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 these alternative avenues are remote sensing data, and more specifically, satellite data. One of the most prominent of these datasets, the Stable Lights product, derived from the Defense Meteorological Satellite Program's (DMSPs) Operational Linescan System (OLS), has been especially popular. However, this product, which captures nightlight luminosity (NL) suffers from some potentially severe accuracy-related challenges, such as the blooming effect and background noise. The following paper explicates on some of these challenges, 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 (RF) algorithm that categorises night-light data as being human-generated. This methodology proves fruitful in addressing some of the challenges related to NL data, and therefore also in preserving data at the lower end of the light spectrum.

Keywords: Remote Sensing, Night Lights, Random Forest

1. Introduction

In recent years, the expansion of the usage of remote sensing data presents potential for new and diverse applications in socioeconomic research. These data sources are often deemed largely objective and free from standard sample selection challenges, and is thereby suitable as a proxy in a broad array of studies that require the usage of potentially unreliable or unavailable data. This advantage is especially pertinent in parts of the developing world, where good data can be relatively hard to come by. One of the most prominent remote sensing data avenues has been the satellite-derived nightlight luminosity data sets, with luminosity often being used as a proxy for important measures such as population counts[1], economic activity and growth [2] and regional inequality indices.[3]

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

[1] See, for instance, Mellander, Lobo, Stolarick & Matheson (2015)

[2] Elvidge, Baugh, Kihn, Kroehl, Davis & Davis (1997); Henderson, Storeygard & Weil (2012) and Chen & Nordhaus (2011) are prominent examples

[3] Such as in Ivan, Holobâcă, Benedek & Török (2020) and Mveyange (2015)

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 and difficult-to-overcome challenge: 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.

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 other sources. 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. Whilst drawing on Määttä & Lessmann (2019), there are distinct differences in approach. Most prominently, Määttä & Lessmann (2019) apply their filter on the entire world, whilst including some inputs that are region specific. The following paper encapsulates data only from South Africa, thereby foregoing the necessity for region-specific adjustments. This, and other deviations, will be explicated on in more detail in section 3, however. The rest of the paper is laid out as follows: section 2 briefly discusses some of the primary challenges with the DMSP-OLS data, whilst section 3 discusses the noise-filtering methodology and the data inputs needed for the RF algorithm. Section 4 presents some results and visualizations, and section 5 concludes.

plot of initial stable lights product:

Figure 1.1: The Raw Stable Lights Image

2. Night Light data and Noise

2.1. Stable Lights and Economic Activity

The Stable Lights Product is a

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.

Common solution to the problem of noise is to either filter out data below a certain threshold, or set those values to 0.1 (Source NB NB NB)

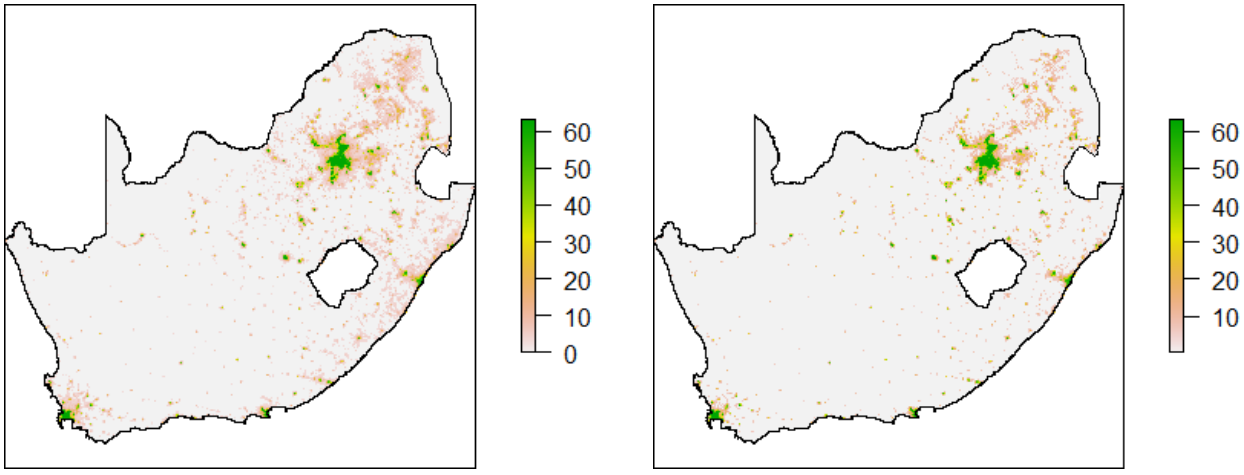

Figure 2.1: Left - Raw Stable Lights; Right - Noise Discarded

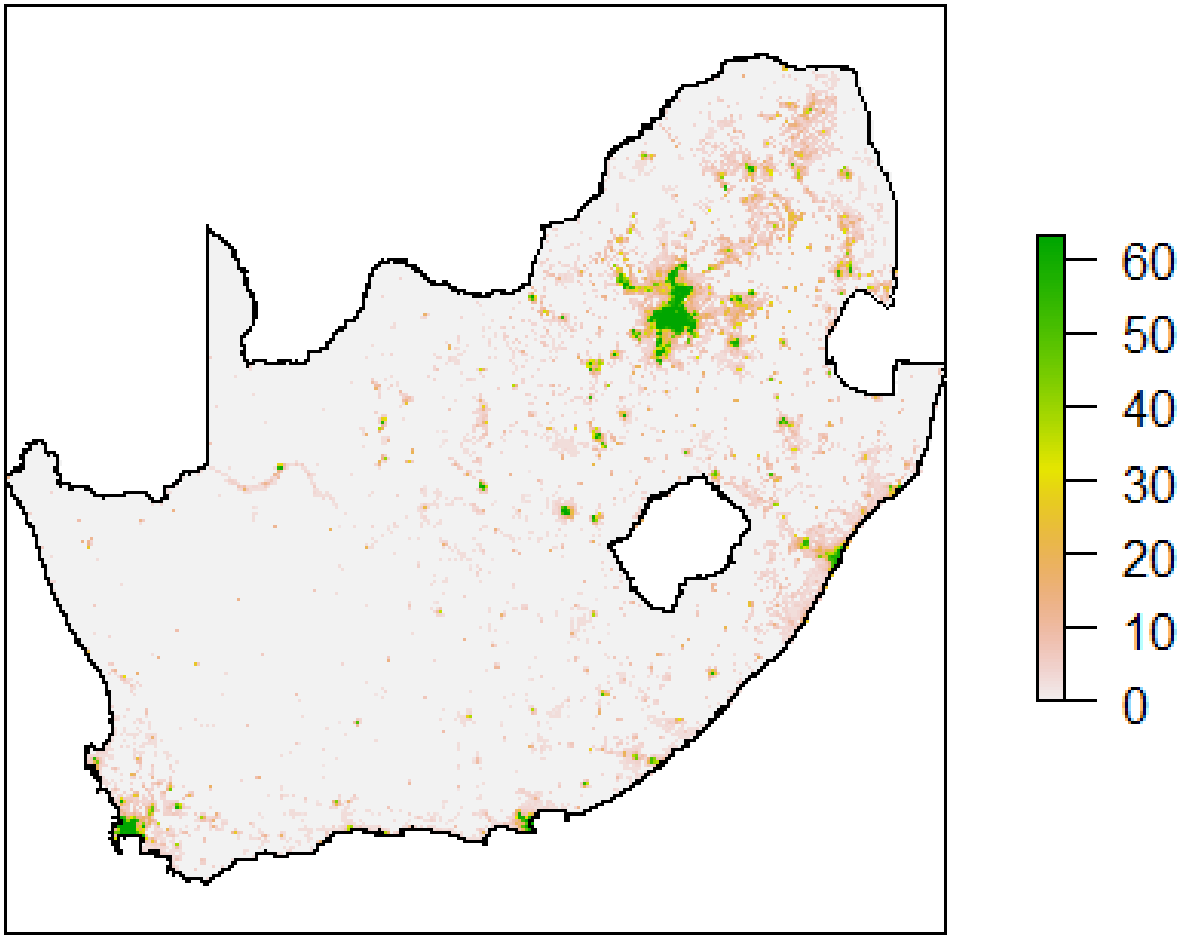

Figure 2.2: Human Lights Image

3. Method and Data

- Short overview of Määttä & Lessmann (2019)

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

3.1. Data

Table 3.1 below gives an overview of the most important variables used as inputs in the machine learning algorithm. The Global Human Settlement Layer (GHSL) *built-up* grid - satellite-derived images of the world based on Landsat satellites - is used to classify whether a certain area is more likely

to contain human generated light.[4]. The second and third variables constitute the two primary DMSP-OLS products released by NOAA, and is also the most crucial inputs in the filtering methodology. This is due to the fact that the subsequent variables are derived from them. The *local noise* variable is used to identify those areas where background noise is systematically larger than some threshold, and consists out of the number of pixels in a window of 499 by 499 pixels below this threshold.[5] The *local detections* counts the number of pixels where zero light is detected in a 399 by 399 window around a pixel. The rest of the variables are generated to be regional input variables that more closely describe the characteristics of luminosity surrounding a specific pixel. By varying the size of the area around a pixel, one more accurately accounts for spatial correlations in light that is human generated.

Input	Source	Description
built up	GHSL	Landsat satellite images of built-up areas
avg light	DMSP-OLS	average light per pixel over a year
freq light	DMSP-OLS	proportion of days where light is detected per pixel
local noise	Derived from DMSP-OLS	number of pixels below threshold in avg light image
local detections	Derived from DMSP-OLS	accounts for regional differences in freq light
lm freq 5	Derived from DMSP-OLS	average of freq light in a 5 by 5 pixel area
lm avg 25	Derived from DMSP-OLS	average of avg light in a 25 by 25 pixel area
lm freq 99	Derived from DMSP-OLS	average of freq light in a 99 by 99 pixel area
lm avg 199	Derived from DMSP-OLS	average of avg light in a 199 by 199 pixel area

Table 3.1: Data Inputs

3.2. Methodology

The variables above constitute the primary inputs for Määttä & Lessmann (2019)'s 'Human Lights' product. Methodologically, the following five steps are necessitated:

First, the regional night light variables are created from the original *avg light* and *freq light* DMSP-OLS products.

Second, the world map is divided in various sub-regions, in preparation that the random forest algorithm be run on each sub-region separately. All the variable images are cropped into 2000-pixel square sub-regions, which equates to about 3.4 million km^2. Each sub-region consists of approximately 4 million rows, whilst the model is trained on a 10% subsample.

[4]For more on the GHSL methodology, consult Pesaresi, Ehrlich, Ferri, Florczyk, Freire, Halkia, Julea, Kemper, Soille, Syrris & others (2016)

[5]We follow Määttä & Lessmann (2019) in choosing the upper bound threshold value of 6 so as to be most strict in what is considered noise

5. Local Human Lights Process
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

Many of the tasks needed for successful implementation of the filtering procedure are intensive both in terms of computation and memory usage. Due to computational limitations, we therefore employed AWS' EC2 cloud computing platform for parallel computation. This proved costly, however, and implementation thereby required some deviations from the methodology presented in Määttä & Lessmann (2019). Most crucially, the random forest algorithm was only applied to a subset of the world data. As our research interests entailed a specific country, South Africa, we limited the regional sub-windows to a one-square window, roughly the size of South Africa. Although this step may carry some costs in that it means that less regional variability is accounted for, these costs are mitigated by the possibility that there would not be large regional variability within a single country in any case.

Other Differences:

- Size of area

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

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 Chen, X. & Nordhaus, W.D. 2011. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*. 108(21):8589–8594.

Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W., Davis, E.R. & Davis, C.W. 1997. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*. 18(6):1373–1379.

Henderson, J.V., Storeygard, A. & Weil, D.N. 2012. Measuring economic growth from outer space. *American economic review*. 102(2):994–1028.

Ivan, K., Holobâcă, I.-H., Benedek, J. & Török, I. 2020. Potential of night-time lights to measure regional inequality. *Remote Sensing*. 12(1):33.

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

Mellander, C., Lobo, J., Stolarick, K. & Matheson, Z. 2015. Night-time light data: A good proxy measure for economic activity? *PloS one*. 10(10):e0139779.

Mveyange, A. 2015. *Night lights and regional income inequality in africa*. WIDER Working Paper.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., et al. 2016. Operating procedure for the production of the global human settlement layer from landsat data of the epochs 1975, 1990, 2000, and 2014. *Publications Office of the European Union*. 1–62.

Appendix

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

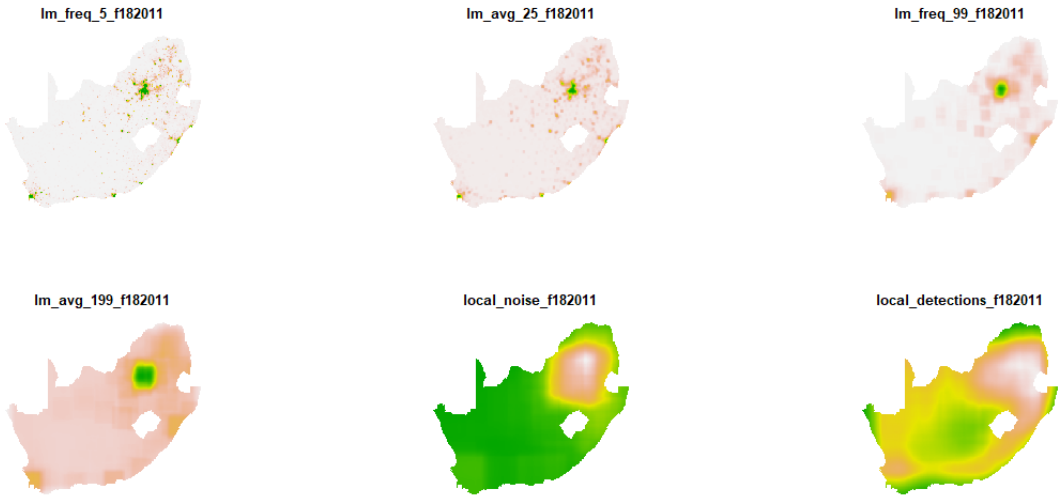

Figure 5.1: Local Image Inputs

overlay plot of the above two with different colours