TRANSCIPT

Combining geospatial data and modelling to understand population distributions

Professor Andrew Tatem, Director, WorldPop, University of Southampton

I'm going to give a quick overview of the work of WorldPop at the University of Southampton.

We collaborate with UN agencies and statistics offices to complement traditional forms of population data with new forms of data from satellites, from cell phones, from geolocated surveys.

Part of some of the challenges we're trying to address is that in many countries, in low- and middle-income settings, the population data that exists can be very outdated, can be many decades since the last census, there can be a lack of registry administrative data systems to keep things up to date - and there have been disruptions to census plans for the 2020 round to fill some of those gaps.

Even with a reliable census, in between those every 10-year censuses there can be changes at small area scales and fertility rates, migration and displacement that can mean that population numbers can quickly become outdated.

We're trying to tackle these kinds of situations where data that we trust may be available but at course resolution, or data that exists is outdated or incomplete or inaccurate.

But of course, what we do have is new forms of data particularly from satellites and the ability to extract information from those datasets. The aim here is to try and build up a picture of the landscape and the types of correlates - characteristics of the landscape - that relate to how populations distribute themselves on the landscape. In the situation where we don't have recent data, or we don't have detailed data we can use these as surrogates.

This may be things like settlements, as well of the location of schools, roads, marketplaces, data that we can get from interpolation from surveys and neighbourhood types, poverty rates. But importantly also these satellite derived building footprints. These are becoming increasingly available and increasingly accurate, and we've been working with datasets across the whole of sub-Saharan Africa recently.

As well as using those data as they are, we can also extract additional value from them. So, we see patterns in these data, and we can measure those by measuring building densities, counts, areas. We can train a computer to recognise patterns in those data to identify automatically different types of settlements, neighbourhood types - which may relate to how population densities vary within cities.

We can also train these models to identify the types of buildings as to whether they are residential or non-residential. We can tell a computer to say these are the types of buildings that are typically non-residential in a city and be able to classify those.

This gives us a set of data to try and improve population estimates at finer scales and more recent data. So, we may have some data that exists, we may have satellite derived maps of settlement or building footprints that can determine where people are and where they're not. And then we have these geospatial covariates that can tell us within those settlements what are the likely variations in population density. These feed into statistical models to produce gridded estimates, 100 by 100-meter grid squares, ideally with measures of uncertainty, because we're never perfect in our estimates.

There are two types of approach that we use. One where we may have census counts or projections that cover an entire area and we want to get smallest area estimates. So, we have to take those aggregate counts and disaggregate those to grid squares. Or in the situation where we don't trust those census data. Maybe they're it's been decades since the last census, and we don't trust the projections but there may be small area estimates and we want to try and fill the gaps and produce estimates.

So, for that first situation - we may have data like this from a census that tells us there are a hundred thousand people in each one of these units, but we don't know where. But we do have much more detailed data from mapping of buildings, from satellites, from things like land cover datasets, the brightness of lights from satellites - each one can tell us a different bit of information about why we may see different population densities in one area and not in another.

And so, we can use this stack of data that we built up, sometimes 50 or 100 different layers, and the relationships that we see at these coarser scales with these finer scale data to disaggregate to 100 by 100 meter grid square datasets.

This is something that we've done using projections census data across the world for multiple time periods to produce these global gridded datasets and estimates, as well as age and sex structures when we're bringing in household surveys and those census data.

Increasingly the possibilities exist to project those into the future. There are increasing amount of efforts to model the settlement growth in between years of observations but also projecting into the future.

This enables then, with projections of population, to disaggregate those projections for official subnational projections, or to produce models to project under things like the shared socioeconomic pathways that we've seen based on some of these datasets. We see a wide use of these types of data in the health field in particular but also increasingly in disaster response. And we see these gridded datasets now in the FAO Hand-in-Hand Initiative and the Geospatial Portal.

But what I've presented so far is this disaggregation, it doesn't tackle the situation where we don't trust those data and we have gaps in those data. This has been an increasing amount of our work recently - working directly with governments and co-developing methods to produce new estimates of population in the absence of a census.

An example comes firstly from Afghanistan, where the last census was 1979. And they're reliant simply on straight line projections from that baseline. But there has been recent collection of data from some areas. And so, we a need to build these models, use the satellite data and the geostatistical models to estimate into those unsampled areas, and produce new estimates. These work pretty well, these types of models. Importantly you're producing predictions so there are uncertainties around those estimates, and we see in those areas a long way from where we have data high levels of uncertainty, but it does produce new estimates to move beyond 1979. And we presented to the previous President - and these datasets have been used in-country ever since.

A similar situation - Burkina Faso - the recent census could not reach certain areas of the country because of instability. But these types of models enable a way to produce estimates, and this feeds into the census numbers.

Then in the situation where we don't trust the most recent census - in this case Nigeria 2006 - a long time since populations have been enumerated, and many problems in those data. But working with field teams, new enumerations can be collected in some areas, used to produce new population estimates and these have since been used in multiple health campaigns.

And with a similar thing in Democratic Republic of Congo, where the last census was 1984 - working with field teams to collect new data, representative samples across the geographies and the demographics and the types of neighbourhoods that exist in the Kinshasa region, to produce these new estimates. And these have then since been used in a range of different vaccination and survey campaigns.

Many of these datasets - we collaborated with multiple countries across sub-Saharan Africa - now are available for you to explore and download. And the aim here is that we can start to integrate these types of geospatial models and datasets in a way towards producing projections, but also updates of those population estimates. So, we may have a baseline 2021 census that we can disaggregate. We need a way to be able to project forward, to be able to produce datasets for preparing the next census. And this is where we can link together data from these satellites, from administrative data sources, together with household surveys, GIS feature data to update and refine these models, to ultimately produce new estimates to prepare and plan for the next in-person or census enumeration.

Many of these approaches are described in UN documents we've collaboratively produced, and there are many training materials as well. I've done a selection here on the top down and the bottom-up types of population modelling.

Thank you.