

But clouds got in my way: Bias and bias correction of VIIRS nighttime lights data in the presence of clouds

Ayush Patnaik Ajay Shah Anshul Tayal
Susan Thomas *

xKDR Forum

8th October 2021

Abstract

The VIIRS nighttime lights dataset constitutes progress in the measurement of night lights radiance, with monthly data at a pixel of roughly $0.5\text{km} \times 0.5\text{km}$. We identify a downward bias in the reported radiance when the number of cloud-free images in a month is low. This bias often takes on large values from -10% to -30%. We develop a cautious bias-correction scheme which partially addresses this problem. This scheme is applied upon the pixel-level dataset to create an improved dataset. The bias-corrected data hews closer to the ground truth as seen in household survey data.

*We are grateful, for help and valuable discussions, with Pramod Sinha, Geetika Palta, Radhika Pandey, Robert CM Beyer of the World Bank and Kim Baugh of NOAA.

Contents

1	Introduction	3
2	Nighttime lights as a measure of economic activity	4
3	Data description	5
4	Conventional methods of cleaning nighttime lights data	9
5	The problem of bias	11
5.1	Two rival explanations	12
5.2	Implications	16
5.3	Four classes of pixels	16
6	A partial bias correction scheme	19
7	The bias-correction scheme in action	20
7.1	A Case D example: Bombay	22
7.2	A Case D example: Johannesburg	24
7.3	A Case B example: Dubai	24
7.4	A Case C example: Singapore	25
7.5	Large-scale evidence for a country: India	27
8	Reproducible research	30
9	Conclusions	32
	Appendices	35
A	Conventional cleaning: detailed methodology	35
B	Summary statistics about radiance at the pixel level: Conventional vs. new	37
C	Comparing the cumulative density of seasonal factor range: conventional vs. new	38

1 Introduction

A remarkable development in economic measurement was satellite imagery that sees nighttime lights, and gives a pathway to measuring economic activity without requiring institutional capacity in the statistical system on the ground. The first wave of research used the DMSP-OLS dataset, which had annual images from 1992 to 2013. An improvement in this field was the launch of Suomi-NPP in 2012 where the pixels are smaller, and the frequency shifted from annual to monthly. This led to the next wave of research papers from 2012 onward which uses NOAA monthly dataset.

In this paper, we demonstrate that in months where NOAA sees a below-median value for the number of cloud-free images, there is a downward bias in the observed radiance. One hypothesis is that economic activity is lower in more cloudy months. A second is that perhaps there is bias when clouds get in the way. We show that there is a bias associated with the number of cloud-free images. This statistical regularity lends itself to the development of a partial bias correction scheme.

We work at the pixel level, and construct a new nighttime lights dataset that is compatible with the NOAA release. With this, researchers can compute the aggregate radiance in a geographical region of interest. The approach is conservative and corrects the bias for a limited class of situations, leaving many pixels and months intact. We identify a class of pixels which are cloudy in most months, where the bias is not identified and is not corrected.

As an example, we apply the method to the nighttime lights data for Bombay. We find that while the deviation from annual mean of conventional nighttime lights data for Bombay in July is -28.06%, this value for the new nighttime lights data is -6.95%. When applied for all the homogeneous regions of India, the median value for the average seasonal factor range drops from 50.10 to 26.14 percentage points. Gains are obtained for almost all the “homogeneous regions” make up the country.

This paper is accompanied with the release of an open-source package through which the results here can be reproduced, and applied to any of the NASA/NOAA monthly data files.

Table 1 Data sources for nighttime lights

There are two widely used datasets for nighttime lights. DMSP-OLS is an annual dataset that is available from 1992 to 2013. This was discontinued after the launch of Suomi-NPP in 2012, which contains the “Visible Infrared Imaging Radiometer Suite” (VIIRS) sensor. Suomi-NPP VIIRS has a spatial resolution of 15 arc-seconds ($\approx 0.5\text{km} \times 0.5\text{km}$ at the equator) which is superior to the value of 30 arc-seconds ($\approx 1\text{km} \times 1\text{km}$ at the equator) of DMSP-OLS. The VIIRS dataset is released by NASA on a daily frequency. NOAA processes the daily data, and releases monthly averages based on the cloud-free days.

Satellite	Source	Availability	Frequency	Resolution
DMSP-OLS	DoD-NOAA	1992-2013	Annual	30''
Suomi-NPP	NASA	April 2012 - Present	Daily	15''
Suomi-NPP	NOAA	April 2012 - Present	Monthly	15''

2 Nighttime lights as a measure of economic activity

From the early 1990s, nighttime lights has been used to generate measures of economic activity such as GDP with accuracy, latency and geographical resolution that are superior to conventional methods of measurement. As an example, Doll, Muller and Morley (2006) found a strong correlation between the nighttime lights, aggregated over a geographical region of interest, with the regional GDP of 11 European Union countries, and US GDP at the state level. A significant research literature has emerged, including unusual applications in situations where conventional measurement of economic activity is infeasible (Gibson, Olivia and Boe-Gibson, 2020). This alternative data is particularly useful in less developed countries, where the institutional capacity for conventional economic data collection tends to have inferior accuracy, latency and geographical resolution.

There are two main sources of nighttime lights data in the public domain (See Table 1). The older system is the DMSP-OLS satellite which generated annual frequency data from 1992 to 2013 at the resolution of a pixel which was roughly $1\text{km} \times 1\text{km}$ at the equator. This was produced by the US Department of Defence and the US NOAA and made available to the public globally. Data from April 2012 onward is available from the US NASA and the US NOAA, which comes from the “Visible Infrared Imaging Radiometer Suite” (VIIRS) on the Suomi-NPP satellite. This dominates the previous dataset by improving the frequency (going up to daily images) and the resolution (a pixel is roughly $0.5\text{km} \times 0.5\text{km}$ at the equator).

Daily images are often obscured by clouds. NOAA performs a certain useful

Table 2 Datasets used in the paper

The starting point for nighttime lights is the monthly aggregated data-set from NOAA which is for 16.7 million pixels within the boundary of India. For each pixel, there are two monthly time series from 2014 to 2019: (1) mean radiance on cloud-free days (measured in $nW\ cm^{-2}\ sr^{-1}$) and (2) the number of cloud-free images which fed into the mean.

Researchers at the World Bank have cleaned the monthly radiance estimator. The raw data from NOAA and aggregated the radiance over the 640 districts of India present in 2011. For each district, the World Bank dataset used in this paper has a monthly time series of radiance from 2014 to 2019.

The income data is from the CMIE Consumer Pyramids Household Survey data-set for 96 ‘homogeneous regions’ (“HRs”) of India that have been defined by CMIE. For each HR, we obtain a 72-month time-series of estimated aggregate income from January 2014 onward.

Dataset	Nighttime lights	Nighttime lights	CPHS
Span	Jan 2014 - Dec 2019	Jan 2014 - Dec 2019	Jan 2014 - Dec 2019
Frequency	Monthly	Monthly	Monthly
Resolution	15-arc second	District	HR
Units	$nW\ cm^{-2}\ sr^{-1}$	$nW\ cm^{-2}\ sr^{-1}$	Rupees
Observations	$(1.67 \times 2 \times 10^7) \times 72$	640×72	96×72
Missing data (%)	2.72	0	1.88
Source	NOAA VIIRS	World Bank	CMIE

set of computations that lead up to the release of a monthly time-series. For each day, NOAA uses the VIIRS Cloud Mask Product to determine whether a given pixel is cloud-free or not. The radiance seen in the cloud-free days of each month is averaged and reported as an estimate of radiance for each month. The number of cloud-free images that were used in a given month is also reported. Most analyses that link to economic activity tend to use this monthly VIIRS dataset.

In this paper, we identify a phenomenon of a downward bias in this dataset that is related to the number of cloudy days. We propose a bias correction scheme which addresses this, and demonstrate how this induces an improved nighttime lights dataset.

3 Data description

Table 2 summarises the three datasets used in this paper. The NOAA nighttime lights dataset used in the paper is a 72-month time-series for each of

1.67×10^7 pixels inside the boundary of India, from January 2014 to December 2019. This raw data has many problems, some of which are documented and corrected by researchers who use the nighttime lights data. In addition to India, we also use nighttime lights data for Dubai, Johannesburg and Singapore.

Most research on India, which uses nighttime lights, uses a dataset at the district level available upon request from the World Bank. Researchers at the World Bank have cleaned the raw data, and aggregated it using the boundaries for the 640 districts of India that were present in 2011, thus creating a compact dataset from April 2012 to December 2019 with a 93-month time-series for each district. We use the World Bank data from January 2014 to December 2019 as the present benchmark for cleaned nighttime lights data.

Centre for Monitoring Indian Economy (CMIE) has a large-scale household panel dataset, the CPHS, which measures about 170,000 households, three times a year, from January 2014 onward. Household income is measured in the survey, alongside many other aspects of the household. CMIE divides India into 102 homogeneous regions (“HRs”) consisting of adjoining districts that have similar agro-climatic conditions, urbanisation levels and female literacy.¹ Since some of the homogenous regions² do not have the span of time-series for the tools used in the paper, this paper uses data for 96 HRs.

For these 96 HRs, we have estimated the aggregate income expressed in real rupees for each month.³ This gives a monthly time-series for aggregate income for each of the 96 HRs of India. We expect that the aggregate radiance and the aggregate income over a given geographical area should be related to, and reflect, the traditional economics interpretation of nighttime lights. Table 3 shows the summary statistics of the estimated aggregate income of HRs.

Table 4 shows summary statistics of the raw nighttime lights data, the radiance as seen in each pixel in the monthly files released by NOAA. Some small negative values are present. The mean and median values range from

¹CMIE does not survey the Andaman and Nicobar islands, Arunachal Pradesh, Dadra and Nagar Haveli, Daman and Diu, Lakshadweep, Manipur, Mizoram, Nagaland and Pakistan Occupied Kashmir.

²The HRs with weaknesses on span are Adilabad-Karimnagar, Bahraich-Gonda, Guntur-Nellore, Meghalaya, Tripura and Sikkim.

³The aggregate income of a homogeneous region is estimated as its median household income $\times$ the number of households in the population. The median is used instead of the mean as it is a more robust estimator, and it avoids the difficulties of household survey data in observing the two tails of the income distribution.

Table 3 Summary statistics of aggregate income of HRs estimated using CPHS

This table shows summary statistics of estimated aggregate income of the 96 HRs. All numbers are expressed in Dec 2019 rupees. As an example, the median income – across the HRs – rose from Rs. 30.28 billion per month in January 2014 to Rs. 45.57 billion per month in December 2019, which is an average annualised real growth rate of 8.50%. The median value of the aggregate household income within an HR is about USD 0.64 billion per month (which is what Rs. 45.57 billion translates into).

(Real rupees, in billion)			
Month	Jan 2014	Dec 2019	
Minimum	5.20	8.57	
25th percentile	18.66	34.67	
Median	30.28	45.57	
Mean	32.47	51.06	
75th percentile	40.78	66.11	
90th percentile	54.98	79.95	
99th percentile	87.83	138.18	
Maximum	100.11	143.71	

Table 4 Summary statistics about radiance and the number of cloud-free images, at the pixel level as seen in the nighttime lights data

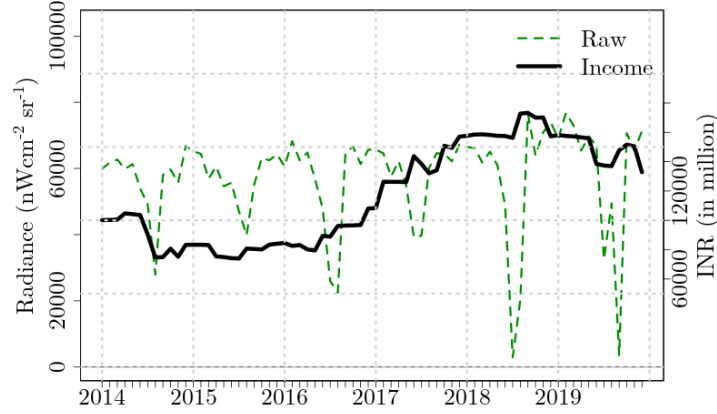
We show summary statistics about nighttime lights radiance (first two columns of data in the Table) for January 2014 and December 2019, each containing 16.7 million pixels. The negative values are a result of cleaning procedures performed by NOAA. The location estimators (mean, median) have values from 0.19 to 0.82.

The last two columns of data in the Table shows the number of cloud-free images for each month that were identified by NOAA, and which fed into the mean value for the radiance for each pixel. NOAA’s cloud identification algorithms are fairly stringent. About 12 to 14 images/month are accepted as being cloud-free for the typical Indian pixel. There are no more than 21 to 22 images cloud-free images, in these two non-monsoon months of January 2014 and December 2019.

$(nW\,cm^{-2}\,sr^{-1})$			Number	
Date	Jan 2014	Dec 2019	Jan 2014	Dec 2019
Minimum	-0.56	-0.52	0.00	0.00
25th percentile	0.06	0.18	11.00	10.00
Median	0.19	0.34	14.00	12.00
Mean	0.60	0.82	13.26	12.19
75th percentile	0.43	0.68	16.00	15.00
90th percentile	1.00	1.38	17.00	17.00
99th percentile	7.76	9.75	18.00	19.00
Maximum	3310.55	1347.49	21.00	22.00

Figure 1 Income juxtaposed with raw nighttime lights, an example: Bombay

The figure is an illustration of the monthly time-series of raw nighttime lights and of the aggregate income, for the city of Bombay. On the y axis, the green dotted line shows the aggregate raw nighttime lights measured in $nW\,cm^{-2}\,sr^{-1}$, and the black line shows the aggregate income measured in million Rupees.



0.19 to 0.82. This raw data is highly skewed. While the 99th percentile has values of 7.76 and 9.75, the maximum values are 3310.55 and 1347.49. This reflects phenomena like gas flares, forest fires, lightning, and meteorite strikes. For example, Zhao et al. (2017) found that extremely high values are nearly always due to gas flaring.

The table also shows summary statistics about the number of cloud-free images which contribute to the computation of the mean radiance. When we look at the location estimates, we find that 13 images per month are classified as being cloud-free. While one may expect that months like April or May could have many cloud-free days in many parts of India, the highest value in any pixel is 21 to 22 cloud-free images per month. This suggests that NOAA’s methods for the identification of cloud-free images are relatively stringent.

Table 4 shows that nighttime lights are concentrated in a small fraction of the well-lit pixels. In terms of land use, the bulk of the Indian sub-continent has farms, water bodies, forests, desert, etc. Human activity is limited to a relatively small amount of land area. Particularly after right tail phenomena like lightning or forest fires are eliminated, the bulk of the aggregate radiance emanates from a small set of pixels. We find that 60% of the pixels add up to 17% of the radiance, and 80% of the pixels account for 30% of the radiance.

Finally, as an example, Figure 1 juxtaposes the aggregate radiance data

against the aggregate income in Bombay.⁴

4 Conventional methods of cleaning nighttime lights data

There are three problems with the standard NOAA–VIIRS release data which are well understood and require the data to be filtered before it can be used in economic analysis:

1. When there are 0 cloud-free images in a month, NOAA codes the radiance value as 0.⁵ These need to be re-coded as ‘not available’.
2. There are outliers with unusual high values that do not reflect high economic activity, they reflect phenomena such as gas flares, forest fires, lightning, and meteorite strikes. These need to be removed.
3. There is a problem of background noise. Places without economic activity often tend to show small positive values of the radiance. This may be because of physical phenomena such as reflection and sensor noise. Ma et al. (2014) found there is conspicuous background noise in VIIRS nighttime lights imagery due to a low detection limit on days with no moonlight. If pixels that have no economic activity carry a background noise value that is as low as $0.2 \text{ nW cm}^{-2} \text{ sr}^{-1}$, this adds up to economically significant values for the aggregate radiance when computed over a large land area.⁶

⁴Bombay corresponds to the districts Mumbai and Mumbai Suburban, and the CMIE defined HR which is Greater Mumbai.

⁵On the download page, NOAA says “*it is imperative that users of these data utilize the cloud-free observations file and not assume a value of zero in the average radiance image means that no lights were observed*”.

⁶As an example, consider the most agriculture-intensive HR in India, Jhabua-Nimar. Jhabua-Nimar consists of 6 districts in Madhya Pradesh i.e. Alirajpur, Barwani, Burhanpur, Dhar, East Nimar (Khandwa), Jhabua, West Nimar (Khargone). In this HR:

- (a) In the month of January 2014, it had an estimated aggregate income of USD 114.3 million, which is about 8.4% of the value seen in Bombay.
- (b) There are 196,928 pixels in this HR. If each of these pixels has a noise value of $0.2 \text{ nW cm}^{-2} \text{ sr}^{-1}$, the aggregate radiance of the HR works out to 39,385.6 $\text{nW cm}^{-2} \text{ sr}^{-1}$. For a comparison, the aggregate radiance of the most prosperous city of India – Bombay – is 63,521 $\text{nW cm}^{-2} \text{ sr}^{-1}$ in January 2014.
- (c) In this example, adding up a large number of small values (of background noise) had a significant impact upon the usefulness of nighttime lights data as an estimator of economic activity.

Table 5 Summary statistics of conventional nighttime lights

This table shows summary statistics of the 8.6 million lit pixels of India in January 2014 and December 2019 after conventional cleaning methods have been applied. The location estimators (mean, median) have values between 0.40 and 1.35. When compared with the summary statistics of the raw nighttime lights in Table 4 we see that the problem of extremely large maximal values has been addressed.

	$(nW\ cm^{-2}\ sr^{-1})$	
Date	January 2014	December 2019
Minimum	0.00	0.00
25th percentile	0.22	0.43
Median	0.40	0.66
Mean	1.00	1.35
75th percentile	0.80	1.15
90th percentile	1.70	2.19
99th percentile	12.85	15.46
Maximum	100.31	98.43

These problems are well known as unavoidable roadblocks in utilising nighttime lights, and a variety of methods have been used to address them. Zhao et al. (2017) has used exponential smoothing to address missing data and outliers. To address background noise, R. C. Beyer, Chhabra et al. (2018) have suggested that the set of pixels with zero radiance in the 2015 annual image (Elvidge et al., 2017) provided by NOAA be used as a background noise mask⁷. But this solution does not take into account the emergence of new locations of economic activity (e.g. the edges of an urban area). As an example, the annual image for 2016 has 13.5% more lit pixels for India as compared with 2015. Any approach to clean this data has to respond to current information and not use just one fixed annual image from the past.

We have implemented conventional cleaning methods, with some refinements. Our methods are shown in Appendix A and released as open source software.

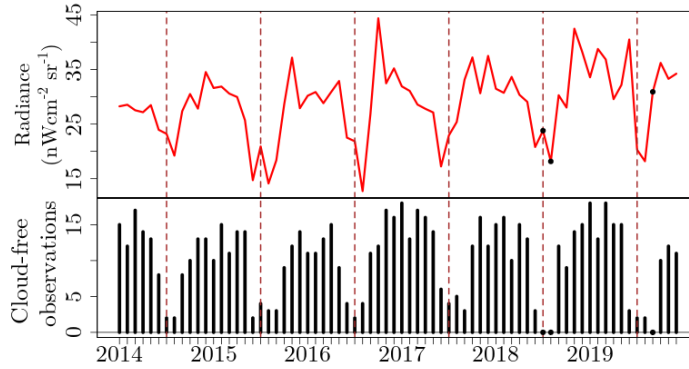
Once the conventional cleaning is done, the downstream analysis is limited to 8.6 million lit pixels. Here also, economic activity is concentrated in a small set of pixels: the 60% dimmest pixels account for 21% of the total radiance. In other words, 20% of the land area of India contains 79% of the total radiance.

Table 5 shows summary statistics of the radiance at the pixel level for the set

⁷NOAA provides the annual images of 2015 and 2016 in which background noise has been removed.

Figure 2 Seasonal dips in nighttime lights, an example: one pixel in Bombay

The figure is an illustration of seasonal patterns in conventional nighttime lights of one pixel in Bombay. For each month ranging from January 2014 to December 2019, the upper pane of the figure shows conventional nighttime lights and the lower pane shows the number of cloud-free observations. The interpolated radiance have been marked with a black dot. Every year around July, which is marked with a dotted line, there is a dip in both of the time-series.



of 8.6 million lit pixels in the 2019 annual image, after conventional methods are applied to the data. When compared against Table 4, we find that the negative values have been eliminated. The mean values in January 2014 and December 2019 have gone up from 0.60 and 0.82 to 1.00 and 1.35. The maximum values in January 2014 and December 2019 have gone down from 3310.55 and 1347.5 to 100.31 and 98.43.

5 The problem of bias

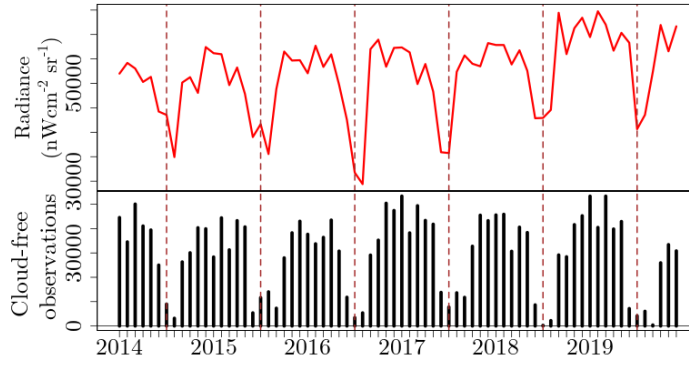
In India, the monsoon runs from June to September. A simple examination of the VIIRS conventional data reveals a peculiar pattern where the radiance, and the number of cloud-free images, is lower in this period.

As an example, Figure 2 shows the (conventional) monthly time-series of radiance for one pixel in Bombay. We see dips in every monsoon. Of particular interest are the points marked with a dot which are an entire month in which there were 0 cloud-free images as defined by NOAA. The vertical lines show July, the month with the most rainfall.

We go from the example of one pixel in Bombay to a larger example: the aggregate radiance in Bombay, in Figure 3. For all pixels in Bombay, we sum

Figure 3 Seasonal dips in nighttime lights, an example: Bombay

In comparison with Figure 2 which shows seasonal dips of conventional nighttime lights in one pixel in Bombay, this figure is an illustration of seasonal patterns in aggregate conventional nighttime lights of one HR, i.e. Bombay. For each month ranging from January 2014 to December 2019, the upper pane of the figure shows the aggregate conventional nighttime lights and the lower pane shows the aggregate number cloud-free observations. Every year around July, which is marked with a dotted line, there is a dip in both of the time-series. The lower pane shows the number of pixels which contribute to the sum for a given month, and there is a dip in each monsoon.



up the number of cloud-free observations of the month (which are showed in the lower pane) and the radiance (which is shown in the upper pane). Here also, we see the pattern, that there are fewer cloud-free observations in the monsoon and the aggregate radiance values are lower.

This reasoning is applied to nighttime lights for all the 96 HRs of India and the kernel density plot of the July dip is reported in Figure 4. We see a small upper tail where there is *more* light in July. For the bulk of the HRs, there is a seasonal dip in July. Thus, the problem that we have identified in the Bombay example – of a seasonal dip during the monsoon – is present in the full country. The modal value of this July dip is at a decline of 28%. The numerical value for the July dip that we see for Bombay (which we have used as an example above) is also the modal value in the overall data.

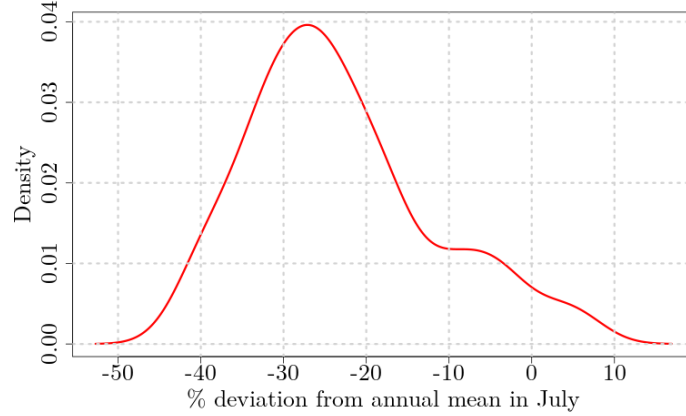
These seasonal fluctuations in the Indian night lights data have also been noticed by other researchers (K Gupta et al., 2017; Hsu et al., 2021).

5.1 Two rival explanations

Why might such dips occur with the monsoon? Possible explanations include:

Figure 4 Large scale evidence of July dip: All HRs

The figure demonstrates that a monsoon dip in aggregate conventional nighttime lights is a large scale phenomenon. For each of 96 HRs, the average seasonal factor of July of aggregate conventional nighttime lights radiance is computed. The estimated kernel density function of this number is plotted in the figure.



A: Seasonality of economic activity One possibility is that economic activity is lower during those months. Then the seasonal decline in radiance is the instantaneous consequence of a seasonal decline in economic activity.

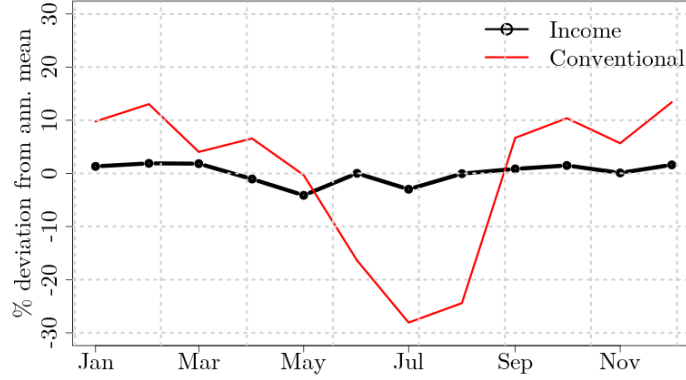
B: A physical phenomenon of signal attenuation The other possibility lies in the physical reality of cloudy days during the monsoon: there may be significant signal attenuation with clouds, even on days that are classified as cloud-free.

In the literature on nighttime lights and economic activity, there is evidence that nighttime lights responds to sustained changes in economic activity and does not respond to high frequency fluctuations. The lights which are on at night are often associated with street lights or commercial establishments, and these respond only slowly to changes in prosperity. Conversely, short-term declines in prosperity do not lead to many of these lights being switched off. As an example, Henderson, Storeygard and Weil (2012) raised the possibility that nighttime lights may not dim as much during economic downturns as they would rise during periods of growth, drawing on evidence from the Rwandan genocide and the Asian Financial Crisis in Indonesia. This runs against Explanation A. Even if economic activity dipped when there was more rainfall, this is unlikely to instantaneously impact upon radiance.

We explore these conjectures in greater detail for the example of Bombay in Figure 5 which superposes the monthly seasonal factors for income in

Figure 5 Economic significance of seasonal patterns, income vs radiance, an example: Bombay

The figure illustrates the disparity between the economic significance of seasonal patterns of aggregate conventional nighttime lights and the seasonal patterns of aggregate income, for Bombay. The average percentage deviation from the mean, owing to seasonality, estimated using X-13-ARIMA-SEATS is shown. On average, nighttime lights in Bombay are 28.06% dimmer in July compared to the annual mean, but there is no such pattern in income.



Bombay against the conventional nighttime lights data.⁸ With income, there is a seasonal dip of about 5% in May and slight positive values in December, January, February and March. We see little to no change in income in the monsoon months of July and August. In the conventional nighttime lights data, there is a seasonal dip of 28.06% in July. The seasonal pattern in conventional nighttime lights deviates a lot from the seasonal pattern in income. Instead, the seasonal factors in June, July and August are more consistent with these months having clouds on account of the monsoon.

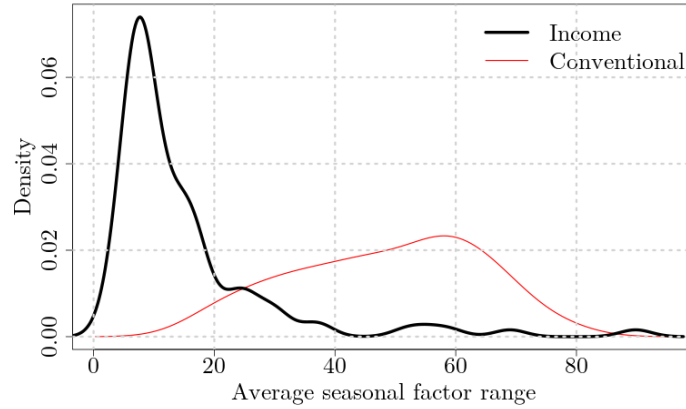
Let $s(x)$ be the 12 average seasonal factors for a time series x estimated using X-13-ARIMA-SEATS. Define $R(x) = \max(s(x)) - \min(s(x))$, the seasonal factor range. In the Bombay example, $R(\text{income})$ is 6.99 and $R(\text{radiance})$ is 41.46: there is little seasonality in income and high seasonality in radiance.

We scale up from the example of Bombay, which is one HR in India, to all 96 HRs in our data. For each region, we compute the seasonal factor range of aggregate nighttime lights and compute the seasonal factor range of

⁸The seasonal factors in a time series are estimated using the `seasonal` package in R (Sax and Eddelbuettel, 2018), which is an interface to the seasonal adjustment software, X-13-ARIMA-SEATS (U.S. Census Bureau, 2013). The constant 100 has been subtracted from each value to make the base 0 instead of 100.

Figure 6 How different are the seasonal factors of aggregate income vs. nighttime lights

The figure demonstrates there is high seasonality in aggregate conventional nighttime lights while there is low seasonality of aggregate income. For each of 96 HRs, the average seasonal factor range of aggregate conventional nighttime lights radiance is computed. The estimated kernel density function of the average seasonal factor ranges is plotted in the figure. The estimated kernel density function of the average seasonal factor ranges of aggregate income of 96 HRs is juxtaposed in the figure.



aggregate income of 96 HRs. The kernel density plots of these are is shown in Figure 6. The problem of a low seasonal factor range with aggregate income data, and a high seasonal factor range with conventional aggregate radiance data, is present in the entire country. The median value of the seasonal factor range for income is 9.34 percentage points, while for the conventional nighttime lights series it is 50.10 percentage points.

Even if there was a seasonal decline of income during the monsoon months, we know that nighttime lights do not rapidly respond to economic fluctuations, and it would be surprising there was a seasonal decline in nighttime lights during the monsoon months as well. In the event, we see that there is no seasonal decline in income during the monsoon. This evidence supports Explanation B over Explanation A: that there is a *bias* in radiance measured during the monsoon months. On the days which are considered cloud-free, there are light clouds which are attenuating the signal picked up from the ground.

5.2 Implications

More cloudy months should not have less economic activity, and even if there was a dip in economic activity in a cloudy month, this should not generate a dip in nighttime lights in that month. This evidence on correlations favours a physical explanation. In months where most days are cloudy, even on the days which the VIIRS data product considers to be cloud-free, there appear to be mild clouds that attenuate the signal, which generate a downward bias in the radiance that reaches that sensor in the satellite.

Some researchers have seen this problem and approached it through the strategy of seasonal adjustment of nighttime lights data before economic analysis (Chodorow-Reich et al., 2020). This procedure is, however, problematic from many points of view. The objective of measurement should be to accurately observe nighttime lights, and if there is a problem with bias, this demands a commensurate bias correction scheme. Imposing one particular theoretical prior – that nighttime lights should have no seasonality – is not desirable. In agricultural areas, income and festivals take place at harvest time, therefore income and nighttime lights *have* seasonal fluctuations, which would be mistakenly removed through seasonal adjustment.

If there is a physical phenomenon of a downward bias when there is more cloud cover in the month, this would arise with nonseasonal rainfall and cloud cover. Sometimes, there is rain in December or in May, even though these are not monsoon months. If this unusual cloud cover generates bias in the radiance, this needs to be corrected; mere seasonal adjustment would not suffice.

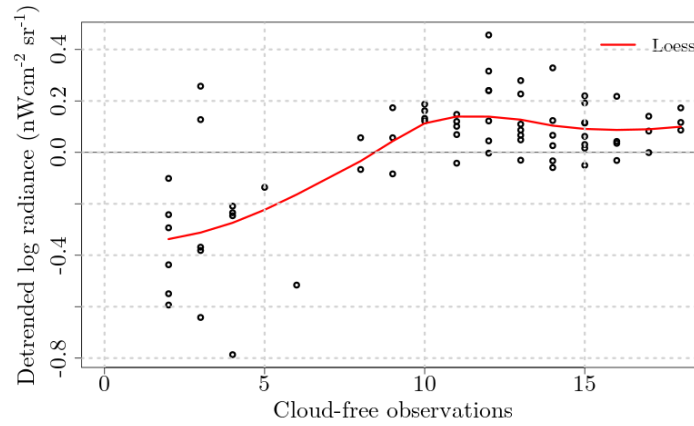
5.3 Four classes of pixels

Figure 7 shows an example for one pixel, with the scatter plot of the number of cloud-free images (on the x axis) vs. the de-trended log radiance (on the y axis). A LOESS curve is superposed, which shows a non-parametric relationship between the two. We see that the number of cloud-free observations and de-trended log radiance are correlated. In months with about 0–5 cloud-free observations, de-trended log radiance is suppressed by about 35% when compared with months which have 10 or more cloud-free images. This is consistent with the seasonal factor of -28.06% that was seen for the aggregate radiance in Bombay for July which is the peak month of the monsoon where there are very few cloud-free images.

In Figure 7, the dataset for the pixel contains a wide variety of values for n_{it} , the number of cloud free observations. There are months at $n_{it} = 3$ and

Figure 7 Nighttime lights vs cloud-free observation, an example: one pixel in Bombay

The figure is an illustration demonstrating the correlation between radiance and the number of cloud-free observations for one pixel. The time trend from the log radiance has been removed and then plotted against the number of cloud-free observations. The locally estimated scatter-plot smoothing function (LOESS) has been drawn to highlight the increasing trend. The LOESS conditional mean is about -0.29 when there are 0–5 cloud-free observation, and this value is 0.14 when there than 11 cloud-free observations. In this pixel, then, there is a bias of about -35% in going from about 11 cloud-free observations in the month to 0–5 cloud-free observations.



there are months at $n_{it} > 15$. The problem of bias, in this pixel, seems to be present in months where $n_{it} < 10$. Under these conditions, our proposed bias correction scheme works well.

Two extreme cases can be envisioned which are not as tidy: a region where most months have high values for the number of cloud-free observations, and a region where most months have low values of cloud-free observations. In the former case (high number of cloud-free observations), there is no significant problem of bias. Such pixels will not have a strong correlation between n and λ_{it} and no bias correction will be done.

The difficulty lies in the other extreme case: a pixel which is mostly cloudy. If only low values of n_{it} are observed, the bias is not identified. Sometimes, there will be a weak correlation between n_{it} and λ_{it} and no bias correction will be done. Even when the null of a zero correlation is rejected, the magnitudes estimated for the bias correction will be small, as there are few sunny months that can reveal the true radiance of the pixel.

We can then usefully classify pixels into four categories:

- Case A** If a region has very low economic activity, the radiance would always be near zero and there should be little to no correlation between radiance and number. These pixels are not important for measuring economic activity. They are likely to be eliminated by conventional cleaning.
- Case B** If a region always has very low clouds, then radiance is observed in all months, and there would be little or no correlation between radiance and number. There is no significant bias that requires correction.
- Case C** If a region is always cloudy, the radiance would be biased downwards in all months, there would be little to no correlation between radiance and number, the magnitude of the bias is not identified.
- Case D** If a region has some bright months and some dark months, there would be some positive correlation between the radiance and the number of cloud-free months, and the bias can be estimated well.

This classification helps to sort and identify the pixels that are vulnerable to the bias discussed in the previous section. As an example, consider the one pixel in Bombay that has been shown earlier. This falls under Case D: it is a pixel where there are both bright and cloudy months. But when faced with a Case C pixel, the bias is unidentified.

We recognise this limitation of our work and take a conservative stance in how it ought to be applied to the data. We will now construct a bias-correction

scheme for the situations which are highly likely to fall under Case D: the pixels where there is strong evidence of bias and where the magnitude of the bias can be relatively precisely estimated. All other pixels are left intact. The methods of this paper improve upon conventionally cleaned data, but do not entirely solve the problem of signal attenuation in cloudy months.

6 A partial bias correction scheme

Figure 7 shows the presence of bias – in this one example of a pixel in Bombay – and immediately suggests a bias correction scheme. The bias is visible in the graph and should be added back into the months with few cloud-free images. We now formally set out a bias correction scheme and measure the extent to which it is able to address the mismatch between the seasonality of aggregate income vs. the seasonality of nighttime lights.

The algorithm operates upon the monthly time series for the pixel. We denote the pixel as i , the month as t , the conventional radiance as x_{it} , and the number of cloud-free observations as n_{it} . The estimated LOESS relationship between the vectors n and x is $\text{LOESS}(n, x)$. The algorithm goes from the inputs x and n to the new radiance vector y which is also y_{it} , a radiance value for pixel i at month t . We use the notation $x' = \log(x)$.

Every scheme that manipulates the incoming data runs the risk of introducing errors into it. We adopt a conservative stance. The Spearman rank correlation between n and de-trended x is $\rho(n, x)$ and we focus on the p -value of the one-tailed test of $H_0 : \rho(n, x) = 0$. This p -value is denoted as $p(n, x)$. For a pixel where H_0 is not rejected by a one-tailed test at a 95% level of significance, no bias correction is done. Further, bias correction is not allowed to generate values bigger than the maximal value in the pixel.

The proposed scheme is:

1. We convert the integer-valued number of cloud free images, n , into a real number from 0 to 1:

$$\lambda_{it} = n_{it}/\max(n_i)$$

2. In the long term, nighttime lights should go up alongside sustained economic growth. To reflect this phenomenon, a linear time trend of the log radiance is estimated:

$$x'_{it} = \alpha + \beta t + \epsilon_{it}$$

using which the residuals ϵ_{it} are the de-trended log radiance:

$$\epsilon_{it} = x'_{it} - \alpha - \beta t$$

3. The LOESS relationship between λ_{it} & ϵ_{it} is estimated:

$$f(\lambda_{it}, \epsilon_{it}) = \text{LOESS}(\lambda_{it}, \epsilon_{it})$$

4. Bias correction is done, only if the rank correlation between ϵ_{it} and n is significantly different from 0, using a one-tailed test:

$$\epsilon'_{it} = \begin{cases} \epsilon_{it} + f(1, \epsilon_{it}) - f(\lambda_{it}, \epsilon_{it}) & : p_i < 0.05 \\ \epsilon_{it} & : p_i \geq 0.05 \end{cases}$$

5. The time trend is added back, and we go from logs to levels:

$$y'_{it} = \exp(\alpha + \beta t + \epsilon'_{it})$$

6. The result is used only if it's below the maximum radiance of the pixel

$$y_{it} = \begin{cases} y'_{it} & : y'_{it} \leq \max(x_{it}) \\ x_{it} & : y'_{it} > \max(x_{it}) \end{cases}$$

This algorithm is applied to all lit pixels. Appendix B shows the summary statistics of the input, the conventional nighttime lights data, versus the new nighttime lights dataset.

7 The bias-correction scheme in action

In this section, we apply the bias-correction method to nighttime lights data. As examples, we will work with four cities – Bombay, Dubai, Johannesburg and Singapore. And, for large-scale evidence, we will work with a large country, India.

Table 6 shows the summary statistics for the number of cloud-free images observed for these five different regions. This helps classify these regions as Cases B, C and D. Since conventional cleaning only picks out lit pixels, there are no Case A regions.

Bombay is Class D and has 25th percentile and 75th percentile values of 8 days and 15 days in a month respectively. This is the case for Johannesburg also. But Dubai is Class B with the 25th percentile and 75th percentile value are 13 and 17 days respectively. Hence, we get only 3.4% of the pixels where the null of a 0 rank correlation is rejected. Singapore is Class C with considerable cloud cover. The 25th percentile is 3 days and the 75th percentile is 7 days, which is why only 36.94% of the pixels feature a rejection of the null.

Table 6 Summary statistics about the number of cloud-free observations in regions used as examples

In this table, we present (a) Summary statistics about the number of cloud-free images and (b) The share of pixels where de-trended radiance is correlated with the number of cloud-free images in the sense of a rejection of $H_0 : \rho = 0$ for the spearman rank correlation coefficient against a one-tailed alternative in order to classify the regions as described in Section 5.3.

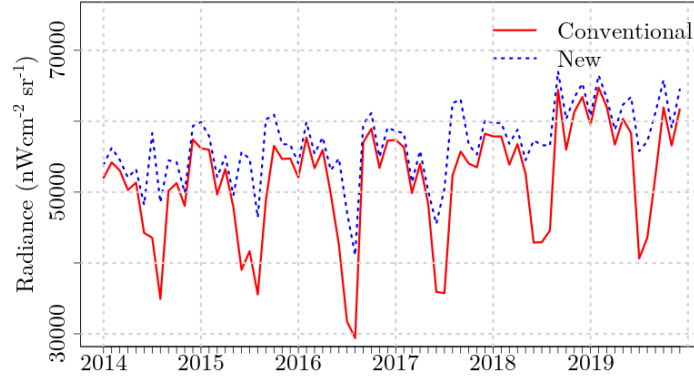
Dubai is classified as Case B: even the 25th percentile month has 13 cloud-free days. Singapore is classified as Case C: the 75th percentile month has just 7 cloud-free days. Bombay and Johannesburg are classified as Case D which have both bright and dark months. For example, the 25th percentile month for Johannesburg has 8 cloud-free days but the 75th percentile month has 15 cloud-free days.

India is made of many pixels, and these overall summary statistics can be consistent with different regions falling into different cases.

	Bombay	India	Johannesburg	Dubai	Singapore
Characteristics of number of cloud-free images					
Minimum	1	1	1	1	1
25th percentile	8	5	8	13	3
Median	12	10	12	16	5
Mean	10.83	9.53	11.65	15	5.37
75th percentile	15	14	15	17	7
90th percentile	16	16	17	19	9
99th percentile	18	19	19	21	12
Maximum	20	23	20	22	15
Pixels with positive correlation					
Share	54.96	59.00	85.65	3.40	36.94

Figure 8 New nighttime lights for a Case D example: Bombay

In comparison with Figure 1 which shows the raw nighttime lights, this figure juxtaposes the aggregate new nighttime lights of one HR, i.e. Bombay. Conventional and new nighttime lights of each pixel inside the region have been added for each month and for both procedures the outcomes have been juxtaposed. The red solid line shows the time series of conventional nighttime lights and the blue dotted line shows the time series of new nighttime lights.



7.1 A Case D example: Bombay

Figure 8 juxtaposes the aggregate radiance seen in Bombay through conventional data and through the proposed bias correction scheme. This shows that the seasonal dips that were previously observed in the conventional data are now lower in the new nighttime lights data.

Figure 9 juxtaposes the seasonal factors of aggregate household income in Bombay against three competing nighttime lights estimators: (a) Conventional data; (b) The cleaned data released by the World Bank and (c) new nighttime lights. The black line has the seasonal factors of aggregate income. In Bombay, there is low seasonality of aggregate income: the maximum and minimum values of the seasonal factors are 1.88% and -4.12%. As had been shown in Figure 5, the red (solid) line, which has seasonal factors for the conventional data, diverges from this seasonality substantially, with a bias of as much as -28.06% for the month of July.

The conventional data released by the World Bank, which is represented by the purple dotted line, has a seasonal factor of -17.59% for July. Spurious seasonality is present in this data also.

The blue dotted line shows the seasonal factors associated with the proposed bias correction scheme. The highest value of the seasonal factor is 5.56%

Figure 9 Improvement in seasonality using the new nighttime lights for a Case D example: Bombay

In comparison with figure 5 which shows the seasonal patterns of conventional nighttime lights and aggregate income, this figure illustrates that the seasonal patterns of the new nighttime lights are closer to the ground truth than that of conventional nighttime lights and World Bank estimate for one HR i.e Bombay. The average deviation from the mean due to seasonal patterns are plotted in the figure for each month. The red solid line shows the seasonal patterns of aggregate conventional nighttime lights, the thin purple shows line the seasonal patterns of World Bank estimates of the aggregate radiance of the region, the black line shows the seasonal patterns of the aggregate income of the Bombay and the thick blue dotted line shows the seasonal patterns of the aggregate new nighttime lights.

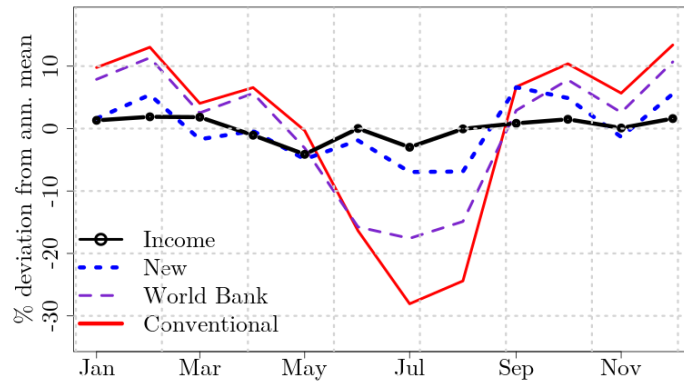
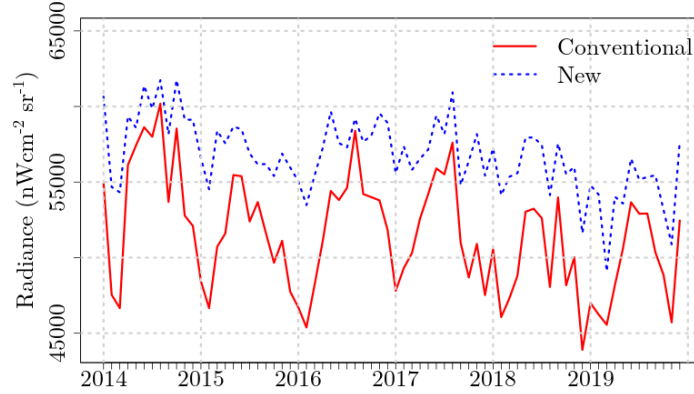


Figure 10 New nighttime lights for a Case D example: Johannesburg

This figure juxtaposes the aggregate new nighttime lights with conventional nighttime lights of Johannesburg. Conventional and new nighttime lights of each pixel inside the region have been added for each month and for both procedures the outcomes have been juxtaposed. The y-axis in the figure measures radiance in $nW\,cm^{-2}\,sr^{-1}$ versus months on the x-axis ranging from January 2014 to December 2019. The blue dotted shows new nighttime lights and red solid line shows the conventional nighttime lights.



and the lowest value of the seasonal factor is -6.95%. In this example, we see that the proposed scheme yields seasonal factors which hew closer to the seasonality of aggregate income.

7.2 A Case D example: Johannesburg

Figure 10 shows the proposed bias correction scheme as applied to Johannesburg. The extreme dips are successfully removed through the bias correction.

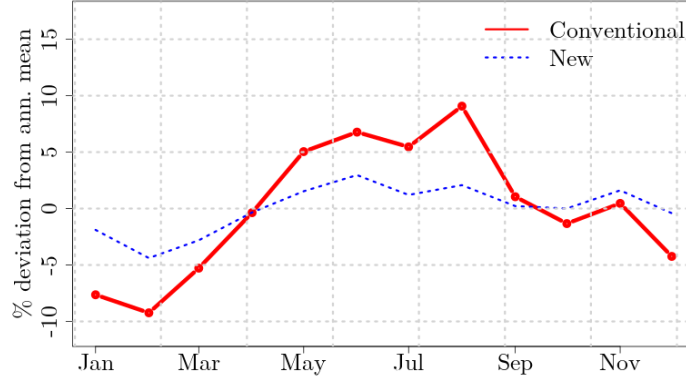
Figure 11 shows the average seasonal factors of conventional nighttime lights vs. new nighttime lights for Johannesburg. In the conventional data, January and February have big dips, and there is much more radiance from May to August. These seasonal factors are much more muted once the bias-correction method is applied to the data.

7.3 A Case B example: Dubai

In Section 5.3, we had suggested that the bias correction methodology should have little impact upon the values of the radiance for a region where most months are sunny. In Table 6, only 3.4% of the pixels in Dubai had a rejection of the null hypothesis of no-correlation, and the 25th percentile value for n_{it} is

Figure 11 Improvement in seasonality using the new nighttime lights for a Case D example: Johannesburg

In comparison with figure 13 which shows the aggregate conventional nighttime lights and new nighttime lights of Johannesburg, this figure shows the average seasonal factors of those series. The average deviation from the mean due to seasonal patterns, estimated using X-13-ARIMA-SEATS, are plotted in the figure for each month. The average seasonal factor range comes down from 18.31 for conventional to 7.35 for new nighttime lights.



13. There is little bias, and little opportunity to identify the bias correction. Hence, the algorithm would leave most pixels intact.

Figure 12 shows the proposed bias correction scheme as applied to Dubai. As expected, the bias correction makes a small difference.

7.4 A Case C example: Singapore

In Table 6, the 90th percentile value for n_{it} in Singapore is 9 days. There are very few months which have low clouds. Hence, there is little opportunity to identify the bias correction. Our algorithm is then likely to do little in Singapore.

Figure 13 shows the proposed bias correction scheme as applied to Singapore. There are relatively small changes in the radiance, reflecting the persistent high level of cloud cover.

Figure 14 shows the average seasonal factors of conventional nighttime lights vs. new nighttime lights for Singapore. As the city is persistently cloudy through the year, there is little seasonality in the conventional nighttime lights radiance. And, as the bias correction strategy of this paper fares poorly for persistently cloudy places, the seasonal factors are not significantly

Figure 12 New nighttime lights for a Case B example 4: Dubai

This figure juxtaposes the aggregate new nighttime lights with conventional nighttime lights of Dubai. Conventional and new nighttime lights of each pixel inside the region have been added for each month and for both procedures the outcomes have been juxtaposed. The y-axis in the figure measures radiance in $nW\,cm^{-2}\,sr^{-1}$ versus months on the x-axis ranging from January 2014 to December 2019. The blue dotted shows new nighttime lights and red solid line shows the conventional nighttime lights.

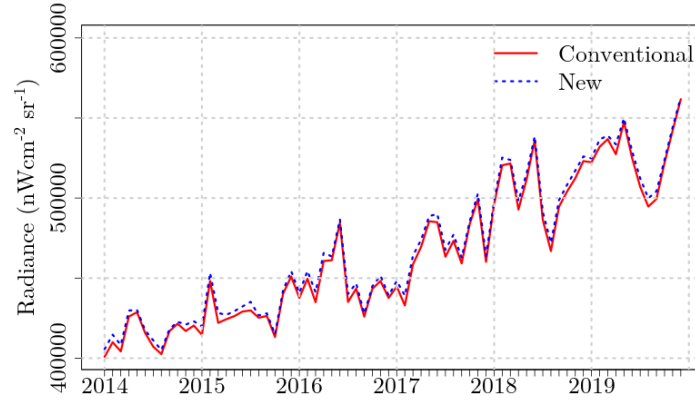


Figure 13 New nighttime lights for a Case C example: Singapore

This figure juxtaposes the aggregate new nighttime lights with conventional nighttime lights of Singapore. Conventional and new nighttime lights of each pixel inside the region have been added for each month and for both procedures the outcomes have been juxtaposed. The y-axis in the figure measures radiance in $nW\,cm^{-2}\,sr^{-1}$ versus months on the x-axis ranging from January 2014 to December 2019. The blue dotted shows new nighttime lights and red solid line shows the conventional nighttime lights.

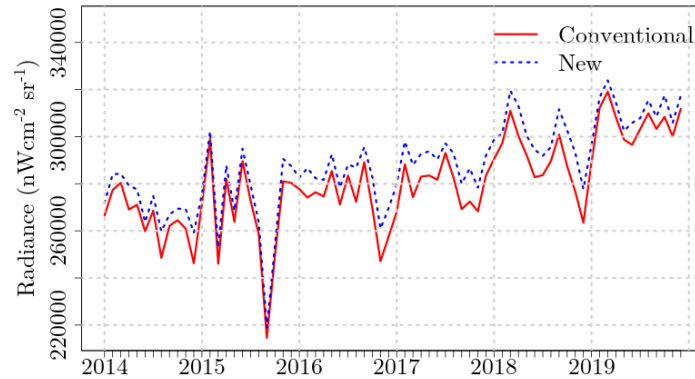
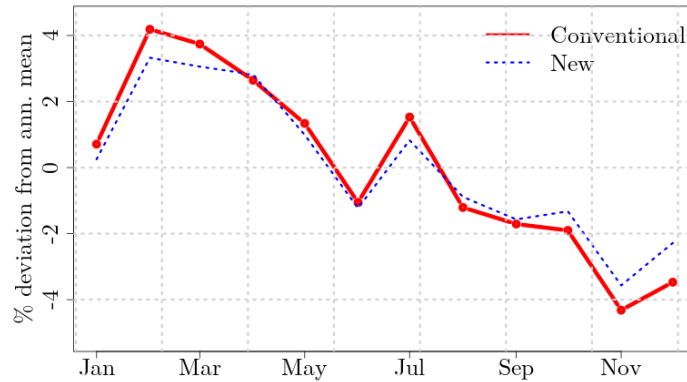


Figure 14 Improvement in seasonality using the new nighttime lights for a Case C example: Singapore

In comparison with figure 13 which shows the aggregate conventional nighttime lights and new nighttime lights of Singapore, this figure shows the average seasonal factors of those series. The average deviation from the mean due to seasonal patterns, estimated using X-13-ARIMA-SEATS, are plotted in the figure for each month.

The average seasonal factor range comes down from 8.50 for conventional to 6.91 for new nighttime lights.



modified.

7.5 Large-scale evidence for a country: India

We now scale up from examples to large-scale evidence for India.

Figure 15 shows the proposed bias correction scheme as applied to India. As expected, the large seasonal fluctuations seen with the conventionally cleaned data are significantly reduced in the bias-corrected series.

Many researchers have worked on Indian questions using the World Bank data. Figure 16 finds that the World Bank data seems to lack a time trend over the period from 2014 to 2020, a period over which aggregate income (the black line) grew significantly. The bias-corrected radiance shows some increase over this period.

Figure 17 shows the average seasonal factors of conventional nighttime lights vs. new nighttime lights for India. Conventionally cleaned and World Bank radiance estimates show a seasonal dip of about 20% during the monsoon. The new bias-corrected data shows a low seasonal factor of about -10% for these months, and is closer to the seasonal factor of income of about -4%.

Figure 15 New nighttime lights for India

This figure juxtaposes the aggregate new nighttime lights with conventional nighttime lights of India. Conventional and new nighttime lights of each pixel inside the region have been added for each month and for both procedures the outcomes have been juxtaposed. The y-axis in the figure measures radiance in $nW\,cm^{-2}\,sr^{-1}$ versus months on the x-axis ranging from January 2014 to December 2019. The blue dotted shows new nighttime lights and red solid line shows the conventional nighttime lights. The large seasonal fluctuations in the red solid are much reduced in the new nighttime lights aggregate radiance.

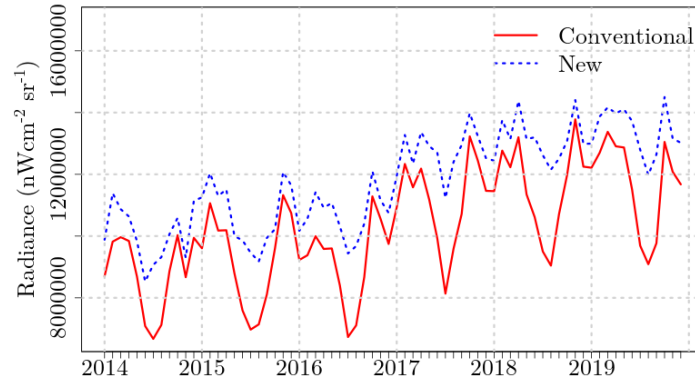


Figure 16 New nighttime lights juxtaposed with World Bank's estimate for India

In comparison with Figure 15 which juxtaposes new and conventional nighttime lights of India, this figure juxtaposes the aggregate new nighttime lights with aggregate income and the World Bank's estimate of the aggregate radiance of India. The blue dotted line shows the time series of new nighttime lights, while the purple dotted line shows the World Bank's estimate, and the black solid line shows income.

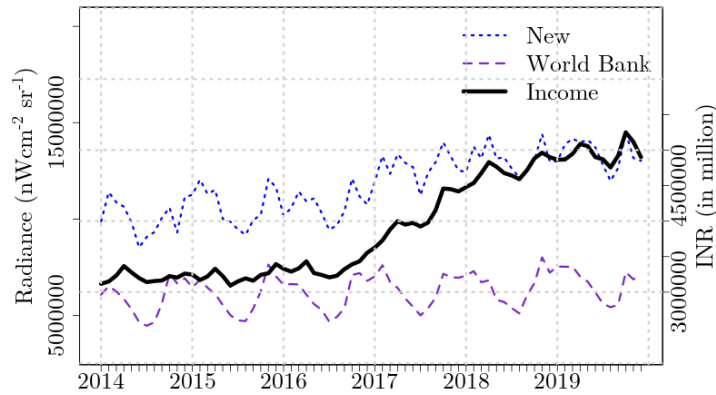


Figure 17 Improvement in seasonality using the new nighttime lights for India

This figure shows that the seasonal patterns of the new nighttime lights are closer to the ground truth than that of conventional nighttime lights and World Bank estimate for India. The average deviation from the mean due to seasonal patterns are plotted in the figure for each month. The red solid line shows the seasonal patterns of aggregate conventional nighttime lights, the thin purple shows line the seasonal patterns of World Bank estimates of the aggregate radiance of the region, the black line shows the seasonal patterns of the aggregate income of India and the thick blue dotted line shows the seasonal patterns of the aggregate new nighttime lights.

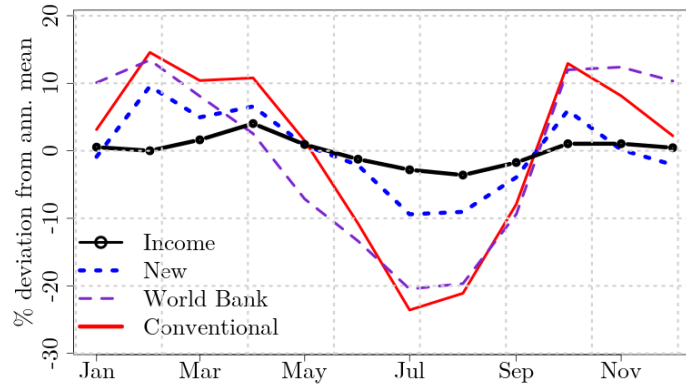
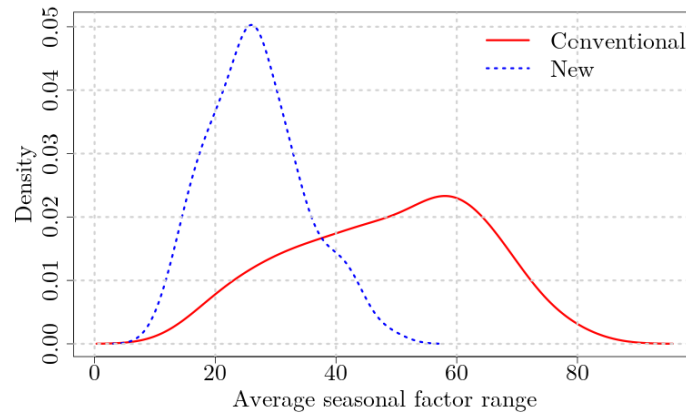


Figure 18 Kernel density plots of average seasonal factor ranges

This figure juxtaposes the estimated kernel density function of the average seasonal factor ranges of new nighttime lights with the estimated kernel density function of the average seasonal ranges of conventional nighttime lights shown in figure 6. For each HR, the average seasonal factor ranges of aggregate radiance are computed for conventional nighttime lights and new nighttime lights. The estimated kernel density functions are plotted in the figure.



For each of the 96 HRs, we obtain the average seasonal factor range of conventional nighttime lights and for nighttime lights data based on the proposed bias correction scheme. Figure 18 juxtaposes the kernel density plot of the conventional nighttime lights (red solid) and the kernel density plot of the new nighttime lights (blue dotted). The figure shows a substantial leftward shift of the distribution; the proposed scheme achieves a lower average seasonal factor ranges. The gains are economically significant. The highest decrease in average seasonal factor range after bias correction is in Bhabua-Patna, of 38.61 percentage points. There are 3 HRs – Kozhikode-Malappuram, Madurai-Dindigul and Coimbatore-Dharmapuri – where the average seasonal factor range of new nighttime lights exceeds that of the conventional nighttime lights data by 3.07, 1.00 and 0.38 percentage points respectively.

Appendix C shows a comparison of the two cumulative distributions. The KS test rejects the null of the two distributions being the same with a p value of 0.2×10^{-15} .

Our proposed bias correction scheme largely removes the rank correlation between (de-trended) radiance and the number of cloud-free images. In the conventional nighttime lights data, this rank correlation was statistically significant for 59.00% of the pixels; after bias-correction, this is present with 5.10% of the pixels.

Table 7 compares the summary statistics, across 96 HRs, of seasonal factor ranges of income and the two radiance estimators. There are gains in all parts of the distribution. The median value of the seasonal factor range in income is 9.35 percentage points. Conventional data has a large median value of the seasonal factor range of 50.10 percentage points. Once bias correction is done, this median value comes down to 26.14 percentage points. While this is a significant decline, it remains higher than the median value of the seasonal factor range of income.

8 Reproducible research

Many researchers use the World Bank nighttime lights data when analysing economic problems (R. Beyer, Jain and Sinha, 2020; R. C. Beyer, Franco-Bedoya and Galdo, 2020; Prakash et al., 2019; Roberts, 2021; Galdo, Lopez-Acevedo and Rama, 2020). While the dataset is generously made available upon request, documentation of the methods used and the underlying source code are not released, which can be a barrier to new research work. For

Table 7 Summary statistics about average seasonal factor ranges of new nighttime lights

For each HR, the seasonal factors are estimated using X-13-ARIMA-SEATS and the average seasonal factor range i.e. the average of the highest - lowest for each year is computed. This table juxtaposes the summary statistics about the average seasonal factor ranges of aggregate new nighttime lights for the 96 HRs with the average seasonal factor ranges of aggregate conventional nighttime lights and the average seasonal factor ranges of aggregate income of 96 HRs. The summary statistics of these seasonal ranges are shown in the table. The location estimators (mean, median) of the average seasonal factor ranges of conventional nighttime lights have values 48.55 and 50.10 while the average seasonal factor ranges of new nighttime lights are 26.78 and 26.14. The location estimators of the average seasonal factor ranges of new nighttime lights are closer to those of the average seasonal factor ranges of aggregate income.

(Percentage points)			
	Aggregate Income	Aggregate nighttime lights	
		Conventional	New
Minimum	3.20	17.05	11.65
25th percentile	7.05	37.03	20.95
Median	9.35	50.10	26.14
Mean	15.17	48.55	26.78
75th percentile	16.86	60.44	31.52
Maximum	89.81	79.07	48.95

methodological research, it is desirable to have a foundation of documented methods and software in the public domain.

Hence, we have built open source software, which implements a coherent and unified approach that addresses these basic questions. The details of the steps that we follow are placed in Appendix A. Both stages of the work of this paper – conventional cleaning and the bias correction scheme – are released as open-source software.⁹ Alongside this, datasets are released through which all the calculations for Bombay (Figure 1, Figure 2, Figure 3, Figure 5, Figure 7, Figure 8, Figure 9) can be replicated by the public. While the package implements the bias correction procedure of this paper, it is also useful as the first open-source implementation of conventional cleaning.

9 Conclusions

Nighttime lights is a novel and important path to estimating economic activity. Particularly after the launch of the new VIIRS data – with daily/monthly release of images and pixels that are about $0.5 \text{ km} \times 0.5 \text{ km}$ – there are many opportunities for important applications in research and in real-world applications. In this paper, we demonstrate that this data, after conventional cleaning methods are applied, suffers from bias. The time-series of radiance varies slowly and large fluctuations from one month to the next are not a reflection of the underlying economic activity. In months with just a few cloud-free images, the radiance values are biased downwards. A careful analysis of the phenomenon of the Indian monsoon helps reveal this problem. The rules adopted for defining a cloud-free image, by NASA and NOAA are relatively stringent, and only a few days in the Indian July or August are considered cloud-free. However, it appears that even on these days, there is light cloud cover, which generates a seasonal dip for July of -28.06% for Bombay that is clearly an artefact.

The magnitude of this bias is substantial and would matter to many research projects. As an example, when studying the impact of Covid-19 in India in 2020, there are questions about the recovery of the economy after April and May. However, the nighttime lights observed in June, July and August suffers from the downward bias associated with cloudy days, and the use of conventional data would generate a downward bias in the perceived recovery of the economy (R. Beyer, Jain and Sinha, 2020).

⁹These materials are at https://xkdr.org/releases/PatnaikShahTayalThomas_2021_bias_correction_nighttime_lights.html

We harness the fact that the *number* of cloud-free images is observed to establish a bias correction scheme at the pixel level. This is convenient for researchers as aggregate radiance can then be calculated, using the new nighttime lights pixel-level dataset, for any geographical region of interest. The construction of the new nighttime lights dataset is done in a conservative way, where many pixels are left intact when the rank correlation between the number of cloud-free images and the radiance is not statistically significant.

The limitation of this work lies in the lack of identification of bias for a pixel which is mostly cloudy (what is termed Case C in this paper). In the scatter diagram of de-trended radiance vs. cloud-free images, if most observations are months with a small number of cloud-free images, the bias is not identified. Hence, this paper constitutes a cautious and partial algorithm, which can always be used with confidence that artefacts are not introduced, but it does not completely solve the problem of clouds that get in the way.

References

- [1] Robert Beyer, Tarun Jain and Sonalika Sinha. ‘Lights Out? COVID-19 Containment Policies and Economic Activity’. In: *Policy Research Working Paper, World Bank* (2020).
- [2] Robert CM Beyer, Esha Chhabra et al. ‘Measuring districts’ monthly economic activity from outer space’. In: *Policy Research Working Papers, World Bank* (2018).
- [3] Robert CM Beyer, Sebastian Franco-Bedoya and Virgilio Galdo. ‘Examining the Economic Impact of Covid-19 in India through Daily Electricity Consumption and Nighttime Light Intensity’. In: *Policy Research Working Papers, World Bank* (2020).
- [4] Gabriel Chodorow-Reich et al. ‘Cash and the economy: Evidence from India’s demonetization’. In: *The Quarterly Journal of Economics* 135.1 (2020), pp. 57–103.
- [5] Christopher NH Doll, Jan-Peter Muller and Jeremy G Morley. ‘Mapping regional economic activity from night-time light satellite imagery’. In: *Ecological Economics* 57.1 (2006), pp. 75–92.
- [6] Christopher D Elvidge et al. ‘VIIRS night-time lights’. In: *International Journal of Remote Sensing* 38.21 (2017), pp. 5860–5879.
- [7] Virgilio Galdo, Gladys Lopez-Acevedo and Martin Rama. ‘Conflict and the Composition of Economic Activity in Afghanistan’. In: *Policy Research Working Papers, World Bank* (2020).
- [8] John Gibson, Susan Olivia and Geua Boe-Gibson. ‘Night lights in economics: Sources and uses’. In: *Journal of Economic Surveys* 34.5 (2020), pp. 955–980.
- [9] J Vernon Henderson, Adam Storeygard and David N Weil. ‘Measuring economic growth from outer space’. In: *American economic review* 102.2 (2012), pp. 994–1028.
- [10] Fengchi Hsu et al. ‘The annual cycling of nighttime lights in India’. In: *Remote Sensing* 13 (Mar. 2021). URL: <https://doi.org/10.3390/rs13061199>.
- [11] Rob J Hyndman and Yeasmin Khandakar. ‘Automatic time series forecasting: the forecast package for R’. In: *Journal of Statistical Software* 26.3 (2008), pp. 1–22. URL: <https://www.jstatsoft.org/article/view/v027i03>.
- [12] P K Gupta et al. ‘Analysis of SNPP-VIIRS-DNB derived nightlights over India’. In: *Asian Association of Remote Sensing* (2017).
- [13] Ting Ma et al. ‘Responses of Suomi-NPP VIIRS-derived nighttime lights to socioeconomic activity in China’s cities’. In: *Remote Sensing Letters* 5.2 (2014), pp. 165–174.
- [14] Anupam Prakash et al. ‘Night-time Luminosity: Does it Brighten Understanding of Economic Activity in India?’ In: *RBI Bulletin* (2019).
- [15] Mark Roberts. ‘Tracking Economic Activity in Response to the Covid-19 Crisis Using Nighttime Lights’. In: (2021).
- [16] Christoph Sax and Dirk Eddelbuettel. ‘Seasonal adjustment by x-13arima-seats in R’. In: *Journal of Statistical Software* 87.1 (2018), pp. 1–17.
- [17] Statistical Research Division U.S. Census Bureau. ‘X-13ARIMA-SEATS Reference Manual’. In: (2013).
- [18] Naizhuo Zhao et al. ‘Improving accuracy of economic estimations with VIIRS DNB image products’. In: *International journal of remote sensing* 38.21 (2017), pp. 5899–5918.

Appendices

A Conventional cleaning: detailed methodology

Our open-source implementation of conventional cleaning for the NOAA-VIIRS nighttime lights data consists of the following steps:

1. We partition the universe of pixels into a pair of disjoint sets, one that is considered dark at all times and removed from all analysis, and another that is considered lit. The lit pixels are those which have an average annual radiance that exceeds a fixed threshold.

This requires choosing a background noise threshold. To do this, we construct annual images for 2015 and 2016 as the weighted averages of their months with weights that are inversely proportional to their number of cloud-free images (K Gupta et al., 2017). We compare the similarity of these weighted averages against the cleaned annual images for 2015 and 2016 provided by NOAA. The numerical value for the threshold of background noise is chosen to maximise the similarity with the NOAA annual images. This value works out to $0.4nW\text{ cm}^{-2}\text{ sr}^{-1}$.

We now focus on the *last* year of the dataset: 2019. Economic activity in this year was the highest, when compared with the other years under examination, and we are likely to obtain the largest possible set of pixels with economic activity for this year. We choose the lit pixels for 2019, as those with a radiance of above $0.4nW\text{ cm}^{-2}\text{ sr}^{-1}$ in the 2019 annual image. With this filter, 8.6 million out of 16.7 million pixels over India, which is 51% of the pixels, are considered lit. This set of lit pixels is used for all previous years in this paper, thus ensuring a homogeneous time-series for the radiance of all lit pixels. There is a slight imprecision in this strategy: pixels which were truly dark in previous years, but obtained economic activity in 2019, are utilised in all years, thus contributing to background noise.

2. The problem of NAs and outliers is addressed through two steps:
 - (a) Observations with 0 cloud-free observations, those with negative values, and those considered to be outliers by the `tsoutliers` function from the R ‘forecast’ package (Hyndman and Khandakar, 2008) are replaced with estimates obtained through linear interpolation.
 - (b) Finally, pixels with high standard deviation are rejected. The standard deviation each of the 8.6 million lit pixels of India is computed. The top 0.1% of the pixels by standard deviation are considered to be outliers and removed from all analysis. That is, 1 in 1000 pixels with the highest value of the time-series standard deviation is rejected as an outlier; these pixels are removed from the set of lit pixels.

We illustrate the working of this implementation of the conventional cleaning methods works for two targets: (1) a single pixel, and (2) the pixels in the district of Bombay.

Figure 19 juxtaposes the NOAA-VIIRS data for one pixel in Bombay against the same data filtered using our methods of conventional cleaning. The cleaned data looks significantly better than the raw nighttime lights.

Figure 19 An example of conventional nighttime lights: one pixel in Bombay

The figure is an illustration of conventional cleaning of one pixel which is a ‘lit pixel’ in the 2019 annual image. The green dotted line shows raw nighttime lights provided by NOAA and the red solid line shows the conventional nighttime lights for the pixel. There are NAs which occur four times that are coded as zero in the raw nighttime lights are replaced by interpolated values.



Figure 20 An example of conventional nighttime lights: Bombay

In comparison with figure 19 which shows raw nighttime lights and conventional nighttime lights for one pixel in Bombay, this figure shows an illustration of aggregate conventional cleaning of one HR, i.e. Bombay. Conventional cleaning has been done on each pixel and the aggregate radiance for all pixels inside the region is computed for each month. This result is juxtaposed with the aggregated raw nighttime lights. The green dotted line shows raw nighttime lights and the red solid line shows conventional nighttime lights

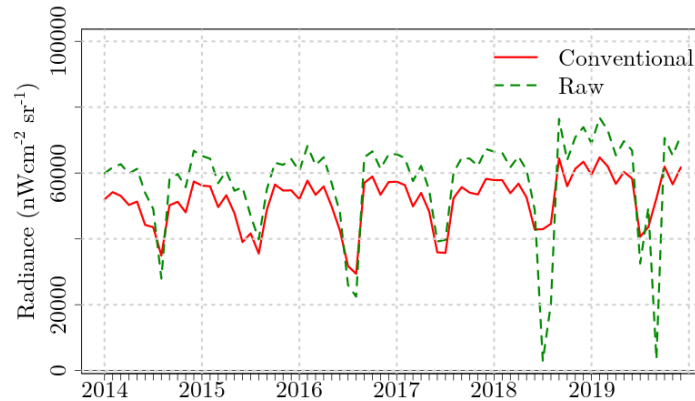


Figure 21 World Bank estimates compared to conventional nighttime lights, an example: Bombay

In comparison with figure 20 which shows raw nighttime lights and conventional nighttime lights for Bombay, this figure shows an illustration of conventional nighttime lights of one HR, i.e. Bombay. Conventional cleaning has been done on each pixel and the aggregate radiance for all pixels inside the region is computed for each month. This result is juxtaposed with the World Bank’s cleaned nighttime lights for Bombay. The red solid line shows conventional nighttime lights and the purple line shows the World Bank estimate for nighttime lights in Bombay.

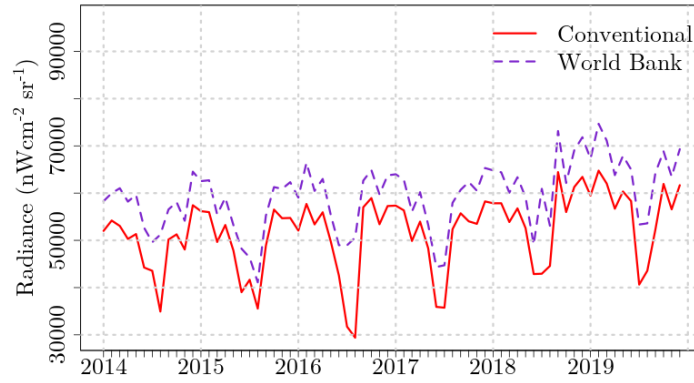


Figure 20 shows raw nighttime lights and conventional nighttime lights for one HR, i.e. Bombay. Here also, we see that the conventional data appears to be more sensible when compared with the raw data.

Finally, we juxtapose our implementation, of conventional cleaning methods, against the cleaned aggregate radiance data released by the World Bank. By and large, these two methods yield similar results.

B Summary statistics about radiance at the pixel level: Conventional vs. new

Table B.1 juxtaposes the summary statistics about 8.6 million lit pixels, for conventional data versus the new data. The location estimators (mean, median) rise slightly from 0.66/1.35 to 0.73/1.51 in December 2019. The key insight is that for 40% of the pixels, there is low economic activity, the absolute level of the radiance is low regardless of the cloud cover, and no bias correction is done as the null of zero rank correlation (between detrended radiance and the number of cloud-free images) cannot be rejected. Bias correction is only attempted for the remaining 60% of the pixels. Further, for many months, there are ample cloud-free images and bias correction has no impact. Hence, the two groups of summary statistics are quite similar.

Table B.1 Summary statistics of conventional nighttime lights and new nighttime lights

In comparison with Table 5 which shows summary statistics of the conventional nighttime lights, in this table, we show summary statistics of both conventional nighttime lights and new nighttime lights for the 8.6 million lit pixels in January 2014 and December 2019.

$(nW\,cm^{-2}\,sr^{-1})$				
Conventionally cleaned			Bias-corrected	
Date	Jan 2014	Dec 2019	Jan 2014	Dec 2019
Minimum	0.00	0.00	0.00	0.00
25th percentile	0.22	0.43	0.24	0.46
Median	0.40	0.66	0.46	0.73
Mean	1.00	1.35	1.15	1.51
75th percentile	0.80	1.15	0.94	1.31
90th percentile	1.70	2.19	1.99	2.50
99th percentile	12.85	15.46	14.68	17.27
Maximum	100.31	98.43	100.31	106.70

C Comparing the cumulative density of seasonal factor range: conventional vs. new

Figure 22 juxtaposes the CDF of the seasonal factor range, across all HRs, for conventional data versus new data. There is a clear gap between the two distributions. Kolmogorov-Smirnov shows the gap between the two distributions is statistically significant with a p -value of 0.2×10^{-15} .

Figure 22 Cumulative density plots of average seasonal factor ranges

In comparison with figure 18 which shows the estimated kernel density functions of the average seasonal factor ranges of aggregate conventional nighttime lights and new nighttime lights, this figure shows the cumulative density plots. For each HR, the average seasonal factor ranges of conventional nighttime lights and new nighttime lights and computed and the cumulative density plots are generated. Kolmogorov-Smirnov test is performed to demonstrate that the average seasonal factor ranges of new nighttime lights are smaller than conventional nighttime lights. The KS statistic is 0.65.

