UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

AST7939: Machine Learning in Astronomy

Limitations of k-NN Environment Recovery in DESI DR1

A Negative, but Insightful Result on Spectrum–Density Regression

James Watt · Department of Astronomy, University of Florida

DESI DR1
Main/Dark Survey
RA = 150°, Dec = 0°
Spring 2026

Abstract

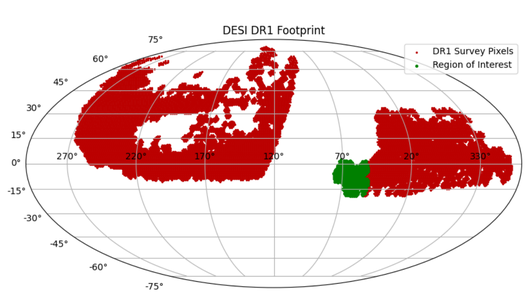
Galaxy spectra encode stellar age, star formation, metallicity, and gas conditions, all of which **correlate with a galaxy's local environment**. We test whether **supervised machine learning** can recover continuous environmental density directly from **DESI DR1 rest-frame spectra**. ~55,000 galaxies at $0.05 < z < 0.40$ were compressed with **principal component analysis (PCA)** and regressed onto **k-nearest-neighbor (k-NN) density** labels with **histogram-based gradient boosting**. The leading PCs recover physically meaningful galaxy-evolution axes, and regression onto raw log n_k appears successful, with **test R^2 rising from 0.32 to 0.53**. However, once the survey's redshift-dependent selection function is removed, **R^2 collapses to ≈ 0.03** for both PCA and full raw-spectrum inputs. The apparent signal is **shortcut learning of the survey selection function**, not local clustering, so within-survey k-NN labels are not a viable environment target in DR1.

Introduction

A galaxy's **environment shapes its evolution**: galaxies in groups, clusters, and filaments differ from field galaxies in **color, star formation rate, age, and metallicity**[1,2]. Those imprints should be spread across the whole spectrum, so a capable model **might infer environment from the spectrum alone**. DESI DR1 provides **~13 million galaxy spectra**[3], yet the spectrum is usually used only for its redshift.

The catch: DESI's observed galaxy density $\bar{n}(z)$ falls ~100× across redshift because of magnitude limits and target selection, so any density label carries a **redshift trend unrelated to local clustering**.

Question: Can supervised regression of DESI DR1 rest-frame spectra recover a galaxy's local k-NN environmental density?



DR1 main/dark footprint (red) and our region of interest (green): 498 HEALPix pixels (NSIDE = 64), ~420 deg² near COSMOS.

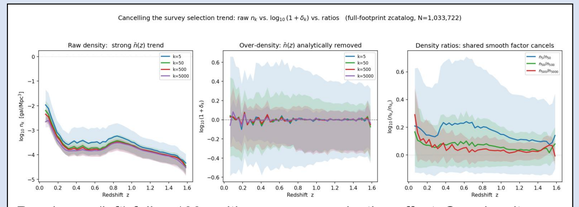
Methodology

Data & reduction: DR1 main/dark coadds from **498 HEALPix pixels**; SPECTYPE = GALAXY, ZWARN = 0, $0.05 \le z \le 0.40$. B/R/Z arms are IVAR-merged, shifted to the **rest frame**, median-normalized, and min–max scaled per bin → **1,844 pixels** per spectrum.

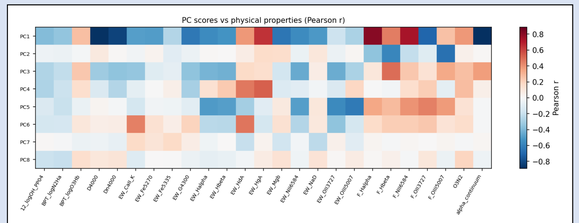
Spectral features: a **40-component PCA** basis is fit on a clean subset (SNR ≥ 8, N = 2,621), then applied to the full SNR ≥ 1 sample (***N* = 54,914**). PCs are checked against 25+ Lick, D_n4000, emission-line and BPT indices.

Environment labels: a k-d tree on the **full DR1 zcatalog (1.03 M galaxies)** gives $n_k = k / (4/3\ \pi\ d_k^3)$ for k = 5 … 10,000, plus density ratios and the z-binned overdensity $\log \delta_k = \log n_k - \text{median}_{z\text{-bin}}(\log n_k)$, which **cancels the survey's $\bar{n}(z)$**.

Regression: HistGradientBoosting, one model per k, 80/20 split, scored by **test R^2**.

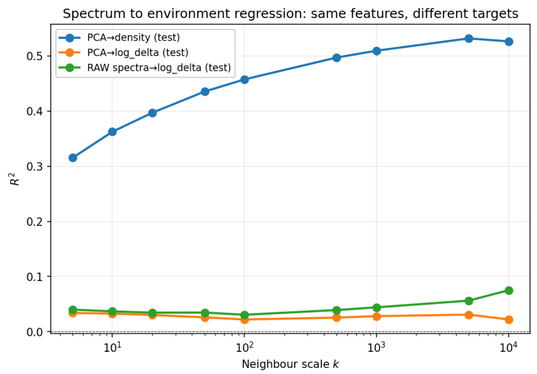


Raw log n_k (left) falls ~100× with z: a survey selection effect. Overdensity (center) and ratios (right) remove it.



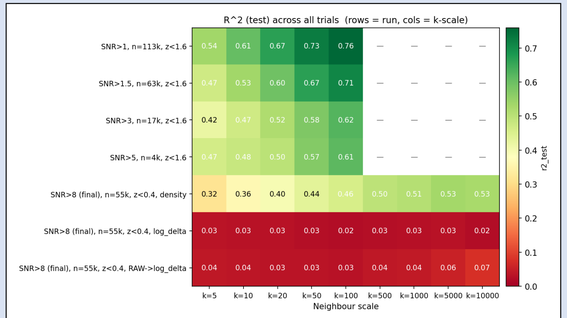
*PC scores vs. physical indices. **PC1 (45.9%)** is the age/SFR axis; PC2 (20.6%) continuum/dust; PC3–5 metallicity, post-starburst, ionization.*

Results & Discussion



Same features, different targets. PCA → raw density (blue) climbs to R^2 = 0.53; PCA → log δ_k (orange) and raw spectra → log δ_k (green) stay flat at ~0.03.

R^2 = 0.53 → 0.03
once $\bar{n}(z)$ is removed



Test R^2 across seven reductions: every raw-density row is green, every overdensity row is red.

- **PCA is physical:** leading PCs recover the classic age, dust, metallicity and post-starburst axes.
- **Raw density looks predictable:** R^2 rises from 0.32 (k = 5) to 0.53 (k ≥ 5,000).
- **But it is redshift:** adding z as a feature lifts R^2 by 0.10–0.22; removing $\bar{n}(z)$ drops it to **~0.03**.
- **Not a compression artifact:** raw 1,844-pixel spectra fail identically.

Conclusion

The apparently strong spectrum-to-density signal is **shortcut learning**: the model infers redshift from the spectrum and reads off the survey's **selection function $\bar{n}(z)$**, not local clustering.

Once $\bar{n}(z)$ is removed, **R^2 collapses to ~0.03 at every scale**, independent of feature representation and **robust across seven reductions**. The spectra still carry real galaxy physics, so the bottleneck is the **label, not the spectrum**: in DR1's sparse, selection-shaped sampling, a **within-survey k-NN density is not a faithful environment label**.

Take-away: before claiming a spectrum predicts environment, check that the signal **survives removing the survey selection function**. Here, it does not.

Future Research

- **Cosmic-web classes** (void / sheet / filament / node from T-web[4]) as categorical targets for a 1D CNN
- k-NN density against an **external photo-z catalog**
- **Selection-corrected** (propensity-weighted) overdensity
- **β-TCVAE** latents to test non-linear features

Acknowledgements

Computing on **UF HiPerGator** (UF Research Computing). This work uses public **DESI DR1** data. Carried out for **AST7939**, Spring 2026.

References

1. Dressler A. 1980, *ApJ*, 236, 351
2. Kauffmann G. et al. 2004, *MNRAS*, 353, 713
3. DESI Collaboration 2025, *DESI Data Release 1*, arXiv:2503.14745
4. Hahn O. et al. 2007, *MNRAS*, 375, 489
5. Cooper M. C. et al. 2008, *MNRAS*, 383, 1058
6. Ke G. et al. 2017, *NeurIPS*, 30 (LightGBM)
7. Pedregosa F. et al. 2011, *JMLR*, 12, 2825