Literature Review - Using Satellite Imagery and Machine Learning to Address Data Scarcity

Cooper Nederhood

2018.04.23

I am exploring using satellite imagery and machine learning to address data scarcity in developing economies. Because it is essentially a pure prediction problem focused on the unique data and computational techniques, I present literature which motivates the question, introduces the novel data source, then present the few and very recent (earliest paper is 2016) papers addressing these techniques.

New technologies and methodologies have led to a recent data revolution for governments, researchers, and private companies in developed nations. However, many developing economies are unable to take advantage of these advancements, resulting in a growing data divide. Many developing countries still lack reliable information on socioeconomic and health outcomes. For example, during the years 2000 to 2010, 39 of 59 African countries conducted fewer than two surveys of national poverty measures. Further, 14 countries conducted no such survey. This data scarcity is often most acute in the poorest and most remote regions, precisely the populations of greatest concern. Even when official measures are conducted, the quality or coverage can be dubious. Finally, informal economies, often of greater significance in developing economies, can remain unseen by official statistics. Making informed policy decisions relies on the presence of high quality data and ambitious initiatives like the UNs Millennium Development Goals highlight the need to develop accurate and cost-effective statistics for even the most under-represented regions. [1]

While satellites have been circling the Earth and acquiring data since the 1970s, only recently have developments in computer vision, engineering, and geography as well as partnerships between governments and private companies led to this rich data being publicly available and accessible to researchers. The richness of the data, in terms of the information encoded in the variety of bands, the spatial and temporal granularity, and the nearly worldwide geographic coverage offers many possibilities for research, including addressing data scarcity issues in developing economies. However, because of the high-dimensionality and unstructured nature of the data, many of these advantages can only be fully attained by exploiting state-of-the-art approaches in computing and algorithmic design. Just this December, the United Nations Task Team on Satellite Imagery and Geospatial Data published a guide for all National Statistical Offices which provides a brief introduction to the use of earth observations (EO) data for official statistics. Satellite imagery can be used to fill in gaps where no data exists and to improve the spatial and temporal resolution of current measures. Because of the different types of bands, satellite imagery can be used to improve measures across a variety of areas including statistics on agriculture, environment, business activity, and transport. Finally, after initial development costs, satellite imagery techniques scale far more efficiently than tradition designed data sources like censuses and surveys, and at a fraction of the cost. [2] [3]

Information provided by satellite imagery extends far beyond simple birds eye views of the Earths surface. Each satellite hosts one or more sensors, which generates independent data stream(s) consisting of energy across one or more bands. Many sensors measure bands beyond what the human eye can see with each band having special use cases. For example, infrared data can be used to measure vegetation and growth, temperature, and (indirectly) precipitation. While most satellites collect data passively, some satellites actively collect data by emitting their own signal and measuring reflectance allowing them to measure additional quantities. For example, the Shuttle Radar Topography Mission (SRTM) provides a digital measure of the Earths topography. Finally, in addition to the raw imagery, many providers calculate derived products like vegetation indices, hydrology information, land classification, biomass burning, and many others. Platforms like Google Earth Engine have greatly increased accessibility by providing a unified platform in which this once cumbersome data can be efficiently accessed, cleaned, and joined. My proposed research relies on data from the Landsat 7 satellite which has 8 bands a spatial resolution of 30m and circles the globe every two weeks. [4] [5]

The most common satellite data used by economists is nighttime luminosity (i.e. nightlights). Since the original study using nighttime luminosity as a proxy for economic activity by Henderson et al nightlights have been used in a host of research, far too much to recall here. [6] The widespread use of nightlights data highlights the potential for fully exploring much richer datasets like Landsat 7. The nightlights data has a 1km resolution, while Landsat has 30m resolution. The nightlights maps each pixel to a 1-63 value while Landsat has 8 continuous bands (not to mention all the other available satellites with different bands). While there have been many novel uses of satellite data, the potential of satellite data to measure economic conditions is incipient. The vastness of the research using nightlights as a proxy for economic activity despite the coarseness of the luminosity data illustrates the potential of using a more fine-grained dataset and modern methods. Most research employing satellite imagery simply uses the pixel band values as a data input, without regard to any surrounding pixels. Thus, approximations of vegetation or urbanization can be calculated over a region which is correlated with economic outcomes, but the actual features of the image are lost. For example, a normalized vegetation index (NVI) cannot distinguish between a gray pixel that is a part of a rock outcrop or part of an interstate highway system. A convolutional neural network (CNN) is a modified deep learning framework which has had recent success in computer vision object detection tasks. Because of the high dimensionality of images, having many fully connected layers would lead to an intractable number of parameters to estimate (even by machine learning standards). Most CNNs begin with a partially connected (convolutional) layer which reduces dimensionality and allows for the location of the target object within the image to vary, critical for analyzing satellite imagery. The first layers of a CNN typically learn to recognize basic features, like edges and color blobs. The features increase in complexity with each layer, until the last layer, a fully-connected layer, maps the resulting feature-vector to a given classification category. Critically, training CNNs require massive amounts of labeled training data. However, because early layers of CNNs identify low-level features like edges models pre-trained on vast datasets like ImageNet can be modified on new tasks such that only later layers need retraining. This transfer learning approach is common practice in computer vision tasks and is central to estimating economic outcomes where ground-truth data is usually in short supply. [7]

Jean et al [8] are leaders in using CNNs to extract economic information from satellite imagery and much of the additional research is at some level an extension of their methods. They use two stages of transfer learning to estimate economic expenditure and wealth indices in five countries in sub-Saharan Africa. They begin with a VGG-F model trained on ImageNet then train a modified model to predict nighttime luminosity (a noisy but data-rich proxy for economic activity) based on daytime satellite imagery for the countries of interest. The CNN learns to identify features like urban areas and roads that are predictive of nighttime luminosity and thus correlated with economic activity. By discarding the final layer of the CNN the model becomes a feature extractor, and these features are then used as explanatory variables in a regression of economic expenditure and wealth indices. They use regression rather than another machine learning method because of data scarcity concerns and to yield R-squared measures which can be directly compared to traditional estimations. They find that their performance approaches the performance of data collected in the field for poverty estimates. However, while their findings are encouraging cross-sectionally, because they use Google Static Maps they cannot assess predictive power across time. By using a time-series of Landsat 7 images my approach will test the time-series properties.

Banergee et all [9], a direct extension of Jean, analyzes development in various sub-regions of India, which provides an interesting test case because of the geographic and economic diversity. They directly challenge Jeans choice to train the CNN on the noisy nightlights data before using the CNN as a feature extractor. Rather, they show that a direct regression of asset indicators gives superior R-squared when compared to a transfer-learning via nightlights approach. They maintain the first transfer learning step, again beginning with the VGG-F model pre-trained on ImageNet. It is unclear whether their ability to omit the nightlight data transfer learning step is due to a richer ground-truth dataset or if it is a function of their CNN structure. Regardless, my current research will test both approaches, albeit using the ground truth data from Jean.

Banergee offers the conclusion their approach can still facilitate monitoring development progress of a region over time. While this suggests the researchers used time-series data, their assertion is in fact based on the different stages of economic development across the different sub-regions of India rather than a true temporal change.

Finally, Goldblatt et al [10] is the only explicit test of satellite datas predictive power through time. They use Landsat imagery, which I too will use, and they use enterprise data, employment, and expenditure from a geo-coded dataset in Vietnam as their ground-truth data. They find that the Landsat data and the coarser nightlights data have similar explanatory power cross-sectionally but that both perform poorly in the time-series which severely limits the usage of remote sensing for predicting economic changes. However, they use only a simple linear model to analyze the imagery rather than the CNN feature-extractor which they deem casts doubt on the usefulness of more advanced yet intractable prediction methods. It is possible that their simplified approach is unable to discern complex changes over time, and that a CNN feature extractor could identify such changes. Or, it is possible that while the economy is changing the physical representation seen by a satellite is changing much slower. Most likely, it is a combination of the two, with certain types of growth like manufacturing development or rapid urbanization leading to better time-series explanatory power for satellite images. My target area will include the city of Lagos, one of the largest and fastest growing cities over the period of interest.

Engstrom et al [11], another direct reply to Jean, use CNNs to measure poverty from high-resolution private imagery of Sri Lanka. Because of their high-resolution images (< 5m) they do not need the second transfer learning on nightlights step. They can also extract extremely fine-grained features, like the number of cars in a parking lot, which is not possible with the medium-resolution Landsat images I intend to use. However, they are unable to attain nationwide coverage because of the high cost of purchasing the data. While their research illustrates the bright future as costs decrease and resolution increases, currently they cannot properly scale their approach, limiting its applicability. Their data also does not extend 40 years into the past like the Landsat data, which remains the gold standard in terms of resolution and spatial and temporal coverage.

Finally, these computer vision methods are also increasingly applied to urbanization questions, which often has development applications. Machine learning excels at classification tasks which is especially applicable when classifying land usages or mapping urban form. For example, Albert et al [12] use Google Static Maps and two different CNN architectures to extract features relevant to land-use classification, comparing urban formation across European cities. Given the importance of urbanization to economic outcomes in the both developed and developing economies, these methods will become increasingly related.

References

[1] L. G. Morales, Y.-C. Hsu, J. Poole, and B. R. I. Rutherford, "A world that counts: Mobilising the data revolution for sustainable development," *United Nations Expert Report*, 2014.

[2] "Big data for development: Challenges and opportunities," *UN Global Pulse*, May 2012.

[3] "Earth observations for official statistics," *United Nations Satellite Imagery and Geospatial Data Task Team Report*, Dec 2017.

[4] "What are the best spectral bands to use for my study?." https://landsat.usgs.gov/what-are-best-spectral-bands-use-my-study, March 2018. Accessed on 2018-04-22.

[5] D. Donaldson and A. Storeygard, "The view from above: Applications of satellite data in economics," *Journal of Economic Perspectives*, Fall 2016.

[6] V. Henderson, A. Storeygard, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[7] N. Jean, M. Xie, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *Association for the Advance of Artificial Intelligence*, 2016.

[8] N. Jean, M. Xie, M. Burke, D. Lobell, S. Ermon, and M. Davis, "Combining satellite imagery and machine learning to predict poverty," *Science*, 2016.

[9] S. Banergee, S. B. Paul, M. Sharma, A. Gupta, and P. K. Suraj, "On monitoring development using high resolution satellite images," *Working Paper*, Dec 2017.

[10] R. Goldblatt, K. Heilmann, and Y. Vaizman, "Properties of satellite imagery for measuring economic activity at small geographies," *Working Paper*, Dec 2017.

[11] R. Engstrom, J. Hersh, and D. Newhouse, "Poverty from space - using high resolution satellite imagery for estimating economics well being," *World Bank Group Poverty and Equity Global Practice Group*, Dec 2017.

[12] A. Albert, J. Kaur, and M. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale," *arXiv pre-print*, Aug 2017.