Teaching Graduate Level Data Science for Atmospheric and Oceanic Sciences

Jen Kay[1,2], Elizabeth Maroon[3], Eleanor Middlemas[4], Gina Josef[1,2], Vineel Yettella[5]

[1]Dept. Atmospheric and Oceanic Sciences (ATOC), [2]CIRES, University of Colorado Boulder

[3]Dept. of Atmospheric and Oceanic Sciences, University of Wisconsin [4]PricewaterhouseCoopers, [5]Apple

Context: How do early career researchers learn best practices in applying data science methods to their research? Does tinkering with real data and applications enhance learning and engagement?? *Since 2018, CIRES researchers (a professor, postdocs Elizabeth/Eleanor and graduate students Vineel/Gina) have developed self-guided application laboratories in jupyter notebooks/python as a core part of a graduate-level data science course offered six times since 2018 (ATOC5860, Figure 1+2). These classroom-tested labs illustrate best practices by applying data science methods to classic datasets in atmospheric and oceanic sciences and beyond. Select examples below - All on github:* https://github.com/jenkayco/ATOC5860_Spring2024

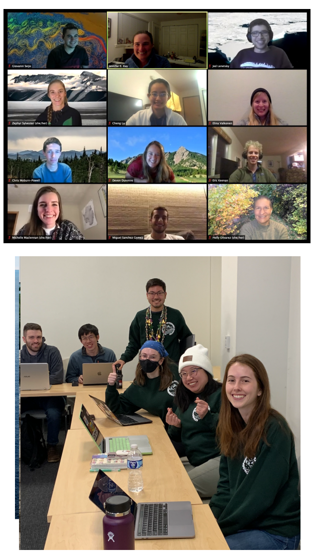


Figure 1. Students in ATOC5860 on Zoom (Fall 2020) and working together on application labs in class (Spring 2023)

	Tuesday	Thursday
January	**January 16** 1. Introductions/Basic statistics/Bayes Theorem (Barnes 1.1-1.2) Complete pre-class survey, Set up Python and Github, HW#1 assigned	**January 18** 2. Statistical Significance Testing /Hypothesis testing/Resampling/Monte Carlo (Barnes 1.3-1.5)
	January 23 Application LAB #1 Basic Statistics and Hypothesis testing	**January 25** Applications LAB #1 cont.
	January 30 3. Compositing/Other distributions/Non-parametric tests (Barnes 1.6-1.8)	**February 1** 4. Regression (Barnes 2.1-2.2) HW#1 due HW#2 assigned
February	**February 6** 5. Autocorrelation/Autoregressive model/Sample Size (Barnes 2.3-2.4)	**February 8** Applications LAB #2 Regression/Autocorrelation
	February 13 Applications LAB #2 cont.	**February 15** 6. EOFs via Eigenanalysis/SVD (Barnes 3.1.1-3.1.4) HW#2 due, HW#3 assigned
	February 20 7. EOFs with actual data (Barnes 3.1.5)	**February 22** *No class*
	February 27 Applications LAB #3 – EOFs	**February 29** Applications LAB #3 cont.
March	**March 5** 8. Harmonic analysis; power spectra (Barnes 4.1.1-4.1.2)	**March 7** 9. Fourier Transforms/Significance testing of spectral peaks/Data windows (Barnes 4.1.3-4.1.5) HW#3 due, HW#4 assigned
	March 12 10. Convolution Theorem, Response function for various windows, Applying overlaps of the windows (Barnes 4.1.5)	**March 14** Applications LAB #4 – Timeseries analysis/Power spectra
	March 19 Applications LAB #4 Continued	**March 21** HOMEWORK PRESENTATIONS: #2, #3 HW#4 due Friday March 22
	SPRING BREAK – NO CLASS	
April/May	**April 2** 11. Filtering (Barnes 4.1.6; Hartmann 7) HW#5 assigned	**April 4** 12 (short). Application Lab #5
	April 9 13. Machine Learning Overview	**April 11** Machine learning with Dr. Jake Gristey NOAA/LASP (guest lecture) HW#5 due, HW#6 assigned
	April 16 14. (short) Applications LAB #6: Machine Learning - Clustering	**April 18** 15. (short) Continue AL #6 – Self Organizing Maps Application Lab #6 links due today
	April 23 *No class*	**April 25** Supervised Machine Learning HW #6 due
	April 30 HOMEWORK PRESENTATIONS: #4, #5, #6	**May 2** Interpretable machine learning with Dr. Kirsten Mayer NCAR (guest lecture) Paper review due Tuesday May 7

Typical Semester ATOC5860 Objective Data Analysis. Meets T/Th 2:30-3:45 in SEEC
Classes in yellow are entirely "learning by doing" application labs in small groups

Figure 2. Typical ATOC5860 Schedule

Empirical Orthogonal Functions (EOF)

What structures explain the most variance in a database of faces?

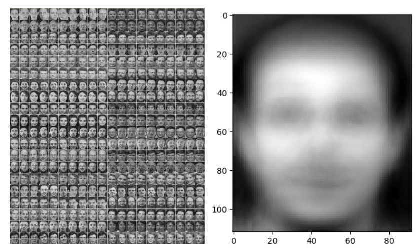


Figure 3. All faces in the database (left), Average face (right). "Look at your data!"

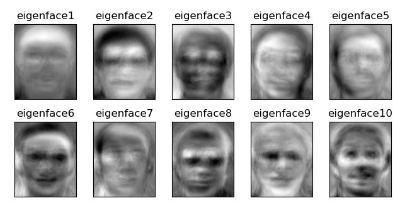


Figure 4. Eigenfaces, i.e., the structures that explain the most variance in sample

Key Result: Hair, Glasses, Eyebrows, Noses explain a lot of variance. Also, eigenfaces are creepy.

Spectral Analysis

Which frequencies have statistically significant power?

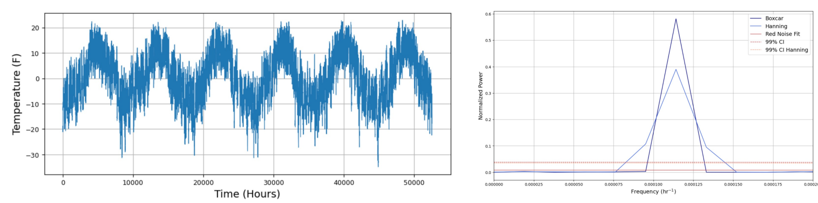
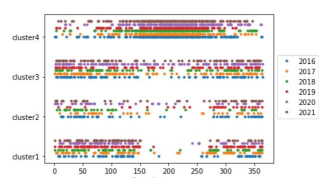


Figure 5. Boulder, CO 2016-2021 hourly surface temperature in the time (left) and spectral (right) domains

Key Result: Power at the annual cycle exceeds the red noise null hypothesis for some but not all variables.

K-means clustering

What happens when we define the seasons based on data, instead of the date?



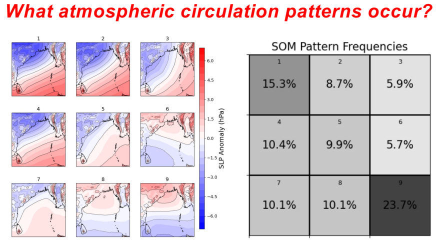
	pres_mb	tdry_degC	rh_percent	wdir	wspd_m_per_s	wspdmax_m_per_s	raina_event_mm
cluster1	809.1	0.1	70.8	146.9	2.4	4.4	0.1
cluster2	807.5	10.5	26.6	267.9	7.5	14.5	0.1
cluster3	808.9	11.9	28.9	234.2	2.7	5.5	0.0
cluster4	814.7	18.7	35.7	79.8	2.0	4.2	0.0

Figure 6. Boulder, CO 2016-2021 clustering to define data-based "seasons": when they occur (left), cluster centroids (above)

Key Result: Date-based and Data-based definition of the seasons differ substantially!

Self Organizing Maps

What atmospheric circulation patterns occur?



Key Result: Large variety of sea level pressure (SLP) anomaly patterns can be found

Supervised machine learning

Predict rain using surface meteorological station observations

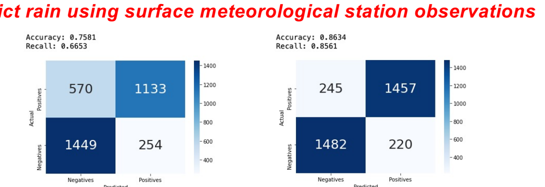


Figure 7. Confusion matrices for random forest (left), and neural network (right) on the "test" data.

Key Result: Comfortingly (!), relative humidity is the most important feature for predicting rain for all methods.