VEXAS:
the VISTA EXtension to Auxiliary Surveys
Data Release 2. Machine-learning based classification of sources in the Southern Hemisphere

C. Spiniello[1,2], V. Khramtsov[3,4], A. Agnello[5] & A. Sergeyev[1,6]

[1]Sub-Dep. of Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, UK
[2]INAF - Osservatorio Astronomico di Capodimonte, Salita Moiariello, 16, I-80131 Napoli, Italy
[3]Institute of Astronomy, V. N. Karazin Kharkiv National University, 35 Sumska Str., Kharkiv, Ukraine
[4]Department of Data Science, Quantum, 20, Otakara Yarosha lane, Kharkiv, Ukraine
[5]DARK, Niels Bohr Institute, Copenhagen University, Lyngbyvej 2, 2100 Copenhagen, Denmark
[6]Institute of Radio Astronomy of the National Academy of Sciences of Ukraine, 4, Mystetstv St., Kharkiv, 61002, Ukraine

Abstract

This document describes the second public data release of the VISTA EXtension to Auxiliary Surveys (VEXAS, Spiniello & Agnello, 2019, A&A, 630, hereafter S19), which is described Khramtsov et al. 2021, (arXiv:2103.09257 hereafter K21).
In this VEXAS DR2, we classify objects into stars, galaxies and quasars using an ensemble of thirty-two different machine learning models, based on three different algorithms and on different magnitude sets, training samples and classification problems (two or three classes). We apply the ensemble learning on the three VEXAS DR1 optical+infrared (IR) tables VEXAS-DESW, VEXAS-PSW and VEXAS-SMW.
The final aim of VEXAS is to build the widest multi-wavelength catalogues (from X-ray to radio), providing reference magnitudes, colours and morphological information for a large number of scientific uses: object classification (e.g., quasars, galaxies, and stars, K21; high-z galaxies, white dwarfs, etc.); photometric redshifts of large galaxy samples; searches of exotic objects (e.g., extremely red objects and lensed quasars).
As of March 2021, the VEXAS catalogue is the widest and deepest public optical-to-IR photometric and spectroscopic database in the Southern Hemisphere.

Overview of Observations

In this second data release, covering the Southern Galactic Hemisphere (SGH) below the Galactic plane (at b<-20 deg), we limit ourselves to the three optical+infrared DR1 tables created cross-matching VISTA (Emerson et al., 2006) near-infrared data (the VISTA Hemisphere Survey, VHS, McMahon et al., 2013; and the VISTA Kilo Degree Infrared Galaxy Survey, VIKING, Sutherland et al., 2012) with WISE far-infrared data and with optical magnitudes from the Dark Energy Survey (DES, Abbott et al., 2018), the SkyMapper Southern Sky Survey (Wolf et al., 2018) and the Panoramic Survey Telescope & Rapid Response System (PanSTARRS DR1, Chambers et al., 2016).

The core requirement set in VEXAS DR1 for the assembly of these tables is a reliable photometry in more than one band. This condition, together with the detection in at least two surveys (via cross-match), should minimize, if not completely eliminate, the number of spurious detections in the final catalogues.

In this VEXAS-DR2, we use a filtered version of the DR1 tables: we remove all sources fainter than 25 mag in each considered band, since below this value the extrapolation of the training sample cannot be tested. In the case of VEXAS-PSW, we also apply a more severe cut on the optical magnitudes and associated uncertainties: we restrict ourselves to magnitudes brighter than the mean 5-sigma point-source limiting sensitivity values given in Chambers et al. (2016) and we filter out all sources with uncertainties >1 mag.

For the VEXAS-SMW table, in DR1 we limited ourselves to declinations<-30 degrees, since above this declination the PSW coverage is uniform and at least two magnitudes deeper. Here in DR2 we consider the whole coverage of Sky Mapper in the SGH, extending the VEXAS-SMW table to ~32M unique objects. This allows us to compare the results obtained from training the pipeline on SMW with that obtained training it on DESW and PSW (see K21 for more details).

The three optical+infrared "cleaned" tables constitute the photometric datasets used for this VEXAS-DR2. Their sky coverage is shown in Figure 1, colour-coded by object density (number of objects per deg^2).

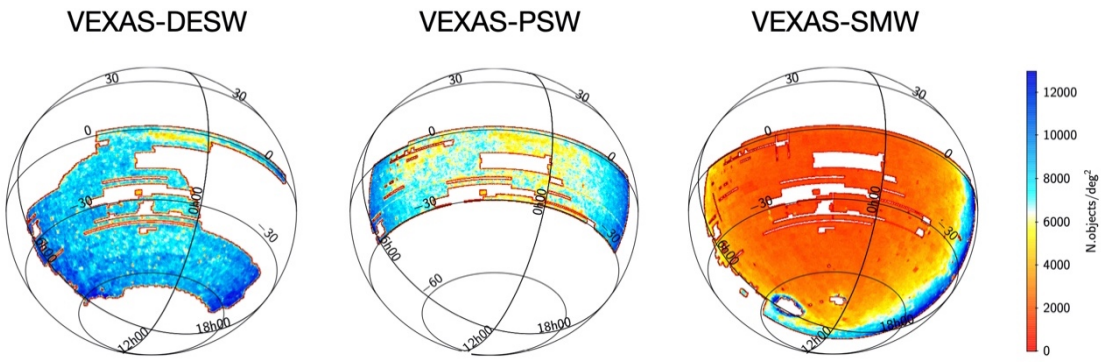


Figure 1: Sky coverage view of the three input VEXAS optical+IR tables. The colours indicate the number of objects per deg^2, as shown by the side bar, obtained using a Hierarchical Equal Area iso-Latitude Pixelation of a sphere (HealPix) with resolution equal 9.

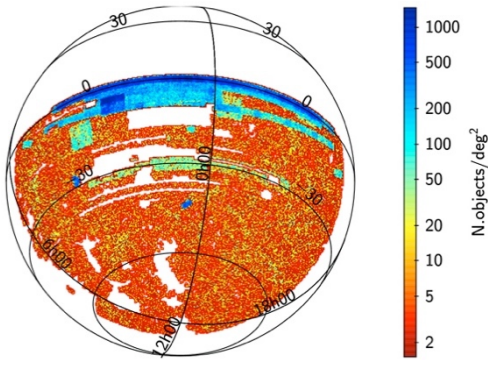
Training sample: The VEXAS-SPEC-GOOD table

As training dataset for the ML process, we combine data from six different spectroscopic surveys (SDSS DR16, Ahumada et al. 2019; WiggleZ, Drinkwater et al. 2018; GAMA DR3 Baldry et al.2018; OzDES, Childress et al. 2017, 2QZ Croom et al. 2003 and 6dFFS DR3, Jones et al. 2009) in order to build a training sample as large as possible and as complete as possible in all the three classes of objects (star, galaxy and qso). We use a matching radius of 1.5" (since the resolution in the optical is better than that in the IR), and perform the match with TOPCAT ("*Tool for OPerations on Catalogues And Tables"*, Taylor, 2005).

A detailed description how we select and combine these different sets into the training, validation and testing samples is provided in K21. The data always include spectroscopic classification of the sources and redshift information for extra-galactic objects. We also apply some selection criteria on the spectra to select only sources passing the quality level thresholds recommended directly by the data providers.

Combining the six `cleaned' spectroscopic tables described above, and cross-matching the resulting table with the three input tables, we assemble a final spectroscopic table for each of the VEXAS tables. In particular, for VEXAS-DESW, VEXAS-PSW, and VEXAS-SMW we find 293584, 328821, and 211092 unique spectroscopic sources, respectively.

Despite the fact that we use these three separate spectroscopic samples for the three VEXAS input tables, we release here, as part of the DR2, the VEXAS-SPEC-GOOD, comprising 415628 unique VEXAS objects with photometry in the optical and infrared and a secure and clean spectroscopic classification. This table, comprising 89222 unique stars, 35179 unique quasars and 291227 unique galaxies update the VEXAS-SPEC-DR1. The footprint of the table, colour-coded by object density (number of objects per deg^2 , is plotted in Figure 2, which also list, on the table to the right, how many objects per each class are found in each survey.



Survey	Class	VEXAS DESW	VEXAS PSW	VEXAS SMW
SDSS	STAR	56 187	74 794	77 100
	QSO	23 820	26 872	4 660
	GALAXY	144 495	162 717	65 937
2QZ	STAR	537	151	701
	QSO	3499	657	699
	GALAXY	368	114	46
OzDES	STAR	1 614	1 216	1 569
	QSO	1 191	998	220
	GALAXY	7 246	4 438	1 778
6dFGS	STAR	506	649	1 544
	GALAXY	25 274	21 792	54 093
GAMA	GALAXY	33 050	34 551	11 578
WiggleZ	GALAXY	7 851	12 759	723

Figure 2: Sky coverage view of the VEXAS-SPEC-GOOD final table and number of objects in each cathegory for each spectroscopic survey. The colour indicates the number of objects per deg^2, in logaritmic scale, as shown by the side bar, obtained as in Fig.1.

Release Content

In this VEXAS-DR2, we release four tables, each updating a table already released in DR1. They are listed below, with the total number of unique sources from the VISTA Surveys contained in each of them. Their file format and structure are given in the **Data Format** section.

The record IDs are fully compatible between the old and new version of the tables for common objects (we reimind the readers that we apply some cuts to the DR1 sources and we extended the sky coverage of the VEXAS-SMW), which means that it is possible to reproduce the results of previous queries using the new version of the data.

We note that the uniqueness of the objects in the matched surveys is not a requirement. In fact, given the different image resolutions of the different surveys, it might likely happen that a single object in e.g., WISE or in the spectroscopic survey result in multiple matches with VISTA.

TABLE NAME	Corresponding DR1 table	Number of unique sources in VISTA
VEXAS-DESW_V2	VEXAS-DESW_V1	35'381'482
VEXAS-PSW_V2	VEXAS-PS1W_V1	21'891'609
VEXAS-SMW_V2	VEXAS-SMW_V1	31'999'995
VEXAS-SPEC-GOOD_V2	VEXAS_SPEC_V1	415'656

Release Notes

All matches were performed with the "*Tool for OPerations on Catalogues And Tables*" (*TOPCAT*, Taylor, 2005) using the `Join - Pair Match'.

We note that all the magnitudes in the tables are provided in their native system of reference (AB for the optical surveys and Vega for VISTA and AllWISE).

Known Issues

Safe ranges and Outliers

One of the most common issue in machine learning based classifications is that the 'depth' of the data is larger than that of the training set. The most common solution is to cut out the input tables limiting the inference to bright objects, and thus avoid any extrapolation to unseen regions in the space of features (e.g., Khramtsov et al. 2019, 2020). However, this approach is not desirable in our case, since it would violate the main purpose of the VEXAS Project: collect as much information as possible in the multi-wavelength sky and thus classify the largest possible number of sources. Thus, in K21, we do not restrict the classification to the brightest sources only, but nevertheless study the completeness of the training samples in the r-band for each VEXAS input table. We define as "safe ranges" the (central) r-band intervals properly covered by the training sample and where the ratio between the training set and the corresponding input table is larger than 0.1 %.

In the Appendix of K21, we demonstrate that the classification outside these safe ranges can still be trusted assuming that the object is not an "outlier", i.e. a source whose colour distribution is completely offset with respect to that of the training sample used to classify it. We therefore add a "warning flag" to the released tables which can take three values: 0 if the source is within the safe ranges and is not an outliers, 1

if the object is outside the safe range but is not an outlier, 2 if the source is an outlier in at least one colour (see the Release content for more details).

Imputation for the VEXAS-SMW

For the VEXAS-SMW table the percentage of missing magnitudes in the optical is much larger than that in the other two tables. Although in K21 we show evidence supporting the fact that the object classification based on the ensemble learning is not affected by the non-optimal imputation, for this table we release also the classification probabilities obtained from a single model (#25) without imputation, in addition to the probabilities from the META_MODEL.

Data Format

Files Types

This release contains four multi-band catalogues, all given as single-file catalogue (monolithic) in the FITS format, and all associated with the VEXAS-AllWISE table released in DR1. The tables are linked together using the SOURCEID_VISTA (unique identifier of the merged detection in VHS or VIKING) to identify the same objects in different tables (UCD: meta.id;meta.main). The identifier is provided in all tables.

In general, each table contains the source ID from one or more surveys, FK5 J2000.0 coordinates and magnitudes in various bands with associated errors, the probability for each source to belong to the class of STARS, GALAXY or QSOs, the Imputation flag and the Warning flag, already described in the Known Issue session.

Catalogue Columns

The following tables list the columns that are present in each of the catalogue, with their unit and a short description of the content they represent.

TABLE 1: VEXAS-DESW

NAME	UNIT	DESCRIPTION
SOURCEID_VISTA		UID of the merged detection in VISTA
Coadd_object_id_DES		Unique ID in DES
SLAVEOBJID_WISE		The unique ID of the neighbour in calSource (=sourceID) for the 10arcsec match
ID_WISE		AllWISE ID for the 3 arcsec match
RA_DES	degrees	FK5 J2000.0 Right Ascension in DES
DEC_DES	degrees	FK5 J2000.0 Declination in DES
mag_auto_g_DES	mag (AB)	Magnitude in the *g*-band from DES
magerr_auto_g_DES	mag (AB)	Magnitude error in the *g*-band from DES
mag_auto_r_DES	mag (AB)	Magnitude in the *r*-band from DES
magerr_auto_r_DES	mag (AB)	Magnitude error in the *r*-band from DES
mag_auto_i_DES	mag (AB)	Magnitude in the *i*-band from DES
magerr_auto_i_DES	mag (AB)	Magnitude error in the *i*-band from DES
mag_auto_z_DES	mag (AB)	Magnitude in the *z*-band from DES
magerr_auto_z_DES	mag (AB)	Magnitude error in the *z*-band from DES
mag_auto_y_DES	mag (AB)	Magnitude in the *y*-band from DES

magerr_auto_y_DES	mag (AB)	Magnitude error in the *y*-band from DES
spread_model_i_DES		Stellarity indicator in the *i*-band as defined by the Dark Energy Survey collaboration
spreaderr_model_i_DES		Error on the stellarity indicator
wavg_mag_psf_i_DES	mag (AB)	Weighted-average of PSF magnitude in *i*-band
wavg_magerr_psf_i_DES	mag (AB)	Error on wavg_mag_psf_i_DES
p_{STAR}		ML probability of being a star (added in DR2)
p_{QSO}		ML probability of being a quasar (added in DR2)
$p_{GALAXY.}$		ML probability of being a galaxy (added in DR2)
CLASS		Object classification according tot he probabilities (added in DR2)
Imputation_flag		String indicating which magnitudes of the nine used in the ML ensemble are imputed (added in DR2)
Warning_flag		Integer which highlight if the objects is within the safe ranges or not and if it is an outlier or not (added in DR2)

TABLE 2: VEXAS-PSW

NAME	**UNIT**	**DESCRIPTION**
SOURCEID_VISTA		UID of the merged detection in VISTA
objID_PS		Unique ID PanSTARRS1 (PS)
SLAVEOBJID_WISE		The unique ID of the neighbour in calSource (=sourceID) for the 10arcsec match
ID_WISE		AllWISE ID for the 3 arcsec match
RA_PS	degrees	FK5 J2000.0 Right Ascension in PS
DEC_PS	degrees	FK5 J2000.0 Declination in PS
gpetMag_PS	mag (AB)	Petrosian magnitude in the *g*-band
gpetMagErr_PS	mag (AB)	Error on the Petrosian magnitude in the *g*-band
rpetMag_PS	mag (AB)	Petrosian magnitude in the r-band
rpetMagErr_PS	mag (AB)	Error on the Petrosian magnitude in the *r*-band
ipetMag_PS	mag (AB)	Petrosian magnitude in the *i*-band
ipetMagErr_PS	mag (AB)	Error on the Petrosian magnitude in the *i*-band
zpetMag_PS	mag (AB)	Petrosian magnitude in the *z*-band
zpetMagErr_PS	mag (AB)	Error on the Petrosian magnitude in the *z*-band
ypetMag_PS	mag (AB)	Petrosian magnitude in the *y*-band
ypetMagErr_PS	mag (AB)	Error on the Petrosian magnitude in the *y*-band
iPSFMag_PS	mag (AB)	PSF magnitude in the *i*-band
iPSFMagErr_PS	mag (AB)	Error on the PSF magnitude in *i*-band

ipsfLikelihood_PS		Likelihood that the *i*-band stack detection is best fit by a PSF (stellarity indicator)
p_{STAR}		ML probability of being a star (added in DR2)
p_{QSO}		ML probability of being a quasar (added in DR2)
$p_{GALAXY.}$		ML probability of being a galaxy (added in DR2)
CLASS		Object classification according to the probabilities (added in DR2)
Imputation_flag		String indicating which magnitudes of the nine used in the ML ensemble are imputed (added in DR2)
Warning_flag		Integer which highlights if the objects is within the safe ranges or not and i fit is an outlier or not (added in DR2)

TABLE 3: VEXAS-SMW

NAME	UNIT	DESCRIPTION
SOURCEID_VISTA		UID of the merged detection in VISTA
Object_ID_SM		Unique ID in Sky Mapper (SM)
SLAVEOBJID_WISE		The unique ID of the neighbour in calSource (=sourceID) for the 10arcsec match
ID_WISE		AllWISE ID for the 3 arcsec match
RA_SM	degrees	FK5 J2000.0 Right Ascension in PS
DEC_SM	degrees	FK5 J2000.0 Declination in PS
u_petro_SM	mag (AB)	Petrosian magnitude in the *u*-band
E_u_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *u*-band
v_petro_SM	mag (AB)	Petrosian magnitude in the *v*-band
E_v_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *v*-band
g_petro_SM	mag (AB)	Petrosian magnitude in the *g*-band
E_g_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *g*-band
r_petro_SM	mag (AB)	Petrosian magnitude in the *r*-band
E_r_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *r*-band
i_petro_SM	mag (AB)	Petrosian magnitude in the *i*-band
E_i_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *i*-band
z_petro_SM	mag (AB)	Petrosian magnitude in the *z*-band
E_z_petro_SM	mag (AB)	Error on the Petrosian magnitude in the *z*-band
Class_star_SM		Stellarity indicator from SkyMapper
ebmv_sfd_SM		The galactic extinction value E(B-V) measured from the Schlegel maps
p_{STAR}		ML probability of being a star according to the meta-model (added in DR2)
p_{QSO}		ML probability of being a quasar according to the meta-model (added in DR2)
$p_{GALAXY.}$		ML probability of being a galaxy according to the meta-model (added in DR2)
p_{STAR_25}		ML probability of being a star according to

		model 25 (no imputation) (added in DR2)
p_{QSO}_25		ML probability of being a quasar according to model 25 (no imputation) (added in DR2)
$p_{GALAXY.}\_25$		ML probability of being a galaxy according to model 25 (no imputation) (added in DR2)
CLASS		Object classification according to the probabilities of the meta-model (added in DR2)
CLASS_25		Object classification according to the probabilities of model 25 (added in DR2)
Imputation_flag		String indicating which magnitudes of the nine used in the ML ensemble are imputed (only valid for the meta-model) (added in DR2)
Warning_flag		Integer which highlights if the objects is within the safe ranges or not and i fit is an outlier or not (added in DR2)

TABLE 4: VEXAS-SPEC-GOOD

NAME	UNIT	DESCRIPTION
SOURCEID_VISTA		UID of the merged detection in VISTA
RA_spec	degrees	FK5 J2000.0 Right Ascension in the spectroscopic survey
DEC_spec	degrees	FK5 J2000.0 Declination in the spectroscopic survey
Z_spec		Redshift measurement from the spectroscopic survey
CLASS_Spec		Object classification
SPEC_survey		Survey in which this object is identified
SPECID_ID		Spectroscopic ID
Separation-VISTA-SPEC	arcsec	Distance between matched objects in VISTA and in the spectroscopic survey

We note that the number of columns in the spectroscopic table has been reduced with respect to DR1, where we also include some SDSS columns relative to stellar quantities measured from ELODIE high resolution spectra. Since we do not have this information for the other spectroscopic samples used here in DR2, we do not report the columns any longer. We note to the interested readers that this information can be obtained cross matching the VEXAS-SPEC table with the SDSS DR16 catalog available from here: https://www.sdss.org/dr16/spectro/spectro_access/

Acknowledgements

Users of VEXAS data should cite *Spiniello & Agnello, 2019, A&A, 630, A146* (doi:10.1051/0004-6361/201936311 full ADS link: https://ui.adsabs.harvard.edu/abs/2019A%26A...630A.146S, A&A link: http://www.aanda.org/10.1051/0004-6361/201936311/pdf)
Users of data from this release should cite Khramtsov et al. 2021, arXiv:2103.09257

This research has made use of the services of the ESO Science Archive Facility.

Bibliography

Abbott T. M. C., et al., 2018, ApJS, 239, 18; Ahumada R., et al. 2019, arXiv, arXiv:1912.02905; Baldry I. K, et al. 2018, MNRAS, 474, 3875 ; Chambers K.C., et al., 2016, arXiv:1612.05560; Childress M. J et al., 2017, MNRAS, 472, 273; Croom, S. M., et al. 2003, MNRAS, 349, 1397; Drinkwater M. J., et al., 2010, MNRAS, 401, 1429; Emerson J., et al. , 2006, Msngr, 126, 41; Jones D. H., et al., 2009, MNRAS, 399, 683; Khramtsov, V., et al. 2019, A&A, 632, A56; Khramtsov, V, et al. 2020, A&A, 644, A69; McMahon R. G. and the VHS Collaboration, 2013, Msngr, 154, 35; Spiniello & Agnello, 2019, A&A, 630, A146; Sutherland W., 2012, sngi.conf, 40; Taylor M. B., 2005, ASPC, 347; Wolf C., et al., 2018, PASA, 35, e024; Wright E.L., et al., 2010, AJ, 140, 1868