Photometric redshifts with Quasi Newton Algorithm (MLPQNA).
Results in the PHAT1 contest.

S. Cavuoti\textsuperscript{1,2}, M. Brescia\textsuperscript{2,1}, G. Longo\textsuperscript{1,2,3}, and A. Mercurio\textsuperscript{2}

\textsuperscript{1} Department of Physics, Federico II University, via Cinthia 6, I-80126 Napoli, Italy e-mail: cavuoti@na.infn.it
\textsuperscript{2} INAF - Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy
\textsuperscript{3} Visiting associate - Department of Astronomy, California Institute of Technology, CA 90125, USA

Received xxxx; accepted xxxxxx

ABSTRACT

Context. Since the advent of modern multiband digital sky surveys, photometric redshifts (photo-z’s) have become relevant if not crucial to many fields of observational cosmology, from the characterization of cosmic structures, to weak and strong lensing.

Aims. We describe an application to an astrophysical context, namely the evaluation of photometric redshifts, of MLPQNA, a machine learning method based on Quasi Newton Algorithm.

Methods. Empirical methods for photo-z’s evaluation are based on the interpolation of a priori knowledge (spectroscopic redshifts or SED templates) and represent an ideal test ground for neural networks based methods. The MultiLayer Perceptron with Quasi Newton learning rule (MLPQNA) described here is a computing effective implementation of Neural Networks and is offered to the community through the DAMEWARE (Data Mining & Exploration Web Application RSource) infrastructure.

Results. The PHAT contest (Hildebrandt et al. 2010) provides a standard dataset to test old and new methods for photometric redshift evaluation and with a set of statistical indicators which allow a straightforward comparison among different methods. When applied to the PHAT1 dataset, MLPQNA obtains very competitive accuracies in terms of bias, RMS (Root Mean Square) and outlier percentage, scoring as the second most effective empirical method among those which have so far participated to the contest.

Key words. techniques: photometric - galaxies: distances and redshifts - galaxies: photometry - cosmology: observations - methods: data analysis

1. Introduction

Estimating redshifts of celestial objects is one of the most pressing technological issues in the observational astronomy and, since the advent of modern multiband digital sky surveys, photometric redshifts (photo-z’s) have become fundamental when it is necessary to know the distances of million of objects over large cosmological volumes. Photo-z’s provide redshift estimates for objects fainter than the spectroscopic limit, and result much more efficient in terms of the number of objects per telescope time with respect to spectroscopic ones (spec-z). For these reasons, after the advent of modern panchromatic digital surveys, photo-z’s have become crucial. For instance, they are essential in constraining dark matter and dark energy studies by means of weak gravitational lensing, for the identification of galaxy clusters and groups (e.g. Capozzi et al. 2009), for type Ia supernovae, and to study the mass function of galaxy clusters (Albrecht et al., 2006; Peacock et al., 2006; Keiichi et al., 2012). The need for fast and reliable methods for photo-z evaluation will become even greater in the near future for the exploitation of ongoing and planned surveys. In fact, future large field public imaging projects, like KiDS (Kilo-Degree Survey\textsuperscript{1}), DES (Dark Energy Survey\textsuperscript{2}), LSST (Large Synoptic Survey Telescope\textsuperscript{3}), and Euclid (Euclid Red Book 2011), require extremely accurate photo-z’s to obtain accurate measurements that does not compromise the surveys scientific goals. This explains the very rapid growth in the number of methods which can be more or less effectively used to derive photo-z’s estimates, and the efforts made to better understand and characterize their biases and systematics. The possibility to achieve a very low level of residual systematics (Huterer et al., 2006; D’Abrusco et al., 2007; Laurino et al., 2011), is in fact strongly influenced by many factors: the observing strategy, the accuracy of the photometric calibration, the different point-spread-function in different bands, the adopted de-reddening procedures, etc. The evaluation of photo-z’s is made possible by the existence of a rather complex correlation existing between the fluxes as measured in broad band photometry, the morphological types of the galaxies and their distance. The search for such correlation (a non-linear mapping between the photometric parameter space and the redshift values) is particularly suited for data mining methods. Existing methods can be broadly divided into two large groups: theoretical and empirical methods. Theoretical methods use templates, like libraries of either observed galaxy spectra or model Spectral Energy Distributions (SEDs). These templates can be shifted to any redshift and then convolved with the transmission curves of the filters used in the photometric survey to create the template set for the redshift estimators. Empirical methods use instead a subsample of the photometric survey with spectroscopically-measured redshifts as a ‘training set’ for the redshift estimators. This subsample describes empirically the redshift distribution in magnitude and color space and it is used to calibrate this relation. Both methods make use of training sets as bases for the redshift estimating routines.

The variety of methods and approaches (for a review see Koo 1999) and their application to different types of datasets, as well

\textsuperscript{1} http://www.astro-wise.org/projects/KiDS/
\textsuperscript{2} http://www.darkenergysurvey.org/
\textsuperscript{3} http://www.lsst.org/lsst/
as the adoption of different and often not comparable statistical indicators, make it difficult to evaluate and compare performances in an unambiguous and homogeneous way. Useful but limited in scope blind tests of photo-z’s have been performed in Hogg et al. (1998) on spectroscopic data from the Keck telescope on the Hubble Deep Field (HDF), in Hildebrandt et al. (2008) on spectroscopic data from the VIMOS VLT Deep Survey (VVDS; Le Févre et al. 2004) and the FORS Deep Field (FDF; Noll et al. 2004, and in Abdalla et al. 2008) on the sample of Luminous Red Galaxies from the SDSS-DR6.

A significant advance in comparing different methods was introduced by Hildebrandt and collaborators (Hildebrandt et al. 2010), with the so called PHAT (PHoto-z Accuracy Testing) contest, which adopts a black-box approach which is typical of benchmarking. Instead of insisting on the subtleties of the data structure, they performed a homogeneous comparison of the performances concentrating the analysis on the last link in the chain: the photo-z’s methods themselves.

As pointed out by the authors, in fact, “it is clear that the two regimes - data and method - cannot be separated cleanly because there are connections between the two. For example, it is highly likely that one method of photo-z estimation will perform better than a second method on one particular dataset while the situation may well be reversed on a different data set.” (cf. Hildebrandt et al. 2010).

The present work follows the same path, by having as its aim the testing and probing of the accuracy of the Quasi Newton based Neural Model (MLPQNA) for the derivation of photometric redshifts. The application of MLPQNA to the photometric redshift estimation of QSO will be presented in Brescia et al. (in preparation).

In Sect. 2 we shortly describe the PHAT contest and the PHAT1 data made available to the contestants and used for the present work. In Sect. 3 we describe the MLPQNA method which was implemented by us and used for the contest, while in Sect. 4 we describe the experiments performed and, in Sect. 5 we present the results derived for us by the PHAT team. Summary and conclusions are wrapped up in Sect. 6.

2. The PHAT dataset

First results from the PHAT contest were presented in Hildebrandt et al. (2010), but the contest still continues at the project’s web site. PHAT provides a standardized test environment which consists of simulated and observed photometric catalogues complemented with additional materials like filter curves convolved with transmission curves, SED templates, and training sets. However, the subsets used to evaluate the performances are still kept secret in order to provide a more reliable comparison of the various methods. Two different datasets are available (see Hildebrandt et al. 2010 for more details).

The first one, indicated as PHAT0, is based on a very limited template set and a long wavelength baseline (from UV to mid-IR). It is composed by a noise-free catalogue with accurate synthetic colors and a catalogue with a low level of additional noise. PHAT0 represents a real case to test the most basic elements of photo-z estimation and to identify possible low-level discrepancies between the methods.

The second one, which is the one used in the present work, is the PHAT1 dataset, which is based on real data originating from the Great Observatories Origins Deep Survey Northern field (GOODS-North; Giavalisco et al. 2004). According to Hildebrandt et al. (2010), it represents a much more complex environment to test methods to estimate photo-z’s, pushing codes to their limits and revealing more systematic difficulties. Both PHAT test datasets are made publicly available through the PHAT website4 while in Hildebrandt et al. (2010) there is a detailed description of the statistical indicators which were used for the comparison of the results provided by the 21 participants who have so far participated by submitting results obtained with 17 different photo-z codes.

The PHAT1 dataset consists of photometric observations, both from ground and space instruments, presented in Giavalisco et al. (2004), complemented with additional data in other bands derived from Capak et al. (2004). The final dataset covers the full UV-IR range and includes 18 bands: U (from KPNO), B, V, R, I, Z (from SUBARU), F435W, F606W, F775W, F850LP (from HST-ACS), J, H (from ULBCAM), HK (from QUIRC), K (from WIRC) and 3.6, 4.5, 5.8 and 8.0 μ (from IRAC Spitzer).

The photometric dataset was then cross correlated with spectroscopic data from Cowie et al. (2004); Wirth et al. (2004); Treu et al. (2005), and Reddy et al. (2006). Therefore, the final PHAT1 dataset consists of 1984 objects with 18-band photometry and accurate spectroscopic redshifts. In the publicly available dataset a little more than one quarter (515) of the objects comes with spectroscopic redshifts and can be used as Knowledge Base (KB) for training empirical methods. While it is clear that the limited amount of objects in the KB is not sufficient to ensure the best performances of most empirical methods, the fact that all methods must cope with similar difficulties makes the comparison consistent.

3. The MLPQNA regression model

MLPQNA stands for the traditional neural network model named Multi Layer Perceptron (MLP; cf. Bishop 2006) implemented with a Quasi Newton Algorithm (QNA) as learning rule. This particular implementation of the traditional MLP’s has already been described in Brescia et al. (2012a), and we refer to that paper for a more detailed description in the classification problem context. MLPQNA is made available to the community through the DAMEWARE (DAta Mining & Exploration Web Application RESource; Brescia et al. 2011, 2012a,b). In the text we also provide the details and the parameters settings for the best performing MLPQNA model so that anyone can easily reproduce the results using the Web Application. User’s manuals are available on the DAMEWARE web site5. A complete mathematical description of the MLPQNA model is available on the DAME web site6. Feed-forward neural networks provide a general framework for representing nonlinear functional mappings between a set of input variables and a set of output variables (Bishop, 2006). One can achieve this goal by representing the nonlinear function of many variables by a composition of non-linear activation functions of one variable:

\[ y_k = \sum_{j=0}^{M} w_{kj}^{(2)} g \left( \sum_{i=0}^{d} w_{ij}^{(1)} x_i \right) \]

A Multi-Layer Perceptron may be represented by a graph: the input layer (\(x_i\)) is made of a number of perceptrons equal to

---

4 http://www.astro.caltech.edu/twiki-phat/bin/view/Main/GoodsNorth
5 http://dame.dsf.unina.it/beta_info.html
6 http://dame.dsf.unina.it/machine_learning.html#mlpqna
the number of input variables \((d)\); the output layer, on the other hand, will have as many neurons as the output variables \((K)\). The network may have an arbitrary number of hidden layers (in most cases one) which in turn may have an arbitrary number of perceptrons \((M)\). In a fully connected feed-forward network each node of a layer is connected to all the nodes in the adjacent layers. Each connection is represented by an adaptive weight which represents the strength of the synaptic connection between neurons \((w_{kj}^{0})\). The response of each perceptron to the inputs is represented by a non-linear function \(g\), referred to as the activation function. Notice that the above equation assumes a linear activation function for neurons in the output layer. We shall refer to the topology of an MLP and to the weights matrix of its connections as to the model. In order to find the model that best fits the data, one has to provide the network with a set of examples: the training phase thus requires the KB, i.e. the training set. The learning rule of our MLP is the Quasi Newton Algorithm (QNA). In general these models are variable metric methods used to find local maxima and minima of functions (Davidon, 1968) and, in the case of MLPs they can be used to find the stationary (i.e. the zero gradient) point of the learning function. The Newton method is the general basis for a whole family of so called Quasi Newton methods.

Among these methods there is also the L-BFGS algorithm (cf. Byrd et al. 1994; Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970) which has been implemented in MLPQNA.

In practice, QNA is an optimization of learning rule based on a statistical approximation of the Hessian by cyclic gradient calculation which, as already mentioned, is at the base of the classical Back Propagation (BP; Bishop 2006) method.

The traditional Newton method uses the Hessian of a function to find the stationary point of a quadratic form. The step of the method is defined as the product of an inverse Hessian matrix and a function gradient. If the function is a positive definite quadratic form, the minimum can be reached in just one step, while in case of an indefinite quadratic form (which has no minimum), we will reach either the maximum or a saddle point. To solve this problem, Quasi Newton methods proceed with a positive definite Hessian approximation. So far, if the Hessian is positive definite, we make the step using the Newton method. If, instead it is indefinite, we first modify it to make it positive definite, and then perform a step using the Newton method, which is always calculated in the direction of the function decrement.

The L-BFGS algorithm tries to perform a step using the Newton method. If it does not lead to a function value decreasing, it lessens the step length to find a smaller function value. The Hessian of a function is not always available and in many cases it is far too complex to be computed. More often we can only calculate the function gradient which can be used to derive the Hessian via N consequent gradient calculations.

The algorithm L-BFGS does not generate the Hessian, but rather its inverse matrix, thus saving computing time by avoiding the inversion of the Hessian. By using a local square approximation of the error function, we can obtain an expression for the minimum position. The gradient in every point \(w\) is in fact given by:

\[
\nabla E = H \times (w - w^*)
\]

where \(w\) corresponds to the minimum of the error function, which satisfies the condition:

\[
w^* = w - H^{-1} \times \nabla E
\]

The vector \(-H^{-1} \times \nabla E\) is known as Newton direction and it is the base for a variety of optimization strategies, such as for instance the QNA which instead of calculating the H matrix and then its inverse, uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which are more and more accurate approximations of \(H^{-1}\).

During the exploration of the parameter space, in order to find the minimum error direction, QNA starts in the wrong direction. This direction is chosen because at the first step the method has to follow the error gradient and so it takes the direction of steepest descent. However, in subsequent steps, it incorporates information from the gradient. By using the second derivatives, QNA is able to avoid local minima and to follow more precisely the error function trend, revealing a “natural” capability to find the absolute minimum error of the optimization problem.

However, this last feature could be a downside of the model, especially when the signal-to-noise ratio of data is very poor. But with “clean” data, such as in presence of high quality spectroscopic redshifts, used for model training, the QNA performances result extremely precise.

4. The experiment Workflow

MLPQNA method was applied by following the standard Machine Learning (ML) workflow (Bishop 2006), which is here summarized: i) extraction of the KB by using the 515 available spectroscopic redshifts; ii) determination of the “optimal” model parameter setup, including pruning of data features and training/test with the available KB; iii) application of the tuned model to measure photometric redshifts on the whole PHAT1 dataset of \(N=1984\) objects, by including also the re-training on the extended KB. We also follow the rules of the PHAT1 contest, applying the new method in two different ways, first to the whole set of 18 bands and then to the 14 non-IRAC bands only.

Details of the workflow are described in the following subsections.

4.1. Extraction of KB

For supervised methods it is common praxis to split the KB in at least three disjoint subsets: one (training set) to be used for training purposes, i.e. to teach the method how to perform the regression; the second one (validation set) to check against loss of generalization capabilities (also known as overfitting); and the third one (test set) to be used to evaluate the performances of the model. As a rule of thumb, these sets should be populated with 60%, 20% and 20% of the objects in the KB, respectively. In order to ensure a proper coverage of the Parameter Space (PS), objects in the KB are split among the three datasets by random extraction and usually this process is iterated several times in order to minimize biases introduced by fluctuations in the coverage of the PS.

In the case of MLPQNA described here, we used cross-validation (cf. Bishop 2006) in order to minimize the size of the validation set (\(\sim 10\%\)). Training and validation were therefore performed together using as training set \(\sim 80\%\) of the objects and as test set the remaining \(\sim 20\%\) (in practice 400 records in the training set and 115 in the test set). In order to ensure a proper coverage of the PS we checked that the randomly extracted populations had a spec-z distribution compatible with that of the whole KB. The automatized process of the cross-validation was done by performing 10 different training runs with the following procedure: (i) we split the training set into 10 random subsets,
each one composed by 10% of the dataset; (ii) at each training run we apply the 90% of the dataset for training and the excluded 10% for validation. This procedure is able to avoid overfitting on the training set (Bishop 2006).

4.2. Model optimization

As known, supervised machine learning models are powerful methods able to learn from training data the hidden correlation between input and output features. Of course, their generalization and prediction capabilities strongly depend by the intrinsic quality of data (signal-to-noise ratio), level of correlation inside of the PS and by the amount of missing data present in the dataset. Among the factors which affect performances, the most relevant is the fact that most ML methods are quite sensitive about the presence of Not a Number (NaN) in the dataset to be analysed. This is especially relevant in astronomical dataset where NaN’s may either be non detections (i.e. objects which in a given band are observed but non detected since they are below the detection threshold) or related to patches of the sky which have not been observed. The presence of features with a large fraction of NaN’s can seriously affect the performances of a given model and lower the accuracy or the generalization capabilities of a specific model. It is therefore a good praxis to analyze the performances of a specific model in presence of features with large fractions of NaN’s. This procedure is strictly related to the so called feature selection or ”pruning of the features” phase which consists in evaluating the significance of individual features to the solution of a specific problem. In what follows we shall shortly discuss the outcome of the ”pruning” performed on the PHAT1 dataset.

4.2.1. Pruning of features

It is also necessary to underline that especially in presence of small datasets there is a need for a compromise: while on the one hand it is necessary to minimize the effects of NaN’s, on the other it is not possible to simply remove each record containing a NaN, because otherwise too much information would be lost.

In table 1 we list the percentage of NaN’s in each photometric band both in the training and full datasets. Poor features, namely the fluxes in the K and m5.8 bands were not used for the subsequent analysis.

The pruning was performed separately on the two PHAT1 datasets (18-bands and 14-bands), respectively. A total of 37 experiments was run on the two datasets: the various experiments differing in the groups of features removed. We started by considering all features (bands), removing the two worst bands, for instance K and m5.8, which outlier quantity was over the 15% of patterns. Then a series of experiments was performed by removing one band at a time, by considering the NaN’s percentage shown in table 1.

4.2.2. Performance metrics

The performances of the various experiments were evaluated (as done in the PHAT contest) in terms of:

- scatter: is the RMS of $\Delta z$
- bias: is the mean of $\Delta z$
- fraction of outliers: where outliers are defined by the condition: $|\Delta z| > 0.15$

Where:

$$\Delta z \equiv \frac{z_{\text{spec}} - z_{\text{phot}}}{1 + z_{\text{spec}}}$$

At the end of this process, we obtained the best results, reported in table 2.

4.3. Application to the PHAT1 dataset

Once the model optimization described in table 2 had been determined, the MLPQNA was re-trained on the whole KB (515 objects) and applied to the whole PHAT1 dataset (1984 objects), which was then submitted to the PHAT contest for final evaluation (see below).

Details of the experiments can be found at the DAME web site, while the parameter settings and the results for the best models are summarised in table 3.

5. The PHAT1 results and comparison with other models

The statistics obtained by the PHAT Team\(^8\) by analysing the photometric redshifts obtained by MLPQNA, are reported in table 3.

The most significative results can be summarized as it follows:

i) 18-band experiment: 324 outliers with $|\Delta z| > 0.15$, corresponding to a relative fraction of 16.33%. For the remaining 1660 objects bias and rms are: 0.000604251 ± 0.0562278

ii) 14-band experiment: 384 outliers with $|\Delta z| > 0.15$, corresponding to a relative fraction of 19.35%. 1600 objects with bias and variance 0.00277721 ± 0.0626341.

A more detailed characterization of the results can be found in the first line of parts A, B and C in the table 3, while figure 1 gives the scatter plots (spec-z’s vs photo-z’s) for the 18 and 14 bands, respectively.

In order to compare our results with other models, we also report in table 3 the statistical indicators for the other empirical methods which competed in the PHAT1 contest. The methods are:

- \textbf{AN-e}: an empirical photo-z code based on artificial neural networks (Collister & Lahav, 2004);
- \textbf{EC-e}: a subclass of kernel regression methods; which mimics a template-based technique with the main difference that an empirical dataset is used in place of the template grid (Wolf, 2009);
- \textbf{PO-e}: a ”nearest neighbour” empirical photo-z method based on a polynomial fit so that the galaxy redshift is expressed as the sum of its magnitudes and colours (Li & Yee, 2008);
- \textbf{RT-e}: based on Random Forests which are an empirical, non-parametric regression technique (Carliles et al., 2010).

More details can be found in the quoted references and in Hildebrandt et al. (2010).

For each of the datasets (18 and 14 bands), statistics in Table 3 refers to several regimes: the first one (A) defines as outliers all objects having $|\Delta z| > 0.15$ and it is divided into two subsections:

\(^8\)http://www.astro.caltech.edu/twiki-phat/bin/view/Main/GoodsNorthResults#Cavuoti_Stefano_et_al_neural_net

\(^7\)http://dame.dsf.unina.it/dame_photoz.html
Table 1. The percentages of Not a Number in the whole dataset (col 3), with 1984 objects and in the trainset (col 4), with 515 objects, for each band. The last column reports the absolute differences between the two NaN percentages. As shown this difference remains always under 3%, demonstrating that the two datasets are congruent in terms of NaN quantity.

<table>
<thead>
<tr>
<th>BAND</th>
<th>Dataset Column ID</th>
<th>% NaN in whole set</th>
<th>% NaN in Training</th>
<th>NaN % Absolute Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>m5.8</td>
<td>17</td>
<td>19.35</td>
<td>17.28</td>
<td>2.07</td>
</tr>
<tr>
<td>K</td>
<td>14</td>
<td>17.14</td>
<td>18.64</td>
<td>1.5</td>
</tr>
<tr>
<td>HK</td>
<td>13</td>
<td>5.65</td>
<td>6.21</td>
<td>0.57</td>
</tr>
<tr>
<td>m8</td>
<td>18</td>
<td>3.48</td>
<td>3.5</td>
<td>0.02</td>
</tr>
<tr>
<td>F435W</td>
<td>7</td>
<td>2.67</td>
<td>1.75</td>
<td>0.92</td>
</tr>
<tr>
<td>H</td>
<td>12</td>
<td>2.37</td>
<td>2.52</td>
<td>0.16</td>
</tr>
<tr>
<td>J</td>
<td>11</td>
<td>1.16</td>
<td>1.55</td>
<td>0.39</td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>1.01</td>
<td>1.17</td>
<td>0.16</td>
</tr>
<tr>
<td>R</td>
<td>4</td>
<td>0.15</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0.1</td>
<td>0.19</td>
<td>0.09</td>
</tr>
<tr>
<td>V</td>
<td>3</td>
<td>0.05</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>F606W</td>
<td>8</td>
<td>0.05</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>m3.6</td>
<td>15</td>
<td>0.05</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>I</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Z</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F775W</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F850LP</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>m4.5</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Description of the best experiments for the 18 bands (Exp. n. 37) and the 14 bands datasets (Exp. n. 26). Column 1: sequential experiment identification code; column 2: features not used in the experiment; columns 3-4: number of input (features) and hidden neurons; column 5–9: parameters of the MLPQNA used during the experiment; column 10: scatter error evaluated as described in the text; column 11: fraction of outliers; column 12: bias.

<table>
<thead>
<tr>
<th>exp. n</th>
<th>missing features</th>
<th>feat.</th>
<th>hid.</th>
<th>step</th>
<th>res.</th>
<th>dec.</th>
<th>MxIt</th>
<th>CV</th>
<th>scatter</th>
<th>outliers%</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>m5.8, K, HK, m8</td>
<td>14</td>
<td>29</td>
<td>0.0001</td>
<td>30</td>
<td>0.1</td>
<td>3000</td>
<td>10</td>
<td>0.057</td>
<td>22.61%</td>
<td>-0.0077</td>
</tr>
<tr>
<td>26</td>
<td>m5.8, K, m3.6, m4.5, HK, m8</td>
<td>12</td>
<td>25</td>
<td>0.0001</td>
<td>30</td>
<td>0.1</td>
<td>3000</td>
<td>10</td>
<td>0.062</td>
<td>17.39%</td>
<td>0.0078</td>
</tr>
</tbody>
</table>

Fig. 1. Results obtained using the analysis described in this paper by the PHAT contest group. In the (a) panel are plotted the photometric vs. spectroscopic redshifts for the whole dataset using 10 photometric bands (Experiment 37). In panel (b) the same but using only 14 photometric bands (Experiment 26).

By analyzing the MLPQNA performance in the different regimes:

- **All objects**: in the 18 bands experiment, QNA scores the best results in term of bias, and gives comparable results with PO-e in terms of scatter and number of outliers. In fact, while in Part A the scatter is slightly larger than those of PO-e method (0.052 against 0.056), the number of outliers is lower
6. Summary and Conclusions

The MLPQNA method described in the previous sections was applied on the whole PHAT1 dataset of N=1984 objects Hildebrandt et al. (2010) after an optimization of the model performed by using as a training set the 515 available spectroscopic redshifts.

The statistics obtained by the PHAT Team by analyzing the photometric redshifts derived with MLPQNA, and the comparison with other empirical models are reported in Table 3.

From a quick inspection of table 3, it descends that it does not exist an empirical method which can be regarded as the best in terms of all the indicators (e.g. bias, scatter and number of outliers) and that EC-e, PO-e and MLPQNA produce comparable results. The MLPQNA method, on average, gives the best result in terms of bias at any regime.

For what the scatter is concerned, by considering the dataset with 18 bands reported in Parts A and B of table 3, MLPQNA obtains results comparable with the PO-e method. In fact, in Part A PO-e’s scatter is better than MLPQNA’s, but with a larger number of outliers; while the trend is reversed in Part B. In the other cases both the scatter and number of outliers are slightly worse than PO-e and EC-e methods.

In general, MLPQNA seems to have better generalization capabilities than most other empirical methods especially in presence of underpopulated regions of the KB. In fact, ~ 500 objects with spectroscopic redshifts spread over such a large redshift interval are by far not sufficient to train most other empirical codes on the data. This was also pointed out also by Hildebrandt et al. (2010), who noticed that the high fraction of outliers produced by empirical methods is on average higher than what is currently found in literature (~ 7.5%) and explained it as an effect of the small size of the training sample, which maps poorly the very large range in redshifts and does not include a large enough number of objects with peculiar SED’s.

In this respect we wish to stress that as it has already been shown in another application (cf. Brescia et al. 2012a) and will be more extensively discussed in a forthcoming paper, MLPQNA enjoys the very rare prerogative of being able to ob-
tain good performances also when the KB is small and thus undersampled (Brescia et al. in preparation).

Acknowledgements.
The authors wish to thank all present and former members of the DAME collaboration and in particular R. D’Abrusco for useful discussions. DAME has been funded through several grants which are mentioned on the collaboration web site. We acknowledge the use of TOPCAT and of other Virtual Observatory tools during some steps of the procedure. GL wishes to thank the California Institute of Technology for the kind hospitality. AM and MB wish to thank the financial support of PRIN-INAF 2010, “Architecture and Tomography of Galaxy Clusters”.

References

Brescia, M., Longo, G., Castellani, M. et al., 2012b, Mem. SAIt Suppl, 19, 324
Davidon, W.C., 1970, Mathematics of Computation, 24, 23
Goldfarb, D., 1970, Mathematics of Computation, 24, 23
Hildebrandt, H., Wolf, C., & Benitez, N., 2008, AA, 480, 703