Night Lights and Noisy Data - Using Machine Learning to Better Detect Human-Generated Light

Johannes Coetsee[a]

[a] *Stellenbosch University*

Abstract

In recent years, researchers in the social sciences have increasingly relied on the usage of alternative sources of data with which to answer research questions. One of the avenues remote sensing data, and more specifically, satellite data, as . One of the most prominent of these datasets, the stable lights product from DMSP-OLS, has been especially helpful. However, this product suffers from some potentially severe accuracy-related problems. The following paper explicates on some of these, and implements an adjusted version of the noise-filtering methodology proposed by Määttä & Lessmann (2019), in which the authors use a variety of derived and externally sourced remote sensing inputs to inform a Random Forest algorithm that categorises night-light data as being human-generated. This methodology proves fruitful in preserving data at the lower end of the light spectrum, which is often discarded due to it being too noisy.

Keywords: Remote Sensing, Night Lights, Random Forest

1. Introduction

The use of remote sensing data, and more specifically, satellite nighttime light data, presents potential for new and diverse applications in socioeconomic research. Nightlight data remains a largely objective measure, and is thereby suitable to use as a proxy in a broad array of studies that require the usage of potentially unreliable or otherwise lacking data. This advantage is especially pertinent in parts of the developing world, where socioeconomic research can prove most beneficial. There exists different night lights products which can be utilized towards this end, the most common of which is the 'Stable Lights' product, derived from the Defense Meteorological Satellite Program's (DMSPs) Operational Linescan System (OLS). This paper emphasises the usage of this product specifically focusing on its shortcomings. Most prominently, DMSP-OLS has difficulty in separating background noise from night lights generated from human-generated light, especially in areas that display lower levels of night light intensity. This presents an obvious problem: analyses that attempt to use Stable Lights as a proxy for economic activity, for instance, would exaggerate or understate economic activity in these low-luminous areas.

Email address: 19491050@sun.ac.za (Johannes Coetsee)

This paper attempts to address the challenge of inaccurate measurement of night lights by applying a filtering technique to identify and separate nightlights emitted by humans from those emitted by anything else. This filtering process is based on the methodology for deriving the 'Local Human Lights' product by Määttä & Lessmann (2019), and relies on a Random Forest (RF) Machine Learning algorithm for classification.

2. Explication of the Problem

2.1. Stable Lights and Economic Activity

The most prominent difficulty, however, relates to the amount of noise in the lower end of the light distribution due in part to the blooming effect mentioned above. Standard practice using the stable lights data set is to discard these values from analysis, thereby removing a large proportion of cell observations.

3. Method and Data

- Short overview of Määttä & Lessmann (2019)

The filtering process used by Määttä & Lessmann (2019) attempts to more accurately detect specifically *human-generated* night lights

3.1. Data

Table ref{variables_table} below gives an overview of the most important inputs

“- Global Human Settlement Layer (GHSL) built-up grid. It is based on Landsat satellite images and prepared following the GHSL methodology

-The most important variables are the average visible band (avg_light) and frequency of light detection (freq_light) images from NOAA. They provide information on average light over one year and the percentage of days with light detections in a pixel.

- In addition, we generate variables that describe light characteristics around a pixel. The regional input variables address concerns that the background noise in light data differs across regions. We use a local_noise variable to identify regions, where the background noise in avg_light is systematically higher. The variable counts the number of pixels below the value 6 in avg_light image in a square window of 499 × 499 pixels for all pixels.

- For the Local Human Lights product, we add more variables on local light characteristics to maximize the classification accuracy. We create these variables by taking averages of avg_light and freq_light on different area sizes around a pixel. For example, lm_avg_199 calculates the average of avg_light in a 199 × 199 pixel area around a pixel. Similarly, we create the variables lm_freq_5, lm_avg_25 and lm_freq_99. The reason for adding local averages on different area sizes is that we do not know the correct radius of spatial correlation in human-generated light. ”

3.2. Methodology

Maatta's methodology necessitates the following steps:

First, a host of regional night light variables are created from two of the original DMSP-OLS products. These variables constitute some of the most important inputs of the RF algorithm.

The second step in Määttä & Lessmann (2019)'s methodology is related to the division of the world map in various sub-regions, with the algorithm

5. Local Human Lights Process

Our second product shifts weight from minimizing regional bias to maximizing the accuracy in human-generated light detection. We utilize the regional light characteristics in the filtering process in two ways.
we divide the world into sub-regions and run the algorithm separately for each region. This allows the algorithm to learn how the light characteristics relate to built-up areas in different sub-regions. Since that relationship most likely differs across the world, we achieve a better prediction accuracy by allowing for variations in the classification rule.

In practice, we crop the variable images into 2000 × 2000 pixel sub-regions after the data preparation phase (Step 1). Accordingly, the size of a sub-region is approximately 3.4 million km2 at the equator, which is close to the size of India. The number of rows in a sub-region is 4 million, and we take a subsample of 10% for the model training. By running different versions, we found this window size to have a good balance between improvements in prediction accuracy and being widely independent of the quality of the built-up data. However, the changes in the classification rule across the sub-regions causes a problem at their borders. The shifts in predicted probabilities, and consequently the number of lit pixels, are visibly clear.

The solution for smoothing the border effects is to move the sub-region window in smaller steps and then average the results. The step size should be as small as possible, but we rapidly run into computational constraints. Therefore, we decide to move the window in steps of 250 pixels so that we receive 64 values for each pixel. Unlike in the global approach, we now set the number of lit pixels

within each window by using a 4% tolerance level. This means that we keep classifying light in pixels as human-generated in the order of highest probability until 4% of avg_light pixels with values below 5 are included. By setting the amount of light with this tolerance threshold, we avoid having to make subjective choices of light amounts across regions or base them on the built-up data. However, we have to include a backstop or we end up adding lit pixels also in areas where there is no human-generated light at all. Therefore, we additionally set all pixels with a predicted probability less than 20% of being human-generated to zero.
As described above, we classify light in each pixel 64 times by the overlapping sub-region windows as human-generated or not. After Step 3 in the filtering process, we mosaic all the windows and count the classifications. We then create the final product in Step 4 by taking the light value from the average visible band image if a pixel is classified as human-generated more than eight times out of 64. Otherwise, we set the light value in a pixel to zero. The requirement of eight human-generated light classifications makes a final global adjustment to the amount of light. We have chosen the light amount settings in a way that balances the improvement in detecting human-generated light while keeping misclassifications at a low level. Howeve

Although our methodology largely follows that presented by Määttä & Lessmann (2019), there are some distinct differences.

- Size of area
- Jitter size
- Probability of human-generated
- only f152001, we do f142001 and f182011 as well

3.3. Random Forest Algorithm

RF

4. Results

- maps comparing stable lights with noise removed vs human-generated light (overlay map)
- also amount/percentage of pixels saved
- accuracy measures?

5. Discussion

This paper extended the usage of a filtering methodology proposed by Määttä & Lessmann (2019) with which to separate background noise from human-generated nightlight data.

Our results echo those by Määttä & Lessmann (2019): the RF algorithm introduces great improvements in classification accuracy and thus greater accuracy in filtering out background noise from the ‘Stable Lights’ product. This allows the researcher to do away with the need for quick-and-easy type fixes to noisy data at the lower end of the luminosity distribution.

However, it is important to note what this method does not achieve. For instance, the ‘Human Lights’ product does not address the issue of blooming or oversaturation at the high end of the luminosity spectrum. Likewise, its spatial resolution remains low in comparison to more modern products such as the Visible Infrared Imaging Radiometer Suite (VIIRS), and it is recommended to use these products rather than the DMSP-OLS ‘Stable Lights’ or ‘Human Lights’ products if possible.

References

10 Määttä, I. & Lessmann, C. 2019. Human lights. *Remote Sensing.* 11(19):2194.

Appendix

- images of all the different local images - refer them back to variables table

plot of initial stable lights product

plot of stable lights product with noise filtered out at lower end

plot of noise-filtered result

overlay plot of the above two with different colours