Big Data (AES 690) Peer Reviews

Reviewer: Mitchell Dodson

March 22, 2024

Final Scores:

1. Good/Fair
2. Very Good
3. Excellent/Very Good

Review of Proposal 1

The proposer plans to use a trained convolutional neural network (CNN) to approximate a sub-hourly time series of analytically-calculated sensible heat flux values over a spatial grid centered on Madison, Alabama. Next, the proposer intends to compare model predictions and calculated values over the domain in order to infer surface properties. The analytic formula the model approximates is given in terms of the vertical temperature gradient, so it probably uses an approximation for vertical heat flux similar to the bulk aerodynamic formula $w't'_c \approx C_{DH}\,U_r\,(T_s - T_a(z_r))$ [1]. Although I'm not sure, the CNN architecture they're referring to probably applies a pixel-wise 1D convolution along the time axis (since the grid size is undetermined, and 2D CNNs generally require a specific input domain shape).

One strength of this approach is the concept of using a neural network emulating a theoretically robust model as an analysis tool. Since neural networks are limited by the biases and expressivity of their input data toward predicting the desired output, they may be indirectly used to study the coupling between variables, and to identify situations where the predictors aren't sufficient for explaining variance of the outputs. There's a wealth of recent literature applying deep learning techniques to anomaly detection tasks, where the general idea is to train a model to predict observations under nominal conditions. Once trained, major spikes in the model's prediction error given new data are an indication that the inputs are anomalous [2]. An interesting potential direction of inquiry could be to investigate whether the model's prediction error correlates with regions' surface roughness, since the drag coefficient C_D in the above formula isn't listed as an input, but can vary by a factor of about 4 [1].

The overall methodology for this project is interesting and reasonable, but the proposal doesn't contain very many details about the datasets, model configurations, or analysis techniques the proposer intends to use. I was left wondering about the spatial and temporal resolution of the domain, whether the data from Google Earth Engine are model-driven or observational, and which land surface properties will be investigated. Nonetheless, the proposed plan is feasible within the given amount of time, and the project seems very exploratory in nature, so I am curious to see how the direction and scope evolve; I believe there can be many creative and novel ways to analyze the results.

Also, consider that the minimum and maximum normalization bounds need to match those used when training the model. Neural networks tend to closely replicate the value histograms of their training data, so small offsets might significantly affect overall prediction accuracy.

[1] D. L. Hartmann, Global Physical Climatology. Newnes, 2015.

[2] K. Choi, J. Yi, C. Park, and S. Yoon, "Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines," IEEE Access, vol. 9, pp. 120043–120065, 2021, doi: 10.1109/ACCESS.2021.3107975.

[3] M. Romaguera, R. G. Vaughan, J. Ettema, E. Izquierdo-Verdiguier, C. A. Hecker, and F. D. van der Meer, "Detecting geothermal anomalies and evaluating LST geothermal component by combining thermal remote sensing time series and land surface model data," Remote Sensing of Environment, vol. 204, pp. 534–552, Jan. 2018,
doi: 10.1016/j.rse.2017.10.003.

Review of Proposal 2

Given the recent advent of high-resolution tropospheric NO_2 data availability from the new geostationary TEMPO spectroradiometer, the proposer suggests to identify spatial clusters of the trace gas using K-means in order to discover consistent emission sources surrounding Chicago during November, 2023. After determining cluster centers, the centroid locations will be directly co-located and compared with near-surface observational data collected by a SeaRey amphibious propeller plane. The procedure involves collecting a series of geolocated hourly tropospheric NO_2 data (partial volume fraction), clustering the data along the three axes, then validating TEMPO observations against the juxtaposed and simultaneous in-situ measurements.

In my opinion, this proposal is quite novel both because the TEMPO and SeaRey data are new and rare, and because I haven't often personally seen K-means used as a technique for clustering data alongside spatial coordinates. This unique approach serves as a general and relatively simple way of identifying regional groupings in data, and it makes sense to use centroid-based clustering to identify spatially sparse point sources of pollutants like NO_x. I think it would be interesting to see how the results change if clustering is run multiple times over several discrete NO_2 bins, or using custom minimum thresholds. In my experience with K-means it sometimes helps to "overshoot" the expected number of clusters, then manually merge the ones that don't correspond to meaningfully distinct centroids.

I wonder if you could sub-classify different emission sources by adding the TEMPO experimental formaldehyde and O_3 from EarthData as predictors?

As a reviewer, I appreciated that you explicitly enumerated your objectives. I think the scope of this project is appropriate for the amount of time available, and I'm very excited to see the results.

Review of Proposal 3

The proposer seeks to investigate the relationship between a land cover map and satellite-derived MODIS (MxD21) land surface temperature (LST) and emissivity (ϵ) data using techniques including clustering and decision trees.

In my opinion, this proposal has a high degree of scientific merit because it makes a plausible argument that emissivity and LST are valid predictors for land cover type, and uses examples in literature to justify the decision to use clustering and decision tree algorithms.

Using the MxD21 product for this task is a great choice since it uses spectral contrast (the TES algorithm) to calculate LST and ϵ independently for each pixel, unlike the common MxD11 product which determines its emissivity values based on an a-priori land cover map. I'm very interested in what the LST/ϵ clusters will look like, since they represent two of the most fundamental radiative quantities. Emissivity is a material property, and temperature is a transient state, so there's a lot of information contained in just two numbers.

I'm curious what it would look like if you acquire 1km MODIS L1b or L2 products and use a few visible or near-infrared bands as additional inputs to your clustering and decision tree algorithms. These bands provide much more information characterizing water, vegetation, and other surface types. The additional inputs might make your clusters and decision trees more expressive without introducing much bias (since the MODIS pixels are directly co-located and empirically derived from basic radiances).

Covariance matrices of pixels within each cluster are useful for investigating the data characteristics that determined the classification result. Also, in my experience with K-means, it sometimes helps to "overshoot" the number of clusters, then manually merge the ones that don't correspond to meaningful differences in surface classes.

The schedule for the project was explained in great detail, with specific tasks bulleted between deadlines. These were very helpful for understanding the methodology, and demonstrate that the proposal is realistic for the alotted time.