Classifying Los Angeles County

Reservoir Capacities

May 17, 2017 | Metis Data Science

PROBLEM & MOTIVATION

California's recent drought has placed unprecedented demands on our freshwater resources, renewing enthusiasm for surface water infrastructure investments such as raising dams to capture more water in wet years.

Reservoir improvements would need to consider the **frequency** and the **extent** to which these dams are depleted.

My analysis looked at time series reservoir storage data around LA county to **classify storage levels at summer's end (Sept 1)**, given data about the rest of the year.

Andrew Tom | Slack @andrew.tom

The Data Landscape

Cleaned data: (~400 Observations, 32 Features)

Data sourced from the Department of Water Resources **California Data Exchange Center (CDEC)**

Reservoir storage data has was the most complete dataset, going back well into the 1970s. For this analysis, I chose to look at monthly storage data from 2000 - 2013.

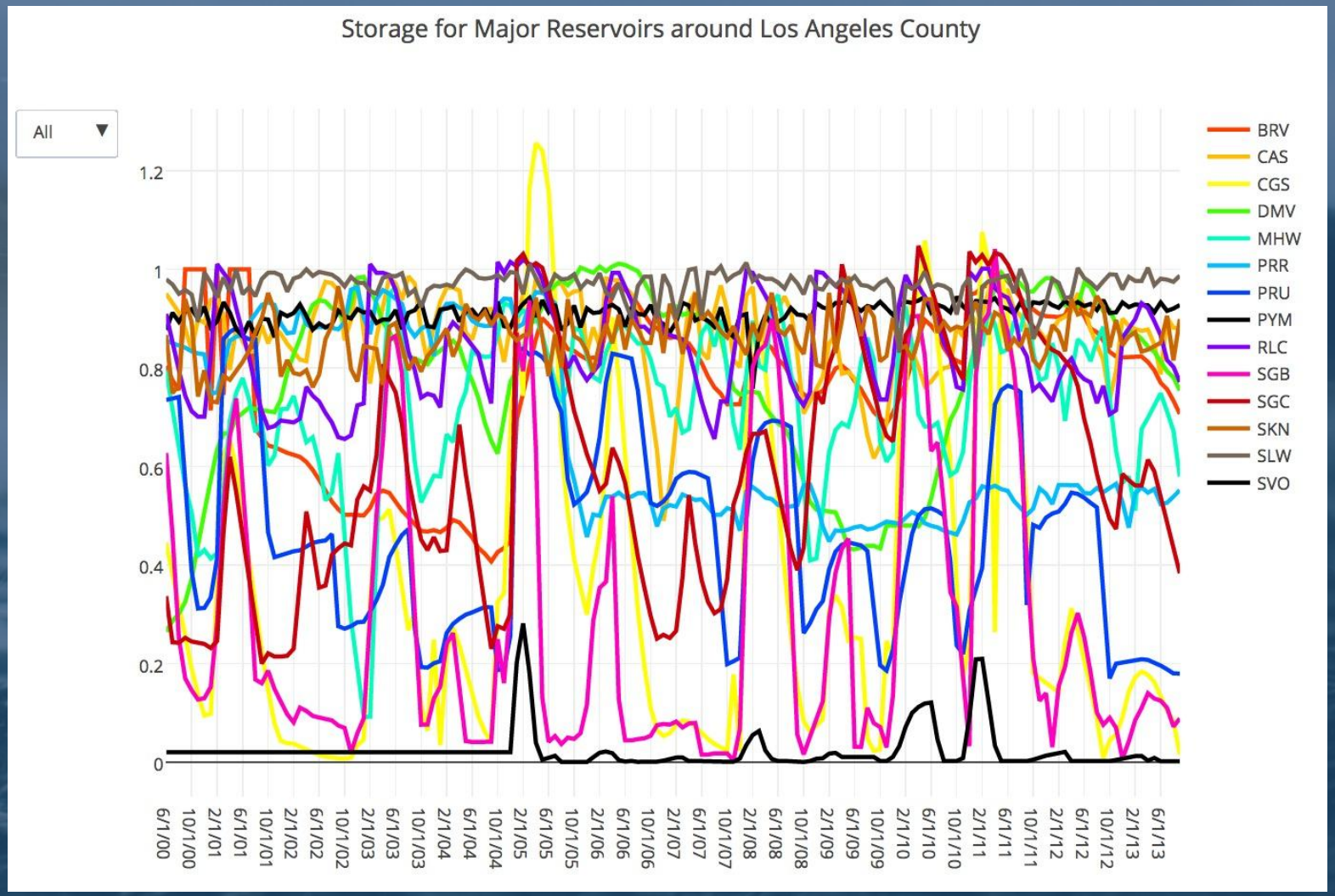
Data sourced from Berkeley Earth - Climate Monthly Average Temperature around Los Angeles.

- **Average Temperature from 2000-2013**
- **Average Temperature error**

Additional information sourced from Wikipedia: Elevation of the dam, year completed, dam type (material), heights (in feet and meters), capacity (in feet and meters.

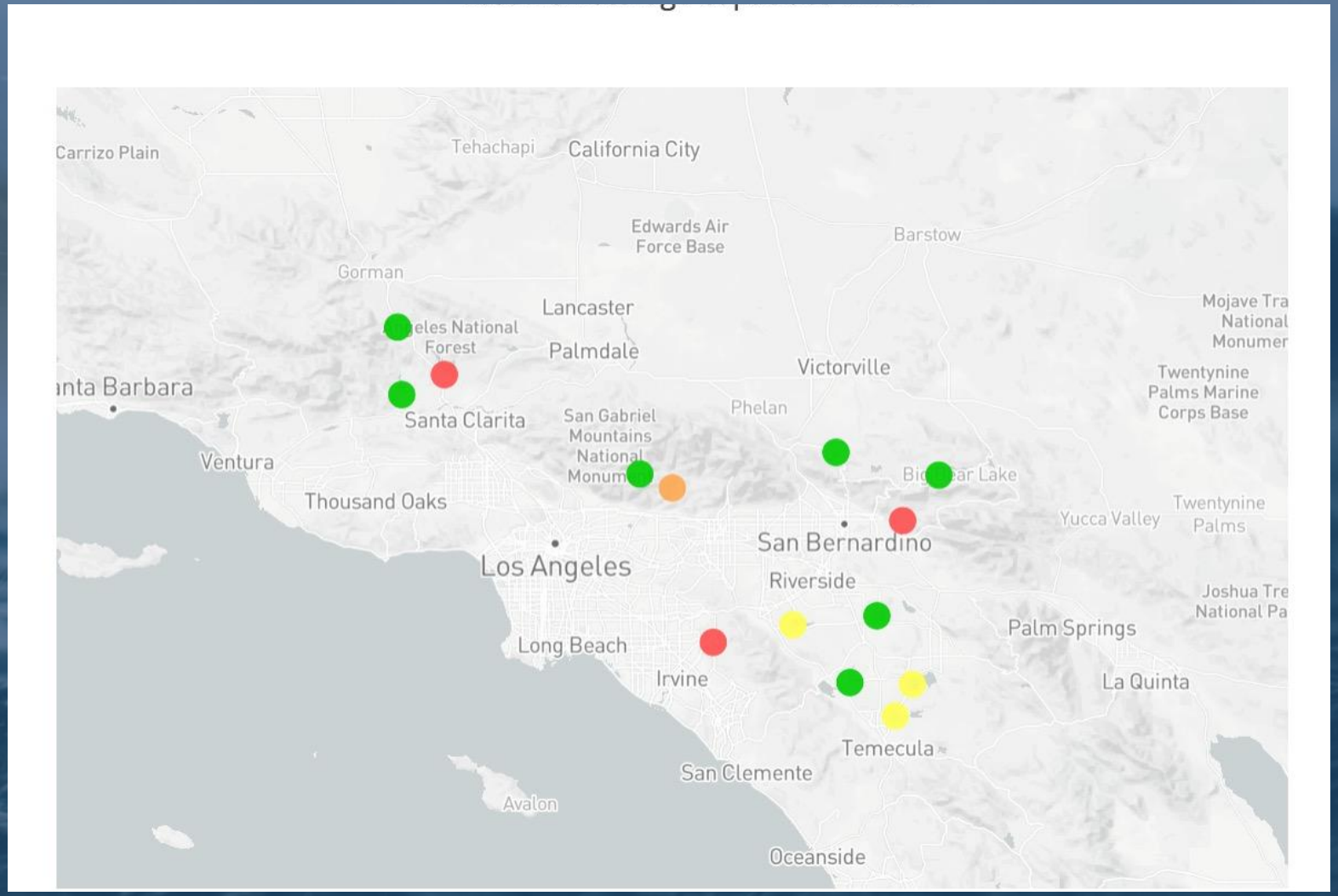
Interactive Visualization 1

https://plot.ly/~atomahawk/48/storage-for-major-reservoirs-around-los-angeles-county/



Interactive Visualization 2

https://plot.ly/~atomahawk/50/reservoir-storage-capacities-in-2001/



Best Performer: Random Forest Classifier

- Max Depth: 3
- Number of estimators = 3

Table 1: Classification Report for each Class of Reservoir

Class	precision	recall	f1-score	support
0 (< 55%)	0.88	1	0.93	7
1 (< 70%)	0.75	0.33	0.46	9
2 (< 85%)	0.77	0.94	0.85	18
3 (> 85%)	1	1	1	67
avg/ total	0.93	0.93	0.92	101

Limitations:

- **Assumes business as usual water demand**
  - No natural disasters (wildfires and earthquakes)
  - A static population size (& consumption rate)
  - Unchanging urban, ag, and environmental uses

- **Limited to reservoirs that had data available on CDEC**
  - Large reservoirs that had recent CDEC data (2000-2017) were used in this analysis (as recent years give context to contemporary population size, consumption, water demand, etc).
    - Not completely inclusive of all reservoirs in the LA area.
  - Many reservoirs had storage data dating back from the 80s to 2001 only - Why did they stop recording monthly storage data?

Issues and Improvements

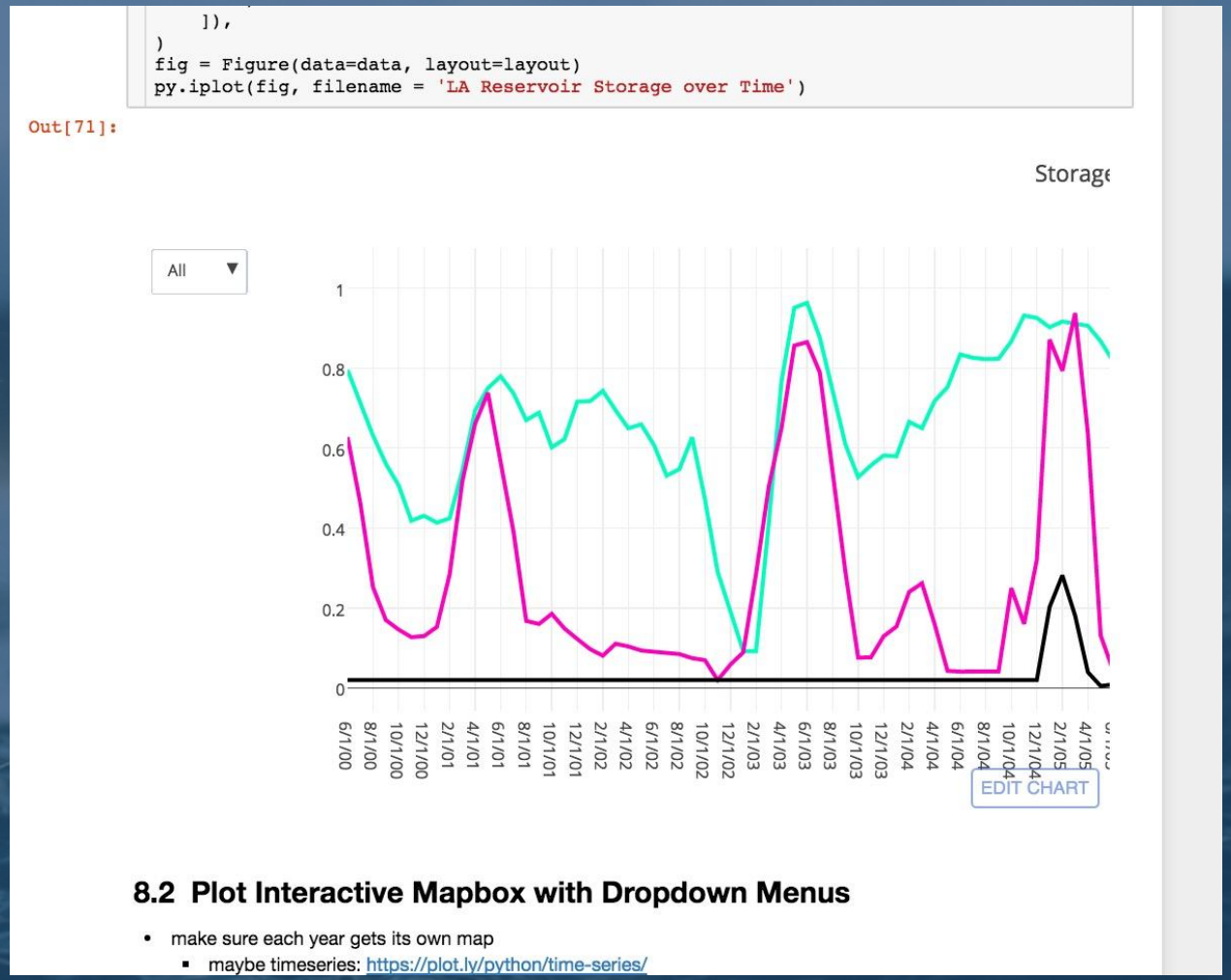
- **Regression?**
- **Validate the model with a holdout set (2013)**
- **Outlier and pattern analyses**
  - Maybe drought has a specific pattern?
- **Investigate class imbalances**
  - Recent data has hit historical lows
- **Investigate data leakage? Eliminate summer data?**
- **Interactive slider for map overlays**
- **D3.js implementation**
  - Web app that predicts classes and overlay them on map

APPENDIX

Thoughts on Plotly:

```
    ]),
)
fig = Figure(data=data, layout=layout)
py.iplot(fig, filename = 'LA Reservoir Storage over Time')
```

Out[71]:



8.2 Plot Interactive Mapbox with Dropdown Menus

- make sure each year gets its own map
  - maybe timeseries: https://plot.ly/python/time-series/

Disclaimer on Plotly:

Limits on Requests to Plotly Servers

The API limits are as follows:

API client	Exceptions	Per Day	Per Hour	API Version
Python, Arduino, R, MATLAB	None	50	30	V1
Plotly Web App (plot.ly)	GET requests are not limited	1000	None	V2