Madison Bradley

MSBR 70260

4 December 2024

Silent Night: Satellite-Retrieved Light at Night as an Indicator of Under-Five Mortality

Introduction

Despite major advances in the first decade of the 21st century, under-five mortality rates in sub-Saharan Africa remain the highest globally[1]. Of the 4.9 million under-five deaths in 2022, 57% occurred in sub-Saharan Africa[2,3].

The under-five mortality rate is widely recognized as an indicator of a country's broader development status[4]. Its causes and covariates are well-documented[5, 6, 7]. In most previous studies, the spatial resolution of these covariates tends to be either (i) coarse, such as Gross Domestic Product, which is normally estimated at the national scale, or (ii) sparse, derived from nationally-representative household surveys only available for select years[8]. The absence of a comprehensive georeferenced indicator prevents uniform analysis across space and may oversimplify important socioeconomic or spatial dynamics[8].

Satellite-retrieved light at night (LAN) data provides measures of the brightness of human-generated light as seen from space[9]. With a spatial resolution of 1km x 1km, LAN data offers significantly greater geospatial specificity and coverage than traditional indicators. These advantages make LAN data an increasingly popular proxy for development analyses[8], including estimations of poverty[10], subnational GDP[11], social indicators[12, 13], electrification[14], population density[15], urbanization[16, 17, 18], and others.

Objective

The objective of this exercise is to determine the association between LAN and the risk of under-five mortality rate using child survival time as the outcome. Given that previous research found a correlation between LAN and child survival using a Cox proportional hazards model[8], I attempt to use an alternative method, XGBoost, to validate those results, establish methodological rigor, and understand method sensitivity.

This exercise does not attempt to replicate the *precise values* of the relationship between LAN and under-five mortality as reported in previous studies; to do so would be computationally prohibitive and would require making assumptions about baseline hazard that is beyond the scope of this assignment. Rather, the objective is to validate the relative strength and direction of the relationship.

Data Description

Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS)

LAN data is provided by the Linear Scanning Operational System (OLS) on the United States Defense Meteorological Satellite Program (DMSP) nightlight satellite. The raw LAN data was processed by the national Geophysical Data Center (NGDC) of the United States Oceanic and Atmospheric Administration (NOAA)[8]. Processed data contains annual brightness of LAN from 1992 to 2013 at a spatial resolution of 1 km × 1 km and global coverage, available at https://eogdata.mines.edu/products/ [8]. Brightness is recorded in Digital Number (DN) with values ranging from 0 to 63, with higher values indicating brighter

areas at night. The DMSP-OLS product was discontinued in 2014. This exercise uses the two latest years available; 2012 and 2013.

LAN data is available in tagged image file (TIF) format, a raster image format commonly used in geographic information systems to support large, layered, high-resolution images. DMSP-OLS offers both raw and preprocessed datasets. For this exercise, the preprocessed (intercalibrated stable lights average DN values) are used, which have been adjusted to remove ephemeral events (i.e. fires) and replace background noise with values of zero.

Demographic and Health Survey (DHS)

Data on child mortality is derived from the Demographic and Health Surveys (DHS). The DHS database provides nationally-representative household surveys used to monitor population, health, and development. It is open access with approval at https://www.dhsprogram.com/.

To avoid excessive computation time and cleaning efforts, I use data from only two sub-Saharan African countries: Rwanda and Kenya. I use the two surveys nearest two, but after, 2013 for each country. Raw data is subset following the inclusive criteria: (i) surveys include information on mothers, child birth records, and GPS data of their survey cluster, and other covariates; and (ii) child birth year was during 2012-2013, aligned with available LAN data.

Preprocessing

Raw DHS surveys are provided in .dta format and include roughly 30,000 to 80,000 rows of over 1,000 variables. The raw surveys were subset by rows to include only records for children born in 2012 and 2013, and by columns to include only relevant identifiers and known covariates.

Raw LAN files are provided in raster format. To avoid memory limitations, the raster files are subset to only the countries of interest using the countries boundaries shapefile available from Natural Earth at https://www.naturalearthdata.com/downloads/10m-cultural-vectors/. Once subset, these smaller raster files are converted to data frames with `x`, `y`, and `values` columns using the rasterToPoints() function.

Child records in the DHS subsets are then matched with LAN values in time by year of birth and in space corresponding to the grid cell closest to the latitude and longitude of the center point of their survey cluster. Precise residential address coordinates are not reported by DHS for privacy reasons.

Additional preprocessing efforts included the conversion of data types (primarily from numeric to factor), deriving an `age_at_death` in days column from an `age_at_death_coded` column, deriving a binary `status` column to indicate death, and the removal of columns used for LAN matching but not required for modeling. The resulting data frame (`model_data`) contains 21,532 child records with 14 features.

To use XGBoost for survival analysis, the `model_data` data frame was split to have all feature columns in one data frame (`features`) and all survival columns in a survival object (`survival`) with `age_at_death` as the time variable and `status` as the event variable. The `features` data frame was then converted to a matrix (`X`). Training (70%) and test (30%) datasets were built for `X` and both the time and event portions of `survival`. These were then converted to DMatrix format to create `dtrain` and `dtest`.

Methods

Model Selection

I selected an XGBoost model for its ability to model non-linear relationships and handle high-dimensional data. Given under-five's mortality's complexity and variety of covariates, I assumed its relationship with

those variables to be non-linear due to contextual factors, threshold effects, diminishing returns, or interaction effects. While 14 features is not exceptionally high-dimensional, it is enough to consider the use of a tree-based method to capture relationships linear regression may miss.

The XGBoost objective function can be adapted to survival analysis by modifying the objective function to handle censored data (the event (i.e. child death) has not occurred at the time of the study) with the Cox proportional hazards (Coxph) model. The Coxph model assumes that the risk of the event occurring at any given time is constant for each individual (a baseline), and the effect of the features is multiplicative on that hazard[19].

Strengths and Limitations

While the Coxph model assumes a linear relationship between features and hazard rate, XGBoost is equipped to handle non-linear data. Their combination would, in theory, provide "the best of both worlds". XGBoost is resilient to overfitting, which is valuable, but is also complex compared to alternative methods and sensitive to outliers (which are not of priority for this analysis).

Results

Parameters

The XGBoost model was adapted to use Coxph as the objective function with the evaluation metric cox-nloglik. The Cox negative log-likelihood is a measure of model accuracy against test observed data; a lower value indicates a better-fitting model[20]. Negative log-likelihood is minimized during model training. The booster `gbtree` is common for survival analysis[21].

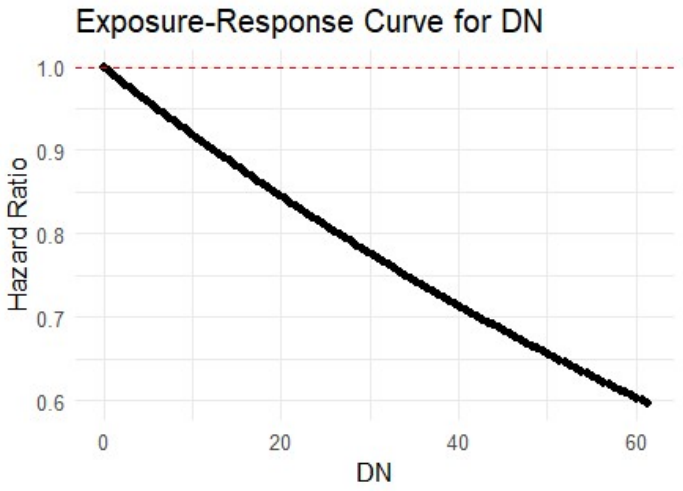
Output

The XGBoost model's predictions were risk; a measure of neither probability of an event occurring (values exceed 1) nor time-to-event (values do not exceed two, as would be expected for predictions in days). Given the difficulty of interpretation, this model was determined to be suboptimal. The objective, consistent with the reference study conducted by reference study by Li, Bachwenkizi, Chen, et al., is to produce an exposure-response curve between the hazard ratio of under-five mortality and LAN. This was not possible with XGBoost.

Model Revision

Following review of XGBoost performance, we revert to a simpler, proven method: the Cox proportional hazards regression from the 'survival' package. Using the model's calculated coefficients for the LAN variable, I produce a hazard ratio value that can be plotted against LAN values. In a Coxph model, the hazard ratio indicates the relative change in the hazard of an event occurring for a one unit increase in a predictor variable. Values of greater than one represent an increased risk of the event occurring, while values of less than one indicate decreased risk.

Figure 1.



This figure suggests that (i) higher light values (brighter lights) are associated with lower risk of under-five mortality, and (ii) that as light values continue to increase, the reduction in risk of under-five mortality becomes smaller. In other words, the benefit of increases in light diminishes at higher values. This is consistent with the reference study's findings in direction, though my model indicates a notably more linear relationship. While the Cox model determined the light variable to not be significant at the 1%, 5%, or even 10% level, its p-value of 0.2 remains worthy of consideration.

Discussion

This exercise established proof of concept for the use of LAN data as a predictor of under-five mortality. However, significantly more analysis would be necessary prior to taking action on these results. The Cox proportional hazards model makes a major assumption in generating variable coefficients by effectively assuming all else constant.

Predicting under-five mortality is a critical public health initiative to appropriately direct funding and resources to at-risk families. Allocation decisions should be made on robust and replicable analysis.

Conclusion

To perform a more robust validation of the results presented by Li, Bachwenkizi, Chen, et al., I would need to expand the dataset to include several additional sub-Saharan African countries and several additional birth years. I may also consider binning categorical covariates such as `water_source`, `toilet_type`, and others into 'undeveloped' and 'developed' to reduce the number of dummy variables and potentially improve variable significance.

Additional stratification analyses would further enrich the analysis by determining if the association between LAN and risk of child mortality varies by subpopulation. Of particular interest is the subpopulation of neonatal deaths, referring to deaths that occur in the first 28 days of life. Sub-Saharan Africa's infant mortality rate stands at 72 per 1000 live births, significantly above the reduction target of fewer than 12 per 1000 established by the United Nations Sustainable Development Goals[22]. Neonatal mortality accounts for 47% of deaths of children under five globally[2].

Citations

[1] World Health Organization (WHO). (2022, December 1). *Africa's advances on maternal, infant mortality face setbacks: WHO report*. WHO Regional Office for Africa. https://www.afro.who.int/news/africas-advances-maternal-infant-mortality-face-setbacks-who-report

[2] World Health Organization (WHO). (2024, March 14). *Newborn mortality*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/newborn-mortality

[3] UNICEF. (2024, March 12). *Levels and trends in child mortality*. UNICEF. https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2024/

[4] World Health Organization (WHO). (n.d.). *Under-five mortality rate (per 1000 live births)*. World Health Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3139#:~:text=Under%2Dfive%20mortality%20rate%20measures,live%2C%20including%20their%20health%20care

[5] Ester, P. V., Torres, A., Freire, J. M., Hernández, V., & Gil, Á. (2011). Factors associated to infant mortality in Sub-Saharan Africa. *Journal of public health in Africa*, *2*(2), e27. https://doi.org/10.4081/jphia.2011.e27

[6] Schell, C. O., Reilly, M., Rosling, H., Peterson, S., & Ekström, A. M. (2007). Socioeconomic determinants of infant mortality: a worldwide study of 152 low-, middle-, and high-income countries. *Scandinavian journal of public health*, *35*(3), 288–297. https://doi.org/10.1080/14034940600979171

[7] Yamamoto, M., & Oshima, K. (2021). *Trends and determinants of under-five mortality in low- and middle-income countries: A systematic review*. *Global Health Action*, 14(1), 1899604. https://doi.org/10.1186/s13690-021-00727-9

[8] Li, X., Bachwenkizi, J., Chen, R., Kan, H., & Meng, X. (2023). Association between light at night and the risk of child death in Sub-Saharan Africa: A cross-sectional analysis based on DHS data. *BMC Public Health, 23*, 2366. https://doi.org/10.1186/s12889-023-17284-1

[9] World Bank. (n.d.). *Introduction to nighttime light data*. World Bank. https://worldbank.github.io/OpenNightLights/tutorials/mod1_2_introduction_to_nighttime_light_data.html

[10] Kaufman, L. M., & Sakamoto, A. (2018). The promise of using satellite data to address the challenges of global child health. *Science, 359*(6376), 999-1001. https://doi.org/10.1126/science.aaf7894

[11] Suleiman, H. (2024). Illuminating the Nile: Estimating subnational GDP in Egypt using nighttime lights and machine learning. *GeoJournal, 89*, 117. https://doi.org/10.1007/s10708-024-11106-6

[12] Andries, A., Morse, S., Murphy, R. J., Sadhukhan, J., Martinez-Hernandez, E., Amezcua-Allieri, M. A., & Aburto, J. (2023). Potential of Using Night-Time Light to Proxy Social Indicators for Sustainable Development. *Remote Sensing*, *15*(5), 1209. https://doi.org/10.3390/rs15051209

[13] Bruederle, A., & Hodler, R. (2018). Nighttime lights as a proxy for human development at the local level. *PLOS ONE, 13*(9), e0202231. https://doi.org/10.1371/journal.pone.0202231

[14] Dugoua, E., Kennedy, R., & Urpelainen, J. (2018). Satellite data for the social sciences: measuring rural electrification with night-time lights. *International Journal of Remote Sensing*, *39*(9), 2690–2701. https://doi.org/10.1080/01431161.2017.1420936

[15] Liu, Q., Sutton, P., Elvidge, C. Relationships between Nighttime Imagery and Population Density for Hong Kong. *Proceedings of the Asia-Pacific Advanced Network, 31,* 79-90. http://dx.doi.org/10.7125/APAN.31.9

[16] Ma, T., Liu, Y., & Li, X. (2015). Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP/OLS satellite data. *Remote Sensing of Environment, 158*, 453–464. https://doi.org/10.1016/j.rse.2015.01.022

[17] Small, C., Pozzi, F., & Elvidge, C. (2005). Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sensing of Environment, 96*(3), 277–291. https://doi.org/10.1016/j.rse.2005.02.002

[18] Pandey, B., Joshi, P., Seto, K. (2013). Monitoring urbanization dynamics in India using DMSP/OLS night time lights and SPOT-VGT data. *International Journal of Applied Earth Observation and Geoinformation, 23,* 49-61. https://doi.org/10.1016/j.jag.2012.11.005

[19] Halabi, Susan. (2008). Proportional hazards model. *Urologic Oncology: Seminars and Original Investigations.* Translational Surgery. https://www.sciencedirect.com/topics/medicine-and-dentistry/proportional-hazards-model#:~:text=The%20Cox%20proportional%20hazards%20model%20is%20a%20frequently%20used%20approach,no%20distributional%20assumptions%20are%20required

[20] Segota, I. (2023, June 6). *Unbox the cox: Intuitive guide to cox regressions*. Medium. https://towardsdatascience.com/unbox-the-cox-intuitive-guide-to-cox-regressions-c485408ae15d

[21] Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific reports*, *11*(1), 6968. https://doi.org/10.1038/s41598-021-86327-7

[22] World Health Organization. (2024, November 19). *Africa's advances in maternal, infant mortality face setbacks – WHO report*. WHO. https://www.afro.who.int/news/africas-advances-maternal-infant-mortality-face-setbacks-who-report