Applied Energy 319 (2022) 119237

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

The electricity scene from above: Exploring power grid inconsistencies using satellite data in Accra, Ghana

Zeal Shah [a,*], Noah Klugman [b], Gabriel Cadamuro [c], Feng-Chi Hsu [d], Christopher D. Elvidge [d], Jay Taneja [a]

[a] *STIMA Lab, Electrical and Computer Engineering Department, University of Massachusetts, Amherst, 01003, MA, USA*
[b] *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 94720, CA, USA*
[c] *Atlas AI PBC, 855 El Camino Real, Ste. 13A 387, Palo Alto, 94301, CA, USA*
[d] *Payne Institute, Colorado School of Mines, 816 15th St., Golden, 80401, CO, USA*

ARTICLE INFO

Keywords:
Electric grid reliability
Remote sensing
Nighttime lights
Smart meters
Distribution grid monitoring
Emerging economies

ABSTRACT

Complicated systems are complicated to monitor. The electric grid is one of the most complicated systems, and subsequently goes under-monitored in many regions around the world that cannot easily afford expensive meters. However, the electric grid is also critical for sustaining a high quality of life, and requires better monitoring than is often available to ensure consistent service is provided. Past work has shown that images taken by satellites during the night, capturing nighttime illumination (“nightlights”), could provide a proxy measurement of grid performance for minimal cost. We build upon earlier work by identifying the pixel z-score – a statistical measurement of a pixel’s illumination relative to its history – as a key method for detecting electricity outages from the often-noisy nightlights dataset. We then train and validate our approach against observations from a network of on-the-ground power outage sensors in our observation area of Accra, Ghana, a dataset representing the largest collection of utility-independent electricity reliability measurements on the African continent. Using multiple machine learning techniques for estimating potential outages from nightlight images, we obtain high performance for predicting outages in Accra at scales as small as a single pixel (0.2 km^2) and with training datasets as small as three months of illumination/sensor data. We further validate our methodology beyond the spatio-temporal coverage of the on-the-ground sensor deployment against a human-labeled dataset of outages by neighborhood throughout Accra. Delving deeper into the applications and limitations of available datasets and our work, we conclude by highlighting questions about the generality of our method vital to understanding its potential for low-cost worldwide measurements of grid reliability.

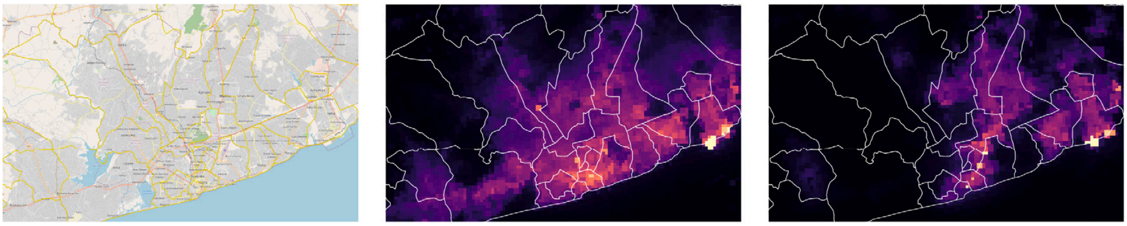
1. Introduction

Access to electricity, widespread but not universal, is a vital input to human and societal development the world over. However, unreliable electricity is the norm for billions of people across many low- and middle-income countries, stifling growth and livelihoods. Frustratingly, a failure to measure where and to what extent electricity grids are unreliable hinders the progress needed for communities to flourish.

While it is typical for electricity system operators in high-income countries to measure grid reliability with a mix of in-network and endpoint sensors, a dearth of regulations to publicly report data result in few if any widely available grid reliability statistics. In low- and middle-income countries, the situation is further exacerbated by utilities that only have automated data collection systems at the high-voltage transmission tier of their electricity grids, with little to no consistent measurement of electricity distribution outages; this results in utilities typically vastly under-counting outages [1,2]. In these settings, outage data usually only arise from phone calls and social media reports from exasperated consumers, methods that suffer from consumer apathy, potential manipulation by outage response personnel, and a persistent inability to associate outage reports with the affected grid components [2]. While many settings encounter outages caused by electricity supply shortages, a high rate of outages persists even in settings with excess electricity supplies, as a result of brittle grid infrastructure [3]. Further, backup systems (e.g., generators and inverters) are neither deployed in enough volume nor are they networked in order to provide public outage data.

[*] Corresponding author.
E-mail addresses: zshah@umass.edu (Z. Shah), nklugman@berkeley.edu (N. Klugman), gabe@atlasai.us (G. Cadamuro), fengchihsu@mines.edu (F.-C. Hsu), celvidge@mines.edu (C.D. Elvidge), jtaneja@umass.edu (J. Taneja).

https://doi.org/10.1016/j.apenergy.2022.119237
Received 15 November 2021; Received in revised form 7 April 2022; Accepted 30 April 2022
Available online 17 May 2022

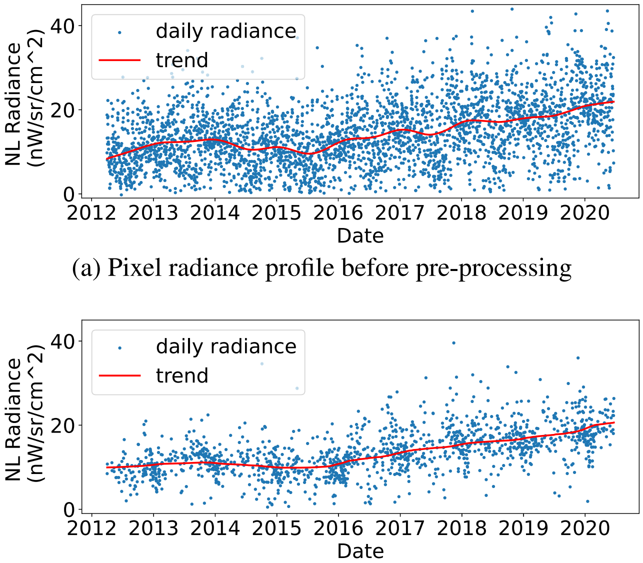


(a) Open street map of Accra

(b) VIIRS nightlight map of Accra May 21, 2020 (no outage)

(c) VIIRS nightlight map of Accra May 22, 2020 (outage)

Fig. 1. The OpenStreetMap [4] view (a) and overlapping VIIRS nightlight [5] view (b) with delineated city borders show the spatial extent of the nightlight sample in Accra. VIIRS nightlight view (b) was captured on an outage-free night (May 21, 2020). VIIRS nightlight view (c) was captured on May 22, 2020, when many parts of Accra experienced a power outage, especially Western Accra [6], which caused a significant drop in the radiance of affected pixels relative to the previous night, i.e., view (b).



(a) Pixel radiance profile before pre-processing

(b) Pixel radiance profile after pre-processing

Fig. 2. Typical daily radiance profile and signal trend for a single pixel taken across the entire nightlight dataset are shown before pre-processing (top) and after pre-processing (bottom). Unit of radiance is *nano-watts per steradian per square centimeter.*

For utilities in low- and middle-income countries to address this problem with sufficient automated in-network and endpoint sensing would typically cost in the billions of USD per country – often too large an investment to sustain in the face of other development priorities. As a low-cost alternative, satellite-based grid reliability measurement – typically derived from nighttime illumination measurements – has long presented the tantalizing prospect of incomparable coverage over space and time, measured independently of electricity utilities.

In this work, we develop a technique for predicting when an area is experiencing a power outage using satellite nighttime lights (NL) data. We validate our technique with what we believe is the single largest ground-measured reliability dataset from a developing region presented in the literature — collected by 721 sensors deployed in Accra, the capital of Ghana, over a period of 27 months. We identify the pixel z-score – a measure of how much a given day's radiance deviated from the pixel's mean radiance level – as a key statistical metric for identifying outages from daily NL data, which are noisy (see Fig. 2). We provide evidence that the pixel z-score is an excellent proxy for outages detected by the ground-based sensors. We use this insight to evaluate for the first time the true fidelity and limits of NL data – showing what scale of outages are and are not detectable with this dataset – and to characterize the volume of temporal and spatial data needed for our approach to be accurate. To sanity-check our approach, we test our developed technique against a human-labeled dataset of outages by neighborhood throughout the city — providing further validation of our method during an out-of-sample period in out-of-sample locations.

Finally, we explain the next steps needed to translate our model into one that is usable by policymakers and researchers more broadly. We caution that there is likely a ceiling to performance at least until more ground truth data are available and spatial and temporal resolution of nightlight data is improved. We also address concerns about generalizing our method and propose experiments needed to evaluate this property.

While our technique cannot provide the type of granular outage data that is needed for enhancing utility system operations, it does offer measurements that are aggregated, historical, and (partially) transferable, enabling *post hoc* analysis of electricity grid supply performance and comparisons among or within cities or regions. This work demonstrates a feasible first step towards a scalable, non-intrusive, and low-cost grid monitoring solution for regions with poorly-monitored grids. We believe that our measurements of electricity supply inconsistency will be of great interest to both regulators and operators as well as policymakers and researchers in the absence of widespread grid reliability sensing. The former will benefit from the ability to assess, monitor, and respond to electricity system performance while the latter will be able to study the historical record of grid reliability to understand its importance and relation to other socioeconomic outcomes of interest.

2. Related work

The Defense Meteorological Satellite Program Operational Line Scan (DMSP-OLS) and the Suomi National Polar Orbiting Partnership (Suomi NPP) Visible Infrared Imaging Radiometer Suite (VIIRS) are the two most prominent sources of NL data with NPP-VIIRS being the latest generation instrument. Processed VIIRS NL data in the form of yearly, monthly and daily composites are publicly available [7–9]. Our work explores the potential of daily VIIRS NL data to estimate power supply inconsistencies using PowerWatch sensors [2] in the city of Accra, Ghana. PowerWatch has produced the largest and highest resolution utility-independent power quality dataset in Sub-Saharan Africa.

Satellite data have long been proposed as a solution for grid reliability sensing [10–17], though often with little to no ground-truth validation.

Closest to our work is a paper by Mann et al. [17], which combines daily VIIRS data and reliability data recorded by voltage meters deployed across 39 different locations on the ground [18] to train a random forest model to detect outages in the state of Maharashtra, India, for a period of 9 months in 2015. While our goals are broadly similar, our work differs in a few key ways: (1) the sensor deployment we make use of has significantly better coverage in both the temporal (9 months vs. 27 months) and spatial (39 pixels coverage vs. 210 pixels) dimensions than this previous work. This allows us to leverage nearly 10 times more NL observations when training and evaluating our models for detecting outages using NL and lets us better explore the relationships between spatial and temporal resolution and predictive power. (2) Mann et al. [17] showed a 62% error rate for the task of correctly identifying individual outages using NL data, while our technique demonstrates 3x improvement on the same task. (3) Additionally, our work evaluates the meta-system which is left out of Mann et al. [17], and in fact a key contribution of this work is our evaluation of the base statistical measures for detecting outages of different types and sizes.

The rest of the literature differs from our study in so far as it relies on monthly or annual VIIRS composites (lower temporal resolution) or on village/city-level reliability estimates (lower spatial resolution) as ground truth for supply inconsistency measurements. Studies have shown that nighttime satellite observations can detect wide-area grid outages that span over multiple days [12–14]. Most studies using NL data have relied on monthly or annual NL composites as ground truth for reliability or supply-inconsistency measurements. Min et al. [16] proposed an annual power supply irregularity index to compare power supply inconsistency across 600,000 villages in India using the standard deviation of radiance levels in DMSP-OLS NL data spanning from 1993 to 2013. Elvidge et al. [15] used monthly VIIRS NL composites to explore power disruption patterns in India's Gangetic Plains. Dugoua et al. in their working paper [11] developed a tool that uses monthly VIIRS NL composites to estimate electricity reliability at the village level in Uttar Pradesh state, India. Authors of [19] proposed that annual cycling patterns in the NL data for different states of India can be indicative of power supply inconsistencies in those states. More recently Elvidge et al. [10] proposed multiple indices to estimate city-level power supply irregularities using daily VIIRS NL data, however, the presence of direct measurements of reliability lets us sidestep the need for these types of heuristics.

Furthermore, researchers have used NL data for a variety of applications in different fields, including non-intrusively analyzing the impacts of war [14,20], identifying damage caused by natural disasters [21,22], monitoring restoration and recovery efforts following a natural or man-made disaster [14,20,22], estimating socio-economic parameters like poverty [23], studying the impacts of political institutions on distribution of electricity in developing regions [24], tracking electrification [25,26], characterizing energy consumption patterns [26,27], and tracking human activity [28], most recently during the COVID-19 pandemic [29,30].

3. Data and pre-processing

This paper primarily employs two large datasets: (1) a nighttime lights dataset collected by satellite and (2) a grid reliability dataset collected by a deployment of plug sensors.

3.1. Nighttime lights data

Since 1992, a constellation of U.S. government weather satellites has recorded daily measurements that capture the illumination of the Earth each night. In particular, since 2012, NPP-VIIRS records daily nighttime lights observations at a resolution of 750 m. [5]. These are then processed and converted into a grid with a spatial resolution of 15 arc-seconds (approximately 450 m at the equator) by the Earth Observation Group (EOG) at the Payne Institute for Public Policy, Colorado School of Mines [9].

3.1.1. Sample size

The NL grid for our study area – Accra, Ghana, in West Africa – is 78 × 128 pixels with a pixel resolution of 15-arc seconds. The upper left corner of our study-area grid is 5∘49′15″N, 0∘30′15″W. Accra's NL images captured on May 21 and May 22, 2020, are shown in Fig. 1(b) and (c). The NPP-Suomi satellite's daily flyover time for Accra varies during the period from 0:00 am and 3:00 am local time (GMT). Accra's NL data consists of daily temporal profiles of radiance for every pixel from April 2012 to September 2020, as seen in Fig. 2. The unit reported for radiance is nano-watts per steradian per square centimeter ($\mathrm{nW/sr/cm^2}$).

3.1.2. Data pre-processing

In addition to pixel radiance, the daily NL dataset provides pixel-level metadata – stray light flag, cloud flag, sensor zenith angle or sample position, and lunar illuminance – recorded by the instruments hosted on the same satellite. As the first pre-processing step, we preserved the NL readings with stray light flag values indicating "no impact" or "corrected for stray light". The majority of pixels in Accra were not affected by stray light and the remaining ones were corrected from the source, and therefore we did not lose any data points due to stray light.

The second pre-processing step is to filter out cloudy pixels using the cloud flag. The cloud flag for each pixel on a given day indicates one of the four conditions – "confidently clear", "probably clear", "confidently cloudy", and "probably cloudy". The presence of thick clouds on moon-free nights tends to attenuate surface lighting observed by the satellite leading to the dimming of NL radiances [10,31]. Such a radiance attenuation effect is observed in both urban (well-lit) and rural (dim-lit) areas on moon-free nights. But cloud contamination gets more complex on moon-lit nights. Researchers have shown that in areas with dim lighting, moon-lit clouds can be brighter than the region's uncontaminated (cloud-free, moon-free) surface lighting resulting in inflated NL radiances for the region [10,31]. The magnitude of inflation or attenuation of a pixel's observed radiance is also dependent on the thickness of cloud cover. To avoid any misleading results due to such cloudy pixels, we only preserved the readings with cloud flag indicating "confidently clear" and filtered out the rest. This obscured 57.6% of Accra's NL readings during our period of study. Moreover, the VIIRS cloud detection technique is ≈85% accurate which makes cloud misclassification a major source of uncertainties in the VIIRS data suite [31]. Therefore, we chose to limit our exposure to uncertainties by utilizing only the "confidently clear" readings. It is also worth noting that we do not apply any spatio-temporal gap-filling techniques to compensate for lost readings to steer clear of uncertainties related to synthetic gap-filled values.

The third pre-processing step is to correct pixel radiance profiles for the satellite's zenith angle. Depending on the satellite's zenith angle (SZA) and the types of buildings in a pixel, substantial differences in a pixel's radiance can be observed as a function of SZA. Authors of [10,31] observed high variance in a pixel's radiance as a function of SZA and showed that a pixel's radiance at 60° SZA could be 1.6–4 times higher than its radiance at nadir. In order to compensate for the SZA effect, we implemented an SZA-radiance normalization approach proposed in [10]. This approach normalizes the radiance of a pixel recorded at different SZA to match with the pixel's radiance pattern at nadir, resulting in the reduction of variance in radiance profiles.

The final pre-processing step is the lunar correction of cloud-free, scan-angle corrected NL radiance. Lunar irradiance has a significant impact on the radiance of dimly lit regions. For example, previous work [31] has shown that radiance of a rural pixel increased from ≈1 $\mathrm{nW/sr/cm^2}$ on moon-free nights to ≈10 $\mathrm{nW/sr/cm^2}$ on full moon nights. Furthermore, authors in [10,31] observed a high correlation between lunar illuminance and radiance of pixels in dim-lit areas. But unlike rural areas, lunar irradiance has minimal to no effect on the radiance of well-lit regions like urban centers. We corrected for lunar illuminance using the technique proposed in [10] which models the contribution of lunar illumination to a pixel's radiance and then subtracts the lunar component from the observed radiance. It lowered the variance observed in the radiance of rural pixels and reduced the correlation between their radiance and lunar illumination whereas the radiance of urban pixels remained broadly unchanged. Detailed descriptions of lunar and scan-angle correction methodologies are omitted for brevity and we refer the reader to [10] for more details.

At the end of the pre-processing stage, every pixel's radiance profile has been filtered for clouds and corrected for stray light, scan-angle effects from the satellite track, and lunar illumination to reduce the noise and mitigate steep discontinuities in NL data. Fig. 2 shows the daily radiance profile of a pixel from Accra before and after pre-processing. Our key claim is that the residual fluctuations observed in filtered and corrected NL data can be attributed to external factors like power outages (discussed in detail in Section 4.1).

3.1.3. NL-based indices

We identify two specific indices for analysis in this work: z-scores and dispersion index. Both indices consider pixel illumination relative to other readings for the same pixel. Z-score for a given pixel is calculated as $Z_{pt} = (rad_{pt} - \mu_p)/\sigma_p$, where rad_{pt} is the radiance of a pixel p on day t, and μ_p and σ_p are the mean and standard deviation of pixel p over a specific timeline (weeks, months, or years). Z-score facilitates consistent comparison of deviations observed in the radiance of pixels irrespective of their mean brightness levels. A pixel's z-score tells you how many standard deviations away that pixel's radiance is from its mean. A positive or negative z-score for a pixel indicates radiance that was above or below the mean, i.e., that the pixel was brighter or darker. Since Accra's NL values do not grow appreciably over a single year, instead of de-trending the data, we simply compute yearly radiance mean and radiance standard deviation and then compute daily z-scores of pixels on a year-to-year basis.

We did not find the dispersion index to be a very influential metric for studying outage experiences at pixel-level as compared to z-scores. Readers can refer to Appendix B for a detailed evaluation of the dispersion index.

3.1.4. Challenges

NL data is highly noisy. Fig. 2(b) shows the daily radiance profile of a pixel in Accra that never reported any outages, on cloud-free nights. It shows that even a relatively reliable NL pixel exhibits high variance in its daily radiance. A fair amount of work is required to get the signal out of NL data because of its noisy nature.

It is also important to note that data scarcity is a common problem in developing regions. NL data is captured once per night, limiting temporal resolution. At present, there are only two satellites that capture daily nightlight data for Accra, and while these two satellites do not often overlap, when they do we use both readings.

3.2. PowerWatch sample

This work depends on a previously existing deployment of power quality sensors in Accra, called PowerWatch, to provide ground truth measurements of power outages and their duration. These sensors were deployed independently of this work to help improve estimates of grid reliability for the monitoring and evaluation efforts of a large investment in the electricity grid in Accra led by the Millennium Challenge Corporation [2]. In this work, we used data from three districts in Accra — Achimota, Kaneshie, and Dansoman collected over 27 months by 721 sensors. Achimota, Kaneshie, and Dansoman districts contain 52%, 16%, and 32% of the total sensors, respectively, and the overall deployment covers an area of approximately 130 km^2 [2].

3.2.1. Data description

PowerWatch devices are installed at outlets in households and businesses. GPS accuracy of the sensors lies within a few tens of meters which enables estimation of outage footprints independent of the underlying grid topology [2]. Whenever the sensor loses power or regains power, the devices send timestamped reports to a central server, where these reports are spatio-temporally clustered. A true positive outage requires at least two sensors to report a loss of power near together in space and time. PowerWatch is described in more detail in [2]. The PowerWatch central service returns a dataset so that each row represents an outage, and contains the list of sensors participating in the outage, the start and stop time of the outage, the number of sensors (size) of the outage, and the area of the outage (convex hull).

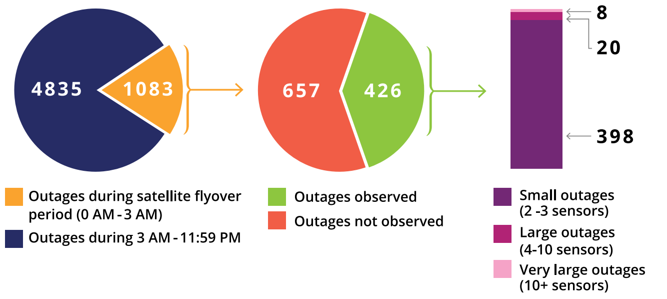


Fig. 3. Left: Outages occurring during the possible fly-over window. Center: Outages that intersected with a specific night's fly-over and did not have cloud cover. Right: Size of observed outages. The more the number of sensors reporting an outage, the larger is the spatial extent of an outage.

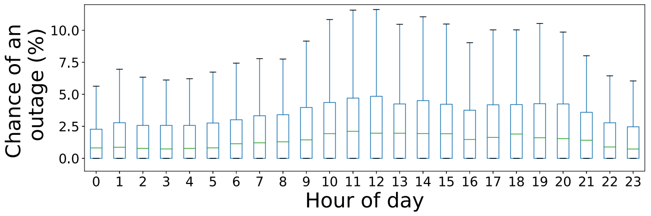


Fig. 4. Percentage chance of an ongoing outage by the hour of the day. An outage is half as likely to occur during the satellite's observation period (0–3 am) as during the grid's peak hours.

3.2.2. Limitations

The PowerWatch dataset comprises the largest and highest-resolution utility-independent measurement of power reliability on the African continent. However, PowerWatch sensors are only placed on a subset of the grid, leading to potential under-sampling of low-voltage outages (which may only impact a small, localized section of the grid). Larger outages tend to impact much of the grid simultaneously and increase the likelihood that a PowerWatch sensor is present to detect the outage.

4. Detecting outages

When the power goes out, lights go out. We, therefore, assume that an unusual dip in a region's (defined as a set of adjacent pixels) illumination levels is most likely caused by an outage occurring in that region. Expanding this observation, we can hypothesize that frequent sudden changes in a region's illumination levels can be attributed to electric supply inconsistency, while consistent illumination levels indicate a consistent electric supply. However, in doing so we must also be mindful that daily NL data can also have minor random variations and periodicity unrelated to outages. In this section, we present a method to detect outages using NL data, which we developed and evaluated based on the presence of high-frequency, high-resolution measurements taken by PowerWatch.

4.1. Overlap with PowerWatch

We only consider NL data that overlaps with the PowerWatch data for the supply inconsistency analysis. Fig. 3 summarizes the 426 PowerWatch outages that overlap with the nightlight dataset, a comparatively small percentage (7.2%) over the full PowerWatch dataset. The following subsection describes the steps taken to determine overlap.

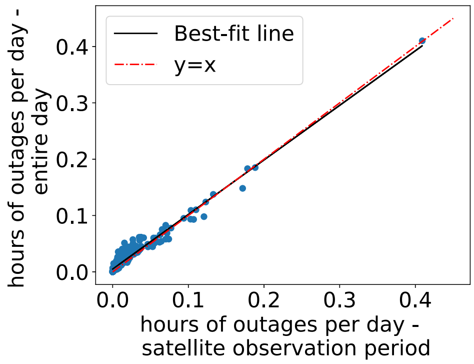


Fig. 5. The proportion of outage duration experienced during an entire day (0 am–11:59 pm) versus only during the satellite's observation period (0–3 am). $R^2 = 0.92$. Each point represents a monitored pixel. From this relationship we see that power supply inconsistency during nighttime is representative of power supply inconsistency during the entire day.

4.1.1. PowerWatch sample

We use data collected from 721 PowerWatch sensors deployed across Accra between June 2018 and September 2020. PowerWatch sensors are present in 210 of the 9984 nightlight pixels that cover Accra (each square pixel is 15 arc-seconds per size, or ≈450 m at the equator). PowerWatch sensors reported a total of 5918 outages over the course of the deployment. We refer to the pixels containing PowerWatch sensors as *monitored pixels* throughout this work and assume that a monitored pixel is experiencing an outage if at least one sensor inside the pixel reports an outage. This may seem strange, as PowerWatch considers space–time clusters before reporting an outage and ignores the reports from a single sensor, however a single sensor in a pixel can be part of a cluster across pixels so its presence still provides coverage. Monitored pixels span over an area of approximately 125 km^2 and there is an average of 4 PowerWatch sensors present in a monitored pixel and a maximum of 10.

4.1.2. Nightlight sample

The satellite captures Accra's nighttime scene between midnight and 3 am local time (GMT) daily as it flies overhead (this window is commonly referred to as the *observation period*), providing a temporal bound for any directly observable outages. PowerWatch recorded 1083 outages (18% of the total) during the satellite's observation period across the length of the deployment. We can characterize how the relatively narrow observation period impacts the coverage of a nightlight-based outage detection system by comparing outages recorded by PowerWatch during peak hours (10 am to 2 pm) to those collected by PowerWatch during the satellite's observation time, an exercise displayed as Fig. 4. We find that during peak hours the grid is likely to experience over twice as many outages as would be expected during the satellite observation period providing a source of error on an absolute basis. These dynamics lower our supply consistency estimates (biasing towards estimating stable grids), but as we show in Fig. 3, we can partially mitigate this by estimating the extent of suppression. At the same time, we also observe that pixels that experience a higher proportion of outage duration during the satellite's observation period tend to experience a higher proportion of outage duration during the entire day, providing confidence that inference based on satellite data collected during the observation period is representative on a relative basis. This strong agreement, shown in Fig. 5, means that if we are able to accurately predict the supply inconsistencies during the flyover hours we can then easily translate that into an estimate of grid unreliability for all hours of the day.

Further limiting our sample, we observe that although the observation period is a number of hours, only outages that are ongoing during the instant of image capture are being sampled (we refer to these outages as *flyover outages*). Additionally, we can only consider satellite observations on cloud-free nights as radiance attenuation due to clouds can produce misleading results, and this further reduces the number of *flyover outages* observed by the satellite. As shown in Fig. 3, *flyover outages* observed on cloud-free nights constitute 39% of the total outages recorded by PowerWatch during the satellite's observation period, leaving 426 outages that fully intersect between PowerWatch and NL. While this sample is somewhat limited, we show in Appendix A that cloud patterns do not excessively bias this dataset towards periods of fewer outages.

4.1.3. Classifying outages

In addition to outage occurrence and restoration timestamps, PowerWatch data also provides the cluster size of an outage, i.e., the total number of sensors that reported an outage within a 90 s period from the first sensor reporting an outage. Outage cluster size provides two important pieces of information that can be useful when searching for outage signals in NL data: (1) Confidence. The more PowerWatch sensors reporting an outage, the more likely the outage as viewed by PowerWatch is a true positive outage, and (2) Scale. As the PowerWatch cluster size increases, the outage will have a larger affected area — we computed the correlation between cluster size and area to be $r^2 = 0.975$, indicating very strong agreement.

Based on the area covered by outages of different cluster sizes, we classify outages into three categories — small outages ($2 \le clustersize < 4$), large outages ($4 \le clustersize < 10$), and very large outages ($clustersize \ge 10$). Mean hull areas of small, large, and very large flyover outages are 0.04 km^2, 1.14 km^2, and 39.23 km^2 respectively. Fig. 3 gives the total number of observed flyover outages belonging to each category. This information allows us to analyze the impact of outages of different scales on the dynamics of NL radiance.

4.1.4. Edge case detection

Some outages in the PowerWatch dataset lasted for multiple days, leading to multiple observations by the satellite. To handle this, we divide a long-lasting outage into daily instances and assign a unique outage label to each instance. For example, if an outage started on March 20, 1 am and power was restored on March 21, 9 pm, the satellite would observe the same outage on March 20 and 21 assuming both days were cloud-free. Since an outage was captured twice in this case, we assign a unique label to each of the two outage instances and treat them as separate outages. Therefore, each outage label will represent a unique outage-day pair. This increases the final dataset used for analysis in this section to contain 977 unique outage-day pairs observed by the satellite.

4.2. Methodology

The primary factors that determine the radiance of a pixel are the population density inside the pixel and whether the pixel is in an urban, rural, or peri-urban environment. The mean NL radiance of monitored pixels in Accra varies from 6 to 30 $\mathrm{nW/sr/cm}^2$. Comparing deviations in radiance between pixels with different radiance levels can be misleading, so we use z-scores of pixel radiance (i.e., pixel radiances are only compared to themselves) instead of raw radiance values. Z-scores are computed for every pixel on a per-day basis. The methodology we use to compute pixels' z-scores is described in Section 3.1. Z-scores allow us to consistently quantify and compare the deviations observed in radiance across different pixels irrespective of their radiance levels. Significantly high or low z-scores can be indicative of unusual events on the ground like power outages and we explore this finding in more detail in Sections 4.3 and 4.4.

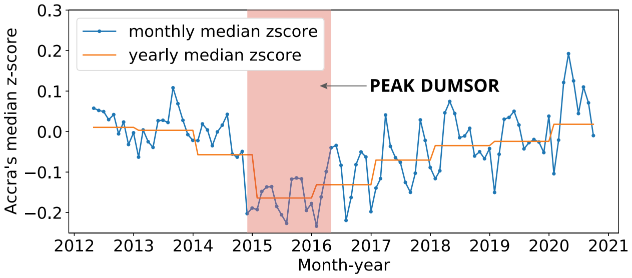


Fig. 6. Median z-score of Accra's NL radiance over time. A significant dip in Accra's median z-score indicates unstable lighting levels during the peak "Dumsor" period — late 2014 to early 2016.

4.3. Sanity check: Dumsor detection

In 2012, increasing demand for electricity and insufficient supply marked the beginning of a sustained period of power instability in Ghana that became known as "Dumsor" (combining the Twi words for off, "dum", and on, "sor") [32,33]. Dumsor peaked from late 2014 through late 2015 or early 2016 [33].

Dumsor consisted of frequent rolling blackouts and was prevalent across Accra, and so we expect the region-wide aggregated values of z-score and dispersion index to indicate a significant difference in NL radiance during the peak Dumsor period relative to their pre- and post-Dumsor values.

Indeed, the median monthly and yearly z-scores, shown in Fig. 6, exhibit a dip from November 2014 to June 2016 closely corresponding to, and likely attributable to, the Dumsor peak period. Further, median z-scores show Accra experienced relatively low illumination levels in 2015 and 2016, corresponding to the Dumsor power crisis.

Accra's average radiance grew significantly from 5.4 $\mathrm{nW/sr/cm^2}$ in 2012 to 10.14 $\mathrm{nW/sr/cm^2}$ in 2020, with a dip observed in 2015. For this specific analysis, in order to avoid the impact of this 8 year long, linear growth trend on z-score calculations we de-trended the radiance time-series data of every pixel. Pixel-level daily z-scores were computed using the mean and variance of the entire 8-year, de-trended time-series. Accra's median monthly and yearly z-score values were obtained by computing the median of all pixels' z-scores over the entire month and year, respectively.

The results obtained using z-scores show that the metric is able to detect large changes in lighting on the ground via satellite data [12–14]. Furthermore, the increasing trend of Accra's median z-score following 2015 indicates improvement in the region's power-supply consistency following peak Dumsor. This observation aligns with the information on progressively more stable power as reported by Ghanaian utilities [33,34].

4.4. Results

We present our key findings on how well our methodology can detect outages. We looked at z-scores of all the monitored pixels under outage and no outage conditions separately. Pixels when experiencing an outage reported z-scores with a mean of −0.98 and a standard deviation of 1.2. The same pixels while not experiencing an outage reported z-scores with a mean of 0.05 and a standard deviation of 0.08. Fig. 7 shows a CDF of z-scores of pixels with different outage experiences, where small outages refer to outages with cluster size $<$ 4 and large outages are outages with cluster size $\geq$ 4. Four important observations to be made based on this CDF plot: (1) The variance of a pixel's z-score when experiencing an outage is significantly higher than when it is not under an outage which is an indicator of higher fluctuations in a pixel's radiance during an outage, (2) the CDF plot moves towards the negative side of the z-score spectrum as we go from smaller outages to larger outages, (3) the distribution of z-scores of pixels under small outages overlaps the distribution of z-scores of pixels under no outage, making it difficult to disambiguate whether a pixel is experiencing a small outage or no outage at all by using its z-scores, and (4) it is easier to differentiate and identify pixels experiencing a larger outage from pixels not experiencing an outage.

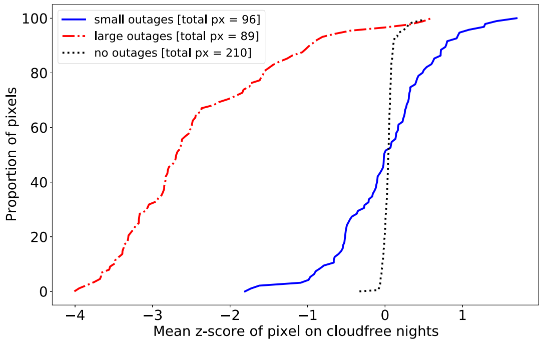


Fig. 7. CDF of z-scores of pixels under no outage, small outage, and large outage conditions. Total px represents the total number of pixels that experienced at least one of the three outage conditions — no, small, and large.

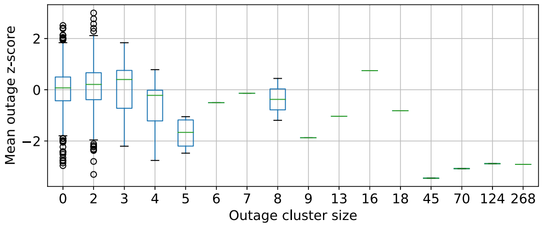


Fig. 8. Distributions of mean outage z-scores by outage cluster size. 0 represents no outage. Using NL, it is difficult to differentiate between regions experiencing smaller outages (cluster size $<$ 4) and regions experiencing no outages at all. Larger outages (cluster size $\geq$ 4) are relatively easier to identify.

It is also important to note that we observe z-scores on both sides of the spectrum – positive and negative – when pixels are experiencing a small outage. Since NL observations are heavily influenced by street lights, pixels tend to exhibit negative z-scores when small outages impact outdoor lighting. But when a small outage does not affect outdoor lighting, it has a negligible impact on the pixel's overall radiance. In this situation, a pixel's brightness is more likely to be influenced by other external factors than the underlying outage. Such a pixel could potentially appear brighter (positive z-score) due to three possible reasons: (1) increased ambient lighting in the region — atypical traffic conditions leading to extra lights, infrequently occurring night market, holiday/event celebrations [25], (2) cross-spillage of light from neighboring pixels [35], (3) misclassified clouds reflecting moonlight [31].

Furthermore, we analyzed the distribution of mean z-scores of all the observed outages grouped by their cluster sizes. The mean z-score of an outage is equal to the mean of z-scores of all the monitored pixels reporting an outage on a given day. Fig. 8 shows how much the mean z-scores of outages of a given size vary. An outage cluster size of 0 represents the mean z-scores of all the monitored pixels that were not reporting an outage. Outages with cluster size $<$ 4 exhibit similar variance in their z-scores as the region not experiencing any outages, which again indicates that it is difficult to identify and differentiate regions experiencing smaller outages from regions not experiencing any outages on a given day. The overall z-score distribution starts moving downwards towards the negative end of the z-score spectrum indicating noticeable dips in radiance during larger outages. However, even though these dips are not directly proportional to outage cluster sizes, the sample size of these outage cluster sizes is small, limiting their predictive value.

5. Outage prediction

In this section we attempt to train an outage predictor that uses the signal extracted in Section 4.4.

5.1. Preparing data

We divided Accra's NL grid into square patches of size $k \times k$ pixels^2 where k varied from 1 to 10 pixels. The goal was to predict whether a given patch on a given day contained a large outage or not. Spatial grouping of pixels allows us to factor in the impact of an outage on the immediate neighbors of an outage reporting pixel. Since we only possessed ground truth data for monitored pixels, we selected patches that contained at least one monitored pixel for training and testing purposes. Furthermore, recalling that z-score is only capable of detecting relatively larger outages, we labeled a patch-day tuple as an "outage" if at least one of the patch's pixels reported an outage of size ≥ 4, else we labeled it as "no outage". The positive training label, in this case, was "outage" and the negative was "no outage".

5.1.1. Data arrangement for random forest and logistic regression

For Accra's NL grid divided into patches of size $k \times k$, the dataset contained $m \times n$ rows and k^2 columns, where m represents the total number of days and n is the total number of monitored patches. Each row basically represented a unique patch-day tuple. For each patch-day tuple, the dataset contained k^2 features (or columns) where each feature represented the z-score of a unique pixel in that patch on that day. Each patch-day tuple also had an associated boolean label, i.e., whether the patch was under an outage or not on that day.

5.1.2. Data arrangement for Z-score threshold-based classifier

Instead of using z-scores of all the unique pixels belonging to a patch on a given day as features, we only used the mean of those z-scores as a feature for a patch-day tuple. The final dataset contained $m \times n$ rows and 1 columns, where the column represented mean z-scores.

5.1.3. Data splitting

We used the first 16 months of data for training our models, the following 6 months for validation, and the remaining 5 months for testing. In later evaluations, we also explored the robustness of these methods under temporally and spatially partitioned data splits.

5.2. Prediction models

5.2.1. Z-score threshold-based classifier (ZT)

Motivated by the results shown in Fig. 7, we implemented a simple classification method to threshold on z-scores. If the mean z-score of a patch on a given day (z_k) lied beyond the threshold, i.e., $z\_k > z\_upper$ or $z\_k < z\_lower$, it got classified as experiencing an "outage" and "no outage" otherwise. We trained one model for every patch size setting and then selected the set of hyperparameters – z_lower and z_upper – that gave the best training results for each setting.

5.2.2. Logistic regression (LR)

Convolutional neural networks are the most popular image classification models and to the best of our knowledge, the smallest images that CNNs have been used for classification are the MNIST images, which are of size 20 × 20 pixels^2 [36]. Since a CNN would be best suited for relatively larger patches – even a 10 × 10 patch in our case is too big as it represents an area of approximately 4.5×4.5 km^2 – a CNN was not the right choice for our work. We decided to use logistic regression since it is the logical choice for binary classification when the data precludes the use of a multi-layer classification CNN. We used an L2 penalty for regularization and implemented grid search to tune the model's hyperparameter using training data.

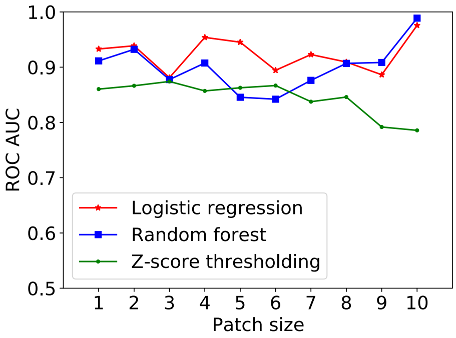


Fig. 9. ROC-AUC by patch size for three different classifiers. Patch size has no significant impact on models' performance.

Table 1
Performance summary table (ZT-z-score threshold, RF-random forest, LR-logistic regression). Positive labels = 40, patch size = 1.

Model	ROC-AUC	TP	FP	TN	FN	prec-ision	Recall (TPR)	FPR
ZT	0.86	36	1206	6692	4	0.03	0.90	0.15
LR	0.93	31	118	7780	9	0.21	0.78	0.01
RF	0.91	36	1130	6768	4	0.03	0.90	0.14

5.2.3. Random forest classifier (RF)

Random forest is an ensemble machine learning technique that combines multiple base classifiers or decision trees; its one major benefit is that it is not very prone to over-fitting [37]. The number of trees in the forest and the maximum depth of a tree were the two hyperparameters tuned using grid search for training the model.

It is important to note that classifiers like LR and RF predict the probability of each class. In a binary classification task, if a reading's probability is < 0.5 it is assigned a particular class, if not it is assigned the other class. The default probability threshold of 0.5 is not optimal for imbalanced datasets like ours. Therefore, the probability threshold was also added to the list of hyperparameters that were tuned using grid search for training LR and RF.

5.3. Results & evaluation

Accuracy is not a strong indicator of a model's performance when the dataset is imbalanced. For example, in our scenario we could simply predict each patch to never contain an outage regardless of the z-scores contained within and, due to the relative infrequency of outages, we would in this way obtain an accuracy of over 95%. So instead we resort to the Receiver Operating Characteristics-Area Under the Curve (ROC-AUC) value for evaluating the performance of our classification models. The ROC curve plots a model's true positive rate (TPR) to its false positive rate (FPR) as a function of the probability threshold value [38]. The larger the area under the ROC curve, the better a model's capability to correctly classify the positive class (outage) as positive and the negative class (no outage) as negative [38]. ROC-AUC varies from 0 to 1 where 0.5 represents random guessing. ROC-AUC can also be thought of as an objective measure to quantify how well a model performs at predicting compared to randomly guessing.

All the models described earlier in this section had their optimal hyperparameters determined via cross-validation for each of the patch size settings. Hyperparameter optimization using cross-validation helped us keep a check on model overfitting. Models were then trained using the determined hyperparameters and ROC-AUC scores were computed on the held-out test sets. Fig. 9 shows the ROC-AUC value for all the classifiers by patch size.

LR and RF showed a modest performance improvement as the patch size increased, while ZT's performance deteriorated. This difference in performances of LR/RF and ZT can be attributed to two reasons: First, LR and RF are capable of learning more complex decision boundaries than ZT. ZT has a strictly linear decision boundary. Second, by treating the z-score of each pixel inside a patch as a unique training feature, LR and RF can perform multi-dimensional (multi-feature) binary classification leveraging the local information of a patch. But the ZT model loses this local information because it only uses a single feature – mean z-score of a patch – for classification.

Nevertheless, all the models perform reasonably well for the smallest patch size setting of 1, i.e., at a pixel level. Pixel level outage estimation is the most desirable due to two important reasons: First, since pixel-level (≈450 m × 450 m) is the most granular, aggregation of pixel-level predictions would allow us to obtain outage estimates at any desired neighborhood level, which would not be possible with, for example, a 10×10 patch (≈4.5 km×4.5 km). Second, there is a potential chance of high label noise at higher patch settings because not every pixel in a patch is being monitored by PowerWatch sensors. Let us take an example of a 10×10 patch that contains a single monitored pixel. In such a case, if the monitored pixel is not experiencing an outage, the patch would be labeled as "no outage", even though all the remaining un-monitored pixels would be experiencing an outage. The likelihood of such label noise occurring grows quadratically with patch side length which makes it even more difficult to trust the prediction results at higher patch size settings. Thus, having discussed the rationale behind favoring pixel-level outage estimations, we perform all the further evaluations only using the patch of size 1.

Table 1 shows the performance summary of the classifiers for a patch of size 1. ZT and RF were able to correctly identify 90% of the actual outages while LR identified 78% of them. ZT and RF show a significantly high number of false positives (FP) relative to LR. FP means "no outage" entries are classified as "outages" by an algorithm. High FPs could represent cases when a large outage was detected by the satellite but since it did not impact the PowerWatch sensors inside a pixel, it was not reported. This is possible when a large outage impacts the circuit different from the one(s) PowerWatch sensors are connected to. True negative (TN) cases represent the algorithm correctly labeling "no outage" events while false negative (FN) cases represent the algorithm incorrectly labeling "outage" events as "no outage" events. FNs could possibly be cases when PowerWatch sensors detect an outage that does not impact the street lights and so the event goes undetected by the satellite. Low precision scores can be attributed to a high imbalance in the dataset and a significantly high number of FP relative to the total number of actual positive cases.

Depending on the context of a problem being addressed, the best-performing model – a model with maximum ROC-AUC – might not always be the right solution. Therefore, in addition to ROC-AUC, the model selection process should also consider the trade-offs between different derivations of the confusion matrix (FP vs FN, precision vs. recall). In this work, for example, we believe that FNs have a higher cost than FPs. An FN is a case the system believes there is not an outage where there is actually an outage, a costly mistake for consumers who may go longer without a service repair, and utilities who cannot learn correct patterns about how to increase operational efficiency. An FP is a case where the system reports an outage and there is none, the cost of which could be largely mitigated by calling the local utility engineer in the region predicted as experiencing a failure. Therefore, we evaluate performance with FN as our metric of choice and find that all three models performed well — exhibiting significantly fewer FNs at the cost of higher rates of FPs.

We chose LR and RF for conducting further experiments and analyzing case studies for two reasons: (1) both the models have high ROC-AUC scores, and (2) each model has a different strong point — LR has the lowest FPR and RF has the highest recall (TPR).

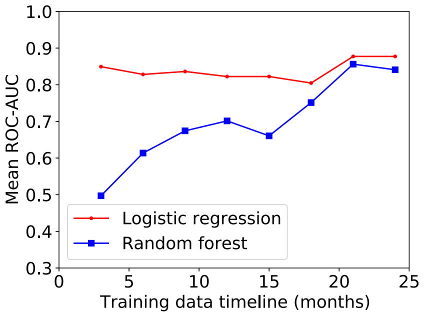


Fig. 10. Mean ROC-AUC value by the amount of historical temporal data used for training RF and LR for patch size = 1.

Table 2
Results of the spatial splitting experiment. Patch size = 1.

Model	ROC-AUC	Recall (TPR)	FPR	TNR	FNR
LR	0.95	0.89	0.10	0.90	0.10
RF	0.92	0.88	0.12	0.88	0.12

5.3.1. *Temporal splitting*

What is the minimum amount of historical NL data (z-scores) required for training a model to make reasonable supply inconsistency predictions for a patch size of 1? We employed a sliding window approach for the temporal splitting of training and testing data. Windows of different sizes were used to vary the temporal range of training data from 3 to 24 months in increments of 3 months. One step of the process involved training RF and LR using the data present inside the window and testing on the NL z-scores during the month following right after the window. We predicted whether a monitored pixel had an outage on a particular day during that month. After every step, the window is moved forward by 1 month and then the training–testing on limited temporal data is repeated until the end of the data timeline. We recorded the ROC-AUC value at every step and report the mean ROC-AUC value for each window size in Fig. 10. We implemented two rules to be able to evaluate the test output using ROC-AUC: (1) training data should contain at least 2 outage instances, and (2) test data should contain at least 1 outage instance. An instance is a pixel-day tuple and an outage instance represents a pixel-day tuple labeled as experiencing an outage. The first rule was set to ensure successful stratified splitting of training data into training and validation sets. Stratified splitting tries to create a train/validation split such that each split has the same proportion of "outage" and "no outage" instances [39], which requires at least 2 outage instances to be present in the original training dataset. Furthermore, TPR and FPR measurements essential for ROC-AUC computations require both – positive ("outage") and negative ("no outage") – examples to be present in the test set. Therefore, the second rule was set to ensure the validity of ROC-AUC computations. Fig. 10 shows that only 3 months of historical data were enough for LR to make accurate predictions and adding any amount of extra data only minimally improved performance. For RF, adding more historical data improved performance almost linearly.

5.3.2. *Spatial splitting*

PowerWatch sensors are deployed in 13 different areas spread across 3 districts in Accra. Unlike temporal splitting where we preserved space and split data by time, here we preserve the entire timeline and split the data by space. RF and LR were trained using z-scores of monitored pixels belonging to 11 areas and were then tested on data from the remaining 2 areas. Trained models predicted when a monitored pixel in the 2 test areas experienced an outage over their entire monitoring timeline. Spatial-splitting experiment was repeated multiple times with a random selection of 11 areas for training and 2

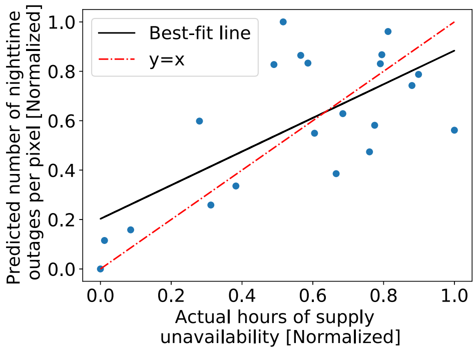


Fig. 11. The number of outages per pixel predicted by LR using NL data versus actual supply unavailability from the Dumsor Report during April–May 2015. Each point represents an area monitored by the Dumsor Report group [32]. Areas reported as experiencing longer duration of outages by the Dumsor Report were also identified as areas experiencing higher frequency of outages per pixel by our NL-based predictors.

areas for testing in every iteration, ensuring that every area is included in the test set at least once. LR and RF exhibited average ROC-AUC values of 0.95 and 0.92, with standard deviations of 0.04 and 0.05 respectively. Table 2 summarizes the results of this experiment. LR and RF exhibited similar performance with high recall and low FNR which shows that models are capable of correctly predicting outages in areas they were not trained for. These results provide an encouraging first step towards validating the applicability of this model in settings beyond Accra.

We observed that both models exhibited better ROC-AUC in spatial splitting experiments compared to temporal splitting experiments. In temporal-splitting, we train the model on a fraction of the data timeline for all geographic areas as discussed in Section 5.3.1, while in spatial-splitting we train the model on the entire data timeline over a fraction of geographic areas. Training models using the entire data timeline, for spatial-splitting experiments, seems to make models learn the region's outage dynamics better, leading to more promising prediction results for the held-out set. It suggests that temporal variations like weather are more influential than the physical differences imposed by the city making temporal-splitting relatively more challenging.

5.4. *Dumsor report*

In this section, we compare our system against one additional dataset from Accra. Local researchers working independently of the utility studied the frequency and duration of power outages in 21 different areas in Accra during the peak Dumsor period. They collected this data over two weeks (April 28 to May 15, 2015) using volunteers and mobile surveys, asking participants to keep a journal of outage start and stop times [32]. Considering the presence of human error in recorded measurements, and the otherwise spatially- and temporally-limited sample, this is not ideal ground truth. Although the Dumsor Report represents a relatively small and likely incomplete dataset, it is one of the only quantitative datasets on grid reliability during Dumsor [40], and is one of the only sources of data to compare model performance against.

LR and RF models trained on NL data of monitored pixels were used to predict outages at pixel level in 21 areas covered by the Dumsor Report. As a metric of predicted supply inconsistency, we measured the total number of predicted outages per pixel for every area and compared it with the observations presented in the Dumsor Report, which we show in Fig. 11. Our data largely agrees with the Dumsor Report — regions that were reported to have experienced a higher duration of supply unavailability according to the Dumsor Report were also the regions predicted to have experienced a higher number of outages per pixel during nighttime. Pearson correlation coefficient (r) values of 0.69 and 0.68 for LR and RF, respectively, show that our models are able to reasonably identify and correctly rank regions based on their power-supply inconsistencies, giving us confidence that our model can uncover large supply inconsistencies across regions in Accra, even over short periods.

It is important to note that obtaining high-quality ground truth reliability data for training and evaluating models is a challenging task. Most utilities use the supervisory control and data acquisition (SCADA) system to capture outage data at high-voltage (HV) and medium-voltage (MV) levels. However, the majority of the utilities, particularly in developing countries, either do not capture or do not share access to low-voltage (LV) distribution grid outage data. HV/MV outages occur rarely and have a wider footprint, while LV outages tend to occur more frequently and have a smaller footprint. Due to these differences in characteristics of HV/MV and LV outages, training models just on HV/MV level outages might not perform well further down the hierarchy of the grid. However, LV-level, time-series outage data sensed using advanced metering infrastructure can be a reliable source of ground truth to train and evaluate our models. Such a dataset can even augment the model training process that could potentially improve the performance and generality of our models.

6. Discussion

6.1. *Potential applications*

NL-based outage detection enables a number of potential applications, including but not limited to:

- Identifying regions experiencing frequent outages in a city or an entire country would require large-scale deployment of smart sensors which can be expensive. This can be achieved at a significantly lower cost by leveraging the granular insights gained from freely available NL data. Furthermore, the global consistency of NL data allows us to perform cross geographical grid reliability comparisons at no cost.
- It can serve as a system monitoring tool that allows electricity regulators, policy-makers, and international donors to track and evaluate grid reliability independently from the utilities on a daily basis without physical intervention.
- International organizations like the World Bank conduct annual infrastructure surveys that contain little information on grid reliability. These surveys are aggregated at country level and are subject to measurement noise [1]. Highly granular and consistent grid reliability estimates (pixel-level) obtained using NL data can complement infrastructure surveys.
- Utilities often use two metrics – SAIDI (System Average Interruption Duration Index) and SAIFI (System Average Interruption Frequency Index) – to quantify grid reliability. SAIDI and SAIFI represent the average duration and frequency of interruption experienced by a customer over a year respectively. The quality of data and methodologies to compute SAIDI and SAIFI vary widely. NL data can be used to verify the utility-reported SAIDI and SAIFI values to ensure the veracity of claims made by utilities.
- For utilities, having better spatio-temporal visibility into supply inconsistencies can help in recommending targeted maintenance and allocating outage response resources. This can help to reduce the number of interruptions, and also enhance revenue from increased electricity sales.
- NL-based outage estimates can complement other research that studies the economic and political effects of variable grid reliability. The creation of a spatially-explicit detailed historical record of outages can help with evaluating long-term grid performance improvements, studying the equity of outages across regions, and identifying disparities causing varied outage experiences.

6.2. Shortcomings and limitations

We also introduce some vital shortcomings and limitations associated with measuring outages via NL data with verification via PowerWatch sensors:

- As stated in Section 3.1, NL data are recorded only once per night per region, limiting the number of outages that could be captured by the satellite to only the outages that persist as the satellite flies over. For Accra, nighttime power supply inconsistency is reasonably indicative of daytime power supply inconsistency, but this may not be true for other cities.
- The number of outages observed by the satellite is further reduced by the loss of daily NL data due to clouds. It is important to note that data cleaning to mitigate clouds is standard practice [5,19] and the need for this filter is fundamental. Throughout this work, we have reported the data filtered to inform researchers considering the suitability of this dataset.
- Accuracy of the cloud cover classification technique used in the VIIRS suite is ≈85% [31]. As a consequence, some cloudy pixels get misclassified as cloud-free and vice versa, which leaves residual noise in the NL dataset even after the pre-processing stage.
- Temporal constraints of the nightlight data source make real-time observations infeasible.
- NL satellite observations are heavily influenced by street lighting in the area. PowerWatch sensors are connected to residential and small commercial customers that often individually do not have enough external lighting to impact the overall radiance of a pixel. Outages reported by PowerWatch sensors can be limited to a small set of houses being supplied by a faulty transformer which has no impact on the street lighting in the area. In such cases, an outage that affects a house or a group of houses may only have a small impact on the overall pixel radiance, going undetected by the satellite.
- PowerWatch sensors report only the outages that affect the houses or shops they are monitoring. There may be outages of varying sizes that do not impact PowerWatch customers but do have an impact on the radiance of a pixel, thereby affecting the results of our analysis. This introduces label noise in an already imbalanced dataset which limits the prediction performance. In this work, we minimize the effect of label noise by focusing only on pixel-level predictions. But as the likelihood of label noise increases with increasing patch size, predictions over larger patches become highly uncertain.
- The overall data distribution is skewed or imbalanced because events like outages are rare and an even smaller number of outages is captured by the satellite. Such imbalance introduces five major challenges: (1) increases the likelihood of model overfitting, (2) training models using conventional metrics like accuracy becomes unproductive, (3) an extra hyperparameter – probability threshold – needs to be tuned, (4) model selection becomes more complex — the best performing model might not be the best solution for a problem, (5) label noise in an imbalanced dataset can deteriorate prediction performance. Throughout Section 5, we have reported the steps taken to address these challenges.

Despite all the limitations, it is important to note that deployment cost limits the coverage of PowerWatch sensors in Accra and other geographies. Our nightlight-based system, which does not depend on expensive deployments, might be a realistic way to take critical grid performance measurements in many countries, even if imperfect due to resolution limitations.

6.3. Future work

6.3.1. Predictive performance

In order to increase the predictive performance of our model, we would expect to need some combination of NL data with a higher spatial resolution (allowing us to see smaller outages), longer sampling periods (allowing us to see more outages), and/or faster samples (allowing us to see shorter outages). Further, if these features are made available, we may be able to characterize the behavior of the grid beyond just outages (i.e., looking for rolling outages, studying household and/or rural energy usage patterns, etc.). However, without improvements, this work remains if not out-of-reach, then at least much more difficult with NL data alone.

Modest improvements using the existing data might be obtained with some research into the machine learning methods. For example, a kernel-based or simple convolutional model could help better identify outages by leveraging spatial proximities of abnormal pixel illumination. The intuition here is that three pixels with all abnormally low NL values are more likely to be an outage if they are right next to each other than if they were distributed randomly in a 5 × 5 grid. Our current method does not have the spatial structure to capture such an insight. Furthermore, our current method does not particularly consider the impact of weather conditions or seasons on the underlying NL distributions that impact z-score computations. Considering a different NL distribution per month/season instead of a year in z-score calculations could potentially help us mitigate the impact of temporal variations in NL data.

While we wait for higher resolution and more frequent satellite images, improving and augmenting ground truth data used to train our models could also improve predictive performance. Existing data on outages collected by grid operators through advanced metering infrastructure should be explored (using similar methods like this work). Further, high-quality simulation of failures in regions without ground truth measurements might additionally bootstrap prediction accuracy.

6.3.2. Generalizing satellite-based outage detection

Having a technique that generalizes to different contexts is a prerequisite to studying grid reliability across multiple regions. A key motivation of this work is a vision of globally-scaled observations of grid reliability, and although our system works well in Accra, we found that it does not trivially cross borders. From a machine learning perspective, this is hardly surprising; related work [41] that estimates economic outcomes of importance in Sub-Saharan Africa from satellite imagery has found that labeled data from many countries is vital for good out-of-sample prediction. However, given the normalizing properties of Z-scores and the relatively high performance of our ML models, it seems plausible that further sensor data from additional cities could rapidly improve the out-of-sample performance of this model. One possibility to note is that the behavior of the grid is not homogeneous across the Earth, and a single global model may be sub-optimal. Instead, we plan on developing different predictors for different regions, raising questions about exactly how local a predictor needs to be, and also potentially increasing dependence on multiple different local ground truth outage datasets.

7. Conclusion

Nightlight data are becoming popular as researchers uncover pieces of a compelling puzzle, and as policymakers and multi-lateral organizations become more aware of the consequences of massive data gaps that otherwise could only be filled by massive spending. However, nightlight data at its current spatial, temporal, and noise characteristics are not a turn-key solution for monitoring energy inconsistencies, and, like any other tool, are useful, but only for the right applications. For the right application, our methodology is powerful. Two of the major international reliability KPIs, SAIDI and SAIFI, are often built from utility-collected data that undersample high-, medium-, and low-voltage outages. If our system can better estimate high- and medium-voltage outages which tend to be larger in size, leaving smaller low-voltage outages on the table until the next generation of satellites, that alone meaningfully raises the quality of these estimates, returning

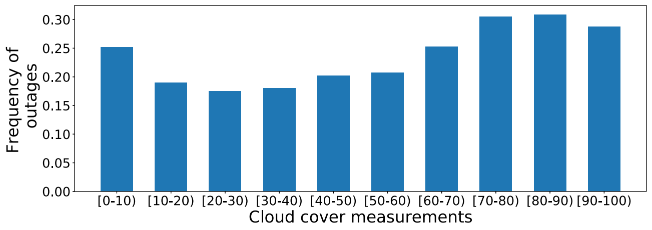


Fig. A.1. Frequency of outages by binned cloud cover measurements. The frequency of outages was calculated by dividing the number of outages triggered for every cloud-cover bin by the total number of times the cloud-cover bin was encountered in Accra. Outages are more common but not exclusive to higher cloud cover levels.

value to regulators, utilities, and researchers. Also interesting is that while we show we can improve satellite-based models by training based on measurements collected by sensors on the ground, we observe the need to generalize our technique. This opens a new direction of research, allowing us to imagine a hybrid system that better anticipates the correct, minimal set of data needed on Earth to best support the efforts from space, which could further improve the accuracy and value of remote measurements of the grid and other complicated, massive, critical, and understudied systems.

CRediT authorship contribution statement

Zeal Shah: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Noah Klugman:** Conceptualization, Methodology, Resources, Writing – original draft. **Gabriel Cadamuro:** Conceptualization, Methodology, Writing – review & editing. **Feng-Chi Hsu:** Data curation, Resources, Writing – review & editing. **Christopher D. Elvidge:** Data curation, resources. **Jay Taneja:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

We thank The Rockefeller Foundation, USA for their generous support of this work under Grant #2018 POW 004.

Appendix A. Clouds and outages

We lose 57.6% of Accra's total NL readings due to cloud cover. However, outage frequency increases during inclement weather [42], so while clouds limit our total number of samples, they also provide us with a signal to test our detector against. Here, we explore if there is a higher rate of outages in Accra on cloudier days (as expected), and estimate the data lost due to clouds on the overall proportion of detectable outages.

We used PowerWatch outage data and Accra's historical hourly cloud-cover data downloaded using the World Weather Online (WWO) API [43] to see if more outages tend to occur on cloudier days. We assumed that the hourly cloud-cover level across Accra remained consistent and equal to the hourly cloud-cover data from WWO API. WWO Cloud-cover readings ranged from 0 to 100 and so for simplicity, we grouped the cloud-cover readings into 10 discrete bins, each of size 10 as shown in Fig. A.1. For every cloud-cover bin, we then calculated the corresponding frequency of outages as the ratio of the total number of outages that started during the presence of that cloud-cover bin to the total number of times that specific cloud-cover bin was encountered in Accra. As shown in Fig. A.1, it was found that, in Accra, more outages tend to occur on cloudier days but they are not exclusive to cloudier days. This finding suggests that the cloud-free outage dataset is not biased towards days with fewer outages.

When relying on information from cloud-free NL readings, we found that *flyover outages* constituted 26% of all outages that occurred during the satellite's observation period (0–3 am local time) on cloud-free nights. However, we would like to account for the number of outage-day pairs that occurred during the flyover period but otherwise went unobserved due to the presence of clouds, which we can do by interpolating from the PowerWatch dataset. We are comfortable with a simple interpolation because, as discussed above, the frequency of outages during high cloud cover and low cloud cover was not significantly different, giving us confidence that the cloud-free NL dataset would not be biased towards periods with low outage frequency. According to PowerWatch readings, a total of 1510 outage-day pairs occurred during the satellite's observation period on cloudy nights, allowing us to estimate that 353 outage-day pairs (26%) would have been captured by the satellite if not for the clouds.

Appendix B. Dispersion index

B.1. Dispersion index computation

We identified the dispersion index of a pixel's radiance as one of the indices indicative of power outages. Dispersion index for a pixel p is calculated as $D_t p = \sigma_t^2 p / h_t p$, where $\mu_t p$ and $\sigma_t^2 p$ are the mean and variance of pixel p's radiance computed over a specific timeline t – weeks, months, or years. This metric was previously proposed [10] as a means of comparing NL-based supply inconsistency estimates across different cities. According to that work, higher dispersion of radiance in a city represents higher fluctuations in its brightness, which can be a potential indicator of inconsistent power supplies.

B.2. Detecting outages using dispersion index

Dispersion index is a metric proposed in [10] for using nighttime illumination levels to compare the consistency of the power supply of different regions. As discussed in Section 3.1, dispersion index is the ratio of the variance to the mean of a pixel's radiance. The monthly dispersion index is computed for every monitored pixel using cloud-free NL data. According to [10], a higher dispersion value for a pixel indicates fluctuating lighting levels, which in turn could be the result of an inconsistent power supply, while a lower dispersion value could potentially indicate a stable supply. The authors of [10] demonstrate that it is possible to identify and compare supply fluctuations in two different cities with comparable ranges of NL radiance – Dhaka and Nairobi – using the dispersion index value. In this work, we applied the same methodology to compare dispersion index values of pixels with different outage frequencies.

B.3. Dumsor detection using dispersion index

Monthly/yearly dispersion index values were calculated using the monthly/yearly mean and variance of a pixel's raw NL radiance. The median monthly/yearly dispersion index values of all the pixels across Accra were computed to obtain the monthly/yearly dispersion index for the entire region of Accra.

Fig. B.2 shows Accra's monthly and yearly median dispersion index over time. A higher dispersion index represents greater fluctuations in the radiance. The monthly and yearly plots in Fig. B.2 show that Accra's dispersion index peaked during 2015 and 2016, matching the Dumsor

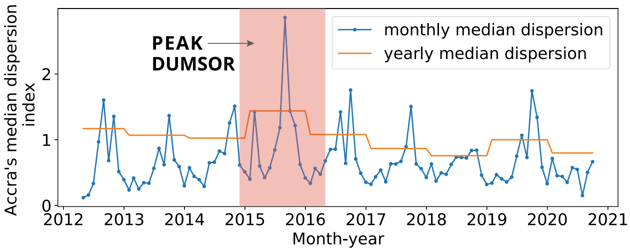


Fig. B.2. Median NL dispersion index of Accra's NL radiance over time. A significant spike in the dispersion index indicates unstable lighting levels during the peak "Dumsor" period — late 2014 to early 2016.

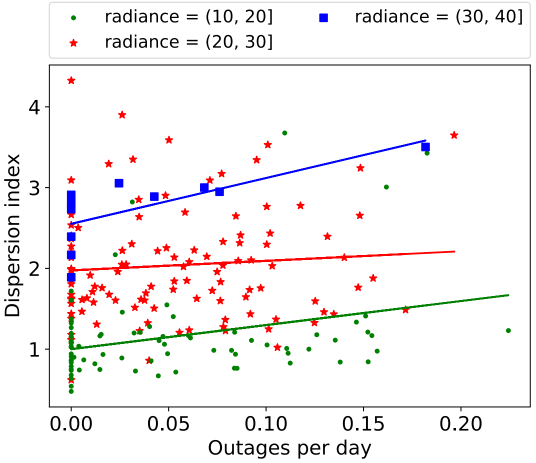


Fig. B.3. A plot of dispersion index by outage frequency. Each point represents a monitored pixel. The legend represents radiance ranges to indicate pixels belonging to groups having a similar range of mean radiances for comparison purposes. r^2 values for radiance groups $(10, 20], (20, 30]$, and $(30, 40]$ are 0.09, 0.006, and 0.51 respectively. This shows that an NL-based dispersion index is not very representative of supply inconsistencies at a pixel level.

timeline and indicating that the Dumsor outages not only reduced the overall mean radiance of the region but also resulted in a high degree of lighting fluctuations.

The results obtained using the dispersion index showed that the index was able to detect large changes in lighting on the ground via satellite data [12–14]. Furthermore, the decreasing trend of Accra's dispersion index following 2015 indicates improvement in the region's power-supply consistency following peak Dumsor. This observation aligns with the information on progressively more stable power as reported by Ghanaian utilities [33,34]. Although the dispersion index passed the sanity check, it did not prove to be an influential metric for studying different outage experiences at pixel level as discussed in the next section.

B.4. Dispersion index and outage frequency

In Section 4.3 of the main paper, we demonstrated that the nightlights dispersion index can be indicative of power supply inconsistencies at the city level. As an extension to that study, we analyzed the effectiveness of an NL-based dispersion index in estimating power supply inconsistencies at a pixel level by attempting to understand if pixels with higher outage frequency experienced higher lighting fluctuations and thereby higher dispersion. Outage frequency for every pixel is equal to the outages experienced by a pixel divided by the total number of days it was monitored by the sensors. Additionally, as stated in [10], dispersion index can only be used to compare cities with similar mean radiances, and so we grouped pixels with a similar range of mean radiances together, creating 3 different groups of pixels: low-range ($radiance \in (10, 20]$), medium-range ($radiance \in (20, 30]$), and high-range ($radiance \in (30, 40]$). Two important observations can be made based on the output shown in Fig. B.3: (1) The higher the mean radiance of the group, the higher the overall dispersion index value is. This observation matches well with results in [10] that indicate that the higher the mean radiance of a country, the higher its dispersion index is. For example, the dispersion index of Shanghai is higher than the dispersion index of Nairobi, and (2) r^2 values for radiance groups $(10, 20], (20, 30]$, and $(30, 40]$ are 0.09, 0.006, and 0.51 respectively, showing that the relationship between outage frequency and dispersion index at a pixel-level is not very obvious and so it is difficult to estimate power supply inconsistencies simply using dispersion index at a pixel-level. Nevertheless, the dispersion index is a valuable metric to compare power supply inconsistencies using nighttime lights across cities with a comparable range of mean radiances.

References

[1] Correa S, Klugman N, Taneja J. Deployment strategies for crowdsourced power outage detection. In: 2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm). 2018, p. 1–6.

[2] Klugman N, Adkins J, Paszkiewicz E, Hickman M, Podolsky M, Taneja J, Dutta P. Watching the grid: Utility-independent measurements of electricity reliability in Accra, Ghana. In: Proceedings of the 20th international conference on information processing in sensor networks. ACM; 2021.

[3] Taneja J. Measuring electricity reliability in Kenya. 2017.

[4] OpenStreetMap contributors. Planet dump. 2017, retrieved from https://planet.osm.org, https://www.openstreetmap.org.

[5] Elvidge C, Baugh K, Zhizhin M, Hsu F, Ghosh T. VIIRS night-time lights. Int J Remote Sens 2017.

[6] Darko KA. Power outages hit parts of Ghana. 2020, URL https://www.myjoyonline.com/power-outages-hit-parts-of-ghana/.

[7] Min B, Baugh K, Monroe T, Glodblatt R, Stewart B, Kosimdou-Bradley W, Crull D. Light every night – new nighttime light data set and tools for development. URL https://blogs.worldbank.org/opendata/light-every-night-new-nighttime-light-data-set-and-tools-development.

[8] World bank - light every night. URL https://worldbank.github.io/OpenNightLights/wb-light-every-night-readme.html.

[9] Earth observation group, the payne institute for public policy, colorado school of mines. URL https://payneinstitute.mines.edu/eog.

[10] Elvidge CD, Hsu F-C, Zhizhin M, Ghosh T, Taneja J, Bazilian M. Indicators of electric power instability from satellite observed nighttime lights. Remote Sens 2020;12(19):3194.

[11] Dugoua E, Kennedy R, Shiran M, Urpelainen J. Assessing reliability of electricity grid services from space: The case of Uttar Pradesh, India. Working paper, 2020.

[12] Aubrecht C, Elvidge C, Ziskin D, Baugh K, Tuttle B, Erwin E, Kerle N. Observing power blackouts from space - a disaster related study. 2009.

[13] Elvidge C, Baugh K, Sutton P, Bhaduri B, Tuttle B, Ghosh T, Ziskin D, Erwin E. Who's in the dark—satellite based estimates of electrification rates. In: Urban remote sensing: Monitoring, synthesis and modeling in the urban environment. 2011.

[14] Shah Z, Hsu F, Elvidge C, Taneja J. Mapping disasters & tracking recovery in conflict zones using nighttime lights.. In: IEEE global humanitarian technical conference. 2020.

[15] Elvidge C, Baugh K, Zhizhin M, Hsu F-C, Ghosh T. Preliminary results on VIIRS detection of power outages in India. In: Proceedings of the Asian Conference on Remote Sensing; 2011.

[16] Min B, O'Keeffe Z, Zhang F. Whose power gets cut? Using high.-frequency satellite images to measure power supply irregularity. Policy research working paper: No. WPS 8131, World Bank Group; 2017.

[17] Mann M, Melaas E, Malik A. Using VIIRS day/night band to measure electricity supply reliability: Preliminary results from maharashtra. Remote Sens 2016.

[18] Prayas electricity monitoring initiative. http://www.watchyourpower.org/the_initiative.ph.

[19] Hsu F, Zhizhin M, Ghosh T, Elvidge C, Taneja J. The annual cycling of nighttime lights in India. Remote Sens 2021;13(6). http://dx.doi.org/10.3390/rs13061199, URL https://www.mdpi.com/2072-4292/13/6/1199.

[20] Witmer F, O'Loughlin J. Detecting the effects of wars in the caucasus regions of Russia and Georgia using radiometrically normalized DMSP-OLS nighttime lights imagery. GISci Remote Sens 2011.

[21] Cao C, Shao X, Uprety S. Detecting light outages after severestorms using the S-NPP/VIIRS day/night band radiances. Geosci Remote Sens Lett 2013.

[22] Gillespie T, Frankenberg E, Chum K, Thomas D. Nighttime lights time series of tsunami damage, recovery, and economic metrics insumatra, Indonesia. Remote Sens Lett 2014.

[23] Elvidge C, Sutton P, Ghosh T, Tuttle B, Baugh K, Bhaduri B, Bright E. A global poverty map derived from satellite data. Comp Geosci 2009;35.

[24] Min B. Power and the vote. Cambridge University Press; 2015.

[25] Min B, Gaba KM. Tracking electrification in Vietnam using nighttime lights. Remote Sens 2014.

[26] Falchetta G, Pachauri S, Parkinson S, Byers E. A high-resolution gridded dataset to assess electrification in sub-Saharan Africa. Sci Data 2019;6(1):1–9.

[27] Chand TRK, Badarinath K, Elvidge C, Tuttle B. Spatial characterization of electrical power consumption patterns over India using temporal DMSP-OLS nighttime satellite data. Int J Remote Sens 2009.

[28] Elvidge C, Cinzano P, Pettit D, Arvesen J, Sutton P, Small C, Nemani R, Longcore T, Rich C, Safran J, et al. The nightsat mission concept. Int J Remote Sens 2007.

[29] Elvidge C, Hsu F, Ghosh T, Zhizhin M. World tour of COVID-19 impacts on nighttime lights. 2020.

[30] Rubin S, Goldblatt R, Park H. Nighttime lights are revolutionizing the way we understand COVID-19 and our world. 2020.

[31] Wang Z, Román MO, Kalb VL, Miller SD, Zhang J, Shrestha RM. Quantifying uncertainties in nighttime light retrievals from Suomi-NPP and NOAA-20 VIIRS day/night band data. Remote Sens Environ 2021;263:112557.

[32] Aidoo K, Dontoh E. Scientific dumsor report: See which areas enjoy or suffer most. 2015, URL https://citifmonline.com/2015/08/scientific-dumsor-report-see-which-areas-enjoy-or-suffer-most/.

[33] Nyarko Kumi E. Is Ghana's dumsor over? URL https://www.energyforgrowth.org/memo/is-ghanas-dumsor-over/.

[34] of Ghana EC. National energy statistics (2009–2018).

[35] Abrahams A. Improving the granularity of nighttime lights satellite imagery. 2015, URL https://blogs.worldbank.org/impactevaluations/improving-granularity-nighttime-lights-satellite-imagery-guest-post-alexei-abrahams.

[36] Deng L. The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag 2012;29(6):141–2. http://dx.doi.org/10.1109/MSP.2012.2211477.

[37] Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. Syst Sci Control Eng 2014;2(1):602–9.

[38] Classification: ROC Curve and AUC. URL https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.

[40] Aidoo K, Briggs RC. Underpowered: Rolling blackouts in africa disproportionately hurt the poor. Afr Stud Rev 2019;62(3).

[41] Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. Combining satellite imagery and machine learning to predict poverty. 2016;353(6301):790–4.

[42] Campbell RJ, Lowry S. Weather-related power outages and electric system resiliency. Congressional Research Service, Library of Congress Washington, DC; 2012.

[43] World weather online local history API. URL https://www.worldweatheronline.com/developer/api/historical-weather-api.aspx.