ray–optical scaling relations, which, in works similar to Kloster et al. (2011) and Rykoff et al. (2008), relies on the tight correlation between X-ray observables, such as temperature and cluster mass, in order to calibrate the GMPhoRCC richness as an optical mass proxy. While this paper illustrates the validity of the GMPhoRCC richness as a proxy for cluster properties, a complete analysis of such richness scaling is left for future work.

The previous subset of 131 clean spectroscopic clusters from XCS are analysed to determine the correlation between GMPhoRCC richness and X-ray temperature, modelled as a power-law scaling relation similar to those used by Rykoff et al. (2008) and defined as

$$\ln(T_{\rm x}) = \alpha + \beta \ln(n_{200}), \tag{12}$$

where α and β are constants. Determination of this relation proceeds by stacking the clusters in richness bins and using the BCES method of Section 3.2 and Akritas & Bershady (1996). While this is a rather simplistic approach for illustration, future analysis is intended using more sophisticated techniques, such as the

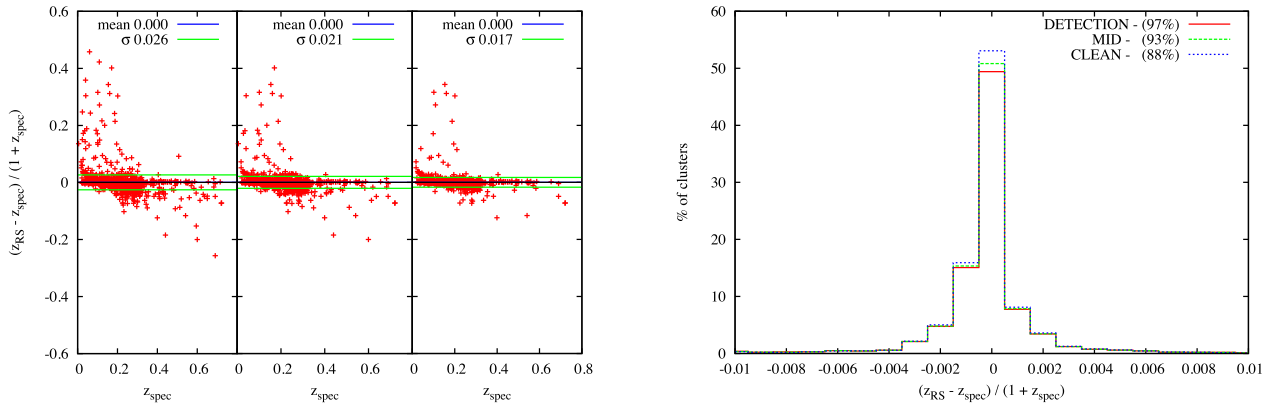

Figure 10. A comparison of GMPhoRCC photometric red-sequence redshifts to spectra using 4501 clusters with DR10 coverage from GMBCG, NORAS, REFLEX and XCS. Left-hand panel: a scatter plot highlighting the quality control, showing from the left- to right-hand side, the detection, mid and clean subsets. While some discrepancies remain, the majority of outliers have been removed in the clean subset. Although they have been correctly identified as problems, very low redshift clusters ($z < 0.05$) are not characterized well by GMPhoRCC due to poor contrast against the background and limitations in the field area. In addition, high-redshift clusters, $z > 0.5$, are subject to large discrepancies due to incompleteness and increasing photometric errors. Right-hand panel: the distribution of redshift comparisons where the results have been normalized and split into the separate quality subsets where the legend shows the fraction of the total clusters in each set. While the majority of all estimates are within $|z_{\rm RS} - z_{\rm spec}|/(1 + z_{\rm spec}) < 0.01$, the clean subset can again be seen to have removed the worst estimates with a greater fraction attaining this bound.

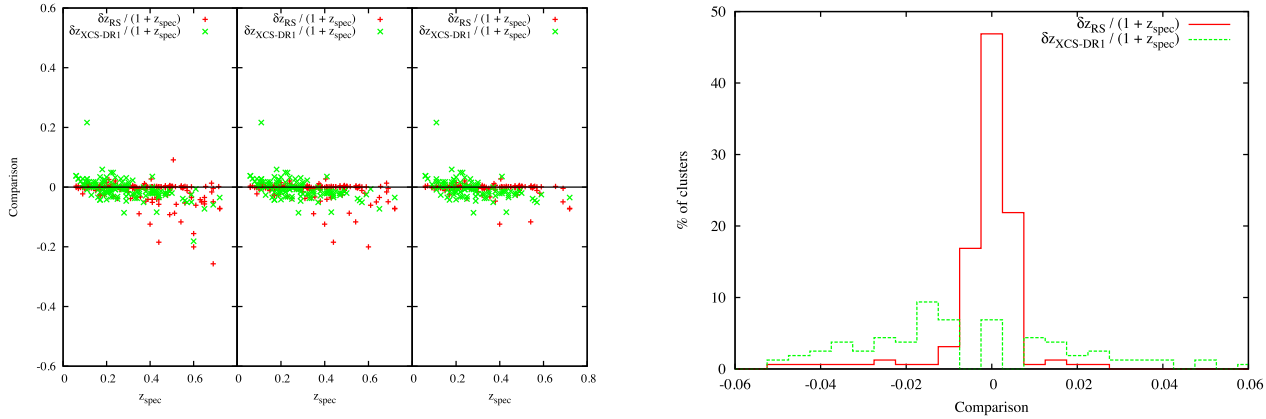

Figure 11. Redshift comparisons for the GMPhoRCC and XCS photometric redshifts to spectra for a subset of 131 XCS clusters where $\delta z_{\rm RS} = z_{\rm rs} - z_{\rm spec}$ and $\delta z_{\rm XCS-DR1} = z_{\rm XCS-DR1} - z_{\rm spec}$. Left-hand panel: a scatter plot highlighting the quality control, showing from the left- to right-hand side, the detection, mid and clean subsets. Right-hand panel: the distribution of the redshift comparisons for the clean subset, highlighting the substantial improvement offered by GMPhoRCC over XCS, providing more accurate estimates with a lower scatter around the spectroscopic redshift. Although the detection subset has a few more extreme outliers, a much greater fraction than from XCS agree within $|z_{\rm phot} - z_{\rm spec}|/(1 + z_{\rm spec}) < 0.01$, with the clean subset attaining the highest fraction in this band. In addition to providing more accurate redshifts, the estimates are independent of any colour–redshift model as employed by XCS.

Bayesian method of Rykoff et al. (2008) and Rozo & Rykoff (2014). Fig. 12 demonstrates the $T_{\rm X}$–n_{200} scaling-relation finding: $\alpha = -0.08 \pm 0.09$ and $\beta = 0.43 \pm 0.03$ with a scatter $\sigma_{\ln T_{\rm X}|\ln n_{200}} = 0.14$.

Again, as an illustration only, a clear correlation can be observed highlighting the validity of the GMPhoRCC richness as an optical proxy for cluster properties.

4.3 SDSS-like mocks

While comparisons with existing catalogues are a useful tool to evaluate a characterization method, these can take us only so far. Existing methods are subject to their own strengths, weaknesses and selection functions; hence, comparisons with a controlled 'truth' are considered with the use of mock galaxy clusters. Mock clusters can be constructed with known redshifts, richnesses and CMRs either through simulations (Cai et al. 2009; Murphy et al. 2012; Song et al. 2012, etc.) or empirically (Koester et al. 2007; Hao et al. 2010, etc.).

SDSS-like empirical mocks are constructed for use with GMPhoRCC by adding artificial clusters to field galaxies, derived from existing cluster detections and SDSS optical data. This has the advantage of producing mocks tailored to match the available photometry, allowing the specific evaluation of GMPhoRCC for the targeted optical data. The evaluation in the following sections relates to using only the intended target data, the SDSS, and while GMPhoRCC is a general algorithm for characterization, evaluation for use in other data sets is necessary before application, where results will depend on, among other things, the quality of the photometric redshifts.

Artificial clusters were generated by resampling galaxies from existing red sequences and BCGs to reproduce five main aspects of real clusters:

(i) a suitable BCG;
(ii) radial profile;

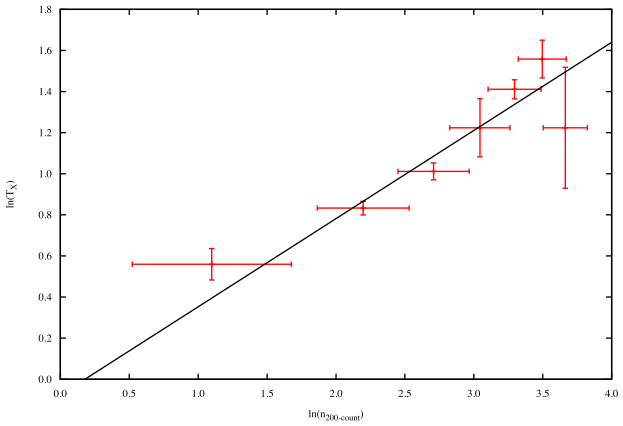

Figure 12. A preliminary analysis of the $T_{\rm X}$–n_{200} scaling relation produced using a clean spectroscopic subset of XCS. By stacking the clusters in richness bins and using the BCES method of Section 3.2 and Akritas & Bershady (1996), a best-fitting linear model is found from the average temperatures and n_{200}. The best-fitting scaling relation of equation (12) is shown with $\alpha = -0.08 \pm 0.08$, $\beta = 0.43 \pm 0.03$ and a scatter of $\sigma_{\ln T_{\rm X}|\ln n_{200}} = 0.14$.

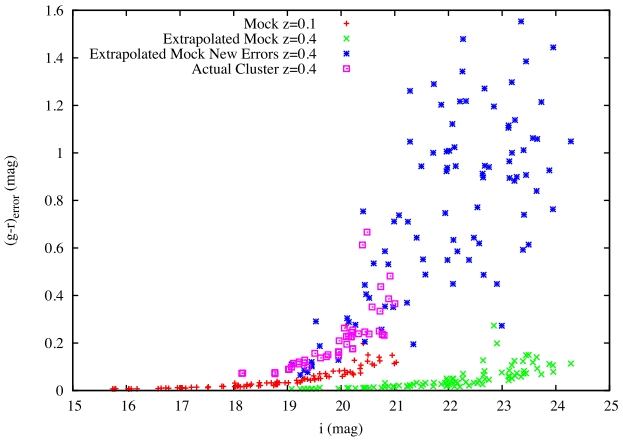

Figure 13. A plot showing how the $g - r$ error compares between the $z = 0.1$ seed cluster, the extrapolated cluster at $z = 0.4$ with and without the new errors, and a real $z = 0.4$ cluster. Without the new photometric errors, it is clear that the extrapolated cluster underestimates the $g - r$ error. The new errors are seen to be in good agreement with an observed cluster at the same high redshift.

(iii) redshift distribution;
(iv) luminosity function;
(v) CMRs/colour distributions.

As these are dependent on the properties of the cluster, it is necessary to resample from red sequences that best match the target mock. Rather than using a small number of well-observed seed clusters, red sequences were identified and stacked in redshift/richness space in order to provide a source of galaxies suitable for a range of mock properties.

Using GMPhoRCC, 10 000 very clean red sequences were identified from the C4, GMBCG, REFLEX, NORAS and XCS catalogues with very good agreement between spectroscopic and GMPhoRCC redshifts, $|z_{\rm RS} - z_{\rm spec}| < 0.005$. By separately stacking the BCG and red-sequence galaxies of these clusters in redshift–richness bins, a larger source is produced to sample cluster properties than from considering these individually. Stacking many clusters in this way ensures that each bin is dominated by the red sequence where the bulk properties are representative of a cluster with the bin redshift and richness.

While the available richness was fixed by the original clusters, extrapolation by adding a fixed Δz to each galaxy allowed a larger redshift range to be sampled. Photometry was then adjusted with K+e-corrections to account for evolution and observations at different redshifts. K-corrections were performed using KCORRECT v4.2 from Blanton & Roweis (2007), with evolutionary corrections taken from Koester et al. (2007). Colour evolution models from Tojeiro et al. (2011) were also considered but provided no significant deviation from the main results of this section.

While extrapolating to much higher redshifts, care is needed to reproduce appropriate errors. This mainly affects high-redshift artificial clusters that should possess higher errors than the low-redshift seed due to the fainter photometry. To reproduce appropriate errors, a sample of ∼500 000 red sequence galaxies are used to model the various error distributions. For photometry errors, the distribution are modelled as a function of band magnitude, whereas the redshift error is modelled as a function of both i-band magnitude and redshift. For high-redshift extrapolation, a new error is drawn from this distribution with magnitudes and redshifts updated by randomly shuffling about this error. Although these errors depend on a number of things, including seeing, this method reproduces sensible results providing good agreement with existing high-redshift clusters, as shown in Fig. 13.

Generation of the artificial clusters now proceeds as follows:

(i) randomly select a redshift and richness;
(ii) select the closest redshift/richness bin;
(iii) resample with replacement, a BCG and red sequence;
(iv) apply a fixed Δz to extrapolate from bin to mock cluster redshift;
(v) K+ e-correct photometry;
(vi) sample suitable errors and reshuffle redshift and photometry.

Finally, to simulate SDSS completeness levels, members were removed with i band >21 mag.

Appropriate backgrounds for these artificial clusters are constructed by removing the red sequence from the original 10 000 fields considered by GMPhoRCC. The backgrounds are binned in the same redshift/richness space as used previously according to the properties of the cluster. Real backgrounds are assigned by randomly selecting from the bin that best matches the properties of the artificial cluster. A total of 8745 mocks were prepared with $0.05 < z < 1.1$ and $5 \leq n_{200} < 75$ by randomly inserting the artificial cluster within 3 arcmin of the centre of the assigned backgrounds. This method has the advantage of modelling the background as a function of the local neighbourhood where the background densities encountered would be typical for the given properties of the artificial cluster.

4.3.1 Richness consistency

With the use of multiple red-sequence bands, it is necessary to ensure that the GMPhoRCC richness estimate is consistent across the large redshift ranges considered, which is indeed confirmed by analysis of the artificial clusters. Fig. 14 shows how the GMPhoRCC estimate richness of artificial clusters, generated at $z = 0.1$, evolves as these are extrapolated across $0.05 < z < 1.1$. While incompleteness results in a loss of richness above $z > 0.45$, the GMPhoRCC estimate is consistent at lower redshifts. In addition, the luminosity

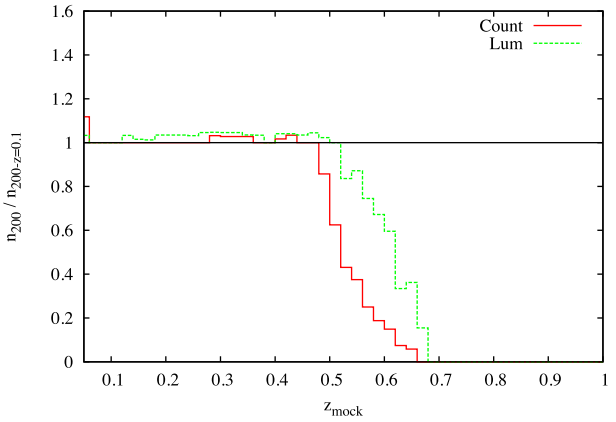

Figure 14. A comparison of luminosity and counting n_{200} as a function of redshift for mocks with an SDSS-complete photometry cut, $i < 21$ mag. While incompleteness results in a loss of richness for $z > 0.45$, at lower redshifts, the GMPhoRCC richness is consistent. In addition, the luminosity richness is shown to be able to reliably extrapolate beyond the $z \sim 0.45$ completeness level up to $z \sim 0.55$.

estimate has been shown to extrapolate into regions with incomplete photometry.

4.4 Comparison with mocks

With the SDSS mocks, direct comparisons of GMPhoRCC estimates to 'true' cluster values are possible. In order to assess the accuracy and bias of GMPhoRCC, the *i* band <21 cut on mock members is not used, where the analysis of the full effect of incomplete photometry is deferred to a study of completeness in Section 4.6.

Of the 7050 mocks, estimates were found for 99.2 per cent and compared to the cluster values, as shown in the left-hand panel of Fig. 15. Redshift comparisons agree with those from real spectroscopic clusters where the GMPhoRCC are unbiased with the majority achieving $|z_{RS} - z_{mock}|/(1 + z_{mock}) < 0.01$. In addition, the clean subset attains a larger fraction within this bound and less contamination with outliers again highlighting the value of the quality control system.

Richness comparisons, presented in right-hand panel of Fig. 15, confirm that the GMPhoRCC estimate is unbiased with $|n_{200-count} - n_{200-mock}| = 0.01 \pm 0.005$ and $|n_{200-lum} - n_{200-mock}| = -0.03 \pm 0.02$. In addition, it is clear that both the counting and luminosity function methods are able to adequately recover cluster richness. An accurate richness estimate is far more challenging to determine than redshift, as evident by the larger scatter. These difficulties arise due to the discreteness of n_{200} and the sensitivity to discrepancies in redshift, r_{200}, n_{gals}, BCG identification, CMR modelling and projection effects. In addition, the luminosity method is subject to a lager scatter as a result of extra complexity and uncertainty introduced by fitting and integrating a luminosity function.

4.5 Purity

Although the target clusters for GMPhoRCC have already been detected in other wavebands (e.g. X-ray), it is important to understand purity when using the code to optically confirm a candidate or in cases where the candidate list may be contaminated. By using random real backgrounds only, purity is estimated as the fraction of fields where no cluster was detected, i.e. detections in this case are impurities. While this tackles only the issue of false detections, the validity of the various GMPhoRCC estimates is assessed further in Section 4.6, which may be incorrect for a number of reasons, including projection effects.

Table 6 presents GMPhoRCC purity results that represent the probability that a candidate is in fact a cluster, given that it was assigned a particular quality marker and richness. Very few spurious characterizations are found with high quality or richness, i.e. these have the highest probability of representing real clusters. Of particular note is the fact that candidates belonging to the clean subset, *quality* ≥ 3, have a negligible probability of resulting from a false

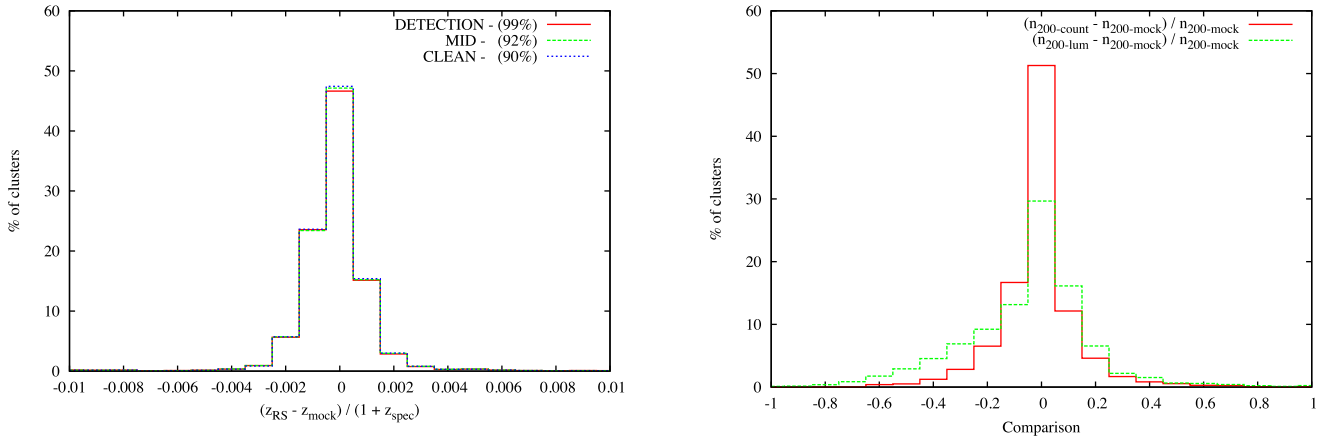

Figure 15. Comparisons of GMPhoRCC estimates to actual values for 7050 SDSS-like mock galaxy clusters. Left-hand panel: the distribution of the redshift comparisons where the results have been normalized and split into the separate quality subsets where the legend shows the fraction of the total clusters in each set. While the majority of all estimates are within $|z_{RS} - z_{mock}|/(1 + z_{mock}) < 0.01$, the clean subset can again be seen to have removed the worst estimates with a greater fraction attaining this bound. This agrees well with results using real spectroscopic clusters showing accurate unbiased estimates. Right-hand panel: the distribution of the n_{200} comparison for the clean subset showing both the counting and luminosity function methods. Both methods are observed to provide accurate and unbiased estimates of cluster richness. The larger scatters highlight the difficulty in recovering richness compared with redshift. The discreteness of n_{200} and the sensitivity to discrepancies in redshift, r_{200}, n_{gals}, BCG identification, CMR modelling and projection effects all contribute to the larger scatter. In addition, the luminosity method is subject to the extra complexity and uncertainty introduced by fitting and integrating a luminosity function, further increasing scatter.

Table 6. A list of the GMPhoRCC purity results based on counting richness and quality marker. Very few spurious characterizations are found with high quality or richness, i.e. these candidates have the highest probability of representing real clusters. This analysis concerns only false detections where the validity of the richness estimate, which may be inaccurate due to projection effects, is further assessed in Section 4.6.

quality	$n_{200} > 0$	$1 \leq n_{200} < 5$	$5 \leq n_{200} < 10$	$n_{200} \geq 10$
>0	79.7 per cent	80.7 per cent	99.1 per cent	100.0 per cent
1	88.6 per cent	89.2 per cent	99.4 per cent	100.0 per cent
2	92.0 per cent	92.1 per cent	99.9 per cent	100.0 per cent
3	99.2 per cent	99.4 per cent	99.8 per cent	100.0 per cent

detection. It is noted that GMPhoRCC attains extremely high levels of purity compared with maxBCG, which attains ∼93 per cent for clusters with $n_{200} = 10$ and ∼99 per cent for $n_{200} = 15$, and with GMBCG, which attains purity levels of ∼75 per cent for $n_{200} > 10$ and ∼97 per cent for $n_{200} > 25$.

4.6 Completeness

One of the most important properties to evaluate is completeness; this gives a measure of how well clusters are characterized across a range of redshifts and richnesses. Completeness is measured as the fraction of mock clusters where the estimated properties agree with the actual value within a given bound. In order to estimate the optical selection function, completeness is considered with respect to redshift, richness and BCG matching.

4.6.1 Redshift recovery

Using the fraction of the clean subset that attains the bound $|z_{rs} - z_{mock}| < 0.03$, comparable to typical SDSS photometric redshift errors, Fig. 16 highlights completeness as a function of both richness and redshift with full results shown in Table 7. For $z_{mock} < 0.5$, the majority of the GMPhoRCC estimates are in very good agreement with the mock value, with high levels of completeness attained. Above this point, photometry incompleteness results in difficulties in modelling the red sequence resulting in the lower completion. In addition to this, limitations in field area and poor contrast against the background for low-redshift clusters, $z < 0.1$, make the red sequence more difficult to isolate and model, resulting in the lower fraction of clusters with good redshift estimates. As expected, low-richness clusters suffer from lower completeness due to difficulties in modelling sparse data sets. In addition, these are seen to be more susceptible to photometry cuts resulting in the earlier reduction in completeness.

Extending this completeness analysis by considering the subset of clusters with a given *quality* and z_{RS}, the accuracy of the GMPhoRCC redshift is estimated. Shown in the right-hand panel of Fig. 16 is the fraction of these subsets attaining the redshift bound. This represents the probability that a cluster, having a specific

Table 7. A summary of the redshift completeness results.

Subset	Richness	Redshift	Completeness
Detection	All	$0.05 \leq z < 0.65$	95.8 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.55$	97.7 per cent
	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.55$	97.2 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.65$	98.3 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.65$	98.1 per cent
Mid	All	$0.05 \leq z < 0.60$	92.6 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.55$	91.7 per cent
	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.55$	93.0 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.65$	93.8 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.65$	94.8 per cent
Clean	All	$0.05 \leq z < 0.60$	89.3 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.50$	89.4 per cent
	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.50$	92.2 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	92.6 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	93.6 per cent

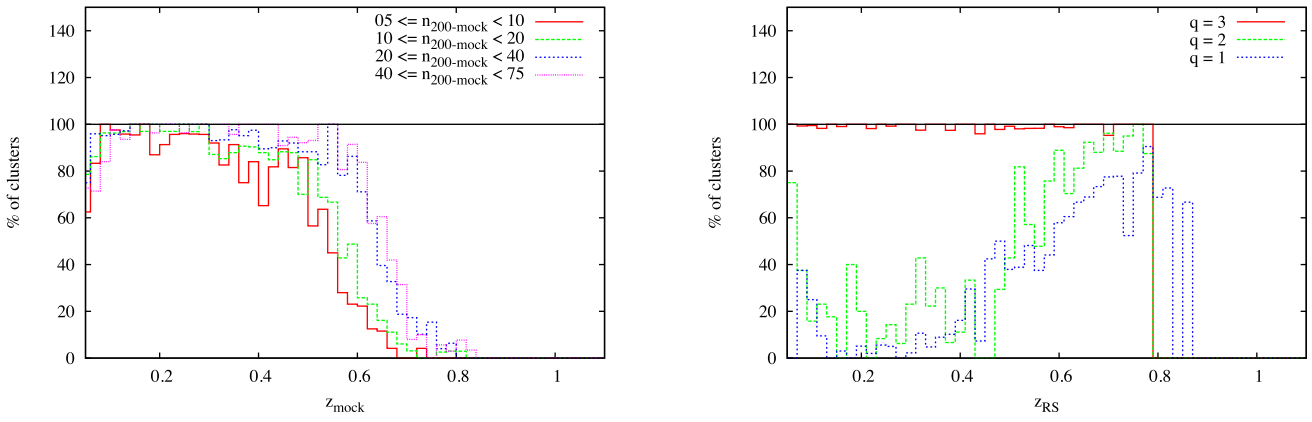

Figure 16. Left-hand panel: the fraction of the clean subset of mock clusters for different richness bands where the redshift estimate is within $|z_{RS} - z_{mock}| < 0.03$. Low-richness clusters are more sensitive to incomplete photometry due to the already low number of galaxies. Isolating the red sequence and estimating redshift is more challenging for low-richness clusters than their high-richness counterparts at the same redshift. Hence, the ability to reliably estimate cluster redshift drops more quickly with redshift for groups than rich clusters. Due to this difficulty, the clean set is also subject to an earlier reduction in completeness and a lower fraction of low-richness clusters across all redshifts. Right-hand panel: the fraction of mock clusters with a given *quality* that achieve the $|z_{RS} - z_{mock}| < 0.03$ bound. In addition to the clean subset, *quality* ≥ 3, achieving a very high probability (>97 per cent) that the redshift estimate is within 0.03 of the mock value, those with lower quality for $z_{RS} < 0.45$ have low probabilities (<25 per cent), again showing the ability of the quality marker to identify and remove potential outliers. The sparse number of galaxies above $z > 0.45$ in the SDSS DR10 and the mock background results in a low chance of spurious high-redshift estimates; hence, given $z_{RS} > 0.45$, there is a larger probability that the redshift is associated with the cluster. In addition, those with good high-redshift estimates are more likely to be flagged as low-richness ones due to the incomplete photometry. This leads to higher probabilities that the estimate is associated with the cluster than expected for *quality* < 3.

Table 8. A list of the probabilities that a redshift estimate is within various bounds of the actual value, given the z_{RS} estimate and *quality* of the cluster, where $\Delta z = |z_{RS} - z_{mock}|$. The increase in probability for low-quality high-redshift clusters is clear. The sparse number of galaxies above $z > 0.45$ in the SDSS DR10 and the mock background results in a low chance of spurious high-redshift estimates; hence, given $z_{RS} > 0.45$, there is a larger probability that the redshift is associated with the cluster than expected for those with *quality* < 3.

quality	Redshift range	$p(\Delta z < 0.01)$	$p(\Delta z < 0.03)$	$p(\Delta z < 0.05)$
1	$0.05 \leq z_{RS} < 0.50$	0.11	0.11	0.12
2	$0.05 \leq z_{RS} < 0.50$	0.16	0.19	0.25
3	$0.05 \leq z_{RS} < 0.80$	0.96	0.99	0.99
1	$0.50 \leq z_{RS} < 0.80$	0.45	0.61	0.66
2	$0.50 \leq z_{RS} < 0.80$	0.68	0.81	0.86

quality and z_{rs}, achieves $|z_{rs} - z_{mock}| < 0.03$. In addition to the clean subset, *quality* ≥ 3, achieving a very high probability (>97 per cent) that the redshift estimate is within 0.03 of the mock value, those with lower quality for $z < 0.45$ have low probabilities (<25 per cent), again showing the ability of the quality subsets to identify and remove potential outliers. The sparse number of galaxies above $z > 0.45$ in the SDSS DR10 and the mock background results in a low chance of spurious high-redshift estimates; hence, given $z_{RS} > 0.45$, there is a larger probability that the redshift is associated with the cluster. In addition, those with good high-redshift estimates are more likely to be flagged as low-richness ones due to the incomplete photometry. This leads to higher probabilities that the estimate is associated with the cluster than expected for *quality* < 3. While adjustments could be made to the quality subsets to take advantage of this increased probability, it is noted that for $z > 0.45$, the lower quality mainly results from low numbers of galaxies due to incompleteness, and hence the current quality subsets are necessary to maintain cleanliness for both redshift and richness estimates. A full set of probabilities for each quality marker and several bounds is presented in Table 8.

4.6.2 *Richness recovery*

While two richnesses are investigated by GMPhoRCC, n_{200} best represents an optical mass proxy, considering galaxies within a characteristic radius, rather than the fixed aperture of n_{gals}, and hence is the subject of this section. With the extra n_{gals} step, additional sources of error are introduced, and with n_{200} highly sensitive to correct CMR modelling, BCG selection and redshift richness attain a much larger spread about the mock value, and thus relatively large completeness bounds are considered.

Fig. 17 highlights richness completeness as the fraction of the clean subset where the counting richness is within 25 per cent of the mock value. Completeness in both the counting and luminosity estimates tails off above $z > 0.45$ due to incomplete photometry where cluster galaxies become too faint for reliable detection. It is noted that the luminosity method is able to extrapolate richness, resulting in a slower reduction with redshift and higher completeness than the counting method for $z > 0.45$.

As stated in the previous sections, low-richness clusters are difficult to model and analyse due to difficulties in fitting distributions to a small number of galaxies, and this is reflected in the lower completeness rates. In addition to this, background fluctuations that, in this case, can cause as much as an 80 per cent discrepancy due to discreteness further reduce completeness. Table 9 summarizes and extends these results to the luminosity richness, $n_{200-lum}$.

Again completeness is considered with respect to subsets with a given *quality* and z_{rs} to estimate the accuracy of GMPhoRCC richness. Shown in the right-hand panel of Fig. 17 is the fraction of these subsets that attain the richness bound, representing the probability that, given a cluster has a specific *quality* and z_{rs}, the richness estimate is within 25 per cent of the mock value. A full set of probabilities for each quality marker and several bounds is presented in Table 10. In addition to the clean subset, *quality* ≥ 3, achieving a very high probability (>80 per cent) that the richness estimate is within 25 per cent of the mock value, those with higher quality markers have low probabilities (<15 per cent), again showing the ability of the quality subsets to identify and remove potential outliers. In addition, while the counting richness has negligible probability of matching the mock at high redshift ($0.5 < z_{RS} < 0.8$), the luminosity method is clearly able to extrapolate,

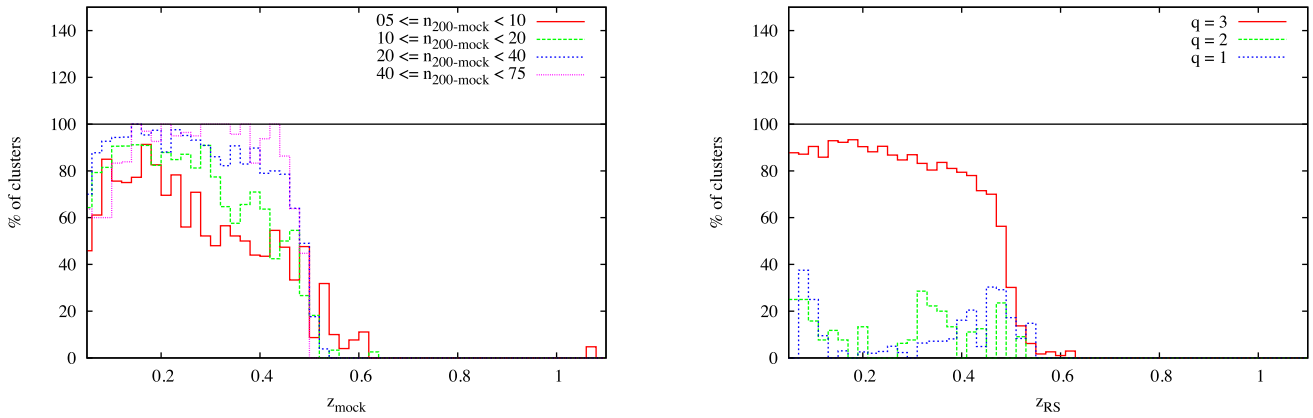

Figure 17. Left-hand panel:the fraction of the clean subset of mock clusters in richness bands where the $n_{200-count}$ estimate was within 25 per cent of the actual value. Above $z > 0.45$, incompleteness in photometry results in the sharp decline in richness recovery rates. In addition, low-richness clusters are more susceptible to this decline due to modelling the already low number of galaxies. In addition to this, n_{200} is highly sensitive to errors in redshift, BCG selection and n_{gals}, resulting in lower completion rates than with redshift. Right-hand panel: the fraction of mock clusters with a given *quality* where the n_{200} estimate was within 25 per cent of the original value. In addition to the clean subset, *quality* ≥ 3, achieving a high probability (>80 per cent) that the richness estimate is within 25 per cent of the mock value, those with lower quality have very low probabilities (<15 per cent), again showing the ability of the quality subsets to identify and remove potential outliers.

Table 9. A summary of results where the top and bottom tables present completeness where the $n_{200-\mathrm{count}}$ and $n_{200-\mathrm{lum}}$ estimates, respectively, are within 25 per cent of the mock value. Due to the extrapolation, the luminosity method attains a deeper redshift range than counting before completeness rates decline.

Subset	Richness	Redshift	Completeness
	(a) The counting richness, $n_{200-\mathrm{count}}$		
	All	$0.07 \leq z < 0.45$	85.0 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	67.5 per cent
Detection	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	81.2 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.45$	92.8 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.45$	93.2 per cent
	All	$0.07 \leq z < 0.45$	83.3 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	65.6 per cent
Mid	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	79.2 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.45$	91.0 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.45$	92.3 per cent
	All	$0.07 \leq z < 0.45$	82.4 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	64.3 per cent
Clean	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	77.9 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.45$	90.7 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.45$	91.3 per cent
	(b) The luminosity richness, $n_{200-\mathrm{lum}}$		
	All	$0.07 \leq z < 0.50$	81.8 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	67.5 per cent
Detection	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	81.2 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.52$	87.5 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.52$	88.0 per cent
	All	$0.07 \leq z < 0.50$	79.5 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	65.6 per cent
Mid	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	79.2 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.52$	85.1 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.52$	86.9 per cent
	All	$0.07 \leq z < 0.50$	78.6 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.45$	64.3 per cent
Clean	$10 \leq n_{200} < 20$	$0.07 \leq z < 0.45$	77.9 per cent
	$20 \leq n_{200} < 40$	$0.07 \leq z < 0.52$	84.8 per cent
	$40 \leq n_{200} < 75$	$0.07 \leq z < 0.52$	86.0 per cent

achieving a 30 per cent probability that the richness is within 25 per cent of the mock value.

4.6.3 *BCG identification*

Identifying the correct BCG is very important not only for subsequent cosmology, but also for calculating cluster richness. This analysis considers two scenarios: one where the BCG is correctly identified and the other where any cluster member is selected as the BCG. While correctly matching the BCG shows that the strongest evidence that GMPhoRCC has suitably modelled the red sequence, even matching to a cluster member suggests that the CMR is a reasonable representation of the cluster.

Mismatching the BCG results from two main issues: background interlopers and poor red-sequence modelling. While mismatching to a background galaxy is easier to find with the quality markers due to inconsistencies in redshift, matching to another cluster member can be more challenging to identify. Fig. 18 shows the fraction of the clean subset of mocks where the BCG has been correctly matched. As photometry becomes incomplete, issues with fitting the red sequence due to the lower number of galaxies give rise to the lower fraction matched above $z > 0.5$. In addition to this, the difficulty in modelling the red sequence at low redshift, $z < 0.1$, due to poor contrast against the background and limitations in the field area, results in a smaller fraction of these mocks with correctly matched BCG.

Table 10. A list of the probabilities that the n_{200} estimate is within various bounds of the actual value, given the z_{RS} estimate and the quality marker of the cluster. The top and bottom tables present results for the counting and luminosity richnesses, respectively, where $\Delta n_{\mathrm{c}} = |n_{200-\mathrm{count}} - n_{200-\mathrm{mock}}|/n_{200-\mathrm{mock}}$ and $\Delta n_{l} = |n_{200-\mathrm{lum}} - n_{200-\mathrm{mock}}|/n_{200-\mathrm{mock}}$. In addition to the clean subset, *quality* ≥ 3, achieving the highest probabilities that the richness estimate is within given bounds of the mock value, those with lower quality have very low probabilities (<15 per cent), again showing the ability of the quality subsets to identify and remove potential outliers. In addition, while the counting richness has negligible probability of matching the mock at high redshift ($0.5 < z_{\mathrm{RS}} < 0.8$), the luminosity method is clearly able to extrapolate, achieving a 30 per cent probability that the richness is with 25 per cent of the mock value.

quality	Redshift range	$p(\Delta n_{\mathrm{c}} < 0.10)$	$p(\Delta n_{\mathrm{c}} < 0.25)$	$p(\Delta n_{\mathrm{c}} < 0.50)$
	(a) The counting richness $n_{200-\mathrm{count}}$			
1	$0.05 \leq z_{\mathrm{RS}} < 0.45$	0.03	0.06	0.08
2	$0.05 \leq z_{\mathrm{RS}} < 0.45$	0.05	0.11	0.20
3	$0.05 \leq z_{\mathrm{RS}} < 0.45$	0.53	0.86	0.97
quality	Redshift range	$p(\Delta n_{\mathrm{l}} < 0.10)$	$p(\Delta n_{\mathrm{l}} < 0.25)$	$p(\Delta n_{\mathrm{l}} < 0.50)$
	(b) The luminosity richness $n_{200-\mathrm{lum}}$			
1	$0.05 \leq z_{\mathrm{RS}} < 0.50$	0.04	0.09	0.12
2	$0.05 \leq z_{\mathrm{RS}} < 0.50$	0.05	0.11	0.20
3	$0.05 \leq z_{\mathrm{RS}} < 0.50$	0.51	0.83	0.96
1	$0.50 \leq z_{\mathrm{RS}} < 0.80$	0.04	0.09	0.22
2	$0.50 \leq z_{\mathrm{RS}} < 0.80$	0.04	0.11	0.24
3	$0.50 \leq z_{\mathrm{RS}} < 0.80$	0.13	0.30	0.58

Again, it is expected that a smaller fraction of low-richness clusters have suitably determined CMRs due to the difficulty in modelling a sparse number of galaxies, and this is reflected in the lower BCG match rates. Table 11 summarizes and extends these results to correct BCG matching and cluster member BCG matching for each of the quality subsets.

Again, considering the BCG matching fractions with respect to subsets with a given *quality* and z_{RS} gives an estimate of the probability that the BCG has been correctly matched, given that the cluster is consistent with the subset. The right-hand panel of Fig. 18 shows these fractions for clusters where the BCG has been correctly identified. A full set of probabilities for each quality marker and the different BCG sources are presented in Table 12 . In addition to the clean subset, *quality* ≥ 3, achieving a very high probability (>90 per cent) that the BCG has been correctly identified, those with lower quality for $z < 0.45$ have low probabilities (<10 per cent), again showing the ability of the quality subsets to identify and remove cases where the red sequence has not been well modelled. The sparse number of galaxies above $z > 0.45$ in the SDSS DR10 and the mock background results in a low chance of spurious high-redshift estimates; hence, given $z_{\mathrm{RS}} > 0.45$, there is a larger probability that the CMR and BCG are associated with the cluster. In addition, those with suitable CMRs at high redshifts are more likely to be flagged as low-richness ones due to the incomplete photometry. This leads to higher probabilities that the CMR and BCG are associated with the cluster than expected for *quality* < 3. Again, no adjustments are made to the quality subsets since incomplete photometry becomes an issue for $z_{\mathrm{RS}} > 0.45$ with the current subsets necessary to maintain cleanliness for both redshift and richness estimates.

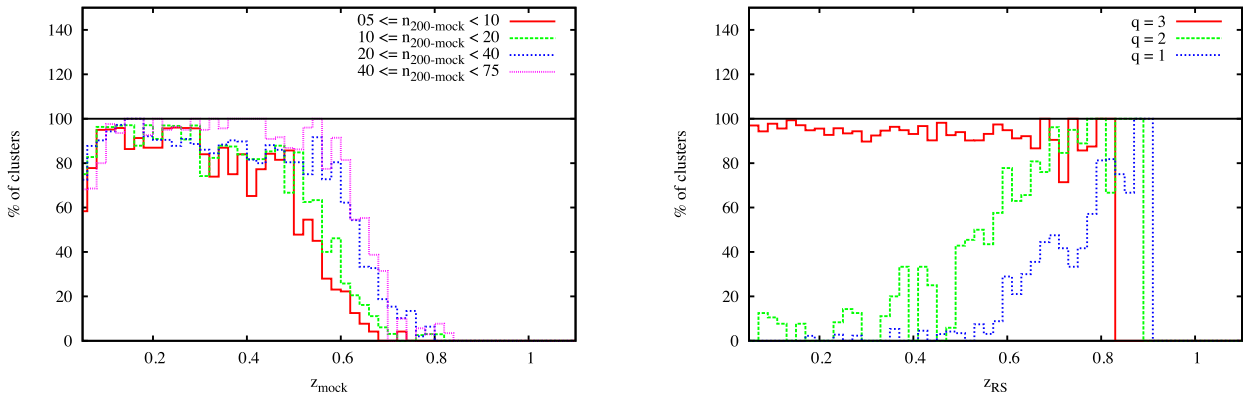

Figure 18. Left-hand panel:the fraction of the clean subset of mock clusters for different richness bands where the BCG has been correctly identified. Above $z > 0.5$, incompleteness in photometry results in difficulties fitting the red sequence, resulting in the reduction of matching rates. In addition to the difficulty in modelling low-richness clusters, these suffer from earlier declines and lower matching rates. Right-hand panel: the fraction of mock clusters with a given *quality* where the BCG has been correctly identified. In addition to the clean subset, *quality* ≥ 3, achieving a very high probability (>90 per cent) that the BCG has been correctly identified, those with lower quality for $z < 0.45$ have low probabilities (<10 per cent), again showing the ability of the quality subsets to identify and remove cases where the red sequence has not been well modelled. Again, the sparse number of galaxies above $z > 0.45$ in the SDSS DR10 and the mock background results in a low chance of spurious high-redshift estimates; hence, given $z_{RS} > 0.45$, there is a larger probability that the CMR and BCG are associated with the cluster. In addition, those with suitable CMRs at high redshifts are more likely to be flagged as low-richness ones due to the incomplete photometry. This leads to higher probabilities that the CMR and BCG are associated with the cluster than expected for *quality* < 3.

Table 11. A summary of the completeness results where the top table presents the completeness where the BCG has been correctly identified and the bottom table presents the completeness where the BCG has been matched to any cluster member. A smaller fraction of low-richness clusters have suitable CMRs due to the difficulty in modelling a sparse number of galaxies, resulting in lower completeness rates in both cases.

Subset	Richness	Redshift	Completeness
	(a) Correctly identified BCG		
	All	$0.05 \leq z < 0.60$	87.9 per cent
	$5 \leq n_{200} < 10$	$0.05 \leq z < 0.55$	86.6 per cent
Detection	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	89.1 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	87.2 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	93.1 per cent
	All	$0.05 \leq z < 0.58$	87.8 per cent
	$5 \leq n_{200} < 10$	$0.05 \leq z < 0.50$	87.3 per cent
Mid	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	88.7 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	86.9 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	92.6 per cent
	All	$0.05 \leq z < 0.58$	86.1 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.50$	86.6 per cent
Clean	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	87.1 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	85.9 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	91.7 per cent
	(b) BCG matched to any cluster member		
	All	$0.05 \leq z < 0.60$	91.2 per cent
	$5 \leq n_{200} < 10$	$0.05 \leq z < 0.55$	88.0 per cent
Detection	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	90.7 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	92.6 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	96.7 per cent
	All	$0.05 \leq z < 0.58$	90.9 per cent
	$5 \leq n_{200} < 10$	$0.05 \leq z < 0.50$	88.8 per cent
Mid	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	90.2 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	92.2 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	95.4 per cent
	All	$0.05 \leq z < 0.58$	88.7 per cent
	$5 \leq n_{200} < 10$	$0.07 \leq z < 0.50$	87.5 per cent
Clean	$10 \leq n_{200} < 20$	$0.05 \leq z < 0.55$	88.3 per cent
	$20 \leq n_{200} < 40$	$0.05 \leq z < 0.62$	91.1 per cent
	$40 \leq n_{200} < 75$	$0.05 \leq z < 0.62$	93.1 per cent

Table 12. A list of the probabilities that the BCG has been matched to various sources, given that the cluster has a specific quality marker and redshift estimate. In both scenarios, the clean set demonstrates the highest probability that the red sequence has been suitably modelled.

quality	Redshift range	*p*(correct BCG)	*p*(cluster member BCG)
1	$0.05 \leq z_{RS} < 0.60$	0.03	0.03
2	$0.05 \leq z_{RS} < 0.50$	0.10	0.16
3	$0.05 \leq z_{RS} < 0.80$	0.94	0.97
1	$0.60 \leq z_{RS} < 0.80$	0.39	0.39
2	$0.50 \leq z_{RS} < 0.80$	0.72	0.75

5 DISCUSSION AND CONCLUSIONS

Presented in this paper is GMPhoRCC, which is designed to take cluster candidates, previously detected, and provides an optical confirmation and characterization based on the red sequence. GMPhoRCC has been designed specifically to attain estimates of redshift, richness and the red-sequence CMR and offers many advantages over existing algorithms, including treatment of multimodal distributions, treatment of a variable-width full CMR red sequence, richness extrapolation and quality control. One of the most important features developed is the flag and quality control procedure. By flagging issues, particularly low-richness and inconsistent red-sequence and BCG redshifts, potential catastrophic failures can be identified and removed from cleaner subsets.

Comparisons with other characterization methods highlight the advantages of GMPhoRCC. Using a sample of 4501 clusters taken from the GMBCG (Hao et al. 2010), NORAS (Böhringer et al. 2000), REFLEX (Böhringer et al. 2004) and XCS (Mehrtens et al. 2012) catalogues, GMPhoRCC redshift estimates are compared to spectra showing low scatter [$\sigma_{\delta z/(1+z)} \sim 0.026$] around the actual value. In addition, applying the quality control to produce a clean subset removes most outliers, giving a much tighter agreement: $\sigma_{\delta z/(1+z)} \sim 0.017$ showing significant improvement over maxBCG, $\sigma_{\delta z/(1+z)} \sim 0.025$, and XCS, $\sigma_{\delta z/(1+z)} \sim 0.050$. The high accuracy of GMPhoRCC is also demonstrated with a significant percentage (∼75 per cent) of all redshift estimates from the clean subset agreeing within $|z_{RS} - z_{spec}| < 0.01$.

While analysing known clusters provides useful feedback, comparisons with those with known properties are far more valuable; hence, the remaining evaluation of GMPhoRCC proceeded with the use of empirical mock galaxy clusters. These mocks were produced by stacking red-sequence galaxies from existing clusters, analysed using data from SDSS, in redshift–richness bins from which new sequences are resampled. This extends the similar approaches of maxBCG and GMBCG, where only rich clusters are used as seeds to generate mocks with a range of properties.

Assessment of the optical selection function proceeded with the consideration of completeness, the fraction of mocks with characterizations within given bounds of the actual value. First, incomplete photometry, simulated by an i band <21 cut, is shown to remove members for clusters with $z > 0.45$. Redshift completeness, i.e. the fraction of clusters within 0.03 of the mock value, is not immediately hindered by the photometry attaining 93 per cent for $0.05 < z < 0.62$ for clusters with a richness greater than 20. With the large scatters in the estimates, richness attains lower completeness rates, mostly due to projection effects and background fluctuations, as also noted by Hao et al. (2010). The fraction of clusters within 25 per cent of the mock value, defining completeness, is measured as 91 per cent for $0.07 < z < 0.45$ for clusters with richness greater than 20, 78 per cent for those with richness between 10 and 20, and 64 per cent for those with richness less than 10.

Additionally, evaluation with mocks had confirmed the value of the quality control system showing a high probability that, given a cluster is in the clean set, the redshift and richness estimates are within a given bound of the mock value. Most importantly, it was shown that those with lower quality markers, indicating less confidence in the characterization, show much smaller probabilities confirming that the quality control is effective in identifying potential catastrophic failures.

The cluster redshifts and richnesses obtained using the methods described in this paper will underpin the scientific exploitation of future XCS data releases. For example, Manolopoulou et al. (in preparation) are currently preparing for release a large new cluster catalogue comprising the subset of the XCS Data Release 2 that falls within the footprint of SDSS using redshifts and richnesses from GMPhoRCC. This will form the basis of improved scaling relations between the X-ray properties of the XCS clusters and the optical properties of their galaxy populations, constraining models of the evolution of the intracluster medium and the effect of the cluster environment on galaxy evolution.

ACKNOWLEDGEMENTS

RJH acknowledges the support of the Science and Technology Facilities Council (STFC) via the award of an STFC studentship. This paper makes use of optical data from SDSS-III. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation and the U.S. Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration, including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

REFERENCES

Ahn C. P. et al., 2014, ApJS, 211, 17
Akritas M. G., Bershady M. A., 1996, ApJ, 470, 706
Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R. et al., 2003, ApJ, 592, 819
Böhringer H. et al., 2000, ApJS, 129, 435
Böhringer H. et al., 2004, A&A, 425, 367
Cai Y.-C., Angulo R. E., Baugh C. M., Cole S., Frenk C. S., Jenkins A., 2009, MNRAS, 395, 1185
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Clerc N., Sadibekova T., Pierre M., Pacaud F., Le Fèvre J.-P., Adami C., Altieri B., Valtchanov I., 2012, MNRAS, 423, 3561
Gladders M. D., Lopez-Cruz O., Yee H. K. C., Kodama T., 1998, ApJ, 501, 571
Hansen S. M., Sheldon E. S., Wechsler R. H., Koester B. P., 2009, ApJ, 699, 1333
Hao J. et al., 2009, ApJ, 702, 745
Hao J. et al., 2010, ApJS, 191, 254
Heymans C. et al., 2012, MNRAS, 427, 146
High F. W. et al., 2010, ApJ, 723, 1736
Hood R. J., 2014, PhD thesis, Univ. Edinburgh
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
Kloster D., Rines K., Svoboda B. E., Arnold R. L., Welch T. J., Finn R. A., Vikhlinin A., 2011, American Astronomical Society Meeting Abstracts, 217. 149.12
Koester B. P. et al., 2007, ApJ, 660, 221
Kraft D., 1988, A Software Package for Sequential Quadratic Programming. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln, Forschungsbericht, Wiss. Berichtswesen d. DFVLR
Lloyd-Davies E. J. et al., 2011, MNRAS, 418, 14
Magnier E. A. et al., 2013, ApJS, 205, 20
Mehrtens N. et al., 2012, MNRAS, 423, 1024
Miller C. J. et al., 2005, AJ, 130, 968
Murphy D. N. A., Geach J. E., Bower R. G., 2012, MNRAS, 420, 1861
Peebles P. J. E., 1980, The Large-scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Planck Collaboration XXIX, 2014, A&A, 571, A29
Reichardt C. L. et al., 2013, ApJ, 763, 127
Romer A. K., Viana P. T. P., Liddle A. R., Mann R. G., 2001, ApJ, 547, 594
Rozo E., Rykoff E. S., 2014, ApJ, 783, 80
Rozo E. et al., 2010, ApJ, 708, 645
Rykoff E. S. et al., 2008, ApJ, 675, 1106
Rykoff E. S. et al., 2014, ApJ, 785, 104
Shanks T., Metcalfe N., 2012, Science from the Next Generation Imaging and Spectroscopic Surveys
Shanks T. et al., 2015, MNRAS, 451, 4238
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Song J., Mohr J. J., Barkhouse W. A., Warren M. S., Dolag K., Rude C., 2012, ApJ, 747, 58
Tinker J. L. et al., 2012, ApJ, 745, 16
Tojeiro R., Percival W. J., Heavens A. F., Jimenez R., 2011, MNRAS, 413, 434
Voit G. M., 2005, Rev. Mod. Phys., 77, 207

APPENDIX A: GMPhoRCC OUTPUTS

A full list of the flags, SDSS quality marker calculations and outputs. Tables A1, A2 and A3 show the various GMPhoRCC flags relating to multi-modal distributions, redshift and richness estimates, and detection respectively. Tables A4 and A5 denote the determination of quality and how the various data subsets are produced. Finally Table A6 shows the full list of outputs produced by GMPhoRCC.

Table A1. A list of the GMPhoRCC flags relating to the detection of multimodal distributions. Unfortunately, as ∼70 per cent of clusters have exhibited multimodal distributions, these are insufficient to identify catastrophic failures.

Name	Value	Description
MULTI_BANDS	0x00000000001	Multiple bands considered
POOR_CMR_FIT	0x00000000002	Poor CMR fit, $\chi^2 > 5$
INAPPROPRIATE_Z	0x00000000004	One or more of the redshift estimates not appropriate for the RS band
MULTI_INITIAL	0x00000000010	Multiple peaks in initial z distribution – relative heights <5
MULTI_INITIAL_AMBIGUOUS	0x00000000020	Multiple peaks in initial z distribution – relative heights <2
MULTI_INITIAL_CLOSE	0x00000000040	Primary and secondary peak within 0.1 of each other
MULTI_COLOUR	0x00000000100	Multiple peaks in the colour distribution – relative heights <5
MULTI_COLOUR_AMBIGUOUS	0x00000000200	Multiple peaks in the colour distribution – relative heights <2
MULTI_COLOUR_CLOSE	0x00000000400	Primary and secondary peak within 0.2 mag of each other
MULTI_ZRS	0x00000001000	Multiple peaks in the RS z fit – relative heights <5
MULTI_ZRS_AMBIGUOUS	0x00000002000	Multiple peaks in the RS z fit – relative heights <2
MULTI_ZRS_CLOSE	0x00000004000	Primary and secondary peak within 0.1 of each other

Table A2. A list of GMPhoRCC flags indicating issues with the redshift or richness estimates that give the strongest indication that an estimate may be erroneous. Here $\Delta z_{\rm cp}$ and $\Delta z_{\rm cs}$ are, respectively, photometric and spectroscopic redshift consistency bounds, where, for the SDSS DR10, $\Delta z_{\rm cp} = 0.035$ and $\Delta z_{\rm cs} = 0.025$.

Name	Value	Description
SPARCE_INITIAL	0x00000010000	<5 galaxies found in the cluster region for the initial z fit
SPARCE_COLOUR	0x00000020000	<5 galaxies found in the cluster region for the colour fit
SPARCE_ZRS	0x00000040000	<5 galaxies found in the cluster region for the RS z fit
LOW_RICHNESS_N200_3	0x00000100000	Low counting richness recovered, $n_{200-\rm count} < 3$
INCONSISTENT_Z_PHOT	0x00000200000	$z_{\rm RS}$ and $z_{\rm BCG-phot}$ are inconsistent with each other, $\lvert z_{\rm RS} - z_{\rm BCG-phot}\rvert > \Delta z_{\rm cp}$
INCONSISTENT_Z_SPEC	0x00000400000	$z_{\rm RS}$ and $z_{\rm BCG-spec}$ are inconsistent with each other, $\lvert z_{\rm RS} - z_{\rm BCG-spec}\rvert > \Delta z_{\rm cs}$
LOW_RICHNESS_N200_1	0x00001000000	Low counting richness recovered, $n_{200-\rm count} < 1$
INCONSISTENT_Z_PHOT_2X	0x00002000000	$z_{\rm RS}$ and $z_{\rm BCG-phot}$ are inconsistent with each other, $\lvert z_{\rm RS} - z_{\rm BCG-phot}\rvert > 2\Delta z_{\rm cp}$
INCONSISTENT_Z_SPEC_2X	0x00004000000	$z_{\rm RS}$ and $z_{\rm BCG-spec}$ are inconsistent with each other, $\lvert z_{\rm RS} - z_{\rm BCG-spec}\rvert > 2\Delta z_{\rm cs}$

Table A3. A list of GMPhoRCC flags relating to the non-detection of a cluster overdensity.

Name	Value	Description
CLUSTER_INSIDE_MASK_0_5_MPC	0x00010000000	Empty apertures found inside $r < 0.5\,h^{-1}$ Mpc of the cluster centre
CLUSTER_INSIDE_MASK_R200	0x00020000000	Empty apertures found inside $r < r_{200}$ of the cluster centre
CLUSTER_INSIDE_MASK_5_AM	0x00040000000	Empty apertures found inside $r < 5$ arcmin of the cluster centre
NO_OVERDENSITY_INITIAL	0x00100000000	No overdensity found in the cluster region for the initial z fit
NO_OVERDENSITY_COLOUR	0x00200000000	No overdensity found in the cluster region for the colour fit
NO_OVERDENSITY_ZRS	0x00400000000	No overdensity found in the cluster region for the RS z fit
NO_CLUSTER_INITIAL	0x01000000000	0 galaxies found in the cluster region for the initial z fit
NO_CLUSTER_COLOUR	0x02000000000	0 galaxies found in the cluster region for the colour fit
NO_CLUSTER_ZRS	0x04000000000	0 galaxies found in the cluster region for the RS z fit
NO_DETECTION_REDSHIFT	0x10000000000	No detection in redshift module
NO_DETECTION_RICHNESS_NGALS	0x20000000000	No detection in richness, $n_{\rm gals} < 0$ for both counting and luminosity
NO_DETECTION_RICHNESS_N200	0x40000000000	No detection in richness, $n_{200} < 0$ for both counting and luminosity
NO_COVERAGE	0x80000000000	No optical coverage

Table A4. A list of the quality markers assigned to clusters based on the GMPhoRCC flags.

quality	Flags value	Description
−1	0x80000000000 ≤ flags	No optical coverage
0	0x01000000000 ≤ flags < 0x80000000000	No characterization found
1	0x00001000000 ≤ flags < 0x01000000000	$n_{200} < 1$, large redshift inconsistencies, field-masking issues
2	0x00000100000 ≤ flags < 0x00001000000	$n_{200} < 3$, small redshift inconsistencies
3	flags < 0x00000100000	Clean

Table A5. A list of cluster subsets based on the GMPhoRCC quality markers used to remove potentially erroneous characterizations.

Subset	*quality*	Description
Detection	≥1	All clusters considered to have been detected, i.e. estimates were found for both redshift and richness
Mid	≥2	An intermediate subset removing the worst outliers, i.e. removing clusters with very low richness or large discrepancies between redshift estimates
Clean	≥3	The cleanest subset removing the majority of outliers i.e. removing cluster with low richness and discrepancies between redshift estimates

Table A6. A list of outputs generated by GMPhoRCC using SDSS DR10 photometry for both the primary and secondary cluster. In the case where GMPhoRCC was unable to determine a property, a default value of −1 is used. While the redshift labels are specific to SDSS DR10, these can be adjusted to match any optical input.

Name	Description
band	The red sequence colour used to detect the cluster. $0 = g - r$, $1 = r - i$, $2 = i - z$.
size	The angular radius in arcmin of the initial aperture used to model the red sequence.
z_initial	The position of the peak of the initial redshift distribution.
z_initial_peak	The size of the peak of the initial redshift distribution (galaxies arcmin^{-2}).
z_initial_errorm(p)	1σ error on the peak in the 'minus' ('positive') direction.
z_initial_info	A flag based on how the error was determined. 0 = no issues, 1 = extrapolation needed due to multiple peaks.
rs_colour_(peak,error,info)	The position, amplitude and error of the peak in the initial red-sequence colour distribution.
z_rs_(peak,error,info)	The position, amplitude and error of the peak in the red-sequence photometric redshift distribution.
BCG_objID	The objID of the BCG.
BCG_dis	The angular distance in arcmin of the BCG from the cluster centre.
z_BCG_best_(err)	The best redshift with error of the BCG: spectra if available, photometric otherwise.
z_BCG_phot_(err)	The photometric redshift of the BCG
z_BCG_spec_(err)	The spectroscopic redshift of the BCG.
z_gals_spec_(err)	A spectroscopic cluster redshift based on the spectra of the five brightest galaxies on the red sequence.
z_gals_spec_no	The number of galaxies available with spectra.
cmr_grad_(err)	The gradient of the red sequence CMR.
cmr_intercept_(err)	The intercept of the red sequence CMR.
cmr_width	The intrinsic width the red sequence CMR.
ngals_count_(err)	$n_{\mathrm{gals-count}}$, the background-subtracted number of galaxies inside 0.5 h^{-1} Mpc on the red sequence with Poissonian error.
ngals_lum_(err)	$n_{\mathrm{gals-lum}}$, the background-subtracted richness inside 0.5 h^{-1} Mpc from integrating an LF with Poissonian error.
r200_mpch-1	r_{200} in h^{-1} Mpc.
n200_count_(err)	$n_{200-\mathrm{count}}$, the background-subtracted number of galaxies inside r_{200} on the red sequence with error.
n200_lum_(err)	$n_{200-\mathrm{lum}}$, the background-subtracted richness inside r_{200} from integrating an LF with error.
flags	A hexadecimal combination of the GMPhoRCC flags.
quality	The quality marker based on the GMPhoRCC flags.

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