Improved selection of extremely red quasars with boxy C IV lines in BOSS

Reza Monadi [ORCID]★ and Simeon Bird

Department of Physics and Astronomy, University of California, Riverside, 92507 CA, USA

ABSTRACT

Extremely red quasars (ERQs) are an interesting sample of quasars in the Baryon Oscillation Spectroscopic Sample (BOSS) in the redshift range of 2.0–3.4 and have extreme red colours of $i - W3 \geq 4.6$. Core ERQs have strong C IV emission lines with rest equivalent width of ≥ 100 Å. Many core ERQs also have C IV line profiles with peculiar boxy shapes which distinguish them from normal blue quasars. We show, using a combination of kernel density estimation and local outlier factor analyses on a space of the $i - W3$ colour, C IV rest equivalent width and line kurtosis, that core ERQs likely represent a separate population rather than a smooth transition between normal blue quasars and the quasars in the tail of the colour-REW distribution. We apply our analyses to find new criteria for selecting ERQs in this 3D parameter space. Our final selection produces 133 quasars, which are three times more likely to have a visually verified C IV broad absorption line feature than the previous core ERQ sample. We further show that our newly selected sample are extreme objects in the intersection of the *WISE* AGN catalogue with the MILLIQUAS quasar catalogue in the colour–colour space of $(W1 - W2, W2 - W3)$. This paper validates an improved selection method for red quasars which can be applied to future data sets such as the quasar catalogue from the Dark Energy Spectroscopic Instrument.

Key words: quasars: emission lines – quasars: general – galaxies: statistics.

1 INTRODUCTION

Quasars are high luminosity active galactic nuclei (AGNs), fuelled by gas and dust accreting on to a supermassive black hole (SMBH). Observations show that the growth of a SMBH is correlated with the physical properties of the host galaxy, such as velocity dispersion, stellar mass, and star formation rate, although the mechanisms which correlate these properties are not completely clear (Gebhardt et al. 2000; Tremaine et al. 2002; Haring & Rix 2004; Gültekin et al. 2009; Shankar, Bernardi & Haiman 2009; Kormendy & Ho 2013; Azadi et al. 2015). The co-evolution of an SMBH and its host galaxy mostly occurs during dusty starbursts resulting in the observation of sub-mm or ultra-luminous infrared galaxies (Sanders et al. 1988; Veilleux et al. 2009; Simpson et al. 2014). Kroupa et al. (2020) proposed a model whereby SMBHs form when the dynamical collapse of star clusters is accelerated by the accretion of gas from a host galaxy, which can naturally explain the correlation between a host galaxy and SMBH mass, as well as explain quasars found at high redshift.

Unobscured quasars which exhibit blue thermal continuum are the majority of optically selected quasars. Red quasars, on the other hand, are a small population of quasars that show a variety of redder near infra-red and optical colours. Several studies have investigated the origin of the red colour in red quasars (e.g. Kim & Im 2018 or Klindt et al. 2019), however the question still remains unsettled (Calistro Rivera et al. 2021).

One possibility is that red quasars have been obscured and reddened by dust during a brief transition phase between dusty starburst galaxies and blue quasars (e.g. Richards et al. 2003; Hopkins et al.

2005; Hopkins et al. 2008; Urrutia, Lacy & Becker 2008; Glikman et al. 2012, 2015; Assef et al. 2015; Banerji et al. 2015; Ishibashi & Fabian 2016; Hickox & Alexander 2018). In this model, a quasar is buried in the starburst dust when the host galaxy is young, making the colour of that quasar red. Many ERQs also exhibit extreme line properties which may indicate unusually powerful outflows occurring in a young evolution phase. Quasar-driven outflows may clear out the observer's line of sight, and so at the end of this evolutionary phase, we observe an optical and/or UV luminous quasar.

There are other models; for example, the unified AGN model (Antonucci 1993; Netzer 2015) suggests that red quasars are viewed with intermediate orientations between Type 1 and Type 2 quasars (for a recent review see Hickox & Alexander 2018). According to this model, unobscured (type 1) AGN are viewed face-on, while obscured (type 2) AGNs are observed edge-on. A red colour is produced by a dusty torus blocking part of the nuclear emission. However, this model has difficulty explaining the extreme line properties seen in some red quasars (Urrutia et al. 2009; Klindt et al. 2019).

Ross et al. (2015) studied a population of red quasars at $0.28 \leq z \leq 4.36$ in the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) of the Sloan Digital Sky Survey-III (SDSS-III; Eisenstein et al. 2011). These red quasars were identified using a simple colour selection originally intended for red galaxies: a magnitude difference of $r_{AB} - W4_{Vega} \geq 14$ between the infra-red band (*W4* in *WISE* with effective wavelength of 12 μm) and the optical band (*r* in SDSS with effective wavelength of 6231Å).

Hamann et al. (2017) (hereafter H17) used the sample of Ross et al. (2015) but narrowed down the redshift range to $2.0 \leq z \leq 3.4$ and changed the colour selection to $i - W3 \geq 4.6$ (∼3 magnitudes redder than the typical colour of BOSS quasars), calling

★ E-mail: reza.monadi@email.ucr.edu

the sample thus identified Extremely Red Quasars, or ERQs.[1] Interestingly, ERQs showed exotic spectral properties, which motivated H17 to define a smaller core ERQ (CERQ) subsample defined by REW(C IV)≥100 Å. This criterion was chosen to better correlate red quasars with other extreme line properties: peculiar boxy profiles, NV>Ly α, a high incidence of blueshifted broad absorption lines (BALs), and [O III] 5007Å outflow speeds reaching >6000 km s^{-1}. ERQs also have an unusually flat UV SED considering their extreme red colour (steep Mid-IR to UV SED), although this may be an artefact of the BOSS selection algorithm, which would not target quasars which are red in all SDSS bands for spectroscopic follow-up.

Perrotta et al. (2019) (hereafter P19) studied a sample of 28 ERQs and found an outflow speed for the [O III] line (see Zakamska N. L. et al. 2016) of 1992–6702 km s^{-1}. This is on average three times faster than those of luminosity matched blue quasars. This outflow speed is highly correlated with $i - W3$ colour but not with radio loudness nor Eddington ratios. However, Fawcet et. al. (2020) and Rosario D. J. et. al. (2020) showed that red quasars have different radio properties with the help of more sensitive radio observations P19 suggests that this correlation may indicate a connection between reddening and the efficiency of energy and momentum injection from ERQs to the interstellar medium. ERQs may produce more effective feedback in their host galaxies, regulating the star formation rate and SMBH growth more effectively. This is again indicative that some ERQs with extreme line values are connected with an early dusty stage of quasar–galaxy evolution where strong quasar-driven outflows provide important feedback to the host galaxies (P19).

We therefore have a working hypothesis identifying ERQs with an intermediate stage of quasar evolution between dusty galaxies and red quasars. This study is an effort to produce quantitative evidence for or against this hypothesis, which is based on the unusual line properties exhibited by the spectra. If such quantitative evidence is forthcoming, we also desire to refine the selection criteria for ERQs in order to better study the outflows connected with this stage of quasar evolution. We will provide selection criteria for objects that exhibit the extreme properties of ERQs, among a sample of quasars with spectroscopic data. We use the existing manually selected sample of ERQs to define a training set, and then provide a modified sample of extremely red quasars in BOSS with more uniform (and more uniformly exotic) properties. In summary, we address the following questions:

1) To what extent are ERQs separated from the main locus of BOSS quasars (Section 4.1, Section 4.2, and Section 4.3)?

2) If they are, which selection criteria best produce quasars connected with this intermediate stage of quasar evolution Section 4.4)?

We acknowledge the possibility that our sample may be affected by the selection criteria of BOSS, which uses a colour selection to find quasar candidates and thus may discard some red quasars. However, in the absence of another equally large spectroscopic quasar survey this is unavoidable. We will thus analyse BOSS quasars and check for evidence that we are affected by colour selection in Section 4.5.

We use a standard cosmology throughout ($H_0 = 67.3$ km s^{-1} Mpc^{-1}, $\Omega_m^0 = 0.315$, $\Omega_\Lambda = 0.685$) (Planck Collaboration XVI 2014).

2 QUASAR SAMPLES

In this section, we introduce our quasar samples their selection criteria, summarized in Table 1. The primary parent sample is similar

[1]ERQs are not the reddest $i - W3$ quasars overall, but the reddest ones in BOSS.

Table 1. The parent samples, sizes, and sample selection criteria for the various quasar samples we use. All subsets are taken from a parent sample made in H17 by custom emission line fits. The first sample, T1, is a superset of all the others. These are: type 1 luminosity matched quasars (T1LM), type 1 extremely red quasars (T1ERQ) and type 1 core extremely red quasars (T1CERQ).

Sample	Selection criteria	Size
T1	FWHM(C IV) $\geq$ 2000 km s^{-1}	–
	$2 \leq$ z_dr12 ≤ 3.4	–
	$i - W3 > 0.8$	–
	SNR(REW(C IV)$\geq$3	–
	SNR(FWHM(C IV)$\geq$4	–
	SNR(AB$_{W3}$) ≥ 3	–
	q_flag $=0$	–
	cc_flag='0000'	–
	nv_flag=0	35 976
T1LM	$10^{46.54}$ erg s^{-1} $\leq L_{bol} \leq 10^{48.00}$ erg s^{-1}	29 072
T1ERQ	$i - W3 \geq$4.6	154
T1CERQ	$i - W3 \geq$4.6	–
	REW(C IV)$\geq$ 100 Å	72

to the emission-line catalogue of H17 which results from custom fits of C IV and N V emission lines performed on spectra in the SDSS-III BOSS quasar catalogue (Paris 2017). The subsamples follow the selections in H17, to which we refer the reader for a detailed explanation of the criteria adopted.

2.1 Type 1 sample

Following the H17 sample selection procedure, we first limit the quasar redshift to $2 \leq z \leq 3.4$. This redshift range encompasses most of the BOSS survey, while ensuring that Ly α and N V λ1240 are within the BOSS spectral range. We require that successful fits to the N V (nv_flag=0) and C IV (q_flag $=0$) emission lines are made at reasonable signal to noise (SNR(REW(C IV)) $\geq$3 and SNR(FWHM(C IV)$\geq$4)). We limit ourselves to Type 1 quasars, defined as FWHM(C IV) $\geq$2000 km s^{-1} (Alexanderoff et al. 2013; Ross et al. 2015). We also require that the quasars have a good detection in the W3 band (SNR(AB$_{W3}$) ≥ 3), do not exhibit artefacts in the *WISE* data (cc_flag='0000'), and are not excessively blue ($i - W3 > 0.8$).

2.2 T1ERQ and T1CERQ samples

Following H17 we use a colour cut of $i - W3 \geq 4.6$ to extract ERQs from the full sample of type 1 quasars. H17 considered several colour cuts, choosing their boundary to produce the most dramatic differences in the median spectral properties of ERQs as compared to blue quasars. H17 also defined a subsample, core type 1 ERQs (T1CERQs), with the additional condition of REW(C IV)≥100 Å, chosen to be more correlated with the unusual line properties found in some ERQs. These conditions define a natural 2D parameter space in $i - W3$ and REW(C IV), which we will use extensively in what follows.

2.3 T1LM sample

ERQs are very luminous, with an average bolometric luminosity for T1CERQs of $10^{47.21 \pm 0.31}$ erg s^{-1}. For comparison, the full quasar sample has an average luminosity of $10^{46.82 \pm 0.21}$ erg s^{-1}. This high

luminosity is a selection effect. SDSS cannot detect faint ERQs, because it does not detect objects with an *i*-band magnitude ≲22.

This large luminosity has implications for our later analysis, as there is an anticorrelation between the REW of the C IV emission line and the continuum luminosity of Type 1 quasars (Baldwin 1977). We made a sample of Type 1 luminosity matched quasars (T1LM) drawn from the T1 sample, but with luminosities between the minimum and maximum luminosities of T1CERQs (i.e. $10^{46.54} \le L_{bol}(T1CERQ) \le 10^{48.00}$).

We derive mean luminosities using the procedure described in P19, using the AB magnitude in the W3 band and the luminosity distance of our standard cosmology.

Bolometric luminosities are difficult to determine for ERQs due to the large and uncertain extinction corrections which must be applied in the rest-frame UV and optical. However, the W3 band is less susceptible to obscuration than the UV and optical bands. We therefore use it as a surrogate for estimating the bolometric luminosity. We add a correction for flux suppression due to obscuration of a factor of 8, following H17. We thus set $L_{bol} = 8\lambda L_\lambda$ at $\lambda = 3.45\,\mu m$ in the rest frame. *WISE* W3 photometry measures $\sim 3.45\,\mu m$ in the rest-frame at the typical redshift of ERQs in our study. We first converted the AB magnitude of W3 to units of flux per frequency $(F_\nu = 10^{-0.4(48.6 + m_{AB})})$ and then to observed flux per wavelength $(F_\lambda = \frac{c}{\lambda^2} F_\nu)$. Intrinsic luminosity is computed using the luminosity distance, D_L, of our standard cosmology by:

$$L_\lambda \lambda_{rest} = 4\pi D_L^2 F_\lambda \lambda_{obs}. \tag{1}$$

Even after luminosity matching, the sample may contain quasars with a range of different black hole masses or Eddington ratios. However, reliable measures of these quantities are not generally available in our data set, and so to avoid the risk of overfitting we rely on simple luminosity matching to homogenize our sample.

3 ANALYSIS METHODS

3.1 Kurtosis of the C IV line: a third parameter

H17 investigated a large number of unusual emission line properties in the T1ERQ and T1CERQ samples. In particular, they found boxy C IV line shapes, a large N V to C IV line ratio (N V/C IV) and moderately reduced FWHM(C IV) compared to a population of normal blue quasars with similar W3 magnitudes to T1CERQs. However, the N V fits done in H17 did not attempt to deblend the nearby Lyman-α line and so the N V strength may be overestimated. The C IV line is uniquely powerful for our analysis as it is the strongest metal line in the quasar spectrum. Other promising lines (e.g. Si IV or He II) lines are much weaker or blended. For example, C III is blended with S III and Al III.

We focus here on the boxy shape of the C IV line, quantified by the kurtosis. Kurtosis (kt_{80}) is defined in H17 as the ratio of the velocity width of the C IV line at 80 per cent of the peak height to the velocity width at 20 per cent of the peak height. A high kurtosis C IV line profile occurs in most ERQs and indicates a boxy line. The median kt_{80} for T1ERQs is 0.35 and for T1CERQs 0.36, while the larger T1LM sample has a median of 0.25.

Fig. 1 show histograms of kt_{80}(C IV). Each panel is labelled by joint thresholds on REW(C IV) and $i - W3$ colour and the number of quasars satisfying these conditions. A redder colour skews the kt_{80}(C IV) distribution towards more boxy C IV line quasars (higher kt_{80}(C IV)). Increasing the REW(C IV) threshold does not change the overall shape of the kt_{80}(C IV) distribution dramatically, nor its most probable value when compared to the unconditioned sample in the

top left-hand panel. However, we see a slightly enhanced population of high kurtosis objects when conditioning on REW(C IV). High kt_{80}(C IV) is thus highly correlated with red colour, but not with REW(C IV), suggesting that it is a good choice for a third parameter, along with $i - W3$ and REW(C IV).

Note that there is a possible confounder in the fitting procedure of H17: weak lines will be fit with a single Gaussian rather than two if the second Gaussian does not improve the fit. A single Gaussian has $kt_{80} = 0.37$. We have checked that this does not significantly affect our results by making a version of Fig. 1 where spectra with kt_{80}(C IV) >0.37 have been removed. This reduces the total size of the sample by 34 780 spectra and the number of core ERQs by 57. In practice, since most of the removed spectra are not ERQs this cut moderately strengthens the trends we report. Fig. 2 shows these high kt_{80}(C IV) objects in the low REW(C IV) and blue part of the parameter space. The median REW(C IV) for kt_{80}(C IV) >0.37 and kt_{80}(C IV) >0.36 are 17 Å and 19 Å, respectively. This indicates that most of the quasars with high kt_{80} are weak C IV line objects, very far from the ERQs in colour space.

3.2 Defining T1CERQs with a wedge or a cone

One of our main goals in this study is to examine variations in quasar spectral properties as one moves in parameter space between the main quasar locus and T1CERQs. We define $\vec{v}_{T1CERQ}$, the vector in the 2D parameter space of colour-REW(C IV) between the median of the T1LM sample and the median of the T1CERQ sample, where the black line in Fig. 2 shows its direction. In order to average over quasar properties, we then define wedges (in 2D) and cones (in 3D), the simplest directional geometric shapes for calculating the median spectra over a region in parameter space. We will use these shapes extensively in the following analysis.[2]

A complexity to our definitions of wedges and cones is that we need a dimension-less parameter space within which to define opening angles. We thus normalize all parameters to a dimensionless unit square (cube), based on the range of each parameter. The normalization procedure we choose is the min-max method,[3] which performs a linear transformation to map each coordinate on to a unit square (cube). The maximum value found in the data set maps to 1 and the minimum maps to 0. For our data set, $i - W3$ ranges between 0.8 and 8.0, log 10(REW(C IV)) between 0.2 and 2.6, and kt_{80}(C IV) between 0.11 and 0.39.

We define a wedge in a 2D space of $i - W3$ and REW(C IV) along the vector between the median of the T1LM quasar sample and the T1CERQ sample. This vector is $v_{T1CERQ}^{2D} = (0.40, 0.26)$ in the normalized space, which corresponds to (2.90,136Å) once the normalization is removed. The opening angle for this wedge is calculated by:

$$\theta = \max_i \{\angle P_i, v_{T1CERQ}\}. \tag{2}$$

P_i is a vector from the median of the T1LM sample and the *i*th quasar in the T1CERQ sample. $\angle$ means the angle between two vectors and is always less than 180°. Equation (2) implies that θ is the maximum angle among all deviation angles of T1CERQs from v_{T1CERQ}. θ is thus the smallest angle for which the wedge

---

[2]It is possible to consider more complex geometries. However this makes the analysis overcomplicated and is not necessarily better than a wedge (or cone) which covers an area (or volume), as long as a variety of directions are included.

[3]We used the `MinMaxScaler` function from `sklearn`
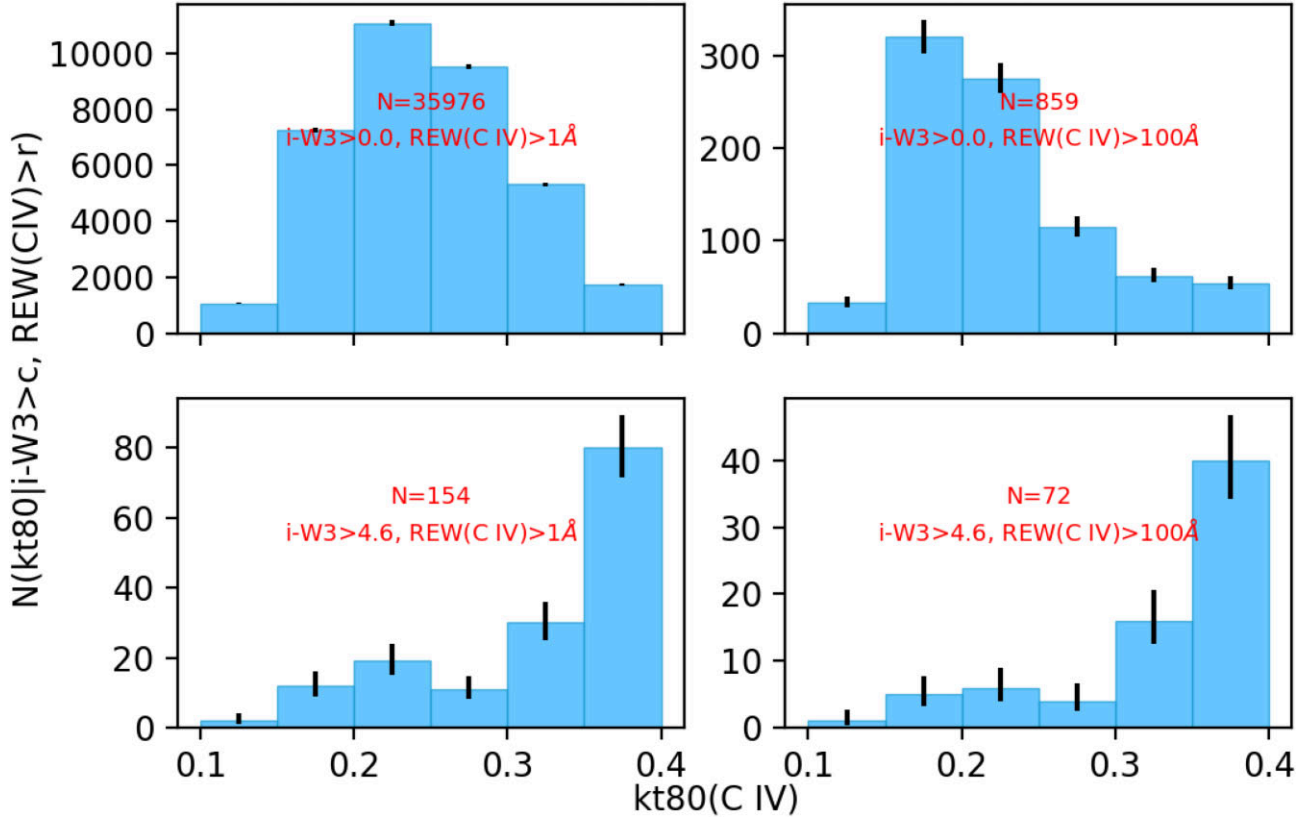
Figure 1. Histograms for distribution of $kt_{80}(C\,IV)$ given conditions on $i - W3$ and $REW(C\,IV)$. Top left-hand panel shows the unconditioned distribution. Bottom left-hand panel is conditioned on $i - W3 > 4.6$, and thus shows T1ERQs. Top right-hand panel is conditioned on $REW(C\,IV) > 100\,\text{Å}$ and bottom right-hand panel is conditioned on both, thus showing T1CERQs. Each panel is labelled by the conditions and the number of quasars satisfying them. c and r in $N(kt_{80}(C\,IV)|i - W3 > c,\ REW(C\,IV) > r)$ are the colour and $REW(C\,IV)$ thresholds shown in each panel.
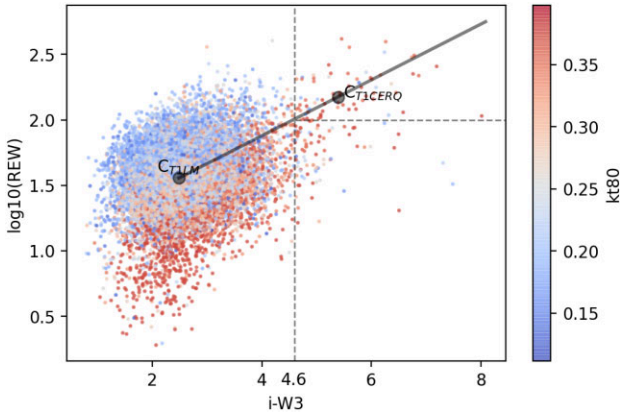
Figure 2. Luminosity matched sample distribution in $(i - W3, REW(C\,IV), kt_{80}(C\,IV))$ space. The redder points show higher $kt_{80}(C\,IV)$ and thus higher kurtosis. The vertical line separates the T1ERQ sample from the rest of T1LM sample and the horizontal line separates T1CERQs from other T1ERQs. The black line is along $\vec{v}_{T1CERQ}$ (see Section 3.2), which connects the median of the T1LM (C_{T1LM}) sample to the median of the T1CERQ (C_{T1CERQ}) sample.

covers the entire T1CERQ sample, quasars with $i - W3 \geq 4.6$ and $REW(C\,IV) \geq 100\,\text{Å}$. The opening angle in the normalized 2D parameter space of $i - W3$ and $REW(C\,IV)$ for v^{2D}_{T1CERQ} is $\theta = 18.5°$.

Equivalently, in the 3D normalized space of $(i - W3, REW(C\,IV), kt_{80}(C\,IV))$ we defined a cone towards the median of T1CERQs. But

we observed that spectra with low kt80 did not generally exhibit the exotic line properties of T1CERQs. Therefore, we impose a minimum $kt_{80}(C\,IV) \geq 0.33$. This cut excludes quasars with low $kt_{80}(C\,IV)$ which have large θ angles from v^{3D}_{T1CERQ} and keeps the cone focused on the v^{3D}_{T1CERQ} direction. This threshold is chosen because it corresponds to the dip in the population distribution of $kt_{80}(C\,IV)$ conditioned on red colour and strong $REW(C\,IV)$ (bottom right panel of Fig. 1). If we do not impose $kt_{80}(C\,IV) \geq 0.33$, the cone opening angle will be large, $35.7°$, and will include many interloper quasars without the extreme line properties of T1CERQs. The corresponding vector between the median of the T1LM quasar sample and the T1CERQ sample is $v^{3D}_{T1CERQ} = (0.40, 0.26, 0.38)$, which is $(2.90, 136\text{Å}, 0.11)$ when the normalization is removed.

The 3D cone with this definition thus includes all quasars with $i - W3 \geq 4.6$, $REW(C\,IV) \geq 100\,\text{Å}$, and $kt_{80}(C\,IV) \geq 0.33$. It has an opening angle of $\theta = 19.6°$ in the normalized 3D space.

Fig. 2 visualizes the boundaries between the T1LM, T1ERQ, and T1CERQ quasar samples in the parameter space of $i - W3$ and $REW(C\,IV)$. Colours denote kt_{80}, showing again the large kt_{80} associated with ERQs. The line in Fig. 2 is along v^{2D}_{T1CERQ}, directed from the median of the T1LM sample to the median of the T1CERQ sample.

3.3 Local outlier factor analysis

Wishing to investigate whether T1CERQs are a separate sub-population of quasars, we applied several clustering methods on our data set. We tried density-based clustering techniques (e.g.

DBSCAN; Ester et al. 1996) and hierarchical clustering algorithms (e.g. agglomerative clustering; William & Edelsbrunner 1984). However, clustering algorithms could not handle the very wide disparity in size between the T1LM sample (29 237 quasars) and the T1CERQ sample (72 quasars). Since T1CERQs are a very small portion (0.25 per cent) of the total quasar sample, clustering methods were either not able to find T1CERQs as a separate cluster or, if they could, the uncertainties in the obtained labels were high.

Instead, we used a Local Outlier Factor (LOF)[4] analysis (Brreuing et. al. 2000) which quantifies the level of distinctness in T1CERQs. The LOF has had other uses in astronomy: for example, detecting unusual spectra in SDSS (Wei et al. 2013) and distinguishing supernovae candidates from massive galaxies (Tu et al. 2010). LOF measures the extent to which a data point is isolated with respect to its neighbours by comparing the local reachability density of an object to the local reachability density of its k-nearest neighbours using the following score:

$$LOF_k(A) = \frac{1}{\rho_k(A)} \frac{\sum_{B \in \mathbb{N}_k(A)} \rho_k(B)}{\|\mathbb{N}_k(A)\|}. \tag{3}$$

This is otherwise called the LOF score for the k-nearest neighbours of point A. $\rho_k(A)$ (or $\rho_k(B)$) is the local reachability density of the k-nearest neighbours of A (or B), defined by:

$$\rho_k(A) = \frac{\|\mathbb{N}_k(A)\|}{\sum_{B \in \mathbb{N}_k(A)} RD_k(A, B)}, \tag{4}$$

where $\mathbb{N}_k(A)$ (or $\mathbb{N}_k(B)$) is the set of all k-nearest neighbours of the point A (or B). $\|\mathbb{N}_k(A)\|$ is the number of objects in $\mathbb{N}_k(A)$. $RD_k(A, B)$ is the reachability distance between point A and B defined by:

$$RD_k(A, B) = \max\{\mathbb{D}_k(B), d(A, B)\}. \tag{5}$$

$d(A, B)$ in equation (5) is the Euclidean distance between point A and B in the normalized 2 or 3D space. $\mathbb{D}_k(B)$ is the set of all distances between point B and $\mathbb{N}_k(B)$.

For example, the density around a data point, deep in a dense cluster of points, is very similar to the density of its neighbourhood; this results in LOF~1. If the data point is located somewhere denser than its nearest neighbours, then it has LOF<1. A point where the average density of the neighbours is higher than that of the point has LOF>1, corresponding to the expected behaviour for a small cluster separate from the main group.

The LOF is defined as a function of the number of nearest neighbours, k, which sets the scale or resolution of the cluster searched for. Thus translated into our analysis, k provides information about the size of the putative T1CERQ cluster.

3.3.1 Mock data analysis in 2D

To better illustrate the behaviour of the LOF(k) score on known distributions of data points and for different k-nearest neighbours, we created 100 mock 2D data sets by making 100 draws from two overlapping Gaussian distributions, G_1 and G_2. To make each mock data set we draw 30 000 data points from (G_1: $\mathcal{N}(\mu = [0, 0], \sigma = [1, 0; 0, 1])$ and 200 data points from (G_2: $\mathcal{N}(\mu = [3, 3], \sigma = [1, 0; 0, 1])$), which in total gives us 100 mock data sets consisting of 30 200 data points each. We chose the same covariance matrix for G_1 and G_2 for simplicity. However, the distance between the centres of G_1 and G_2 imitates the distance between the median of $i - W3$ in T1LM and the median of $i - W3$ in T1CERQs. Similar to the i

---

[4]We used the LOF implementation in `sklearn` (Pedregosa et al. 2011).
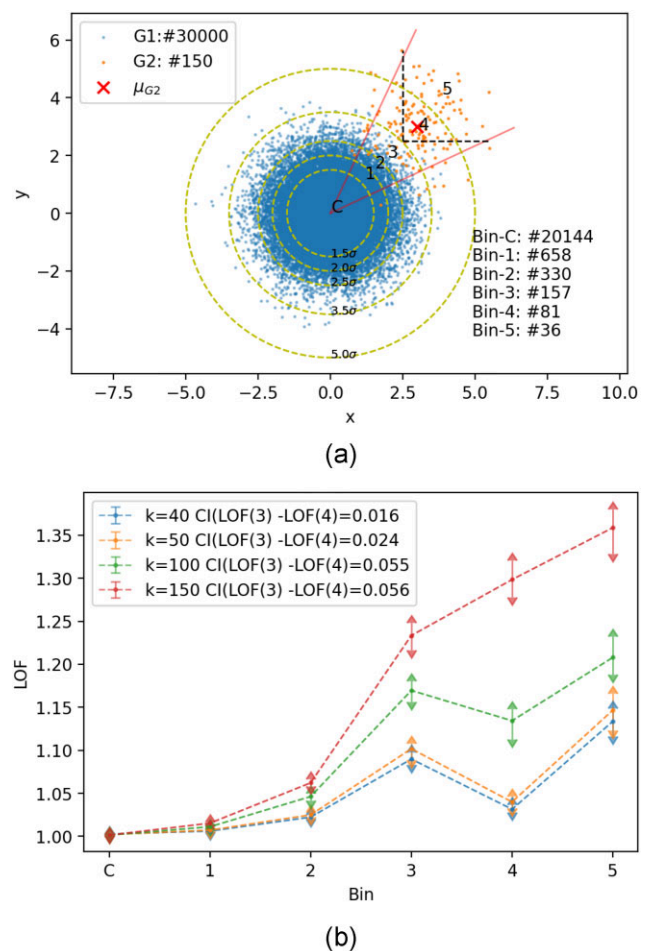
Figure 3. Top (a): 2D density binning on a mock data set composed of two Gaussian populations: G_1: $\mathcal{N}(\mu = [0, 0], \sigma = [1, 0; 0, 1])$ with 30 000 points (blue dots) and G_2: $\mathcal{N}(\mu = [3, 3], \sigma = [1, 0; 0, 1])$ with 100 points (orange dots). The red cross shows the centre of G_2 sample. The yellow circles are at constant distances from the centre of G_1 of 1.5σ, 2σ, 2.5σ, 3.5σ, and 5σ. The data points inside the dashed lines ($x, y > 2.5$) are those which would be selected as mock T1CERQs following the procedure outlined in Section 3.2. Bottom (b): Median LOF scores and their uncertainties in the bins shown in the top panel for nearest neighbours of $k = 40, 50, 100, 150$.

$- W3 \geq 4.6$ and REW(C IV)≥ 100 Å cuts which define T1CERQs, we define a core G_2 sample (cG_2) with $x, y > 2.5\sigma$. The average population of cG_2 among our 100 mock data sets is 96 (see Fig. 3a). Moreover, on average only 1 data point from G_1 belongs to cG_2, while on average 104 data points from G_2 lay outside of the $x, y > 2.5\sigma$ cuts, showing the level of blending between the G_2 and G_1 populations.

We create a wedge, following the same procedure as Section 3.2, for our mock data. We are interested in the behaviour of data in a wedge directed from the median of the bigger population (G_1 in the mock 2D data, T1LM in the real data set) towards the smaller population (cG_2 in the mock 2D data, T1CERQs in the real data set). Note that the centre of G_2 is 3σ away from the centre of G_1. The opening angle for the mock wedge (see Section 3.2) is $20°$ to be close to the opening angle in 2D of the real data ($18.5°$). The corresponding unit vector that is directed from the centre of G_1 to the centre of G_2 is $\hat{v}_{cG_2} = (1/\sqrt{2}, 1/\sqrt{2})$. To see how LOF$_k$ changes along $\hat{v}_{cG_2}$, we binned the wedge as a function of distance from the centre of G_1, with bins shown in Fig. 3(a).

We then calculated median LOF scores in each bin of the mock 2D data set using nearest neighbours (i.e. cluster size) $k = 40$, 50, 100, and 150 for each of our 100 mock data sets. We plotted the corresponding 68 per cent confidence intervals for median LOF scores within each bin in Fig. 3(a). The LOF scores for the mocks have a local minimum in bin 4, well beyond the 68 per cent confidence intervals of bin 3 and bin 5 and also consistent with the confidence interval of the difference between LOF scores of bin 3 and bin 4 (CI(LOF(3))–CI(LOF(4)) when $k = 40$ or $k = 50$, but not when $k = 100$ or 150.

This local minimum in LOF score is caused by the local overdensity in bin 4 from cG_2. A point in bin 6 close to the centre of G_2 is located in a denser region compared to the average point located in bin 3, where the transition between G_1 and G_2 happens; thus the average LOF score in bin 4 is smaller than in bin 3. In the terminology of the literature, bin 4 includes locally less outlier data points. Data points in bin 5 are far from the centre of G_2; as a result, data points in bin 4 are on average also locally less outlier than data points in bin 5. These two observations explain the dip in the LOF score of bin 4 for $k = 40$ and 50.

However, the local overdensity in bin 4 (i.e. local minimum in median LOF score) is less significant when we consider more neighbours (i.e $k = 100$). This is because the population size of cG_2 in the sample is 96 points. A larger cluster includes much of G_1 in addition to G_2 (see equation 3). Thus LOF_{100} and LOF_{150}, by incorporating more nearest neighbours for a data point in bin 4, do not show a local minimum in the median LOF score. A significant local minimum in the LOF scores in a specific region of parameter space can therefore be used to find the boundary between two populations, even though they have dramatically different sizes.

We confirmed that this local minimum did not occur in other mock data sets without two clearly separated populations. We tested the LOF score variation in a single Gaussian population (G_1) with a normal distribution of $\mathcal{N}(\mu = [0, 0], \sigma = [3, 0; 0, 3])$. We generated LOF_k scores for $k = 40$, 50, 100, and 150 in 100 draws from G_1 each with 30 000 data points, and used the same binning as for the mock data set containing two Gaussian distributions. As before we looked at the median LOF_k score and 68 per cent confidence intervals around it. The corresponding plot to Fig. 3(b) never showed a local minimum in the LOF_k score.

3.3.2 Mock data analysis in 3D

To confirm that this dip also occurs in a mock 3D data set, we performed a similar analysis for 3D Gaussian distributions $G_1 : \mathcal{N}(\mu = [0, 0, 0], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 30 000 points and $G_2\mathcal{N}(\mu = [3, 3, 3], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 150 data points. We generated 100 mock data sets and used the same cuts for building cG_2 $(x, y, z \geq 2.5\sigma)$. On average the population of cG_2 in our 100 mock data sets is 50. For comparison, the population size of T1CERQs with the additional $kt_{80}(C\,IV) \geq 0.33$ cut is 52. We never have a point from G_1 in the $x, y, z \geq 2.5\sigma$ region, but on average 100 data points from G_2.

We used a simple binning procedure similar to the one used for our mock 2D data. We defined a cone along $\hat{v} = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]$, directed from the centre of G_1 to the centre of G_2 with an opening angle of $20°$. We defined a central bin C where $r \leq 1\sigma$. Bins 1–5 are within the cone and between two spheres as follows: bin 1, $1\sigma \leq r \leq 1.5\sigma$; bin 2, $1.5\sigma \leq r \leq 2.5\sigma$; bin 3, $2.5\sigma \leq r \leq 4.8\sigma$; bin 4, $4.8\sigma \leq r \leq 7\sigma$; bin 5, $r \geq 7\sigma$.

Fig. 4 shows the median LOF score in each bin for different nearest neighbours of $k = 70$, 100, 150, and 200. The 68 per cent
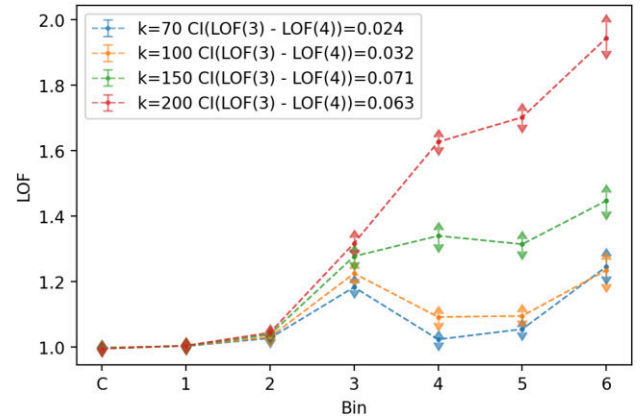
Figure 4. Mock 3D with 2 Gaussian population of $G_1 : \mathcal{N}(\mu = [0, 0, 0], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 30 000 points and $G_2\mathcal{N}(\mu = [3, 3, 3], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 150 data points for the number of nearest neighbours: $k = 70$, 100, 150, and 200. CI(LOF(3)-LOF(4)) in the legend refers to the the 68 per cent confidence interval for the difference of LOF score between bin 3 and bin 4 for each k.

confidence intervals for each median LOF score is shown as the error bar. The decrease in the LOF score from bin 3 to bin 4 is more than the 68 per cent confidence intervals of each bin and also more than the 68 per cent confidence interval of the difference between the average LOF score in bin 3 and bin 4. As a result the local dip in the LOF score of bin 6 is significant to at least the 68 per cent level.

Having demonstrated that a signature of two mixed populations (in 2D and 3D) is a dip in the LOF_k computed along the vector towards the smaller population, at a nearest neighbour value smaller than the size of the smaller population, we continue to analyse our real quasar sample.

4 RESULTS

4.1 Density in the 2D parameter space

Our first objective is to gather evidence to determine whether T1CERQs are part of a separate population or extreme examples of T1 quasars, probing the tail of the main distribution. As a first, simple attempt to answer this question, Fig. 5 visualizes the quasars in the 2D parameter space of $i − W3$ colour, REW(C IV), normalized as explained in Section 3.2. It is visually apparent that the T1CERQs are overdense compared to other regions of parameter space at a similar distance from the main quasar locus. To quantify how much, we computed the density of quasars in parameter space using kernel density estimation (KDE) with a Gaussian kernel. We want to compare the density of the parameter space to the high density region near the median of T1LM sample, and so we plotted density contours relative to the maximum density. For the KDE smoothing bandwidth we applied Silverman's rule of thumb to obtain the bandwidth for each dimension separately:

$$h_i = \sigma_i \left(\frac{4}{N(d + 2)} \right)^{\frac{1}{d+4}}. \tag{6}$$

Here, σ_i is the standard deviation for the i-th dimension of our normalized parameter space, N is the number of objects (29 237 for the T1LM sample), and d is the number of dimensions; 2 in 2D parameter space and 3 in 3D parameter space. Given $\sigma_{i − W3}$
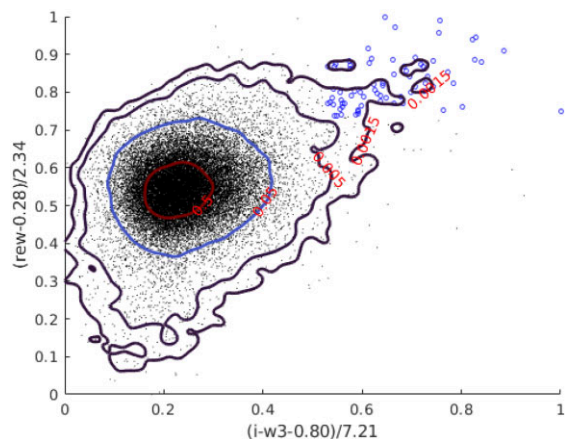
Figure 5. Density of quasars in a normalized $i - W3$, REW(C IV) space. Density contours are shown relative to the maximum density at $\rho/\rho_{max} = 0.5, 0.05, 0.005, 0.0015$. The blue circles are T1CERQs. The black dots are the other T1LMs.
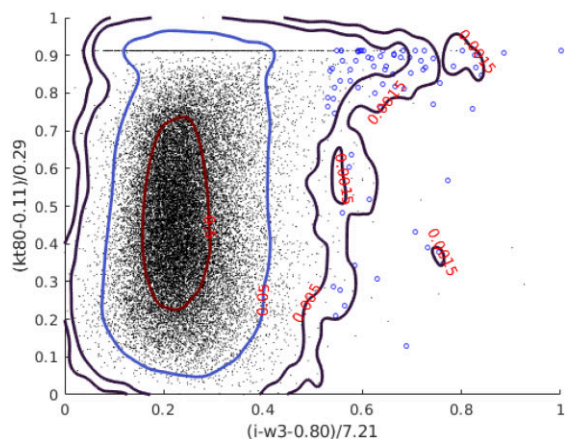


Figure 6. Density of quasars in a normalized $i - W3$, kt$_{80}$ space. Density contours are shown relative to the maximum density at $\rho/\rho_{max} = 0.5, 0.05, 0.005, 0.0015$. The blue circles are T1CERQs. The black dots are the other T1LMs.

$= 0.077$ and $\sigma_{REW(C iv)} = 0.084$, we obtained $h^{2D}_{i-W3} = 0.014$ and $h^{2D}_{REW(CIV)} = 0.015$.

The density contours with $\rho > 0.05\rho_{max}$ in Fig. 5 are similar in shape, and show the shape of the main quasar locus. However, the contours at lower densities are elongated in the direction of the T1CERQs (blue circles in Fig. 5), including some mild local density maxima caused by T1CERQs. The low number of samples in this region means that the density contours are somewhat noisy, but it is apparent that the T1CERQs are an overdensity in this parameter space. Fig. 6 shows a similar density trend in the $i - W3$, kt$_{80}$ plane. Here the overdensity near the T1CERQs is even more apparent: the lowest density contour is extended at kt$_{80} \sim 0.35$ towards high $i - W3$.

A possible explanation for these outer contours is the non-linear effect of dust reddening on the colour distribution of quasars (Richards et al. 2001). However, fig. 11 of H17 shows that the typical SED of ERQs is very different from the SED of dust-reddened quasars without the strong C IV line characteristic of core ERQs. Core ERQs have SEDs which are much flatter in the rest-frame UV than suggested by their red $i - W3$ colours, while non-core ERQs
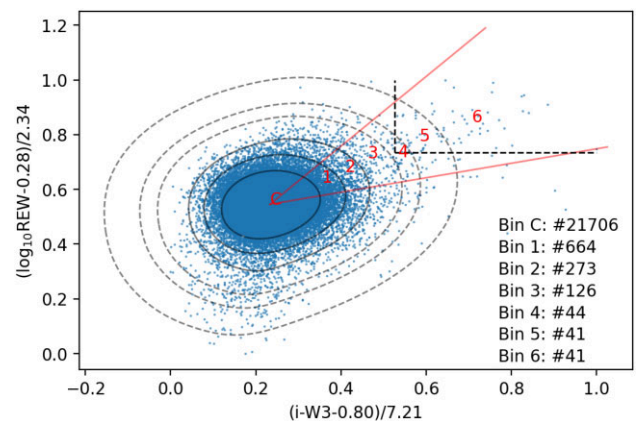
Figure 7. A binned wedge along v_{T1CERQ} towards the T1CERQ sample, with bins defined by density contours. The population of each bin is provided. Bin-C is enclosed by the innermost solid line contour at $0.3\rho_{max}$. 2nd and 3rd contours are at the levels of 0.03 and 0.01 of ρ_{max}. The three outer dashed line contours are $\times 1.35$, $\times 1.55$, and $\times 1.95$ enlarged version of the biggest solid line contour.

exhibit a sharp decline in the near UV with only moderately red colours across the near-IR, similar to type 1 QSOs reddened by dust extinction (H17).
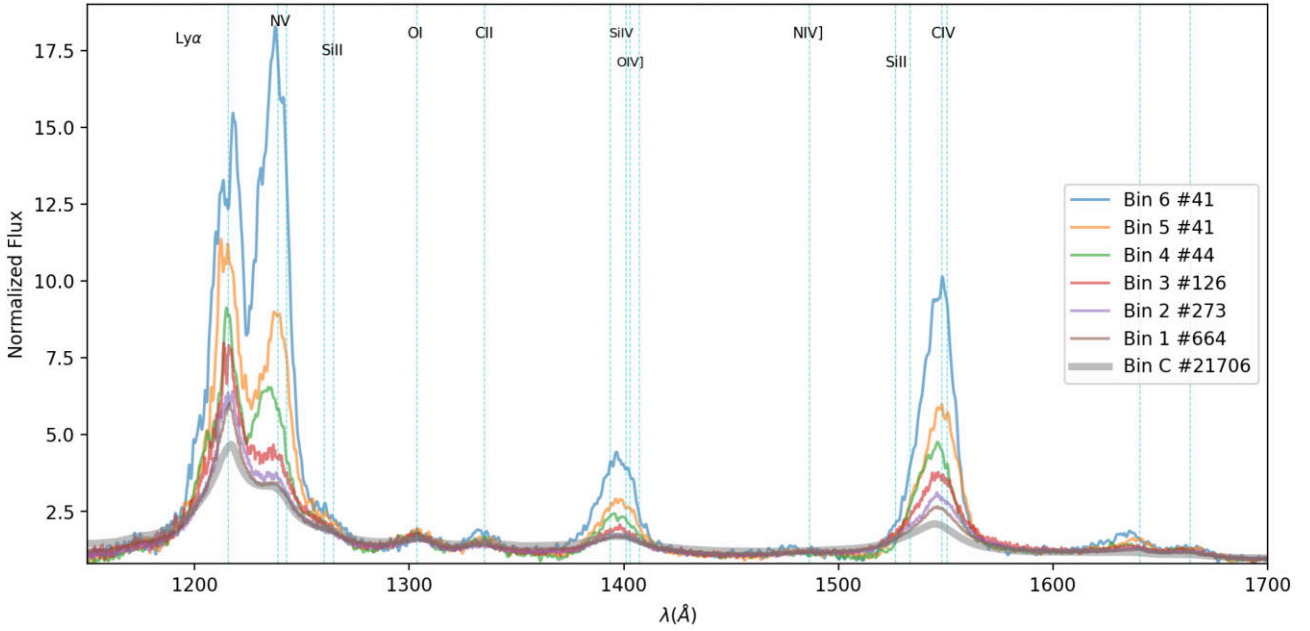
4.2 Median spectra

A second intuitive way to examine exotic line properties is to make median spectra for the T1CERQ sample. Since we have, in Section 3.2, constructed vectors towards the T1CERQ sample, together with geometric cones which contain the T1CERQ quasars. We are now able to bin the cones and make median spectra within these bins and examine how the spectra of T1CERQs change in 2D and 3D parameter space. Median spectra are made by stacking, after the following pre-processing steps:

(1) Shift the observed flux into the quasar's rest frame.
(2) Normalize the spectrum by the median flux between 1680 Å and 1730 Å in the rest frame. This region was chosen as the quasar spectrum is mostly free from significant line features.
(3) Interpolate all fluxes on to a logarithmic grid defined between 800 Å and 3000 Å.

Since visualization is easier in 2D, we perform our first median spectrum analysis by binning in the normalized parameter space of $i - W3$ and $\log_{10}$REW(C IV) along the v_{T1CERQ} direction described in Section 3.2. Fig. 7 shows the 2D wedge towards v_{T1CERQ}, together with the density contours around which we define the bins for our median spectra. The bin boundaries are chosen to bring out specific features of the median spectra. Three inner density contours at $0.3\rho_{max}, 0.1\rho_{max}, 0.03\rho_{max}$ show the shape of the core of the T1LM sample. The three outermost contours scale the contour at $0.03\rho_{max}$ by 1.35, 1.55, and 1.95, as the number of quasars this far from the main locus is too low to accurately estimate density. The choice of the 1st (1.35) and 2nd (1.55) scale factors gives a bin that covers the boundary of T1CERQs suggested by H17 (the lower left-hand corner of REW(C IV) > 100 Å and $i - W3 \geq 4.6$ box in Fig. 7). The 3rd scale factor (1.95) is chosen so that bins 5 and 6 are equally populated.

Fig. 8 and Table 2 show that there is an evolution in the line properties of quasars along v_{T1CERQ}. The median C IV emission line in bins 1 through 3 is symmetric, close to the shape of the median

Figure 8. Median spectra for bins in the direction of $\vec{v}_{T1CERQ}$. The bin number and the number of quasars in each bin are shown. As a reminder, T1CERQs are found in bins 5 and 6 and part of bin 4.

Table 2. List of median physical properties in each bin from Fig. 7. 'C' in the 2nd column denotes the central bin. Column No. shows the number of quasars in each bin. f_{BAL} is the fraction of quasars which contain a visually verified BAL feature near the C IV line. Other columns show the median values and their standard deviations in each bin.

Bin	No.	$i - W3$	REW(C IV)	FWHM(C IV)	kt_{80}(C IV)	N V/C IV	f_{BAL}	Luminosity
C	21706	2.45 ± 0.35	35 ± 10	5400 ± 1500	0.25 ± 0.05	1.54 ± 0.55	0.14	46.83 ± 0.21
1	664	3.37 ± 0.15	55 ± 8	4700 ± 1700	0.24 ± 0.06	1.04 ± 0.46	0.22	46.76 ± 0.19
2	273	3.76 ± 0.18	68 ± 13	4300 ± 1700	0.23 ± 0.07	0.95 ± 0.55	0.29	46.79 ± 0.22
3	126	4.14 ± 0.24	89 ± 22	3900 ± 1500	0.28 ± 0.07	0.99 ± 0.60	0.32	46.86 ± 0.24
4	44	4.65 ± 0.17	92 ± 30	3500 ± 1400	0.35 ± 0.07	1.66 ± 0.78	0.52	47.02 ± 0.28
5	41	5.02 ± 0.26	125 ± 42	3300 ± 1200	0.35 ± 0.06	1.28 ± 0.76	0.29	47.09 ± 0.32
6	41	5.90 ± 0.57	181 ± 80	3100 ± 900	0.36 ± 0.06	1.74 ± 0.58	0.12	47.30 ± 0.29

spectrum in bin C, the main T1LM quasar locus (shown by the thick grey spectrum in Fig. 8).

Table 2 shows that there is a jump in the C IV line kurtosis in bin 4: median kt_{80}(C IV) is 0.28 in bin 3 and 0.35 in bins 4–6. Bin 4 also has an unusually large BAL fraction (0.52). To ensure that these properties are due to the T1CERQ vector and not a function of distance from the quasar locus, we also checked the line properties along $-v_{T1CERQ}$ and confirmed that kt_{80}(C IV) remained similar to bin C, while the BAL fraction dropped to 0.07. We confirmed these trends by making median spectra along vectors both clockwise and anticlockwise of v_{T1CERQ}, again finding that the line kurtosis remained low and confirming that the v_{T1CERQ} direction is unique.

There is also a relatively large jump in N V/C IV from ∼1 in bins 1–3 to 1.66 in bin 4. We found that N V/C IV also increased along $-v_{T1CERQ}$, perhaps indicating that this is not intrinsic to T1CERQs, but in this direction the N V line is weak and the fit is likely to suffer severely from blending with the Lyman-α line.

4.2.1 3D parameter space

Motivated by the success of our 2D analysis, we made median spectra in the 3D normalized parameter space of $i - W3$, $\log_{10}$REW(C IV), and kt_{80}(C IV). The median spectra bins were made within the 3D

cone defined in Section 3.2. The central bins were again defined by density iso-surfaces relative to the maximum density and computed using a KDE as in Section 4.1. The central quasar locus, bin C, was $\rho > 0.5\rho_{max}$. Bin 1 has $\rho = 0.5\rho_{max} - 0.05\rho_{max}$ and bin 2 is $\rho = 0.05\rho_{max} - 0.01\rho_{max}$. As for our 2D binning, we did not use the density iso-surfaces at lower density levels, because the low numbers of spectra in these bins make local density estimates too noisy. Instead, we uniformly enlarged the iso-surface of $0.01\rho_{max}$ by factors of 1.5, 2.1, and 2.5 to make three extra surfaces and used these enlarged surfaces for bins 4 though 6. The expansion factors of 1.5 and 2.1 were chosen so that bin 4 covers H17's boundary for ERQs (the $i - W3 \geq 4.6$ plane in Fig. 9a). The scale factor of 2.5 was chosen so that the last two bins had an equal sized population (23 quasars in each).

All 3D bins are colour coded in Fig. 9(a). The transparent box in Fig. 9(a) shows $i - W3 \geq 4.6$, $\log_{10}$(REW(C IV)) ≥ 100Å, and kt_{80}(C IV)≥ 0.33. Median spectra of these 3D bins are plotted in Fig. 9(b) and Table 3 summarizes the median physical properties in each bin of Fig. 9(a). As for the 2D analysis, the kurtosis increments in each bin from 1 to 3 and saturates at 0.33 in bin 4, suggesting bin 4 is a good candidate for a boundary separating a population of red quasars from the main T1LM sample. N V/C IV is larger in bin 4–6 compared to bin 1 though bin 3. It is again large in bin C, but
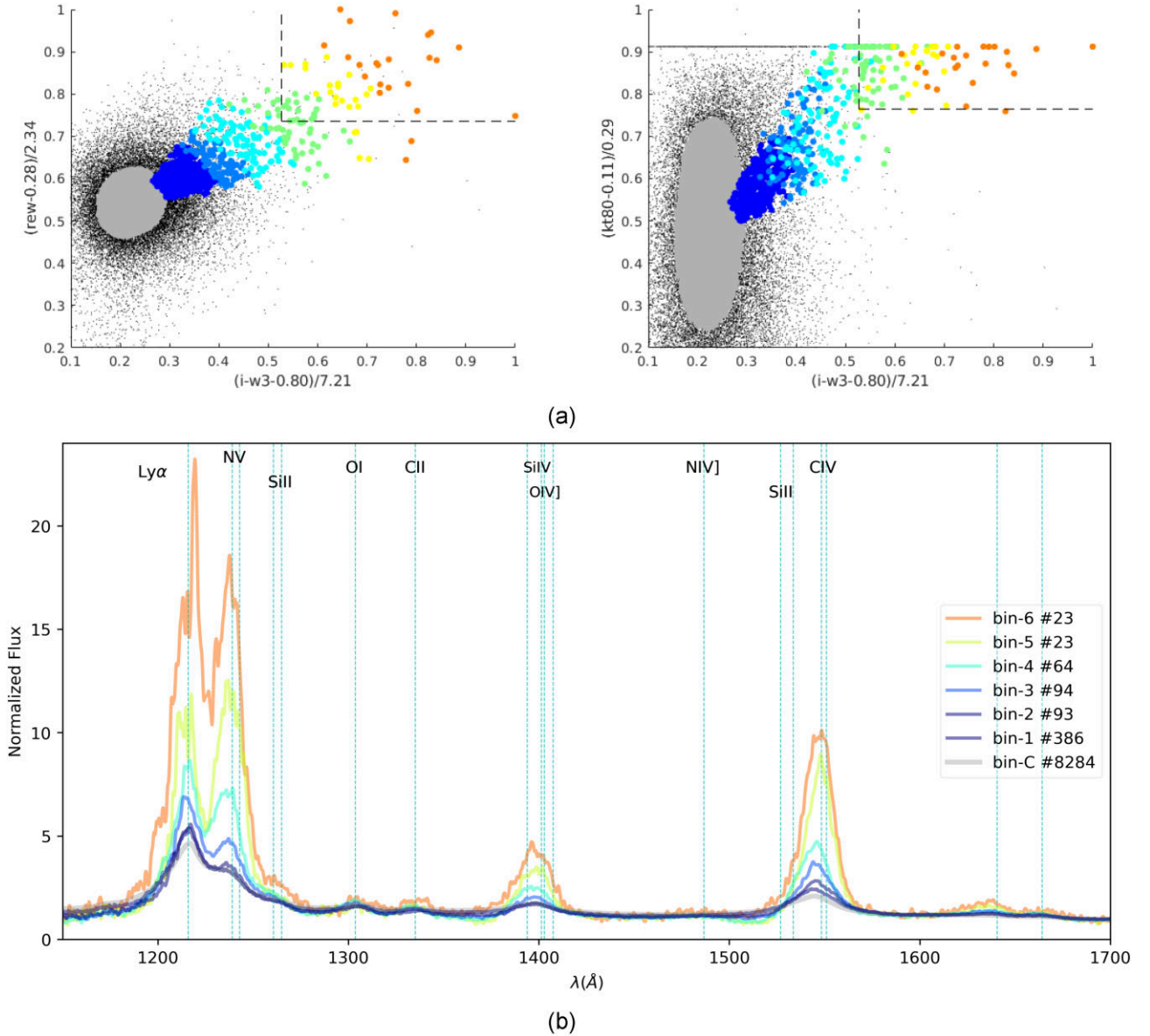
Figure 9. Top (a): 3D bins along a cone around v^{3D}_{T1CERQ}. The central bin is shown by grey points at the centre. Each bin, separated by density iso-surfaces as described in Section 4.2.1, is painted a different colour. The dashed lines show the region of $i - W3 \geq 4.6$, $\log_{10}$REW(C IV)≥ 2, and kt$_{80}$(C IV)≥ 0.33 in the min-max normalized space. Bottom (b): Median spectra for the corresponding coloured objects in each bin of the top panel. Spectra colours for each bin match those in Fig. 9(a).

Table 3. List of median physical properties in the bins of Fig. 9(a). Columns are named as in Table 2.

Bin	No.	$i - W3$	REW(C IV)	FWHM(C IV)	kt$_{80}$(C IV)	N V/C IV	f_{BAL}	Luminosity
C	8284	2.40 ± 0.21	35 ± 6	5600 ± 1300	0.25 ± 0.03	1.60 ± 0.50	0.11	46.85 ± 0.21
1	386	3.10 ± 0.20	47 ± 7	5400 ± 1400	0.28 ± 0.01	1.22 ± 0.44	0.16	46.79 ± 0.18
2	93	3.69 ± 0.21	58 ± 10	4500 ± 1800	0.30 ± 0.02	1.09 ± 0.55	0.30	46.80 ± 0.22
3	94	4.08 ± 0.33	78 ± 18	3900 ± 1600	0.32 ± 0.03	1.29 ± 0.58	0.52	46.91 ± 0.22
4	64	4.74 ± 0.33	93 ± 24	3500 ± 1300	0.36 ± 0.02	1.68 ± 0.72	0.55	47.19 ± 0.27
5	23	5.43 ± 0.35	142 ± 44	3000 ± 1400	0.37 ± 0.01	1.78 ± 0.71	0.30	47.20 ± 0.33
6	23	6.17 ± 0.61	209 ± 92	3500 ± 1100	0.36 ± 0.01	1.74 ± 0.65	0.04	47.49 ± 0.27

this may again be due to blending with Lyman-α. As in 2D, bins 3 and 4 have a high fraction of BALs, although this is not true of bin 6. In general the trends in 3D are similar to those in 2D: this extra parameter, however, will be useful in the next sections.

4.3 Local outlier factor analysis

We showed in Section 4.1 that T1CERQs are an overdensity when compared to other quasars at a comparable distance from the centre
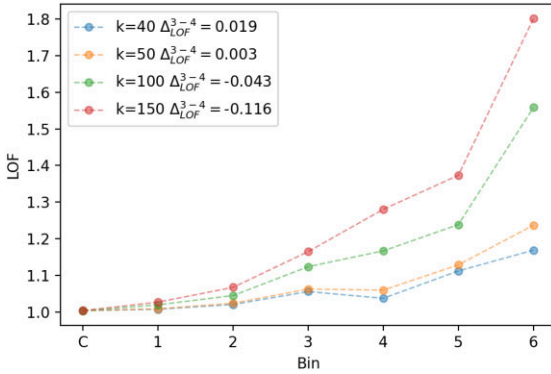
Figure 10. Median LOF score in each bin within the wedge of Fig. 7, along the vector $\boldsymbol{v}_{T1CERQ}$ between the centroid of T1LM and T1CERQ, for different numbers of nearest neighbours (k).

of the main population, and in Section 4.2 we found that there was an increase in kurtosis around the fourth bin from the central T1LM sample. In this section, we quantify the distinctness of T1CERQs from the main T1LM sample using LOF, and examine the 2D and 3D candidate boundaries we found in Section 4.2. As a reminder, in Section 3.3 we showed that a signature of two distinct populations is a dip in the LOF score.

4.3.1 LOF analysis in 2D

We now proceed to analyse the full T1LM sample with LOF along a vector directed towards the T1CERQ, $\boldsymbol{v}_{T1CERQ}$. We use the bins depicted in Fig. 7, and compute a median LOF score in the normalized 2D space of $i - W3$ and $\log_{10}REW(C\,IV)$). There is a (small) dip in the median LOF score for bin 4. This is interestingly consistent with the results from median spectra in Section 4.2, where we saw that bin 4 was also associated with unusual spectral properties. The magnitude of the dip for $k = 40$ is 0.019, which is somewhat less than the 68 per cent uncertainty of the LOF score in our 2D mock data analysis (σ(LOF(bin 3)-LOF(bin 4)) = 0.022). The 2D LOF analysis thus provides indications that the T1CERQs may be a separate population from the main T1LM, but is by no means definitive.

Fig. 10 also shows the dependence of LOF score on neighbour number, k. for $k = 40$, 50, 100, and 150. The LOF score falls from bin 3 to bin 4 when the number of nearest neighbours is 40 or 50, and monotonically increases for $k = 100$ or 150. The LOF score thus suggests that a putative separate population of T1CERQs would have a population between 50 and 100, in good agreement with H17, who identified a population of 72 T1CERQs.

4.3.2 LOF analysis in 3D

We performed a similar LOF analysis in 3D, adding $kt_{80}(C\,IV)$ to $i - W3$, $\log_{10}(REW(C\,IV))$ and using a normalized space. Fig. 11 once again shows a dip in the median LOF score for bin 4. This fall in the LOF score from bin 3 to bin 4 is 0.123 for $k = 70$, significantly larger than the 68 per cent uncertainty of the similar bin in the mock 3D data analysis (σ(LOF(bin 3)-LOF(bin 4)) = 0.025), as Fig. 4 shows. The LOF analysis in 3D is thus evidence that there is a separate population of quasars along the vector to T1CERQs. Interestingly, Fig. 4 shows that the dip in the LOF score persists for $k \leq 150$, suggesting that the separate population may be somewhat larger than that found in H17.

4.4 Selecting T1BERQs in 3D

H17 found a subpopulation of quasars in 2D space, the T1CERQs. Our median spectra and LOF analysis has provided quantitative evidence that this subpopulation is distinct from the general trend of the T1LM sample, especially when viewed in the 3D parameter space of $i - W3$, $\log_{10}(REW(C\,IV))$ and $kt_{80}(C\,IV)$. There are also indications in the LOF score that the subpopulation is moderately larger than the T1CERQ set found by H17. In this section we will design 3D criteria which optimizes the selection of these objects. We call our new subset of quasars Type 1 boxy C IV emission line extremely red quasars (T1BERQs).

The choice of our $kt_{80}(C\,IV)$ parameter space is also motivated by Fig. 1, which suggests that there are a small number of low $kt_{80}(C\,IV)$ objects within the T1CERQ class and that a minimum kt_{80} condition will produce a purer sample. Here we outline our recipe for selecting T1BERQs, summarizing steps introduced in earlier sections:

(1) Normalize the parameter space of ($i - W3$, $\log_{10}(REW(C\,IV))$, $kt_{80}(C\,IV)$) with a Min-Max scaler (discussed in Section 3.2).

(2) Define $\boldsymbol{v}_{T1CERQ}^{3D}$, a vector from the median of the T1LM sample to the median of those points satisfying $i - W3 \geq 4.6$, REW(C IV)≥ 100 Å, and $kt_{80}(C\,IV)\geq 0.33$), in the normalized space.

(3) Find a cone along $\boldsymbol{v}_{T1CERQ}^{3D}$, with a tip located at the median point of T1LM, and an opening angle so that the cone includes all quasars satisfying $i - W3 \geq 4.6$, REW(C IV)≥ 100Å, $kt_{80}(C\,IV)\geq 0.33$.

(4) Using KDE, find density iso-surfaces and bin the cone in the previous step by successive iso-surfaces. One of the bins passes through an initial guess about the boundary of the desired population (here $i - W3 \geq 4.6$, REW(C IV)≥ 100Å, and $kt_{80}(C\,IV)\geq 0.33$).

(5) Calculate the LOF score in each bin.

(6) Find the bin showing a local minimum in LOF score.

(7) Repeat steps (4) to (6) varying the inner and outer boundaries of the candidate bin found in step (6) and find the bin which shows the largest decrease in LOF scores as compared to the neighbour bin located closer to the centre of the T1LM sample.

(8) Find a plane perpendicular to $\boldsymbol{v}_{T1CERQ}^{3D}$ and tangent to the inner boundary of the optimum bin in step (7).

(9) Define the boundaries of the T1BERQs using the common region between the plane of step (8) and the cone of step (3).

Following this procedure, we found all quasars in a cone with a tip at (in our normalized 3D parameter space) (0.23, 0.54, 0.47) and an opening angle of $19.6°$, the same cone as in Section 3.2. The bin with a minimum LOF score was bin 4, as in Section 4.4, and the optimized, adjusted boundary between bins 4 and 3 was expanded by a factor of 1.5 from the bin boundaries of Section 4.2.1. The change in the LOF score across this bin boundary increased moderately to LOF(bin 3)–LOF(bin 4) =0.130 for $k = 70$. Thus, the plane in step (8) of our procedure passes though a point (0.50, 0.72, 0.75) in the normalized space of ($i - W3$, $\log_{10}(REW(C\,IV))$, $kt_{80}(C\,IV)$) with a normal vector of $\hat{\boldsymbol{n}} = (0.64, 0.41, 0.64)$. We thus define T1BERQs by the following inequalities in the 3D normalized parameter space:

$$0.64(i - W3) + 0.41 \log_{10} REW + 0.64kt_{80}(C\,IV) - 1.10 \geq 0. \tag{7}$$

$$\theta \leq 19.6°, \tag{8}$$

where θ is defined in equation (2).

Fig. 12 visualizes the resulting set of quasars, T1BERQs, in 2D projections of the 3D space. Quasars are colour coded to show those which would be selected by both the T1CERQ and the T1BERQ
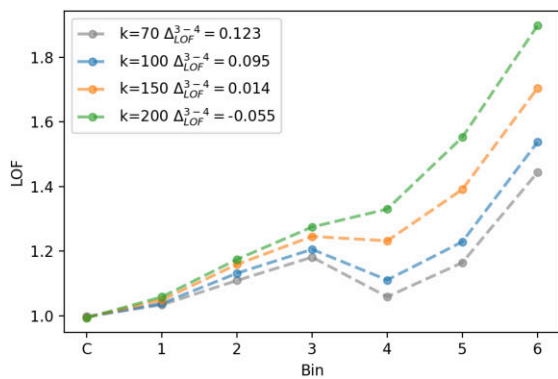
Figure 11. Median LOF score in each 3D bin of Fig. 9a, along the 3D vector v_{T1CERQ}^{3D} between the centroid of T1LM and T1CERQ, for different numbers of nearest neighbours (k).

criteria, by one but not the other, or by neither. Quasars selected by T1CERQ but not T1BERQ (15 quasars) are those with low kt_{80}: they are red and possess strong but not boxy C IV lines. Quasars selected by T1BERQ but not T1CERQ (76 quasars) are those which our local outlier factor selection algorithm judged to be closer to the ERQ subset than the main quasar locus. They are generally somewhat less red than the other ERQs and have weaker C IV lines, but exhibit the same extreme line properties. Overall, the T1BERQ selection produces 133 quasars.

Fig. 13 compares the median spectra of T1BERQs to T1CERQs. As expected given the selection criterion, the T1BERQ sets have higher average kt_{80}, but lower average $i - W3$ and lower REW(C IV). However, they also exhibit the other unusual line properties associated with T1CERQs, to a stronger extent. In particular, the 76 quasars in T1BERQs but not T1CERQs have a high BAL fraction of $f_{BAL} = 0.62$, roughly three times larger than T1CERQs (see Fig. 14 for a clearer comparison). The low kt_{80} quasars which were removed were also those with the lowest BAL fraction. The FWHM of the C IV line is larger in the newly selected T1BERQs, strengthening the general trend shown in Table 3. Finally the N V line is strong, as shown by the high N V/C IV, and visually in the median spectra, where N V strength is comparable to the Lyman-α emission line.

4.5 T1BERQs in *WISE* AGN catalogue

To determine whether our sample of T1BERQs are extremely red only within the SDSS colour selection criteria or are extreme also as a part of other quasar samples, we performed a parallel analysis using the MILLIQUAS catalogue (Flesch 2021). We cross-matched the quasars in MILLIQUAS with the *WISE* AGN catalogue (Assef et al. 2018), as the infrared flux measurements of *WISE* are well-suited to studying red quasars like ERQs (H17). MILLIQUAS is a compendium of extant spectra with a high likelihood of being quasars. As SDSS is the largest spectral survey in existence, most, but not all, spectra in MILLIQUAS come from SDSS. If the colour selection function of SDSS were truncating the ERQ distribution, we would expect that the set of objects in MILLIQUAS but not in SDSS would extend substantially further towards the locus of ERQs in *WISE* colour space.

Fig. 15 shows our selected sample of T1BERQs in the ($w1 - w2$, $w2 - w3$) colour–colour space[5] of the *WISE* catalogue. We

also show our comparison sample (i.e. cross-matched *WISE* AGN and MILLIQUAS quasars that have spectroscopic redshift but are not listed in SDSS). MILLIQUAS does extend the quasar locus moderately towards low $w2 - w3$, but this is in the opposite direction to the T1BERQs. There is no evidence that colour selection effects are skewing our sample. We also show the histogram of T1BERQs in colour–colour space, as well as the histogram of the MILLIQUAS catalogue. These two histograms clearly have separate centres, indicating that even though T1BERQs are originally selected in the SDSS, they are extreme even in the MILLIQUAS catalogue after excluding SDSS quasars.

5 CONCLUSIONS

We have studied the phenomenon of extremely red quasars (ERQs), found by Hamann et al. (2017) (H17) to have red colour, large REW(C IV) and unusual emission line properties, including boxy C IV line profiles a high incidence of blue-shifted BALs, and high [O III] 5007 Å outflow speeds up to 6702 km s^{-1} Perrotta et al. (2019). These properties are consistent with ERQs being consistent with an early dusty stage of quasar–galaxy evolution, where strong quasar-driven outflows provide important feedback to the host galaxies. In this paper, we have used data-driven techniques to understand whether the ERQs, when mapped into spectral parameter space, represent a separate population and, if so, where the boundaries of this population lie.

We applied a kernel density estimation in the space of $i - W3$ colour and C IV rest equivalent width identified by H17 to show that the ERQs produce overdensities. We computed the local outlier factor, previously calibrated on mock data, to assess whether these overdensities could be explained as statistical fluctuations at large distance from the median sampled quasar. The signature of two separate populations, as we showed using mock data, is a dip in the local outlier factor at the boundary between the populations. In 2D there was a dip near the boundary of the ERQs, but it was not strong enough to rule out a statistical fluctuation. We therefore considered higher dimensionality space, adding a third parameter, kt_{80}, defined by H17 as a measurement of line kurtosis or boxiness, and correlated with the presence of an ERQ. In the 3D space of $i - W3$ colour, C IV rest equivalent width and kt_{80}, we found a strong dip in the local outlier factor around the boundaries of ERQs, with a cluster size of 100–150. The dip in the local outlier factor provides evidence that ERQs are connected to a distinct phase of quasar formation, rather than being part of a smooth transition from normal blue quasars to the tail of colour-REW distribution towards redder colours and larger REW(C IV)s.

We refined the selection criteria for T1CERQs, resulting in a new sample of 'boxy' ERQs (T1BERQs). The idea behind these selection criteria is to use line emission properties to better align the boundary of the T1CERQ sample with the onset of the special phase of quasar formation that leads to these exotic quasar properties. To do this, we made use of the 'boxy' shape of the C IV line, defining a boundary in 3D which maximized the depth of the dip in the LOF score. Our final sample defined T1BERQs by the inequalities (7) and (8), which refer to the common region between a plane and a cone obtained by finding the largest minimum of an LOF score in a bin.

There are 15 quasars in the sample of T1CERQs which are not in T1BERQs. Despite having very red colour and extremely strong

---

[5]Unfortunately, the i-band magnitudes for the quasars in our comparison sample are not available; otherwise it would be very illustrative to compare

T1BERQs with our comparison sample in a 3D colour–colour–colour plot of ($i - w3$, $w1 - w2$, $w2 - w3$).
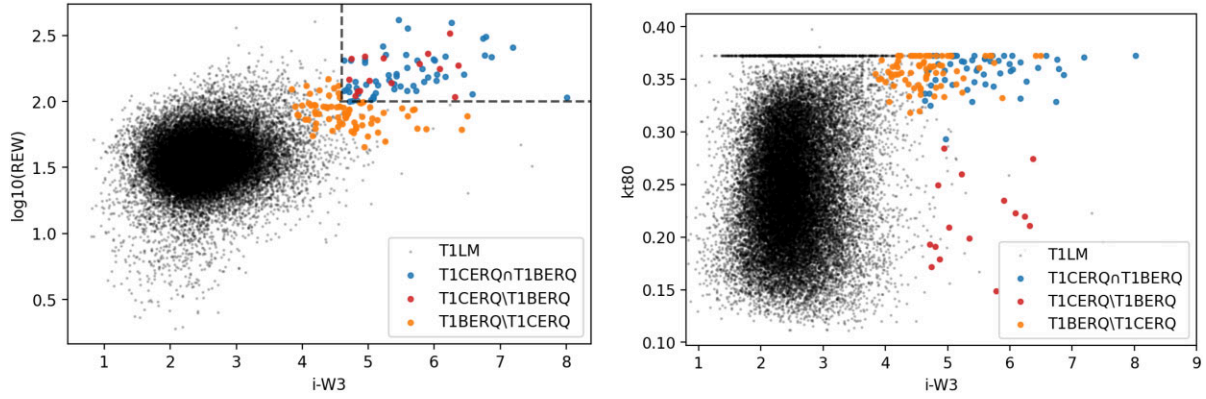
Figure 12. Left-hand panel: Projection of the 3D selection of T1BERQs into $(i - W3, \log_{10}(\text{REW}(\text{C IV}))$ space. Right-hand panel: Projection of the 3D selection of T1BERQs into $(i - W3, \text{kt}_{80}(\text{C IV})$ space. The blue dots belong to the intersection of T1CERQs and T1BERQs. The red dots are T1CERQs which are not T1BERQs. The orange dots are T1BERQs which are not T1CERQs. T1LMs which are not T1CERQS or T1BERQs are shown by black dots.
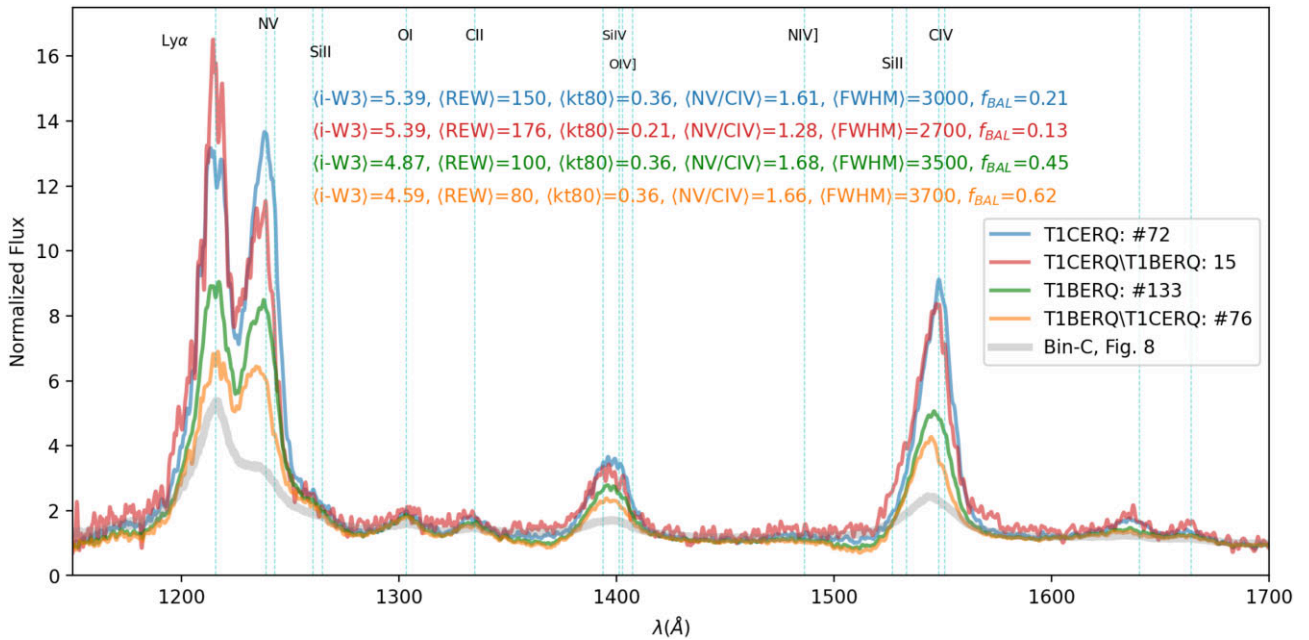

Figure 13. The median spectrum of T1CERQs is shown by the blue curve. The median spectrum of T1CERQs which are not among T1BERQs is plotted with the red curve. The median spectrum of those T1BERQs which are not among T1CERQs is shown in orange. The thick grey curve shows the median spectrum of all quasars in the dense region within bin C in Fig. 9(a).
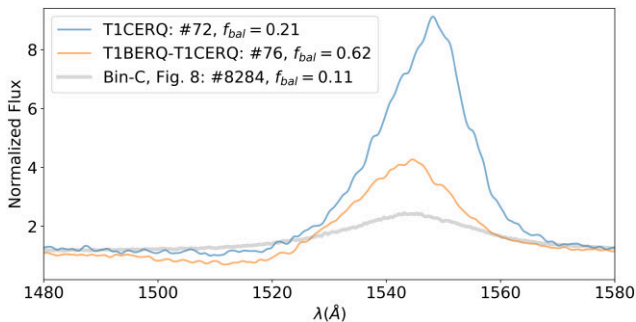
Figure 14. This figure shows a zoom in view of the median spectra around C IV BAL region. Our newly classified objects (i.e. T1BERQs which are not among T1CERQs) are compared to T1CERQs and to the quasars in the central bin of Figs 9(a) and (b). Our 76 newly classified quasars have a higher visually verified BAL fraction.

C IV lines, these quasars have much lower $\text{kt}_{80}(\text{C IV})$ compared to the average T1CERQ and are thus excluded on the basis of their non-boxy line profile. On the other hand, there are 76 quasars which are within the T1BERQ sample, but are not T1CERQs. Selected on the basis of their $\text{kt}_{80}(\text{C IV})$ as well as their red $i - W3$ and high REW(C IV), these quasars have more extreme spectral properties, exclusive of the selection criteria, than the T1LM sample. Our T1BERQ selection criteria produced N V lines which were strong compared to the C IV and Lyman-α, and C IV lines with a greater FWHM than expected for the quasars' colour. The T1BERQs also had a high BAL fraction of $f_{\text{BAL}} = 0.62$, roughly three times larger than T1CERQs. If ERQs are associated with an early dusty stage of quasar formation, we would expect strong metal lines and a high fraction of BAL, associated with a dense accretion disc. The final result of our paper is thus improved selection criteria which produce a purer sample of these interesting objects. This will help us to identify ERQs more efficiently in up-coming large quasar surveys such as the
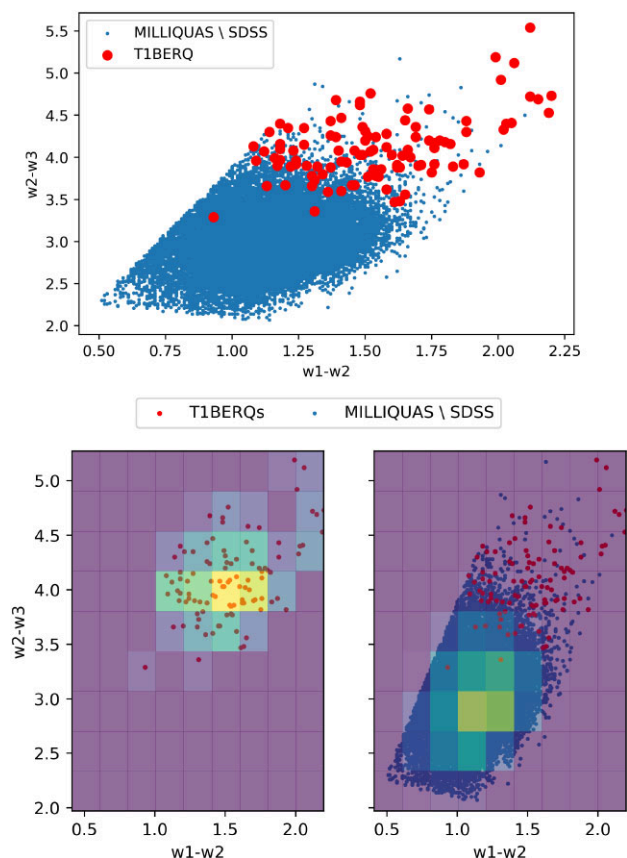
Figure 15. Upper: blue dots are all quasars in the MILLIQUAS catalogue, but not in the SDSS catalogue, with a spectroscopic redshift. The red dots are T1BERQs. Lower left-hand panel: 2D histogram of T1BERQs in this colour space. Lower right-hand panel: 2D histogram of all quasars in the upper panel, with T1BERQs shown as red dots.

Dark Energy Spectroscopic Instrument (DESI) DESI Collaboration (2016) or HETDEX Hill et al. (2008), and select the best targets for follow-up observations investigating quasar and galaxy evolution.

ACKNOWLEDGEMENTS

DATA AVAILABILITY

Our underlying quasar line catalogue is from Hamann et al. (2017) and is available as the Supplemental BOSS Emission Line Catalogue.[6] Our analysis scripts and the sample of T1BERQs as a fits table are publicly available in `GitHub`.[7]

[6]https://datadryad.org/stash/dataset/doi:10.6086/D1H59V
[7]https://github.com/rezamonadi/ExtremelyRedQuasars

REFERENCES

Alexandroff R. et al., 2013, MNRAS, 435, 3306
Antonucci R., 1993, ARA&A, 31, 473
Assef R. J. et al., 2015, ApJ, 804, 27
Assef R. J., Stern D., Noirot G., Jun H. D., Cutri R. M., Eisenhardt P. R. M., 2018, ApJS, 234, 23
Azadi M. et al., 2015, ApJ, 806, 187
Baldwin J. A., 1977, ApJ, 214, 679
Banerji M., Alaghband-Zadeh S., Hewett P. C., McMahon R. G., 2015, MNRAS, 447, 3368
Breunig M. M., Kriegel H. P., Ng R. T., Sander J., 2000, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, New York, p. 93
Calistro Rivera G. et al., 2021, A&A, 649, A102
Dawson K. S. et al., 2013, AJ, 145, 10
DESI Collaboration, 2016, preprint (arXiv:1611.00036)
Eisenstein D. J. et al., 2011, AJ, 142, 72
Ester M., Kriegel H. P., Sander J., Xu X., 1996, Proceeding of "Knowledge Discovery in Databases-1996", 34
Fawcett V. A. et al., 2020, MNRAS, 494, 4802
Flesch E. W., 2021, preprint (arXiv:2105.12985)
Gebhardt K. et al., 2000, ApJ, 539, L13
Glikman E. et al., 2012, ApJ, 757, 51
Glikman E., Simmons B., Mailly M., Schawinski K., Urry C. M., Lacy M., 2015, ApJ, 806, 218
Gültekin K. et al., 2009, ApJ, 698, 198
Hamann F. et al., 2017, MNRAS, 464, 3431 (H17)
Haring N., Rix H. W., 2004, ApJ, 604, L89
Hickox R. C., Alexander D. M., 2018, ARA&A, 56, 625
Hill G. J. et al., 2008, in Kodama T., Yamada T., Aoki K., eds, ASP Conf. Ser. Vol. 399, Panoramic Views of Galaxy Formation and Evolution. Astron. Soc. Pac., San Francisco, p. 115
Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Martini P., Robertson B., Springel V., 2005, ApJ, 630, 705
Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008, ApJS, 175, 356
Ishibashi W., Fabian A. C., 2016, MNRAS, 463, 1291
Kim D., Im M., 2018, A&A, 610, A31
Klindt L., Alexander D. M., Rosario D. J., Lusso E., Fotopoulou S., 2019, MNRAS, 488, 3109
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Kroupa P., Subr L., Jerabkova T., Wang L., 2020, MNRAS, 498, 5652
Netzer H., 2015, Annual Review of Astronomy and Astrophysics, 53, 365
Paris I. et al., 2017, A&A, 597, A79
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Perrotta S. et al., 2019, MNRAS, 488, 4126 (P19)
Planck Collaboration XVI, 2014, A&A, 571, A16
Richards G. T. et al., 2003, AJ, 126, 1131
Richards G. T. et al., 2001, AJ, 121, 2308
Rosario D. J. et al., 2020, MNRAS, 494, 3061
Ross N. P. et al., 2015, MNRAS, 453, 3932
Sanders D. B., Soifer B. T., Elias J. H., Madore B. F., Matthews K., Neugebauer G., Scoville N. Z., 1988, ApJ, 325, 74
Shankar F., Bernardi M., Haiman Z., 2009, ApJ, 694, 867
Simpson J. M. et al., 2014, ApJ, 788, 125
Tremaine S. et al., 2002, ApJ, 574, 740
Tu L., Luo A., Wu F., Zhao Y., 2010, Science China Physics, Mechanics and Astronomy, 53, 1928
Urrutia T., Lacy M., Becker R. H., 2008, ApJ, 674, 80
Urrutia T., Becker R. H., White R. L., Glikman E., Lacy M., Hodge J., Gregg M. D., 2009, ApJ, 698, 1095
Veilleux S. et al., 2009, ApJS, 182, 628
Wei P. et al., 2013, MNRAS, 431, 1800
William H. E. D., Edelsbrunner H., 1984, J. Classif., 1, 7
Zakamska N. L. et al., 2016, MNRAS, 459, 3144

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