Matching satellite data to surveillance site data to investigate service delivery

NASA/NOAA

The challenge

Despite significant progress over the last 20 years, basic access to electricity in South Africa is by no means stable or guaranteed, and remains one of the largest development issues faced by post-apartheid South Africa. The reasons for inequality in public service delivery, particularly access to electricity by poorer communities, in rural as well as urban settings, are not as well understood as they could be. Knowing more about domestic energy consumption within households can better inform strategies to address fuel demand and poverty. A richer, more policy-relevant picture can be gained by investigating dynamics of electricity access and consumption using reconstructed, complex, large-scale longitudinal data linked to other sources of big data. In their efforts to scale up research opportunities, data services will need to consider the methodological, technical and ethical challenges of preparing, linking and hosting such large-scale data.

> "This work delivers fantastic insights into the application of Nightlight data, and how it can be used to compliment and validates analyses drawn from HDSS-type databases.

Research example

For developed regions, the use of satellite data in mapping urbanization has been widely tested and validated. Measuring the electrification of rural areas, using the same method, however, is a much newer phenomenon.

In this vein, Machemedze, Dinkelman, Collinson, Twine and Wittenberg implemented a research project focused on using satellite data to investigate trends in rural electricity access and use within North-Eastern South Africa. The Wits/MRC Rural Public Health and Health Transitions Research Unit (Agincourt) project has been monitoring demographic and health changes in the Agincourt area since 1992. Since 2000, information on households' access to and use of electricity has also been recorded. The team therefore aimed to use Nightlight satellite data and match it to this Health and Demographic Surveillance System (HDSS) data to validate and test satellite data's potential for providing reliable insights on rural electrification in South Africa.

The research team focused specifically on exploring whether the satellite data would pick up the temporal patterns of rural electrification in the rural area, to allow for analyses that could be corroborated by the data collected on the ground. Specifically, they focused on: 1) identifying whether the Nightlight data conveyed similar electrification trends to those observed in the HDSS data; 2) identifying whether these trends were an artefact of satellite measurement error or not, and thus whether trends differed across different types of rural and urban sites; and 3) whether the use of on-the-ground insights and data matching methods could overcome any issues of measurement error or periods of missing data in either data source.

This work consequently delivers fantastic insights into the application of Nightlight data, and how it can be used to compliment and validates analyses drawn from HDSS-type databases.

Data and data issues

Agincourt HDSS data overview

The Agincourt Health and Demographic Surveillance System (HDSS) monitors key demographic events and socio-economic variables in the Agincourt sub-district in north-eastern Mpumalanga Province, South Africa. A baseline census was conducted in 1992 with annual census rounds being conducted since. Key variables measured routinely by the HDSS include: births, deaths, in- and out-migrations, household relationships, resident status, refugee status, education, antenatal, and delivery health-seeking practices. Additional modules have also been added at different points in time – such as the collection of a household asset module, which has been implemented every second year since 2000, and includes information on household access to services, such as electricity. 'Temporary migrants' are defined as non-resident members who retain significant contact and links with the rural home and 'share a common pot', and are included on the household grid.

Nightlights data overview

Nighlights data was sourced from the DMSP-OLS Nighttime Lights Time Series, and was matched to the study site, and specific villages, using shape files supplied by the MRC/Wits University Rural Public Health and Health Transitions Research Unit. These data are annual cloud-free composites of average digital brightness values for the detected lights, filtered to remove ephemeral lights (e.g. fires) and background noise. The data come from six satellites and span a period of 21 years. Each pixel within the area represents around a square kilometre on the ground. The data per pixel gives the annual average brightness level, with digital numbers (DN) ranging from 0 to 63: 0 representing the absolute absence of light, and the top value representing saturated light.

Using an intercalibration methodology, data values are converted from individual satellite products into a common range defined by a reference year. Calibration coefficients are then applied, before comparing data over the years. Negative values are treated as zero or no light. After calibration, the 'sum-of-lights' (SOL, the sum of all pixel values for a particular region) was derived, in this case as the sum of the digital values for the Agincourt area for each year and satellite. Where there two readings for a year, the intercalibrated values were averaged to produce one estimate per year

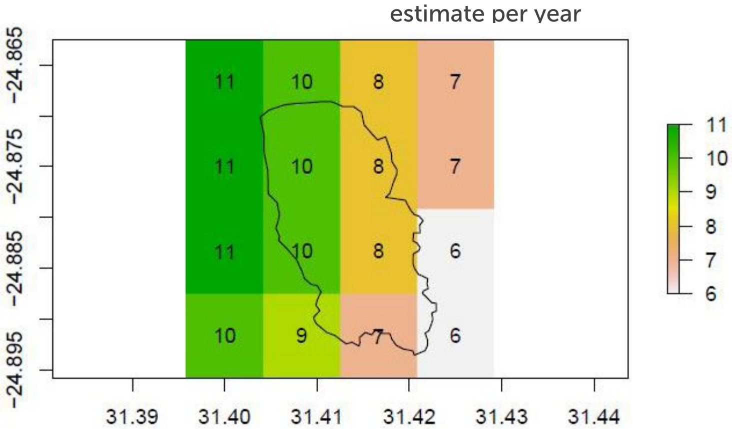

Figure 1. Comparison of irregular village shape to the typical square pixel grid

Issue 1: size of areas

The 'sum of lights' (SOL) is a measure used commonly in developed regions to analyse satellite data on lighting and urbanisation. However, in rural regions, this measure can be problematic. Firstly, villages are all different sizes which means the SOL measure reflects not only the level of electrification, but also the size of the village (e.g. a village with higher proportional electrification may have a lower SOL measure than a larger, less electrified counterpart).

Another immediate issue is that our villages are irregularly shaped and do not fit neatly into the square pixel grid of the satellite data for the SOL calculations, as shown in Figure 1.

Issue 2: measurement errors

There is considerable measurement error at the scale that the researchers are interested in, i.e. the village level. Some of the light that emanates from a village will be lost to adjoining areas. In fact, the measurement error is worse than this, given how the original data are constructed. While the Nightlight data is accessible at a nominal resolution of a square kilometre, the pre-averaging in the data preparation means the actual resolution is not as accurate as that.

Issue 3: limits in detecting light sources

Another drawback of using satellite data to measure electrification is that it only detects external light. Even when one measures a decline or stagnation in the SOL measure (a fall in lighting) for a given area, it could very well have been that electrification was still occurring in the area, but, people/households were opting to use internal lights – so their households were not detected.

Other issues

Confidentiality and ethics

The confidentiality, legal and ethical issues that come from access to administrative data, local level census data and longitudinal data from Health and Demographic Surveillance Site (HDSS) data are a concern with these data sources.

Shape and size of data

The size and complexity of the data are a concern. The nightlight data are of a size that social scientists in South Africa are not used to analysing, and the added challenges that come with combining data from different sources raise various analysis issues.

Quality of data

Data quality is also a concern. Missing data, data errors and uncertain or unknown provenance need to be identified and dealt with. Spurious results emerge from the Nightlight data, which need to be checked on the ground to iron out mistaken bright spots.

The methodological solution

The Nightlight data, along with the matching of this data to survey/census data, has substantial potential to provide new insights and details on the dynamics of electricity consumption in rural areas. In the Agincourt area specifically, the authors validated trends in electricity consumption/use across the two big data sources (the HDSS and Nightlight data), and matched the data to identify the impact of new connections on village/area brightness. They have therefore explored the reliability of satellite data in a local context where they had access to substantial ancillary information. On the whole, they conclude, the nightlight data seems to have captured the electricity roll-out in the Agincourt study site: showing marked increases in brightness over time and capturing the broad differences between "developed" and "undeveloped" parts of the site.

While using Nightlight satellite surveillance data to analyse electricity use trends in a rural location such as Agincourt presented certain measurement issues which contaminated the relationships and trends measured, these issues could be circumvented by using novel technical methods and on-the-ground insights, and by drawing on additional information sourced from the HDSS data.

To address the shortcomings of an SOL measure in the rural context, Machemedze calculated the proportion of the village that fits into a pixel and added that proportion of its light measure to the SOL score for the village/area. While this approach resulted in some loss of brightness, it gave the most accurate indication of electrification in the area. In addition, to deal with the confounding effect of village size, for most of our analyses, the team instead worked an 'average sum-of-lights' measure (Figure 2).

There were also various measurement error concerns noted in the Nightlight data – including the fact that non-domestic lighting, such as light from train stations, police stations and supermarkets, were also picked up by the satellite. To address this, the authors note the value of using available survey/census data and on-the-ground insights to distinguish between sources of non-domestic and domestic lighting, and to isolate specific communities/areas that were perhaps confounding results. In the authors' own analysis, the HDSS data played a key role in this respect: helping them to identify village type and the timeline/process of village development, thus contributing to the validation of the satellite data in light of this measurement error. The team were accordingly able to identify a key under-developed village that appeared as a 'bright-spot' in the Nightlight data, but had in fact spuriously picked up light from a nearby commercial area. Once identified, this village was excluded from the analysis to prevent it from confounding the results.

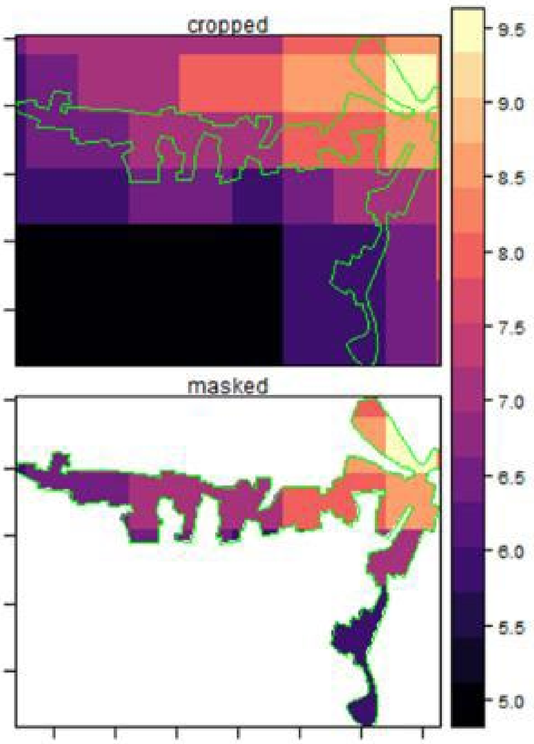

Figure 2. Applying the SOL in a rural context
Note: top panel shows how one village is positioned relative to the pixel grid, bottom panel shows how light measures are allocated to the village

UK Data Service

Research outcomes

The team aimed to corroborate electricity consumption trends across the two data sources. In previous work, Wittenberg, Collinson and Harris noted stable progress and a steady, upward trend in the direction of electrification between 2000 and 2012, as indicated by the HDSS data (as well as national-level survey data) (2017). However, periods of decline and deviation from this trend were also observed – most notably a large decline in electricity access in 2008. After Machemedze advanced the actual technique of linking satellite and local HDSS data, the team identified a definitive link between the two data sources. They corroborated the trends identified in the HDSS data within the Nightlight data. Most notable was the fact that the satellite data for Agincourt demonstrated a decline in brightness in 2008, thereby verifying the potential decline in electricity consumption that had been identified in the HDSS data. This decline was hypothesised to have come as a result of the Eskom collapse over the period: when the capacity of electricity failed to meet the demands of the growing economy, leading to load shedding and electrical rations, and the declaration of a state of electrical emergency. The satellite data also indicated that, even in the wake of the collapse, the levels of electricity use failed to return to the pre-failure levels. This anomaly was attributed to the massive tariff increases that followed.

The research team also investigated whether the satellite data had picked up the rural electrification programme over a more extended period. Rather than solely mapping trends in lighting data in Agincourt itself, which were thought may demonstrate increases in brightness that resulted from changes in satellite instruments, they decided instead to compare within-site trends with similar trends in two counterfactual sites. The take up of lighting for Agincourt was compared to trends in satellite lighting data for The Kruger National Park (KNP) (which is, of course, largely unelectrified) and the nearby city of Nelspruit, using a Difference in Difference (DID) approach. As expected, the electrification levels in Agincourt and the KNP were considerably lower than those evident for Nelspruit at the beginning of the analysis (in 1992) as shown in figure 3 (below). Furthermore, the data showed that Agincourt had a significantly higher tendency to get brighter than the both the KNP or Nelspruit. In 1992, Agincourt displayed almost total darkness, lighting up incrementally to 2007 before dulling in the midst of the Eskom collapse of 2008. The authors' regressions suggest that increases in household connections (measured in the HDSS data) help to predict the rise in brightness.

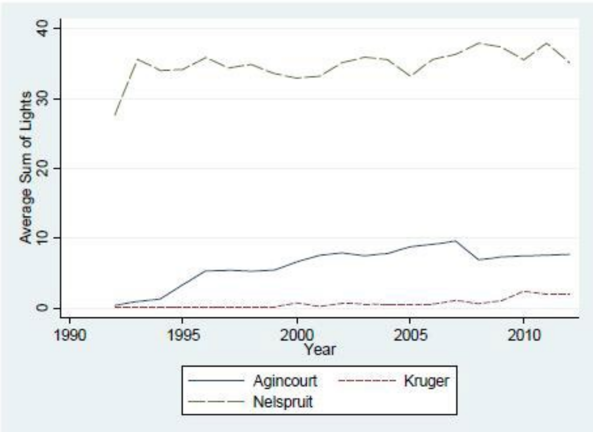

Figure 3. Average sum of lights (SOL), by year and location

In further research, the authors conducted a focused investigation on electrification in villages and homes within the rural site using a standard typology of village – to identify the direct importance of new connections. Their results suggest that the Nightlight data shows spatial variation in brightness – with spatial variation being definitively associated with the pattern of local development and electrification. More specifically, there is a clear two-way split in brightness: with "central" and "established" communities having a brightness level around 8.8 and "undeveloped" and "refugee" settlements three points below that. They also match the HDSS data on household connections to the Nighlight data for villages, with their results suggesting that 200 new connections in a village (per square km) would increase the brightness level of that village by 1.4 units – or a 20% increase on the baseline in 2001.

Conclusions

This work highlights the substantial insights that can be gained by linking restructured complex large-scale longitudinal data to other big data sources such as satellite data. More specifically, it demonstrates that satellite imagery data for the last two decades has not only truthfully captured the electrification of rural areas, but also displays the differences between the developed and undeveloped areas of the site. Even from a non-technical perspective, the gravity of the research being done is evident and it is clear the potential exists to do so much more.

The observed data suggest a clear correlation between local survey estimates of electricity use/access and the Nightlight data – which raises the possibility that, with due attention to the measurement issues raised above, satellite surveillance data could be used as a powerful proxy measure of connection and development. More importantly, this work demonstrates in particular that this can be done even in rural areas and at a spatial resolution where this has hitherto not been attempted. What might researchers and policy-makers gain from establishing this relationship? Firstly, one can use these estimations to derive imputed measures of access in years and for contexts where there is no corroborating survey evidence. Secondly, by complementing different data sources in the manner demonstrated (e.g. by linking survey data to satellite data), one will be able to track usage patterns (e.g. load shedding) where the connection data is uninformative.

Authors:

Takwanise Machemedze and Martin Wittenberg, Cape Town University, Taryn Dinkelman, Dartmouth, and Mark Collinson and, Wayne Twine. MRC Wits Rural.

Full paper:

Machemedze, T., Dinkelman, T., Collinson, M., Twine, W., Wittenberg, M. (2017). Throwing light on rural development: using nightlight data to map rural electrification in South Africa. A DataFirst Technical Paper 38. Cape Town: DataFirst, University of Cape Town.

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

WITS UNIVERSITY

NRF National Research Foundation

DataFirst

UK Data Service