## Introduction

In 2015, when the United Nations envisioned the world in 2030, eradicating poverty was listed the first among its 17 Sustainable Development Goals ([SDGs](https://sdgs.un.org/goals)) that the world should work towards.

“We resolve, between now and 2030, to end poverty and hunger everywhere,” the 2030 Agenda for Sustainable Development, adopted in 2015, [reads](https://sdgs.un.org/2030agenda). But a goal as meaningful and ambitious as this is almost certainly easier said than done. And to work towards this goal, we first need to know where poverty lies as well as how it differs from one place to another, in order to allocate resources in accordance with local needs.

## What makes an ideal poverty map and why past efforts were less than ideal

An ideal poverty map would be both (1) spatially granular and (2) temporally up-to-date. It other words, it captures the slightest nuances across both time and space — nuances that are easy to ignore but meaningful and necessary for making informed and tailored poverty alleviation policies. Perhaps one of the best examples of such application in Sri Lanka is the reform of the Samurdhi (or Prosperity) Subsidy Program in 2005, when the ministry in charge of this national social assistance program used the map to identify the poorest 119 District Secretariat Divisions (DSDs or DS Divisions) that deserved the most attention (##2015 Report).

Historically, however, this has been no easy task. High-quality poverty maps that are both granular and up-to-date are hard to produce. This can be largely attributed to the difficulty of collecting detailed data on poverty how — and when — data on poverty are collected.

Administrated typically every ten years, censuses remain the most detailed and granular source of statistical knowledge on a country, but in most developing countries, information on income and expenditure — indicators strongly suggestive of the presence of poverty — are rarely collected (## Tarozzi & Deaton 2009).

Sri Lanka is no exception. The Sri Lankan census (called Census of Population and Housing, or CPH, last collected in 2012) makes available statistics on arguably the most granular level possible for each of the 14,022 Grama Niladhari Divisions (GNDs) in Sri Lanka. Nevertheless, none of the 109 variables it collected directly addresses income or consumption — indicators that have most often been used to capturing poverty in past research.

In its stead, the Sri Lankan government relies on a different source to understand the spatial distribution of poverty (albeit on a much less granular level): the Household Income and Expenditure Survey (HIES), which takes place every three years, a much shorter cycle than the decennial census. It collects information on the economic wellbeing of nationally representative households, such as their source of income and expenditure structure. It then calculates and presents the percentage of population falling below the official poverty line in a given district (known as Headcount Index, or HCI) as well as a common measure of the severity of poverty known as Poverty Gap Index.

For those familiar with administrative geography in Sri Lanka, however, the way HIES aggregates and makes available data lends itself a fatal shortcoming (despite the comprehensiveness of its statistics). While the survey is conducted at a relatively granular level of District Secretariat Divisions (DSDs or DS Divisions) — of which there are 331 in Sri Lanka — the survey’s sample size is not large enough to accurately estimate statistics below the level of Districts, a higher-level administrative unit of which there are only 25 across the island (## DCS 2015 Report). In other words, by relying on this survey itself, we cannot possibly gain any concrete knowledge of the economic disparity that expectedly and demonstrably exists within each District.

## What we are doing instead

This is where our current approach comes in.

Thanks to the rapid development of remote sensing technologies as well as growing prevalence of mobile phone usage, we now have access to a new stream of high-resolution data derived from satellite imageries (produced through remote sensing, thereafter called RS data) as well as call detail records collected by telecommunication operators (thereafter called CDR data). Taken together, they offer us a rare look into the ever-changing spatial pattern of socio-economic wellbeing on a much more granular and up-to-date level, beyond the limits of traditional sources like censuses and surveys.

The rationale behind using these two novel sources as a proxy for predicting poverty is relatively straightforward. RS data capture the physical landscape of a place such as the presence of agricultural land, the compactness of buildings, the connectivity of roads, among many others. Many of these characteristics determine or reflect the ease to conduct productive economic activity, and are therefore strongly correlated with the overall income level and socioeconomic well-being of an area.

Likewise, the pattern of mobile phone usage revealed by CDR data would be also indicative of a region’s overall possession of socioeconomic resources. Patterns of interests include not only the prevalence of mobile phone usage in a given region, but also the intensity at which people make calls, the distance their calls reach, or the frequency at which they travel. All of these could reflect the caller’s socioeconomic status, from their ability to afford a cell phone, pay the phone bills, or travel and appear at multiple locations captured by cell towers within a short period of time.

## How we are doing it

We adopt a traditional, widely practiced machine learning approach to examine the validity of our alternative mapping method.

Our plan is to fit three types of models: (1) non-spatial frequentist models; (2) spatial frequentist models; and (3) spatial Bayesian models. All of the models use RS and CDS data as the predictors and the alternative socioeconomic status as the referenced ground truth (##). They differ in underlying assumptions about what better explain the differences and minimize our inevitable uncertainty of the accuracy of our modeling.

What differentiates a spatial and non-spatial model is the varying belief about whether location matters. That is, for example, whether the poverty of one region can be partially explained by its mere adjacency to another low-income region, and vice versa. A simple non-spatial (OLS) model insists that it might not, while all the spatial models assume that it may.

The difference between frequentist and Bayesian models, in turn, boils down to an interesting disagreement in statistical philosophy: what is the nature of probability, and what makes a better model anyway?

Simply put, Bayesians understand probabilities as degrees of uncertainty in one’s personal belief, while frequentists consider probabilities as objective frequencies produced by some unknown process. Crucially, Bayesians always start with a prior belief about the event of interests. In our case, it could mean us starting with a prior belief in the level of poverty in each GND, based on previous survey and census data (even though it does not necessarily have to be true). As we make new observations (e.g. from the RS and CDS data), we systematically update our beliefs and develop a comprehensive understanding of the full spectrum of probabilities associated with different estimates.

In a frequentist manner, however, we do not start with any particular prior belief. Instead, we compare multiple competing hypotheses — without favoring any one in particular to begin with — until we conclude the most likely answer while rejecting the rest.

## Preliminary Results

At this stage, we have finished fitting all the frequentist models – spatial or non-spatial – which we started with given the relative simplicity of computation and the relative abundance of information available to us.

The results of the four frequentist models do not significantly differ from one another, but statistical tests show us that a Spatial Durbin Error Model (SDEM) performs comparatively better than a Spatial Error Model (SEM), a Spatially Lagged X Model (SLX), or a non-spatial Ordinary Least Squares (OLS) Model.

One of the characteristics of SDEM is that it acknowledges the existence of certain variations that are spatially related but not captured by observed variables. Another is that it understands the spatial effects as local, not global – that is, neighbors only impact each other, without propagation throughout the system. This should align with the nature of our predictors, which are physical landscape captured by remote sensing satellite imagery as well as pattern of mobile phone usage captured by telecommunication operators.

The following are the maps of the SDEM fit as well as original ground truths.

Map

Description automatically generated