Socioeconomics of Light:

Geospatial analysis associating nights lights and consumer sophistication

Saifullah Rais, Nic Vogler, and Heather Koo

W205 Final Project Report

Overview of Project

Building on the relationship between night lights and economic development, we have attempted to build a real-time economic monitoring dashboard using Google's Earth Engine Platform. This project was inspired by the work done by NASA in building the Night Lights Development Index (NLDI), and we have extended the association to different dimensions of consumer sophistication.

Geodemographics tackles ecological fallacy arising from regional averages, especially in developing nations which usually witness high levels of inequality and geo-diversity. Information gained from the use of these frameworks has enabled policy makers to understand on-the-ground conditions in near real-time, capture inflection points, and determine more clear results of policy actions.

We have created SoLi, a real-time economic monitoring dashboard, utilizing Google's Earth Engine Platform. SoLi is trained using Tensor Flow to associate spending patterns and night lights, to offer a resource for agile policy decision-making. The aim of the project is to determine real-time measures of India's economic activity - because such data is typically difficult to obtain at regular, frequent intervals.

Acquisition and Organization of Data

We acquired and used the following data sets as the basis for our predictive model:

National Sample Surveys (MOSPI, India)
This data includes samples of households across India, including household characteristics such as location (state, district), and size, as well as average monthly household consumption. We uploaded this data set onto Google BigQuery. This is a structured data set which is collected every 3-4 years. We collected the latest year (2014) for our model.

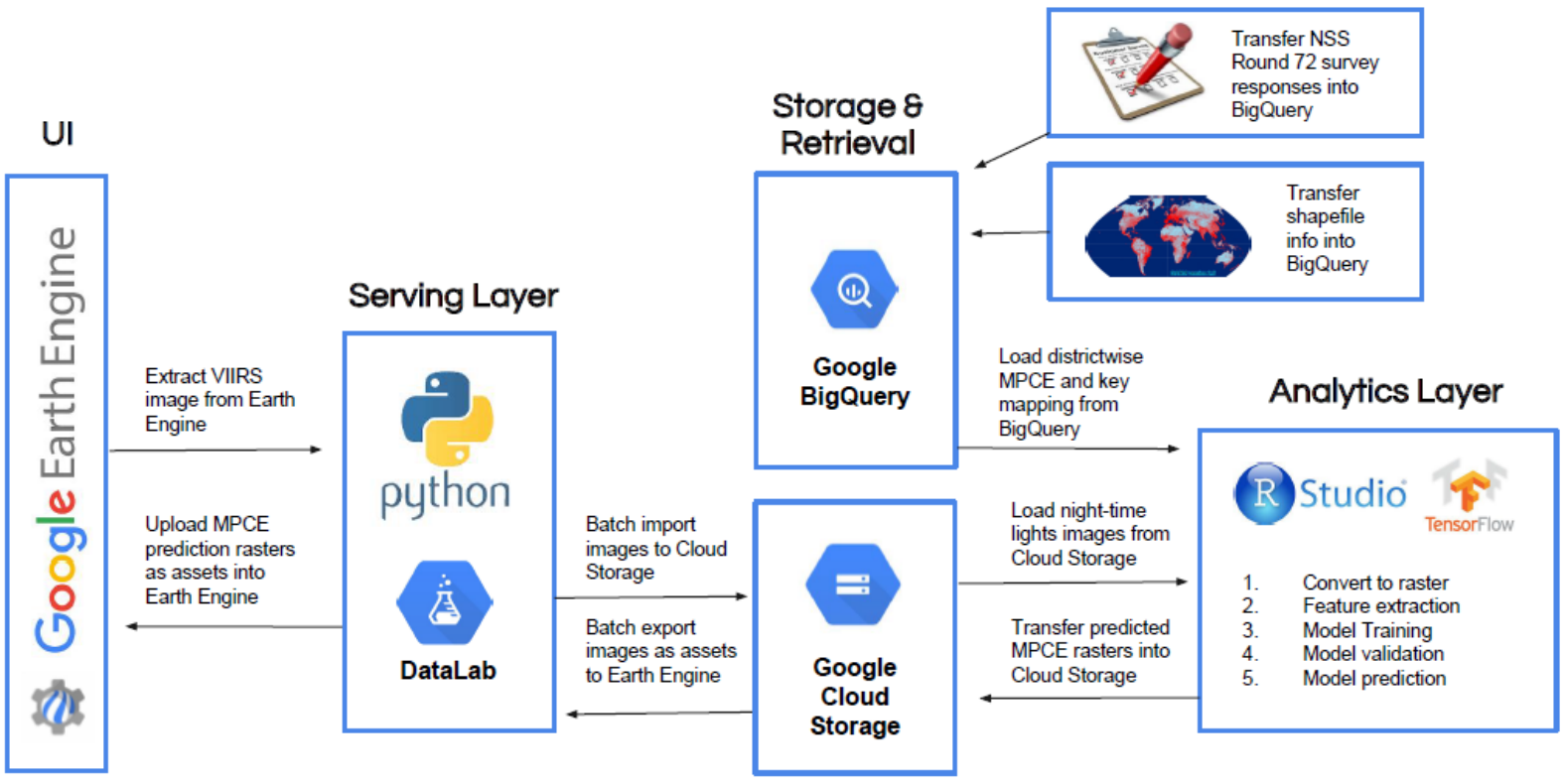
Satellite Images from VIIRS
Extracted from Google Earth Engine, we extracted TIFF files containing monthly average radiance composites of nighttime lights captured by the Visible Infrared Imaging Radiometer Suite (VIIRS). This data is semi-structured, and we collected all available months of images from January 2014 through September 2017 (there is a 2 month lag in data available).

Shapefiles from GADM, the database of Global Administrative Areas

We downloaded publicly available files through ArcGIS which store geographical information and features. The Indian shapefile was used to aggregate average radiance composites across states and districts. This structured data was stored both on Google BigQuery, and on the Google Cloud Storage drive.

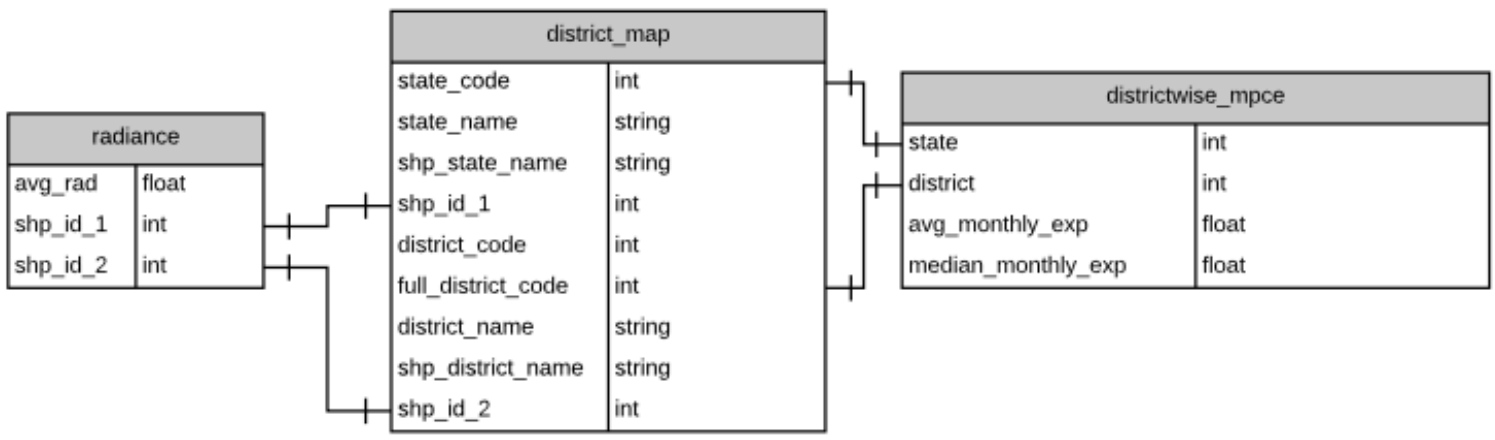
Overall Architecture and Implementation Details

The overall architecture and process flow of the project is summarized in the diagram below:



Data Ingestion

- **NSS survey (MOSPI)**: We ingested the household characteristics survey into BigQuery and focused on the *median and average monthly consumer expenditure, and Gini coefficients* on a district level. The Gini coefficient, a measure of income dispersion (with consumption being a proxy for household income), was calculated in BigQuery using individual household consumption data. The initial data relationship is shown below:



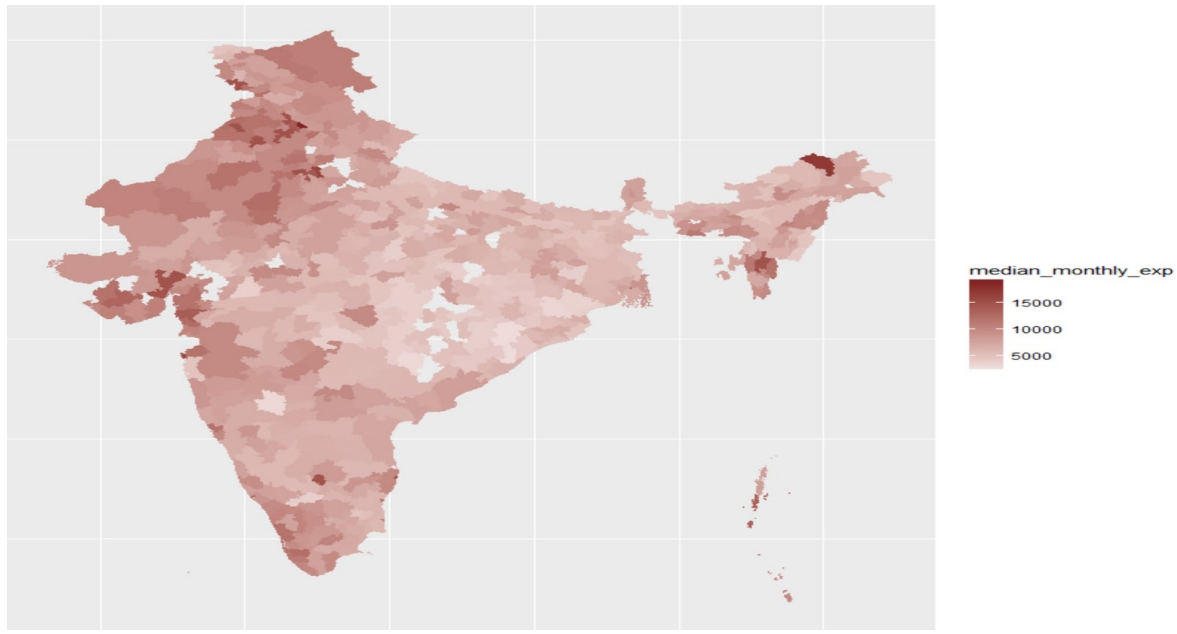
We created a crosswalk between the shapefile and consumption data by joining on state name and district name. Additional matching had to be performed for ~21 districts which had varying spellings, or differing naming conventions between the two data sets.

- **Extraction of VIIRS Satellite Images from Earth Engine:** We used an iPython notebook on a Google Datalab virtual machine to access the Google Earth Engine API and pull all of the data from the VIIRS Stray Light Corrected Nighttime Day/Night Band (Version 1). Our initial pull was for all available data; subsequent pulls are for only the most recent data set. These datasets are processed (applied a gamma correction and converted to 32 bit and 8 bit pixel formats), and then transferred to the project's Google Cloud Storage.

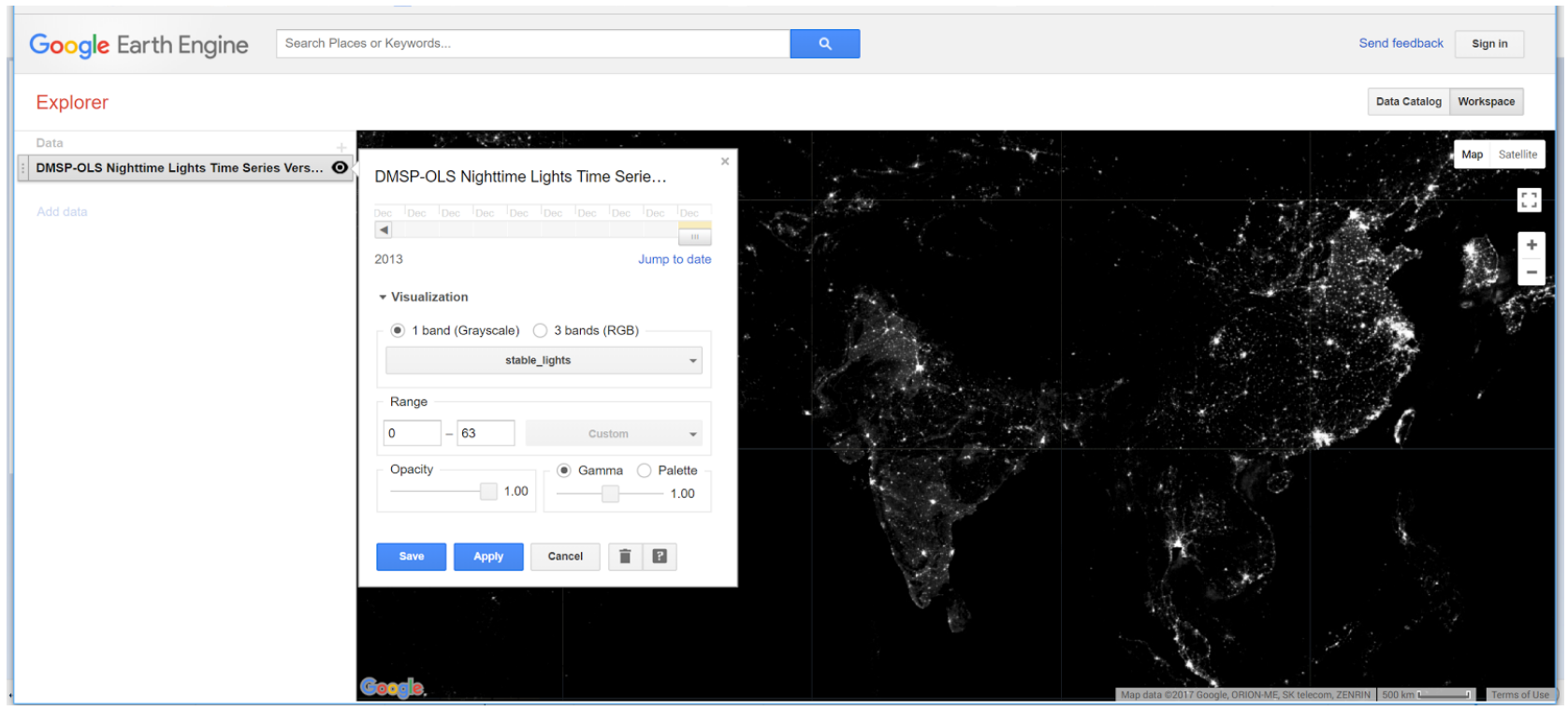
Data Processing

The project required working with 2 different data formats (i.e. images and tables) and 3 different sources of data (i.e. GADM, NOAA and MOSPI). One of the main challenges we faced after ingesting the data onto storage platforms for use was processing the data transformation:

Tables to Images: We converted survey data (MOSPI) to raster images after linking district consumption data to the shapefile and training our predictive model. This was loaded onto the Earth Engine platform using the Python API. The distribution of median consumer expenditure across 642 districts can be visualized as:



Images to Tables: We also built features for analysis within the model by aggregating pixel intensities for specific districts. We extracted, processed, and aggregated the pixel information using R and stored it in Google Cloud Storage. Below is a sample of a monthly satellite image available from the Earth Engine Platform.



Predictive Modelling

We used Tensor Flow Estimators in our model, an open source software library for numerical computation using data flow graphs. This gave us the flexibility to scale and deploy computation across servers with an API. The model was trained using a Deep Neural Network regressor using satellite imagery and the 2014 NSS Consumption Survey. This model gives us the capability to predict and estimate temporal variations in monthly expenditure.

The model's current feature set comprises of 16 descriptive statistics of average radiance within a district, and is trained for ~30,000 steps. The output is stored as

Geotiff files, which are uploaded as assets into Google's Earth Engine Platform.

Upload to Earth Engine
After images have been processed through the predictive modelling application in R, the results are stored in a post-transformation folder, staged for being uploaded into Google Earth Engine through it's Python API.

Project Results, and Next Steps

Using the consumption data and VIIRS satellite Nighttime Lights images, we were able to create a model which explained ~35% of the variance before cross-validation deteriorated. Using this model, we are able to estimate each district's average household consumption for years without survey data.

This project created a process to use night light imagery to nowcast consumption characteristics by district. Further improvements to the application can be made with additional feature extraction. The current model may also be improved by converting our continuous consumption variable into categories, in accordance with Development Economics literature. Additionally, we can expand the analyses by examining other measures of economic activity, such as asset ownership. While the basic framework is now in place, further research and work can add functionality and predictive power for a variety of socioeconomic characteristics, allowing for better and more informed policy decisions.