Mon. Not. R. Astron. Soc. **367,** 879–904 (2006) doi:10.1111/j.1365-2966.2006.10024.x

Identification and redshift determination of quasi-stellar objects with medium-band photometry: application to *Gaia*

J.-F. Claeskens,[1⋆] A. Smette,[1,2⋆] L. Vandenbulcke[1⋆] and J. Surdej[1⋆]

[1]*Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 17, B-4000 Sart Tilman (Liège), Belgium*
[2]*European Southern Observatory, Alonso de Cordova 3107, Vitacura, Casilla 19001, Santiago 19, Chile*

Accepted 2005 December 19. Received 2005 December 7; in original form 2005 October 26

ABSTRACT

All-sky, multicolour, medium deep ($V \simeq 20$) surveys have the potentiality of detecting several hundred thousands of quasi-stellar objects (QSOs). Spectroscopic confirmation is not possible for such a large number of objects, so that *secure* photometric identification and precise photometric determination of redshifts (and other spectral features) become mandatory. This is especially the case for the *Gaia* mission, in which QSOs play the crucial role of fixing the celestial referential frame, and in which more than 900 gravitationally lensed QSOs should be identified.

We first built two *independent* libraries of synthetic QSO spectra reflecting the most important variations in the spectra of these objects. These libraries are publicly available for simulations with any instrument and photometric system.

Traditional template fitting and artificial neural networks (ANNs) are compared to identify QSOs among the population of stars using broad- and medium-band photometry (BBP and MBP, respectively). Besides those two methods, a new one, based on the spectral principal components (SPCs), is also introduced to estimate the photometric redshifts. Generic trends as well as results specifically related to *Gaia* observations are given.

We found that (i) ANNs can provide clean, uncontaminated QSO samples suitable for the determination of the reference frame, but with a level of completeness decreasing from $\simeq$50 per cent at the Galactic pole at $V = 18$ to $\simeq 16$ per cent at $V = 20$; (ii) the χ^2 approach identifies about 90 per cent (60 per cent) of the observed QSOs at $V = 18$ ($V = 20$), at the expense of a higher stellar contamination rate, reaching $\simeq$95 per cent in the galactic plane at $V = 20$. Extinction is a source of confusion and makes difficult the identification of QSOs in the galactic plane and (iii) the χ^2 method is better than ANNs to estimate the photometric redshifts. Due to colour degeneracies, the largest median absolute error ($|\Delta z|_{\rm Median} \simeq 0.2$) is predicted in the range $0.5 < z_{\rm spec} < 2$. The method based on the SPCs is promisingly good at recovering the redshift, in particular for $V < 19$ and $z < 2.5$ QSOs. For bright ($V \lesssim 18$) QSOs, SPCs are also able to recover the spectral shape from the BBP and MBP data.

Key words: methods: data analysis – quasars: general.

1 INTRODUCTION

Large surveys of quasars are important, not only for the study of the quasar phenomenon itself but also for the numerous astrophysical applications they offer. The quasi-stellar object (QSO) phenomenon itself is related to the galaxy history and evolution, to star formation and possibly to the interaction with other galaxies. Indeed, the mechanism which feeds the central black hole and triggers the nucleus active during a given period remains unclear (e.g. Sánchez et al. 2004). Being very bright, QSOs do also trace large-scale matter structures up to high redshifts (e.g. Croom et al. 2005; Yahata et al. 2005). Last, but not least, about 0.5 per cent of the bright distant QSOs are lensed by foreground galaxies, and gravitational lensing is an important tool to probe cosmological parameters as well as the dark matter distribution in distant galaxies (see e.g. Claeskens & Surdej 2002, for a review).

The number of objects and the degree of completeness in the QSO sample are of course important in these studies. The number

⋆E-mail: claesken@astro.ulg.ac.be (JFC); asmette@eso.org (AS); Luc.Vandenbulcke@ulg.ac.be (LV); surdej@astro.ulg.ac.be (JS)

of known objects is dramatically increasing: only 8609 QSOs were identified in 1996, while they are 48 921 in the present version of the Véron-Cetty & Véron (2003) QSO compilation including the set of 23 338 QSOs from the last 2-degree Field (2dF) release (Croom et al. 2004). When completed, the Sloan Digital Sky Survey (SDSS; York et al. 2000) will provide about 100 000 QSOs located at high latitudes in the North Galactic hemisphere.

However, reaching a high degree of completeness is more difficult. Indeed, the Véron and Véron QSO catalogue is very heterogeneous, and detailed studies of traditional QSO surveys are hampered by the large differences between quasar spectra [due to redshift, variable galactic reddening, variable absorption by intervening clouds along the line of sight or due to intrinsic properties, such as weaker emission lines (BL Lac objects), broad absorption lines (BAL) QSOs, red continuum QSOs and type II objects with narrow emission lines] as well as by similarities with stellar colours in some redshift ranges.

During its 5-yr mission, *Gaia*[1] will observe about 500 000 quasars down to $G \simeq 20$,[2] most of them being in the redshift range 1.5–2 (see Section 4.3). However, it is not just the number of QSOs which is important. *Gaia* has, indeed, unrivalled advantages.

(i) *Gaia* will scan the whole sky and potentially provide the most *homogeneous* QSO survey (including at low galactic latitudes where no optical survey has ever been carried out).

(ii) For each detected object, the end-of-mission data base will contain information on parallax, proper motion, photometric variability and colours in the broad-band and medium-band photometric (BBP and MBP) systems, consisting of five and 14 filters, respectively. Three criteria can thus be simultaneously used to identify QSOs among stars: (i) their lack of parallax and proper motion (such as for all distant extragalactic sources); (ii) their flux variations on a long time-scale (Veron & Hawkins 1995; Eyer 2002; Rengstorf et al. 2004) whose amplitudes are possibly anticorrelated with their luminosity (Hawkins 2000); (iii) their specific location in the photometric colour space, either due to the presence of strong emission lines at low redshift [e.g. ultraviolet (UV) excess] or due to the Lyα break at larger redshifts (e.g. Fan 1999). Since this photometric signature is a function of the redshift, the latter can also be derived from the *Gaia* data base, without requiring any spectroscopic follow-up. Other spectral features, such as the continuum slope, the extinction, etc., might also be derived photometrically.

(iii) The end-of-mission angular resolution is excellent: indeed, *Gaia* will be able to resolve a pair of point-like sources that includes a $G = 20$ primary and a secondary fainter by 2 mag and separated by 50 mas (Söderhjelm 2002).

[1] *Gaia* is an astrometric satellite designed by the European Space Agency (ESA) to carry out a stereoscopic census of the Milky Way. It is currently scheduled to be launched in 2011. Similarly, to its predecessor *Hipparcos*, *Gaia* will continuously be scanning the sky during 5 yr in order to produce an all-sky astrometric catalogue. However, *Gaia* will reach a much better accuracy for much fainter objects than *Hipparcos*, without any input catalogue: the *Gaia* final catalogue is foreseen to be complete down to the magnitude $G \sim 20$ and include about one billion stars. The mean accuracy on the parallax, the mid epoch position and the yearly proper motion will be about 20 μas at $G = 15$ and about 160 μas at $G = 20$. In addition, *Gaia* will also perform high-precision broad- and medium-band photometry (BBP and MBP, respectively) over the wavelength range 2400–10 500 Å. For more details on the *Gaia* mission, see Perryman et al. (2001).

[2] The G band corresponds to the unfiltered CCD response curve of the astrometric field detector; it is roughly equivalent to the V band.

Clearly, the three methods described in *Gaia*'s advantage number 2 will help in identifying the most peculiar types of QSOs, study their populations and get a sample as *complete* as possible.

It is also worth mentioning here that, thanks to advantages 1–3, it will be possible to identify with *Gaia* more than 900 multiply imaged, *gravitationally lensed* quasars with an angular separation $\Delta\theta \leqslant 1$ arcsec between the images (provided that an image of that size is telemetered to the ground). This number is one order of magnitude larger than the presently known number of lensed QSOs. *Gaia* will thus significantly contribute to the field by providing the largest and most uniform optical survey of lensed QSOs.

Coming back to the mission, one of its main tasks is the definition of the *Gaia* Celestial Referential Frame (GCRF), the *Gaia* realization of the International Celestial Reference System (ICRS; Mignard 2002). To reach that purpose, high-accuracy astrometry of a large number of QSOs (5000–10 000) *uniformly* distributed over the sky is required. Advantage number 1 is here obvious. However, the QSO (sub)sample dedicated to this task should exhibit absolutely no proper motion and be completely free from any contaminant. Existing QSO samples present in the Véron and Véron compilation may contain contaminants, while tiny apparent astrometric motions due to flux variations or (micro)lensing are always possible (Treyer & Wambsganss 2004).

Again, *Gaia*'s advantage number 2 is critical to build such a *secure* QSO sample. Indeed, none of the criteria taken individually would be 100 per cent sure: many stars detected by *Gaia* will have no measurable parallax and a fraction of them will not even have a detectable proper motion; the QSO variability is only a statistical behaviour; the QSO photometric signature may a priori not be sufficient to distinguish white dwarfs from low-redshift QSOs, M dwarfs from high-redshift QSOs and early F stars from $2 \lesssim z \lesssim 3$ QSOs (for which the Lyα line falls in the B filter).

However, the photometric signature is of particular importance to identify a clean set of QSOs for the GCRF since it is not biased from the astrometrical point of view and it is available before the end of the mission. Of course, the brightest QSOs ($G \lesssim 16$) are important to determine the GCRF. However, they will only represent about 0.1 per cent of all the detected QSOs and, on the other hand, since the determination of the *Gaia* parameters is iterative, after 1 yr of the mission, the signal-to-noise ratio (S/N) in a 16-mag object will be close to the one in a 18-mag object at the end of the mission.

For those reasons, we shall concentrate in this paper only on the photometric resources of *Gaia* and for objects with magnitudes $G = 18$, 19 or 20. We postpone the proper motion (and variability) analysis to another paper. On the other hand, we only deal with *point-like* objects because most of the faint resolved objects will be filtered on board.[3] On this ground, our present *main objective* is to answer two questions: (i) how well can we separate QSOs from stars and (ii) what is the expected accuracy in determining QSO astrophysical parameters (APs), such as the redshift, by using the colour information to be provided by *Gaia*.

Concerning question (i), the reader should bear in mind that QSOs only represent 0.05 per cent of the objects detected by *Gaia*. Getting a clean sample of quasars thus requires a method not only capable to select most QSOs but also capable to discard contaminants with a very high rejection rate. Indeed, averaged over the whole sky, a 99.95 per cent success rate in rejecting contaminants will still produce a catalogue of QSO candidates with 50 per cent of

[3] Faint, close by QSOs for which the host contribution is important might thus be lost.

contaminants (i.e. 500 000 objects!). In other words, the classification must be wrong for less than five stars per million to keep a fraction of contaminants below 1 per cent. For comparison, the sample of colour-selected SDSS QSO candidates only contains 66 per cent of true QSOs (Schneider et al. 2003). Consequently, the number of 'false positives' has to be minimized by the classification algorithm which is to be used to process the *Gaia* data. Of course, the pain is a function of galactic latitude: it is alleviated close to the galactic pole, while crowding and extinction make the quest for QSOs in the galactic plane a close to impossible task.

In the following, we first show in Section 2 how we have generated several libraries of QSO synthetic spectra in the range 2400–10 500 Å, how we have selected existing stellar synthetic spectra and how we have simulated the expected measurements in two different *Gaia* BBP and MBP systems. After introducing in Section 3 three different analysis techniques [the minimum distance method, neural networks and a method based on the spectral principal components (SPCs)], we compare the efficiency of the two former to correctly identify QSOs among the stellar contaminants in Section 4 and we test the ability of all three methods to determine photometric redshifts and other APs in Section 5. Finally, we summarize our conclusions in Section 6.

2 SPECTROSCOPIC LIBRARIES AND PHOTOMETRIC DATA BASES

2.1 Introduction

Contrary to other large QSO surveys (SDSS: York et al. 2000; 2dF: Croom et al. 2004), *Gaia* will not provide spectroscopy for any of the QSO candidates. On the other hand, the large number of medium-band filters will in effect provide a low-resolution spectrum which can be combined with photometry of unprecedented accuracy to identify QSOs. However, we immediately realized in this study that the diversity existing among QSO spectra must be taken into account to realistically estimate the fraction of correctly identified QSOs and the errors on the redshift determination. Indeed, the observed spectrum of a QSO strongly depends on its redshift z, the slope of its continuum, the individual line strengths (peak-to-continuum fluxes, line widths), the possible presence of BALs, absorption by intervening intergalactic clouds or extinction by dust in the QSO host galaxy or in the Milky Way. In addition, some classification methods, such as the artificial neural networks (ANNs) (see Section 3.2), can be very sensitive to the inputs: they can 'memorize' a single typical QSO spectrum but be unable to recognize slightly different QSO spectra. Diversity must be learnt.

Thus, an ideal data base consists of a large sample of observed spectra displaying the whole range of intrinsic diversity. Unfortunately, very few observed QSO spectra are flux calibrated over the full wavelength range covered by *Gaia* ($\lambda\lambda$ 2400–10 500 Å; see Section 2.5). Therefore, we had to rely on synthetic spectra. We have generated *two independent* QSO libraries, in order to avoid overoptimistic results occurring when test objects are selected by simply interpolating the same family of spectral models as the one used to create the reference data bases.

This section first describes those two independent methods used to build large samples of realistic QSO spectra – namely the *'modified template*' (hereafter MT) and the *SPCs* methods. Next, the adopted stellar libraries used as contaminants are presented. Finally, we briefly explain how the expected *Gaia* end-of-mission number of photoelectrons has been computed.

2.2 QSO-modified template

2.2.1 Brief description

The first step is to build a QSO composite spectrum over the widest possible rest-frame spectral range (i.e. covering all strong emission lines) and based on averaged observations available in the literature. Then, an analytical 'continuum spectrum', $C(\lambda;\alpha)$, characterized by the slope α, is removed from the final composite, leaving a pure 'emission-line spectrum', $E(\lambda)$.

A MT spectrum $I(\lambda)$ is then built by multiplying the emission-line spectrum by a factor of β, and adding it to the analytical continuum spectrum – considered over the spectral range 350–10 500 Å – and with a slope $\alpha\prime$ possibly different from the template slope α:

$$I(\lambda) \propto [C(\lambda;\alpha') + \beta(W)E(\lambda)]. \quad (1)$$

Note that because the detection of the emission line depends on their contrast to the continuum rather than on their intensity, we related the β factor to W, the *total* equivalent width of the emission lines according to the following integral relation:

$$W = \int_{350\ \text{Å}}^{10\,500\ \text{Å}} \frac{\beta E(\lambda)}{C(\lambda;\alpha')} \mathrm{d}\lambda. \quad (2)$$

The resulting spectrum $I(\lambda)$ is then redshifted and intergalactic absorption and extinction by Milky Way dust are accounted for.

The following paragraphs detail the ingredients of this receipt, and discuss the validity of the approach.

2.2.2 The composite spectrum

In order to build the input composite QSO spectrum, we have combined four averaged observed spectra available in the literature, each covering smaller spectral ranges (see Table 1). We followed the method described by Royer et al. (2000) in the range 310–6000 Å and we similarly added another template between 6000 and 8000 Å.

2.2.3 The continuum and emission-line spectra

An analytical function was then adjusted to regions of the composite spectrum thought to represent the true continuum. This function, called the *continuum spectrum*, is inspired by the active galactic nucleus continuum spectrum generated by Cloudy (Ferland 1996)

$$C(\lambda;\alpha) \propto \left[\left(\frac{\lambda}{\lambda_0}\right)^{-\alpha} \exp(-\lambda_1/\lambda)\exp(-\lambda/\lambda_2) + a\right], \quad (3)$$

where α is defined as the continuum slope and λ is the wavelength expressed in Å. For the composite spectrum, we find $\alpha = 0.03$. Adequate values for the other parameters are $\lambda_0 = 10\,500$ Å, $\lambda_1 = 240$ Å, $\lambda_2 = 1250$ Å and $a = 0.021$.

Table 1. Contributions to the rest-frame composite QSO spectrum. γ is the single power-law exponent characterizing the continuum over the considered wavelength range.

λ range (Å)	γ	Reference
310–1050	−1.96	Zheng et al. (1997)
1050–2000	−0.99	Zheng et al. (1997)
2000–6000	−0.32	Francis et al. (1991)
6000–8000	−0.32	Cristiani & Vio (1990)

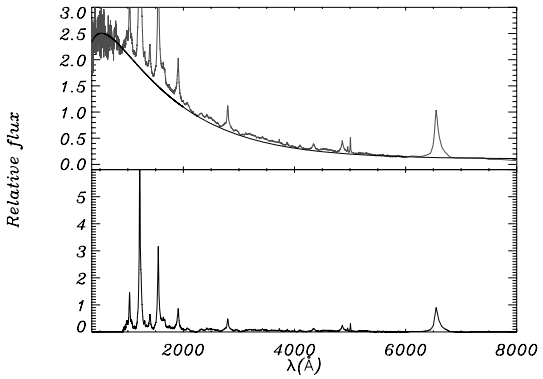

Figure 1. Top: the continuum spectrum $C(\lambda)$ (solid line) with $\alpha = 0.03$ is shown with the composite spectrum used to define it. Bottom: emission-line spectrum $E(\lambda)$. The main lines are Lyβ/O VI at 1025 Å, H I Lyα at 1216 Å, C IV at 1549 Å, C III] at 1909 Å, Mg II at 2800 Å and H I Hα at 6563 Å.

This fitted continuum spectrum is then subtracted from the composite spectrum. The rest of this subtraction defines the *emission-line spectrum*, $E(\lambda)$, for which we set a flux equal to 0 for $\lambda < 912$ Å and $\lambda > 6900$ Å. Fig. 1 shows the composite spectrum, the continuum adjusted to the composite spectrum and the resulting emission-line spectrum.

2.2.4 Intergalactic absorption

The line of sight towards a distant QSO encounters a number of absorbers at different redshifts, resulting in a large number of H I Lyα lines which are collectively known as the Lyα forest. Their H I column density is $\log N_{\rm H\,I} < 17$. Their large number towards high-redshift QSOs leads to significant absorption in the wavelength range defined by the H I Lyβ and Lyα lines at the QSO redshift (zone A): the overall decrement is usually referred to as $D_{\rm A}$. Between the Lyman limit and the Lyβ lines (zone B), lower redshift Lyα absorbers mix with the Lyman series lines associated with the highest redshift Lyα lines and also lead to a decrement $D_{\rm B}$. Finally, at wavelengths smaller than the Lyman limit at the QSO redshift (zone C), continuous absorption caused by the large numbers of absorbers results in the Lyman valley. This region is also primarily affected by larger $N_{\rm H\,I}$ absorbers found at random redshifts in front of the QSOs: the Lyman limit systems if $17 < \log N_{\rm H\,I} < 20.3$ and the damped Lyα systems if $\log N_{\rm H\,I} > 20.3$. While the effect of the H I Lyα lines themselves is relatively minor (except for exceptionally rare, very large $N_{\rm H\,I}$ absorbers), these systems indeed lead to a complete lack of flux below the wavelength of the Lyman limit at the absorber redshift.

Following Royer et al. (2000), we have adopted in the present study the averaged observed values of $D_{\rm A}$ and $D_{\rm B}/D_{\rm A}$ given by Irwin, Mc Mahon & Hazard (1991) and Warren, Hewett & Osmer (1994), respectively. The absorption in zone C is modelled following Madau (1995) on the basis of the work by Møller & Jakobsen (1990). We note that the treatment of zone C corresponds to the mean absorption over a large number of sight lines. At the photometric precision that *Gaia* will reach, the presence of a $2.2 < z_{\rm abs} < 3.5$ Lyman limit or higher $N_{\rm H\,I}$ system in the spectra of $z_{\rm abs} < z < 3.5$ QSO will very significantly decrease the observed fluxes in the first or first two bluest filters. The presence of a Lyman limit system in $z > 3.5$, $G > 18$ QSOs will significantly affect at most one filter compared to the mean absorption model that we have adopted as the mean intergalactic medium (IGM) already correspond to a decrease of about 2 mag for most of the zone C. Quantifying the consequences of the presence of a high $\log N_{\rm H\,I}$ with respect to the identification and determination of APs of QSOs, and alternatively determining the probability that such a system can be detected using *Gaia* photometry will be a part of a coming study for which a stochastic approach to include large $N_{\rm H\,I}$ will be used. However, since $G_{\rm lim} = 20$, most of the QSOs detected by *Gaia* will be at $z \lesssim 2.2$ (see Section 4.3) and their colours will not be much affected by the exact amount of absorption.

2.2.5 Extinction by dust

Finally, extinction by dust can also play an important role. Dust can be associated with different objects located along the line of sight to the background QSO: (i) at the QSO redshift (located either in the host galaxy or very close to the accretion disc, such as the molecular torus), (ii) in intervening absorbers, such as large $N_{\rm H\,I}$ systems or (iii) in the Milky Way.

Optical surveys are usually designed in such a way as putting the intrinsic blue colour of unreddened QSOs in evidence so that they are probably biased against reddened QSOs. Therefore, at large and low galactic latitudes, red QSOs are often discovered through radio surveys.

In order to generate dust extinction in the Milky Way, we used the standard Milky Way extinction law characterized by $R_v = 3.1$ and the free parameter A_V (Cardelli, Clayton & Mathis 1989). In order to minimize the size of the spectroscopic libraries, this step is performed when generating *photometric* measurements (see Section 2.5). We have not carried out simulations with dust located at higher redshifts. On the other hand, we have not considered a possible change of the extinction law (i.e. the R_v value; Fitzpatrick 1999). However, at high galactic latitudes where the reddening is weak (i.e. $E_{B-V} < 0.1$), the exact value of R_v in the range $2.3 \leqslant R_v \leqslant 5.5$ leads to a change in the $U - V$ colour index smaller than $+0.02/-0.09$ mag around the mean. At lower galactic latitude, where substantial reddening may occur, it is very likely that the numerous stars to be detected by *Gaia*, even in a relatively small field around a QSO, will be sufficient to determine the extinction law independently (Jordi et al. 2006).

2.2.6 Validity of the modified template

The MT method allows us to create a wide range of QSO spectra, including red continuum objects or BL Lac objects (i.e. with $W \simeq 0$). However, it cannot account for the broad absorption seen in the blue side of the C IV and Mg II emission lines of the so-called BAL QSOs, which represent more than 10 per cent of the QSO population (e.g. Reichard et al. 2003b). Even for normal QSOs, this method does not take into account any known correlation, such as the Baldwin effect (Baldwin 1977), or the range of observed fluxes and equivalent widths of *individual* emission lines.

Given these potential shortcomings, it is important to test the validity of the MT approach and quantify its limitations. Therefore, we have confronted MT spectra with spectra from the Data Release 1 (DR1) (beta version) of the SDSS QSO catalogues. Compared to other surveys, the SDSS offers the advantage of providing well-documented, flux-calibrated spectra corrected from Milky Way dust extinction. In addition, the available spectral coverage is relatively large (3800–9000 Å) compared to the range covered by *Gaia*.

Specifically, a MT spectrum was fitted to each of the SDSS 'QSO' spectra using the least-squares method. The free parameters were α

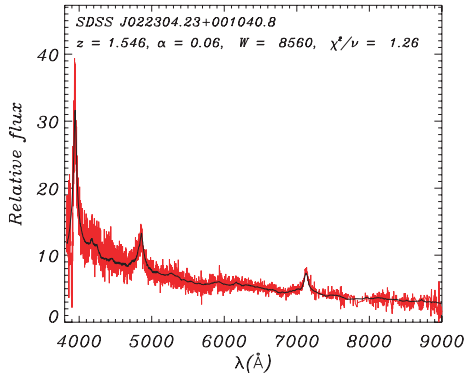

Figure 2. An example of an excellent fit using the MT method.

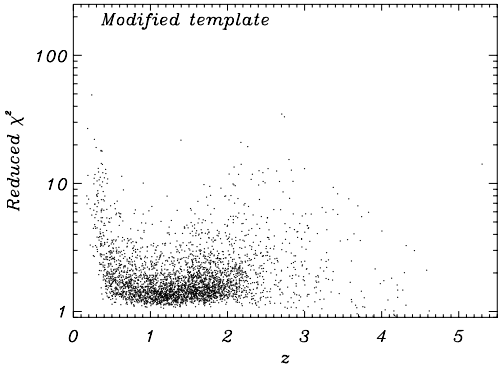

Figure 3. Reduced χ^2 as a function of z using the MT approach.

and W as well as a constant factor to account for the relative difference in intensity between the data and model spectra. After some trials, we found that the best fits were obtained if α was limited to vary in the (still large) range $-4 < \alpha < 4$. On the other hand, W could only take positive values. The redshift z was also fitted but within a range limited to 1 per cent of the value given in the Z keyword of the SDSS file header. The sampling of each template spectrum was matched to the sampling of the spectrum to be fitted. Normal weighting based on the 1σ error array was used. For the fitting process, we considered as useful the pixels with the following characteristics: (i) the corresponding MASK array value (octal numerotation) is below 100 or equal to 4 000 000 (detected emission line), (ii) the pixels were not within 2 pixels of a pixel affected by a sky emission line, (iii) the error array value is larger than 0.1 to avoid meaningless small values and (iv) the wavelengths are larger than 3830 Å as the first pixels in the spectra were often affected by problems. The procedure was performed in IDL using Craig B. Markwardt's MPFITFUN routine[4], based on the Levenberg–Marquardt method.

In the following, we discuss the results of this comparison. However, to make it meaningful, we had to limit the sample of SDSS spectra to those for which (i) the keyword SPEC_CLN is set to 3 or 4 (i.e. QSO or high-redshift QSO), therefore excluding peculiar stars; (ii) the keyword Z_STATUS set between 3 and 10 inclusive (redshift measurement succeeded with the two methods – emission line and cross-correlation – giving consistent redshift estimates) and (iii) a number of useful pixels – as defined in the previous paragraph –larger than 3000. These restrictions reduce the size of the sample of SDSS spectra by 10 per cent, from 3814 to 3429. The median S/N per pixel over the sample is 9.5.

In general, the overall fit is remarkably good given the small number of fitted parameters. An example of an excellent fit is shown in Fig. 2.

The median value of the reduced χ^2 is 1.67. Although the reduced χ^2 values range from 1.04 to 121, 2/3 of the fitted spectra have a reduced χ^2 smaller than 2. However, the distribution of reduced χ^2 strongly depends on the redshift, as shown in Fig. 3. Most of the large χ^2 values occur when at least one of the strongest emission lines (H I Hα, C IVλ1550 and H I Lyα) appears in the observed wavelength range: indeed, (i) their width often deviates from the one in the composite spectrum and (ii) their intensity does not scale well with other emission lines. In this respect, the Baldwin effect (Baldwin 1977) is well known, i.e. the equivalent width of the C IV line is anticorrelated with the QSO luminosity. Other broad emission lines show a similar although less pronounced effect, while some do not show any statistically significant (anti)correlation with the QSO luminosity (see Yip et al. 2004). At $z > 2.1$, part of the increase in the mean reduced χ^2 value is due to the Lyα forest, as the large number of absorption lines contributes to a larger dispersion around the mean.

Finally, high values of the reduced χ^2 are also associated with BAL quasars; this confirms that the latter are badly represented by such MTs. High S/N BAL composite spectra computed from SDSS data have been published by Reichard et al. 2003a. However, it is not straightforward to use the MT technique in this case because individual BAL spectra may be very peculiar (see e.g. Weymann et al. 1991). Yip et al. (2004) have found that BAL features are progressively recovered with the SPC analysis only when high-order modes are included (up to 50 modes for an accurate reconstruction). On the other hand, the BAL spectrum composite should also be dereddened from intergalactic absorption and should cover a wider spectral range to further create BAL QSO spectra at different redshifts. Ideally, a representative set of BAL spectra *observed* in the *Gaia* spectral range should be included in the library.

However, the most important quantities to determine are the differences in integrated fluxes between the SDSS spectra and the modelled spectra, as measured through the adopted filters (the latter are introduced in Section 2.5 and Fig. 9). Fig. 4 shows these quantities for the filters of the 2B BBP and 1X MBP sets that cover the wavelength range of the SDSS spectra. Table 2 summarizes the mean and standard deviation in each filter. Part of the dispersion is actually due to the low S/N of the SDSS spectra, in particular, in the reddest filters (X825, X860) and somewhat in the bluest one (X410). However, systematic magnitude differences observed as a function of redshift reflect the inaccuracy of the modelled spectra to reproduce real spectra. For that reason, we found necessary to create an independent spectral library to check the performance of our algorithms (see Section 2.3).

2.2.7 *Range of parameters and MT spectroscopic libraries*

For the SDSS QSOs, the range covered by the parameters α and W is shown in Fig. 5. A correlation between the two parameters is conspicuous. The lack of QSOs with either large α and small W values or small α and large W values may be due to an observational bias (however, checking whether those objects could have been detected by the SDSS is beyond the scope of this paper), but a residual parameter correlation is quite likely. Indeed, in steep blue continuum spectra ($\alpha \gtrsim 1$), the total W is dominated by the Hα line, so that

[4] Available at http://cow.physics.wisc.edu/craigm/idl/idl.html

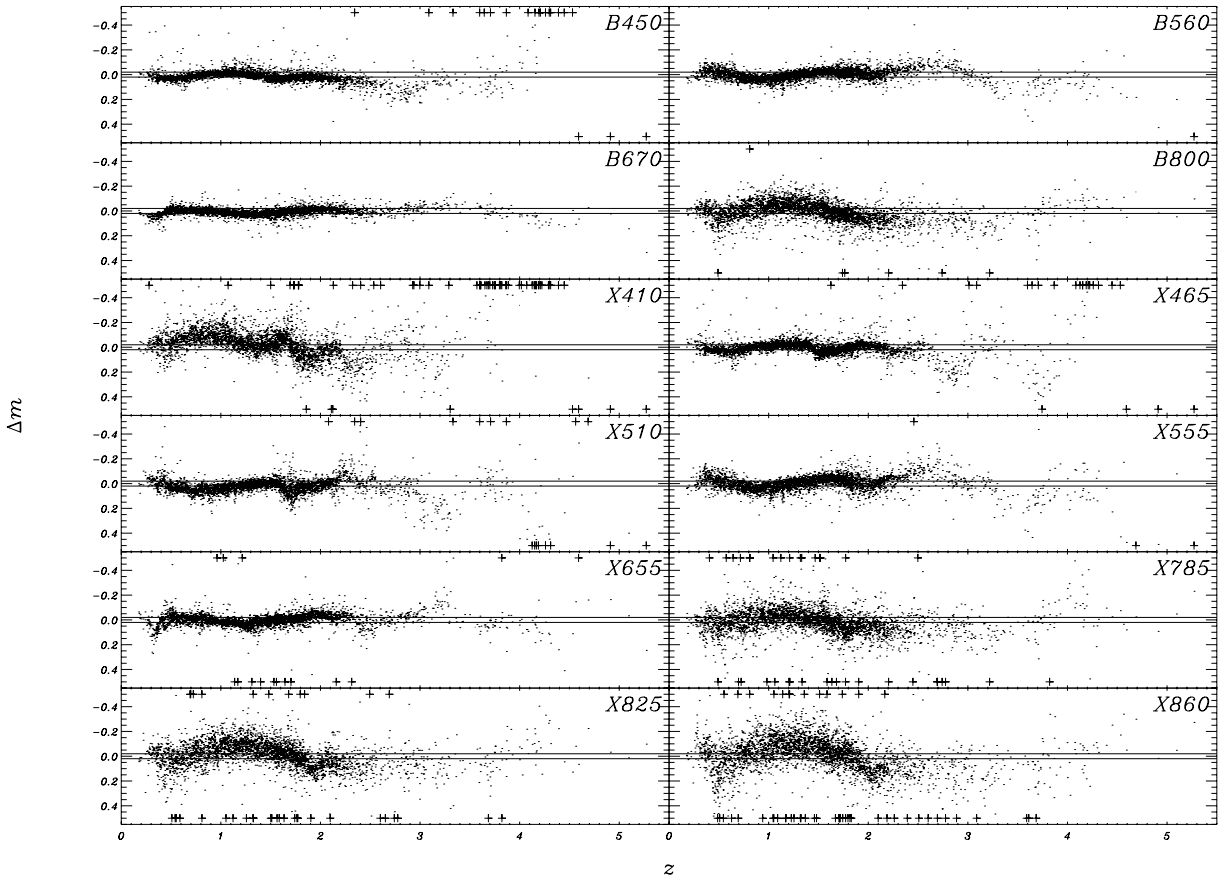

Figure 4. Difference of magnitudes between the SDSS spectra and the corresponding fitted spectra using the MT method in each of the relevant filters of the 1X+2B system. Crosses indicate an absolute difference of magnitudes larger than or equal to 0.5. The horizontal lines indicate the precision on a $G = 20$ A0 star at the end of the mission.

Table 2. Mean and standard deviation of the magnitude difference between SDSS spectra and their modelled spectra.

Filter	Modified template	Principal components (no constraint)	Principal components (non-negativity)
B450	0.01±0.14	0.00±0.11	0.01±0.16
B560	−0.00±0.05	−0.00±0.06	−0.00±0.04
B670	0.01±0.03	0.00±0.05	−0.01±0.04
B800	0.01±0.09	0.01±0.11	−0.03±0.12
X410	−0.05±0.26	−0.04±0.31	−0.03±0.34
X465	0.00±0.16	−0.01±0.15	−0.01±0.17
X510	0.02±0.09	0.02±0.10	0.02±0.08
X555	−0.01±0.07	−0.00±0.10	−0.00±0.06
X655	0.02±0.35	0.02±0.35	0.00±0.35
X785	0.03±0.34	0.03±0.37	0.00±0.38
X825	0.01±0.43	0.02±0.51	−0.03±0.51
X860	0.02±0.45	0.06±0.69	−0.02±0.68

blue lines are only seen for high values of W; at the opposite, for flat or red continuum spectra ($\alpha \lesssim -1$), blue emission lines can be seen even with a low value of W.

Finally, and despite the fact that most (i.e. ∼91 per cent) of the SDSS QSOs are characterized by $-1 \leqslant \alpha \leqslant 1$ and $2000 \leqslant W \leqslant 20\,000$ Å, we decided to explore the widest but reasonable range of parameter values. Therefore, we built with the MT method two libraries of QSO synthetic spectra with parameters z, α and W in the range [0, 5.5], [−4, 3] and [10^2, 10^6] Å, respectively. They contain 17 325 (respectively 20 000) spectra with uniformly and regularly (respectively randomly) distributed values of the parameters. The random set of QSOs is primarily aimed as reference input data base, while the regular one is intended for testing the influence of each individual parameter (see also Section 2.5 and Tab. 3). These two spectroscopic libraries are available and can be queried via the *Gaia* spectral library web page hosted by ESA.[5]

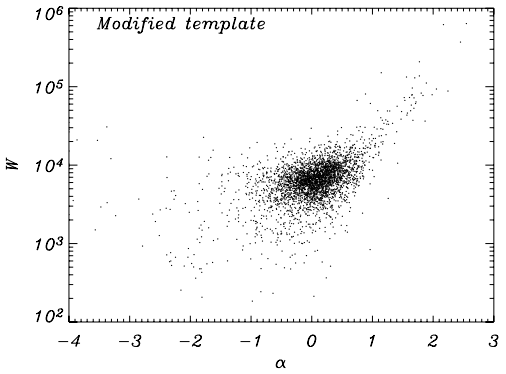

Figure 5. Range of α and W values for the SDSS QSOs.

[5] http://gaia.esa.int/spectralib/

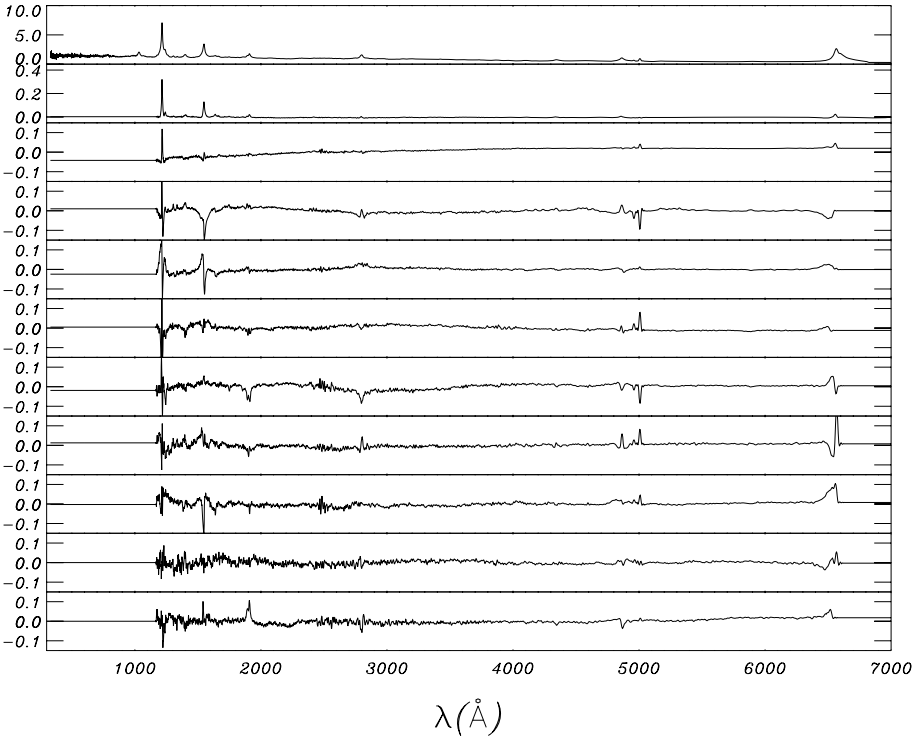

Figure 6. SPCs used to model the SDSS spectra. The spectral ranges between Lyα and Hα were provided by Shang. Their extension to the red and blue is described in the text. Regions redwards of 7000 Å are featureless and not shown.

2.3 QSO spectral principal components

2.3.1 Introduction

Francis et al. (1992) have carried out a (spectral) principal component analysis of the Large Bright QSO Survey (LBQS) spectra. They identify a set of 10 SPCs which account for more than 99 per cent of the intrinsic variance observed amongst rest-frame quasar spectra. We refer to Francis et al. (1992) for a detailed description of this statistical technique. Here, we will use the results of their study in an original way. Indeed, following their results, a QSO spectrum $S(\lambda)$ can be represented as a linear combination of a mean spectrum $M(\lambda)$ and the principal components $P_i(\lambda)$:

$$S(\lambda) \propto \left[M(\lambda) + \sum_{i=1,10} w_i P_i(\lambda) \right], \quad (4)$$

where w_i is the weight of the ith principal component P_i; in particular, all non-linear effects affecting the appearance of a QSO spectrum are included within the principal components. Consequently, given an observed spectrum $S(\lambda)$ and a set of mean spectrum and principal components, the weights w_i and the proportionality factor can be obtained by resolving a system of linear equations.

Unfortunately, the wavelength range of the principal components provided by Francis et al. (1992) is only about 1000 Å and cannot be used to build spectra over the entire wavelength range covered by *Gaia*. However, Shang et al. (2003) recently determined the first set of SPCs covering an extended wavelength range, from shortward of H I Lyα to just redwards of H I Hα. Although the mean spectrum and SPCs were derived from a (too) small set of 18 QSO spectra, we found that they can be used to build a library of QSO spectra based on the observed SDSS QSOs one but extended to cover a wider redshift range

2.3.2 Principal components

Fig. 6 illustrates the mean spectrum and the first 10 principal components (SPCs), as kindly provided by Zhaohui Shang. However, in order to use them for the whole wavelength range covered by *Gaia* and redshift range $0 < z < 5.5$, we have extended them in the following way. (i) In the blue, the mean spectrum was extended using Zheng et al. (1997) composite spectrum, while each SPC was set to a fixed value equal to a representative intensity at the bluest part of each spectrum (often the intensity of the first pixel). (ii) The spectra provided by Shang only cover a part of the red wing of the H I Hα line; therefore, we completed the wing with ad hoc values. (iii) Redwards of the H I Hα line, the mean spectrum was completed by a power law with an index $\gamma = -0.3$ (Cristiani & Vio 1990; table 1), while each SPC was completed by a suitable constant value. As the mean spectrum and the SPCs were derived from only a small set of 18 medium S/N spectra, the SPCs themselves are noisy. Therefore, we slightly smoothed them with a Savitzky–Golay filter which preserves even the narrowest emission lines.

2.3.3 Validity of the principal component approach

As for the MT approach, we checked the validity of the SPC approach to reproduce the SDSS spectra. The useful pixels in the SDSS spectra were selected in the same way as above. Then, given a redshift (see below), we rebinned the redshifted mean and SPC spectra so as to match the wavelength scale for each SDSS spectrum $S(\lambda)$. Note that taking into account the absorption due to the intergalactic absorption is simple, it is sufficient to multiply the mean and all the SPC spectra by the transmission spectrum of the IGM at the given redshift. For each redshift, a two-dimensional matrix $\mathbf{A}(i, j; \lambda)$ is then built, whose first and successive rows contain the intensities

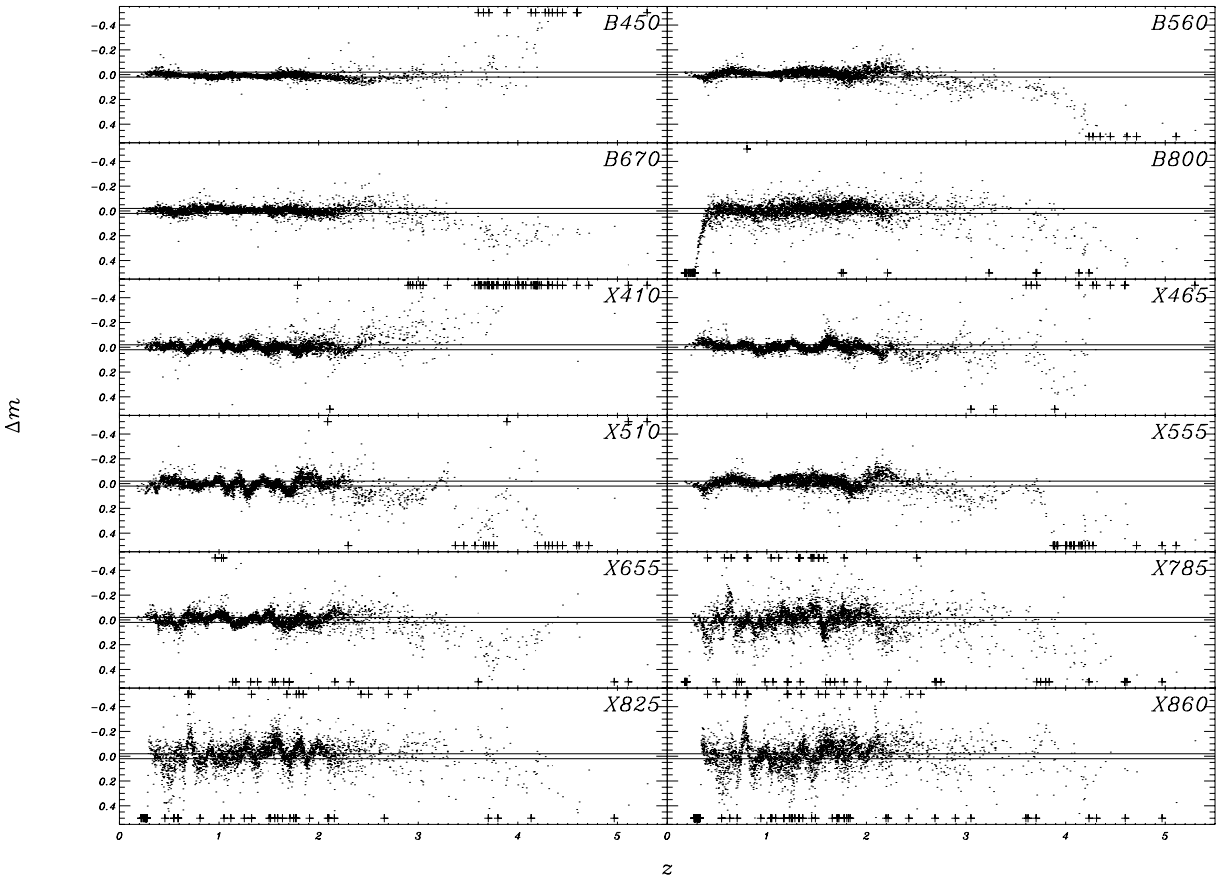

Figure 7. Difference of magnitudes between the SDSS spectra and the corresponding fitted spectra using the principal component approach (see the caption of Fig. 4).

of the mean spectrum and SPC spectra as a function of the wavelength of S, corrected for intergalactic absorption. We then solved the overdetermined system of linear equations

$$\mathbf{A}(\lambda)\boldsymbol{w} = \boldsymbol{S}(\lambda) \tag{5}$$

by singular value decomposition.[6] The solution vector $\boldsymbol{w}$ contains the weights $w_{i>0}$ of each SPC multiplied by a constant normalization factor. This method is elegant as it only involves matrix operations. In particular, the matrix $\mathbf{A}$ is unique for a given redshift z.

In order to estimate the goodness of fit, we also computed a reduced χ^2. As for the MT approach, we searched for the redshift which leads to the smallest χ^2 within a 1 per cent range of the value provided by the Z keyword in the SDSS spectrum header.

As for the MT approach, the SPC one is also remarkably good at reproducing the SDSS spectra (median $\chi^2 = 2.11$). This was not obvious since the SPCs were built from a base of 18 spectra only. Fig. 7 shows the magnitude differences between the SDSS spectra and their model obtained using the SPC approach. Table 2 gives a summary table. Compared to the MT approach, the SPC approach is found to be better at reproducing the diversity of emission lines. On the other hand, the SPC-built spectra tend to introduce wiggles in regions far from strong emission lines. The match will surely improve when a set of SPCs built from a larger data base is available. Indeed, there is obviously a problem in the B800 filter for $z < 0.4$, as the red wing of the Hα line is only poorly approximated.

Unfortunately, the method described above suffers from a drawback. Indeed, at first sight, the weights $\boldsymbol{w}$, solution to equation (5), appear valid for the whole wavelength coverage of the SPCs, and not only over the wavelength range of S. However, this statement would be true if (i) the SPCs that we used were representative of the QSO population as a whole and (ii) the S/N of each SPC and of S would be large enough.

Fig. 8 clearly indicates that the weights w_i are not independent of each other. In addition, over a large redshift range, the distribution of w_i is not centred at 0, and can actually have values much larger, in absolute terms, than the ones found by Shang et al. (2003) for their data base spectra. Therefore, Fig. 8 shows that the SPCs we used are not representative of the QSO population as a whole, although they provide a good starting point.[7] On the other hand, the S/N/pixel of the SDSS spectra rarely exceeds 12 by design, which is barely sufficient to apply equation (5). Therefore, we were not surprised to find that applying the solution $\boldsymbol{w}$ of equation (5) to the SPCs over their whole wavelength range sometimes produces spectra that can reach negative flux.

Requiring the solution spectrum to be non-negative for the whole (redshifted) wavelength range covered by the SPCs forced us to

[6] A linear rebinning of the SDSS spectra had to be carried out before in order to have identical steps in wavelength for both the SDSS, mean and SPC spectra.

[7] While most of the present work had been completed, Yip et al. (2004) described SPCs directly obtained from the SDSS DR1; the latter might alleviate the problems mentioned here.

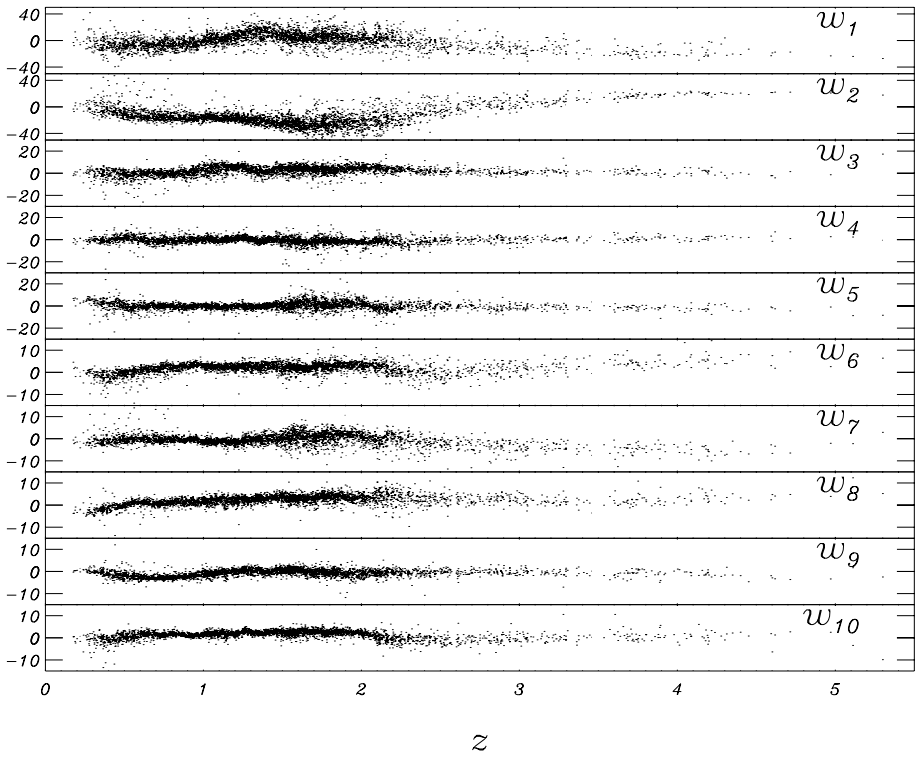

Figure 8. Dependence of the w_i on z for the SDSS spectra. These w_i are solution of the overdetermined system of equations given in equation (5) without non-negativity constraints. They were obtained by a method using singular value decomposition.

use a somewhat less elegant, but more robust way to solve equation (5). We chose the L1 estimation algorithm with degeneracy and linear constraints by Shi & Lukas (2002) designed to find the solution of overdetermined systems of linear equations in the L1 sense with linear constraints. To compare the effect of this added constraint on the solution spectra, we calculated their χ^2 and found it, as an average, 5–10 per cent larger than without non-negativity constraints, although more than 20 per cent of the spectra have their χ^2 actually improved. On the other hand, comparison between Fig. 8 and an equivalent figure using the L1 estimation algorithm indicates that the weights tend to have a smaller spread in the latter.

2.3.4 Range of parameters and SPC spectral library

A library of 20 000 QSO synthetic spectra has been generated with the SPCs method and using the sets of weights w_i obtained by fitting the SDSS DR1β QSO spectra (see above). The redshifts have been randomly drawn from a uniform distribution in the range $0 < z < 5.5$. This data base represents a second, *totally independent* test set (TS), intended to make robust estimates of the algorithm performances with totally independent data. It is also available on the ESA web page.

2.4 Stellar libraries

Given the *Gaia* limiting magnitude $G = 20$, stars represent the most important source of unresolved contaminants when searching QSOs. Indeed, most of the compact emission-line galaxies (CELGs) with $G \leqslant 20$ should be resolved by *Gaia*. In the present study, remaining, faint unresolved CELGs may contaminate (at a low level) the QSO sample, but with *no consequence* for the determination of the GCRF.[8]

We considered three kinds of stars: single 'normal' stars, binary stars and white dwarfs, all described with synthetic spectra.

More specifically, normal stars are represented in the Basel Stellar evolution Library (BaSeL2.2) (Lejeune, Cuisinier & Buser 1998). The latter covers the whole Hertzsprung–Russell diagram with effective temperature $T_{\rm eff}$ from 2000 to 50 000 K, surface gravity in the range $-1.02 < \log g < 5.5$ and metal abundance [M/H] from -5 to 1 dex, but with solar α-element abundances. Dust extinction $A_v < 9$ is added when computing the expected fluxes in the filters (see Section 2.5), assuming a Galactic, $R_v = 3.1$, extinction law. Interpolations have been done for two sets of 10 000 stars with randomly distributed values of $T_{\rm eff}$, $\log g$, [M/H] and A_v.

Next, the fluxes of 1000 unresolved binary stars have been computed from pairs of individual spectra covered by the BaSeL2.2 library by Malkov (*Gaia* blind testing cycle 2, UB-PWG-014.txt). The magnitude difference between the stars is in the range 0–5 mag. No constraints were given on the nature of the spectra within the pairs. Again, dust extinction is added.

Finally, the population of hydrogen white dwarfs (DA class) being the dominant one, the colours of those contaminating stars have been computed from (possibly reddened) pure hydrogen atmosphere spectral models, with regular values of the parameters in the ranges $7000 < T < 80\,000$ K and $7.5 < \log g < 8.5$ (Koester, private communication).

We did not model very peculiar stars, like Wolf–Rayet stars whose strong emission lines can make confusion with quasars. However,

[8] Unresolved CELGs are identified as an individual class of objects to be dealt with by the general *Gaia* classifier and the determination of their APs consists in a specific task, unrelated to the QSOs.

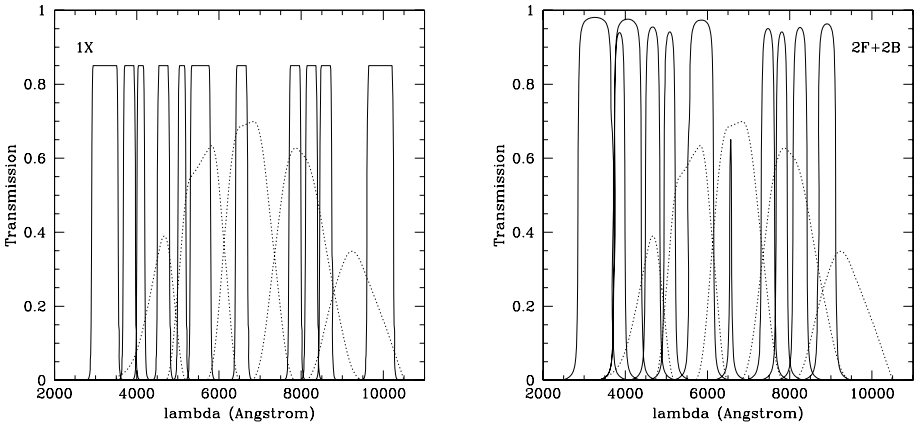

Figure 9. Adopted *Gaia* photometric systems: 1X+2B (left-hand panel) and 2F+2B (right-hand panel). Full lines: MBP photometric systems; dotted lines: BBP photometric systems, already multiplied by the response curve of CCD 1.

those populations of objects are very small compared to that of the QSOs, and are essentially confined to the galactic disc so that they do not represent a big reservoir of contaminants.

2.5 Adopted filter systems and photometric data bases

For our subsequent calculations, data bases must contain the number of photoelectrons with associated errors expected at the end of the *Gaia* mission in the two filter sets selected for *Gaia*. A first set of five broad-band filters (BBP) is optimized for chromaticity correction on the astrometry and a second set of 14 medium-band filters (MBP) will be used to perform efficient astrophysical diagnosis on stars (Jordi et al. 2006). However, the choice of the BBP and MBP was not yet finalized at the time of the simulations presented in the present paper. At that time, the MBP was only consisting of 11 filters. We thus selected two systems of five and 11 filters, namely the 2B (Lindegren 2003) + 1X (Vansevicius & Bridzius 2002) and 2B + 2F (Jordi et al. 2003) systems (see Fig. 9). The final design (C1B + C1B; Jordi, Hoeg & Bailer-Jones 2004) is not drastically different. There are three more MB filters, but they are narrower. The qualitative conclusions of this work should not be modified when adopting the final system. We made the calculations with the Barcelona simulator for the *Gaia*-II revised design (as of 2003 June) and kindly provided by C. Jordi (see also URL http://gaia.am.ub.es/PWG/index.html).

We have generated realistic data by (i) estimating photometric errors from *noisy* counts and (ii) introducing limiting magnitudes in filters, where the magnitude error is larger than 0.5 mag. Dust extinction is added at this stage, with a uniform distribution in the range $0 < A_v < 9$ for normal and binary stars, and in the range $0 < A_v < 2.5$ for white dwarfs (WDs) and QSOs. The data set QSO-REG contains 86 625 objects because it is built from the spectroscopic library of 17 325 QSOs with regular grid of parameters and for five different values of $A_V (= 0; 0.5; 1; 1.5; 2)$.

Finally, for all sets of spectra listed in Table 3 and described in Sections 2.2.7, 2.3.4 and 2.4, data bases have been generated without noise as well as for $G = 18$, 19 and 20, in the two adopted filter systems (2B+1X and 2B + 2F). For TSs, $N_{\rm sim} = 25$, i.e. 24 realizations of the noise have been made from each first noisy realization, in order to determine classification probabilities and uncertainties on the APs, as it would be done from real data.

Table 3. Summary of the photometric data bases (see the text for details).

Name	Object	Origin	$N_{\rm obj}$	$N_{\rm sim}$
QSO-RAN	QSOs	MT	20 000	1
QSO-REG	QSOs	MT	86 625	25
QSO-PCA	QSOs	SPC	20 000	25
STAR-RAN1	Stars	BaSeL	10 000	1
STAR-RAN2	Stars	BaSeL	10 000	25
BIN-RAN1	Binaries	Malkov	1000	1
BIN-RAN2	Binaries	Malkov	1000	25
WD-REG1	W.Dwarfs	Koester	972	1
WD-REG2	W.Dwarfs	Koester	162	25

3 DATA ANALYSIS TECHNIQUES

In this section, we describe the data analysis techniques used (i) to identify QSOs among stars and (ii) to determine their redshift and other APs.

3.1 Nearest neighbours

The search in a multidimensional reference table for the k-nearest neighbours (k-NN) of an observed data point is equivalent to the determination of minimum distances. The most simple and most popular minimum 'distance' is the χ^2 and has been used for a long time to determine galactic photometric redshifts by means of template fitting (e.g. Koo 1999). It consists in selecting in a reference data base the object j with flux vector $\boldsymbol{F}_j$ which matches at best the observed flux vector $\boldsymbol{f}$, i.e. for which

$$\chi^2 = \sum_{\text{filter}_i=1}^{n} \left(\frac{f_i - F_{ij}}{\sigma_i} \right)^2 \qquad (6)$$

is minimum. σ_i is the photometric error in filter f_i. This is a *local* fitting in the sense that the result only depends on the local information in the colour space.

The principal advantages of this method reside in its capacity (i) to provide complementary physical information associated with the reference object (e.g. continuum slope, strength of the emission line in the case of QSOs); (ii) to easily identify possible degeneracies between the members of the reference table leading to wrong associations; (iii) to work with missing data (e.g. corrupted data in one or several filters) and (iv) to identify *unknown* objects from their

large distance to the known loci. However, the minimum distance method is *slow* as it requires comparison of each entry object with a potentially large number of reference objects.

Template fitting with the χ^2 method could be refined in two ways. (i) By using several nearest neighbours (k-NN) and performing interpolation and (ii) by defining a set of principal directions in the filter space and working in this subspace (e.g. Cabanac et al. 2002).

3.2 Artificial neural networks

ANNs have been used in astronomy for more than 10 yr in order to classify stars and galaxies (Bertin 1994; Andreon et al. 2000), galaxy types (Storrie-Lombardi et al. 1992; Naim et al. 1995; Lahav et al. 1996) and stellar spectral types (Bailer-Jones, Irwin & von Hippel 1998; Weaver 2000). The application for redshift estimation is quite new (see Vanzella et al. 2004). In the present study, ANNs are used both to identify QSOs among stars (i.e. classification networks whose outputs contain the probabilities of the object types, see Section 4.2) and to estimate their photometric redshift (i.e. regression networks, see Section 5.3).

A supervised ANN can schematically be considered as a black box able to learn a non-linear relation between known inputs ($I =$ fluxes here) and outputs ($O =$ astrophysical information), contained in a *learning set* (LS). Once the relation is learnt, new outputs can quickly and reliably be derived for new inputs if the latter are statistically compatible with the known inputs of the LS. Within the black box is an architecture A of one (or more) layer(s) of n nodes (or neurones). Each node performs a weighed combination of the outputs of all the nodes of the previous layer, which is then passed through a non-linear 'activation function' (e.g. threshold, tanh, sigmoid,...) and the result is sent to each node of the next layer. A first layer is fed with the physical inputs and the last layer contains the required output(s). In summary, the relation between the outputs and the inputs of an ANN can simply be written as

$$O = F(I; \omega_1, \ldots, \omega_n, A), \tag{7}$$

where $\omega_1, \ldots, \omega_n$ are the weights, which are determined during the learning stage by a minimization algorithm. ANNs are thus complementary to the k-NN approach in the sense that it is a *global* fitting method; the ANN is optimized to recognize most of the objects in the LS so that the response of the ANN to a peculiar value of the input depends on *all* the inputs present in the LS. The interested reader will find a complete description of ANNs in Bishop (1995).[9]

In the present work, we have made use of networks called multilayer perceptrons (MLPs), which code has been developed by A. Munoz and incorporated in a more general data mining software called GTDIDT (GNU Top Down Induction of Decision Trees), kindly made available to us by Prof. L. Wehenkel (Wehenkel 1997). The optimal weights are found by the minimization of a squared error function, using the Levenberg–Marquardt conjugate gradient technique. The weights are updated each time the whole LS is processed by the network (batch learning).

ANNs perform quickly on data, but they suffer from the so-called *generalization problem*. The latter has a statistical and a physical side. The statistical aspect is related to the overfitting problem. The ANN architecture must sufficiently be complex to map the diversity of the inputs to the output targets, thus reducing the bias, but also, it should not be too complex in order to avoid the overfitting of the LS, which makes the model unable to cope with the statistical noise present in the data, i.e. having a large variance. The model variance and the overfitting are minimized with LS as large as possible (built from synthetic data bases), including the population cosmic variance. The complexity of the ANN is then increased to reduce the bias, while overfitting is avoided by stopping the minimization when the variance starts to increase in a small subsample of the LS. The latter is randomly extracted from the LS and is not used in the learning process. We found that architectures with two hidden layers of 10 neurones each are sufficient in most of the cases.

The physical aspect of the *generalization problem* resides in the inability of the ANNs to extrapolate out of the learnt parameter space. For example, if the LS consists of QSOs with $0 < z < 5$, any object – even a star – will be assigned a redshift between $0 < z < 5$. More generally, supervised ANNs are unable to deal with *unknown objects*. For the same reason, ANNs are also poor interpolator on the edges of the intervals of the input parameters and when data are missing. Thus, a generic and empirical guideline to create a LS is to uniformly cover a range of the APs *wider* than expected among true objects and/or to duplicate in the LS the objects located on the edges of the parameter space, in order to give them a stronger weight.

The *generalization problem* also implies that rather than solving a big problem (characterized by a large intrinsic variance) with a complex neurone architecture, a priori knowledge should be used whenever possible to break the problem into several simpler ones, which would then be more efficiently solved with several 'specialized' networks. Thus, the definition of the (several) LS is a critical step in the efficiency of a network (see Sections 4.2 and 5.3, for more specific considerations related to QSO identification and redshift estimations).

Finally, let us mention here that, since ANN inputs are fluxes, the *generalization problem* sharpens the need of an accurate flux calibration.

3.3 Spectral principal component

We also studied the use of the SPCs to determine the shape of QSO spectra as well as their redshift, given by the photometric measurements provided by *Gaia*. We made no attempt to use this method to identify QSOs amongst stars.

Given a redshift z, the shape of a QSO spectrum can be determined by a method basically identical to the one described in Section 2.3.3 We try to find the solution $\boldsymbol{w}^{\mathrm{S}}$ of the system

$$\mathbf{A}^{\mathrm{F}}\,\boldsymbol{w} = \boldsymbol{f} \tag{8}$$

under the constraint that the spectrum corresponding to the solution of this system is positive for all λ, which is mathematically written by the system

$$\mathbf{A}\,\boldsymbol{w} > 0. \tag{9}$$

As above, $\boldsymbol{f}$ is the vector containing the photometric measurements f_i. The ith row of the matrix $\mathbf{A}^{\mathrm{F}}(i, j)$ contains the integral over the passband of the ith filter of the (relative) flux of the mean ($j = 1$) and principal component spectra ($j = 2$–11), each of them being redshifted and multiplied by the transmission spectrum of the IGM for that redshift. Therefore, the matrix $\mathbf{A}^{\mathrm{F}}$ is similar to matrix $\mathbf{A}$ defined in Section 2.3.3 for the considered redshift, but instead of containing one set of principal components per wavelength, it contains one set of principal components per filter.

Coming back to the matrix $\mathbf{A}$, one realizes that its large size (one row per wavelength bin) hampers a fast numerical solution.

[9] There is also an excellent website maintained by Warren S. Sarle and answering many FAQs: ftp://ftp.sas.com/pub/neural/FAQ.html

However, one can see that the constraints represented by equation (9) are always met provided

$$\mathbf{A}^{\rm ext}\,\boldsymbol{w} > 0, \tag{10}$$

where $\mathbf{A}^{\rm ext}$ is a subset of $\mathbf{A}$ that only contains rows corresponding to a wavelength for which either the mean spectrum or at least one principal component spectrum shows either a minimum or a maximum.

Applying $\mathbf{A}^{\rm F}$ to the solution $\boldsymbol{w}^{\rm S}$, we obtain the 'modelled' vector of photometric measurements $\boldsymbol{f}^{\rm S}$, for which we can calculate the χ^2: $\chi^2 = \sum_{i=1}^{n}[(f_i^{\rm S} - f_i)/\sigma_i]^2$. Therefore, this method can be expanded to also determine the QSO redshift by searching for the minimum of the $\chi^2(z)$ as a function of z. Note that equation (8) has even no solutions for a large number of redshift values which are thus excluded.

4 QSO IDENTIFICATION

Whatever the method, the photometric identification of quasars among stars relies on the signature of either their strong emission lines (UV excess like, for moderate redshifts) or the Lyα break absorption [large value of a specific colour index: $(B - V)$ for $3 \lesssim z \lesssim 4$, $(V - R)$ for $4 \lesssim z \lesssim 4.5$ and $(R - I)$ for $4.5 \lesssim z \lesssim 6$].

In the following, and excepted when stated otherwise, the TS is built with the data bases STAR-RAN2+BIN-RAN2+WD-REG2+QSO-REG (see Table 3) and contains 97 787 objects for each magnitude $G = 18$, 19 and 20. It should be noted that this TS does not reflect the true proportion of objects, and is only representative of the diversity of the objects.

From the 25 noise realizations of the TS, the probabilities $P_{\rm QSO}$, $P_{\rm STAR}$ and $P_{\rm WD}$ of an object to be a QSO, a star (without distinction between singles and binaries) or a white dwarf given the Poissonian noise are derived from the relative frequency of attribution of their respective classes. The final classification rule is as follows.

(i) QSO if $P_{\rm QSO} > P_{\rm STAR} + c$ and $P_{\rm QSO} > P_{\rm WD}$;
(ii) STAR if $P_{\rm STAR} > P_{\rm QSO} - c$ and $P_{\rm STAR} > P_{\rm WD}$;
(iii) WD if $P_{\rm WD} > P_{\rm QSO} - c$ and $P_{\rm WD} > P_{\rm STAR}$;

where $c\,(<1)$ is a *contrast parameter* intended to minimize the stellar contamination.

We now present the classification performances of k-NN and ANNs.

4.1 Nearest neighbours

The traditional criterion to define the colour excess of a QSO with respect to stellar colours is in fact similar to a minimum distance in a two-dimensional space. The generalization to higher dimensions can be done through χ^2 template fitting (Hatziminaoglou, Mathez & Pelló 2000). However, the template grid should not only cover the observed variety of QSO spectra, but it should be dense enough so that a good matching can be performed. Indeed, the emission-line signature of a QSO typically appears in only a couple of nearby filters, and it should be strong enough not to be smeared out in a global fitting over many filters, and choosing the relevant filters is almost an impossible task since the QSO signatures depend on its (a priori unknown) redshift.

Of course, once a QSO is correctly identified with the k-NN method, its redshift and astrophysical properties may be derived. However, we only discuss here the classification issue, and we postpone the determination of the other parameters to Section 5.2 because both tasks could in principle be performed by separate algorithms.

Table 4. Confusion matrix with the χ^2 method and $c = 0.95$ (results in per cent; 1X+2B/2F+2B).

	G	True STAR	True QSO	True WD
Class STAR	18	99.99/99.97	07.24/07.78	00.00/00.00
	19	99.98/99.96	15.21/14.11	00.00/00.62
	20	99.26/99.56	28.94/26.61	11.11/16.05
Class QSO	18	00.00/00.01	92.41/91.60	00.00/00.00
	19	00.01/00.01	83.94/84.84	00.00/00.00
	20	00.39/00.22	68.39/71.07	00.00/00.00
Class WD	18	00.01/00.02	00.35/00.62	100.0/100.0
	19	00.01/00.03	00.85/01.06	100.0/99.38
	20	00.35/00.22	02.67/02.32	88.89/83.95

Here, we test the classification performances of the k-NN method precisely in the sense of a simple χ^2 template fitting. The LS (or the reference data base) consists of the *noiseless* versions of data bases STAR-RAN1 + BIN-RAN1 + subset of 81 white dwarfs from WD-REG1 + subset of 4000 QSOs from QSO-RAN (i.e. 15 081 objects). Since stars are considered as the main contaminant, they are over-represented in the LS in order to define (and exclude) at best their locus in the high-dimensional colour space. Note, however, that in this study, the LS and TS do *not* reflect the true distribution of objects in the sky. The latter will be included a posteriori in Section 4.3.

Table 4 presents, for several G mag and both adopted photometric systems, the confusion matrix obtained with the χ^2 method, i.e. the percentage of objects of each type (stars, white dwarfs and QSOs) presents in the TS (defined in Section 4) to be found in each class. The most important point is to check what objects are going to be classified as QSOs.

Four general results can be derived from Table 4.

(i) The theoretical 'completeness' of the QSO sample is good: between 90 and 70 per cent of all QSO types can be retrieved, depending on the G mag.

(ii) The stellar contamination is low: even at $G = 20$, only 0.4 per cent of all stellar types are confused with QSOs. How critical is this contamination depends on the population ratio between those stellar types and QSOs.

(iii) White dwarfs – which are traditional contaminants in the UV excess method – are very efficiently rejected by means of the *Gaia* photometry.

(iv) The two adopted photometric systems have similar performances. We have selected the 1X+2B system in the remainder.

The next step in the analysis is to find which kind of QSOs are correctly identified and which kind of stars are contaminating.

Fig. 10 shows that the completeness level of QSOs is not uniform as a function of their redshift and APs. The left-hand panel explores the influence of W and A_v for typical continuum slope ($-0.5 \leqslant \alpha \leqslant 0.5$), while the right-hand panel shows the influence of α for typical values of W in the range $2500 \leqslant W \leqslant 10\,000$ and $A_v = 0$. We can infer the following results.

(i) The completeness is the highest for $z \gtrsim 3.5$, when the Lyα break signature is detected in the bluest filters.

(ii) The completeness is the lowest for $2 \lesssim z \lesssim 3$, when QSOs look the more similar to normal stars (this is a tribute to pay the high level of rejection of stars).

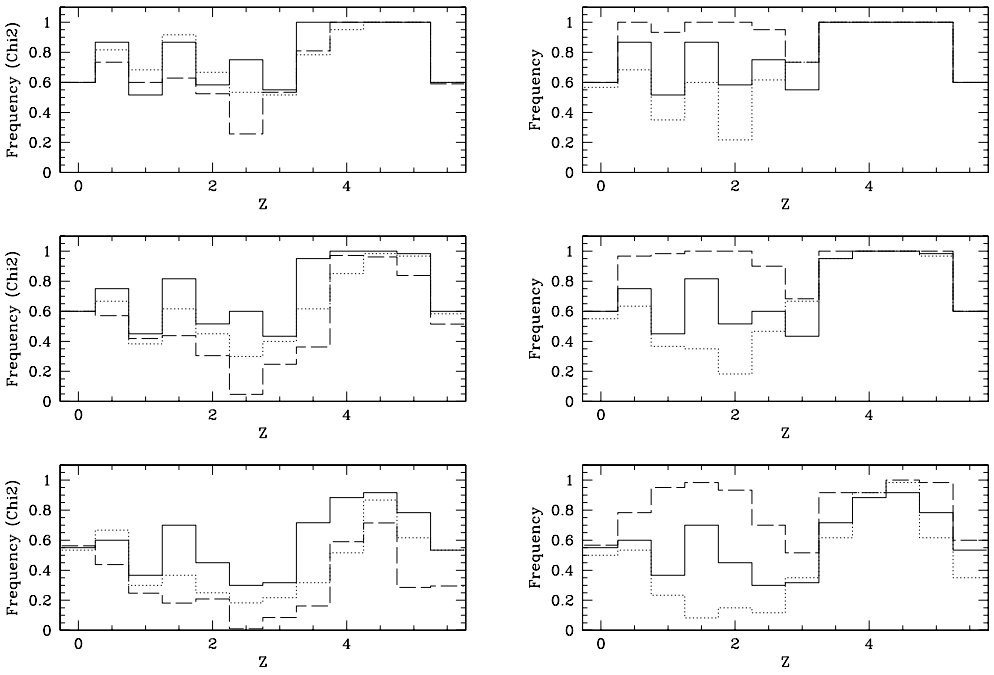

Figure 10. Redshift distributions of the QSOs found with the χ^2 method for $G = 18$ (top), $G = 19$ (middle) and $G = 20$ (bottom). Left-hand panel: $-0.5 \leqslant \alpha \leqslant 0.5$ and solid: $2500 \leqslant W \leqslant 10\,000$, $A_V = 0$; dotted: $2500 \leqslant W \leqslant 10\,000$, $A_V = 2$; dashed: $W < 2500$, $A_v = 0$. Right-hand panel: $2500 \leqslant W \leqslant 10\,000$, $A_V = 0$ and solid: $-0.5 \leqslant \alpha \leqslant 0.5$; dotted: $1 \leqslant \alpha \leqslant 2$; dashed: $-2 \leqslant \alpha \leqslant -1$.

(iii) Even for $G = 18$, the completeness of typical, unreddened low-redshift QSOs is not larger than ∼70 per cent. Most lost QSOs are confused with *low metallicity, reddened* ($1 \leqslant A_V \leqslant 2$) hot stars.

(iv) The extra loss of weak emission-line objects becomes important at $G = 19$.

(v) The extra loss of reddened QSOs becomes important at $G = 20$.

(vi) Steep blue continuum QSOs ($\alpha > 1$) are preferentially lost, while the completeness of flat or slightly red continuum QSOs is much higher than the mean. Indeed, blue continuum QSOs look more similar to hot stars.

Conversely, most of the stellar contaminants at $G = 20$ are *low metallicity, reddened* stars ([Fe/H] < -1; $A_V > 2$) confused with reddened QSOs whose intrinsic continuum is flat or red ($\alpha < -1$).

As expected, the smoothing effect of the reddening is harmful, especially when the S/N degrades. Identifying reddened QSOs among reddened stars, as required within the galactic plane, will thus be difficult on the basis of photometric data only. On the other hand, at high galactic latitudes where reddening is negligible, the stellar contamination is expected to be low (see Section 4.3).

Finally, we made a crosscheck of the performances of the χ^2 with the independent *QSO-PCA* TS of 20 000 QSOs (see Table 3). The completeness level is found to range between 87 and 60 per cent from $G = 18$ to 20.

4.2 Neural network analysis

The most important step when dealing with ANN(s) is the definition of the optimal LS(s). Of course, it depends on the task assigned to the ANN(s). After testing several approaches, we identified the following empirical 'golden rules' to properly train the ANNs in view of our goal, i.e. selecting QSOs with the highest rejection rate of stellar contaminants.

(i) To identify QSOs among n kinds of contaminants, it is more efficient to design n specialized networks, rather than defining one big ANN performing the global classification. The ith ANN is trained to perform a binary classification between QSOs and the ith type of contaminants and yields the probability $p_{i,\mathrm{QSO}}$ that an object is a QSO. The final probabilities P_{QSO} and $P_{i,\mathrm{c}}$ to belong to the class QSO or to the i th contaminant class are given by

$$P_{\mathrm{QSO}} = \frac{\sum_{i=1}^{n} p_{i,\mathrm{QSO}}}{n\mathrm{Norm}}, \tag{11}$$

$$P_{i,\mathrm{c}} = \frac{1 - p_{i,\mathrm{QSO}}}{\mathrm{Norm}}, \tag{12}$$

with

$$\mathrm{Norm} = \frac{\sum_{i=1}^{n} p_{i,\mathrm{QSO}}}{n} + \sum_{i=1}^{n} (1 - p_{i,\mathrm{QSO}}). \tag{13}$$

(ii) The rejection rate is higher when the proportion of contaminants is larger in the LS. Indeed, the ANN should learn at best what it has to reject.

(iii) The classification of objects located on the edges of the parameter space is improved by duplicating the corresponding objects in the LS. This extra weight forces the ANN to better interpolate on the edges.

(iv) A better classification is reached when the ANN is trained on a LS with more noise than expected in the data (this again improves the interpolation). For classification, we systematically applied ANNs trained with objects one or two magnitude(s) fainter than the targets (of course, the magnitude offset is corrected for, but not the noise).

Consequently, for $G = 19$, 20 and 21, and for 1X+2B and 2F+2B photometric systems, one group of *LSs* is built to perform the QSO/STAR classification and another one to perform the QSO/WD

Table 5. Confusion matrix with ANNs. The contrast factor $c = 0.95$, except for 1X+2B and $G = 18$ (respectively $G = 19$), where $c = 0.8$ (respectively $c = 0.85$) (results in per cent; 1X+2B/2F+2B).

	G	True STAR	True QSO	True WD
Class STAR	18	100.0/99.95	30.07/25.12	62.96/34.40
	19	99.92/99.93	35.05/35.60	28.40/29.63
	20	99.92/99.82	63.92/55.74	69.14/43.83
Class QSO	18	00.00/00.05	68.54/72.40	00.00/00.00
	19	00.00/00.03	62.48/59.23	00.00/00.00
	20	00.00/00.02	31.79/40.37	00.00/00.00
Class WD	18	00.00/00.00	01.38/01.13	37.04/63.60
	19	00.08/00.04	02.47/05.17	71.60/70.37
	20	00.08/00.16	04.29/03.89	30.86/56.17

classification. Each LS of the first group contains 36 976 objects (*STAR-RAN1* + *BIN-RAN* + 4000 QSOs randomly selected from *QSO-RAN* + 21 976 duplicated stars from *STAR-RAN1* – a star maybe duplicated several times if more than one parameter has an extreme value). In the second group, each LS consists of 4972 objects (4000 QSOs from *QSO-RAN* + *WD-REG1*). ANNs have been trained several times to check that the classification results were not depending on the peculiar values of the neural weights obtained after the minimization process.

Table 5 presents, for several G mag and both adopted photometric systems, the confusion matrix obtained with ANNs, i.e. the percentage of objects of each type (stars, white dwarfs and QSOs) presents in the TS (defined in Section 4) to be found in each class. The most important point is to check what objects are going to be classified as QSOs.

When comparing with Table 4, several results are striking.

(i) Properly trained ANNs are better than χ^2 to reject stellar contaminants in the QSO catalogue. This is especially true with the 1X+2B photometric system with which virtually *no* stellar spectrum is classified as a QSO (error rate smaller than 10^{-4}), even at $G = 20$. With the 2F+2B photometric system, the error rate is also smaller than with the χ^2 technique.

(ii) The optimization of the ANNs to reject stars has a price: the fraction of lost QSOs is much larger than with the χ^2 method. Indeed, from $G = 18$ to $G = 20$, 30 to 64 per cent of the QSO spectra are classified as stars.

(iii) WDs do not contaminate QSOs, but more WDs are classified as stars than with the χ^2 method. This is a consequence of the binary tests 'against' the QSOs in the situation where $P_{\rm WD} = P_{\rm STAR} = 0.5$. The result could be improved with the application of a third ANN trained to distinguish between WDs and stars, but we are not interested in such a clean selection here.

(iv) The 1X+2B system is slightly better in terms of stellar rejection; this could be due to the discriminating power of the reddest filter at 10 000 Å.

Fig. 11 shows the redshift distributions of QSOs found with the ANN method and should be compared with Fig. 10. Although the general influences of A_v, α and W are similar, the loss of QSOs with $z \simeq 2$ and $G = 20$ (i.e. typical of most of the QSOs to be detected) is much more pronounced. This is the result of our star-rejection policy. However, this is such an extreme situation that we should wonder whether we could relax our selection criteria a little bit. When the contrast parameter c is decreased, entering contaminating stars are mostly *cold, very low metallicity red dwarfs*, which are on the edge of the parameter space. Besides the fact that those stars are faint (so they must be close and their proper motion will be detected by *Gaia*), they are also rare. Therefore, as we shall see in Section 4.3, even with $c = 0$, the observed contamination level remains extremely low.

Finally, we applied the ANNs to the independent set of 20 000 QSO present in the data base QSO-PCA (see Table 3). The fraction of correctly identified QSO with the 1X+2B (respectively the 2F+2B)

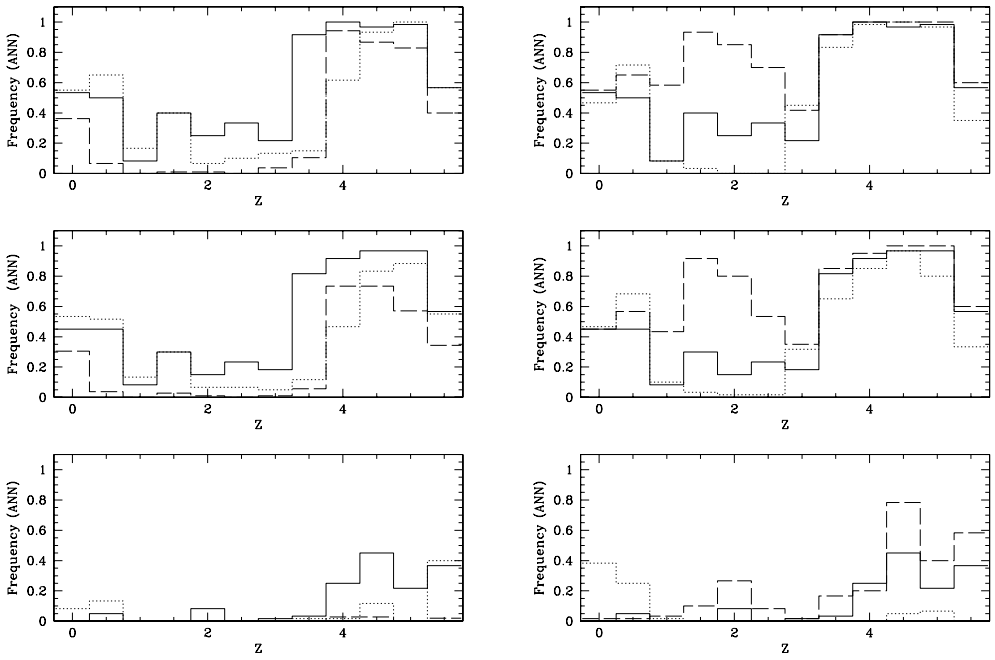

Figure 11. Redshift distributions of the QSOs found with the ANN method for $G = 18$ (top), $G = 19$ (middle) and $G = 20$ (bottom). Left-hand panel: $-0.5 \leqslant \alpha \leqslant 0.5$ and solid: $2500 \leqslant W \leqslant 10\,000$, $A_V = 0$; dotted: $2500 \leqslant W \leqslant 10\,000$, $A_V = 2$; dashed: $W < 2500$, $A_v = 0$. Right-hand panel: $2500 \leqslant W \leqslant 10\,000$, $A_V = 0$ and solid: $-0.5 \leqslant \alpha \leqslant 0.5$; dotted: $1 \leqslant \alpha \leqslant 2$; dashed: $-2 \leqslant \alpha \leqslant -1$.

photometric system is 61.6 per cent (respectively 56.2 per cent) for $G = 18$, but only 10.3 per cent (respectively 17.0 per cent) for $G = 20$. Relaxing the contrast from $c = 0.95$ to 0.2 (i.e. slightly contaminating with red stars) at $G = 20$ increases the completeness only to 29.3 per cent (respectively 37.6 per cent), i.e. about half the value obtained with the χ^2. Therefore, clearly, while ANNs can be tuned to nearly perfectly eliminate contaminating stars, they more often fail to correctly interpolate in the larger family of QSO spectra. This also indicates the danger of including only synthetic libraries in the LS.

4.3 Completeness and contamination with observed populations

In order to compute the completeness of the QSO catalogue to be built by *Gaia* and to estimate the stellar contamination rate (CR), the efficiency of the photometric survey as a function of the APs (presented in the two last sections) must be weighed by the observed distribution of objects as a function of their APs. Of course, this distribution does only exist for *known* populations of objects. On the other hand, the CR also depends on the stellar to QSO population ratio $(N_*/N_{\rm QSO})(b, G)$, which is a function of the galactic latitude b and magnitude G.

Let us first assume that each object i in the TS can be associated with a probability of observation $p(AP_i)$, given its APs AP_i. The completeness of the QSO catalogue, $P(\mathrm{QSO}|\mathrm{QSO})$, is simply given by the ratio between the sum over the probabilities of observation of correctly identified QSO and the sum over the probabilities of observation of all QSOs in the TS:

$$P(\mathrm{QSO}|\mathrm{QSO}) = \frac{\sum_{i,\,\text{identified QSOs}} p(AP_i)}{\sum_{i,\,\text{all QSOs}} p(AP_i)}. \quad (14)$$

Similarly, the CR by stars, *CR*, at galactic latitude b is given by

$$CR(b, G) = \frac{P(\mathrm{QSO}|\mathrm{STAR})\frac{N_*}{N_{\rm QSO}}(b, G)}{P(\mathrm{QSO}|\mathrm{STAR})\frac{N_*}{N_{\rm QSO}}(b, G) + P(\mathrm{QSO}|\mathrm{QSO})}, \quad (15)$$

where

$$P(\mathrm{QSO}|\mathrm{STAR}) = \frac{\sum_{i,\,\text{unidentified stars}} p(AP_i)}{\sum_{i,\,\text{all stars}} p(AP_i)}. \quad (16)$$

The population ratio has been estimated as a function of magnitude and galactic latitude from the predicted stellar number counts (Bahcall & Soneira 1980) and from the observed QSO number counts (Hartwick & Schade 1990).

Results are summarized in Table 6. Not surprisingly, compared to the mean population ratio $\langle N_*/N_{\rm QSO}\rangle \simeq 2000$ expected in the *Gaia* mission, Table 6 confirms that the local population ratio is much more favourable for the identification of QSOs at high galactic latitude, while it represents a huge obstacle to the realization of a clean QSO catalogue in the galactic plane. The steeper QSO number count relation favours fainter magnitudes, but this is in competition with a worse classification efficiency due to lower S/N.

Table 6. Stellar to QSO population ratio as a function of galactic latitude b and magnitude G.

$N_*/N_{\rm QSO}$	$G = 18$	$G = 19$	$G = 20$
$b = 90$	725	175	75
60	1150	280	120
30	3075	765	330
10	33 000	10 500	5600

The probability distribution of observation of the QSO APs has been derived from the observed distributions in the SDSS DR1β. Although a slight correlation is seen between W and z (i.e. the Baldwin effect), the redshift distribution has been assumed to be independent of the (α, W) distribution that we obtained from template fitting (see Fig. 5). Let us note that the depth of the SDSS survey being comparable with what *Gaia* will reach, the peak of the redshift probability distribution at $z \sim 1.7$ is quite realistic.

In order to predict the probability distribution of observation of the stellar APs, we made use of the Besançon Galactic model (Robin et al. 2003). We selected one line of sight at $l = 0^\circ$, $b = 10^\circ$, as being representative of the bulge population, and one line of sight at $l = 200^\circ$, $b = 60^\circ$, as being representative of the halo population. In the latter case, the apparent population is strongly dependent on the magnitude (late-type stars dominate only at faint magnitude) so that we computed two halo probability distributions: one for $G = 20$ and the other for $G \leqslant 19$.

We did not adopt any peculiar probability law to describe the extinction; instead, we kept a uniform distribution.

Applying equations (14) and (16) in conjunction with classification algorithms described in Sections 4.1 and 4.2 yields the results shown in Table 7. The values of the contrast parameter c have been relaxed with respect to their previous values as long as contaminating stars consist in marginal objects according to the Besançon model. The main conclusions are as follows.

(i) ANNs are better than χ^2 in terms of CR. They enable to reach a rejection level better than 99.999 per cent, virtually leading to 0 per cent contamination even at $G = 20$.

(ii) χ^2 is better than ANN in terms of completeness. A better completeness should certainly be achieved with ANNs by including information on the redshift distribution in the LS, but we did not want to bias the searching algorithm at this level. However, a secure sample containing 16 per cent of the QSOs detected by *Gaia* would be valuable for a precise determination of the referential frame.

(iii) To reach the highest efficiency, the classification algorithm could be chosen as a function of galactic latitude. For high b, χ^2 would offer the best completeness while close to the galactic plane, the confusion would be minimized with ANNs.

We would like to mention that, in the case of the χ^2 method, the estimate of the QSO completeness could take into account the probability of observation of the stars with which QSOs are confused. This a posteriori test would only increase the completeness. However, it cannot be made with the ANN method.

5 QSO PHOTOMETRIC REDSHIFTS AND APS

This section is devoted to the determination of the photometric redshift (and APs when possible), with k-NN, ANN and SPC. The first part introduces special features of QSO spectra which may disturb the photometric determination of the redshift.

5.1 Introduction and intrinsic correlations

The determination of photometric redshifts has a rather long story when dealing with *galaxies* (see Koo 1999, for a review). However, QSO candidates were so rare in the past that spectroscopic confirmation could be afforded for most of them. This situation is changing now with the advent of big automated sky surveys. Studies of QSO photometric redshift have been made in the frame of the SDSS (Budavári et al. 2001; Richards et al. 2001), on the basis

Table 7. Predicted completeness and stellar CR in the *Gaia* QSO catalogue, including observed distributions of APs. Left-hand side: with χ^2; right-hand side: with ANNs.

	QSO completeness (per cent) (1X+2B/2F+2B)	Stellar CR (per cent) (1X+2B/2F+2B)	QSO completeness (per cent) (1X+2B/2F+2B)	Stellar CR (per cent) (1X+2B/2F+2B)
	χ^2		ANN	
$G = 18; b = 90$	90.2/87.3	0.00/0.00	46.6/41.1	0.00/0.00
$G = 19; b = 90$	73.7/78.1	0.00/0.00	46.9/53.7	0.00/0.00
$G = 20; b = 90$	54.6/49.1	01.4/00.5	16.0/16.2	0.00/0.00
$b = 60$	54.6/49.1	02.3/00.7	16.0/16.2	0.00/0.00
$b = 30$	54.6/49.1	06.5/02.0	16.0/16.2	0.00/0.00
$b = 10$	54.649.1	95.0/98.1	16.0/16.2	0.00/0.00

of observed colours of QSOs with known redshifts. Very recently, QSO photometric redshifts have been obtained for the *Chandra* Deep Field South with COMBO-17 by Wolf et al. 2004.

Unlike galaxy spectra, QSO spectra do not fall into specific classes but exhibit a wide range of shapes, making more difficult the definition of reference templates. The redshift should a priori be easy to estimate since it affects the apparent wavelengths of strong emission lines and of the Lyα break (for $z \gtrsim 3$): it is the first source of variance among QSO spectra. However, the colours of a pure power-law spectrum (typical of the QSO continuum) are insensitive to the redshift. Even when emission lines are seen, wrong redshift estimates can be made, especially when the Lyα break is not seen through the adopted photometric system. This is due to the interplay between the limited wavelength coverage of the photometric system and intrinsic wavelength shifts between strong emission lines (e.g. $\lambda_{\rm C\,IV}/\lambda_{{\rm Ly}\alpha} \simeq \lambda_{\rm C\,III]}/\lambda_{\rm C\,IV} \simeq 1.25 \pm 0.2$ and $\lambda_{\rm C\,III}/\lambda_{{\rm Ly}\beta} \simeq \lambda_{\rm Mg\,II}/\lambda_{\rm C\,IV} \simeq \lambda_{{\rm H}\beta}/\lambda_{\rm Mg\,II} \simeq 1.80 \pm 0.05$) which can lead to local degeneracies.

In order to quantify this effect, we computed the correlation C expected between the colours of two QSOs at redshifts z_i and z_j observed in a given photometric system:

$$C(z_i, z_j) = \frac{\sum_{k=1}^{n_{\rm f}} F_{z_i}(k) F_{z_j}(k)}{\sqrt{\sum_{k=1}^{n_{\rm f}} F_{z_i}(k)^2 \sum_{k=1}^{n_{\rm f}} F_{z_j}(k)^2}}, \qquad (17)$$

where $n_{\rm f}$ is the number of filters and $F_{z_i}(k)$ is the noise-free expected flux in filter k of a QSO at redshift z_i. Fig. 12 shows the contours of C in the plane $(z_i, z_j - z_i)$ for the 1X+2B and 2F+2B photometric systems. A correlation larger than 0.99 is expected in the former filters for spectra with $0.55 < z < 2$, i.e. when Hα is out of the reddest filter while the strong contrast between Lyα and the break is not yet visible in the two bluest filters. In the 2F+2B system, the last MBP filter is centred at a shorter wavelength and a confusion with $C > 0.99$ spreads down to $z \simeq 0.3$. Fig. 12 also confirms the intuitive prediction that the stronger the emission lines the more different are the spectra.

Removing the BBP filters decreases a little bit the correlation but at the expense of local degeneracies due to individual lines; removing the MBP filters leads to a catastrophic increase of the correlation up to $z \simeq 3$.

Therefore, independent of the adopted analysis technique (χ^2, ANN or SPC), Fig. 12 predicts that the determination of the photometric redshift will be most sensitive to noise in the range $0.5 < z < 2$. This result is nearly independent of the adopted MBP+BBP photometric system.

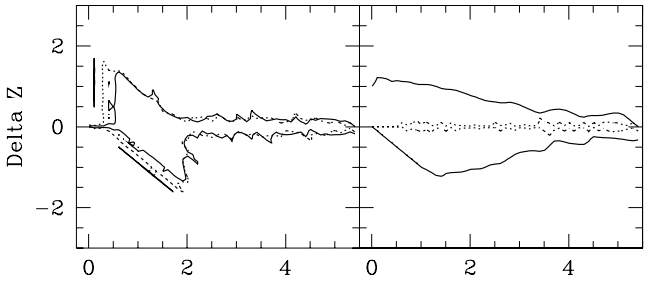

Figure 12. Contours of correlation at $C = 0.99$ (see equation 16). For each value of z, the contours give the maximal shifts in redshift for which the spectra has a correlation index higher than 0.99. Left-hand panel: QSO with $(\alpha, W) = (0.03, 10\,000)$ for the 1X+2B system (full lines) and for the 2F+2B system (dotted lines). Right-hand panel: influence of W in the 1X+2B system ($W = 0$: full lines; $W = 50\,000$: dotted lines).

5.2 *k*-NN analysis

First, we have performed the nearest neighbour analysis in the sense of minimum χ^2 (see Section 3.1) to determine the photometric redshift (and other QSO APs). The LS, or reference template grid, consists of the 20 000 QSOs in the noiseless QSO-RAN data base while the QSO-REG data bases of 86 625 QSOs at $G = 18$, 19 and 20 represent the different TS (see Table 3).

Fig. 13 shows the value of $\Delta z = z_{\rm phot} - z_{\rm spec}$ between the redshift estimated from the *Gaia* photometry, $z_{\rm phot}$, and the 'true' redshift, $z_{\rm spec}$, as a function of $z_{\rm spec}$, for selected values of W of a typical QSO with $G = 19$, $\alpha = 0$ and $A_v = 0$. The error bars represent the 1σ dispersions of 25 simulations (see Section 2.5) and are due to the influence of Poissonian noise. The central values of Δz show the mean error as a function of the spectroscopic redshift. As expected, the smallest errors on the photometric redshift occur for large W ($\gtrsim$10 000) and also for large redshifts ($z_{\rm spec} > 2.5$), i.e. when the Lyα break is dominant. Some points are deviant in the sense that the redshift error is much larger than the error bar. This reflects the influence of the limited resolution of the template grid.[10] Although the observed AP distributions in the QSO population are not used, a general idea of the achievable precision on the photometric redshift is obtained for different z ranges, W ranges and $A_v \neq 0$ after taking the median and quartiles over typical QSOs with $-1 < \alpha < 1$. The results are shown in Table 8. They exhibit the same trend with

[10] It should be noted that the value of the minimum χ^2 does *not* give a robust estimate of the confidence on the redshift estimation. Indeed, good χ^2 may hide wrong photometric redshifts due to spectral similarities (see Section 5.1), while large χ^2 may simply reflect the lack of interpolation at this stage within the finite template grid.

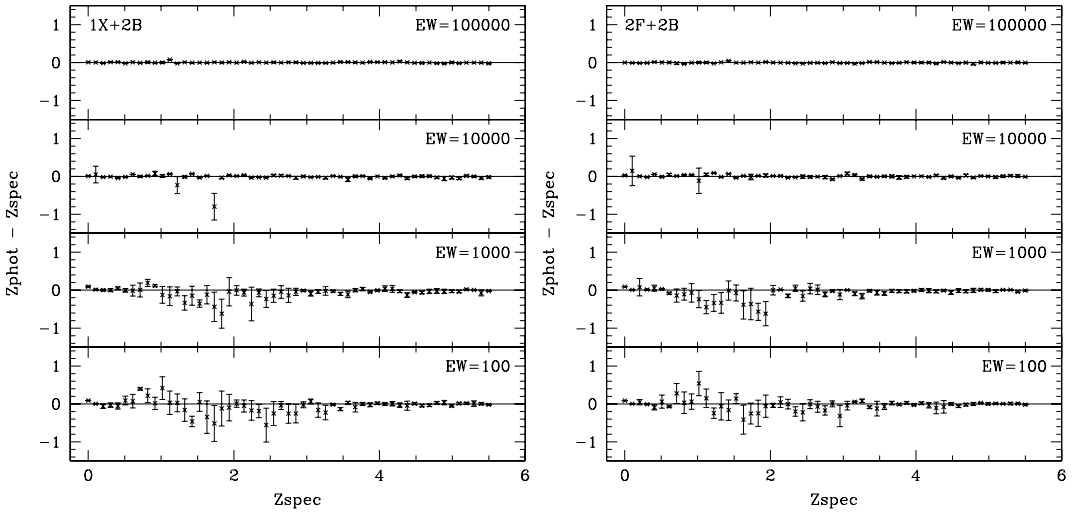

Figure 13. Difference between the photometric redshift $z_{\rm phot}$ obtained with minimum χ^2 and the spectroscopic redshift $z_{\rm spec}$ as a function of $z_{\rm spec}$ for $G = 19$, $\alpha = 0$ and $A_v = 0$ and for different values of W. Left-hand panel: 1X+2B photometric system; right-hand panel: 2F+2B photometric system.

Table 8. Median of the absolute error $|\Delta z|_{\rm med}$ on the photometric redshifts determined with minimum χ^2 for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.08/0.09	0.02/0.02	0.03/0.04	0.02/0.02	0.01/0.01	0.007/0.008
	$A_v = 1$	0.08/0.09	0.02/0.02	0.02/0.03	0.01/0.01	0.006/0.007	0.005/0.005
$G = 19$	$A_v = 0$	0.11/0.11	0.03/0.02	0.03/0.04	0.02/0.02	0.01/0.01	0.007/0.007
	$A_v = 1$	0.13/0.13	0.03/0.02	0.02/0.03	0.01/0.01	0.005/0.006	0.005/0.005
$G = 20$	$A_v = 0$	0.22/0.21	0.07/0.06	0.07/0.08	0.03/0.03	0.009/0.01	0.009/0.008
	$A_v = 1$	0.28/0.26	0.08/0.06	0.11/0.10	0.03/0.02	0.005/0.006	0.008/0.007

respect to $z_{\rm spec}$ and W as in Fig. 13. The median is not sensitive to outliers and gives a more robust estimate of the central value of the redshift error. As seen in Table 8, the latter ranges from 0.2–0.3 to less than 0.01 from low W, $z < 2.5$ QSOs (the worst cases actually correspond to $1.5 < z < 2$) to high W, high-z QSOs. A reasonable amount of extinction does not significantly affect the accuracy on the photometric redshift. Therefore, except for BL Lac and weak emission-line objects at $z \sim 2$, the results are rather encouraging. They do not significantly depend on the adopted filter system.

A more thorough inquiry on the APs of the QSOs for which the estimate of the photometric redshift is not good, i.e. $|\Delta z| > 0.2$ can be read off in Fig. 14 ($G = 19$). Besides the QSOs with $z < 2.5$ and small W, the red solid triangles indicate that QSOs with a very steep blue continuum $\alpha > 2$ also suffer from a bad redshift estimation. This is not surprising since the emission lines are swamped by the huge continuum variation. Nevertheless, such objects would rather be exotic and are only included in the data base to increase the diversity of the LS.

When an object is classified as a QSO, other APs may possibly be retrieved. Tables 9, 10 and 11 show the median of the absolute errors on α, A_V and W, respectively, for $-1 \leqslant \alpha \leqslant 1$ and $G = 18$, 19, 20. Several remarks can be made as follows.

(i) For all parameters, the sensitivity to the magnitude is the strongest for weak emission-line objects ($W \leqslant 3000$), where the absolute errors on α and A_v increase by a factor of 2 and the relative errors on W inflate by a factor of more than 10 from $G = 18$ to 20.

(ii) The parameter determination is systematically better for $A_v = 1$ than for $A_v = 0$. This is probably due to the fact that $A_v = 0$ is on the edge of the parameter space in the reference table and thus can only be overestimated.

(iii) The error on the determination of α is particularly correlated to the determination of A_v. When A_v is overestimated (e.g. when $A_{v,\rm spec} = 0$), α is also overestimated since the colours of the original spectrum may be restored by a bluer QSO affected by more extinction (see Fig. 15, for $G = 18$).

(iv) The strongest the emission lines, the smaller the accuracy on the determination of the spectrum slope.

(v) The overall accuracy on the APs is low with respect to that achieved on the redshift. This is not surprising since their imprints in the spectra are second-order features and are sometimes degenerated.

(vi) The accuracy on the parameter determination is similar for both adopted photometric systems.

In the second step, we have checked the robustness of the χ^2 method by looking for the photometric redshifts in the QSO-PCA-independent TS of 20 000 QSOs (see Table 3) and Section 2.3.

The error on their photometric redshift as a function of the spectroscopic redshift is given in Fig. 16. In order to get a comprehensive view of the dispersion, again, we used the median and quartiles. About 9 per cent of the QSOs have $|\Delta z| > 0.5$, but, as expected from Section 5.1, most of the dispersion occurs for $0.5 \lesssim z_{\rm spec} \lesssim 2$. In this redshift interval, the median of the absolute error on the redshift

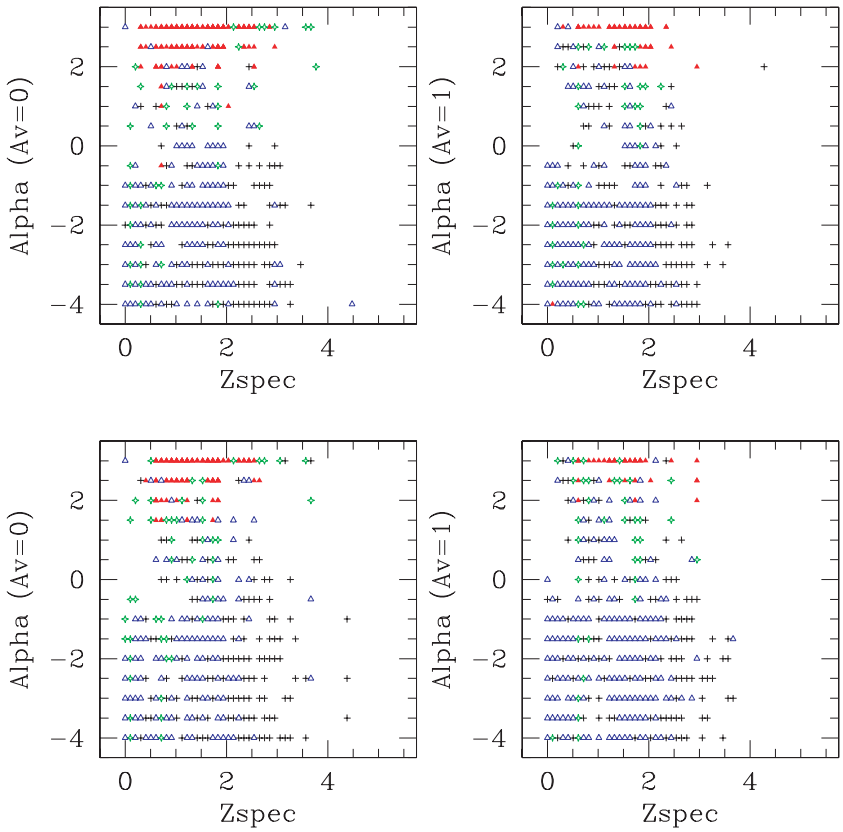

Figure 14. Regions of the $\alpha - z_{\rm spec}$ plane, where $|\Delta z| > 0.2$ with the χ^2 method, for QSOs with $W = 100$ (black +), 1000 (blue open triangles), 10 000 (green losanges) and 100 000 (red solid triangles). Top: 2F + 2B system; bottom: 1X + 2B system; left-hand panel: $A_v = 0$; right-hand panel: $A_v = 1$. ($G = 19$).

Table 9. Median of the absolute error $|\Delta\alpha|_{\rm med}$ on the QSO spectrum slope determined with minimum χ^2 for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.31/0.33	0.64/0.75	0.69/0.59	0.95/1.01	1.11/1.07	0.89/0.97
	$A_v = 1$	0.21/0.39	0.36/0.56	0.44/0.40	0.58/0.65	0.86/0.79	1.01/1.07
$G = 19$	$A_v = 0$	0.41/0.41	0.72/0.80	0.72/0.65	0.96/0.96	1.13/1.06	0.84/0.91
	$A_v = 1$	0.28/0.29	0.40/0.44	0.36/0.43	0.54/0.54	0.79/0.80	1.01/1.11
$G = 20$	$A_v = 0$	0.83/0.76	0.89/0.94	0.89/0.86	0.98/0.96	1.09/1.01	0.82/0.83
	$A_v = 1$	0.42/0.39	0.51/0.56	0.47/0.40	0.60/0.65	0.73/0.79	1.03/1.07

Table 10. Median of the absolute error $|\Delta A_v|_{\rm med}$ on the interstellar extinction (mag) determined with minimum χ^2 for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.32/0.32	0.66/0.71	0.60/0.50	0.75/0.77	0.20/0.20	0.20/0.20
	$A_v = 1$	0.22/0.24	0.30/0.35	0.37/0.40	0.30/0.30	0.20/0.20	0.17/0.16
$G = 19$	$A_v = 0$	0.43/0.41	0.78/0.81	0.67/0.58	0.78/0.80	0.21/0.20	0.20/0.20
	$A_v = 1$	0.28/0.29	0.31/0.31	0.30/0.33	0.28/0.27	0.20/0.18	0.13/0.14
$G = 20$	$A_v = 0$	0.82/0.73	1.03/1.06	0.80/0.72	0.81/0.85	0.28/0.25	0.23/0.24
	$A_v = 1$	0.32/0.32	0.32/0.31	0.30/0.30	0.24/0.26	0.16/0.15	0.18/0.16

is $|\Delta z|_{\rm med} \simeq 0.2$, nearly independent of the G mag, while $|\Delta z|_{\rm med} \simeq 0.03$–$0.05$ for $z_{\rm spec} > 2.5$ and $G = 18$–20. The latter results are expected from Table 8, but the former are significantly worst.

The expected dispersion due to photon noise ranges from $\sigma_z = 0.06$–0.24 in the former redshift interval to $\sigma_z = 0.02$–0.16 in the latter (for $G = 18$–20). This tells us that for $G < 20$, the interplay between QSO spectral diversity and spectral degeneracies dominates the influence of the photon noise on the determination of QSO photometric redshifts with *Gaia* in the $0.5 < z_{\rm spec} < 2$ interval. Unfortunately, this redshift interval is also the one

Table 11. Median of the absolute error $|(W_{\rm phot} - W_{\rm spec})/W_{\rm spec}|_{\rm med}$ on the equivalent width of the emission lines, determined with minimum χ^2 for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.73/0.76	0.95/1.12	0.61/0.78	2.03/2.30	0.62/0.66	0.72/0.70
	$A_v = 1$	0.66/0.65	0.53/0.58	0.45/0.45	0.40/0.45	0.48/0.48	0.55/0.56
$G = 19$	$A_v = 0$	1.83/1.38	2.43/2.58	0.74/0.82	2.97/3.11	0.61/0.71	0.80/0.74
	$A_v = 1$	1.60/1.23	0.88/0.77	0.40/0.43	0.45/0.44	0.48/0.47	0.55/0.54
$G = 20$	$A_v = 0$	19.22/12.17	13.07/13.29	2.12/2.12	6.37/6.20	0.69/0.76	1.09/0.89
	$A_v = 1$	13.40/9.90	5.40/4.10	0.90/0.64	0.92/1.05	0.51/0.47	0.61/0.62

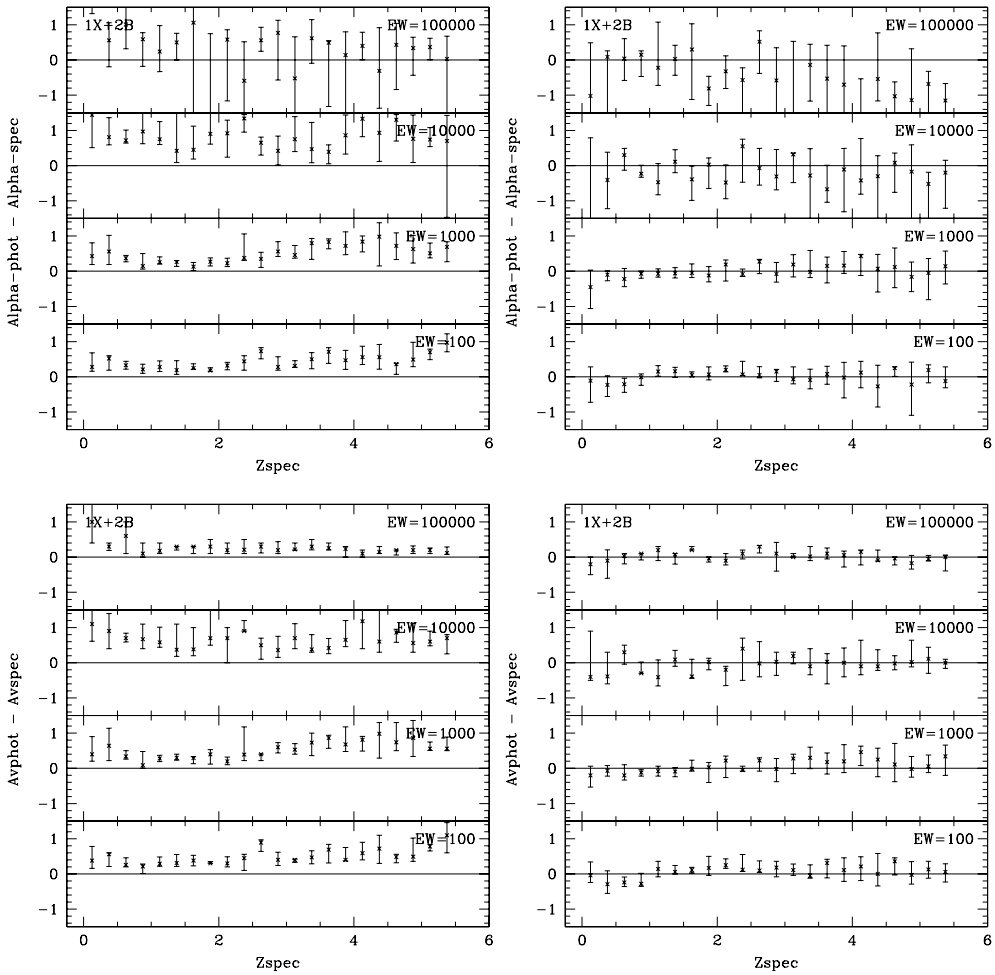

Figure 15. Median of the absolute errors $|\Delta\alpha|_{\rm med}$ (top) and $|\Delta A_v|_{\rm med}$ (bottom) obtained with the χ^2 method, as a function of redshift for $A_v = 0$ (left-hand panel) or $A_v = 1$ (right-hand panel); 1X+2B; $G = 18$.

where *Gaia* will find the majority of QSOs, given the limiting G mag.

5.3 Neural network analysis

After some tests, we found that the better photometric redshift determination was obtained with the help of *three* specialized ANNs: the first one *classifies* a QSO as a *low*-redshift one ($z < 2.2$) or a *high*-redshift one ($z \geqslant 2.2$); the task of the second (respectively third) ANN is to find the redshift within the low (respectively high) redshift domain. In order to minimize the effect of a possible wrong classification for QSOs with $z \sim 2.2$, the two redshift intervals overlap: $0 < z_{\rm low} < 2.4$ and $1.9 < z_{\rm high} < 5.5$. The limit $z \sim 2.2$ corresponds to the entrance of the Lyα break in the bluest MBP filter.

On the other hand, allocating a specialized network for $z < 2$ allows us to deal at best with the intrinsic degeneracies present in that range (see Section 5.1); such a network with *three* hidden layers of 10 neurones instead of two hidden layers has been found to give better results.

For each magnitude $G = 18$, 19, 20 and for the 1X + 2B and the 2F + 2B photometric systems, three ANNs have thus been built (with the same amount of noise as expected in the data). The LS is

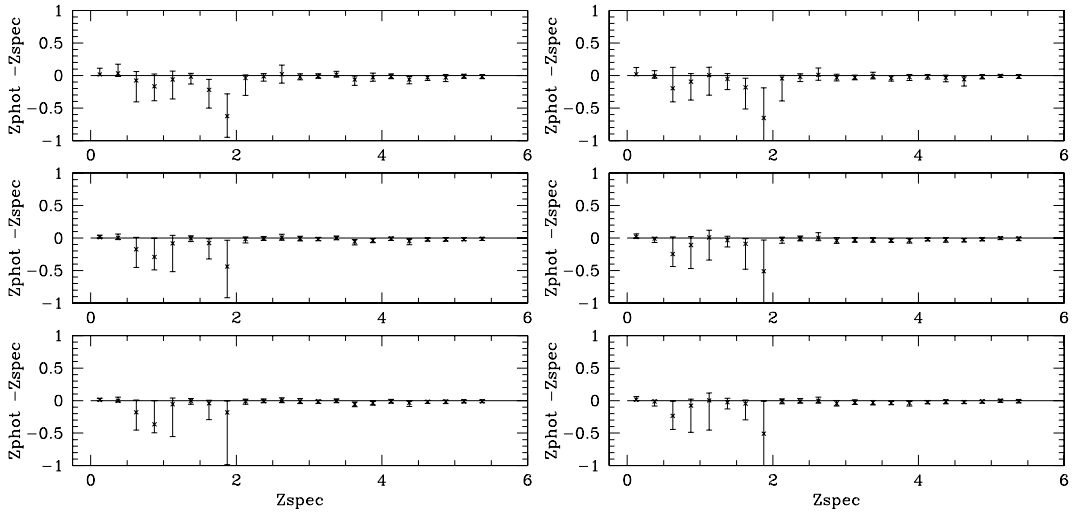

Figure 16. Difference between $z_{\rm phot}$ and $z_{\rm spec}$ determined with the χ^2 method, as a function of $z_{\rm spec}$ for 20 000 objects in the QSO-PCA-independent TS. The redshift bin is 0.25, the crosses represent the median value and the upper and lower error bars are derived from the lower and upper quartiles. 1X +2B and $G = 18$ (bottom), 19 (middle) and 20 (top); 1X+2B (left-hand panel), 2F+2B (right-hand panel).

the QSO-RAN data base of 20 000 QSOs (see Table 3), but instead of a uniform distribution on A_v, 20 per cent of the QSOs have $A_v = 0$ to oversample this important parameter edge. Subsets of 8833 QSOs with $z < 2.4$ and 12 646 QSOs with $z > 1.9$ have been extracted to train the low- and high-redshift regression networks.

Again, the TS consists of the QSO-REG data base of 86 625 QSOs, for $G = 18$, 19 and 20 (see Table 3).

Fig. 17 shows the value of $\Delta z = z_{\rm phot} - z_{\rm spec}$ as a function of $z_{\rm spec}$ for selected values of W of a typical QSO with $G = 19$, $\alpha = 0$ and $A_v = 0$. The error bars represent the Poissonian noise and are computed from the dispersion of 25 simulations. As with the k-NN analysis, the accuracy of the photometric redshift is the highest for large W ($\gtrsim$10 000) and also for large redshifts ($z_{\rm spec} > 2.5$) where the Lyα break is dominant. Again, the 'degeneracy region' below $z_{\rm spec} = 2$ is visible. The error bars at $G = 19$ are generally smaller than the offset between $z_{\rm phot}$ and $z_{\rm spec}$, so that systematic errors due to the imperfect neural modelling are dominant. They increase when only one neural network is used to determine $z_{\rm phot}$. Similar trends are also shown in the median of the absolute error $|\Delta z|_{\rm med}$ for typical QSOs with $-1 < \alpha < 1$ (including ∼90 per cent of the SDDS QSOs) for selected A_v, G mag and W range (see Table 12). As expected, the error on the redshift increases at low redshift and for weak emission lines, where it is also more sensitive to the G mag. However, equivalently good estimates of the redshift will be performed for moderately reddened QSOs.

Those general trends are similar to what was found with the k-NN method, but the absolute errors on the photometric redshift are systematically larger with ANNs (compare Tables 8 and 12).

Finally, it appears that with ANNs, the 2F photometric operates slightly but systematically better to determine the photometric redshift.

The APs of the QSOs for which the absolute error on the photometric redshift is *larger than* 0.2 can be read off in Fig. 18 ($G = 19$). As with the k-NN method, red solid triangles indicate that QSOs with a very steep blue continuum $\alpha > 2$ systematically suffer from a bad redshift estimation. However, they also show here that some peculiar redshifts (e.g. $z_{\rm spec} = 0.71$, 1.73, 3.36) cannot be retrieved accurately with ANNs for the 1X + 2B system, even for QSOs with strong emission lines. Again, those results suggest that the 2F+2B system is slightly better than the 1X+2B system in terms of redshift estimation from the point of view of the ANNs. This might be due to the fact that 1X MBP filter have sharp edges and this makes the interpolation more difficult.

In order to avoid any systematic effect and to really validate the ANN technique, the three 'specialized' ANNs as presented above have been applied to estimate the photometric redshifts in the totally independent TS of 20 000 QSOs present in the QSO-PCA data set (see Table 3 and Section 2.3), for $G = 18$, 19 and 20 in both the 1X+2B and 2F+2B photometric systems.

The error on the photometric redshift as a function of the spectroscopic redshift is given in Fig. 19. About 11 per cent of the QSOs have $|\Delta z| > 0.5$, and again most of the dispersion occurs in the interval $0.5 < z_{\rm spec} < 2$. In this interval, the median of the absolute error on the redshift is $|\Delta z|_{\rm med} \simeq 0.3$, nearly independent of the G mag, while $|\Delta z|_{\rm med} \simeq 0.06$–$0.16$ for $z_{\rm spec} > 2.5$ and $G = 18$–20. Exactly the same trends are observed as with the χ^2 method, except that the overall dispersion is 50–100 per cent larger.

Finally, we have also tried with the same data bases to estimate the α parameter with ANNs. Again, we checked that an architecture more complex than two hidden layers of 10 neurones does not improve the results. We also checked that taking into account the *observed* redshift does not significantly improve the determination of α (a degradation is even observed for $z_{\rm spec} < 2.5$ because of the errors on $z_{\rm obs}$). The results are summarized in Table 13. Again, the determination of α is worst when $A_v = 0$. For small W, results are significantly less accurate than with the χ^2 method.

A good determination of A_V and especially of W could not be achieved with ANNs. The limited performances of the ANNs with respect to the χ^2 approach in terms of photometric redshift and APs determination are probably due to the difficulty to interpolate through highly variable inputs, while in the case of classification, this variability could be used to distinguish them from stars.

5.4 Spectral principal component analysis

In this section, we first present results obtained using the SPC method to recover the shape of QSO spectra from the *Gaia 1X+2B* photometric data alone (cf. Section 3.3), assuming the redshifts are

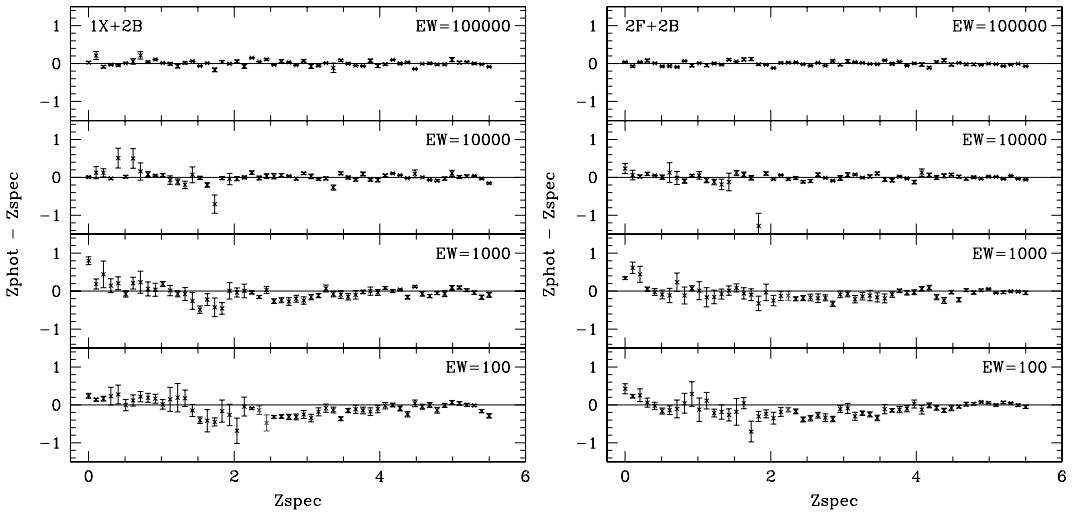

Figure 17. Difference between the photometric redshift $z_{\rm phot}$ obtained with ANNs and the spectroscopic redshift $z_{\rm spec}$ as a function of $z_{\rm spec}$ for $G = 19$, $\alpha = 0$ and $A_v = 0$ and for different W. Left-hand panel: 1X+2B photometric system; right-hand panel: 2F+2B photometric system.

Table 12. Median of the absolute error $|\Delta z|_{\rm med}$ on the photometric redshifts determined with ANNs for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.13/0.10	0.06/0.04	0.09/0.07	0.03/0.03	0.06/0.04	0.04/0.02
	$A_v = 1$	0.14/0.13	0.05/0.04	0.08/0.06	0.05/0.03	0.05/0.03	0.03/0.02
$G = 19$	$A_v = 0$	0.17/0.16	0.08/0.06	0.10/0.08	0.06/0.04	0.05/0.05	0.05/0.03
	$A_v = 1$	0.21/0.18	0.06/0.05	0.10/0.09	0.05/0.04	0.04/0.04	0.04/0.03
$G = 20$	$A_v = 0$	0.33/0.29	0.14/0.11	0.17/0.16	0.09/0.08	0.06/0.06	0.06/0.04
	$A_v = 1$	0.36/0.32	0.11/0.08	0.22/0.18	0.09/0.07	0.05/0.05	0.05/0.04

Table 13. Median of the absolute error $|\Delta\alpha|_{\rm med}$ on the QSO spectrum slope determined with ANN for $-1 \leqslant \alpha \leqslant 1$ (1X+2B/2F+2B).

		$W \leqslant 3000$		$3000 < W \leqslant 11\,000$		$W > 11\,000$	
		$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$	$z_{\rm spec} < 2.5$	$z_{\rm spec} > 2.5$
$G = 18$	$A_v = 0$	0.66/0.57	0.73/0.74	0.68/0.52	0.70/0.72	0.71/0.79	0.92/0.88
	$A_v = 1$	0.34/0.34	0.55/0.48	0.47/0.39	0.56/0.51	0.80/0.67	1.07/0.94
$G = 19$	$A_v = 0$	1.13/0.90	1.00/0.71	1.27/0.93	0.81/0.64	0.91/0.84	0.81/0.90
	$A_v = 1$	0.50/0.56	0.49/0.59	0.50/0.50	0.50/0.56	0.67/0.77	1.00/1.05
$G = 20$	$A_v = 0$	0.82/0.96	0.66/0.63	0.78/0.94	0.68/0.64	0.73/0.72	0.94/0.89
	$A_v = 1$	0.57/0.50	0.60/0.56	0.57/0.43	0.61/0.57	0.81/0.87	1.07/1.13

known. Then, we show the capacity of the method to also recover the QSO redshifts, based on the same photometric data sets.

This exploratory study has been performed with the set of 20 000 QSO spectra of the SPC spectral library (see Section 2.3.3). As shown below, promising results are found, but they should be confirmed on an independent set of QSO spectra also including correlation between emission lines (ideally a library of *observed* spectra).

5.4.1 Spectral shape with known redshift

Fig. 20 graphically summarizes the goodness of recovery for the 20 000 QSO spectra of the SPC spectral library, at $G = 18$, 19 and 20, assuming that their redshifts are known. The y-axis represents the average of the absolute value of the relative difference between the flux measured in a 20 Å wide bin of the reconstructed QSO spectrum (as observed by *Gaia*) and the input QSO spectrum. The average was made over the wavelength range covered by the *Gaia* photometric systems. The bin width was chosen sufficiently large to smooth out the inherent noise already present in the individual SPCs, while being not wider than any of the *Gaia* narrow-band filter.

Fig. 20 as well as Table 14 clearly show that the shape of the QSO spectrum can be accurately recovered for a very large fraction of QSOs, even at $G = 20$, provided that the redshift is not too large ($z < 3$). Indeed, at $G = 18$ for $z < 2.5$ QSOs, 90 per cent of the absolute relative errors in the flux over 20 Å bins is smaller than 10 per cent.

A nice example of a $z = 1.797$ QSO spectrum whose shape is very successfully recovered at $G = 18$ is shown in Fig. 21. The degradation of the reconstruction is also shown at fainter

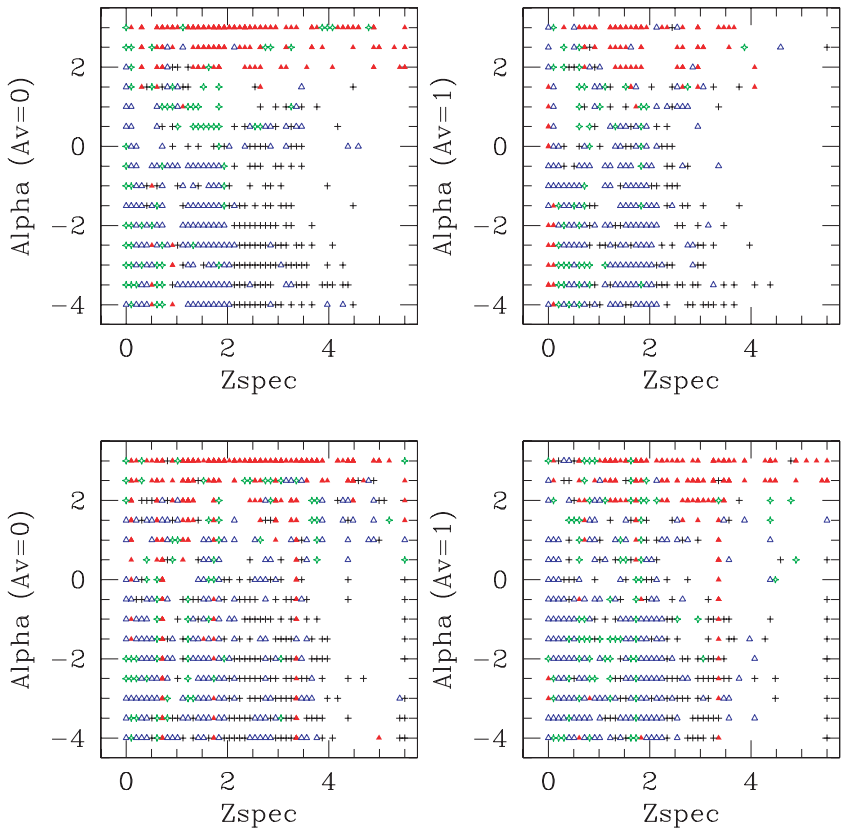

Figure 18. Regions of the α–$z_{\rm spec}$ plane, where $|\Delta z| > 0.2$ with ANNs, for QSOs with $W = 100$ (black), 1000 (blue), 10 000 (green) and 100 000 (red). Top: 2F + 2B system; bottom: 1X + 2B system; left-hand panel: $A_v = 0$; right-hand panel: $A_v = 1$. ($G = 19$).

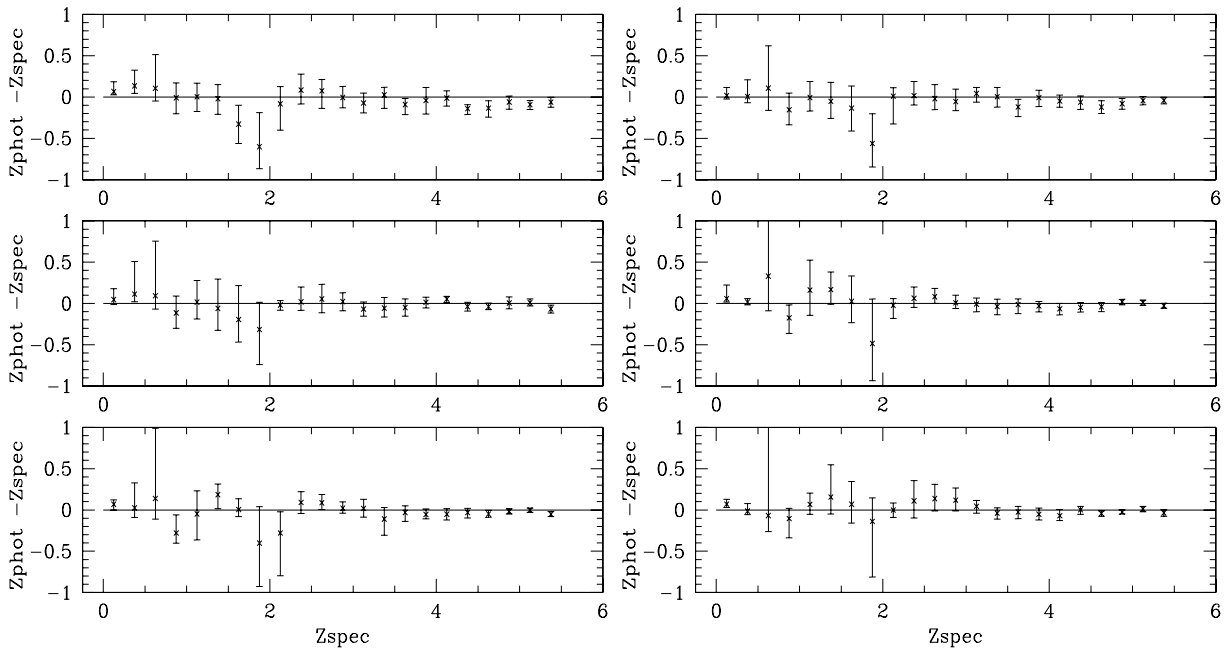

Figure 19. Difference between $z_{\rm phot}$ and $z_{\rm spec}$ determined with the ANN method, as a function of $z_{\rm spec}$ for 20 000 SPC-QSOs. The redshift bin is 0.25, the cross represents the median value and the upper and lower error bars are derived from the lower and upper quartiles, respectively. $G = 18$ (bottom), 19 (middle) and 20 (top); 1X+2B (left-hand panel), 2F+2B (right-hand panel).

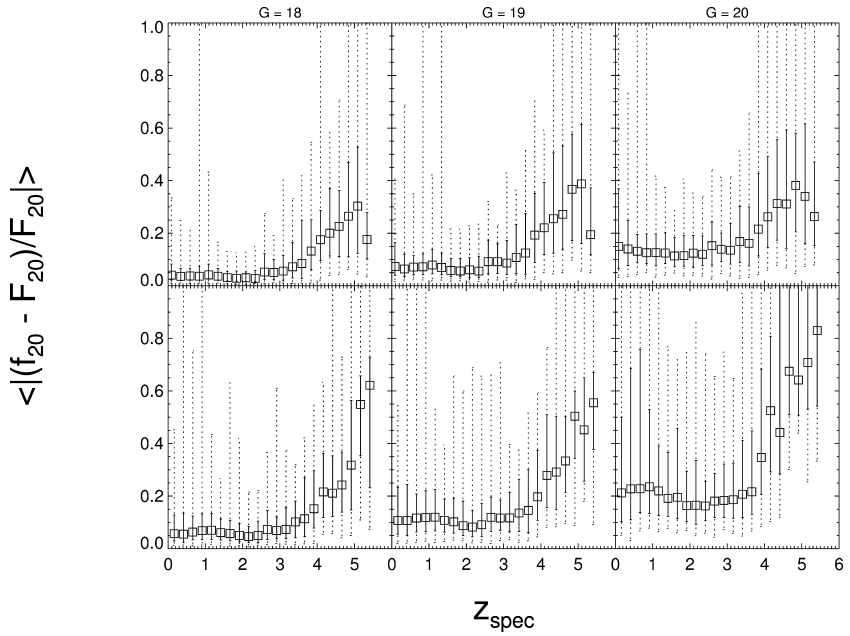

Figure 20. The SPC method using 1X+2B photometric data. Distribution of the average of the absolute value of the relative difference between the reconstructed and input QSO fluxes integrated in 20 Å wide bins, as a function of the (true) QSO redshift. Squares represent the median values for all the QSOs in the corresponding bin in redshift. Solid error bars represent the 10 and 90 percentile of the distribution. Dashed error bars correspond to the minima and maxima of the distribution. The top row graphs assume that the QSO redshifts are known. On the bottom graphs, the results correspond to QSO *photometric* redshifts, for which $\chi^2(z)$ is minimum in the SPC method.

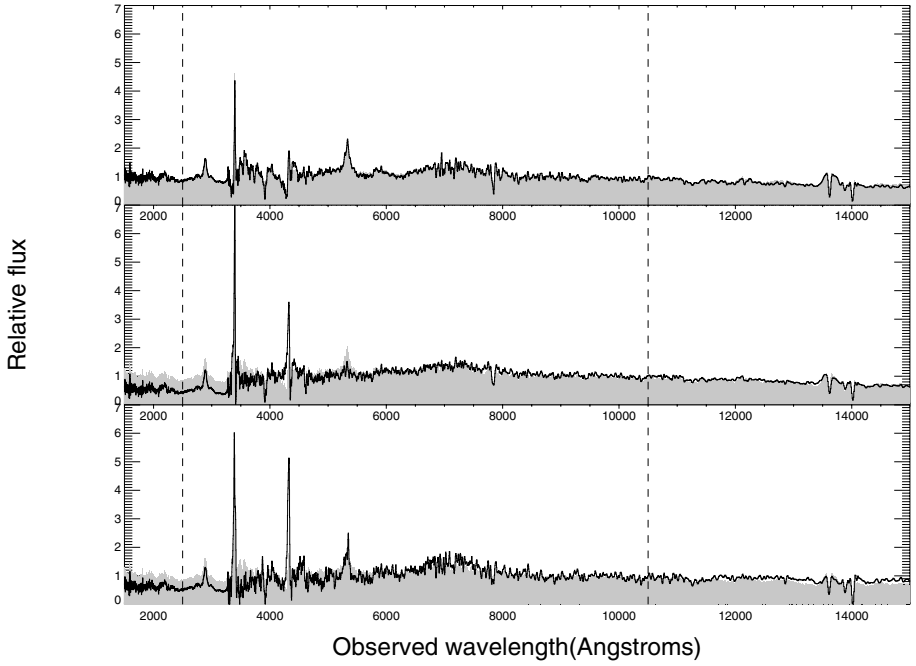

Figure 21. An example of a successful spectral recovery with the 1X+2B photometric data, using the SPC method. In each of the panels, the grey area corresponds to the input spectrum (identical for each panel), in this case a $z = 1.797$ QSO. The vertical dashed lines indicate the wavelength range covered by the *Gaia* photometric system. The top, middle and bottom panels correspond to a $G = 18$, 19 and 20 QSO, respectively.

magnitudes. Small features (e.g. the C IV line) as well as the UV continuum are lost when the S/N is lower, but the global shape is still rather well recovered. Note also that even features located outside the wavelength range covered by the *Gaia* photometric systems can actually be recovered; thanks to the properties of the principal components.

The quality of the recovered spectral shapes is of course dependent on the accuracy with which the redshift is known. If, instead of the exact value of the redshift, we use the one derived from the *Gaia* photometry in the next section, the difference between the true and recovered spectral shapes increases, as illustrated in the bottom graphs of Fig. 20 and in Table 14.

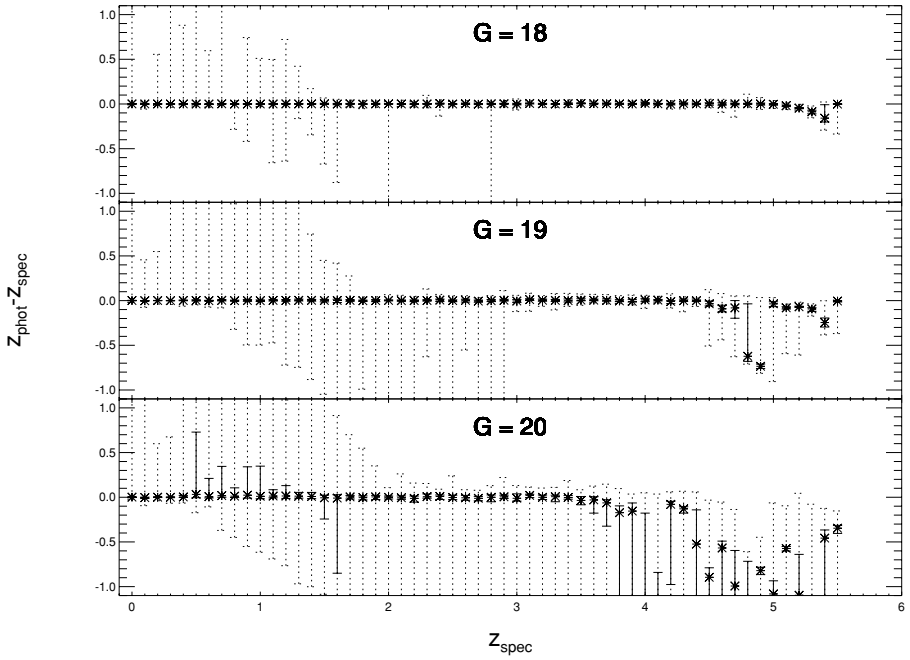

Figure 22. Distribution of errors, ($z_{\rm phot} - z_{\rm spec}$), in the recovery of redshifts using the SPC method and 1X+2B photometric data. Stars indicate the median value for the given redshift bins, solid error bars give the 25 and 75 percentiles of the distribution, while dashed error bars indicate the minimum and maximum of the distribution.

Table 14. The SPC method. Ranges of the 10 and 90 percentiles of the absolute value of the relative difference between the recovered and the input spectra integrated in 20 Å bins: $|(f_{20} - F_{20})/F_{20}|$. Results are given for two redshift intervals, for $G = 18$, 19, 20 and depending on whether the QSO redshift is precisely known ($z_{\rm spec}$) or derived with the SPC method ($z_{\rm phot}$).

		$z_{\rm spec} < 2.5$		$z_{\rm spec} > 2.5$	
		10 percentile	90 percentile	10 percentile	90 percentile
$G = 18$	$z_{\rm spec}$	0.02–0.03	0.05–0.10	0.02–0.12	0.10–0.47
$G = 18$	$z_{\rm phot}$	0.02–0.04	0.08–0.14	0.02–0.35	0.10–0.73
$G = 19$	$z_{\rm spec}$	0.03–0.04	0.10–0.16	0.04–0.17	0.16–0.61
$G = 19$	$z_{\rm phot}$	0.04–0.07	0.14–0.24	0.05–0.38	0.17–0.67
$G = 20$	$z_{\rm spec}$	0.07–0.09	0.18–0.37	0.08–0.21	0.20–0.62
$G = 20$	$z_{\rm phot}$	0.09–0.14	0.26–0.76	0.11–0.54	0.26–3.58

5.4.2 Redshift determination with the SPC method

Prompted by these encouraging results, we have explored the possibility to recover the redshift using the SPC with *Gaia* data alone (see Section 3.3). Fig. 22 shows that even at $G = 19$, a very large majority of redshifts are recovered with very high accuracy. At $G = 20$, a larger dispersion is seen in the degeneracy region, i.e. between $z_{\rm spec} \simeq 0.5$ and 1.6 (see Section 5.1).

The accuracy on the photometric redshift is lost essentially for high-redshifts objects, especially when the S/N is low (at $G = 20$). This is because the number of filters, where the flux is significantly different from zero, decreases when the redshift increases, and so the number of available equations in System 8 becomes too small to provide useful constraints on the w_i. For the same reason, recovering of the spectral shape is expected to fail at large redshift (see also Fig. 20).

These results are also summarized in Table 15.

Table 15. Ranges of the 25 and 75 percentiles of the distribution of $z_{\rm phot} - z_{\rm spec}$, using the SPC method.

	$z_{\rm spec} < 2.5$		$z_{\rm spec} > 2.5$	
	25 percentile	75 percentile	25 percentile	75 percentile
$G = 18$	−0.008 – −0.001	0.002–0.010	−0.202–0.002	−0.070–0.015
$G = 19$	−0.210 – −0.001	0.001–0.018	−0.759–0.001	−0.710–0.026
$G = 20$	−0.850 – −0.000	0.001–0.729	−4.803 – −0.003	−1.119–0.038

6 CONCLUSIONS

This study was motivated from the importance of creating the largest and purest sample of QSOs from the data to be collected with *Gaia* (see Section 1). In this first paper, we quantified the likelihood to identify QSOs and determine their redshift only on the basis of their combined BBP and MBP. To these ends, we first built QSO synthetic spectral libraries, which have been made publicly available.[11] Then, we compared the ability of the χ^2 template fitting and of the ANN method, using the 1X+2B or 2F+2B filter systems. We also used these methods as well as a novel method based on the SPCs to estimate their redshift and other APs. The results we obtained do not strongly depend on the adopted photometric system.

First, we found that building a *secure* QSO catalogue based on photometry alone is possible, although the incompleteness can be severe, especially in the galactic plane where the reddening is expected to be higher. To that aim, ANNs are more efficient than the χ^2 approach, but the latter has a better completeness ($\simeq$60 per cent at $G = 20$) and could be used at high galactic latitudes, where the $N_*/N_{\rm QSO}$ population ratio is more favourable to

[11] See the *Gaia* Spectral Library web page: http://gaia.esa.int/spectralib/

QSOs. With both methods, white dwarfs are well discriminated. The completeness could certainly be improved by rejecting very unlikely kinds of stars. It is, however, clear from this study that photometry alone is not sufficient to reach a high degree of completeness, in particular for QSOs with $2 \lesssim z \lesssim 3$ located close to the galactic plane, and it should be complemented with the variability and astrometric constraints to be provided by *Gaia*. However, adopting $\sigma_\pi = 160$ μas at $G = 20$, a 3σ measurement of the parallax is possible only within a distance of about 2 kpc. Since all stars hotter than M stars can be detected beyond that limit, their lack of apparent parallax will not distinguish them from QSOs. The constraint on the proper motion looks more promising since most of the stars following the galactic rotation have a proper motion which will be detectable by *Gaia*. A model of the Milky Way will thus be necessary to quantify the probability for a given star to have a given apparent proper motion, which depends on the direction of the line of sight. This aspect as well as variability will be the aim of the second-step study.

In the present study, BAL QSOs are not taken into account in the synthetic libraries. However, a preliminary test simply based on the SDSS low-ionization BAL composite spectrum (Reichard et al. 2003a) shows that BAL QSOs would not totally be lost, especially the bright ones. Of course, to improve the sensitivity to this non-negligible – and interesting – fraction of QSOs, *observed* BAL spectra should be included in the final version of the QSO spectral library.

Secondly, this study has shown that the photometric redshift of quasars can reasonably be well retrieved from the BBP plus MBP photometry. The median absolute error on $z_{\rm phot}$ is the largest for $0.5 \leqslant z_{\rm spec} \leqslant 2$ ($|\Delta z|_{\rm Median} \simeq 0.2$ using the QSO-PCA-independent TS). Unfortunately, this is the redshift range where most of the QSOs are expected. The reason is to be found in the interplay between the limited wavelength coverage and the existence of casual multiplication factors between the wavelengths of several QSO strong emission lines. This degeneracy should be alleviated by adding photometric data further in the UV. Such data are being obtained over the whole sky down to $m_{\rm AB} = 20.5$ by the *GALEX* satellite (Martin et al. 2003). They should thus be included into the analysis of the *Gaia* photometry. We also noted that the other QSO APs (α, W, A_V) are much less constrained by the *Gaia* photometry.

Finally, promising results show that, except for very high redshift QSOs ($z \gtrsim 3$), it should be possible to retrieve the photometric redshift together with the weights of the SPCs matching at best the *Gaia* photometry. This technique seems to be less sensitive to the colour degeneracy, perhaps because the emission-line strengths are allowed to vary independently. Another advantage of the SPC method is that for the highest S/N objects ($G \lesssim 18$), it should also be possible to reconstruct the spectrum shape and to get some hints about spectral features located even outside the spectral range directly probed by *Gaia*. Nevertheless, this method should be tested in the future with an independent QSO library.

ACKNOWLEDGMENTS

It is a pleasure to thank Prof. L. Wehenkel, for granting free access to the GTDIDT software. We also thank P. Francis and Z. Shang for kindly providing us with their QSO SPCs and Mingren Shi for the L1 algorithm.

Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the US Department of Energy, the Japanese Monbukagakusho and the Max Planck Society. The SDSS web site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory and the University of Washington. Our research was supported in part by PRODEX HST/GL and contract IUAP P5/36 'Pôle d'Attraction Interuniversitaire'.

REFERENCES

Andreon S., Gargiulo G., Longo G., Tagliaferri R., Capuano N., 2000, MNRAS, 319, 700
Bahcall J. N., Soneira R. M., 1980, ApJS, 44, 73
Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, MNRAS, 298, 361
Baldwin J. A., 1977, ApJ, 214, 679
Bertin E., 1994, Ap&SS, 217, 49
Bishop C. M., 1995, Neural Networks for Pattern Recognition, Oxford Univ. Press, Oxford
Budavári T. et al., 2001, AJ, 122, 1163
Cabanac R. A., de Lapparent V., Hickson P., 2002, A&A, 389, 1090
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Claeskens J.-F., Surdej J., 2002, A&AR, 10, 263
Cristiani S., Vio R., 1990, A&A, 227, 385
Croom S. M., Smith R. J., Boyle B. J., Shanks T., Miller L., Outram P. J., Loaring N. S., 2004, MNRAS, 349, 1397
Croom S. M. et al., 2005, MNRAS, 356, 415
Eyer L., 2002, Acta Astron., 52, 241
Fan X., 1999, AJ, 117, 2528
Ferland G. J., 1996, Hazy: A Brief Introduction to CLOUDY 90, Univ. of Kentucky, Physics Department Internal Report
Fitzpatrick E. L., 1999, PASP, 111, 63
Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., Weymann R. J., Morris S. L., 1991, ApJ, 373, 465
Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., 1992, ApJ, 398, 476
Hartwick F. D. A., Schade D., 1990, ARA&A, 28, 437
Hatziminaoglou E., Mathez G., Pelló R., 2000, A&A, 359, 9
Hawkins M. R. S., 2000, A&AS, 143, 465
Irwin M., Mc Mahon R. G., Hazard C., 1991, in Crampton D., ed., ASP Conf. Ser. Vol. 21, The Space Distribution of Quasars. Astron. Soc. Pac., San Francisco, p. 117
Jordi C., Grenon M., Figueras F., Torra J., Carrasco J. M., 2003, Internal ESA document UB-PWG-011
Jordi C., Høg E., Bailer-Jones C., 2004, Internal ESA document GAIA-CUO-161
Jordi C. et al., 2006, MNRAS, in press
Koo D. C., 1999, in Weymann R., Storrie-Lombardi L., Sawicki M., Brunner R., eds, ASP Conf. Ser. Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies. Astron. Soc. Pac., San Francisco, p. 3
Lahav O., Naim A., Sodré L., Storrie-Lombardi M. C., 1996, MNRAS, 283, 207
Lejeune Th., Cuisinier F., Buser R., 1998, A&AS, 130, 65
Lindegren L., 2003, Internal ESA document GAIA-LL-045
Madau P., 1995, ApJ, 441, 18
Martin C. (GALEX Science Team), 2003, Am. Astron. Soc. Meeting, 203, 96.01
Mignard F., 2002, in Bienaymé O., Turon C., eds, Proc. GAIA: A European Space Project, Vol. 2, held on 2001 May 14–18, Les Houches, France. EDP Sciences, EAS Publications Series, p. 327
Møller P., Jakobsen P., 1990, A&A, 228, 299
Naim A., Lahav O., Sodre L., Storrie-Lombardi M. C., 1995, MNRAS, 275, 567
Perryman M. A. C. et al., 2001, A&A, 369, 339

Reichard T. A. et al., 2003a, AJ, 125, 1711
Reichard T. A. et al., 2003b, AJ, 126, 2594
Rengstorf A. W. et al., 2004, ApJ, in press (astro-ph/0310916)
Richards G. T. et al., 2001, AJ, 122, 1151
Robin A. C., Reylé C., Derrière S., Picaud S., 2003, A&A, 409, 523
Royer P., Manfroid J., Gosset E., Vreux J.-M., 2000, A&AS 145, 351
Sánchez S. F. et al., 2004, ApJ, in press (astro-ph/0403645)
Schneider D. P. et al., 2003, AJ, 126, 2573
Shang Z., Wills B. J., Robinson E. L., Wills D., Laor A., Xie B., Yuan J., 2003, ApJ, 586, 52
Shi M., Lukas M. A., 2002, Comput. Stat. Data Anal., 39, 33
Söderhjelm, 2002, Internal ESA Document DMS-SS-01
Storrie-Lombardi M. C., Lahav O., Sodre L., Storrie-Lombardi L. J., 1992, MNRAS, 259, 8
Stoughton Ch. et al., 2002, AJ, 123, 485
Treyer M., Wambsganss J., 2004, A&A, 416, 19
Vansevicius V., Bridzius A., 2002, Internal ESA document GAIA-VIL-008
Vanzella E. et al., 2004, A&A, 423, 761
Veron P., Hawkins M. R. S., 1995, A&A, 296, 665
Véron-Cetty M.-P., Véron P., 2003, A&A, 412, 399
Warren S. J., Hewett P. C., Osmer P. S., 1994, ApJ, 421, 412
Weaver W. B., 2000, ApJ 541, 298
Wehenkel L., 1997, Automatic Learning Techniques in Power Systems. Kluwer, Dordrecht
Weymann R. J., Morris S. L., Foltz C. B., Hewett P. C., 1991, ApJ, 373, 23
Wolf C., Meisenheimer K., Kleinheinrich M., Borch A., Dye S., Gray M., Wisotzki L., Bell E. F. 2004, A&A, submitted
Yahata K. et al., 2005, PASJ, 57, 529
York D. G. et al., 2000, AJ, 120, 1579
Yip C. W. et al., 2004, AJ, 128, 2603
Zheng W., Kriss G. A., Telfer R. C., Grimes J. P., Davidsen A. F., 1997, ApJ, 475, 469

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