Mapping Poverty and Wealth:
an Alternative Socioeconomic Index for Sri Lanka

Viren Dias, Lasantha Fernando, Tharaka Amarasinghe, Yudhanjaya Wijeratne
LIRNEasia, 12 Balcombe Place, Colombo, Sri Lanka
(viren, lasantha, tharaka, yudhanjaya)@lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people"s lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

Introduction

One of the most interesting documents ever to come out of our part of the world is the *Arthashatra:* a primer for good rule written by the philosopher-statesman Kautilya, a contemporary of Aristotle. In it, Kautilya describes an extremely sophisticated census system, run by a chain of key personnel from the Collector-General of Revenues to the Village Accountants, which ultimately creates, summarizes and synthesizes an inventory of all wealth, incomes and expenditures of households, merchant activities and revenues, and so on - in effect, combining the functions of the U.S. Internal Revenue Service and the Bureau of the Census. [1]

The logic behind government censuses remains very much the same as it was in Kautilya's time: a government needs to know where the wealth of its dominion lies - and, conversely, where it doesn't lie. Patterns of wealth and poverty, when revealed and mapped, allow more precise targeting of taxation, anti-poverty measures, infrastructure planning, and even exploring linkages between the wealth of neighboring areas. One of the innovations modern states have added to the process is the socioeconomic index - also known as a deprivation or poverty index - a single numerical figure that takes the strongest signals from multiple types of data to try and gauge the overall socioeconomic status of a predefined area. It makes it easier to do many things Kautilya would have struggled with - including making simple, direct comparisons between regions, and understanding the ripple effects of different types of measurements on wealth and poverty.

The Household Income and Expenditure Survey, performed by the Department of Census and Statistics, Sri Lanka, calculates both headcounts and measure of the number of households (at a district level) living below the poverty line, for 331 geographical divisions in the country. It also calculates as a Poverty Gap Index that displays a commonly used alternate indicator of poverty. Using the 2012 Census of Population and Housing (the latest available census, as of the time of writing) combined with simple mathematical modelling, we at LIRNE*asia* created an alternative socioeconomic indicator for Sri Lanka, more granular in nature, and based on physical living conditions instead of purely monetary attributes.

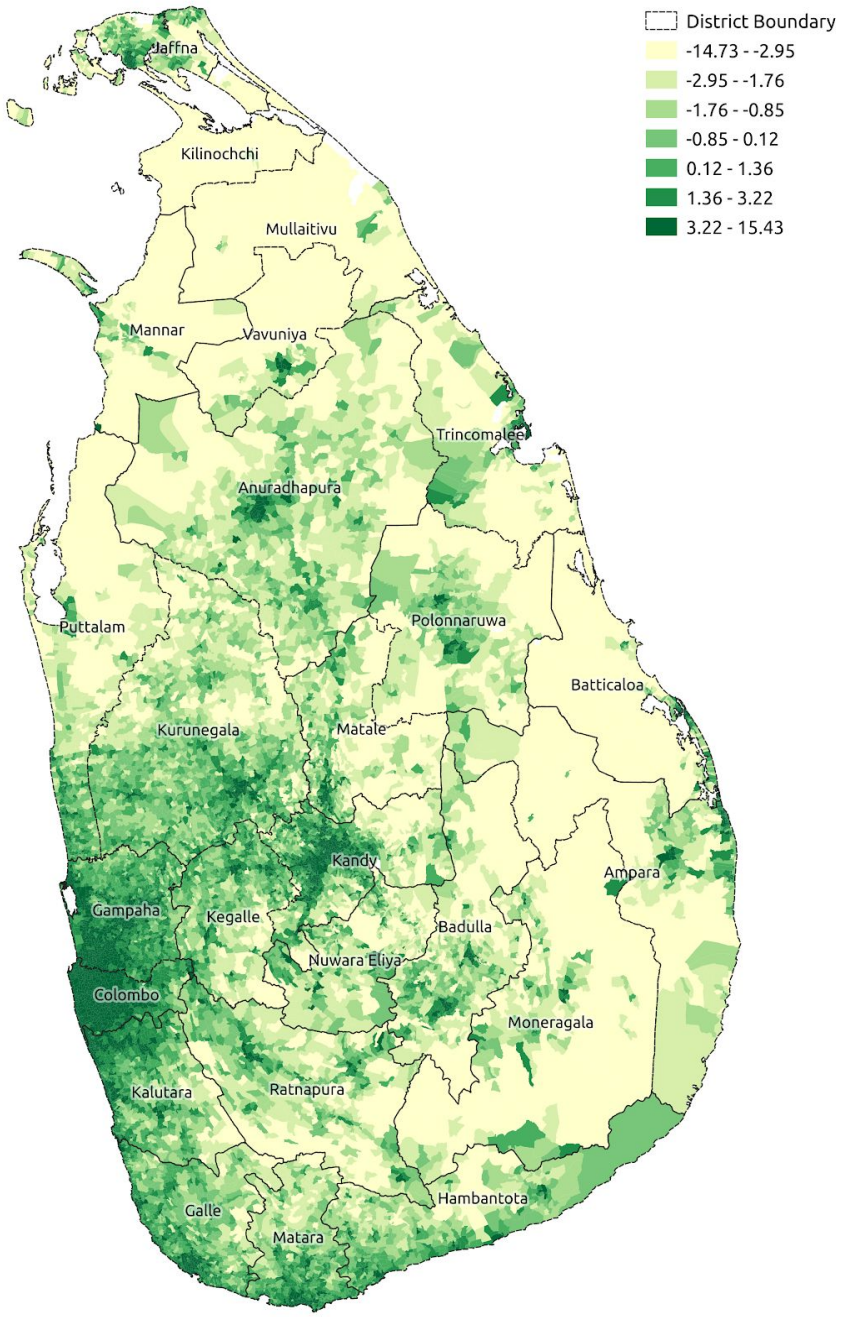

Figure 1: *The socioeconomic index plotted as a choropleth map. Each shade represents one of seven quantiles, with darker shades representing a higher socioeconomic level.*

The result, when mapped out, is fascinating - you can see Colombo, the financial center of Sri Lanka, in this green; Gampaha, with quite a distribution of wealth, but not as large; Kandy, smaller but extending outwards; smaller cities like Galle, Matara, Anuradhapura, Polonnaruwa, Vavuniya. And if you look closely, you'll see how this imputed 'wealth' seems to stick to roads.

Correlation or causation? It's difficult to say using this data alone; such analysis is beyond the scope of this writeup. What we want to do here is discuss *how* we arrived at this indicator, the map, and what choices went into it.

The Dataset

A socioeconomic index requires data. We selected, as our input, the 2012 national census, which is available as a summary of counts at the Grama Niladhari Division (GND) level. The GND is the smallest administrative unit in Sri Lanka: there are 14,022 of them for a landmass that is 65,610 km², which meant that analysis at the level of Kautilya's village-account-level analysis is possible - higher than that of the data released in the 2016 HIES report.

The census gave us a total of 109 variables to curate. Our objective was to first discard anything that did not seem to indicate socioeconomic status, and to keep only ones that could.

Table 1: *An overview of the 2012 national census datasets.*

Dataset	Category	Variables
Household	Cooking Fuel	6
	Floor Material	7
	Housing	4
	Lighting	6
	Roof Material	7
	Structure	9
	Tenure	6
	Toilet Facilities	6
	Wall Material	8
	Waste Disposal	6
	Water Source	13
Population	Age	20
	Education	6
	Employment	3
	Gender	2

In this data, observations made at a household level have been converted to binary variables and aggregated for each GND. This would not necessarily be a problem if each GND was homogeneous with respect to socioeconomic status, but reality is not as organized as we would like it to be.

This obscures certain correlations between variables - consider a GND where half the houses have granite flooring and tile roofing, and the remaining half have cement flooring and

asbestos roofing. When combined, the nature of the data would mean that granite flooring would be equally (and artificially) correlated with tile roofing, cement flooring and asbestos roofing.

As such, we discarded the following variables:

- All variables in the Housing subcategory (not to be confused with Household). The Housing subcategory describes the type of housing the structures fall into, built using an internal classification based on floor, roof and wall materials. As the dataset already contained the core data, this secondary analysis was deemed irrelevant.
- All variables in the waste disposal - this indicated whether waste was collected by the government, burned, buried, composted, thrown into a road/river/canal/sea, and so on. This we removed on the basis that how to dispose waste is a function of how well the government garbage collection services work in your area, and in Sri Lanka both the rich and the poor burn and 'dispose into the environment'. This is due to inefficiencies in the garbage collection systems in both rural and urban areas. [2]
- We removed age and gender categories on the basis that they are not going to be useful for determining socioeconomics at this 30,000-foot view.
- All "other" variables from all categories, because "other" is inherently ambiguous, ultimately shaped by the data collection and the forms involved, and could potentially cover a wide range of data that might be unexplainable - whether in success or failure.

In the end, we are left with 61 variables pertaining to households and 9 variables describing aspects of population. We then went on to build the index.

Building the index

Firstly, even with all this reduction, there are still many indicators to choose from - income, expenditure, education, occupation, durable assets. The actual methods for performing the task differ greatly. Social scientists often use different measures to create a composite indicator, often using different types of data and different methods of combining them to impute something meaningful. Several governments and organisations have developed socioeconomic indices for their respective regions that have been widely accepted:

- The National Statistics Socio-Economic Classification (NS-SEC) for the United Kingdom
- The European Deprivation Index (EDI) for Europe
- The Socio-Economic Indexes for Areas (SEIFA) for Australia
- The New Zealand Deprivation Index (NZDep) for New Zealand
- The Global Multidimensional Poverty Index (MPI)

Each of these differ in the datasets used and the analysis performed on top. Attempting to replicate such qualitative methodology would take us years and many arguments to justify their relative importance in a way that held for the entire country.

Which is where principal component analysis (PCA) comes in. The construction of a socioeconomic indicator requires that we carefully identify which combinations of signals tell us something meaningful. PCA goes about this by combining different variables and checking for which of these combinations account for most of the variation in the data[3]. It algorithmically reduces complex datasets to the signals that are most critical, and has the advantage of being previously used for used in public policy and generating satisfactory results[4].

We make the assumption that the first principal component resulting from the application of PCA on a dataset of socioeconomic indicators is the socioeconomic index. However, the reliability of this index is contingent on the careful selection of variables to include in the PCA.

Ideally, we would have run the PCA on a household level dataset of binary variables. For given household in such a dataset, only a single variable within each category would have a value of 1, with the remaining variables having a value of 0. In order to emulate this, we normalized the variables within each category so that they represented the proportion of households in a GND possessing a certain attribute of a category. We then standardized each variable and ran PCA on the dataset.

Table 2: *A sample dataset.*

Category	**Cooking Fuel**			**Lighting**		
Variable	**Kerosene**	**Gas**	**Electricity**	**Grid**	**Solar**	**Wind**
GND 1	400	520	80	350	230	420
GND 2	560	200	240	730	80	190
GND 3	280	320	400	120	600	280

Table 3: *After normalization.*

Category	Cooking Fuel			Lighting		
Variable	**Kerosene**	**Gas**	**Electricity**	**Grid**	**Solar**	**Wind**
GND 1	0.40	0.52	0.08	0.35	0.23	0.42
GND 2	0.56	0.20	0.24	0.73	0.08	0.19
GND 3	0.28	0.32	0.40	0.12	0.60	0.28

Table 4: *After standardization.*

Category	Cooking Fuel			Lighting		
Variable	**Kerosene**	**Gas**	**Electricity**	**Grid**	**Solar**	**Wind**
GND 1	-0.09	1.07	-1.00	-0.16	-0.27	1.06
GND 2	1.04	-0.91	0.00	1.07	-0.83	-0.92
GND 3	-0.95	-0.16	1.00	-0.91	1.11	-0.14

We multiplied the weights of the resulting first principal component with the standardized dataset and summed each row to produce a score for each GND. This score was to serve as the socioeconomic index.

We can then plot this index as a choropleth map (a type of thematic map where areas are colored in proportion with a given variable; in this case, the strength of the index for each GN division), with each shade representing one of seven divisions along the index. The resultant indicator and the choropleth map creates a picture that, as we stated in the introduction, matches real-world observations. World Bank analysts have long posited that economic growth is positively affected by infrastructure[5]; reports from the OECD highlight socioeconomic benefits to people having access to transport, and to each other [6-7]. It is no surprise, therefore, that the socioeconomic index for Sri Lanka shows higher socioeconomic levels along road networks, nor that darkest skews appear in the Western Province, which accounts for close to 42% of Sri Lanka's GDP, according to the Central Bank of Sri Lanka.[8] Moreover, the thickest clusters appear around well-established cities.

The relevant datasets, code and results can be found at this GitHub repository.

Discussion and further research

The utility of such mapping can not just be understood from ancient texts, but in more recent times from the World Bank's *More than just a pretty picture*[9], which examines a number of mapping exercises around the world and their impact. One of those case studies happens to be Sri Lanka: maps generated as a 2003 collaboration between the World Bank and the Department of Census and Statistics were used in the Samurdhi poverty alleviation measures.

Naturally, the key question is: why use outdated census data when the Household Income and Expenditure Survey [10], which happens at more frequent intervals (every 3 years) is available? The answer lies in the classes of data used for this analysis - externally visible facets of housing, such as roofing types and construction materials; sources of water, energy; and aspects of education. Longstanding efforts in digital photogrammetry have made significant leaps forward in imputing these classes of data using satellite imagery [11-12], as seen in applications such as Facebook's Population Map, which detects buildings before estimating their occupancy.
What we have illustrated here is a method that with relative ease can turn selected classes of data into a socioeconomic indicator, which can then be used for mapping at fairly granular levels. It even has a key advantage over the Household Income and Expenditure Survey, in that it is more granular. The HIES operates at units called District Secretariat Divisions (DSD)s - of which there are 331 in Sri Lanka. The data released in the is aggregated at the level of districts, whereas the index calculated here returns a value for each of the 14,022 GN divisions in sri Lanka. Much additional research is required, but it is not infeasible to imagine a near future where the necessary data is available at a much faster rate than an HIES survey and socioeconomic facets can be imputed at higher frequency, with significantly less cost than the methods of today.

References

[1] Waldauer, C., Zahka, W. J., & Pal, S. (1996). Kautilya's Arthashastra: A neglected precursor to classical economics. Indian Economic Review, 101-108.
[2] Kumaranayake, G. (2013) Waste Management in Sri Lanka: Effective Approaches. Parliamentary Research Journal, 1:03.
[3] This is a non-technical explanation. For the full explanation, see Jolliffe, I. (2011). Principal component analysis. Springer Berlin Heidelberg.
[4] Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis. Health policy and planning, 21(6), 459-468.
[5] Calderón, C., & Servén, L. (2004). The effects of infrastructure development on growth and income distribution. The World Bank.
[6] ITF (2017), Quantifying the Socio-economic Benefits of Transport, ITF Roundtable Reports, No. 160, OECD Publishing, Paris.
[7] Stone, S., Strutt, A., & Hertel, T. W. (2010). Assessing socioeconomic impacts of transport infrastructure projects in the Greater Mekong Subregion.
[8]https://www.cbsl.gov.lk/sites/default/files/cbslweb_images/press_20190925_Provincial_Gross_Domestic_Product_-_2018_e.pdf
[9] Bedi, Tara; Coudouel, Aline; Simler, Kenneth. 2007. More than a pretty picture : using poverty maps to design better policies and interventions (English). Washington, DC: World Bank.
[10] http://www.statistics.gov.lk/HIES/HIES2016/HIES2016_FinalReport.pdf
[11] Chaudhuri, D., Kushwaha, N. K., Samal, A., & Agarwal, R. C. (2015). Automatic building detection from high-resolution satellite images based on morphology and internal gray variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1767-1779.
[12] Zhang, A., Liu, X., Gros, A., & Tiecke, T. (2017). Building detection from satellite images on a global scale. arXiv preprint arXiv:1707.08952.