Detecting Bias in a Self-Organizing Map of Galaxy Photometry Data

Zach Claytor[1], Peter Capak[2], Dan Masters[2]

[1]Ohio Wesleyan University, [2]Infrared Processing and Analysis Center, California Institute of Technology

Abstract

High redshift (very distant) galaxy surveys record broad-band photometry for billions of galaxies in order to measure distances in a faster, more cost-effective way than spectroscopy. Knowing these distances (more specifically, the redshifts) helps cosmologists learn more about the early universe and how it evolved, but calibrating the redshifts from photometry requires a color-selection technique. One such technique is the Self-Organizing Map (SOM), a machine-learning algorithm that projects high-dimensional photometry data onto a visual, two-dimensional map. High-redshift galaxies can be identified efficiently in such a mapping. However, there is a chance that high-redshift galaxies are lost to low-redshift regions and vice versa due to photometric error; this effect was analyzed. A Monte Carlo simulation was run on objects selected from along the boundary between high- and low-redshift regions. Roughly 18% of selected objects scattered from high- to low-redshift, and about 16% scattered the other way. Further research will design a better metric of the scattering percentages based on the number density of galaxies in the map, and future work should use these analysis techniques on other high-redshift data.

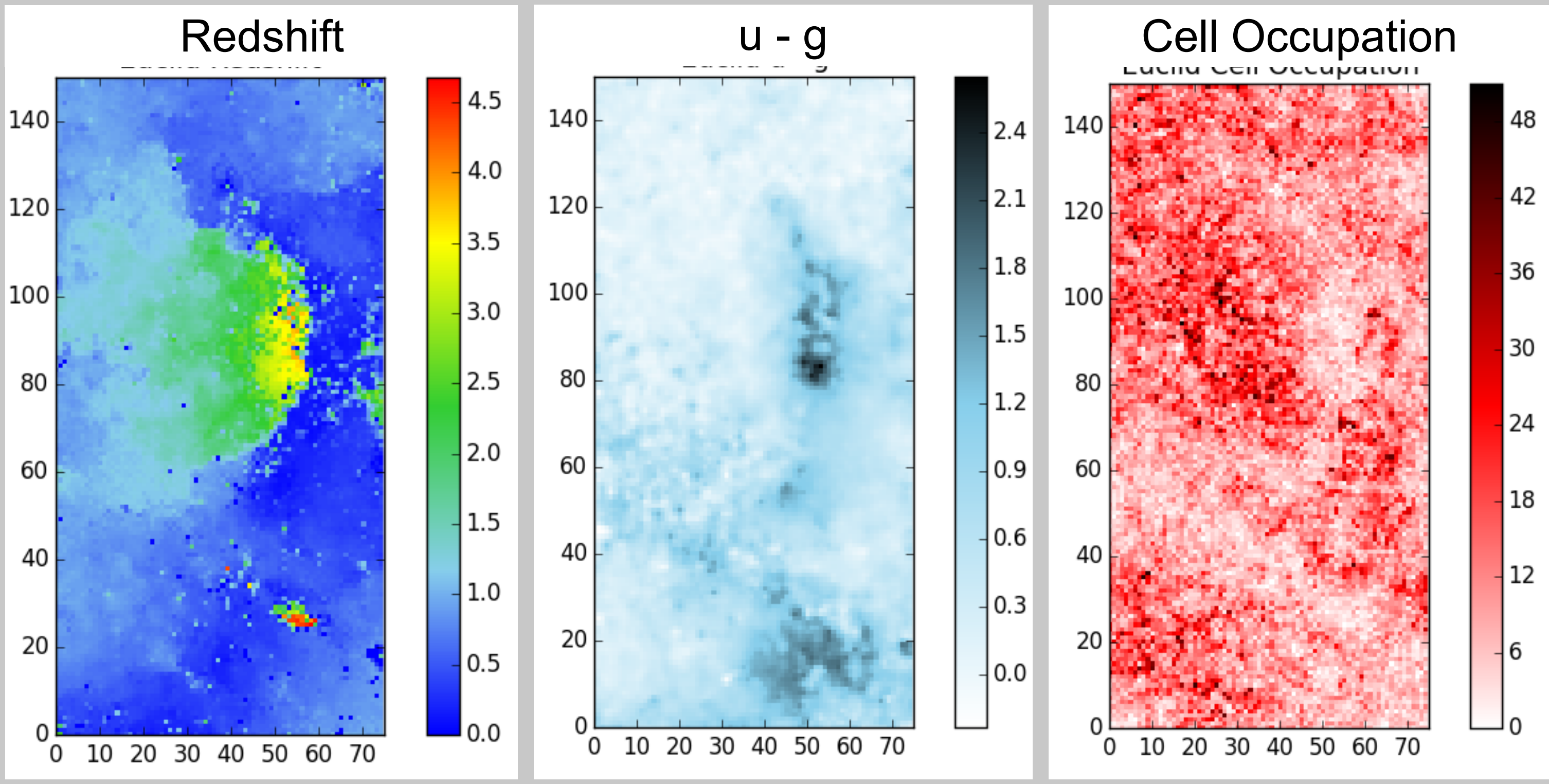
The Self-Organizing Map (SOM)

High-redshift (or "high-z") surveys record photometry of billions of galaxies, using as many as thirty broad band filters per galaxy. Organizing that much data can be difficult, especially when sorting between high-redshift and low-redshift galaxies. The SOM is not only a sorting technique, but it also provides a visual representation of the data.

The SOM represents a 2D assortment of "bins" that hold galaxies with similar photometry. After randomly placing a few sample galaxies on a blank map, each remaining galaxy is placed. This operates via a machine-learning algorithm in which the photometry measurements are treated as coordinates in "color-space," forming a "color vector" for each galaxy. A galaxy is compared with each cell in the map, and the reduced chi-square is measured:

$$X_k^2 = \frac{1}{n}\sum_i^n \frac{(x_i - w_{k,i})^2}{\sigma_{x_i}^2},$$

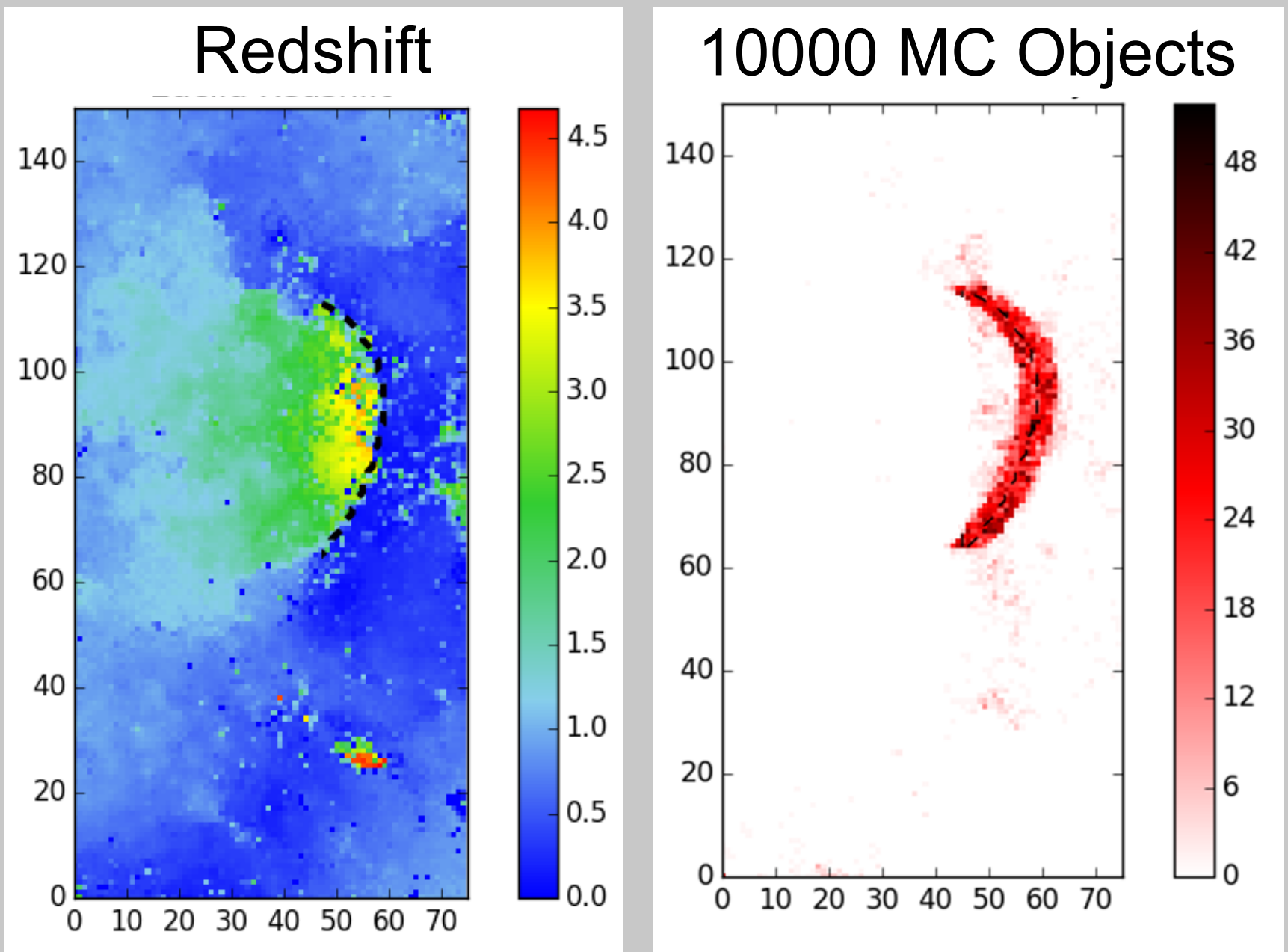
where x is the galaxy's color vector, w_k is the color vector in the kth comparison cell, and n is the number of coordinates in the color vector. The cell that minimizes the chi-square is kept as the best-matching unit (BMU), and the galaxy is placed into that cell. The cell's color vector as well as the vectors of nearby cells are updated to include the new galaxy. Galaxies with similar spectra are placed close together, resulting in a smooth appearance in color-related quantities of the map.



A color-redshift relation can be calibrated using a small subset of spectroscopy, which allows the map to display redshifts associated with each bin (left). Individual colors can be displayed as well (center). Other qualities can be shown that aren't related to color, such as the number density of galaxies in a given cell (right).

Scattering Galaxies

Since neighboring cells have similar color distributions, they are also expected to have similar redshifts. For this reason, the redshift map should appear smooth. Notice, however, that there is a sharp divide on the right edge of the high-redshift region. Since the cells on either side of the dividing line have similar photometry, it is possible for galaxies near that line to be misplaced. In essence, high-z galaxies are lost to low-z regions, and low-z galaxies contaminate the high-z region.



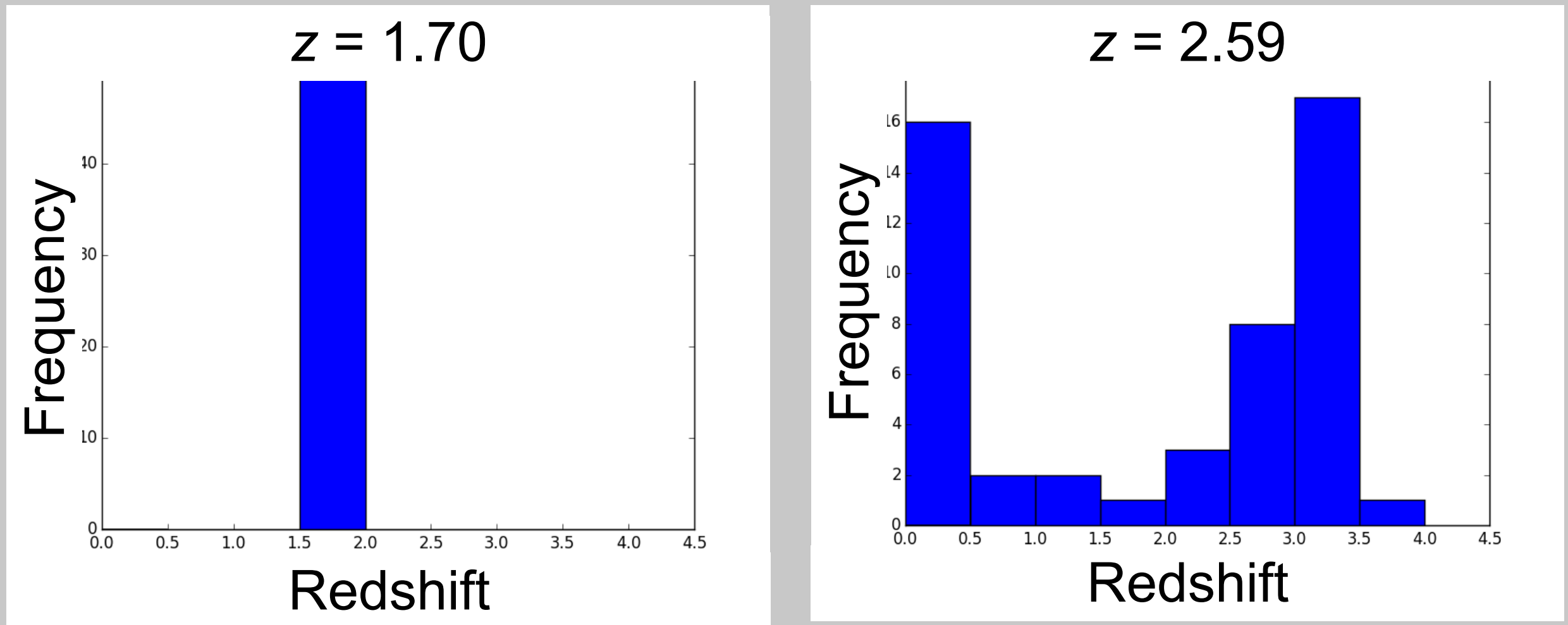
Tracking Photometric Error

To see how galaxies scatter in the SOM, a Monte Carlo simulation was performed on cells along the dividing line between the high- and low-redshift regions. The dividing line was first fit with a fourth-order curve (shown left), and then the bins within three pixels to the left or right of the curve were used as the selected region. This way, both high- and low-z cells were used for the simulation.

Each cell's color vector was randomly scattered about its error, and then the newly scattered vector was treated as a galaxy and rerun through the BMU algorithm to see where it would land on the map. The right image displays a number density of where Monte Carlo objects landed in the map.

Results & Conclusion

The redshift scattering was tracked for each cell in the selected region. Two histograms are below, one showing one of the many cases where the object landed in the same cell through every iteration in the simulation (left), and one showing a cell that scattered evenly between high- and low-z regions (right).



A rough calculation was made for a percentage of galaxies scattering both into and out of the high-z region. About 18% of the objects scattered from high- to low-redshift, and about 16% scattered from low- to high-redshift. This means that around 34% of galaxies involved in the simulation are potentially being misread by the redshift assignment.

Realistically, the number of galaxies that scatter will depend on the number of galaxies in each cell (i.e., cells with fewer galaxies will contribute less to the scatter percentage). Improvements to the Monte Carlo simulation should take the number density per cell into account. Ideally, the algorithm creating the SOM should also be refined to place galaxies more carefully.

References

- Masters, et al. *Mapping the Galaxy Color-Redshift Relation: Optimal Photometric Redshift Calibration Strategies for Cosmology Surveys.* (unpublished)
- Le Fevre, et al. *The VIMOS Ultra-Deep Survey: ~10000 galaxies with spectroscopic redshift to study galaxy assembly at early epochs 2 < z < 6. Astron. Astrophys.* **576**, A79 (2015).

Acknowledgements

I wish to thank my mentors, Peter Capak and Dan Masters, as well as the Infrared Processing and Analysis Center at the California Institute of Technology.
This research was funded by Southern California Edison.