Overview

In an age of economic disparity and cynicism, we believe the dignity of work and upward mobility are vital to individual happiness, economic growth, sustainable development, and community building. We are Philipp, Helen, and Hunter, #Satellite, #Commerce, and #TheForce. We are data scientists and engineers, here answering the burning question: is it even possible, that we can use satellite imaging data, to preempt lagging economic indicators reliably? We are proud to work together to help geolocate, discover, and predict leading indicators of economic and labor trends. We believe that such objective information and preemptive knowledge - without socioeconomic and demographic assumptions prevalent in surveys and census - would help communities to respond to upcoming challenges and opportunities for their economic well-being.

Mainly using random forest for regression, we predicted employment data at the granular county-level. After collecting labor force statistics, county-area data, and energy production maps, we built and validated our model, using luminosity as a high-resolution estimate for employment data. Employment data, as a leading indicator of economic activities, precedes booms and busts throughout history – and now we precede employment data!

You can see some of our final results here:

- Interactive prediction of employment:
- Visualization of historical luminosity:

We truly hope to continue work on this prompt, and aid policy-makers at the Department of Commerce make more informative decisions.

Algorithm

We focused on a two fold system:
Initially, we built quick visualizations of locations and temporal employment data by county. Going further, we investigated ensemble learning and predictive algorithms to test our initial intuition.

Geolocation of Light and Labor

To geolocate night lights, we first extracted county names from NACIS naming system, then matched them to image luminosity and GeoJSON information. We then gathered county-level employment data and area data, adjusting for light intensity and labor force density.
Early on, we noticed some image-processing challenges: pixel over-saturations at high-luminosity level, projection distortions at higher latitudes, and missing data. After consulting experts from Department of Commerce, confirmed that these challenges are inherent to satellite imaging.

As the county-level data was structured specifically for exact county-state match, we utilized it for high-resolution match. We used both R and Python for quickly check for images and data points, and smoothed and normalized both night-light luminosity and employment data based on county areas and changes over time.

To predict employment level from pixel-related values and luminosity on county level, we tested two methodologies. First, a 5-fold cross-validation model of random forest regression, built with 150 trees and selected features. Second, we leveraged a CNN approach to construct more latent features for prediction. Due to the time constraint and lack of computing power, the second method we did not finish completely.
Work can be seen here.

Employment data prediction

Our strategy was to start with all original features and engineered features, such as delta in luminosity and employment data. Then we use a random forest to help select our features and generalize our discoveries with regression. This allows us to flexibly explore the feature space, generating feature sets – informative and essential – to get a broad sense of applicability in our cases.
The computation was quite intensive, so we ran a local 3-core cluster, each of 3.2 GHz for 2 hours. Our dataset was split up for a 5-fold cross-validation, both correlation R^2 and MSE are consistent, with correlations high and MSEs low.
For final analysis, we present interactive visualization and output predicted employment data.

Results

- Our dataset has approximately 3,000 counties, 17 months of imaging data, and 18 months of employment data. With this dataset, we predicted employment density on a county-by-county level by luminosity, with correlation score of $R^2 = 0.96$ on average, with very little fluctuations across validation tests – meaning our selected features are informative and essential to our predictions.
- Our predictions reflect both employment and energy production, as we anticipated from the major sources of night-time light(human settlement, fires, gas flares, and fishing boats).
- Our **interactive visualization at:** shows interactively with historical records, the employment density level and luminosity level

Our validation and tests convinced us: our model is a well-generalized predictive algorithm, representing leading changes to lagging economic activities.

Future Work

The impact of this project extends beyond this hackathon. Building on our basic algorithm to predict leading indicator for lagging economic and labor trends, Bayes Impact and Department of Commerce have the opportunity to empower and impact many communities and people.

Moving forward, we recommend a two-fold approach for improving prediction.

1. For efficiency, data sources should be integrated and then automated in mapping. Our ad-hoc integrations were good for quick starts, but not scalable.
2. For accuracy, 1) image pixels can be re-sampled and re-calibrated to avoid bias and over-saturation; 2) projections can be updated for better representation. Sometimes accentuated low-intensity pixels are indicative of upcoming human activities, which can be surfaced by our light intensity normalization. 3) we would like to finish our convolutional neural network model