

From Missing Censuses to Smarter Decisions: How AI Is Reshaping Population Data

Transcript

Vissého Adjiwanou: We start with Professor Andrew Tatem, who is Professor of Spatial demography and Epidemiology at the University of Southampton and Director of WorldPop. His research combines satellite imagery, surveys, mobile phone data, and censuses to map where people live and how they move particularly in low-income countries. Working with governments and UN agencies, he applies this data this approach to health, disaster response, and development. Andrew, 12 minutes, the floor is yours.

Andy Tatem: Thank you, Vissého, and yeah, thanks to Ridhi and Emilio for setting so nicely. I guess I feel like I'm a bit of a fraud today because I don't really have formal training in demography and I don't really have formal training in AI. So, to be on this session talking about the two together seems strange, but what I guess what I have done is spent a lot of time in national statistics offices, in ministries of health, with UN agencies, seeing the challenges they faced with small area population data to guide decision-making, to guide resource allocation, and thinking about the ways that new technologies, new data sets can try and fill some of those gaps. So, that prompted me about 15 years ago to start WorldPop. We're now a research group at the University of Southampton of about 40 people, largely supported by Gates Foundation and Wellcome Trust. And if I share my screen - we produce data sets like this. It makes for pretty pictures that also potentially useful and talk about how AI can help with some of these things. So, this just shows - obviously the brightness is population density across Europe and you can view this in different ways. You can view it as kind of mountain peaks of population to see the big inequalities in population density that exist or try and break these down by different key population vulnerable population groups. And yeah, this is something that our team does to bring together and try and stitch together a whole range of different data sets. But increasingly we do this in co-production with statistics offices with ministries of health so that this can be locally owned, locally understood and the decision makers can feed into the process and understand why decisions are made, why different data sets can be valuable and how they can help move on from just a straight-line projection from a census 20 or 30 years ago. And the value of these kinds of data sets are not so much in the grid square level but in the ability to be flexible in things like this. If you have a vaccination campaign and need to estimate numbers of children in a health unit if you have a settlement extent and you think about the placement of a health facility so that it's reaching those women of childbearing age, those children at need and these gridded data sets can help. And I'm

going to start on a positive note and then move to the kind of warnings. So, starting with how we are seeing I guess, a love for GeoAI, but also go on to how we're not. And just to say GeoAI, it's a term that really covers the kind of integration of artificial intelligence in all its forms into geographic information science and systems. So, how we use it is in different ways. So, this is a picture of our WorldPop team and brewing the population data sets. And there's different ingredients that that feed in the population data, mapping of where settlements are, so where people are likely to be, and then within those settlements different geospatial data that can help us understand why we may see population numbers higher in one area and lower in another. These feed into different types of models that then produce population estimates at grid square scales. So, we're dividing up the world into either 1 km or 100 m grid cells. And in each of these components, we've found or finding AI

in all its different forms to have some value, but I will go into the warnings behind those. So, if you think about the input data sets, there are now incredible amounts of work going on from tech companies, from academic groups, in terms of as we Ridhi talked about foundation models that can take satellite images and extract from those individual buildings and do that across time, across space, to produce huge amounts of data. Obviously incredibly valuable if we are thinking about trying to understand where populations are and how those are changing in a situation where a census might not have taken place for 20 or 30 years. There are frontier research areas going on with things like street view images where an AI algorithm can be trained to detect and classify buildings. So, this is example here, the whole of Toronto's buildings mapped and classified into whether they are apartments, churches, garages, houses, industrial areas. Again, incredibly useful data set if we want to try and get more accurate estimates of population. Not put people into office buildings or into garages. Then, as we've heard about already as well, these kind of foundation models are appearing. This is a Google DeepMind one called Alpha Earth where they've taken millions of satellite images and applied models to kind of generalize them to pick out features that exist on the landscape. And then, again as Ridhi has just talked about, there are these kind of digital trace data sets that may come from mobile phones, may come from on the right here from Meta AI for Good team, applying all kinds of algorithms to their Meta data sets to look at things of activity spaces of the Meta users. And this can help us in terms of maybe some of the dynamics of those populations at smaller area scales. And then this can help us in terms of if we have census count data that we trust, and maybe there's a million people in each one of these administrative units, it's not ideal for decision-making if we're wanting to place health facilities, we're wanting to understand allocation of resources. But what we do have is this set of geospatial data sets, and AI algorithms can help us learn relationships between these different data sets to disaggregate the grid squares. Or in a situation where a country has not done a census for a long time and we don't trust long-term projections at small area scales, but they have collected

data from a sample enumerations from say a household listing or a health campaign, and we do trust that data, we can use that to train a model again using different types of approaches, which can be machine learning AI type approaches to develop a model and predict into those unsampled grid squares. Another area where we found AI to be potentially valuable is to lower the bar in being able to access these types of data sets. So, this is a small portal we put together with an AI assistant. You can see I'm typing in there, "How many children under five live in Nairobi in 2026?" It goes to our data and extracts from our data and summarizes. And you can ask questions in different languages and lowers the bar to needing geographic information system skills to be able to access those data. So, that's all the positive side, but I want to also add in, of course, some warnings and things that we often see and for all of you to think about. Firstly, there's mapping of buildings. we see incredible variations across the world in the different building data sets that are produced. I have to think about how these are produced. Sometimes there is a huge amount of training data in the global north. Train models and then predict them into the global south and then unsurprisingly don't function so well. Or there are satellite images only available to one group. or these are trained in Silicon Valley using prior knowledge there. And you end up with problems. And we started to look at some of these data sets, particularly across sub-Saharan Africa, you see here this ideally these are four different global building mapping data sets coming from Google, from Microsoft, from Ecopia, and from OpenStreetMap. Ideally, everything should be black here because all four should agree on where the buildings are. But you can see Cairo, Lagos, Kinshasa all see very different patterns. So, the next question is does that actually matter? And so we worked with a company that does planning for anti-malarial indoor residual spraying and looked at okay, if we switched in and out different building data sets, how do the priority settlements you classify change? And it's a huge variation, about 250% difference depending on which building data set you use. So again, a big warning some of these data sets. And then of course the digital trace data, we've already talked about the biases, and we've already talked about AI contamination risk coming in now as well. So, depending on which mobile phone operator you might be working with, whether you're working with Facebook Data for Good, I know if this is in the UK, Facebook users would be people like my parents going to and folk music festivals instead of representing the general population. So, you end up with some big variations that need to be thought through. And then if we're thinking about are we trying to predict or are we trying to explain? Certainly in some of the governments we work with, when ministers ask, why do we see a higher population in this district than another district? You want to be able to explain, want to make sure the stats office can explain to the public and to ministers why this is the case. And if you show them a DeepMind Alpha Earth Foundations and say the reason is that metric number 641 was higher than metric number 642 and we don't know how that was produced. It was just a foundation model then it causes problems. So there's the importance of explainability as well in in these outputs. And then the biases in the input

data. So this is just one example from Mapillary street view imagery and you can see the dark green is where those images are available from. It says street level images for all but actually there's massive differences between Sub-Saharan Africa and Europe in terms of that data. And so if large scale models have been trained on these data, you can already see the biases that exist again. And yeah, if you play with our AI tool as well and you - this is this morning I put in this I said, "How what's the population of Baoro in Central African Republic?" There basically is no geographical boundary that exists that we might expect in the UK here to have a boundary for major towns to be able to extract the numbers of people that in many countries across the world that just doesn't exist. And so again there's this geographical mismatch in some of the use of these large language models. And yeah, just I think finish thinking about what we what we really want to be able to train some of these models to bridge some of these gaps is high-quality demographic data. And some of those unfortunately are disappearing. We see in the last round of censuses about a third of the world's population were not counted. That's a big bias towards India not doing a census, but there's some big gaps that still exist. We have around 30 different countries coming to us asking for support for the next census, for filling gaps, for planning for where the budgets have been cut. So, in a situation where we want to ideally train AI models with the best available demographic data so that we get the best answers coming out. We're in a situation where many of these foundational data sets are in decline. So, that's something I want to leave you with to think about and prompt the debate. Thank you.

Vissého Adjiwanou: Andrew and Matthew, you have both touched this same nerve. You both produce numbers for places and for moments where the traditional census fails. Where the census is old or missing or where the event has not happened yet. So, here is my question to you both. How are your estimates validated when there is no gold standard? Andrew, let us start with you. In your presentation, the answers seem to be do not look for a gold standard. Propagate the uncertainty instead. Is that still how you see it?

Andy Tatem: Thanks, that's a great question and my boring scientific answer is it depends on what really you are wanting to produce these estimates for. We have, for instance, situations where there's a need for new estimates for understanding the amount of vaccine needed to be distributed in different parts of rural areas of the country and there, by producing estimates that we know are uncertain but have a confidence interval, have that uncertainty around it and we can understand the need to maximize the coverage to make sure that as many children as possible are vaccinated. And there having an upper level of uncertainty to say that we are based on the input data, based on the sampling approaches, based on the model uncertainty, we are pretty confident you won't find more than this many children in that area. And that can be a number an actionable number to use for that kind of approach. But yeah, there are a huge number of different ways to try and build trust and confidence in these data. So

with some national statistics offices we have produced estimates together and then gone out into the field together to go around different neighbourhoods and count the number of people as an independent source and then go back and see how well did that line up compared to what is currently being done which is like a cohort component projection from census 15 or 10 years ago. And so things like that build up slowly that trust and producing the data sets together with those decision makers who have that ingrained local knowledge and expertise on what is going on in those communities that becomes very important I think in terms of not just a statistical number but a real validation and ownership and trust in these kinds of data sets and methods but yeah it's a difficult one. Thank you.

Vissého Adjiwanou: Matthew. Underestimates come with credible intervals, confidence interval, and the wider where the model knows less. Yours come with a track record instead. So, let me ask you the harder question of the question, I think. What would an uncertainty interval even mean for estimated population? Where would it come from? And where when there is no sampling frame in your in your working?

Matthew Hauer: Yeah, no, this is a really good question. It's something that my research team grapples with on a daily basis, gold standard and otherwise. So, I think honestly, there's two different sources of uncertainty that exist in the vast majority of estimates that we make. Of course, sampling uncertainty is one of them, but there's also the design uncertainty that we have in the instrument that we're actually creating. Those also have a level of uncertainty. So, a simulation has design uncertainty just like, you know, a Bayesian network might or anything else. Why didn't you choose a Bayesian network versus anything else? Like, why didn't you choose something else? So, you have like the garden of forking paths or the multiverse in terms of the different decisions that we can make for producing a given estimate. And a simulation is just one of those paths, but it does allow us to explore lots of different paths that we could have as well. So, what about all the different specifications that we could have run and whether or not we're actually going to be robust to our given output.

Vissého Adjiwanou: Thank you. And thank you for following this presentation. Before we move to the Q&A questions, we have maybe 10 minutes here. I have a question for all of you, but I would just want you to respond just for 1 minute. Here is a talk that gave by Kramer. Kramer, yes. Who quote Kramer and colleague written in Nature. And he say he will not...Like everything that you said here is that you can trust the prediction may be the best thing that you have. If you are able to predict how many people are in the area, how many children will be dying the in the next time. But the question of understanding is very important for us as well as a demographer. Some criticism that we have is that we don't have theory. And so that in this area of AI, I look for the search for theory in demography. And so on. Where are we in the development of this AI? We have six

pattern that can emerge. Decision, exploration, data analysis, computational microscope, resource and agent of understanding. The agent of understanding is the most important for me in this particular talk that you gave here. If I go to Matthew, he said that what matter is the result. But like Kramer and other said, you will not be satisfied with the oracle itself. You want to comprehend how the oracle conceive this prediction. So, what is the most important? The understanding or just the result? Matthew.

Matthew Hauer: Oh, me? Yeah. Oh, no, this is a great question. So, what is most important, right? The actual results or understanding the results? So, we're in a very unique position with simulation. I can only speak of her for Aaru here and in Well, not for Aaru, but I guess as a conduit. If we were to be a perfect oracle and produce something that is an exact perfect prediction where the actual inference, the inferential statistic component is just pure noise, we wouldn't have any clients. We wouldn't have any customers because you can't do anything with the output. We have to actually have both prediction be correct and the inference part also be correct. So, in a classic regression equation where you've got your y s and your $\hat{y}$ s on one side and you've got your β s and your $\hat{\beta}$ s on the other side, we actually have to have both of those in balance or else a client is going to say, "Why did you predict this this given way?" And if we shrug our shoulders, we actually won't have a client. They'll be very angry with us. So, I think both of those components are extremely important, and simulation has to do both.

Vissého Adjiwanou: Ridhi, do you want to you think about that?

Ridhi Kashyap: Yeah, I think I agree that, you know, to some extent so, I want to make three points. First, I think we sometimes tend to overstate the difference between prediction and explanation vis-à-vis, and sort of the role of both in theory. I mean, in a way, a good theory is a good theory not just because it explains well, but also because of its explanatory potential can tell us something about whether something will happen in some conditions or not, right? So, in a way, a good theory is also about good prediction. And I think we sometimes tend to overstate that kind of dichotomy between prediction and explanation. To me, you know, good theories are trying to do both. And so, I think in a way I think what we need to do more or we need to do better is we need to think a little bit harder about, you know, how we can - and there is methodological development here to be done as well. Which I think will inform the work that population scientists do, you know, and how do we sort of interrogate the interpretability of these new approaches and models? And I think there's a lot more work to be done in in doing that. Both from the perspective of modelling, but also if we think about the social impacts of AI, right? If we're going to have large-scale deployment of AI type systems in our societies, within our institutions, then we also need to be able to ask questions of why they made decisions the way they did and audit them both from the perspective of transparency, but also from the perspective of justice and equity. So, I think we need

kind of auditing of these systems both from a measurement and modelling perspective, but we also need them in thinking about the social impacts and, you know, as AI enters our environments to think about it as also a kind of a social shift. So, I think I wouldn't want to overstate the difference between prediction and explanation. I think both are important for theory, and I also would like to say that I think, you know, we should care about, you know, interpretability and explainability. But that's also something that I think is important not just for these models as models, but also for them as systems within our societies.

Vissého Adjiwanou: Thank you. If anyone else want to jump on, jump in and say something? I'm going to ask the question that got the most up votes, which is by Corey Sparks. And he writes, "I'm very interested in synthetic data and have used AI to make some examples of different types of administrative data with statistically representative marginal distributions. I'd love to hear from the group about their opinions on generating more useful synthetic data that preserve more types of statistical associations. Yeah, go ahead Emilio first and then attempt and go to the direction.

Emilio Zagheni: Well, well, thank you Corey for this question. It's a very, very important question. Like there are many data sets that are sort of synthetic data sets that could preserve sort of like pairwise relationships, but when you put them together, they do not necessarily maintain some sort of multi-variate type of relationships. So, I don't have a good answer for this except that perhaps my sense is that the key is to understand for which purpose we want the data to be useful. So, if the synthetic data and coming up with a framework for like what is the purpose of this synthetic data. There could be a simple like reproducibility aspect of one type of analysis or there are some properties that may be that we may need to sort of maintain for this data set. This data set may sort of generate estimates of certain parameters that shouldn't be biased, or this data set that is synthetic should maintain a multi-variate relationships that are consistent with the original data across a certain set of variables at the very least. If the data is longitudinal, it becomes even more complicated because you also have to take into consideration the privacy aspects. Like you have a longitudinal data, you may want to maintain some of the longitudinal structures but also be aware that like if not done right, you may still expose certain types of data or individuals by creating these or synthetic database. So, it's a very complex problem to address. And my sense is that yeah, the first goal is to like set up some frameworks and then so like optimize for those variables that are most important for that specific application and then keep the other ones as a sort of second order type of importance that may not necessarily be met with a synthetic database. This will be my sort of first reaction.

Vissého Adjiwanou: And Matthew, probably want to add something to that because you use those data.

Matthew Hauer: Yeah, we make a lot of synthetic data. Like a lot of it. I think philosophically, what should synthetic data look like? It should look like it was a draw from a real data, right? Like it should theoretically be indistinguishable from real data. Which real data is very high dimensional. We don't know the data generating process. It's noisy. It's often times messy. So, when you're looking at something and you're trying to use AI to make synthetic data, it's just often times it's going to be very difficult to make something that looks like real data in my opinion with any level of satisfaction without having a lot of technical advancements in order to do so. Like Amelia said with the multivariate relationships, so I'll just use a simple example. The US decennial census (not the decennial) the ACS itself has 286 variables. So, you're looking at a 286-dimensional object. If you want to think about a draw from that if you were to create a new individual from there and assuming a binary relationship, so you have AB for all 286, the combinatorics is just ludicrous. It's two to the power of 286. There's not enough probability mass in order to actually, you know, pump into the tails of that distribution to create a synthetic data frame that would actually look like real US census data. Unless that data exists on a lower dimensional manifold, which is what the entire point of LLMs at this point rests on this idea.

So, if you were to make this, you'd have to go about figuring out what that low rank or low dimensional manifold actually looks like and even then, it's still generally very poor. And this is a constant problem with like the image generators. You can use something like MNIST, which is the handwritten data set that's been around for almost 30 years. And that's a very high dimensional 784 and yet even those synthetic generators don't make good digits. It's just very difficult.

Vissého Adjiwanou: Thank you both of you and for the questions. We are right on time and I know that some of you here have to answer but we answer in another session that we will have. There are one question that we didn't ask here is about ethics, but this question also we will move it to the next our next session in one week.