Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **448,** 1305–1313 (2015) doi:10.1093/mnras/stv063

Star–galaxy separation strategies for *WISE*-2MASS all-sky infrared galaxy catalogues

András Kovács[1,2]★ and István Szapudi[3]

[1]*Institut de Física d'Altes Energies, Universitat Autónoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain*
[2]*MTA-ELTE EIRSA 'Lendület' Astrophysics Research Group, 1117 Pázmány Péter sétány 1/A Budapest, Hungary*
[3]*Institute for Astronomy, University of Hawaii 2680 Woodlawn Drive, Honolulu, HI 96822, USA*

Accepted 2015 January 12. Received 2015 January 12; in original form 2013 December 31

ABSTRACT

We combine photometric information of the *Wide-Field Infrared Survey Explorer* (*WISE*) and Two Micron All Sky Survey (2MASS) all-sky infrared data bases, and demonstrate how to produce clean and complete galaxy catalogues for future analyses. Adding 2MASS colours to *WISE* photometry improves star–galaxy separation efficiency substantially at the expense of losing a small fraction of the galaxies. We find that 93 per cent of the *WISE* objects within $W1 < 15.2$ mag have a 2MASS match, and that a class of supervised machine learning algorithms, support vector machines (SVM), are efficient classifiers of objects in our multicolour data set. We constructed a training set from the Sloan Digital Sky Survey PhotoObj table with known star–galaxy separation, and determined redshift distribution of our sample from the Galaxy and Mass Assembly spectroscopic survey. Varying the combination of photometric parameters input into our algorithm we show that $W1_{WISE} - J_{2MASS}$ is a simple and effective star–galaxy separator, capable of producing results comparable to the multidimensional SVM classification. We present a detailed description of our star–galaxy separation methods, and characterize the robustness of our tools in terms of contamination, completeness, and accuracy. We explore systematics of the full sky *WISE*-2MASS galaxy map, such as contamination from moon glow. We show that the homogeneity of the full sky galaxy map is improved by an additional $J_{2MASS} < 16.5$ mag flux limit. The all-sky galaxy catalogue we present in this paper covers 21 200 deg^2 with dusty regions masked out, and has an estimated stellar contamination of 1.2 per cent and completeness of 70.1 per cent among 2.4 million galaxies with $z_{med} \approx 0.14$. *WISE*-2MASS galaxy maps with well controlled stellar contamination will be useful for spatial statistical analyses, including cross-correlations with other cosmological random fields, such as the cosmic microwave background. The same techniques also yield a statistically controlled sample of stars as well.

Key words: catalogues – large-scale structure of Universe.

1 INTRODUCTION

In recent years, sky surveys have been producing astronomical data with a rapidly accelerating pace resulting in what is commonly called the 'data avalanche'. The large quantity of data necessitates automated algorithms for filtering, photometric selection, and estimation observables such as redshifts. Each object in a catalogue has multiple properties, thus algorithms have to explore high-dimensional configuration spaces in large, often connected, data bases. Such high-dimensional spaces can be effectively explored with machine learning techniques, such as the support vector machines (SVM) used in our work.

When analysing an object catalogue, the most fundamental, and often most challenging task is star–galaxy [possibly Quasi Stellar Object (QSO)] separation. A simple separator between stars and galaxies is a morphological measurement, where extended sources are classified as galaxies (Vasconcellos et al. 2011). Morphology, however, looses its power at fainter magnitudes, a problem for wide-field surveys, e.g. Pan-STARRS (Kaiser, Burgett & Chambers 2010), Euclid (Amendola et al. 2013), BigBOSS (Schlegel et al. 2011), DES (The Dark Energy Survey Collaboration 2005), and LSST (LSST Science Collaboration et al. 2009). At the fainter end, the most widely used tools for object classification are colour–colour diagrams: different types of objects will appear in different regions

★ E-mail: andraspankasz@gmail.com

according to the shape of their spectral energy distribution. Classification methods based on infrared colour–colour selection were employed for star–galaxy separation (Pollo, Rybka & Takeuchi 2010) or for finding special classes of sources, such as high-/low-redshift QSO, AGN, starburst galaxies, or variable stars (e.g. Richards et al. 2002; Chiu et al. 2005; Brightman & Nandra 2012; Stern et al. 2012, and references therein).

Applications of SVMs are widely used for data mining and analysis. SVMs are relatively easy to implement, simple to run, and they are very many-sided, thus several astronomical problems set out to use such methods to classify objects, or to perform regression analysis. Among others, Woźniak et al. (2004) analysed variable sources with an SVM in a five-dimensional parameter space including period, amplitude and three colours, Huertas-Company et al. (2009) analysed morphological properties of infrared galaxies using SVM with 12 parameters, while Solarz et al. (2012) created a star–galaxy separation algorithm based on mid- and near-infrared colours. Recently, Małek et al. (2013) used VIPERS and VVDS surveys to perform object classification into three groups: stars, galaxies, and AGNs. A similar study by Saglia et al. (2012) also used an SVM as three-type-classifier. Their method was developed for the Photometric Classification Server for the prototype of the Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1; Kaiser et al. 2010).

While star–galaxy separation can be performed from *Wide-Field Infrared Survey Explorer* (*WISE*) colours alone (Goto, Szapudi & Granett 2012; Kovács et al. 2013) it is at the expense of severe cuts that still are sensitive to contamination to from the moon and necessitate complex masks. As we show later, adding Two Micron All Sky Survey (2MASS) observations removes artificial features from the *WISE* data. With several open source implementations and computationally modest cost (Fadely, Hogg & Willman 2012), we set out to use the SVM algorithm for separating stars and galaxies in the matched *WISE*-2MASS photometric data. Our principal goal is to create a clean catalogue of galaxies observed by *WISE* and 2MASS suitable for large-scale structure and cross-correlation studies. At the same time, we will show that our selection algorithms are suitable for producing clean stellar samples as well. The galaxy maps we create are useful for cross-correlation studies, such as integrated Sachs–Wolfe measurements, and galaxy–cosmic microwave background lensing correlations, while the large data sets of stars may constrain stellar streams and Galactic structure in general.

The paper is organized as follows. Data sets and algorithms are described in Section 2, while our results are presented in Section 3, with detailed discussion, comparisons, and interpretation.

2 DATASETS AND METHODOLOGY

We combine measurements of two all-sky surveys in the infrared, *WISE* (Wright et al. 2010) and the Point Source Catalog of the 2-Micron All-Sky Survey (2MASS-PSC; Skrutskie et al. 2006). We use photometric measurements of the *WISE* satellite, which surveyed the sky at four different wavelengths: 3.4, 4.6, 12, and 22 μm (*W*1−*W*4 bands). Following Goto et al. (2012) and Kovács et al. (2013), we select sources to a flux limit of $W1 \leq 15.2$ mag to have a fairly uniform data set.

We then add 2MASS *J*, *H*, and K_s magnitudes conveniently available in the *WISE* catalogue, where 93 per cent of the *WISE* objects with $W1 \leq 15.2$ mag have 2MASS observations. We find, however, lower matching rates for a *WISE*-2MASS sample at fainter *W*1 cuts.

Note that the *WISE* *W*1 magnitude limit we define is lower than the 5σ detection limit for *W*1. However, this selection cut makes comparisons to previous *WISE* catalogues easier, and helps to avoid large-scale inhomogeneities (caused by moon-glow effects) which potentially contaminate deeper *WISE* galaxy catalogues (Kovács et al. 2013).

Table 1. Star contamination and galaxy completeness as a function of flux limits and analysis methods (CC = '$W1 - J < -1.7$ colour cut'). We apply a $W1 > 12.0$ lower flux cut in every cases. See text for details. The last column shows the expected galaxy number counts assuming a mask which leaves 21 200 deg^2 unmasked.

Method	*W*1	*J*	$\frac{F_S}{T_S+F_S}$(%)	$\frac{T_G}{T_G+F_S}$(%)	N_{gal}
SVM	<14.5	–	3.4	92.1	1.2×10^6
CC	<14.5	–	1.6	84.4	1.1×10^6
SVM	<15.0	–	2.6	93.4	2.3×10^6
CC	<15.0	–	1.4	82.7	2.1×10^6
SVM	<15.2	–	3.1	93.6	6×10^6
CC	<15.2	–	1.8	78.6	5×10^6
SVM	<15.2	<16.5	2.8	85.9	3×10^6
CC	<15.2	<16.5	1.2	70.1	2.4×10^6

We note that these choices allow us to produce a catalogue deeper than the 2MASS Extended Source Catalog (2MASS; Jarrett et al. 2000), as proper identification of fainter 2MASS objects becomes possible.

To apply machine learning techniques, one needs to identify a 'training set', a set of objects with known classification. We choose a smaller region of Stripe 82 in the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7; Abazajian et al. 2009), deeper than our catalogue and located at $327.5 < \mathrm{RA} < 338.5$ and $-1.25 < \mathrm{Dec.} < 1.25$. We performed the cross-matching with the KD-Tree (Bentley 1975) algorithm as implemented in the PYTHON package SCIPY. We found an SDSS match for 99.4 per cent for the 46 749 *WISE*-2MASS objects using a 1 arcsec matching radius. We have found multiple counterparts for 0.03 per cent among 46 463 objects, and chose the nearest neighbour as a real counterpart. We estimated the accidental matching rate by generating random positions of *WISE*-2MASS source density, finding 0.1 per cent. This suggests no meaningful effect on the training sample. As a further refinement, we applied a $W1 \geq 12.0$ magnitude cut to avoid potentially problematic SDSS classification, and to mark out the galaxy locus in colour–magnitude space. See Fig. 5 for ratification.

As an exploratory test, we downloaded 2MASS XSC data (Jarrett et al. 2000) from the same coverage, finding 1195 galaxies. A deeper *WISE*-2MASS catalogue without extra flux cut in *J* band contains 5922 objects classified as a galaxy in SDSS PhotoObj table. We will show that the fraction of the properly identified galaxies reaches ∼78.6 per cent with ∼1.8 per cent star contamination (see Table 1), even with our simplest algorithms. We thus are able to broaden 2MASS XSC significantly.

The redshift distribution of the matched *WISE*-2MASS-SDSS objects classified as galaxies is provided by matching with the Galaxy and Mass Assembly (GAMA; Driver et al. 2011) spectroscopic data set, at the full GAMA coverage of 144 deg^2. We performed a nearest neighbour search using 1 arcsec as a matching radius, and found a pair for 97 per cent of the *WISE*-2MASS-SDSS galaxies in GAMA Data Release 2. We have found multiple counterparts for 0.15 per cent among 8493 objects, and chose the nearest neighbour as a real counterpart. The un-matched 3 per cent might consist of predominantly massive early-type galaxies at $z > 1$ (Yan

et al. 2013). Another possibility is that the remaining 3 per cent is populated by objects with bad SDSS classification, that actually indicates the purity of our training sample.

Next we will use the resulting multicolour catalogue for object classification. Note that the training set we use is a realistic sample of the multicolour *WISE*-2MASS data base, since we applied the same flux cuts for all subsamples in the analysis.

2.1 Support vector machines

SVM designates a subclass of supervised learning algorithms for classification in a multidimensional parameter space. These methods include extensions to non-linear models of the generic (linear) algorithm developed by Cortes & Vapnik (1995). SVMs carry out object classification and/or regression by calculating decision hyperplanes between sets of points having different class memberships. A central concept of SVM learning is the training set, a special set of objects that supplies the machine with classified examples. Based on its properties, the classifier is tuned, and the hyperspace between different classes is determined. A training set of a few thousands of objects is usually suitable for simpler classification problems (see e.g. Solarz et al. 2012).

The algorithm includes a non-linear kernel function, which is used to find a hyperplane with maximum distance from the boundary to the closest points belonging to the separate classes of objects (Cortes & Vapnik 1995). The kernel is a symmetric function that maps data from the input space X to the feature space F. For our analysis, we chose a Gaussian radial basis kernel (RBK) function, defined as $k(x_i, x_j) = \mathrm{e}^{-\gamma||x_i - x_j||2}$, where $||x_i - x_j||$ is the Euclidean distance between x_i, and x_j. The product of the kernel function is a non-linear representation of each parameter from the input to the feature space. The RBK kernel is often used as SVM kernel function to make the non-linear feature map. We decided to use it because of its effectiveness and simplicity.

SVM offers a whole set of parametrization choices. We chose '*C*-classification' because of its good performance and only two free parameters. C is the cost function, i.e. a trade-off parameter that sets the width of the margin separating classes of objects. A small margin of separation can be set with larger C, but high C values often lead to overfitting. Reduced C values, however, smooth the hyperplane, which can be a source of misclassifications (Małek et al. 2013). The second parameter, γ, determines the topology of the decision surface. A low value of γ sets a rigid, and structured decision boundary, while high γ values indicate a very smooth decision surface with many misclassifications.

3 DISCUSSION

3.1 SVM outputs

We used a free software environment for SVM in PYTHON package SCIKIT-LEARN. First, we performed tests to tune both the C and γ parameters, and found the lowest classification errors with $C = 10.0$, and $\gamma = 0.1$. Then, we proceeded to determine the optimal number of parameters for the optimal classification efficiency experimentally. We used 8000 objects as a 'training set', and 2000 objects for control, i.e. testing the efficiency of our algorithms. The training sample contains 13 per cent galaxy data, in order to preserve the star–galaxy ratio that we have originally found in our SDSS cross-matching. We evoke the terminology of machine learning, and use 'True' (T) and 'False' (F) labels to distinguish between objects that are classified correctly, and the ones have false identification.

We also define five measures of SVM performance.

(i) Star contamination $= \frac{F_S}{T_S+F_S}$.
(ii) Galaxy contamination $= \frac{F_G}{T_G+F_G}$.
(iii) Star completeness $= \frac{T_S}{T_S+F_G}$.
(iv) Galaxy completeness $= \frac{T_G}{T_G+F_S}$.
(v) Accuracy $= \frac{T_G+T_S}{T_G+F_G+T_S+F_S}$.

We used the following set of colours/magnitudes as input parameters: $W1$, $W1-W2$, $W2-W3$, $W3-W4$, $J-H$, $H-K_s$, $W1-J$, $W2-H$, and $W3-K_s$. Initially, we supplied SVM with all possible pairs of this set, and obtained contamination, completeness, and accuracy. The parameter $W1-J$ is an astoundingly good star–galaxy separator, as shown in Fig. 1. Either alone or combined with any other parameter, $W1-J$ guarantees the lowest stellar contamination, the highest galaxy completeness, and the highest accuracy. For instance, the stellar contamination for the combination of $W1$ and $W1-J$, or $H-K_s$ and $W1-J$ is as low as 3.1 per cent, while the galaxy completeness is 93.6 per cent (see Table 1).

Next, we supplied the SVM with more parameters. We started with $W1-W2$ alone, then added one more parameter in each step. Our findings are summarized in Fig. 2. We qualitatively confirmed our former results, namely that the combination of *WISE* and 2MASS parameters increases the SVM performance. For *WISE*

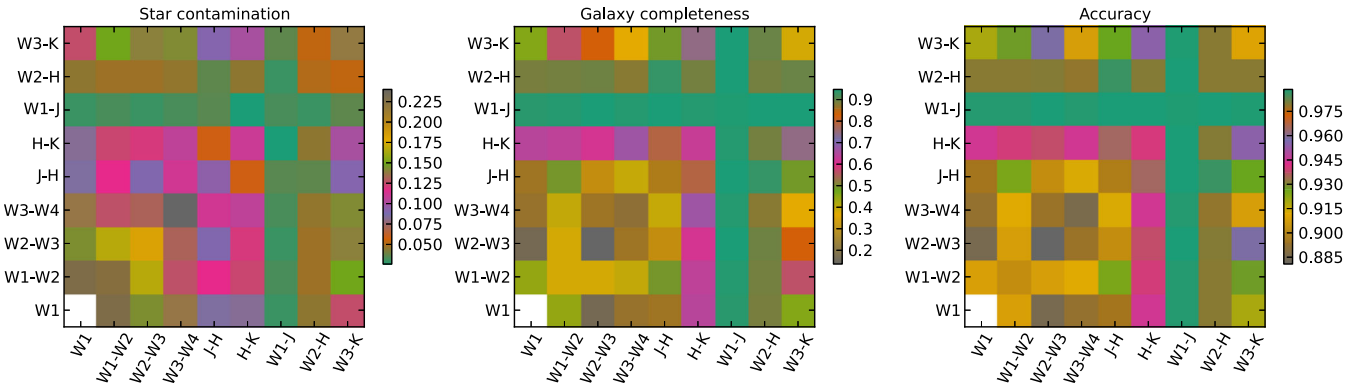

Figure 1. Measures of SVM performance are presented in the case of pairwise and single usage. Colour-coded maps illustrate contamination, completeness, and accuracy for every combinations. All subfigures suggest that $W1-J$ is a dominant potency in star–galaxy separation. We note, however, that SVM failed to produce precise results using $W1$ alone. Every object was classified as a star with that choice, thus we excluded the $W1$-only case from the analysis. Combinations of $W1$ and other parameters, however, are preserved, as they produce valuable results.

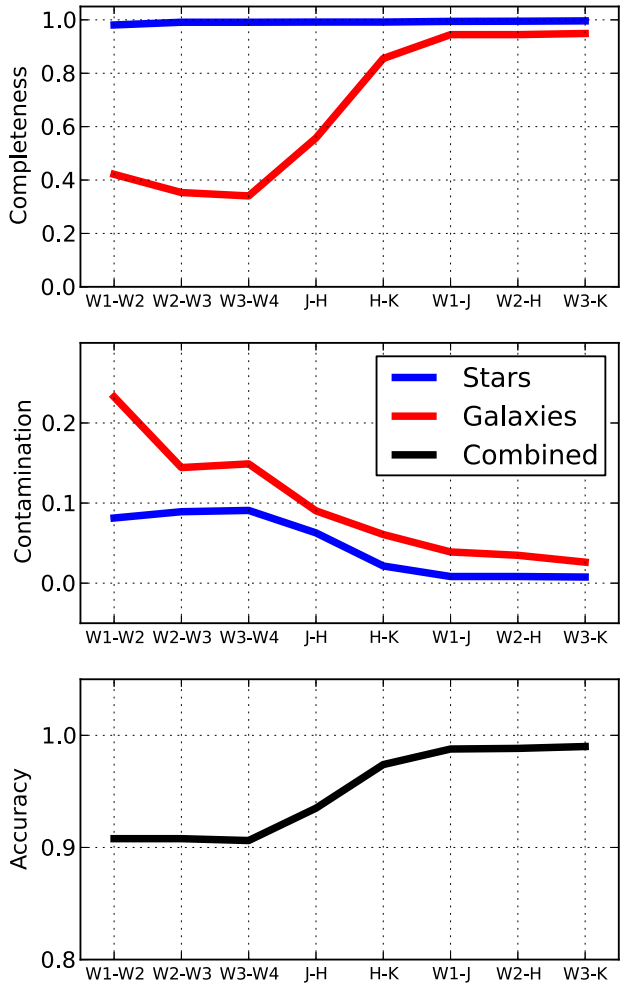

Figure 2. Measures of SVM performance are shown as a function of SVM parameters. We observed upgrading trends in contamination, completeness, and accuracy for both stars and galaxies. High completeness values for the star sample can be explained by the fact that the sample is dominated by stars, thus False galaxies cannot affect star completeness significantly.

colours only, galaxy completeness is at the level of ≈40 per cent, while with all parameters it reaches ≈93 per cent. At the same time, stellar contamination decreased from ≈15 to ≈3 per cent. Finally, similar trends are seen for the accuracy parameter, that incremented by ≈10 per cent by adding 2MASS parameters.

As a test of possible impurities in the training sample, we randomly flipped the classification of 1, 3, 5, and 10 per cent of the training objects, and repeated our SVM analysis with the artificially contaminated training sets. We analysed the case of the combination of $W1 - J$ and $W1$, finding 3.7 per cent, 4.8 per cent, 7.7 per cent, and 11.2 per cent for the star contamination, respectively, while the original value with the unchanged training set was 3.1 per cent.

3.2 SVM versus colour–colour and colour–magnitude cuts

Findings of the previous subsection suggest that separation of stars and galaxies can be achieved a simple cut on the $W1$–$W1-J$ colour–magnitude plane. Stellar contamination and galaxy completeness are then comparable to that of the multicolour SVM algorithm, but with a faster method. Fig. 3 shows the estimated stellar contamination and the ratio of properly identified and lost galaxies in the

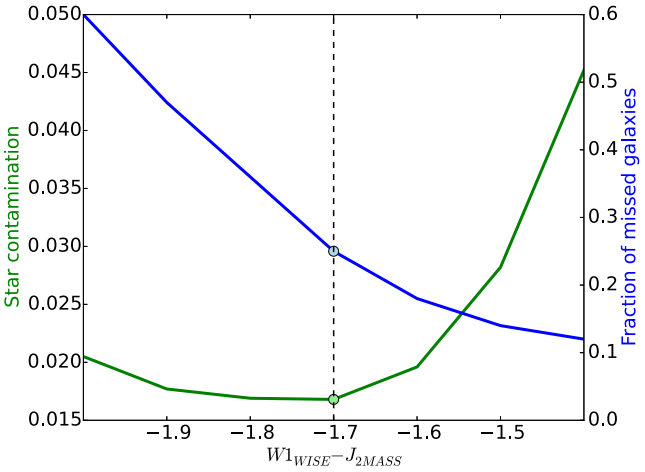

Figure 3. Star contamination (blue curve, values on the left axis) and the fraction of the lost galaxies (green curve, values on the right axis) is shown. We applied a $W1-J \leq -1.7$ cut, and now show the consequences for the parameters of the resulting catalogues.

case of different $W1 - J$ cuts. We choose $W1 - J \leq -1.7$ for our purposes, as it guarantees the lowest stellar contamination, while 78.6 per cent of the galaxies can be classified as galaxies.

Visualizing this parameter choice, we show a $W1$–$W1-J$, and *WISE* colour–colour diagrams for *WISE*-2MASS objects from subsample of 20 000 objects (with 13 per cent galaxy content) in Fig. 4. Classes are indicated by SDSS in this comparison. We note that a remarkable separation of stars and galaxies can be seen in the upper left of Fig. 4.

However, the patterns we found in Fig. 5 enforce a $W1 \geq 12.0$ magnitude cut, as a larger subsample of SDSS 'galaxies' imbedded in the definite stellar locus of this plot. This fact suggests that these objects might have been misclassified by SDSS, and their usage is unsafe in a training set. We investigated the actual SDSS image of these 'galaxy' objects, and found that they are indeed really bright stars with bad classification. We emphasize, that neither our SVM methods nor the $W1$–$W1-J$ based simple galaxy selection are not affected, since we removed these brightest objects from our sample.

3.3 Further comparisons

Next, we compare our galaxy sample to that of Goto et al. (2012), and Kovács et al. (2013). While these works used all four observations of *WISE*, we only need $W1$, i.e. the one with the best quality observations. The stellar contamination we estimate for Goto et al. (2012), and Kovács et al. (2013) using the SDSS classification is 7 per cent, while only 1.8 per cent for our present sample applying the $W1 - J \leq -1.7$ cut. At the same time, while previous *WISE*-only methods produced 21 per cent galaxy completeness, presently we achieved a 78.6 per cent with the new galaxy selection criteria. SVM results reach ≈93 per cent completeness for galaxies, with similar stellar contamination as the $W1 - J$ cut. Fig. 6 summarizes our findings.

We note that the stellar contamination may be higher where the number density of stars is above the average, e.g. close to the Galactic plane, or at the Small and Large Magellanic Clouds. Among others, these regions should be masked out in order to avoid misclassification problems.

There are other object separation algorithms in the literature, but either they are optimized for QSO-AGN selection (Yan et al. 2013), or limited to bright magnitude cuts (Jarrett et al.

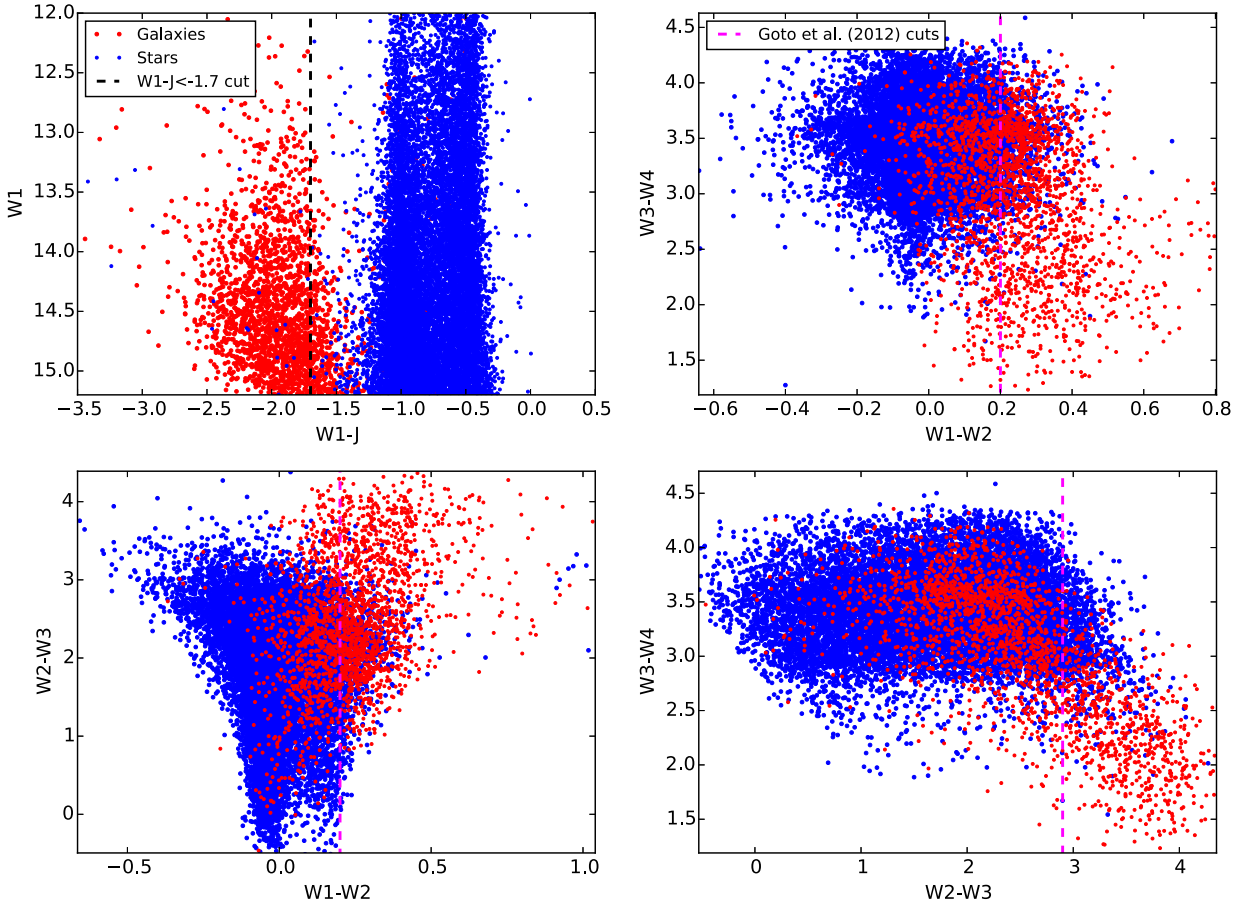

Figure 4. Top left: a simple star–galaxy separator which uses only $W1$, and $W1-J$ colour. The separation of stars and galaxies is remarkably strong in this parameter space. Other subfigures show colour–colour plots of the four *WISE* bands. We show the special galaxy separator cuts applied by Goto et al. (2012). This result illustrates that star–galaxy separation on traditional colour–colour planes with linear cuts is challenging, if one wants to use a large fraction of the achievable galaxy sample.

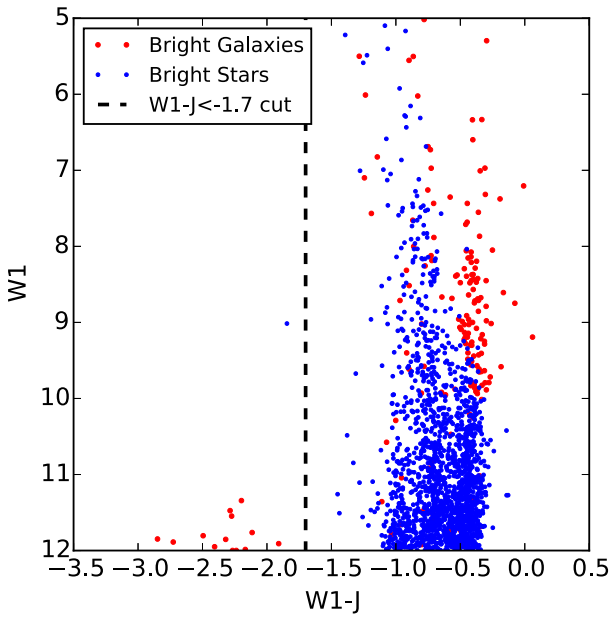

Figure 5. The bright 'galaxies' of $W1 \geq 12.0$ mag with potentially bad SDSS classification are located in the stellar locus. We excluded all bright objects from the analysis, since their presence can alter the training efficiency. See text for details.

2011). We argue, therefore, that a direct and detailed comparison is only possible with the results of Goto et al. (2012), and Kovács et al. (2013).

3.4 All-sky galaxy catalogue

$W1 - J$ cut appears to be a powerful tool for separating stars and galaxies, providing a fast and simple option to create a full-sky *WISE*-2MASS galaxy map. The simple cut can be realized by a query into the *WISE*-2MASS data base. As $W1-J \leq -1.7$ gives the lowest contamination according to our tests, we selected galaxies with the following query:

```
w1mpro between 12.0 and 15.2 and
n_2mass > 0 and
w1mpro - j_m_2mass < -1.7 and
glat not between -10 and 10
```

where 'w1mpro' is the $W1$ brightness, 'n 2mass' is the number of the associated 2MASS sources, 'j m 2mass' corresponds to the brightness in the J band, and 'glat' is the Galactic latitude coordinate.

We downloaded ∼5 million *WISE*-2MASS objects from the IRSA website.[1] The data set contained $W1$, $W2$, $W3$, and $W4$ for *WISE*,

[1] http://irsa.ipac.caltech.edu/

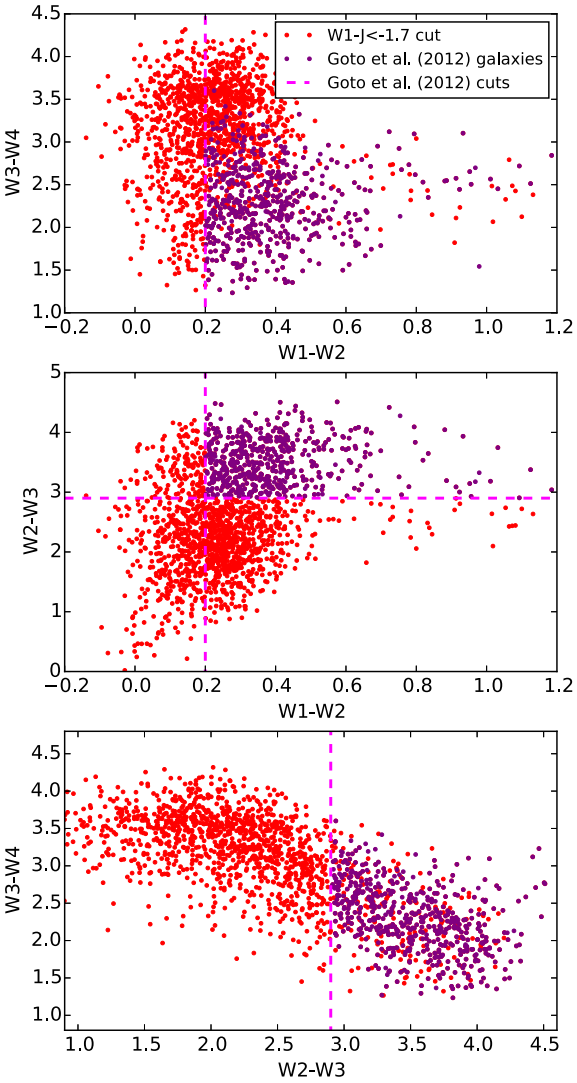

Figure 6. Distributions of galaxies on colour–colour plane is shown for (Goto et al. 2012) cuts and galaxies selected in this work. Significant amount of galaxies is identified properly with the $W1-J \leq -1.7$ galaxy cut. We used 1785 galaxies for making this figure, i.e. all galaxies that we identified properly using our simple CC tool (see Fig. 4 and Table 1 for details). This represents a 78.6 per cent galaxy completeness.

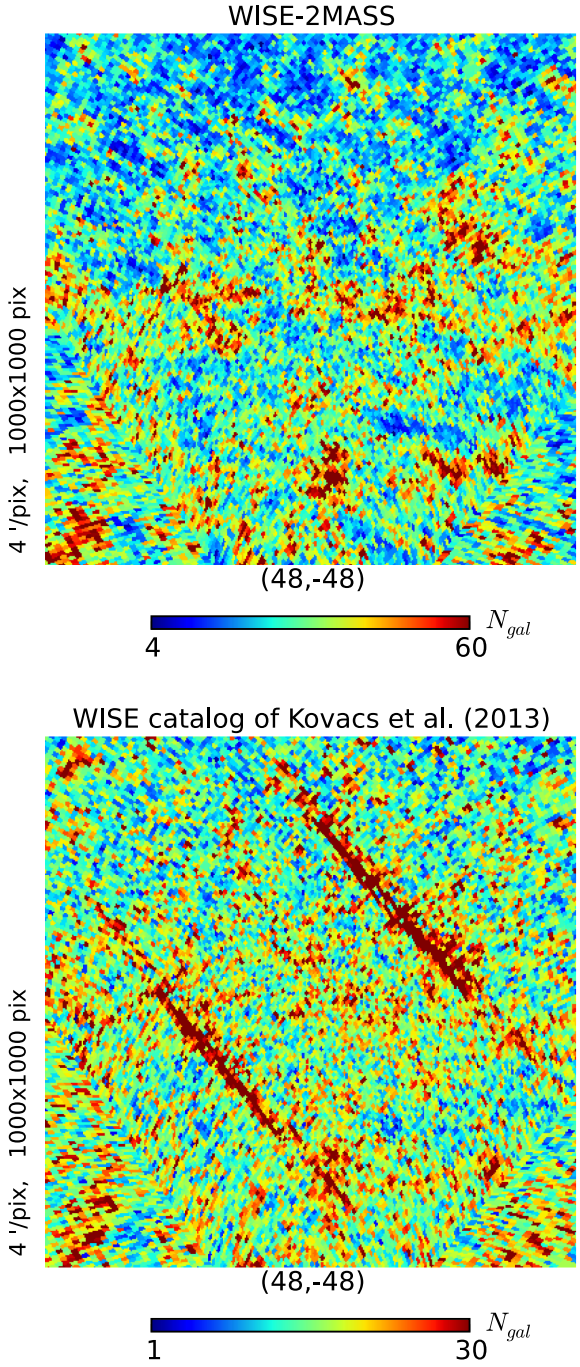

Figure 7. Gnomonic projection of galaxy number counts in HEALPIX pixels at $N_{\rm side} = 128$ is shown for the sample of (Kovács et al. 2013, top) and present approach (bottom). Figures are centred on $\ell, b = 48.0, -48.0$.

and J, H, and K_s for 2MASS as photometric parameters, and we also downloaded 'cc flag' values to uphold the possibility of further restrictions.

Next, we test for possible biases and systematic problems that may affect our selections. We have demonstrated that our SVM algorithms and $W1 - J$ colour cuts are capable of separating stars and galaxies, but the accuracy of the galaxy selection does not guarantee for instance spatial homogeneity. We have shown in Kovács et al. (2013), that observational strategies of surveys may be harmful for the galaxy samples we wish to create, and inhomogeneities might show up as consequences of varying sensitivity or other observational effects. First, we investigate the stripe-shaped overdensities at several locations across the sky we found in Kovács et al. (2013). These artefacts caused by moon glow are significantly reduced in the new catalogue, as shown on the upper panel of Fig. 7. As it was pointed out by Kovács et al. (2013), stripes are associated with the scanning strategy of the *WISE* survey. Different *WISE* bands have different sensitivities and sky coverage, therefore affect the uniformity of a full sky sample through moon-glow contamination. Kovács et al. (2013) handle this issue with special moon-contamination mask using the 'moonlev' flag of the *WISE* data base. The stripes are not present in a $W1 - J \leq -1.7$ selected data set.

The next probe is the test of homogeneity across the sky, in particular the possible gradient in the density field as a function of Galactic latitude, as reported in Kovács et al. (2013). We assume that the limit of $W1 < 15.2$ for *WISE* is conservative enough, although

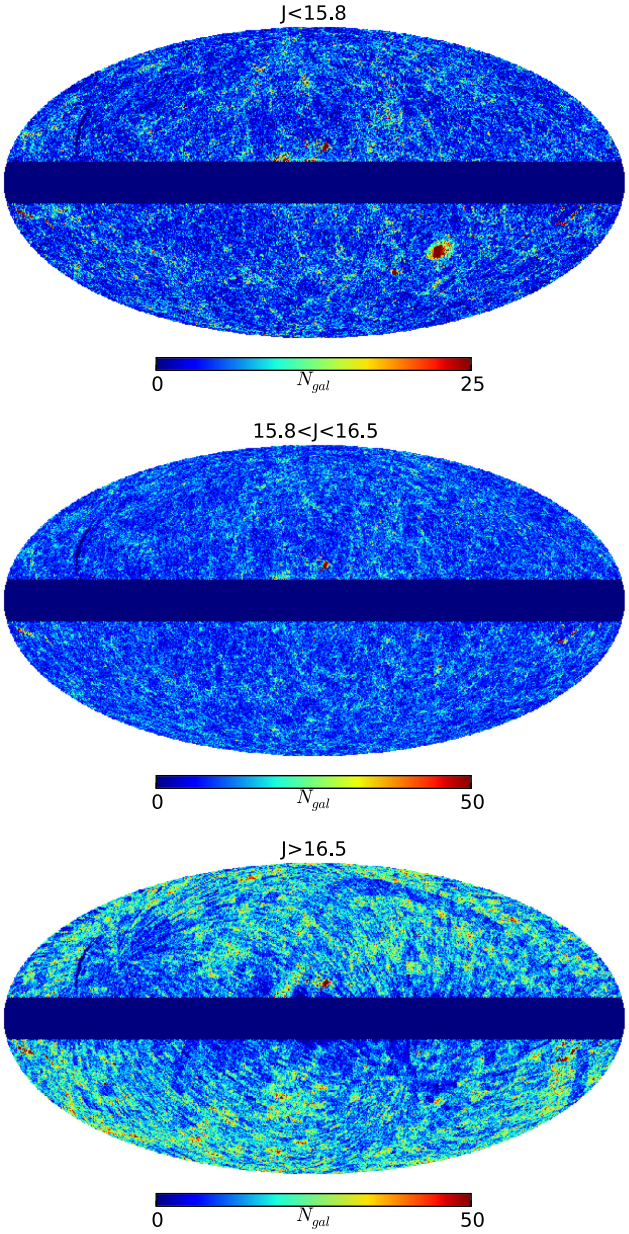

Figure 8. Slices of 2MASS J magnitude as probes of spatial homogeneity. We use HEALPIX at $N_{\rm side} = 128$ for making these plots.

we do not apply any limitations to 2MASS data. Naturally, there is a correlation between the magnitudes of objects observed in *WISE* and 2MASS, but the cut of $W1 - J < -1.7$ does not automatically guarantee high signal-to-noise ratio for 2MASS as well. Therefore, we divide our sample into three slices of 2MASS J brightness. We use $J < 15.8$ as a first limit that is the corresponding 10σ photometric detection limit for 2MASS-PSC objects. This would result a dramatic loss of objects, as only ∼800 000 galaxies remain in the sample. As our goal is to broaden the relatively shallow 2MASS XSC galaxy sample, we empirically obtain a higher magnitude limit for J with experiments. We experimentally found that a $J < 16.5$ cut effectively removes a significant amount of inhomogeneous data from our catalogue. The $J > 16.5$ map, however, is strongly affected by spatial variations of the sensitivity of the 2MASS-PSC data, as shown in Fig. 8. We thus remove these objects from the analysis.

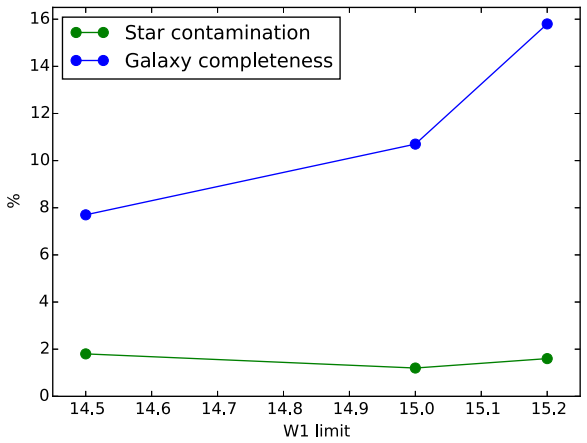

Figure 9. Star contamination and galaxy completeness differences for SVM and CC methods as function of $W1$ magnitude limit (we use the case of an extra $J < 16.5$ mag cut for $W1 = 15.2$). SVM methods result in higher contamination, but increased completeness in all cases.

We test the SVM performance and colour cuts applying different flux limits for $W1$ and J. In SVM analysis, we test for the case of the $W1$ and $W1 - J$ parameter pair in order to use the same information as used in the case of colour cuts. Our findings are summarized in Table 1, and in Fig. 9.

We note that we observe many apparent overdensities near the Galactic plane, and Large and Small Magellanic Clouds are visible in the lower panel of Fig. 8. This finding reflects the fact that our star–galaxy separation tools fail more frequently at regions of high stellar density, as estimated and expected. We argue, however, that both the $J < 15.8$ map, the $15.8 < J < 16.5$ map, and their union effectively trace large-scale structure, and do not contain significant large-scale inhomogeneities. The resulting galaxy map does not suffer from strange stripe-shaped overdensities, and confirms our finding presented in Fig. 7. We note that this additional brightness cut shifts the median redshift of the sample from $z \approx 0.17$ to 0.14.

We further probe the uniformity our catalogue by performing tests on possible gradients in galaxy number counts as a function of Galactic latitude, finding no such effect in the *WISE*-2MASS map. Fig. 10 illustrates our findings.

Finally, we construct a mask to exclude potentially contaminated regions near the Galactic plane using the dust emission map of Schlegel, Finkbeiner & Davis (1998). We mask out all pixels with $E(B - V) \geq 0.1$, and regions at galactic latitudes $|b| < 10^{\circ}$, leaving 21 200 deg^2 for our purposes. Unmasked regions in the galaxy map are corrected for Galactic extinction using the same dust map provided by Schlegel et al. (1998). We use $A_{WISE}/E(B - V) = 0.18$ and $A_{\rm 2MASS}/E(B - V) = 0.72$ coefficients estimated by Yuan, Liu & Xiang (2013). The final all-sky galaxy map is shown in Fig. 11.

We note that the star–galaxy separation methods we developed are useful for selecting stellar samples as well. For instance, a $W1 - J \geq -1.3$ colour cut should result a clean sample of stars. However, a detailed selection of specific types of stars needs further refinements.

CONCLUSIONS

We focused on creating large area galaxy maps of low stellar contamination and high galaxy completeness based on the joint analysis of *WISE* and 2MASS photometric data sets. Using 2MASS colours

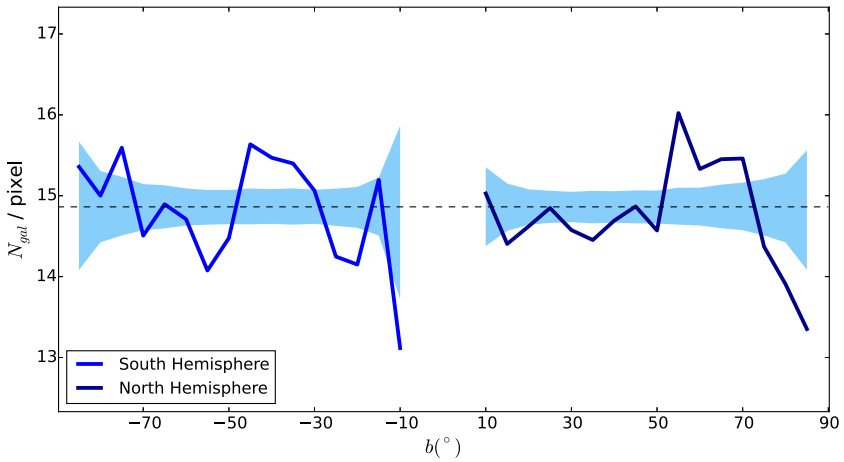

Figure 10. Galaxy number count measurements in rings as a function of Galactic latitude b for the $J < 16.5$ galaxy map. Horizontal dashes indicate the average number of galaxies per pixel in the full map, outside the mask we constructed. Light blue shaded regions indicate Poisson errors for the measurement. We attribute the largest fluctuations to the presence of the largest superstructures in the local Universe. Note that these fluctuations are at the $\sim 2\sigma$ level, except the one at $b \approx 60^\circ$. Possibly these fluctuations can be lowered using a more sophisticated mask which effectively removes further potentially contaminated regions.

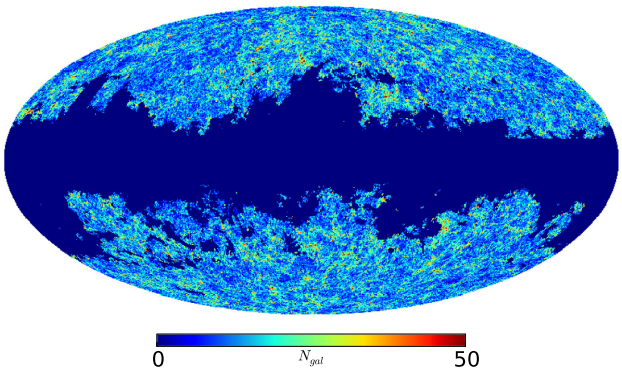

Figure 11. Our example *WISE*-2MASS galaxy number count map applying the extra $J < 16.5$ mag cut, and the corresponding mask that we constructed. We use HEALPIX at $N_{\rm side} = 128$.

add useful information, while ∼93 per cent of the *WISE* objects with $W1 < 15.2$ mag have 2MASS pairs. We performed star–galaxy separation using a class of wide-spread machine learning tools, SVM. *WISE*-2MASS objects were cross-identified with SDSS objects, and available SDSS PhotoObj classification data were used as training and control sets. Exhaustive testing of the SVM algorithm with different parameters and inputs revealed that a simple $W1-J$ photometric colour cut produces similarly clean data set as the SVM classification, at the expense of losing <15 per cent of the galaxies (Table 1).

We produced a clean galaxy sample with 1.8 per cent stellar contamination reaching 78.6 per cent completeness using our basic $W1 \leq 15.2$ and $W1-J \leq -1.7$ cuts. The SVM techniques for the same data yield 3.1 per cent stellar contamination, and 93.4 per cent galaxy completeness. We found at all magnitude limits that SVM algorithms result in slightly higher stellar contamination (an extra ∼1.5 per cent), but with notable gain in the total number of galaxies identified properly (∼10 per cent).

Contamination and completeness are, however, not the only relevant properties of a high-quality galaxy catalogue. We probed the isotropy and homogeneity of the resulting galaxy maps, and empirically found that the faintest 2MASS objects at $J > 16.5$ mag show inhomogeneities in their density on the sky, presumably due to the survey strategy. We thus supplemented our standard flux and colour selection criteria by a $J < 16.5$ mag cut in order to produce a more uniform catalogue. The resulting catalogue contains $N_{\rm gal} \approx 2.4$ million objects, with an estimated star contamination of 1.2 per cent, and 70.1 per cent galaxy completeness. SVM estimates show 2.8 per cent contamination, and 85.9 per cent completeness for this shallower data. These examples demonstrate that the computationally expensive SVM approach creates a more complete catalogue with higher stellar contamination.

Regardless the methods applied, the resulting galaxy catalogues represent significant improvement over previous samples using *WISE* colours only for selection (Kovács et al. 2013). We not only trace a much larger region in the *WISE* colour–colour space than previous catalogues, but also restrict the galaxy selection for only the least noisy *W*1 band for *WISE*.

The need for high completeness while minimizing contamination is clear, although we argue that there is no optimal galaxy map in general, since specific science drivers will require different balance between these two effects and thus the cuts that control them. Nevertheless, when clean catalogue is needed with a simple definition (Kovács et al. 2013; Finelli et al. 2014; Szapudi et al. 2014), we recommend $W1-J \leq -1.7$ colour cut as a compromise between simplicity, completeness, and the lowest possible stellar contamination. Other applications, e.g. galaxy cluster counting, gravitational wave source follow-up, etc. might benefit even from denser galaxy maps with higher contamination, and/or non-uniform coverage.

In the near future, we will add photometric redshifts to our catalogue matching with SuperCOSMOS (Hambly et al. 2001), extending Bilicki et al. (2014). While we plan to make this value added *WISE*-2MASS-SuperCOSMOS catalogue public, at present the *WISE*-2MASS catalogues can be easily downloaded from the IRSA website[2] using the queries quoted earlier in this paper.

[2] http://irsa.ipac.caltech.edu/

ACKNOWLEDGEMENTS

AK takes immense pleasure in thanking the support of OTKA through grant no. 101666. In addition, AK acknowledges support from Campus Hungary fellowship programme. IS acknowledges support from NASA grants NNX12AF83G and NNX10AD53G. We thank István Csabai for useful comments improving the SDSS-*WISE*-2MASS matching properties, and target selection. We thank the constructive comments by the reviewer of our paper, Katarzyna Malek. Funding for this project was partially provided by the Spanish Ministerio de Economa y Competitividad (MINECO) under project Centro de Excelencia Severo Ochoa SEV-2012-0234. We use HEALPIX (Gorski et al. 2005). This publication makes use of data products from the *WISE*, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration, and data products from the 2MASS, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This work also makes use of data products of the GAMA survey,[3] and the SDSS.[4]

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Amendola L., Appleby S., Bacon D., Baker T., Baldi M., Bartolo N., Blanchard A., Bonvin C., 2013, Living Rev. Relativ., 16, 6
Bentley J. L., 1975, Commun. ACM, 18, 509
Bilicki M., Jarrett T. H., Peacock J. A., Cluver M. E., Steward L., 2014, ApJS, 210, 9
Brightman M., Nandra K., 2012, MNRAS, 422, 1166
Chiu K. et al., 2005, AJ, 130, 13
Cortes C., Vapnik V., 1995, Mach. Learn., 20, 273
Driver S. P. et al., 2011, MNRAS, 413, 971
Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15
Finelli F., Garcia-Bellido J., Kovacs A., Paci F., Szapudi I., 2014, preprint (arXiv:e-prints)
Gorski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Goto T., Szapudi I., Granett B. R., 2012, MNRAS, 422, L77
Hambly N. C. et al., 2001, MNRAS, 326, 1279
Huertas-Company M. et al., 2009, A&A, 497, 743
Jarrett T. H., Chester T., Cutri R., Schneider S., Skrutskie M., Huchra J. P., 2000, AJ, 119, 2498
Jarrett T. H., Cohen M., Masci F., Wright E., Stern D., Benford D., Blain A., Carey S., 2011, ApJ, 735, 112
Kaiser N., Burgett W., Chambers K., 2010, Proc. SPIE, 773 Ground-based and Airborne Telescopes III, 77330E
Kovács A., Szapudi I., Granett B. R., Frei Z., 2013, MNRAS, 431, L28
LSST Science Collaboration et al., 2009, preprint (arXiv:e-prints)
Małek K. et al., 2013, A&A, 557, A16
Pollo A., Rybka P., Takeuchi T. T., 2010, A&A, 514, A3
Richards G. T. et al., 2002, AJ, 123, 2945
Saglia R. P. et al., 2012, ApJ, 746, 128
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Schlegel D. et al., 2011, preprint (arXiv:e-prints)
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Solarz A. et al., 2012, A&A, 541, A50
Stern D. et al., 2012, ApJ, 753, 30
Szapudi I. et al., 2014, preprint (arXiv:e-prints)
The Dark Energy Survey Collaboration, 2005, preprint (arXiv:e-prints)
Vasconcellos E. C., de Carvalho R. R., Gal R. R., LaBarbera F. L., Capelato H. V., Frago Campos Velho H., Trevisan M., Ruiz R. S. R., 2011, AJ, 141, 189
Woźniak P. R., Williams S. J., Vestrand W. T., Gupta V., 2004, AJ, 128, 2965
Wright E. L. et al., 2010, AJ, 140, 1868
Yan L. et al., 2013, AJ, 145, 55
Yuan H. B., Liu X. W., Xiang M. S., 2013, MNRAS, 430, 2188

[3] http://www.gama-survey.org/
[4] http://www.sdss.org/

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