Publications of the Astronomical Society of Australia (2024), **41**, e110, 11 pages
doi:10.1017/pasa.2024.118

Research Article

Selecting variable sources with median colours using a self-organising map

Thomas Venville[1,2], Peter L. Capak[3], Andreas L. Faisst[4], Bomee Lee[4,5], Karun G. Thanjavur[6], and Chris Flynn[2,7]

[1]Research School of Astronomy and Astrophysics, Australian National University, Canberra 2611, A.C.T., Australia, [2]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, Victoria, 3122, Australia, [3]Cosmic Dawn Center (DAWN), Denmark, [4]Caltech/IPAC, 1200 E. California Blvd. Pasadena, CA 91125, USA, [5]Korea Astronomy and Space Science Institute, 776 Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Korea, [6]Department of Physics and Astronomy, University of Victoria, 3800 Finnerty Road, Victoria, BC V8P 5C2, Canada and [7]ARC Centre of Excellence for Gravitational Wave Discovery (OzGrav), Mail H29, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia

Abstract

A key objective for upcoming surveys, and when re-analysing archival data, is the identification of variable stellar sources. However, the selection of these sources is often complicated by the unavailability of light curve data. Utilising a self-organising map (SOM), we demonstrate the selection of diverse variable source types from a catalogue of variable and non-variable SDSS Stripe 82 sources whilst employing only the median $u-g$, $g-r$, $r-i$, and $i-z$ photometric colours for each source as input, without using source magnitudes. This includes the separation of main sequence variable stars that are otherwise degenerate with non-variable sources $(u-g,g-r)$ and $(r-i,i-z)$ colour-spaces. We separate variable sources on the main sequence from all other variable and non-variable sources with a purity of 80.0% and completeness of 25.1%, figures which can be modified depending on the application. We also explore the varying ability of the same method to simultaneously select other types of variable sources from the heterogeneous sample, including variable quasars and RR-Lyrae stars. The demonstrated ability of this method to select variable main sequence stars in colour-space holds promise for application in future survey reduction pipelines and for the analysis of archival data, where light curves may not be available or may be prohibitively expensive to obtain.

Keywords: Methods: data analysis; techniques: photometric; surveys

(Received 12 June 2024; revised 6 November 2024; accepted 18 November 2024)

1. Introduction

With the advent of several large astronomical surveys in the near future, for example, Legacy Survey of Space and Time (LSST, Ivezić et al. 2019) and the Nancy Grace Roman Telescope (formerly WFIRST) time-domain surveys (Foley et al. 2019), the volume and dimensionality of data produced by astronomical facilities is scheduled to increase rapidly. Machine learning (ML) and artificial intelligence algorithms, already extensively used to detect and classify variable sources (Fluke & Jacobs 2020; Baron 2019; Soares-Santos et al. 2017; Foley et al. 2019; Rhodes et al. 2017; Bailey et al. 2022; Foley et al. 2019), will be essential to achieve similar science goals for these upcoming survey programmes (Ivezić et al. 2019; Foley et al. 2019). Searches for variable and transient searches in archival data using ML also hold significant potential for scientific return (e.g. Pérez-Díaz et al. 2024; Webbe & Young 2023).

A self-organising map (SOM) (Kohonen 1982, 1990) is an unsupervised machine learning algorithm which performs a non-linear dimensionality reduction of an N-dimensional input data set. The SOM alogrithm represents the data in arbitrary two dimensional space, which preserves the distance between points in the corresponding high dimensional space, forming a two dimensional representation where clusters of similar objects are preserved from the input data space. This facilitates direct mapping of the high dimensional data to the two dimensional space. As the SOM preserves the topology of the input data set on small scales, this two dimensional representation of the SOM space is subsequently useful for identifying groups of similar objects and the relationship between all objects in the high dimensional space. For a more extensive overview of how the SOM algorithm functions, see Kohonen (1982, 1990), Masters et al. (2015). Example prior applications of SOM to astronomy include Faisst et al. (2019), who utilise the SOM to find variable AGN using 8 non-parametric variability indicators, and Masters et al. (2015), who utilise the SOM to predict the areas of Euclid photometric colour-space that lack spectroscopic galaxy redshifts for future Euclid galaxy photometric redshift calibration. Unsupervised algorithms such as the SOM have special relevance for many next generation facilities as they are ideally suited for discovering ‘unknown unknown’ transient/variable sources (Baron 2019) and are thus ideal tools to leverage the unprecedented capabilities of upcoming surveys to detect previously unseen classes of objects Fluke & Jacobs (2020).

This proof-of-concept study demonstrates the use of SOMs for detecting and classifying time varying sources in $(u-g, g-r, r-i, i-z)$ colour-space, through extending the method of Faisst et al. (2019) to identify optically variable main sequence sources

Corresponding author: Thomas Venville; Email: thomas.venville@anu.edu.au

Cite this article: Venville T, Capak PL., Faisst AL., Lee B, Thanjavur KG and Flynn C. (2024) Selecting variable sources with median colours using a self-organising map. *Publications of the Astronomical Society of Australia* **41**, e110, 1–11. https://doi.org/10.1017/pasa.2024.118

in Sloan Digital Sky Survey. To the knowledge of the authors, the SOM has not been utilised to detect or classify variable sources in colour-space. Detecting variable sources in photometric colour-space could use significantly less data (e.g. just four median colour values per object) than classification utilising spectra or source light curves, which is advantageous when processing future planned photometric surveys. Our method also has the significant potential to facilitate the selection of variable sources where light curve information is not available or prohibitively expensive to obtain, for example, during the analysis of archival colour data.

The paper layout is as follows. Section 2 describes the sample of variable and non-variable sources used in this study and their distribution in $(u-g, g-r, r-i, i-z)$ colour-space. Following Sesar et al. (2007), this section also partitions these sources into six regions of a dominant source type, based on their position in $(u-g, g-r)$ colour-space. In Section 3, we demonstrate the selection of the variable sources occupying Region V of the $(u-g, g-r, r-i, i-z)$ colour-space (75% of all variable sources). These sources are overwhelmingly variable main sequence sources and are degenerate with non-variable main sequence sources in various 2D projections of the $(u-g, g-r, r-i, i-z)$ colour-space, for example, $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces. However, we illustrate these variable main sequence sources reside in a distinct array of cells on the four-colour SOM representation, and we utilise this clustering to separate a a sample of these variable sources with a purity of 80.0% and completeness of 25.1%. These purity and completeness values can be modified depending on application. Section 4 repeats this methodology to briefly explore the ability of the SOM to separate and classify variable sources in other regions of colour-space. Lastly, we summarise this study in Section 5.

2. Creation of a heterogeneous SDSS Stripe 82 source catalogue

This study utilises a heterogeneous sample of variable and non-variable objects from the SDSS Stripe 82 region (Ivezić et al. 2007). We start with version 2.6 of the SDSS Stripe 82 standard star catalogue (Ivezić et al. 2007), containing 1006849 sources with an assigned SDSS classification of ‘STAR’ (i.e. unresolved point sources, including stars and quasars). These sources were classified as non-variable in Ivezić et al. (2007) utilising the χ^2 value computed from the source light curves; specifically, these sources had a χ^2 value per degree of freedom in each passband *ugriz* of less than 3. This catalogue was utilised, as opposed to the revised standard star catalogue presented in Thanjavur et al. (2021), for consistency with the employed SDSS Stripe 82 variable source colour catalogue (described later in this section).

Following Ivezić et al. (2007), we first removed sources that had not been observed for more than 4 epochs in the u and z-bands to eliminate objects with unreliable u-band and z-band photometry. Secondly, again following Ivezić et al. (2007), sources that did not satisfy the criterion $\sigma\sqrt{N} < 0.03$ in the g, r, and i-bands were rejected, where σ is the standard deviation of the observed photometric magnitudes in the given band, for which N observations were taken. This avoided biased photometry in these passbands. This resulted in a final sample of 425 546 standard star sources. These standard star sources were observed for an average of 9 observations in each of the g, r, and i bands. The median SDSS colours $u-g$, $g-r$, $r-i$, and $i-z$ were computed for each standard star source by subtracting the relevant median *ugriz* photometric magnitudes provided in version 2.6 of the SDSS Stripe 82 Standard Star catalogue (Ivezić et al. 2007). We compute the median colour by subtracting median photometric magnitudes for each source, given the varying number of observations in each passband prevents computation of the median colour as the median of colours computed for individual observations. However, this computation methodology does not result in incorrect median colours for the standard star sources, as subtracting the median photometric magnitudes of two passbands, for sources of constant magnitude, is equal to computing the median of subtracted individual photometric measurements from the two passbands.

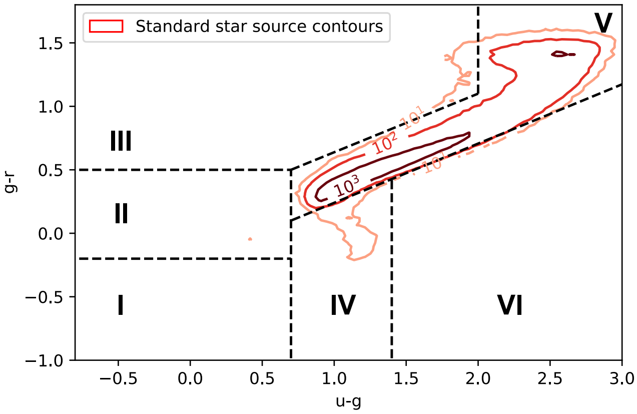

Figure 1. The $(u-g,g-r)$ colour-space distribution of the 425546 SDSS Stripe 82 standard star sources used in this study (as detailed in Section 2). Also shown for comparison with Fig. 2 are the six colour-space regions detailed for *variable* Stripe 82 sources in Sesar et al. (2007). 418 772 main sequence standard star sources are located in Region V, which contains 98.4% of all standard star sources. The number and type of sources in the other regions is described in Section 2.

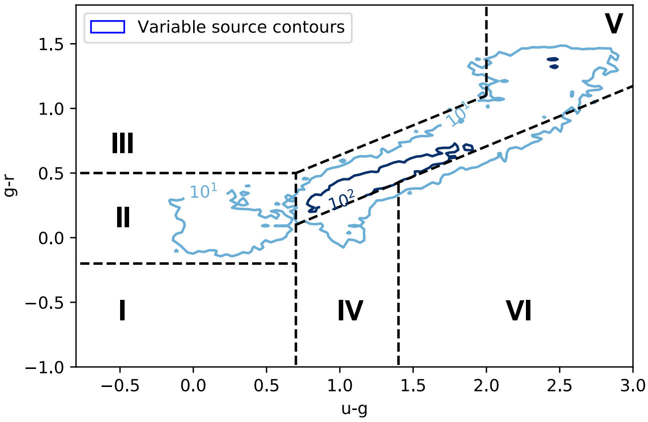

Figure 2. The $(u-g,g-r)$ colour-space of the 67 507 variable SDSS Stripe 82 sources used in the sample used in this study (as detailed in Section 2), with the six colour-space regions detailed for Stripe 82 variable sources in Sesar et al. (2007) overlaid. 75% of all variable sources are located in Region V, which is overwhelmingly dominated by main sequence sources. Region II contains 8735 sources (12.9% of the total) and is dominated by low-redshift quasars. Region IV contains 2725 sources (4% of the total) and is dominated by RR-lyrae stars.

The $(u-g,g-r)$ colour-space of these sources is detailed in Fig. 1, where the contours indicate the number of sources in each pixel. Also displayed on this plot, for the purpose of comparison with Fig. 2, are the six colour-space regions detailed for *variable* Stripe 82 sources in Sesar et al. (2007). These regions utilise colour cuts to divide the $(u-g,g-r)$ colour-space of SDSS Stripe 82

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

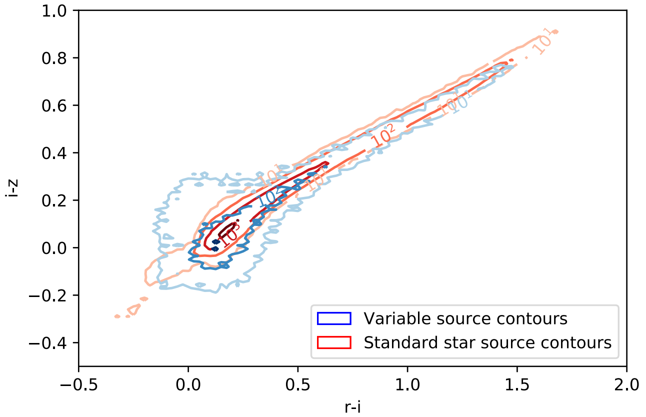

Figure 3. The $(r-i,i-z)$ colour-space of the Stripe 82 variable and standard star sources used in this study (detailed in Section 2). It is clear that the vast majority of the variable sources (in blue contours) are entirely degenerate with the standard star sources in this colour-space, inhibiting straightforward colour-space selection. Degenerate sources in this $(r-i,i-z)$ colour-space include the variable and standard star sources in the main sequence 'Region V'; this is further illustrated in Fig. 5.

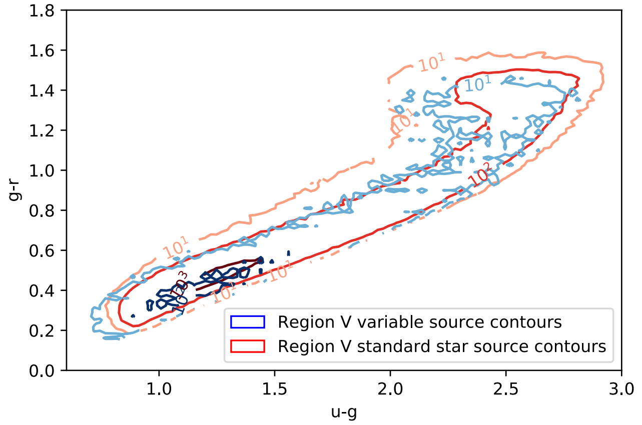

Figure 4. The $(u-g,g-r)$ colour-space contours of all variable (blue) and standard star (red) sources located in Region V of $(u-g,g-r)$ colour-space (see Figs. 2 and 1, respectively). It is clear that the vast majority of variable sources from this region are within in the red contours, illustrating the degeneracy between these variable and standard star sources in $(u-g,g-r)$ colour-space. As depicted in Fig. 5, these Region V sources are also largely degenerate in $(r-i,i-z)$ colour-space.

variable sources into regions dominated by a given kind of variable source (Sesar et al. 2007). It is important to note that these regions are designed to separate distinct types of variable sources and thus are indicated only for comparison with Fig. 2. The (red) $(r-i,i-z)$ colour-space contours of all 425 546 standard star sources are detailed in Fig. 3.

We produce a heterogeneous sample of sources by combining the standard stars sources detailed above with all 67 507 sources in version 1.1 of the SDSS S82 variable source catalogue (Ivezić et al. 2007). Like the standard star sources described above, these sources again have a STAR classification and consist of unresolved point sources (both stars and quasars). However, in contrast to the standard star sources described above, these variable sources have a χ^2 per degree of freedom of >3 in one (or more) of the passbands g, r, and i (Ivezić et al. 2007). These variable sources were observed for an average of 36, 36, and 37 observation epochs in each of the g, r, and i bands. The combined sample contained some 493 053 sources, with 13.7% of these sources variable. The combined sample of sources were observed for an average of 23 observation epochs each of the g, r, and i bands.

The final sample of 67 507 variable sources are detailed in $(u-g,g-r)$ colour-space in Fig. 2 and $(r-i,i-z)$ colour-space in Fig. 3. We illustrate the diverse types of variable sources in our sample by partitioning this sample into six regions of $(u-g,g-r)$ colour-space, as detailed for Stripe 82 sources in Sesar et al. (2007). The partitions are displayed in Fig. 2. Of particular interest to this study is Region V, which contains 50 157 variable and 418 772 non-variable standard star sources. This region contains 74.2% (98.4%) of all variable (standard star) sources and 10.7% of all sources in this region are variable. We have verified that these standard star sources are almost all main sequence stellar sources utilising parallax measurements from the third Gaia data release (DR3, Gaia Collaboration et al. 2023), the SDSS $(u-g,g-r)$ colour-space and SDSS colour magnitude diagrams. Importantly, as detailed in Figs. 4 and 5, these Region V variable and standard star sources are almost entirely degenerate in both $(u-g,g-r)$ and $(r-i,i-z)$ colour-space, inhibiting selection of either variable or standard star sources in this region through a straightforward partition of the heterogeneous sample colour-space. However, we show later in Section 3 that it is feasible to separate these variable and standard star sources using a SOM trained on the $(u-g,g-r,r-i,i-z)$ coordinates of the complete heterogeneous sample.

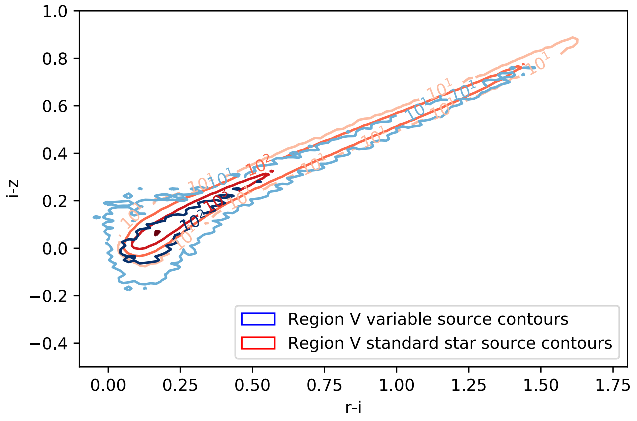

Figure 5. The $(r-i,i-z)$ colour-space contours of all variable (blue) and standard star (red) sources located in Region V of $(u-g,g-r)$ colour-space (see Figs. 2 and 1, respectively). Note that the outermost variable source density contour contains few variable sources. Accordingly, the vast majority of the variable Region V sources are within in the red contours and entirely degenerate in $(r-i,i-z)$ colour-space with the Region V standard star sources.

Of the remaining regions, Region I contains 198 variable sources, predominantly white dwarf stars (Sesar et al. 2007), in addition to 332 standard star sources which also are likely white dwarf stars. Region II is dominated by variable low-redshift quasar sources (Sesar et al. 2007), with 6 307 of the 8 735 variable sources in this region spectroscopically confirmed as variable quasars. Region II contains 594 standard star sources. Matching these to Gaia DR3 (Gaia Collaboration et al. 2023), around half have significant parallaxes (>1 mas) and can be reliably identified utilising colour magnitude diagrams as either single white dwarfs or dM/WD pairs. The remaining sources are consistent with a negligible parallax to within the measurement error and are likely quasars. The 1 417 variable sources in Region III are predominantly dM/WD pairs (Sesar et al. 2007), residing near

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

the edge of the main sequence, and share this region with 2 112 (predominantly main sequence) standard star sources. Region IV contains 2 725 variable sources, predominantly RR-Lyrae stars (Sesar et al. 2007) and variable stars at the edge of the main sequence, and 2 762 standard star sources, including sources on the the main sequence, non-variable quasars, and horizonal branch stars. Lastly, Region VI contains 4 202 variable sources, predominantly high-redshift QSO's (Sesar et al. 2007), and 974 standard star sources which are predominantly are stellar sources on the edge of the main sequence. The utility of selecting Region I, II, II, IV, and VI variable sources using the same SOM trained with the $(u-g, g-r, r-i, i-z)$ coordinates of the complete heterogeneous sample is discussed in Section 4.

3. Separating variable main sequence stars utilising the colour-space SOM

A frequent use-case for supervised and unsupervised ML algorithms is to select a certain type of variable source from a heterogeneous dataset (Fluke & Jacobs 2020). In this section, we illustrate the use of a SOM trained with the $(u-g, g-r, r-i, i-z)$ colours of the heterogeneous sample detailed in Section 2 to separate variable main sequence sources from Region V of $(u-g, g-r)$ colour-space (as defined in Section 2) from both all standard star sources and all variable sources occupying different regions of $(u-g, g-r)$ colour-space. As discussed in Section 2, these variable Region V sources are degenerate with Region V standard star sources both $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces, inhibiting straightforward selection using a simple partition of colour-space.

We begin by initialising a SOM with the PYMVPA package (Hanke et al. 2009), and training it with the complete $(u-g, g-r, r-i, i-z)$ colour-space of the heterogeneous source sample detailed in Section 2. It is important to note this sample was not divided into two distinct training and test sets, as this is not required for this type of unsupervised machine learning algorithm. The optimum size of this SOM, namely (33,106) cells, was calculated based on eigenvectors and eigenvalues of the four-dimensional colour-space, utilising an adapted variant of the `calculate_map_size` method from the SOMPY package. Following Faisst et al. (2019), the SOM was trained over 200 iterations, utilising an initial learning rate of $L_0 = 0.05$ (which decreases with each iteration i). The learning radius σ_i, which also decreases with each iteration i, was initially set to the longest dimension of the specified 2D map size. As noted in Faisst et al. (2019), each SOM is randomly initialised, with different initialisations of the SOM potentially affecting results. We have verified that the quantitative results detailed in this study show variance of $<1\%$ between different SOM initialisations.

The binning of all 493 053 sources in the heterogeneous colour-space sample onto the 2D representation produced by the trained colour-space SOM is depicted in Fig. 6(a). It is important to note that the axes of this representation are aligned with the eigenvectors of the four-dimensional $(u-g, g-r, r-i, i-z)$ colour-space to maximise the retention of high dimensional structure. Thus, these axes are unique to the training sample being utilised.

As seen in Fig. 6(a), though the sample as a whole is distributed fairly evenly across the SOM representation, there is clear structure in the four-dimensional colour-space. The distribution of all 67 507 variable sources in the heterogeneous sample across this representation are depicted in Fig. 6(b), whilst the distribution of the 425 546 standard star sources are shown in Fig. 6(c). It is clear through comparison of these figures that the variable sources and standard star sources inhabit different areas of the SOM representation; to emphasise this, we display the purity of variable sources (from any region) $\mathcal{P}$ (i.e. the total number of variable sources divided by the total number of sources) in each SOM cell in Fig. 6(d). Note that this is not the purity of variable Region V sources detailed below in Equation (1).

Fig. 7(a) shows the distribution of variable Region V sources upon this SOM representation. It is evident through comparison of this figure with Fig. 6(c) that the cells containing variable Region V sources contain very few standard star sources – crucially, *including* standard star sources otherwise degenerate in $(u-g, g-r)$ and $(r-i, i-z)$ colour-space (see Figs. 4 and 5). Fig. 7(a) illustrates the purity of *variable Region V sources* ($\mathcal{P}_{\rm V}$) in each SOM cell (i.e. the fraction of sources in the given cell that are variable Region V sources). In cells with high $\mathcal{P}_{\rm V}$ values, the SOM separates variable Region V sources from variable sources occupying other regions of $(u-g, g-r)$ colour-space *and* from all standard star sources, including those otherwise degenerate $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces.

To quantify the observed success of separating the variable Region V sources from *both* variable sources occupying other regions of the colour-space and all standard star sources, we follow the method of Faisst et al. (2019) and define the group of (not necessarily contiguous) cells where, in each cell, the purity of variable sources from Region V ($\mathcal{P}_{\rm V}$) exceeds a given value $\mathcal{P}_{\rm V,min}$. We then calculate the purity of variable Region V sources $\mathcal{P}_{\rm V}$

$$\mathcal{P}_{\rm V} = \frac{\rm TP}{\rm TP + FP} \tag{1}$$

and the completeness of variable Region V sources $\mathcal{R}_{\rm V}$

$$\mathcal{R}_{\rm V} = \frac{\rm TP}{\rm TP + FN} \tag{2}$$

for the group of cells cells utilising the number of true-negative (TN), false-positive (FP), false-negative (FN) and true-positive (TP) sources. In this case, the number of true-positive sources TP is the number of variable Region V sources within the defined group of cells. The number of false-positive sources FP is the number of sources within the group of cells that are not variable sources from Region V; specifically, all variable sources from other regions of $(u-g, g-r)$ colour-space and all standard star sources. The number of false-negative sources FN is the number of variable Region V sources outside of the group, whilst the number of remaining sources, namely the sources outside of the group that are either standard stars or variable sources from other regions of $(u-g, g-r)$ colour-space, is the number of true-negative sources TN. $T = {\rm TP + TN + FP + FN}$ is the total number of sources on the SOM, in this case the 493053 sources in the heterogeneous colour-space sample detailed in Section 2. Using these definitions, the overall purity $\mathcal{P}_{\rm V}$ and completeness $\mathcal{R}_{\rm V}$ of variable Region V sources, for each group of cells where $\mathcal{P}_{\rm V} > \mathcal{P}_{\rm V,min}$ in every cell, is displayed in Fig. 7(c). The $\mathcal{P}_{\rm V}$ and $\mathcal{R}_{\rm V}$ values calculated for a group of cells defined by a given $\mathcal{P}_{\rm V,min}$ value vary by $<1\%$ between different SOM initialisations, even though the location of the given group cells on the SOM representation does vary between initialisations due to the locally topological nature of each SOM mapping.

Fig. 7(c) indicates that, in the group of cells where $\mathcal{P}_{\rm V} > \mathcal{P}_{\rm V,min} = 60.0\%$ in each cell, $\mathcal{P}_{\rm V} = 80.2\%$ of all sources are variable Region V sources, whilst these cells contain $\mathcal{R}_{\rm V} = 25.1\%$ of all Region V variable sources (some 125 177 sources). This group of

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

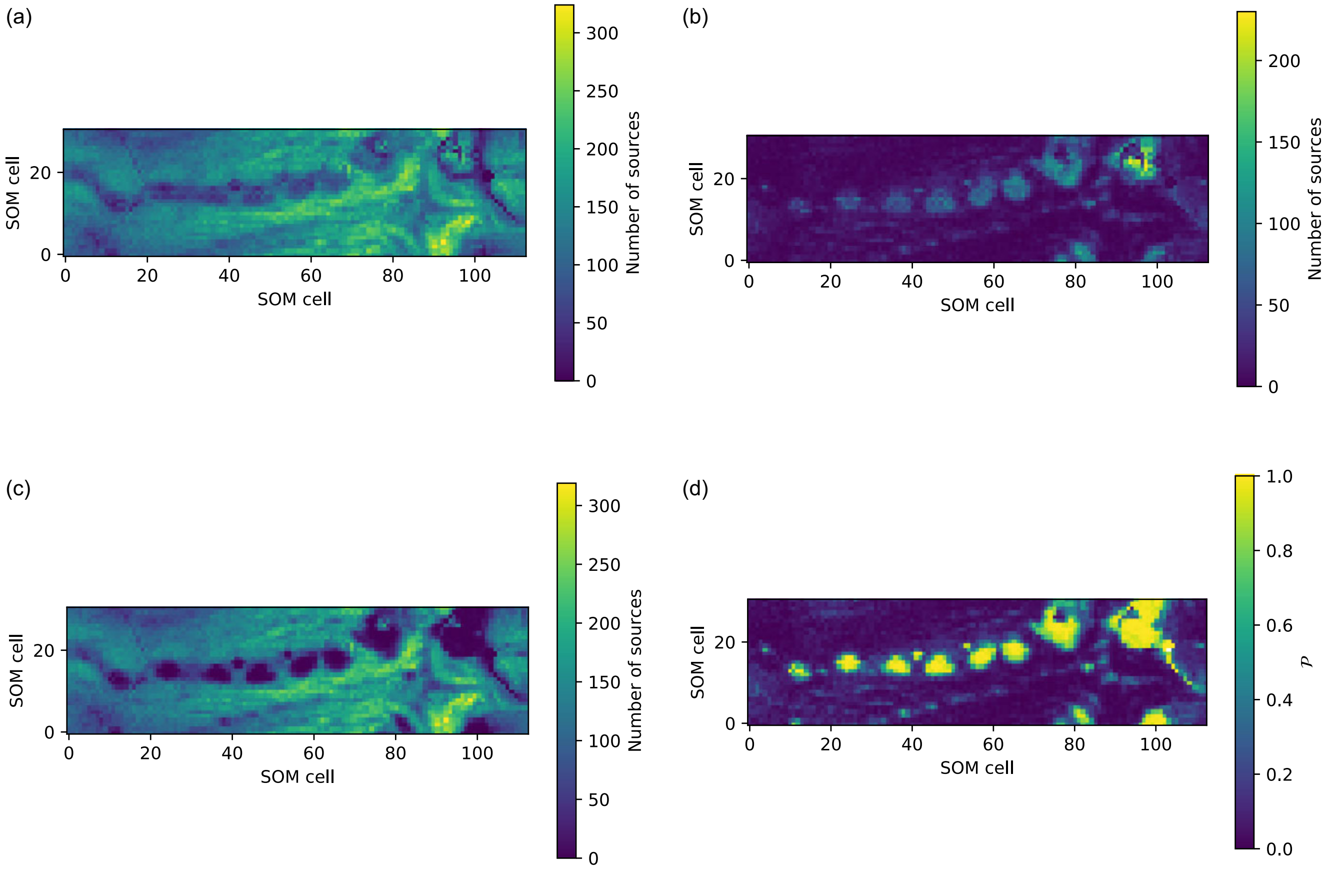

Figure 6. The two dimensional SOM representation of the complete four dimensional SDSS S82 colour-space sample. (a) displays the total number of sources in each cell of the SOM representation. This is equal to the sum of the number of variable sources (b) and the number of standard star sources (c) in each cell. It is evident that the variable and standard star sources generally inhabit different regions of the SOM representation. As discussed in Sections 3 and 4, this indicates that these four median photometric colours are sufficient for separating variable and standard star sources that are otherwise often degenerate in $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces. This is further emphasised by the the variable source purity $\mathcal{P}$ of each cell is displayed in (d).

cells is depicted on the SOM representation in Fig. 7(d) and is the largest group of cells where $\mathcal{P}_{\rm V} > 80.\%$. As is indicated in Fig. 7(c), the $(u-g, g-r, r-i, i-z)$ SOM can also be used to separate variable Region V sources with a variety of other $\mathcal{P}_{\rm V}$ and $\mathcal{R}_{\rm V}$ values depending on the use-case, including $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V}) = (48.5\%, 48.5\%)$ and $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V}) = (75.4\%, 29.1\%)$.

As aforementioned, only 10.7% of Region V sources are variable sources, and as detailed in Figs. 4 and 5, the Region V variable and standard star sources are almost entirely degenerate in $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces, inhibiting selection of either variable or standard star sources in this region through a straightforward partition of the $(u-g, g-r, r-i, i-z)$ colour-space. This degeneracy is further emphasised by an analysis of the Kohonen layers of the SOM, which indicates that, for a given SOM cell, there is no observed correlation between the median value of any source colour and the purity of variable Region V sources $\mathcal{P}_{\rm V}$ – that is, that no single colour is responsible for the observed separation of variable Region V sources on the SOM representation. Accordingly, the remarkable capacity of the SOM to separate variable Region V sources from all other variable and standard star sources using defined groups of cells with high $\mathcal{P}_{\rm V}$ values is *facilitated* by a multi-dimensional analysis of the four input dimensions.

3.1 *Investigating the separation of Region V sources utilising the four-dimensional colour-distance*

To investigate this observed separation of variable and standard star sources in Region V, we use the methodology of Covey et al. (2007) to quantify the distance of the variable and standard star sources from a defined stellar locus in four dimensional colour-space. Following Covey et al. (2007), we first define the median stellar locus in each colour k, as a function of $g-i$ bin (each of width 0.02 mag), as the median colour $X_k^{\rm locus}$ of the standard star sources in that $g-i$ bin. In each $g-i$ bin and each colour, we define the colour error $\sigma_{k,X}({\rm locus})$ as the quadrature subtraction of the median colour error from the standard deviation of colours about the median colour. Using these quantities, we can then define the four-dimensional colour distance (4DCD) for a given 'target' source from this locus as (Covey et al. 2007):

$$4{\rm DCD} = \sum_{0}^{3} \frac{\left(X_k^{\rm target} - X_k^{\rm locus}\right)}{\sigma_{k,X}^2({\rm locus}) + \sigma_{k,x}^2} \quad (3)$$

where, as per Covey et al. (2007), $X_k^{\rm target}$ is the median colour k of the target source and $\sigma_{k,x}$ is the colour error for the target source. In the case of standard star sources, we compute $\sigma_{k,x}$ as

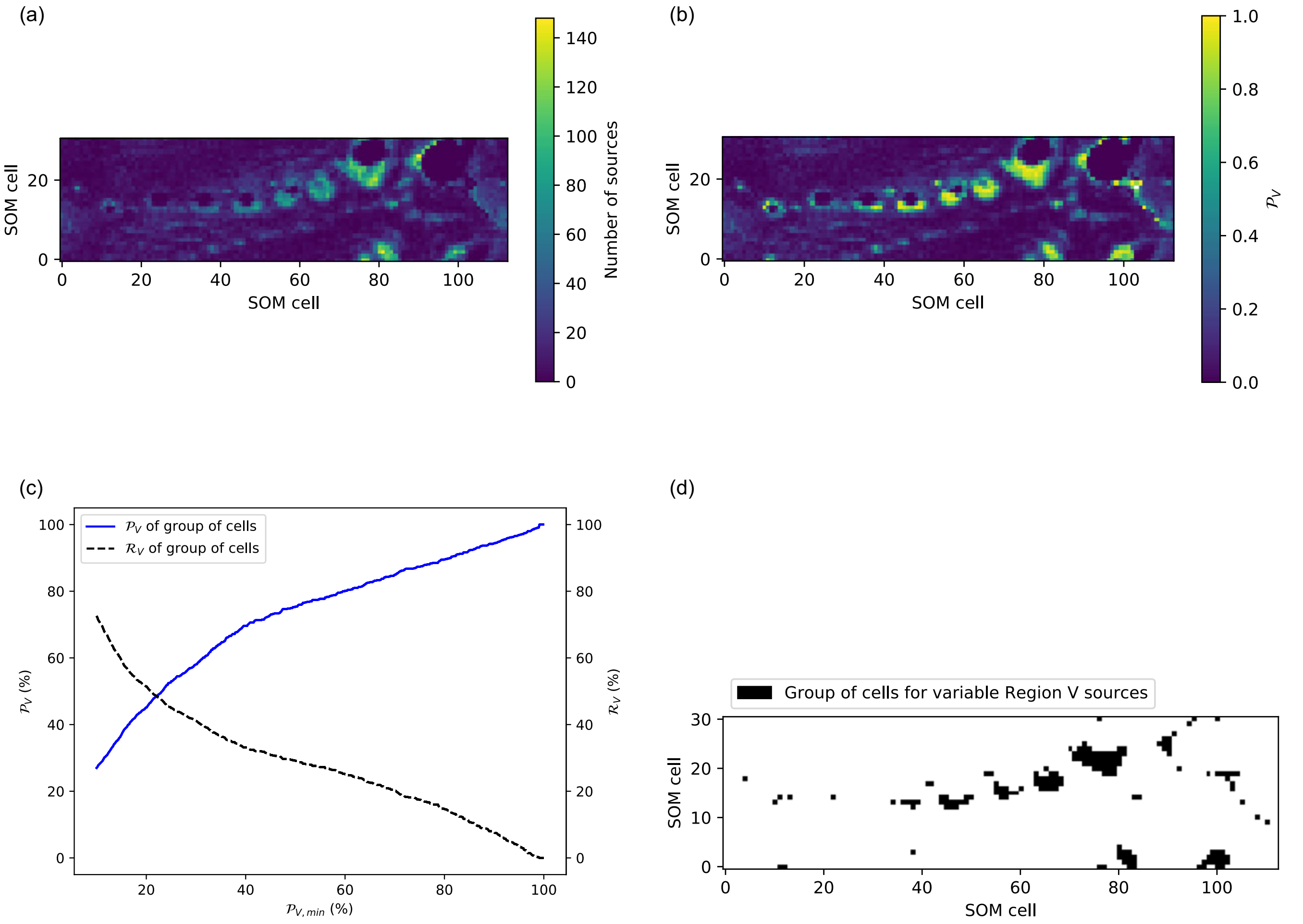

Figure 7. (a) The number and (b) the purity $\mathcal{P}_V$ of of variable Region V sources upon the SOM representation depicted in Fig. 5. The cells containing variable Region V sources contain primarily variable Region V sources with few variable sources from other regions or standard star sources – crucially, *including* standard star sources otherwise degenerate in $(u-g, g-r)$ and $(r-i, i-z)$ colour-space (see Figs. 4 and 5). (c) The overall purity $\mathcal{P}_V$ (in blue) and completeness $\mathcal{R}_V$ (in black dashes) of variable Region V sources in each group of cells defined by a given cell minimum variable Region V source purity, $\mathcal{P}_{V,min}$. The $(u-g, g-r, r-i, i-z)$ SOM can be used to separate variable Region V sources with a variety of $\mathcal{P}_V$ and $\mathcal{R}_V$ values depending on the use-case, including $(\mathcal{P}_V, \mathcal{R}_V) = (80.2\%, 25.1\%)$, $(\mathcal{P}_V, \mathcal{R}_V) = (48.5\%, 48.5\%)$ and $(\mathcal{P}_V, \mathcal{R}_V) = (75.4\%, 29.1\%)$. (d) The group of cells where the variable Region V source purity $\mathcal{P}_V$ of each cell exceeds $\mathcal{P}_{V,min} = 60.0\%$. $\mathcal{P}_V = 80.2\%$ of all sources in this group of cells are variable Region V sources. This group of cells contains $\mathcal{R}_V = 25.1\%$ of all Region V variable sources (some 125177 sources). This group of cells, discussed in detail in section 3, is the largest group where $\mathcal{P}_V > 80.0\%$.

the median single-observation colour error for the source by summing in quadrature the median single-observation photometric error of the passbands composing the colour. In the case of variable sources, we assume that the colour errors $\sigma_{k,x}$ for each source are equal to the median colour error of standard star sources with the same g-band magnitude.

The fraction of variable and standard star sources from Region V with a given 4DCD value is shown in Fig. 8. As is evident, the variable and standard star sources have different distributions of 4DCD values. The larger fraction of variable sources with high 4DCD values indicate that these variable sources, on average, are further from the median stellar locus than the standard star sources. Furthermore, it is evident in Fig. 8 that though both variable and standard star sources have low 4DCD values, very few standard star sources have a 4DCD value of above 15. Together, these differences in 4DCD value distributions emphasise the fact that the variable and standard star sources from Region V often occupy different regions of the four-dimensional colour-space, providing the basis for the successful separation and selection of these source populations using the SOM detailed in Section 3.

4. Separating variable and non-variable sources in other regions of colour-space

As detailed in Section 2, the $(u-g, g-r)$ colour-space of variable sources can be divided into several regions, with different predominant variable source types in each region. Following Section 3, which illustrated the successful selection of variable Region V (main sequence) sources from the heterogeneous colour-space sample, we now briefly explore the separation of variable sources from other regions of $(u-g, g-r)$ colour-space. Firstly, following the method of Section 3, we define, for each of Region R (either III, IV, V or VI), the largest group of cells where the overall purity $\mathcal{P}_R$ of variable sources from the region exceeds 80.0%. This purity was selected as it is suitable for creating samples of variable sources for use in other studies. For variable Region V sources,

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

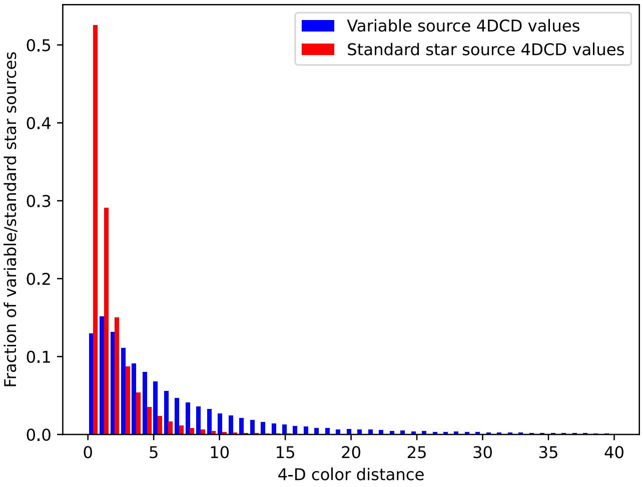

Figure 8. The fraction of variable (blue) and standard star (red) sources from Region V with a given four-dimensional colour-distance (4DCD) value. The larger fraction of variable sources with high 4DCD values indicate that these variable sources, on average, are further from the median stellar locus (defined using standard star source colours) than standard star sources, emphasising the fact that the variable and standard star sources from Region V often occupy different regions of the four-dimensional colour-space.

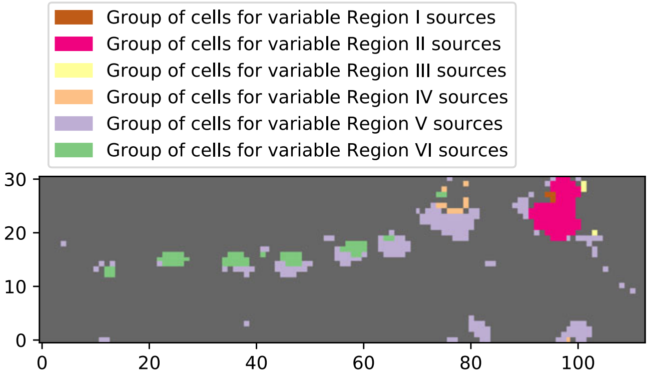

Figure 9. The groups of cells on the SOM containing variable sources from each of the six regions described in Section 2. As detailed in Section 4, the groups of cells depicted for Regions III, IV, V, and VI are the largest groups of cells where the overall purity of variable sources from the given region exceeds 80.0%. The group of cells containing variable sources from Region V is also discussed in detail in Section 3. The group of cells for Region I, as justified in Section 4.1, is defined with a Region I variable source purity of $\mathcal{P}_{\rm I} = 27.1\%$ and completeness of $\mathcal{R}_{\rm I} = 71.7\%$. This group consists of only three cells, located at (94,26), (95,27), and (94,27) on the depicted axes. The (magenta) group of cells displayed on Fig. 9 is dominated by variable Region II sources, with a variable Region II source purity of $\mathcal{P}_{\rm II} = 94.5\%$, and contains $\mathcal{R}_{\rm II} = 96.5\%$ of all variable Region II sources. Described in Section 4.2, it is the largest group where all cells predominately contain variable Region II sources. It is important to note that the groups of cells defined for each region do not overlap. The variable source purity and completeness for each group of cells is detailed in Table 1.

this is the same group of cells described in detail in Section 3 and depicted in Fig. 7(d). It is important to note that, for each defined group of cells, the purity $\mathcal{P}_{\rm R,min}$ of variable sources from Region R exceeds 50% for each cell. Thus, for each group of cells, all cells predominantly contain the variable sources from the given Region R. Accordingly, the groups of cells defined for each region do not overlap.

These four groups of cells are depicted on the SOM representation in Fig. 9, whilst Table 1 lists the (overall) purity $\mathcal{P}_{\rm R}$ and completeness $\mathcal{R}_{\rm R}$ of variable Region R sources in each group. Also displayed on Fig. 9 is the (brown) group of cells with a variable Region I source purity of $\mathcal{P}_{\rm I} = 27.1\%$ and completeness of $\mathcal{R}_{\rm I} = 71.7\%$. As discussed in Section 4.1, the ability to select variable sources from this region is limited. Lastly, the (magenta) group of cells displayed on Fig. 9 is dominated by variable Region II sources. This group, discussed in Section 4.2, has a variable Region II source purity of $\mathcal{P}_{\rm II} = 94.5\%$ and contains $\mathcal{R}_{\rm II} = 96.5\%$ of all variable Region II sources. This group is defined by $\mathcal{P}_{\rm II,min} = 50.1\%$ and is the largest group where all cells predominately contain variable Region II sources. As aforementioned, the groups of cells defined for each region do not overlap.

Table 1. The purity and completeness of variable sources from each region of $(u-g, g-r)$ colour-space in each corresponding group of cells depicted in Fig. 9. As explained in Section 4, the groups of cells for Regions III, IV, V, and VI are defined as the largest groups where the overall purity of variable sources from the given region exceeds 80.0%. As discussed in Section 4.1, the low variable source purity and completeness for the group of cells defined for variable Region I sources reflects the inability of the SOM to separate these variable sources. The group of cells defined for variable Region II sources the largest group where all cells predominately contain variable Region II sources and is discussed in Section 4.2.

Region	Cell group purity (%)	Cell group completeness (%)
I	$\mathcal{P}_{\rm I} = 27.1$	$\mathcal{R}_{\rm I} = 71.7$
II	$\mathcal{P}_{\rm II} = 94.5$	$\mathcal{R}_{\rm II} = 96.5$
III	$\mathcal{P}_{\rm III} = 81.8$	$\mathcal{R}_{\rm III} = 3.18$
IV	$\mathcal{P}_{\rm IV} = 81.4$	$\mathcal{R}_{\rm IV} = 20.3$
V	$\mathcal{P}_{\rm V} = 80.2$	$\mathcal{R}_{\rm V} = 25.1$
VI	$\mathcal{P}_{\rm VI} = 80.9$	$\mathcal{R}_{\rm VI} = 70.1$

4.1 Selecting variable sources from Region I of (u-g,g-r) colour-space

The highest purity of variable Region I sources in any cell on the SOM representation used in this study is 31.25%, and this cell contains only 3% of all variable Region I sources. It is not possible to define a group of cells with a higher purity of variable Region I sources. Accordingly, depicted on Fig. 9 (in brown) is the group of cells where the purity of Region I sources for the group $\mathcal{P}_{\rm I} = 27.1\%$. The completeness of variable Region I sources in the group is $\mathcal{R}_{\rm I} = 71.7\%$. This group consists of only three cells located at (94,26), (95,27) and (94,27) on the depicted axes. The low purity of Region I variable sources in any SOM cell indicates that these sources are not effectively separated from other sources.

4.2 Selecting variable sources from Region II of (u-g,g-r) colour-space

As mentioned in Section 2, the variable sources in Region II are dominated by low-redshift variable quasars, which have previously been successfully selected in colour-space without using a SOM (e.g. Stern et al. 2005; Assef et al. 2013; Peters et al. 2015). The (magenta) group dominated by variable Region II sources depicted in Fig. 9 has a variable Region II source purity of $\mathcal{P}_{\rm II} = 94.5\%$ and contains $\mathcal{R}_{\rm II} = 96.5\%$ of all variable Region II sources. This group is defined by $\mathcal{P}_{\rm II,min} = 50.1\%$ and is the largest group where all cells predominately contain variable Region II sources. Other groups can be defined by imposing a different minimum purity $\mathcal{P}_{\rm II,min}$ of variable Region II sources in the group cells, following the

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

method detailed for Region V in Section 3. $\mathcal{P}_{\mathrm{II}}$ and $\mathcal{R}_{\mathrm{II}}$ values possible for these other groups include $(\mathcal{P}_{\mathrm{II}}, \mathcal{R}_{\mathrm{II}}) = (88.5\%, 99.6\%)$, $(\mathcal{P}_{\mathrm{II}}, \mathcal{R}_{\mathrm{II}}) = (95.4\%, 95.4\%)$, and $(\mathcal{P}_{\mathrm{V}}, \mathcal{R}_{\mathrm{V}}) = (96.8\%, 92.3\%)$.

However, the apparently successful selection of these variable Region II sources needs to be carefully assessed. As discussed in Section 2, 93% of sources in this region are variable, with the 594 standard star sources in this region largely constituted of white dwarf and extra-galactic sources, that is, non-variable quasars. Further analysis of the SOM representation indicates that, of the 82 SOM cells containing these non-variable standard star sources, some 65 of them (containing 176 standard star sources) are within the same magenta group of cells dominated by variable Region II sources discussed in this section. Accordingly, it is unclear whether the SOM is successfully separating the variable and standard star sources in this region, which would otherwise be indicated by the selection of variable Region II sources with a much higher purity $\mathcal{P}_{\mathrm{II}}$ than the fraction of variable sources in Region II, namely $\mathcal{P}_{\mathrm{II}} > 93\%$. Based on the attained results, it is also unclear if the SOM will be successful in separating the variable and non-variable sources from Region II for a sample where the fraction of variable sources in Region II is lower. We leave the detailed assessment of this to future work utilising a different source sample.

4.3 Selecting variable sources from Region III of (u-g,g-r) colour-space

The 1 417 variable and 2 112 standard star sources in Region III of $(u-g, g-r)$ colour-space are dominated by sources near the main sequence, and these sources are largely degenerate in both $(u-g, g-r)$ and $(r-i, i-z)$ colour-space. The SOM does separate the variable sources in this region from other standard star and variable sources. However, the first group of cells where the purity of variable Region III sources $\mathcal{P}_{\mathrm{III}} = 81.8\%$ exceeds 80% contains only $\mathcal{R}_{\mathrm{III}} = 3.18\%$ of all variable Region III sources. Accordingly, the utility of selecting these variable sources in $(u-g, g-r, r-i, i-z)$ colour-space with the SOM is severely limited.

4.4 Selecting variable sources from Region IV of (u-g,g-r) colour-space

Region IV contains 2 725 variable sources, predominantly RR-Lyrae stars and variable stars on the edge of the main sequence, and 2 672 standard star sources. As detailed in Table 1, the SOM is able to separate the variable sources in this region into a group of cells with a variable Region IV source purity of $\mathcal{P}_{\mathrm{IV}} = 81.4\%$ and a completeness of $\mathcal{R}_{\mathrm{IV}} = 20.3\%$. The ability of the SOM to separate these variable sources is not surprising, given RR-Lyrae stars occupy a unique locus of colour-space (Ivezić et al. 2005; Sesar et al. 2007) and have previously been selected using the same SDSS I photometric colours with an efficiency (purity) of 60% and completeness of 28% (Ivezić et al. 2005). However, it is important to note that the variable and standard star sources in this region will predominantly be different types of sources, with the standard star source types known to inhabit this region including main sequence stars, horizontal branch stars and non-variable quasars (e.g. Ivezić et al. 2005). Accordingly, the variable Region III sources being separated by the SOM due to the intrinsically distinct colours of RR-Lyrae stars. This is in contrast to the separation of main sequence variable and standard star sources from Region V detailed in Section 3, where the SOM separation reflects differing variability amongst sources of the *same type* that are degenerate in both $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces.

4.5 Selecting variable sources from Region VI of (u-g,g-r) colour-space

Region VI of $(u-g, g-r)$ colour-space contains 4 202 variable sources (dominated by high-redshift quasars, according to Sesar et al. 2007) and 974 standard star sources predominantly on the edge of the main sequence; accordingly, 81.2% of all sources in this region are variable sources. The SOM does separate the variable sources in this region from both variable sources occupying other regions of $(u-g, g-r)$ colour-space and all standard star sources. The group of cells depicted on Fig. 9 has a variable Region VI source purity of $\mathcal{P}_{\mathrm{VI}} = 80.9\%$ and contains $\mathcal{R}_{\mathrm{VI}} = 70.1\%$ of all variable Region VI sources. Other groups can be formulated, following the methodology of Section 3, can be defined with $(\mathcal{P}_{\mathrm{VI}}, \mathcal{R}_{\mathrm{VI}}) = (95.3\%, 46.8\%)$ and $(\mathcal{P}_{\mathrm{VI}}, \mathcal{R}_{\mathrm{VI}}) = (99.1\%, 23.4\%)$. This good separation is expected, given a large number of these variable sources are not degenerate in $(u-g, g-r)$ colour-space with standard star sources (as illustrated through a comparison of Figs. 1 and 2). It is also important to recognise that, given the variable and standard star sources in this region (as per Region IV) are often different types of sources, this separation is again not reflective of differing variability amongst sources of the same type. Rather, it reflects the intrinsically different colours of the variable and standard star sources inhabiting this region of colour-space. This is in contrast to the separation of main sequence variable and standard star sources detailed in Section 3.

4.6 Analysing the four-dimensional colour-space position of sources in each group of cells

The clustering of variable sources from Regions II to VI into specific groups of cells, as detailed in Section 4, implies that the variable and standard sources in each group of cells occupy distinct regions of the heterogeneous sample's four-dimensional median colour-space. These distinct regions are each dominated by variable sources. To explore the four-dimensional median colour-space distribution of sources in these groups of cells, the $(u-g, g-r)$ and $(r-i, i-z)$ colour-space distributions of the variable sources within each group of cells detailed in Section 4 are, respectively, shown in Fig. 10(a) and (b). Similarly, the $(u-g, g-r)$ and $(r-i, i-z)$ colour-space distributions of the standard star sources within each group of cells detailed in Section 4 are, respectively, shown in Fig. 10(a) and (d). It is evident that the $(u-g, g-r)$ and $(r-i, i-z)$ colour-space distributions of variable and standard star sources in each group of cells are markedly different, with the exception of sources from the group of cells containing the highest purity of variable Region I sources. This again illustrates that the SOM is successfully isolating variable sources in Regions II to Region VI from both variable sources in other regions and standard star sources in all regions, and often mapping variable and standard star sources from the same regions into different SOM cells. As aforementioned in Sections 4.4 and 4.5, distinct $(u-g,g-r)$ and $(r-i, i-z)$ colour-space distributions are expected for the variable and standard star source populations within the groups of cells dominated by variable Region IV and VI sources, given the variable sources from these regions are known to occupy different regions of $(u-g, g-r, r-i, i-z)$ colour-space to the standard star sources (and variable/standard star sources from other regions

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

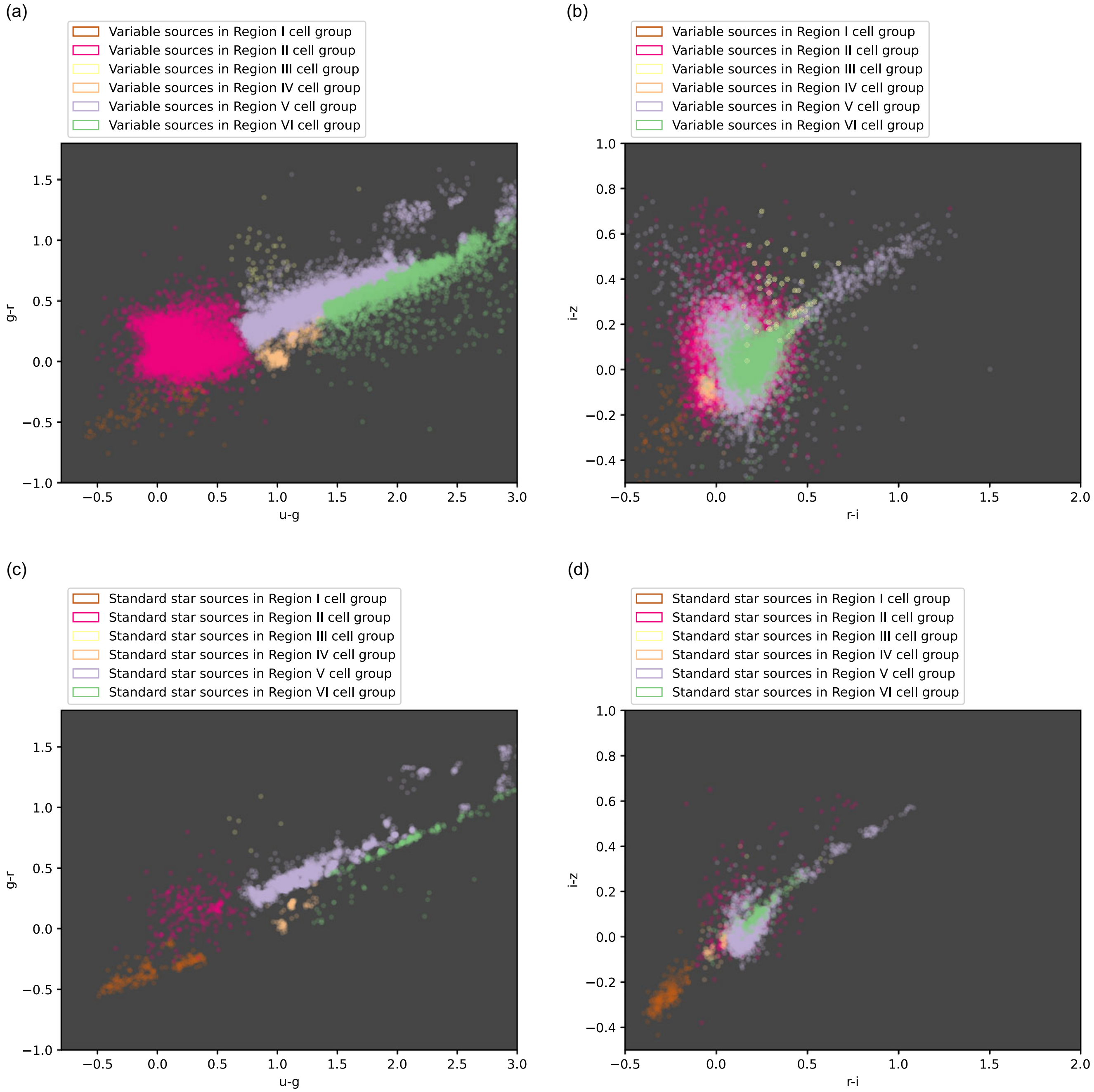

Figure 10. The distributions of median colours for all sources within the six groups of cells detailed in Section 4. For the sources in the cell groups dominated by Region II to Region VI variable sources, the variable source median (a) $(u-g, g-r)$ and (b) $(r-i, i-z)$ colour distributions differ markedly from, but encompass, the standard star median $(u-g, g-r)$ (c) and (d) $(r-i, i-z)$ colour distributions. This reflects the ability of the SOM to successfully select these variable sources with a high purity. In contrast, the variable sources from the cell group identified as containing the highest purity of variable Region I sources (described in Section 4.1) occupy a subset of the four dimensional colour-space spanned by standard star sources from the same group of cells. This is not unexpected given the inability of the SOM to select variable sources from Region I (also described in Section 4.1).

of $(u-g, g-r, r-i, i-z)$ colour-space, by definition). This is also expected in the case of sources from the groups of cells dominated by variable Region II sources (as discussed in Section 4.2), given that this region has an intrinsic variable source purity of 93%.

However, it is particularly remarkable to attain different $(u-g, g-r)$ and $(r-i, i-z)$ colour-space distributions for the variable and standard star sources mapped onto the group of cells dominated by variable Region V sources. As detailed in Figs. 4 and 5, the variable and standard star source from Region V are degenerate $(u-g,g-r)$ and $(r-i, i-z)$ colour-spaces. Thus, this demonstrated ability of the SOM to isolate variable Region V sources, and form a population of sources with distinct $(u-g, g-r)$ and $(r-i, i-z)$ colour-space distributions from the

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

standard star sources in the same group of cells, is not a predicted outcome.

In contrast to the variable sources occupying the other identified groups of cells, the variable sources from the cell group containing the highest purity of variable Region I sources (described in Section 4.1) occupy a subset of the four dimensional colour-space spanned by standard sources from the same group of cells. This is not a surprising outcome, given these cells are dominated by standard star sources, as detailed in Section 4.1.

5. Summary and conclusions

We studied the selection of variable main sequence sources from the SDSS Stripe 82 variable source catalogue (Ivezić et al. 2007) utilising SOMs, from a heterogeneous sample of variable and non-variable SDSS Stripe 82 sources. This paper begins with the assembly of a $(u-g, g-r, r-i, i-z)$ colour-space sample of 493 053 sources (including 67 507 variable sources), following the procedure outlined in Ivezić et al. (2007). Following Sesar et al. (2007), the sources in this sample were divided into six regions (I, II, III, IV, V, and VI) of $(u-g, g-r)$ colour-space, each with a different predominant variable source type.

In Section 3, we then explored the use of a SOM, trained with the $(u-g, g-r, r-i, i-z)$ colours of the complete heterogeneous sample of variable and standard star sources, to select variable sources occupying Region V of $(u-g, g-r)$ colour-space from the entire heterogeneous sample. As mentioned in Section 2, Gaia DR3 Gaia Collaboration et al. (2023) parallax measurements, the SDSS $(u-g, g-r)$ colour-space, and SDSS colour magnitude diagrams indicate these variable sources are almost all main sequence stellar sources. These variable Region V sources are almost entirely degenerate in $(u-g, g-r)$ and $(r-i, i-z)$ colour-spaces with non-variable main sequence sources (which also occupy Region V), and constitute only 10.7% of all Region V sources. Nevertheless, we illustrated that these variable sources occupy a distinct group of cells on the SOM representation, and following Faisst et al. (2019) computed the purity $\mathcal{P}_{\rm V}$ and completeness $\mathcal{R}_{\rm V}$ of variable Region V sources in this group of cells. We showed that it was possible to select these variable Region V sources with a purity of $\mathcal{P}_{\rm V} = 80.2$ and completeness of $\mathcal{R}_{\rm V} = 25.1$ through isolating a group of cells on the SOM representation with a defined minimum purity of variable Region V sources $\mathcal{P}_{\rm V,min} = 60.0\%$ in each cell. Given the aforementioned degeneracy between the variable and non-variable main sequence sources occupying this region of $(u-g, g-r)$ colour-space, the ability to select these variable main sequence sources from all standard star sources (and variable sources occupying different regions of the colour-space) using only these median photometric colours is a significant result. We also illustrated that $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V})$ values can be altered depending on the application; additional examples of the selection purity and completeness values that are possible using the method demonstrated in this study include $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V}) = (48.5\%, 48.5\%)$ and $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V}) = (75.4\%, 29.1\%)$.

We repeat the methodology of Section 3 in Section 4 to briefly explore the ability of the SOM to select variable sources from Regions I, II, III, IV, and VI of $(u-g, g-r)$ colour-space, from the entire heterogeneous sample of all variable and standard star sources, by defining groups of cells dominated by these variable sources on the same SOM representation analysed in Section 3. Selecting variable sources from Regions I and III, respectively, detailed in in Sections 4.1 and 4.3, is not successful. In the case of Region I (dominated by white dwarf sources, according to Sesar et al. 2007), the maximum purity of selected Region I sources is only 31.25%, and in the case of Region III (dominated by dM/WD pairs, according to Sesar et al. 2007), selecting variable sources with a purity of $\mathcal{P}_{\rm III} > 80\%$ is only possible with a very low completeness of $\mathcal{R}_{\rm III} = 3.18\%$. Selecting variable Region II sources (predominantly low-redshift quasars) is at first glance promising: the SOM isolates these sources with a purity of $\mathcal{P}_{\rm II} = 88.5\%$ and completeness of $\mathcal{R}_{\rm II} = 99.6\%$. It is also possible for to select these sources with $(\mathcal{P}_{\rm II}, \mathcal{R}_{\rm II}) = (95.4\%, 95.4\%)$ and $(\mathcal{P}_{\rm V}, \mathcal{R}_{\rm V}) = (96.8\%, 92.3\%)$. However, as fully detailed in Section 4.2, the non-variable sources from this region are also present in this group of cells, and it is unclear if the SOM successfully separates the variable and standard sources from this region. As discussed in Section 4.4, selecting sources from Region IV (dominated by RR-Lyrae stars and variable main sequence sources) is successful, with variable sources from this region selected with a purity of $\mathcal{P}_{\rm IV} = 81.4\%$ and a of completeness of $\mathcal{R}_{\rm IV} = 20.3\%$. However, this separation is reflective of the distinct types and intrinsically different colours of variable and standard star sources in this region, not differences in variability between sources of the same type. Lastly, the selection of variable sources from Region VI (discussed in Section 4.5) was also predictably successful given the intrinsically high variable source purity of this region and that the distinct types of variable and standard star sources in this region also often occupy different areas of $(u-g, g-r)$ colour-space. The variable sources from this region of colour-space are separated from the heterogeneous sample with a purity of $\mathcal{P}_{\rm VI} = 80.9\%$ and completeness of $\mathcal{R}_{\rm VI} = 70.1\%$

This study illustrates that variable Region V sources, otherwise degenerate in both $(u-g, g-r)$ and $(r-i, i-z)$ colour-space with non-variable sources, can be selected using an SOM. It is important to note that the selection of variable sources detailed in this paper does does not use the photometric errors associated with each variable (and standard star) source. However, the errors on each source colour could be incorporated by weighting each point by the inverse of the photometric error during the SOM training process, thus penalising sources with large colour errors. This would result in clusters preferentially defined by sources with accurate photometry, likely increasing the accuracy of variable source selection. We leave the demonstration of this weighting and analysis of the resulting weighted SOM mapping to future work.

Given the costs of acquiring median photometric data is significantly lower than time-series data this method demonstrated in this work has significant potential for classifying variable main sequence sources in upcoming survey data. Furthermore, this method could facilitate the re-analysis of archival data, where light curves are often unavailable. Future and archival surveys which have both common median photometric colour dimensions and identical survey depth in each passband to the SDSS Stripe 82 dataset used in this study can be mapped directly onto the same trained SOM utilised in this study to select and classify variable sources. The method detailed in this study is also applicable to future or archival survey data with differing photometric depth in the four-dimensional colour-space (e.g. LSST survey data). The subset of sources from these surveys that within the magnitude limits of the Stripe 82 Survey (thus occupying the same region of four-dimensional colour-space mapped in this study) will again be able to be mapped directly onto the trained SOM detailed in this work, facilitating the instantaneous separation and classification of variable sources. The remaining subset of variable sources,

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press

which occupy a region of four-dimensional colour-space that is *not* mapped in this study may not be separated successfully from standard star sources using the specific SOM detailed in this study. In this case, the further investigation through the trial application of the method detailed in this paper is needed to determine if variable sources from the given region of colour-space are still separated from other variable and standard star sources with a sufficient purity. We leave the selection of variable sources from other heterogeneous survey datasets, using the method detailed in this paper, to future work.

Acknowledgements. T. V. would like to acknowledge the supervision, support and encouragement of P. L. Capak. T. V. acknowledges the support of the Caltech Visiting Undergraduate Research Program. T. V. would also like to acknowledge the encouragement and support of A. R. Duffy.

Data availability statement. Data sharing is not applicable to this article as no new data were created or analysed in this study.

Funding statement. This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT), NRF-2022R1C1C1008695. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Competing interests. None.

References

Assef, R. J., *et al.* 2013, ApJ, 772, 26
Bailey, A., *et al.* 2022, arXiv e-prints, arXiv:2211.01206
Baron, D. 2019, arXiv e-prints, arXiv:1904.07248
Covey, K. R., *et al.* 2007, AJ, 134, 2398
Faisst, A. L., Prakash, A., Capak, P. L., & Lee, B. 2019, ApJ, 881, L9
Fluke, C. J., & Jacobs, C. 2020, WIREs DMKD, 10, e1349
Foley, R., *et al.* 2019, BAAS, 51, 305
Gaia Collaboration, *et al.* 2023, A&A, 674, A1
Hanke, M., *et al.* 2009, Neuroinformatics, 7, 37
Ivezić, Ž., Vivas, A. K., Lupton, R. H., & Zinn, R. 2005, AJ, 129, 1096
Ivezić, Ž., *et al.* 2007, AJ, 134, 973
Ivezić, Ž., *et al.* 2019, ApJ, 873, 111
Kohonen, T. 1982, BC, 43, 59
Kohonen, T. 1990, Proc. IEEE, 78, 1464
Masters, D., *et al.* 2015, ApJ, 813, 53
Pérez-Díaz, V. S., Martínez-Galarza, J. R., Caicedo, A., & D'Abrusco, R. 2024, MNRAS, 528, 4852
Peters, C. M., *et al.* 2015, ApJ, 811, 95
Rhodes, J., *et al.* 2017, ApJS, 233, 21
Sesar, B., *et al.* 2007, AJ, 134, 2236
Soares-Santos, M., *et al.* 2017, ApJ, 848, L16
Stern, D., *et al.* 2005, ApJ, 631, 163
Thanjavur, K., *et al.* 2021, MNRAS, 505, 5941
Webbe, R., & Young, A. J. 2023, RAS TI, 2, 238

https://doi.org/10.1017/pasa.2024.118 Published online by Cambridge University Press