Where is everyone? Mixing and matching modelling methods for mapping populations

Transcript

I'm going to give an overview of the work that we've done in the WorldPop team, but also wider collaborations with GRID3, Countdown, UNFPA and many other partners.

First of all, that's the question - why do we need population data? I guess most of you know the reasons. Thinking about the kind of questions we may be tackling in the global development field. Is it a major outbreak? What is our denominator? Do we have the workforce to address it? Where is our target demographic group? How to target surveillance and campaigns to reach those groups? If we're running a vaccination campaign, how much vaccine do we actually need? Where do we need it? How should we plan our delivery? And then - a disaster - how many people are affected by this disaster? Where should we deliver aid?

There are many, many other questions and uses of population data in government, but the challenge we face is often this kind of situation. So, this is the situation in 2020. If we're using census data as our basis, done every ten years, there are many countries back in 2020, and still today, that haven't done a census in the latest round.

And it's not just the challenge of having a recent census, but even with the recent census populations can change fast. There are spatially varying fertility and mortality rates that make small area projections difficult to do over the time periods between census years. There's displacement and migration going on that are changing those populations all the time.

And then when we do have reliable, recent data sometimes the data that we're working with can be like this - matched to administrative boundaries but maybe 100,000 or more people in each one of those units. And we don't really know where those people are to be able to do targeting of resources or more fine scale planning. What ideally we'd like is perhaps something like this - gridded population estimates that really show those distributions of the populations within those administrative units. It may look like this - close up and gridded datasets producing an estimate, maybe with uncertainty measures as well. Each one of those hundred by hundred metre grid squares. And ideally some measure of how these populations breakdown by male/female, age groups, and over time.

The value of having these gridded data or many and varied. We can then aggregate those data to different decision-making units. Whether those are administrative, they may be health units, they may be enumeration areas for planning for the next census or planning a survey. They could be numbers of people by settlement. We can delineate settlements from satellite imagery, estimate those populations within those - like adding up those grid cells. We can look at the integration with other types of data. So where are health facilities and how many people are living within 5 kilometres of a health facility.

Then we can start to integrate those to answer questions like identifying where vulnerable populations are in relation to vaccination centres. Here showing those estimated populations over 65 years of age that are within areas that are settled, and then adding in here travel times to settlements, health facilities and population datasets to understand what's the travel time to vaccination centres. And then highlighting those areas that are a long way from those vaccination sites to try and target resources and increase coverage.

Broadly the kind of approaches that have been used to produce these types of gridded datasets involve 3 main ingredients: demographic population data, some count of population, settlement data (that can often come from satellite imagery) to define where are the people living and where they're not. And then some form of geospatial data that tell us within those settlements and across the country, how do those population densities vary within settlements, within region. And it's a statistical model that brings those together. To understand the relationships, to look at where we have data, how do those settlement and geospatial datasets vary, and can we then use the geospatial data to predict population numbers for grid cell by grid cell - ideally with uncertainty.

This looks like we have a solution that works everywhere - we have the model that can just apply everywhere. Unfortunately, it's often not the solution that it's that simple.

There can be a wide range of demographic datasets. Why use a 20-year-old census when we have reliable projections or have registry data that we know is more reliable and up to date? What about where field teams have gone and collected small area enumeration data from a household survey? We may have a demographic surveillance system. We have health campaign data - all of which take different shapes and forms.

Then there are all types of different models, that can sometimes give you a headache reading them. There's not many people in the world who even know what all of these mean, and applying them produces different outputs.

Then, to complicate things further, there are all kinds of new geospatial datasets coming - mapping individual building footprints, building heights from satellites, classifying them into neighbourhoods, identifying residential/non-residential populations, the changes overtime, images of the Earth at night, of land cover, of rates of poverty, of mobility, and mobile phone movements.

Often when we're talking to non-technical specialist, this is the kind of result that causes. Multiple different types of demographic data, multiple types of models, multiple geospatial covariates.

So what are the overriding questions here? We need to take a step back - think about who needs these data, and what do they need it for. That may be a situation of a National Statistics office example, where data is used to support the census process, fill gaps in official statistics, maybe produce intercensal estimates or an updated sample framing. There are some basic key needs that ultimately, ideally, they should be based on data that they have collected routinely through a census or surveys. There needs to be ownership of those outputs to be able to understand and explain those methods, to be able to argue them within government and importantly to the public as well.

Ministry of Health may have slightly different uses - denominators for health metrics, targeting interventions, developing micro plans, strategic plannings. So, there's a need to integrate with other forms of health data, to be broken down by age and sex groups is important for targeting resources, targeting vaccinations, bed nets, and be able to (ideally) update those quite regularly as disease and health situations change.

Then a third use maybe humanitarian response agency. Assessing those populations affected by a major impact, major event. Targeting aid, monitoring changes as a situation unfolds. So, a need for rapid updates, simple outputs, and the ability to adapt to often recently collected messy data.

How do we ensure those needs are met? Methods can be understood by those people who are producing those outputs? The data can be accepted by governments and actually used. We don't want them just sitting on the shelf.

This is where, importantly, co-development with those end users of data are pretty vital. This should take place at the very start of any kind of process to understand what are the needs, what data exists to meet those needs, and their recent survey data - so, recent census. What skills exist within a Ministry of Health, or national statistics office, or a local university, to design and take forward solutions. And then coming up with plans for co-development training, plans, working together, and all the examples are going to present here are ones that have started off with some kind of workshop, or online meeting that have asked these questions and come up with a plan.

There are very different types of solutions depending on where we are in the census cycle. We may have a census where a few years after the enumerated totals can be the most accurate. But as time goes on and projections are used, things become increasingly uncertain there may be need for different types of approaches. Then there can be new forms of data collected and planning for a census, until you go back to that process at the start of the census enumeration. The big question of course is when was the last census? Which for certain countries can be 20, 30 even 40 years ago.

So, this first part, the census enumeration, has often interest from National Statistics offices in having small area estimates that match the census counts. Those data that can be available can be at aggregate level - aggregate counts by district, by province - and so this is where top-down disaggregation to take in those counts at the administrative units. Using the more detail from the geospatial datasets in a random forest model (in this case) to estimate those populations at finer spatial scales.

Here are three examples: work with statistics offices in Kenya, Mozambique, Sierra Leone, to produce these gridded estimates from the most recent censuses. Producing much finer scale information.

There are open training materials. We are getting to a stage now where there are many statistics offices or ministries of health who can run and update these types of models themselves.

The second stage - maybe this point where we're getting beyond the census data into producing projections. Here we may be in a situation where we have 2015 census - we want to project forward using different types of projection models (whether it's a cohort component, a mathematical one, an economic one) to produce 2022, 2023 estimates and then disaggregate those to grid squares to produce those small area estimates that can be so useful for different applications.

Here are examples of working with the Senegal statistics office - their official projections forward to 2020, and then disaggregating those. In Niger projecting forward from the 2012 census to 2021 and disaggregating those to grid squares.

The situation of Ukraine - the last census was 2001. Here working with the Ukraine statistics office and UNFPA to produce common operational datasets based on registry data. But these are only available at province level, so disaggregating those again the grid square scale for use in humanitarian response.

In the situation where there has been lots of displacement, being able to incorporate other forms of data into those projections. In this case displacement surveys from the UN Organisation for Migration - to then disaggregate and also map out the locations of refugee camps to account for those.

Then we get to the situation where maybe we don't trust those projections at all. We’re a long way from the last census, we don't trust the projections because things have changed so much, and we need a way of estimating populations. And here from the toolbox, maybe working towards something like a bottom-up model, where we only have small area enumerations that have been collected recently that we trust, and we want to build a statistical model to estimate populations in the areas that we haven't sampled. Here we can develop types of approaches that adapt the different types of data that may be available. In this case of Cameroon here, the last census is just 2005 and projections become very uncertain when we get to 2022-2023. But recent surveys have been undertaken where they've counted populations, and so gathering together all of that data gives over 2000 different locations where populations have been counted recently. So, these can be used to produce estimates and a prediction model that provides uncertainty as well. In each case in collaboration with the statistics office.

A situation similarly in Papua New Guinea where malaria surveys were undertaken and counted populations in those locations - but not across the entire country, and there are gaps here. Again, training a statistical model to predict and estimate those populations in collaboration with the statistics office.

A part of the GRID3 programme - working with field teams to collect individual areas of micro census enumeration to count populations in small areas and use those to train a model and estimate population numbers. And the situation in the Democratic Republic of

Congo - the last census 1984 - and again, working with field teams to collect training data to predict those population numbers.

In each case there are training materials and collaborations along the way to ensure that there is an understanding and a building up of skills to be able to undertake this kind of modelling and being able to explain it within government and to the public.

Then reaching the stage where a new census is about to be undertaken, or maybe in a situation where census cartography data is taken. They can be collected as part of the census preparation phase, but it can often be incomplete. There's a lot of value in having full coverage of recent estimated population for the census process. Here we're making use of that incomplete data to predict numbers of people per administrative unit and then disaggregate. So, combining different models - another type of tool we can use. Here's an example from working with INSTAT in Mali where census cartography was completed in those light grey areas, but not in the dark grey areas. So, the model was developed to fill in those gaps.

Finally, we're back again to the census enumeration. But some cases there can be very hard-to-access regions in that census. Conflict, just difficulty in access, and natural disasters can mean that there are certain areas of the country that cannot be enumerated. So, here a similar type of approach to the census cartography can be undertaken. Here's an example of working with Burkina Faso - last census in 2019 - where these areas in grey or in dark green were under-numerated. And here those areas that could be counted in A were combined with geographical covariates to predict those areas that were not sampled. Those predicted areas were then combined with the full census and used to disaggregate to grid squares in B.

A similar thing in Colombia, where large areas were inaccessible for the 2018 census and the statistics office developed approaches here to estimate those population numbers.

If we can implement these different types of tools, what does this all enable? Some examples here - a gridded population dataset, in this case in Benin, has enabled them, with the development of a preEA Tool, to identify enumeration areas in planning for their next census (which was being undertaken towards the end of last year).

It can be a way to produce official statistics. The Burkina Faso example formed part of those model estimates - formed part of the national statistics. South Sudan has just validated model estimates in their parliament to produce new estimates across the country.

It can be used as a way of planning and developing infrastructure. So, part of the GRID3 programme - engagement with Sierra Leone here – used gridded relation datasets to identify demand for schools and populations within catchment areas.

They can form part of vaccination micro plans - here in Nigeria, producing ways of integrating both population data with boundaries, place names, and settlement delineations as part of GRID3, to estimate those populations in need of vaccination.

Similarly in the Republic of Congo here. Paper maps identifying those areas in need of vaccination, used then by field teams to go and deliver those health interventions.

These types of dataset are feeding now into health information systems, so DHIS2 here has a maps component and is undertaking training in this to be able to integrate health facility locations, understand populations living in different areas of aggregation.

In health financing – here's an estimation of population for the first time since 1979 in Afghanistan, and were used in analysis to justify extra funding - and extra million kids are now vaccinated because of this population data.

With disaster preparedness, having these data based on official projections and official statistics means more likely acceptance of numbers impacted when a disaster comes along. So, this is working with UNFPA to gather together all the official projections and identify those areas at risk of coastal flooding when the hurricane season comes along. And then finally in disaster response, having these data readily available to overlay with other geospatial datasets, like the mapping of flooding, enables rapid assessment of populations affected.

A very exciting component of all this that we're seeing happening is the development and use of these types of approaches by highly skilled statistics offices, with lots of resources. Then their experts are helping to train other statistics offices, who are supporting those with less resources. So, here DANE in Colombia has supported Brazil, and both DANE and Brazil have supported Haiti, and plans to support other countries with model population estimates.

Of course, we see this type of thing happening as expertise builds across West Africa – Data Science, Nigeria there - training undertaken in Thailand - and hopefully can build up more of this expertise.

I'll finish here, but hopefully I've highlighted there's a wide variety of needs, data and changing situations means that often a one-size-fits-all model is often not the most appropriate. Local ownership, co-development, are really important for developing trust in the data if we want to see those outputs used by National Statistics offices, ministries of health. Developing that expertise regionally and in networks can help move us towards sustainability to be able to do this.

Thank you - and just acknowledge that this is the work of a much wider team and many different initiatives.

Professor Andy Tatem
4th April 2023