Monitoring Electricity

Project Technical Description

# Project Description

## Background

Previous studies have increasingly shown that night lights data from satellite imagery can be used as a proxy to measure lights emissions and reflections from the earth surface. This in turn can be used to monitor long term changes in energy caused by electrification, urbanization and other human activities. Also, the same technique can help to measure power outages caused by disasters as well as due to grid failure which essentially can help in measuring in grid performance. There is a promise that night lights satellite data will become more available and reliable (with better processing) due to efforts from a team at NASA[[1]](#footnote-1). Currently, the most reliable free satellite nights data comes from the NOAA NPP mission. The data come from the VIIRS instrument onboard this satellite which collects images at night. Since this data comes at no cost, in this project we will first attempt to do the augmentation using this data.

## Project Goals

1. **Clean data and impute missing values:** I have made reasonable progress in exploring the data but haven’t yet run any models to check whether ML can be use to impute the missing values.
2. **Augmenting satellite night lights data with electricity monitoring:**  The premise is that satellite nigh lights data could become more useful and accurate if we complement it with ground truth data.

# Description of approach for augmenting night lights data

From here onwards, the document describes how the task of augmenting the satellite night lights data will be carried out:

## What will be measure?

The key goal is to be able measure reliability of the electricity grid in Tajikistan from satellite night lights. The following indicators will be used to measure reliability:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index name** | **Unit** | **Data type** | **Definition** | **Comment** |
| Light intensity | Radiance/sq.km | Night time satellite image or already calculated light intensity values | The idea here is to get an indication of brightness of an area as a proxy for extent of power usage | This seems like an easy indicator to measure but the actual interpretation is open for discussion. |
| Number of power outages per unit of time | Outages/(day, wk, month) |  | As suggested by the indicators, the intention is to measure how often outages occur. | Need to clearly define an outage　from the image side. For example, we can use a threshold on light intensity. |
| Proportion of of light area | 0-1 |  |  |  |

## Required Data

### Ground truth data

This data will come from the survey-the data being collected from households. Roughly speaking, the final output data which will be required for this process should look like this:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Area-id** | **x** | **y** | **Time (with resolution ‘hourly’)** | **Power-status** | **other** |
| A grid (Easiest is to use Enumeration Areas) representing area where a single data point is representing because from the satellite data side, we will have data up to say 750-meter resolution. In some cases, the data is available at specific Lat/Lon. | Longitude | Latitude | This can be changed but hourly seems to be a reasonable choice since satellite is collected once a day. Also, this is where most computations will occur in order to match survey data with satellite | Whether power is on or not | For example, whether the power status is actual or extrapolation (prediction) |

### Satellite data

This is the data we need to augmenting. They seem to be two types of data from the NOAA VIIRS site as follows:

* Point (with lat/lon ) data of light emission measurements derived from nightlights imagery: The density of these points isn’t high enough to guarantee that the we find data for all our areas of interest. I have explored these files but haven’t downloaded them yet.
* Satellite imagery: This is just the raw imagery captured at night. My current understanding is that the highest possible resolution for the VIIRS is 750 meter. This is the best type of data because we would be able to compute our own measurements of light emissions as well decide the spatial resolution (how to grid the data). The free data can come as monthly composites or daily

## How it will be done?

The project is presented as to have 5 phases/stages as shown in Figure 1 below. These stages are briefly explained below:

### Data capture

This refers to the ongoing process of monitoring electricity in Tajikistan which started in November 2016 and is ongoing

### Data preparation

This phase is illustrated in detail in Figure 2. Before any kind of advanced modelling is done, we need to prepare the data so that it’s in the kind of format where we can do modelling on. The

Figure 1: Process flow for the whole project

Raw data from field

Regularization, imputations, cleaning

-Imagery

-Processed point data

Download data

Data download parameters

1. Time: monthly, daily
2. Area of interest
3. Type: Raw image/preprocess csv
4. Period: Specific dates

NOAA servers

Preprocess images (e.g., extract Total Nighttime Light (TNL)

images

Process text data

Base geographic data

-unit of analysis: Enumeration area/grid (e.g., 1km)

-Administrative boundaries

Join/Harmonize ground truth and image data

Ground truth data at chosen unit of analysis (e.g., EA)

Text data

Extrapolation engine

Image/text data with light intensity values at chosen unit of analysis

Data ready for model building

Figure 2: Data Preparation phase

output from this stage is two types of datasets both at same geographic level. This work is in progress.

### Model building

In this phase, we will use Machine Learning (ML) models to see how best the data from satellite imagery predicts the ground truth data for some selected task. One way to do this is to create a training dataset where for each unit of analysis where we have both (the measurements from satellite imagery) and the ground truth data and then train a model based on this data. This model will be tested and used to come with predictions in other areas where data isn’t available.

#### The power outage model

This is perhaps the easiest type of model I can think of right now.

**Objective:** Given satellite night lights data (e.g., in terms of light intensity- Total Nighttime Light (TNL)) can we predict from this data (image) whether there was a power outage in an area of interest x (defined by a geographic unit of analysis (e.g., enumeration area boundary or grid).

**Application:** JP should assess whether such a model could be useful for policy

**Usage:** If properly trained and accuracy is acceptable, we could deploy this model to other regions of Tajikistan where there is no ground truth data. The typical use is that assuming we want to measure independently (how many outages occur per unit of time without consulting the power company), this type of model could give us such answers

**How to come up with the model**: This process is illustrated inFigure 1**.** While inFigure 2presents a single record in the training dataset.

Figure : Model building process

Preprocess (any specific preprocessing as required)

Data

Generate training data

Build and test model

Figure : Dummy training data record. Note that in this case it seems reasonable to drop the time dimension since we are interested mapping association between pixels and some results which doesn’t essentially depend on time.

1. Location: z
2. Time-stamp:
3. Total night light: 780
4. Outage: Yes/No (from survey data)

z

# Anticipated challenges

This process is RD in nature and so there is no guarantee that the results we will find will have reasonable enough accuracy to be useful.

# Work plan, timeline and milestones

The work plan is presented in Table 1 below.

Table 1 : Work plan

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Phase/Task** | **Responsible** | **Status** | **2016** | **2017** | | | | | |
| **Nov-Dec** | **March** | **April** | **May** | **June** | **July** | **August** |
| **Data capture** | | | | | | | | | | |
|  | Project management | Trevor | On-going |  |  |  |  |  |  |  |
|  | Technical planning | Dunstan | On-going |  |  |  |  |  |  |  |
|  | Data capture | JP | ongoing |  | Started-ongoing | | | | | |
| **Data preparation** | | | | | | | | | | |
|  | Data extrapolation | Dunstan | started | April 21 |  |  |  |  |  |  |
|  | Deploy extrapolation engine | Dunstan/JP |  |  |  |  |  |  |  |  |
|  | Consultations with NOAA on imagery pricing | Alexei | started |  |  |  |  |  |  |  |
|  | Imagery specifications (time periods, wavelengths (bands), cloud cover, moon light, levels of preprocessing, areas of interest etc.) | Dunstan | started |  |  | April 20 |  |  |  |  |
|  | Imagery acquisition (Actual downloading of images) | Dunstan (include Alexei?) | Not-started |  |  |  |  |  |  |  |
|  | Imagery checks to ensure correct imagery was providing (check for time periods, resolution, processing etc) | Dunstan | Not-started |  |  |  |  |  |  |  |
|  | Processing imagery to get light intensity related variables | Alexei? /Dunstan/NOAA? | Not-started |  |  |  |  |  |  |  |
|  | Acquisition of base maps from Tajikistan | JP | ? |  |  |  |  |  |  |  |
| **Model building and evaluation** | | | | | | | | | | |
|  | Generating training g data | Dunstan | Not-started |  |  |  |  |  |  |  |
|  | Try out different models | Dunstan | Not-started |  |  |  |  |  |  |  |
|  | Evaluate models | Dunstan | Not-started |  |  |  |  |  |  |  |
| **Tool packaging** | | | | | | | | | | |
|  | Publish and open source code on Github |  | Not-started |  |  |  |  |  |  |  |
| **Raise awareness about tools** | | | | | | | | | | |
|  | Organize events (lighting talks and other | Dharana | Not-started |  |  |  |  |  |  |  |
|  | Publish article(s) | Dharana | Not-started |  |  |  |  |  |  |  |

# SMS Workflow for Data Preprocessing

Figure : Workflow for data preprocessing

Google-Drive

sms.xml

Convert to CSV

Use in STATA

./SMSBuckupRestore

Daily backups

sms\_”date”.csv

Pre-process

Fix timestamps, remove duplicates etc.

Regularize (time-interval = 1 hr.)

sms\_”date”\_v1.csv

Identify missing events (type of event = message, method = ‘strict’, ‘relaxed’, etc.)

sms\_”date”\_v2.csv

sms\_”date”\_v2.csv

Regularized dataset with label for missing

OneDrive:

../processed\_sms/

OneDrive:

../raw\_sms/

Evaluate imputation model (model = ‘moving average, ‘random forest’ etc)

sms\_”date”\_v2.csv

Accuracy >= threshold (e.g. 80%)

NO

Impute data (model = ‘model’, input\_data = sms\_”date”\_v2.csv)

Yes

sms\_”date”\_v3.csv

Regularized with missing events imputed

1. https://www.nasa.gov/feature/goddard/2017/new-night-lights-maps-open-up-possible-real-time-applications [↑](#footnote-ref-1)