Draft version May 2, 2026
Typeset using LaTeX **twocolumn** style in AASTeX631

The limitations of kNN Environment Recovery in DESI DR1: A Negative, but Insightful Result on Spectrum–Density Regression

James Watt[1]

[1] *Machine Learning in Astronomy (AST7939), Spring 2026*

ABSTRACT

Galaxy spectra encode information about stellar populations, star-formation histories, metallicities, and gas excitation conditions – all measures which are known to correlate with a galaxy's environment. In this work, we test whether rest-frame DESI DR1 spectra can be used to recover continuous local environment density using supervised machine learning. A continous DESI DR1 subset was reduced to a common rest-frame wavelength grid, compressed using principal component analysis (PCA), and regressed onto k-nearest-neighbor environmental density labels with histogram-based gradient boosted regression. Environment labels were computed on the the full DESI DR1 zcatalog population, yielding measures of raw k-NN density, redshift-binned overdensity, and inter-scale density ratios. The leading principal components recover physically meaningful galaxy-evolution axes, including age/SFR, continuum shape, metallicity, Balmer absorption, and ionization-state modes. Regression onto the raw $\log n_k$ seems initially successful with test R^2 increasing from 0.32 at $k = 5$ to $\simeq 0.53$ for $k \gtrsim 5000$. However, this reconstruction signal collapses when the target is replaced by either the redshift-normalized overdensity or the density ratios, yielding test $R^2 \simeq 0.03$ for both PCA-scored spectral and full raw spectral inputs. This shows that the apparent reconstructive power seen in $\log n_k$ is dominated by shortcut learning of the survey selection function, $\bar{n}(z)$, instead of the true underlying local clustering. We therefore conclude that for the DESI DR1 survey and within-spectrum k-NN labels, direct spectrum-to-environment regression is not viable for environment reconstruction. We propose that future work should use external or selection-corrected density fields, group or halo labels, explicitly modeled survey selection functions, or perhaps alternative modes of more incisive representational learning.

1. INTRODUCTION

1.1. *Galaxy environment and the cosmic web*

The environment around a galaxy is one of the most important determining factors in its evolution. Since the discovery of the morphology–density relation by A. Dressler (1980), more than four decades of spectroscopic and photometric work have established that galaxies in dense environments, i.e., groups, clusters, filaments, etc., differ greatly from those in less-dense environments in nearly every observable property: morphological type (M. Postman et al. 2005), broadband color (D. W. Hogg et al. 2003), star formation rate (I. Lewis et al. 2002; P. L. Gómez et al. 2003), stellar age, metallicity, and the strength of stellar absorption features (G. Kauffmann et al. 2004; M. L. Balogh et al. 1999). These properties are all tied to a galaxies environment because primary environmental effects like gas stripping, mergers, harassment, or quenched star formation produce effects across nearly all of a gaalxy's spectra.

Under the modern ΛCDM model, this dependence is categorized by the "cosmic web": a network of voids, sheets, filaments, and nodes. The cosmic web was initially predicted by J. R. Bond et al. (1996) and has now been mapped in great detail by N-body simulations (O. Hahn et al. 2007; M. A. Aragón-Calvo et al. 2010; M. Cautun et al. 2014) and large spectroscopic surveys (M. Tegmark et al. 2004). These simulations and surveys revealed that the color, mass function, and star formation history of a galaxy is highly dependent on its position within the web (Y.-j. Peng et al. 2010; M. C. Cooper et al. 2008; O. Cucciati et al. 2006). For our purposes, this implies that the spectrum of a galaxy should encode information about not only its intrinsic stellar population but the environmental history around it, and that a capable model should be able to infer the latter from the former.

1.2. *The DESI Year 1 spectroscopic sample*

The Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016), a ground-based multi-fiber spectrograph, is creating the largest spectroscopic galaxy survey to date. In just its first year (Data Release 1) (DR1; DESI Collaboration 2024), they delivered

13,049,402 useful galaxy-class spectra across roughly 14,000 deg^2 of sky, with the main/dark survey program spanning ∼13,494 `HEALPix` pixels (NSIDE = 64, nested) and covering $0 \lesssim z \lesssim 1.6$. The DESI survey contains fully co-added, flux-calibrated spectrum from 3600 to 9824 Å at a resolving power of $R \sim 2000$–5000 across three spectrograph arms for each target. This observing range samples many rest-frame features from the near-UV to the red Ca II triplet across most of the large scale structure (LSS) relevant redshift range. DESI's depth, footprint, and relatively uniform target selection make it uniquely advantageous over its predecessors (M. A. Strauss et al. 2002; M. Colless et al. 2001) as it finally has the number of samples needed to test environment-dependent spectral signatures across the intermediate redshift range.

1.3. *An understudied dimension of DR1*

The most insightful scientific results of DR1 to date have been those on baryon acoustic oscillations (DESI Collaboration et al. 2024), growth-of-structure measurements, and constraints on the dark energy density parameter. Environment analyses of the survey have likewise focused on the observable spatial distribution of galaxies: (P. M. Sutter et al. 2015), cosmic-web morphological classifiers (M. Cautun et al. 2014; O. Hahn et al. 2007; M. A. Aragón-Calvo 2019), and two-point clustering statistics. In these analyses, the spectrum serves only to give a redshift estimate and the actual spectral information is discarded.

The complementary question, whether the rest-frame spectrum of a galaxy carries inferable information about a galaxy's local environmental density, has not, to our knowledge, been investigated in the DESI data. Prior work on this question has focused almost exclusively on classical analyses: a small number of hand-engineered indices (color, $D_n(4000)$, EW(Hα), [O II]λ3727, etc.) correlated against environmental classifications in stratified samples (G. Kauffmann et al. 2004; M. L. Balogh et al. 1999; M. C. Cooper et al. 2008). Modern machine-learning approaches to galaxy environment analysis have primarily operated in either photometric regimes, in N-body simulated cosmic-web classification (M. A. Aragón-Calvo 2019), or as unsupervised compressions of spectra without an environmental target (S. K. N. Portillo et al. 2020; P. Melchior et al. 2023). Notably, the direct supervised regression of full rest-frame spectra onto continuous environmental density appears untested to our knowledge.

This approach represents both a methodological and a scientific gap. Methodologically, modern dimensionality reduction and gradient boosted tree methods enable multivariate inference from entire spectra without the strong assumptions that hand-engineered indices bring. Scientifically, the amount of environmental information a single galaxy spectrum could carry is itself a constraint on the coupling between galaxy evolution and the cosmic web, and, if calibrated, could allow single spectra to be probes for galaxy environments.

1.4. *Statistical learning for spectrum–environment inference*

The main method used to relate galaxy spectra to environmental properties have been classical and index-based. Spectra are commonly reduced to a small number of scalar measures–$D_n(4000)$, EW(Hα), EW([O II]), broadband $u - r$ color, Lick absorption indices– and each scalar is correlated against environmental classifiers (e.g., void vs. wall or k-NN density). This approach has been very productive in establishing the morphology-density and color-density relations (A. Dressler 1980; I. Lewis et al. 2002; G. Kauffmann et al. 2004); however, it carries two assumptions that are hard to ignore. First, the choice of which spectral features are investigated is fixed in advance and thus any environmental information that is outside the standard set of indices is invisible to such methods. Second, statistical inference methods are commonly preformed at a survey-wide scale which may limit the resolution of any per-galaxy environmental conclusions.

Modern machine-learning methods improves greatly on these issues but have not yet been applied to this specific question. Principal Component Analysis (PCA) on galaxy spectra has a long history of unsupervised spectral compression (A. J. Connolly et al. 1995; C. W. Yip et al. 2004) and persists as a powerful tool for spectral classification. On the regression side, supervised methods like gradient–boosted decision trees (J. H. Friedman 2001; G. Ke et al. 2017), random forests, and deep regression networks have been widely applied to photometric redshift estimation, galaxy classification, and morphological inference. However, their use to predict large-scale environmental density directly from rest-frame spectra at scale is largely untested. Cosmic web classifiers built on N-body simulations (M. A. Aragón-Calvo 2019) provide an encouraging precedent but rely on particle positions, not observed spectra.

In this work, we investigate PCA as a linear, interpretable basis for rest-frame spectra, and Gradient Booting regression (GBR) (F. Pedregosa et al. 2011) onto continuous environmental quantities. PCA imposes no feature engineering as the bases are learned from the data, allowing the leading components to be interpreted post hoc against cataloged physical indices to confirm

they contain real galaxy-evolution structure. GBR is an ensemble decision tree method that is well adapted to handle non-parametric and non-linear regression and can directly be applied to multiple target regression–in this case onto multiple k-NN neighbor density scales. PCA and GBR have been specifically selected as a viability test, as if they cannot recover environmental density from DR1 spectra, more elaborate methods (deep learning or other statistical ML methods) are unlikely to do so either.

1.5. *This work*

This work asks the questions *can supervised regression of DESI DR1 rest-frame galayy spectra recover the local k-nearestneighbor environmental density?* As we will show, the answer thus far for the DR1 main/dark sample under its current sampling is likely no for these methods–but the way it fails is informative. The remainder of this paper details our findings.

2. METHODS

The methodology we use exists in four stages: (i) acquisition, selection, and reduction of DR1 spectra (§2.1); (ii) PCA feature extraction and correlational feature interpretation (§2.2); (iii) building k-NN density, density ratios, and z-normalized overdensity labels from the full DESI zcatalog (§2.3); and (iv) the supervised regression of the spectral features onto the environmental labels via gradient boosted regression (§2.4). §2.5 lists the outside dependencies used to implement this pipeline.

2.1. *Data acquisition, selection, and reduction*

We selected a contiguous subset of the DESI survey centered on (RA, Dec) = (150°, 0°), selected for its high sampling density. The footprint is made up of 500 `HEALPix` pixels (NSIDE = 64, nested) within 12° radius around the central pixel (HEALPix index 27238). After removing pixel which didn't contribute sources after reduction, 498 pixels contribute spectra to the sample. For each pixel, we obtained the DR1 coadded spectrum file `coadd-main-dark-{pixel}.fits`. These each contained per-fiber flux, inverse variance, and bitmask arrays for the three DESI spectrogrpah arms (B, R, Z; covering 3600–5800 Å, 5760–7620 Å, and 7520–9824 Å respectively). Each file was also accompanied with metadata in the `FIBERMAP` HDU where we retrieved `TARGETID`, `TARGET_RA`, `TARGET_DEC`, and `COADD_FIBERSTATUS`. Additionally, the redshift and its quality were retrieved for each target from their matching entry in the pre-computed redshift catalog `zall-pix-iron.fits` (DESI DR1 Release).

Targets in the subset were filtered by the following conditions: `SPECTYPE` = `GALAXY`; (ii) `ZWARN` = 0 (redshift failure flag); (iii) `COADD_FIBERSTATUS` = 0 (fiber degredation indicator); and (iv) measured spectroscopic redshift in $0.05 \leq z \leq 0.40$. This redshift range was chosen because it contained the densest and most uniform, per-redshift bin LSS sampling (Fig. 2). The lower cut was selected to ensure selected galaxies rest-frame grid falls above 1500 Å (all galaxies retain rest-frame Hα/[N II] in band). The upper cut was selected to avoid the sparser sampeling of galaxies in the higher redshift regime (observations in that regime become emission line dominated and SNR-limited). Lastly, only entries flagged `ZCAT_PRIMARY` = True are kept, yielding the single best spectroscopic measurement for each target.

For each fiber that survives the filters above, we read `B_FLUX`, `R_FLUX`, `Z_FLUX` alongside their inverse-variance arrays and bitmasks. Any pixels with `IVAR` ≤ 0 or non-zero `MASK` were flagged as bad and assigned $f_\lambda = 0$, IVAR $= 0$ before further processing. Each arm was interpolated linearly onto a observed frame, log-space grid of 4096 bins covering 3600–9824 Å. The three arms are combined by a inverse-variance weighted average of their overlapping regions($\sim$ 5660–5930 Å for B/R and $\sim$ 7470–7720 Å for R/Z). This follows,

$$f_\lambda^{\text{merged}} = \frac{\sum_{a\in\{B,R,Z\}} \text{IVAR}_\lambda^{(a)}\, f_\lambda^{(a)}}{\sum_{a\in\{B,R,Z\}} \text{IVAR}_\lambda^{(a)}}, \tag{1}$$

with $\text{IVAR}_\lambda^{\text{merged}}$ defined by the sum.

To compute the quality of each spectra, a per-galaxy median signal-to-noise (SNR) ratio is calculated following,

$$\text{SNR} = \text{median}\left[f_\lambda^{\text{merged}} \sqrt{\text{IVAR}_\lambda^{\text{merged}}} \right] \tag{2}$$

for pixels with IVAR > 0. For PCA fitting, only high-quality spectra satisfying SNR ≥ 8 were kept to encourage principal components to encode astrophysiccal properties rather than noise (§2.2). The SNR ≥ 8 cut-off was chosen to balance the visibility of spectral features (the 4000 Å break, the Balmer series,[O II] λ3727, the Hα/[N II] complex, and the Mg/Fe metal-absorption series) while ensuring a large enough sample to maintain the statistical distribution of the target population. A larger regression sample was constructed by relaxing the cut to SNR ≥ 1, providing a more complete target set for the regressor to learn on (§2.4). This separation of PCA and GBR subsets was preformed to ensure the representational training was done on clean, high-quality data before further analysis was done on the larger set of transformed spectra.

Earlier reduction variants explored alternative SNR thresholds (SNR > 1, > 3, > 5, > 10) and we report

their effect on the regressor performance in §3.6. In addition to SNR rejection, we also reject spectra in which more than 40% of valid rest-frame pixel have $f < 0$ and whose per-bin median is not positive (both indicative of poor quality sky subtraction).

For each of the merged surviving spectra, we shift the galaxies into the rest frame following $\lambda_{\rm rest} = \lambda_{\rm obs}/(1+z)$ and re-interpolate them onto a common rest-frame log-space grid of 4096 bins covering 1500–7000 Å. The rest frame range was selected to include both [O II] λ3727 and the Hα/[N II] complex. Additionally, any pixels that fell outside the observed frame coverage were set to zero and tracked through subsequent steps.

We normalize each rest-frame spectrum by the median of the valid pixels,$f \to f/\mathrm{median}(f \,|\, \mathrm{IVAR} > 0)$, removing the variability in absolute flux (a function of SNR) across spectra and isolating the real spectral shape. To determine the common rest-frame coverage range we find the intersection of the non-zero ranges across all surviving spectra and then keep only columns containing valid signal. The final dimensionality after clipping in $M = 1844$ rest-frame pixels. Finally, each of the final wavelength bins is independently rescaled following standard feature scaling onto [0,1] (3.

$$x_{i,j} \;\to\; \frac{x_{i,j} - \min_{i'} x_{i',j}}{\max_{i'} x_{i',j} - \min_{i'} x_{i',j}}, \qquad (3)$$

where i is the galaxy and j is the wavelength. The per-bin minima and maxima are saved alonside the data to allow the data to be inverted for future analysis. This per-bin scaling is critical for PCA as without it leading eigenvectors unequally weight larger amplitude flux values, masking finer-scale spectral structure.

Earlier iteratiosn of the pipelie applied an archsinh stretch or $c-$tanh compression to the rest-frame spectra ($f' = \mathrm{arcsinh}(f)/\mathrm{arcsinh}(5)$, clipped to $[-1, 1]$) in order to controllably weight emission/absorbtion or continuum and to support the activation frunction of an abandoned deep learning approach. We ultimately abandonded these reductions in favor of the standard per-bin min–max feature scaling since PCA traditionally behaves best on linearly scaled feature vectors. Our empirical obsevrations are consistent with this as the transofmration left a non-zero continuum floor of ~ 0.38 that biased the leading loadings, and the tanh compression interacted poorly with the per-bin min-max step for spectra with low SNR. Our final pipeline thus relied on simply the linear per-bin min-max standardization alone.

Our final reduction yields an M-dimensional rest-frame spectrum on a common wavelenth grid alongside a metadata table containing target identifiers, RA and DEC coordinates, redshift, redshift quality, median SNR, and number of valid rest-frame pixels. After all of the selection filters, the PCA training subset contains 2,621 galaxies from 426 `HEALPix` pixels with SNR $\in$ [8.0, 82.7], and the regression subset contains $N = 54{,}914$ galaxies drawn from 498 pixels. An example of the rest-frame shift step is shown in Fig. 1.

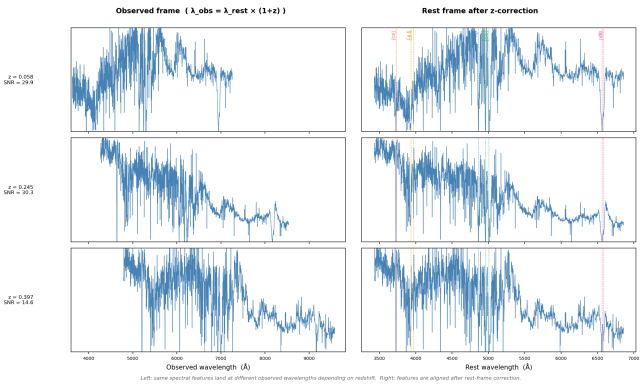


Figure 1. Rest-frame correction step in the reduction pipeline. Left: representative spectrum samples were selected at $z \simeq 0.06$, 0.25, 0.40 and plotted on the full DESI observed wavelength range. Right: rest-frame corrected spectra plotted with notable emission and absorption features identified.

2.2. *Spectral feature extraction*

We compressed each M-dimensional rest-frame spectrum into a low-dimensional score vector via PCA. PCA on galaxy spectra has a long history as an interpretable, assumption-free basis (A. J. Connolly et al. 1995; C. W. Yip et al. 2004) and we use an implmentation in F. Pedregosa et al. (2011) (`sklearn.decomposition.PCA`). The basis was fitted on only the high-SNR trainign subset where the SNR is sufficient to ensure leading components encode real astrophysical variation instead of noise modes. We searched the n-components hyperparameter space up to 95% of explained variance and kept the leading 40 components for use in the regression. The choice to keep the first 40 components was result of analyzing the explained-variance scree plots and identifying where the additional components bring diminishing returns (new components explain $\lesssim 0.1\,\%$) and low physical importance.

The fitted PCA model and its rest-frame wavelength grid were saved and then used to transform the larger SNR ≥ 1 regression sample. When the regression sample had different rest-frame coverage than the PCA training sample, columns were matched by wavelength to within 0.5 Å, and any training-grid bins that were in the application sample were filled with the PCA train-

ing mean (replacing with the training set mean ensures unmatched bins contribute zero to the PCA scores).

PCA always results in an orthonormal basis that doesn't inherently convey qualitative information about the different principlan components. For this reason, we verify that leading components are encoding true galaxy-evolution structre (rather than spectral noise) by computing the literature-defined spectral indices on the rest-frame reduced spectra and correlating them against each PC score. These indices fall into four families:

***Stellar absorption (Lick) indices*:** $H\delta_A$, $H\gamma_A$, G4300, Mg b, Fe 5270, Fe 5335, Na D, and Ca II K, measured as the continuum-relative equivalent widths (EW) over the bandpasses of G. Worthey (1994) and G. Worthey & D. L. Ottaviani (1997). These are sensitive to and act as a indicator of stellar population age and metallicity. Positive equivalent widths correspond to absorption features.

***The 4000 Å break*:** $D_n(4000)$, the narrow-band ratio of f_ν in 4000–4100 Å to 3850–3950 Å (M. L. Balogh et al. 1999), and the wider $D(4000)$ from G. Bruzual A. (1983). These both indicate the integrated stellar age through the build-up of metal absorption lines on the blue side fo the break (old passive populations show strong breaks, $D_n(4000) \gtrsim 1.7$, and young populations show weak, $D_n(4000) \sim 1.2$).

***Emission-line EW and integrated line fluxes*:** $H\alpha$, $H\beta$, [O III] λ5007, [O II] λ3727, and [N II] λ6584 were measured against the linearly-interpolated pseudo-continua (negative sign denote emission). These are indicators of the ionized-gas star formation rate, ionization parameter, and the gas-phase metallicity.

***Composite ionization-state diagnostics*:** $\log_{10}([O\,III]/H\beta)$ and $\log_{10}([N\,II]/H\alpha)$ following J. A. Baldwin et al. (1981); the O3N2 ratio $\log[O\,III/H\beta] - \log[N\,II/H\alpha]$ and the gas-phase oxygen abundance $12 + \log_{10}(O/H)$ via the O3N2 calibration of M. Pettini & B. E. J. Pagel (2004).

Aside from the well documented spectral indices, we also fit a continuum slope α between 4500 and 6500 Å(with $f_\lambda \propto \lambda^\alpha$) as a rough proxy for dust reddening and continuum color.

For each pair of PC and the spectral property we compute both Pearson and Spearman rank correlations. The Pearson coefficient is reported as the main indicator for linear dependence and the Spearman value adds robustness against heavy-tailed distributions. We discuss farther the sign and ranking patterns of these correlations for the physical interpretation of each leading principal component in §3.2.

2.3. *Environmental labels*

Environment labels are computed from the entire DESI DR1 zcatalog, an intentionally broader sample than the PCA or GBR training subsets. Specifically, we selected the entire `ZCAT_PRIMARY` galaxy subset of the DESI DR1 zcatalog within our angular footprint, keeping only targets satisfying `SPECTYPE` = `GALAXY`, `ZWARN` = 0 and entries with z in $0.05 \leq z \leq 1.6$. This yielded $N_{\rm ref}$ = 1,033,722 galaxies within our working footprint which were used to calculate the environmental labels. We decouple this reference population from the PCA or GBR sample intentionally, since building a k-d tree from the SNR-cut spectral sample along introduces a selection bias (pollutes the environmental labels with our selection geometry). For this reason, using the entire zcatalog is the cleanest way to approximate environmental parameters of the true galaxy field.

For every catalog galaxy we compute the comoving distance $\chi(z)$ under ΛCDM cosmology with density parameters $\Omega_m = 0.315$, $\Omega_\Lambda = 0.685$, $h = 0.674$. $\chi(z)$ is calculated by the numerical integration of,

$$\chi(z) = \frac{c}{H_0} \int_0^z \frac{\mathrm{d}z'}{\sqrt{\Omega_m (1+z')^3 + \Omega_\Lambda}}. \tag{4}$$

Since the DESI-BAO+SN poterior chains found evidence for a non-constant dark energy equation of state, the environmental label generation pipeline was build to take w_0, w_a from CPL cosmology as a parameter, however, initial results yielded $\Delta\chi(z)$ values on the order of $< 3 \times 10^{-3}$ across $0 < z < 1$–far below the errors from spatial resolution. For this reason, we use ΛCDM through out the remainder of this investigation, acknowledging that other cosmologies may be a intriguing future line of inquiry.

Each galaxy is then projected to its 3D Cartesian position,

$$\mathbf{r} = \chi(z)\big(\cos\delta\cos\alpha,\ \cos\delta\sin\alpha,\ \sin\delta\big), \tag{5}$$

where α is right ascension in radians, δ is declination in radians, and distance has units Mpc h^{-1} or in physical Mpc for $h = 0.674$.

Following on, a balanced k-d tree is constructed from the aforementioned reference Cartesian positions using `scipy.spatial.cKDTree`. For each galaxy in the regression sample, we then query the tree for its kth nearest neighbor for $k \in$

$\{5, 10, 20, 50, 100, 500, 1000, 5000, 10000\}$, returning the Euclidean distance d_k to that neighbor. The k-NN density estimator follows,

$$n_k = \frac{k}{\frac{4}{3}\pi\, d_k^3} \quad [\text{galaxies}\,\text{Mpc}^{-3}], \tag{6}$$

This is the standard non-parametric estimate of local galaxy number density at some scale set by d_k (e.g. M. C. Cooper et al. 2008). In this estimate, smaller k samples the sub-Mpc to few-Mpc structures like groups and halos, whereas larger k samples ten-of-Mpc structure like filaments or sheets. By reporting performance across a wide range of k, we test which physical length scale, if any, can be recovered from the spectra.

In addition to the straight forward density measurements, we compute the ratios between every pair of density scales. More precisely, for every ordered pair (k_a, k_b) with $k_a < k_b$, we calculate the ratio,

$$R_{k_a/k_b} = n_{k_a}/n_{k_b} = (k_a/k_b)\left(d_{k_b}/d_{k_a}\right)^3. \tag{7}$$

Density ratios.—For every ordered pair (k_a, k_b) with $k_a < k_b$ we additionally form the ratio

$$R_{k_a/k_b} = n_{k_a}/n_{k_b} = (k_a/k_b)\left(d_{k_b}/d_{k_a}\right)^3. \tag{8}$$

By combining k at different scales in R_{k_a/k_b}, we, by construction, rely only on neighbor distances to the target galaxy, and therefore avoid any independent scale factors or gradients that exist at both scales (i.e., an attempt to combat sampling inconsistencies).

Within any magnitude- and fiber-completeness-limited spectroscopic survey, the observed local density at any scale can be described as,

$$n_k(z, \mathbf{x}) \;\approx\; \bar{n}(z)\left(1 + \delta_k(\mathbf{x})\right), \tag{9}$$

Where $\bar{n}(z)$ is the number density of observed galaxies at some redshift z (primarily a results of the apparent magnitude limit, redshift dependent target selection, and fiber assignment efficiency, not local clustering). $\delta_k(\mathbf{x})$ is then the residual local overdensity at scale k. In surveys with a non-flat selection function (most are non-flat), $\bar{n}(z)$ varies by orders of magnitude across the redshift range. This variability means that a model often begins to learns a density-redshift trend from the spectrum (since redshift imprints on nearly all spectra features even when rest-frame corrected), therefore explaining a substantial amount of $\log n_k$ and *missing* any genuinely local information. Taking the logarithm and subtracting the conditional redshift median adresses this factor analytically. We define the redshift-binned log-overdensity as,

$$\begin{aligned}\log_{10}\delta_k(\mathbf{x}, z) \equiv \log_{10} n_k(\mathbf{x}, z) \\ - \text{median}_{z' \in B(z)}\left[\log_{10} n_k(\mathbf{x}', z')\right]\end{aligned} \tag{10}$$

where $B(z)$ is a thin redshift bin containing z. Substituting Eq. (9) and using median$\big[\log(1+\delta_k)\big] \approx 0$ for a sample with median zero overdensity yields

$$\begin{aligned}\log_{10}\delta_k &\approx \log_{10}\big[\bar{n}(z)(1+\delta_k(\mathbf{x}))\big] - \log_{10}\bar{n}(z) \\ &= \log_{10}\big(1+\delta_k(\mathbf{x})\big).\end{aligned} \tag{11}$$

In this, the redshift-binned $\log_{10}\delta$ keeps only the local overdensity component, analytically removing any survey selection bias that may be present. We binned the regression sample into 30 equal count quantile bins of z. Moreover, we use the median of $\log_{10} n_k$ in each bin to define the subtraction in Eq. (10). This construction is the central diagnostic environmental label we present in this work and allows us differentiate learning the survey sampling instead of the galaxy field itself (high n_k performance and low $\log_{10}\delta$ performance is indicative of such).

2.4. *Regression and interpretation*

The supervised regressor we use in this work is the histogram-based gradient bossted decision tree (HGBR) as implmented by `sklearn.ensemble.HistGradientBoostingRegressor` (G. Ke et al. 2017). HGBR wsa selected for three reasons: it is non-parametric meaning it can capture non-linear and non-monotonic responses inherently; it is interally robust to feature scaling and mixed inputs (i.e., the histogram binning); and it is easily scalable to order of magnitude greater than our $\sim 5 \times 10^4$-row training set without problem. The hyperparameters of the HGBR remained at the defaults of`max_iter` = 500 boosting rounds, `learning_rate` = 0.05, `max_depth` = 8, `min_samples_leaf` = 20, and L2-regularization $\lambda = 0.0$. The nature of the failure mode discussed in the later part of this work highlights why we did not pursue exhaustive hyperparameter optimizaiton.

Becuase we are evaluating the predictive performance simultaneously across all scales, k, the regressor was wrapped in `sklearn.multioutput.MultiOutputRegressor`, which fits one HGBR model per target column efficiently. The PCA-score features are identical across target columns, and only the response vector differs between the environment label types outlined in §2.3.

Each regression sample is randomly split into 80% training and 20% test sets with some fixed random seed. Utilizing these sets, we report the predictive performance of a model as the coefficient of determination following,

$$R^2 \;=\; 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{12}$$

on the test set separately for each target. In addition to the test R^2, we report the training as the train-test gap itself is a identifier of overfitting.

For the trained models, we also compute the permutation feature importance with `sklearn.inspection.permutation_importance`, which measures the drop in test-set R^2 when a single feature column is randomly shuffled and the model retested. For each importance, we average over $n = 10$ shuffles of the column. We elected to use permutation importance over the native gain-based importance available in the HGBR as it is only computed on the test set and thus cannot reveal any training set biases. For the multi-output regressor, per-target importances were computed and kept as $(n_{\text{features}} \times n_{\text{targets}})$ matrix I. We then L_1-normalize each target column, $\hat{I}_{f,t} = |I_{f,t}| / \sum_{f'} |I_{f',t}|$, so the per-target importance share the same sum and can be compared across targets that have different predictive scales.

2.5. *Software*

The reduction, feature-extraction, and regression pipelines were all written in `Python` (Python Software Foundation 2024), with `NumPy` (C. R. Harris et al. 2020). `SciPy` (P. Virtanen et al. 2020) was used for KD-trees.`Astropy` (Astropy Collaboration 2022) was used for fits operations, sky pixel and coordinate transformations, and cosmological distance integration.`scikit-learn` (F. Pedregosa et al. 2011) was used for PCA, gradient-boosted regression, multi-output wrapping, and permutation importance, and `pandas` (W. McKinney 2010) was used for catalog joins. `Matplotlib` (J. D. Hunter 2007) was used for figures. Lastly, all analyses were run on the UNiversity of Florida HiPerGator supercomputing cluster.

3. RESULTS

3.1. *Reduced sample and the $n_k(z)$ trend*

The regression sample, `SNR` ≥ 1, after the reduction contained $N = 56{,}068$ galaxies in $0.05 \leq z \leq 0.40$ drawn from 498 `HEALPix` pixels around $(\text{RA}, \text{Dec}) = (150°, 0°)$ (Fig. 2, bottom). The redshift values of this sample were roughly even in distribution with peaks near $z \sim 0.10$ and $z \sim 0.38$ with a median $z = 0.185$. The apparent bimodality of the redshift distribution reflects the contributions from both the bright galaxy and luminous red galaxy populations in the DESI survey. The per-pixel target density is high and reasonably uniform across the footprint of the subset with a median sources per `HEALPix` pixel of 124. The higher snr cutoff, PCA training subset (SNR ≥ 8, $N = 2{,}621$ in 426 pixels; Fig. 2, top) covers the bright tail of the same footprint. It shows the same low z peak at $z = 0.095$ and is centered around $z \sim 0.09$ and has a comparably low median sources per pixel of 5.

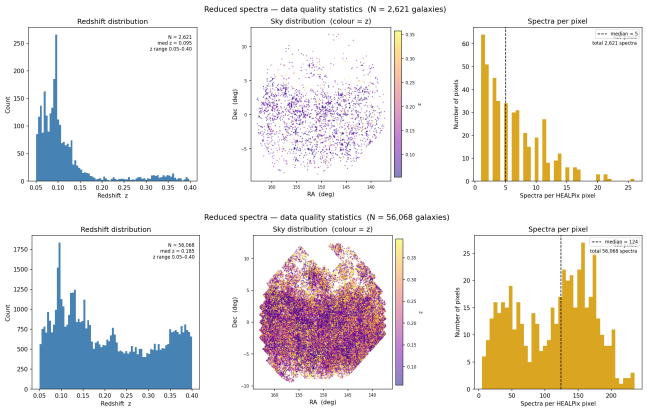


Figure 2. Sample overview for the two working samples. Top: high-SNR PCA training subset (SNR ≥ 8, $N = 2{,}621$, $0.05 \leq z \leq 0.40$, 426 `HEALPix` pixels). Bottom: full SNR ≥ 1 regression sample ($N = 56{,}068$, $0.05 \leq z \leq 0.40$, 498 pixels). For each sample the left shows the redshift distribution, the center shows the footprint, and the right shows the spectra-per `HEALPix`-pixel histogram.

The k-NN density estimator (calculated against the full footprint zcatalog §2.3 of $N = 1{,}033{,}722$ galaxies over $0.05 \lesssim z \lesssim 1.6$) revealed telling structural features that motivate the rest of this analyssi. The left panel of Fig. 3 shows a running median count of $\log_{10} n_k$ as a function of four representative scales ($k \in \{5, 50, 500, 5000\}$). In this we can clearly identify the uneven sampling of sources across the surveyed redshift range. Specifically, we see the sharp monotonic decline from $\log_{10} n_k \sim -1$ at $z \sim 0.1$ down to ~ -3 at $z \sim 0.6$, followed by a secondary peak near $z \sim 0.85$–around the reported peak target density of luminous-red-galaxies in the survey–with the denity tailing off towards higher z. Both of these features are properties of the catalog variable sampling and different target populations in different redshift regimes, not the underlying galaxy field. These are the redshift-dependent observed number densities $\bar{n}(z)$ that motivated the binned and ratio constructions.

The middle and right panels of Fig. 3 display how this redshift dependent trend is absent in the other two targets. As seen in the middle panel, the running median is centered on zero across the redshift range, confirming that the analytic cancellation in Eq. (10) worked as intended. Observing the inter-scale ratios, we can see a roughly flat running median across the redshift range offset by an excitingly scale dependent constant. Both of these constructions thus remove the unwanted

$\bar{n}(z)$ residual and yield signals dominated by true structural variance reflective of local clustering, not sampling. As we will later see, any models with access to redshift information (as a direct feature or implicitly through rest-frame imprinting) will recover the intense redshift dependent signal in $\log n_k$, but encouragingly the same redshift information bares no benefit to the $\log_{10}\delta_k$ or R_{k_a/k_b} targets.

3.2. *The PCA basis captures known galaxy-evolution physics*

Before we could implement the regression, we verified that the spectral compression via PCA is preserving real physical structures rather than noise modes. Two independent analyses support this. First, by observing the explained variance scree (Fig. 4) we can see the very steep concentration of variance in the leading few components: PC1 along accounts for 45.9 % of the total variance and PC2 further accounts for another 20.6 % percent. Together, PC1+PC2 captures 66.5 % of the explained variance and the subsequent PC3-PC8 each explain less than 4 % (beyond PC8 per-component contributions are subpercent and they continue to decay smoothly). We find the curve of the cumulative explained-variance approximately saturates around 77 % by $n_{\rm PC} = 40$, only ever crossing the 90 % or 95 % with hundreds of components. Encoding onto the leading 40 principal components is thus a lossy compression as ~ 33 % of the explain variance likely contained in noise modes is discarded.

The second analysis that supports the encoding of physical structure is the statistical analysis of the leading PC eigen spectra loadings against reference physical indicators. Figure 5 displayes the first five eigenspectra with rest-frame line markers. Excitingly, none of the first 5 principal components exhibit the patterns indicative of pure noise modes, instead showing organized structure tied to known emission/absorption features and the broad-band continuum slope. Projecting each PC against the literature indices §2.2 (Fig. 6) clarifies this relationship with quantiative terms. PC1 is the classical age/star-formation axis, correlating positively at $r \gtrsim +0.7$ with the Hα, Hβ, [N II] λ6584, and [O III] λ5007 line fluxes and negatively at $r \lesssim -0.7$ with $D_n(4000)$ and the Lick metal-absorption indices Mg b, Fe 5270, and Fe 5335. Galaxies with high PC1 scores are thus interpreted to be young, emission line dominated galaxies with weak 4000 Å breaks and metal absorption. The corresponding low PC1 score thus correlates with old, metal-rich galaxies. This is the same primary axis that the standard $D_n(4000)$–EW(Hα) plane isolates (G. Kauffmann et al. 2004; M. L. Balogh et al. 1999). PC2 shows weaker, broad-band correlations in both flux- and EW-type indices consistent with a residual continuum-shape or dust-reddening mode. PC3 anti-correlates with the Lick metal-absorption series and can be thought of as a secondary metallically axis for a fixed age. PC4 correlated most strongly with the Balmer absorption EW Hδ_A and Hγ_A and hence can be considered as a post-starburst/A-star measure. PC5 correlates with both the emission-line fluxes and the O3N2 ratio, suggesting that it encodes ionization-sate/metallicity-of-the-ionized-gas. Between these leading 5 principal components, we have a basis that reproduces the known multivariate structure of the Lick, Balmer, and emission line index space.

Between these two checks, we have established a baseline for the following regression that any observed failure cannot be attributed to non-physical PCA compressions.

3.3. *PCA → raw k-NN density: an apparently strong signal*

The fitting of the GBR pipeline on the PCA scores with raw $\log_{10} n_k$ as the target yields the result seen in Fig. 7. Specifically, the test-set R^2 rises quickly with k scale ($R^2 = 0.32$ at $k = 5$ to $R^2 = 0.46$ at $k = 100$), and then saturates at $R^2 \simeq 0.53$ for $k \gtrsim 5{,}000$ (Fig. 7). The corresponding predicted versus actual scatter plots for every k are shown in Fig. 8. There is a persistent horizontal feature near the sample median $\log_{10} n_k$ which is visible at the low-density end of every panel where the model appears to revert to the mean for the under-represented void tail.

At first glance, this result seems to be a quite predictive and significant result: a $\sim$0.5 R^2 solely by regressing galaxy spectrum onto their local environment scales. However, as we discussed in the methods and expand on the in latter analysis, this interpretation is incorrect. The next two subsections evidentially expand on this as we prove the structure the regressor is recovering is in fact the redshift dependence of $\log_{10} n_k$. Moreover, the climb in R^2 as a function of k is a signature of this, since, at small k, the k-NN sphere samples the local clustering where variability is high and the regressor struggles to identify $\bar{n}(z)$; at high k the sphere samples the smoother general local galaxy density, as seen in from Fig. 3, and hence has an easier time collapsing onto $\bar{n}(z)$.

3.4. *The signal is the redshift trend*

By adding the spectroscopic redshift as a explicit feature alongside the PCA feature vector, we can isolate whether or not the model is already utilizing some interpreted z information by characterizing how R^2 changes with its addition. When we add in z (Fig. 9) we find that the test R^2 increases from 0.39 to 0.55 at $k = 10$

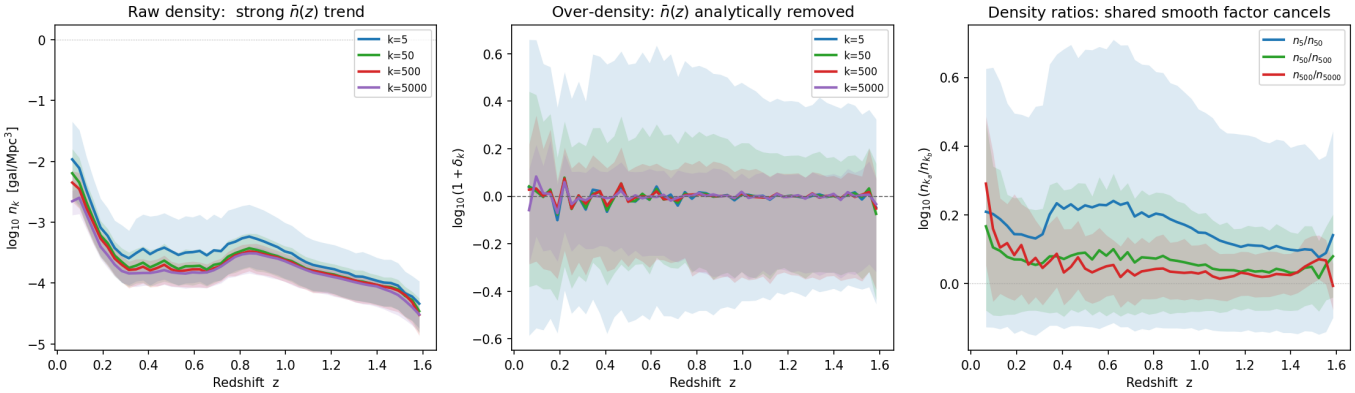


Figure 3. Running median (solid lines) and 16/84-percentile envelopes of three environmental quantities versus redshift on the full footprint zcatalog ($N = 1{,}033{,}722$ galaxies, all k-NN densities computed against the same reference set). Left: $\log_{10} n_k$ for $k \in \{5, 50, 500, 5000\}$ with steep initial monotonic decline and later LRG bump, highlighting redshift-dependent artifact $\bar{n}(z)$ residuals. Middle: redshift-binned overdensity $\log_{10}(1+\delta_k)$ (Eq. 10) which appears centered around 0 indicating sucessful survey artifact cancellation. Right: inter-scale density ratios $\log_{10}(n_{k_a}/n_{k_b})$ (Eq. 8) which are likewise nearly flat, showing only the encouraging offset of a ratio dependent scalar.

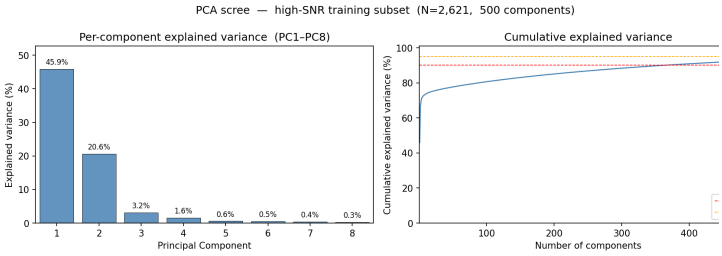


Figure 4. Per-component (left) and cumulative (right) explained variance for PCA on the high-SNR training subset (N = 2,621, 40 components). PC1 and PC2 together capture 66.5 % of the total variance; PC3–PC8 each contribute < 4 %, and the cumulative curve saturates near 77 % by $n_{\rm PC} = 40$ without reaching the 90/95 % reference levels (dashed). The left panel is restricted to PC1–PC8 since beyond PC8 componnets contribute sub-percent variance that is visually indistinguishable from zero.

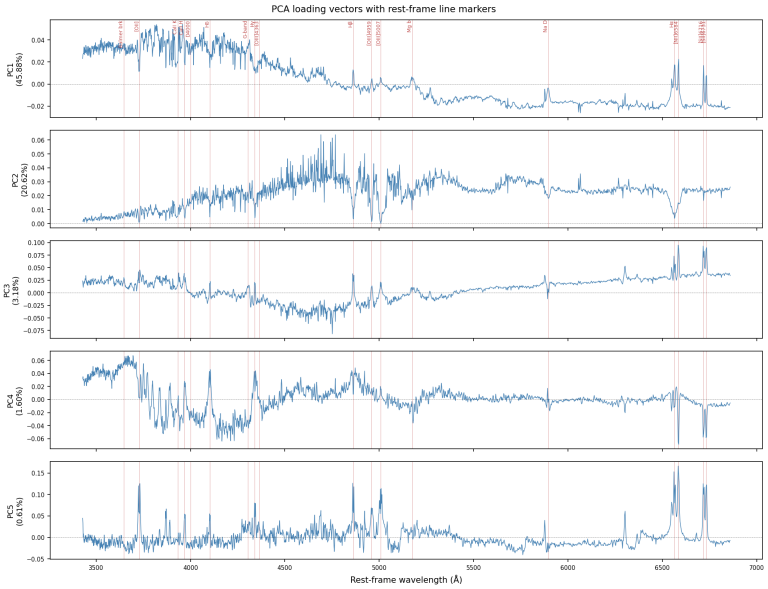


Figure 5. First five PCA eigenspectra with rest-frame emission and absorption lines marked. PC1 captures the dominant age / star-formation axis. PC2 is a residual continuum-shape / dust mode. PC3 is a secondary metallicity axis for a fixed age. PC4 isolates the Hδ_A/Hγ_A Balmer-absorption (post-starburst) direction. PC5 is an ionization-state mode coupling emission-line fluxes to the O3N2 ratio.

and from 0.64 to 0.86 at $k = 500$. This persistent gap implies there is a substantial fraction of the variance in $\log n_k$ that is recoverable directly from z in a way that the spectrum along only imperfectly recovers.

3.5. *The collapse:* $\log_{10}\delta_k$*, ratios, and raw spectra*

Changing the target of the regression to the redshift-binned overdensity $\log_{10}\delta_k$ (Eq. 10), a target which analytically cancels the survey selection bias, we expect to not find this z dependence and hope for real structural regression. Refitting the same GBR pipeline onto the $\log_{10}\delta_k$ target and continuing to use the headline PCA model, the R^2 collapses entirely. Specifically, these test R^2 uniformly plateaus on $\sim 0.02 - 0.03$ at every k (Fig. 10, orange curve). The predicted verses actual scatter plot at $k = 100$ is down in Fig. 11 ($R^2 = 0.03$), and we can easily visually identify the lack of any learned structure as the resultant prediction distribution is tightly centered around the median with not tendency towards diagonal structure. This structure is consistent across k.

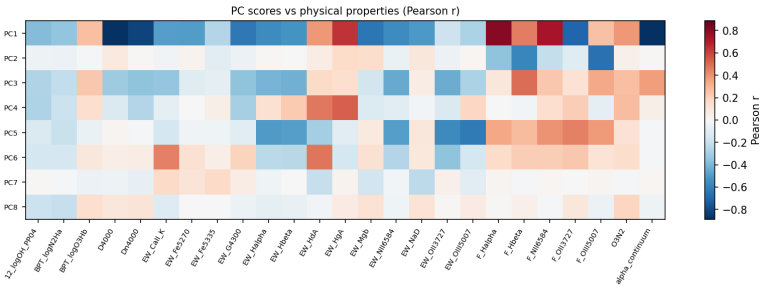


Figure 6. Pearson correlation between each of the leading eight PC scores and each literature spectral index. These were computed on the rest-frame reduced spectra of the high-SNR training subset. PC1 (age/SFR) shows the strongest, broadest correlation pattern while PC2–PC5 isolate progressively narrower sub-axes (continuum/dust, metallicity, post-starburst Balmer absorption, and ionization state). The structure excitingly matches known standard galaxy-evolution axes (G. Worthey 1994; M. L. Balogh et al. 1999; G. Kauffmann et al. 2004).

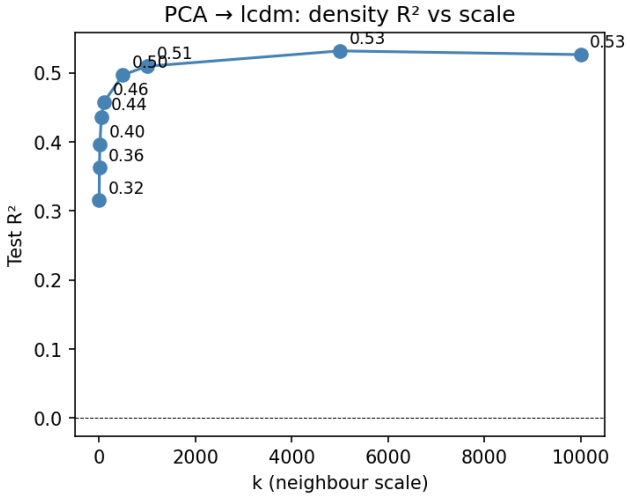


Figure 7. Test R^2 for PCA→ $\log_{10} n_k$ as a function of neighbor scale k on the main GBR test reduction (SNR ≥ 1, $z \leq 0.4$, $N = 54{,}914$, full-zcatalog tree).

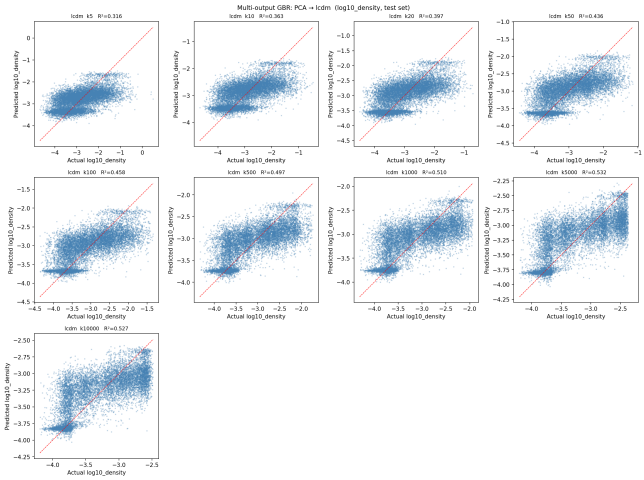


Figure 8. Predicted vs. actual $\log_{10} n_k$ on the test set of the headline PCA→ $\log_{10} n_k$ run (SNR ≥ 1, $z \leq 0.4$, full-zcatalog tree), one panel per neighbor scale $k \in \{5, 10, 20, 50, 100, 500, 1000, 5000, 10000\}$. The distribution tightens onto the diagonal at large k where the smooth $\bar{n}(z)$ trend becomes increasingly easy to recover from the spectrum.

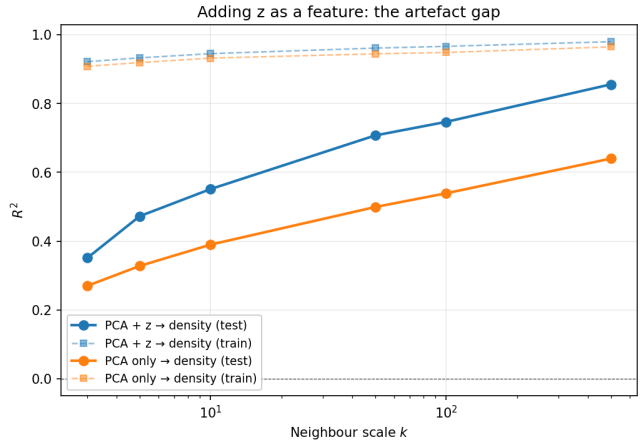


Figure 9. Test and train R^2 as a function of k for PCA+z → $\log n_k$ (red) and PCA→ $\log n_k$ (blue).

We again shift the regression target towards the density ratios $R_{k_a/k_b} = n_{k_a}/n_{k_b}$, in hope that the more clear scale variability may perhaps encourage more significant learning. For all 36 ($\frac{|R|^2-R}{2}$) ordered pairs in the $k \in \{5, 10, 20, 50, 100, 500, 1000, 5000, 10000\}$ grid, we find the test R^2 to be ~ 0.05 for most scales, only showing some semblance of learning at the $R_{k_a/k_{10000}}$ scale with a best R^2 of 22 % (Fig. 12). The growth seen in the bottom right of the heatmap Fig. 12 has a stright-forward interpretation that unfortunately is the same as the earlier regressions onto $\log n_k$: observing Fig. 3 we can see the ratios $R_{k_a/k_{10000}}$ show consistent decline as a function of z over the range $0.05 \lesssim z \lesssim 0.4$. This is logical as in this regime the $\log n_k$ density falls off monotonically with z and the ratio $R_{k_a/k_{10000}}$ thus reflects that correspondence most strongly at higher k ratios. Thus, that high R^2 is simply a symptom of the target being insensitive to true local environment changes, not evidence of environmental recovery.

Raw spectra as features.—A potential concern with the $\log_{10} \delta_k$ collapse is that PCA might be discarding precisely the high-frequency spectral detail that carries the local-environment signal (in which case collapse would identify feature compression as the culprit). We test this by refitting the GBR pipeline on the full rest-frame

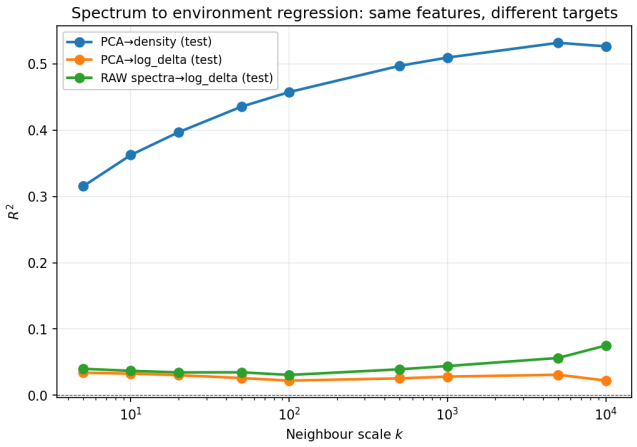


Figure 10. Test R^2 vs. k for the three central regression configurations on the final reduction ($N = 54{,}914$, full-zcatalog tree, no z features). PCA$\rightarrow \log n_k$ (blue) climbs to 0.53. PCA$\rightarrow \log_{10}\delta_k$ (orange) is flat at ~ 0.03. RAW$\rightarrow \log_{10}\delta_k$ (green), using the full rest-frame spectrum ($M = 1844$ pixels) as features, is flat at 0.03–0.07.

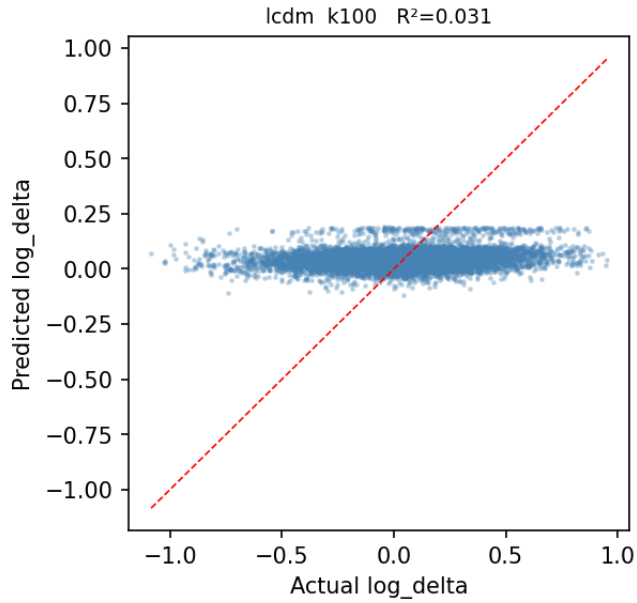


Figure 11. Predicted vs. actual $\log_{10}\delta_k$ at $k = 100$ for PCA features ($R^2 = 0.03$). The distribution is a horizontal cloud at the sample median; the model has no resolving power on residual local overdensity. Lower-k panels are visually identical and quantified by the uniform ~ 0.02–0.03 R^2 plateau in Fig. 10.

spectra ($M = 1844$ rest-frame pixels) as the input feature vectors without PCA compression. The resulting test R^2 on $\log_{10}\delta_k$ (Fig. 10, green curve) shows an almost identical result to the PCA result within 0.02–0.04 at every k with only slight deviation at higher k (all deviation is within the possible train/test gap of potential

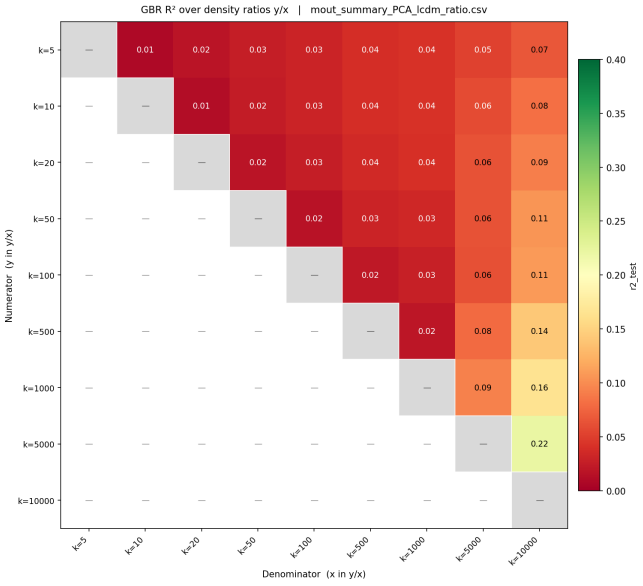


Figure 12. Test R^2 for each ordered density-ratio target $\log_{10}(n_{k_a}/n_{k_b})$, laid out as a numerator × denominator grid (row = k_a, column = k_b).

overfitting). Thus, the representation-independence of the collapse rules out this hypothesis.

3.6. *Robustness across reduction trials*

The conclusions shown above could in principle be specific to the selected reduction range SNR ≥ 8, $z \leq 0.4$. To test this, we created additional reduced subsets of varying SNR cuts (> 1, > 1.5, > 3, > 5) which all spanned the entire $z \leq 1.6$ range. Before getting to the results of those regressions, its worth clarifying that the headline snr and z cuts of SNR ≥ 8, $z \leq 0.4$ were selected *not* because the $\bar{n}(z)$ is flat in that window– it certainly isn't, but becuase that regime is where the zcatalog is most densely populated and best samples the LSSs. This means in that area the k-NN balls that calculate the density have more targets to sample across the k scales. This was done to favor the per-galaxy overdensity and density ratio measures, not the $\log n_k$ measures, and give the regressor the best change to learn some environmental structure. The results of the alternative reductions are detailed in Fig. 13. Here we can see that when we increase the population (lower snr limit) and expand the range to the entire z range, we find the expected z-dependence characterized by increased R^2 on $\log n_k$ as scale gets larger (z dependence is smoother), and the range is fully sampled (more z-dependece is sampled). Furthermore this is confirmed as when tested against the $\log_{10}\delta_k$ and ratios we find the same characteristic null, uniform R^2. Thus, the negative result adds robustness that our specific selection of reduction

configuration was not the reason for poor environmental recovery.

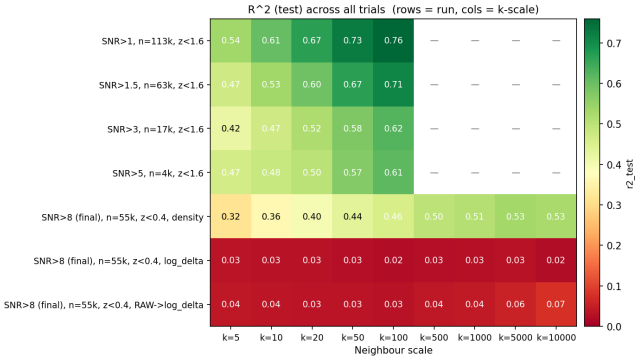


Figure 13. Trial × k-scale heatmap of test R^2 across seven regression runs. Top four rows: exploration reductions on the full $z \leq 1.6$ footprint with looser SNR cuts and different input transforms, raw $\log_{10} n_k$ target, $k \in \{5, 10, 20, 50, 100\}$ only. Bottom three rows: the headline SNR ≥ 8, $z \leq 0.4$ reduction.

In summary, the patterns identified across §§3.3–3.6 are internally consistent showing an apparently strong predictive signal on raw $\log n_k$ for the survey's $\bar{n}(z)$ trend, with no residual local-clustering signal recoverable from the spectrum at present sampling. We turn to §4 to discuss what this implies about the structural reasons for the result and the conditions under which spectrum-driven regression-based environment-inference might become feasible.

4. DISCUSSION

So what does the data support? Our analysis yielded four pieces of evidence converge on a single multi-faceted conclusion. First, GBR on PCA scored galaxy spectrum predicts raw $\log_{10} n_k$ with monotonically rising R^2 that plateaus near 0.53 at $k \geq 5000$ (§3.3). Second, running the same model pipeline with a k-d tree built from only a small SNR-cut subset of the full zcatalog strengthens this recovery. Third, swapping the targets of the regression to the redshift-binned overdensity $\log_{10} \delta_k$, or k-NN density ratios collapses the test R^2 to ~ 0.03 uniformly for all scales, k. Four, this collapse holds when the features are switched from 40 principal components to the full $M = 1844$ dimensional raw, rest-frame spectra (ruling out lossy compression). Together these state that recoverable environmental signal in this configuration is dominated by the shortcut learning of the survey's $n(z)$ selection function, not by local clustering, and that the bottleneck present is not an issue of spectral feature representation.

Importantly, this conclusion should be understood in conjunction with what the data does not support. Specifically, it does not claim that galaxy spectra are uninformative about local environments. The color–density and SFR–density relations are, after all, well observed and verified at low z (A. Dressler 1980; G. Kauffmann et al. 2004; P. L. Gómez et al. 2003; M. L. Balogh et al. 1999), and the leading PCs in our basis recover exactly those axes. What collapses in this study is that predictive mapping from a spectrum to a *within-survey* k-NN environmental density measure once $n(z)$ trends are removed. The negative result, is likely even specific to (i) the DESI DR1 main/dark fiber-spectroscopic sampling, (ii) the k-NN family of density estimates, and (iii) the choice to fix the labels at the time of training. Other survey geometries, denser galaxy fields (perhaps the soon-to-come DESI DR2 survey), and different environmental operators may behave very differently.

For clarification, the sources of the limitations we face in this work are as follows.

- **Sparse sampling.** For intermediate k, the k-NN distances already approach a scale of tens-of-Mpc–larger than the size of typical halo and filment structures–so n_k measures a smoothed cosmic field rather than the true local environment driving spectral galaxy evolution. Smaller k from a denser survey in principal should resolve halo scales better, however, the variance of n_k at those scales in our current surveys drowns out any true environmental learning signal.

- **Spectroscopic selection.** The DR1 fiber-assignment algorithm innately favors regions of high target densities–an operation that is not flat across environments. Same-sample k-NN densities therefore receive the overlying selection fluctuations.

- **Self-referential density.** When building a k-d tree from any type of survey which inherently cannot perfectly sample the galaxy field, we encode that sub-sample spatial distribution onto density labels. Moreover, we saw that when a model has even weak access to some notion of z it can, and will prefer to recover it over any possible underlying environmental signal.

- **Continuous-density target.** A continuous density target label inherently smooths the distinction between a "void" and "cluster" galaxy, a thought to be discrete or quasi-discrete distinction, thus perhaps making the challenging to differentiate between them.

4.1. *Implications for spectrum-driven environment inference*

For a predictive "spectrum → environment" regression on imperfect fiber complete surveys like what is presented in this study, the right target is unlikely to be a within-survey k-NN computed density measure. Some clearer alternatives might be computing the k-NN densities from a more complete photometric survey or photo-z catalog, explicit modeling and compensation of the fiber selection bias function, environmental labels defined on a different length scale (perhaps host-halo membership from N-body mathching or group-finder catalogs), or categorical cosmic-web classifications obtained from classical analytic structure-finding tools.

Expanding on the last of those ideas, perhaps cosmic-web classification (which has already shown possible at very low z) could act as a supervised training sample on which spectrum-environment regression could expand conclusions to previously unreached z regimes. Consider the combinatorial simplicies of vertices, edges, tetrahedra, clusters, and hyperedges as defined by J.-Y. Lee & F. Villaescusa-Navarro (2025). If a topological structure could be fit to some small, low z large-scale structure sample, and the computational machinery (beneficial and efficient because of its combinatorial nature) of J.-Y. Lee & F. Villaescusa-Navarro (2025) could perhaps be adapted to regress spectral features instead of cosmological density parameters. This conceptually seems intriguing, as we may find a cleaner result since the structure is concretely defined and present in the training data. If training on this conception were successful, this could possibly be tested on spectra of any sampling to infer the simplical structure of the cosmic web across a very vast redshift regime previously unreachable.

Beyond entirely different approaches, one change to the representational learning side of this study that may boost the weak environmentally coupled signal is an unsupervised approach. Specifically, one might choose to use a β-TCVAE trained on rest-frame spectra to generate non-linear disentangled latent variables which could highlight weak environmental information that PCA simply misses or discards.

5. CONCLUSIONS

1. Rest-frame DESI DR1 spectra compressed with PCA recover the standard age, SFR, and metallicity axes of galaxy variation. Specifically, the leading PCs correlate with D_n4000, Hα, and Mg b in directions consistent with the literature.

2. HGBR predicts the raw k-NN density $\log n_k$ from PCA scores effectively with test R^2 growing monotonically from 0.32 at $k = 5$, through 0.46 at $k = 100$ where it saturates around $\simeq 0.53$ for $k \gtrsim 5{,}000$ (Fig. 7).

3. Replacing the target of the regressor with redshift-binned overdensity or k-NN density ratios collapsed the test R^2 entirely to ~ 0.03 at all scales of k for both PCA and full raw spectral input features.

4. The high raw-density R^2 is hence attributable to the $n(z)$ selection function inherent to the survey and the self-referential nature of within-survey k-NN density calculations.

5. In the DESI DR1 main/dark spectroscopic sample, a within-survey k-NN density is not a sufficient environmental label for direct spectrum-environment regression. Future spectrum-environment regression based studies should thus either change the label categorically, change the parent catalog, or model the selection function itself.

We acknowledge the use of the University of Florida HiPerGator computing cluster as well as Prof. Jaehan Bae for his excellent instruction throughout the year. This work was carried out as a term project for AST7939 "Machine Learning in Astronomy" Spring 2026 at the University of Florida.

Software: `Python` (Python Software Foundation 2024), `NumPy` (C. R. Harris et al. 2020), `SciPy` (P. Virtanen et al. 2020), `scikit-learn` (F. Pedregosa et al. 2011), `Astropy` (Astropy Collaboration 2022), `Matplotlib` (J. D. Hunter 2007), `pandas` (W. McKinney 2010)

REFERENCES

Aragón-Calvo, M. A. 2019, Classifying the large-scale structure of the universe with deep neural networks, MNRAS, 484, 5771

Aragón-Calvo, M. A., van de Weygaert, R., & Jones, B. J. T. 2010, Multiscale phenomenology of the cosmic web, MNRAS, 408, 2163

Astropy Collaboration. 2022, The Astropy Project: sustaining and growing a community-oriented open-source project, ApJ, 935, 167

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, Classification parameters for the emission-line spectra of extragalactic objects, PASP, 93, 5

Balogh, M. L., Morris, S. L., Yee, H. K. C., Carlberg, R. G., & Ellingson, E. 1999, Differential galaxy evolution in cluster and field galaxies at $z \sim 0.3$, ApJ, 527, 54

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, How filaments of galaxies are woven into the cosmic web, Nature, 380, 603

Bruzual A., G. 1983, Spectral evolution of galaxies. I. Early-type systems, ApJ, 273, 105

Cautun, M., van de Weygaert, R., Jones, B. J. T., & Frenk, C. S. 2014, Evolution of the cosmic web, MNRAS, 441, 2923

Colless, M., et al. 2001, The 2dF Galaxy Redshift Survey: spectra and redshifts, MNRAS, 328, 1039

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, Spectral Classification of Galaxies: an Orthogonal Approach, AJ, 110, 1071

Cooper, M. C., et al. 2008, The DEEP2 Galaxy Redshift Survey: the relationship between galaxy properties and environment at $z \sim 1$, MNRAS, 383, 1058

Cucciati, O., et al. 2006, The VIMOS-VLT Deep Survey: the build-up of the colour–density relation, A&A, 458, 39

DESI Collaboration. 2024, The Early Data Release of the Dark Energy Spectroscopic Instrument, AJ, 168, 58

DESI Collaboration, Adame, A. G., et al. 2024, DESI 2024 III: Baryon Acoustic Oscillations from Galaxies and Quasars, JCAP

DESI Collaboration, Aghamousa, A., et al. 2016, The DESI Experiment Part I: Science, Targeting, and Survey Design, arXiv e-prints

Dressler, A. 1980, Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies, ApJ, 236, 351

Friedman, J. H. 2001, Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics, 29, 1189

Gómez, P. L., et al. 2003, Galaxy Star Formation as a Function of Environment in the Early Data Release of the Sloan Digital Sky Survey, ApJ, 584, 210

Hahn, O., Porciani, C., Carollo, C. M., & Dekel, A. 2007, Properties of dark matter haloes in clusters, filaments, sheets and voids, MNRAS, 375, 489

Harris, C. R., et al. 2020, Array programming with NumPy, Nature, 585, 357

Hogg, D. W., et al. 2003, The Overdensities of Galaxy Environments as a Function of Luminosity and Color, ApJ, 585, L5

Hunter, J. D. 2007, Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90

Kauffmann, G., White, S. D. M., Heckman, T. M., et al. 2004, The environmental dependence of the relations between stellar mass, structure, star formation and nuclear activity in galaxies, MNRAS, 353, 713

Ke, G., Meng, Q., Finley, T., et al. 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Advances in Neural Information Processing Systems (NeurIPS), Vol. 30

Lee, J.-Y., & Villaescusa-Navarro, F. 2025, Cosmology with Topological Deep Learning, https://arxiv.org/abs/2505.23904

Lewis, I., et al. 2002, The 2dF Galaxy Redshift Survey: the environmental dependence of galaxy star formation rates near clusters, MNRAS, 334, 673

McKinney, W. 2010, Data Structures for Statistical Computing in Python, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 56

Melchior, P., Liang, Y., Hahn, C., & Goulding, A. 2023, Autoencoding Galaxy Spectra. I. Architecture, AJ, 166, 74

Pedregosa, F., et al. 2011, Scikit-learn: Machine Learning in Python, J. Machine Learning Research, 12, 2825

Peng, Y.-j., et al. 2010, Mass and environment as drivers of galaxy evolution in SDSS and zCOSMOS and the origin of the Schechter function, ApJ, 721, 193

Pettini, M., & Pagel, B. E. J. 2004, [O III]/[N II] as an abundance indicator at high redshift, MNRAS, 348, L59

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, Dimensionality Reduction of SDSS Spectra with Variational Autoencoders, AJ, 160, 45

Postman, M., et al. 2005, The Morphology–Density Relation: The Group Connection, ApJ, 623, 721

Python Software Foundation. 2024, Python Language Reference,

Strauss, M. A., et al. 2002, Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample, AJ, 124, 1810

Sutter, P. M., et al. 2015, VIDE: The Void IDentification and Examination toolkit, Astronomy and Computing, 9, 1

Tegmark, M., et al. 2004, The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey, ApJ, 606, 702

Virtanen, P., et al. 2020, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods, 17, 261

Worthey, G. 1994, Comprehensive stellar population models and the disentanglement of age and metallicity effects, ApJS, 95, 107

Worthey, G., & Ottaviani, D. L. 1997, Hδ and Hγ absorption features in stars and stellar populations, ApJS, 111, 377

Yip, C. W., et al. 2004, Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey, AJ, 128, 585