Finding the Unconnected Population: The First Step Towards Addressing Rural Connectivity

Emre Tepedelenlioğlu, Sudeep Srivastava

Facebook

César Hernández Pérez, Carlos Salle Moreno, Raul Garcia

Telefónica



Connecting people who lack internet access is one of the top priorities for Facebook and Telefónica, and we are working together on solving this challenge.

To that end, high quality population and coverage data are critical resources for addressing rural connectivity.

A primary objective of our collaboration is building intelligence around areas outside of internet mobile coverage: i.e. rural and remote areas where available information is scarce and outdated. Situating and quantifying the number of people who lack internet access becomes the first challenge of any set of data-driven tools that are aiming to:

1. ***Prioritize network deployments:*** *where to deploy what technology first, and*
2. ***Improve network performance:*** *through tools that monitor user experience.*

In this paper, we will introduce methodologies that help measure and locate two key outputs, namely:

1. *High resolution population density , and*
2. *Mobile coverage footprint.*

These datasets are then used to locate and count the number of people who lack internet access in a given country to get an accurate picture of the connectivity landscape.

Introduction

The Challenge of Connecting Unserved Populations

Extending mobile internet coverage is a top priority for both Telefónica and Facebook, and it is a challenge that many technology and telecommunication companies (telcos) are trying to solve together.

Internet connectivity has been recognized by the United Nations as one of the key sustainable development success factors [1]. Being connected to the internet leads to a better quality of life. A report by the World Bank estimates that for every 10% increase in high-speed Internet connections in a given country, there is a corresponding 1.3% increase in gross domestic product (GDP). But today over 4 billion people globally do not have reliable Internet access and are cut off from information on employment and education, in addition to social, governmental, and health services.

Bringing internet to unserved areas (which are mainly rural) will have a major impact on local communities, in particular, and on society as a whole. Internet connectivity would bring access to a wide range of services such as e-health, e-learning, e-commerce and online banking. These services have a great impact on the Human Development Index (HDI) and can benefit populations who are most in need.

Which Factors are Preventing the Deployment of Rural Connectivity?

High-quality population and coverage data are critical resources for addressing rural connectivity. Most people who lack internet access live in rural and remote areas, where the population is very sparse and where the terrain morphology (e.g., jungle or high mountains) makes it challenging or economically unsustainable to deploy and maintain network infrastructure. For instance, in Peru, the country that we will be using as benchmark for this analysis, at least 20% of the total population (6.3M/31M) lacks mobile internet access. They are spread over 77K different settlements (80% of the total settlements in Peru), averaging less than 100 people per town.

Accessing these places is often difficult. Deploying and operating telecommunications infrastructure in such areas requires significant investment in capital, logistics and work force. Due to these constraints, the return on investment (ROI) for deploying and operating rural networks is usually not financially viable for telcos, especially compared to urban areas, where covering large populations is typically more cost effective.

Additionally, today's cellular networks and technologies have mostly been focused on evolving to better adapt to large urban areas where there is a higher population density. This is not at all the case for most of the remaining unserved and underserved areas, where solutions need to adapt to cover large areas that contain fewer users. Deploying connectivity to rural areas represents a unique technological and economic challenge.

For these reasons, connecting remote and dispersed populations is not economically feasible at scale unless new technologies and operational models are developed and introduced to the market.

To address these barriers to rural connectivity, the industry needs innovation in the following areas:

- network (access and backhaul) technologies for low density population scenarios,
- new distribution models to commercialize the service,
- efficient operational and maintenance models for remote areas, and
- data-centric decision and operating models, enabled by data science to reduce risk and costs.

Finding Unconnected Populations: The First Step Towards Addressing Rural Connectivity

In this paper, we discuss a collaboration between Telefónica and Facebook, which is focused on data-centric decision models for network deployments. We are combining data from varied sources to better assess network deployment strategies in remote unconnected areas. In particular, we will present methodologies to measure and locate the people who lack internet access in a given country, using Peru as a benchmark to calibrate the model.

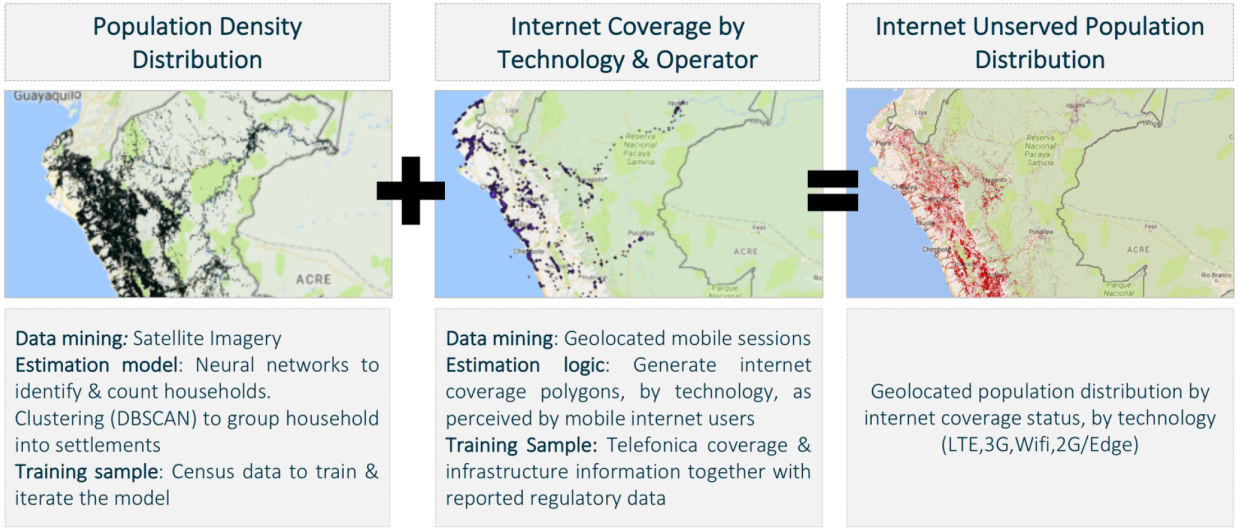


Image 1: Finding the mobile internet unconnected population - Methodology Snapshot

Deployment Criteria - ROIs and Data

Telcos have largely focused on connecting urban areas. The majority of their CAPEX is spent in high-density areas because that is where they see the greatest ROI.

Greater demand for fast connectivity, i.e. for video streaming, social networks, instant messaging, etc. has led telcos to focus on upgrading existing infrastructure in order to keep up

with customer needs and stay competitive. Most of the tools used to determine network planning & investment are made for urban areas and based on urban data.

These network planning tools are calibrated to maximize ROI in the existing network. To prioritize network deployments, telcos gather large amounts of data sessions records (XDRs) from all their sites, mainly 3G and 4G traffic information. They base their network deployment intelligence on this information. Sites that have higher traffic volumes take priority and are more likely to be upgraded. But XDRs (3G and 4G network session events) only apply for places where mobile internet connectivity already exists. Hence, traditional ROI metrics and a lack of broadband network sessions, leave places with no coverage or even 2G-only out of any budget prioritization picture.

Data-Driven Tools for Rural Areas

To address the limitations of the current deployment methods, the "*Internet para Todos*" program (Telefónica's R&D team working to connect unserved populations) has created new products and methodologies based on data for the unserved remote rural areas of its Latin American footprint.

As a part of this program, Telefónica has created the Rural Planner tool to streamline the development of new methodologies for rural deployments. This tool analyzes and produces different data sources that can be classified in three main categories:

- **Settlement information**: Location, population, HDI, energy accessibility, terrain morphology, etc. from public sources of any given country (i.e., National Institutes of Statistics).
    - Settlement definition: Cluster of permanent households in a given proximity of each other, identified with a single "settlement" name. Can vary from two to thousands of households. It is also the most granular measurement of populations where information is available, and varies from one country to another.
- **Coverage information**: Areas with voice and/or internet connectivity, type of technology (2G, 3G, 4G), quality of the service provided (e.g., average throughput, uptime vs downtime) Sources: Telefónica internal and Telecommunications Regulatory authority of any given country.
- **Network infrastructure**: Tower locations, tower height, closest Transport point available, fiber nodes, satellite availability, prediction of line of sight between towers etc. Sources: Telefónica internal information.

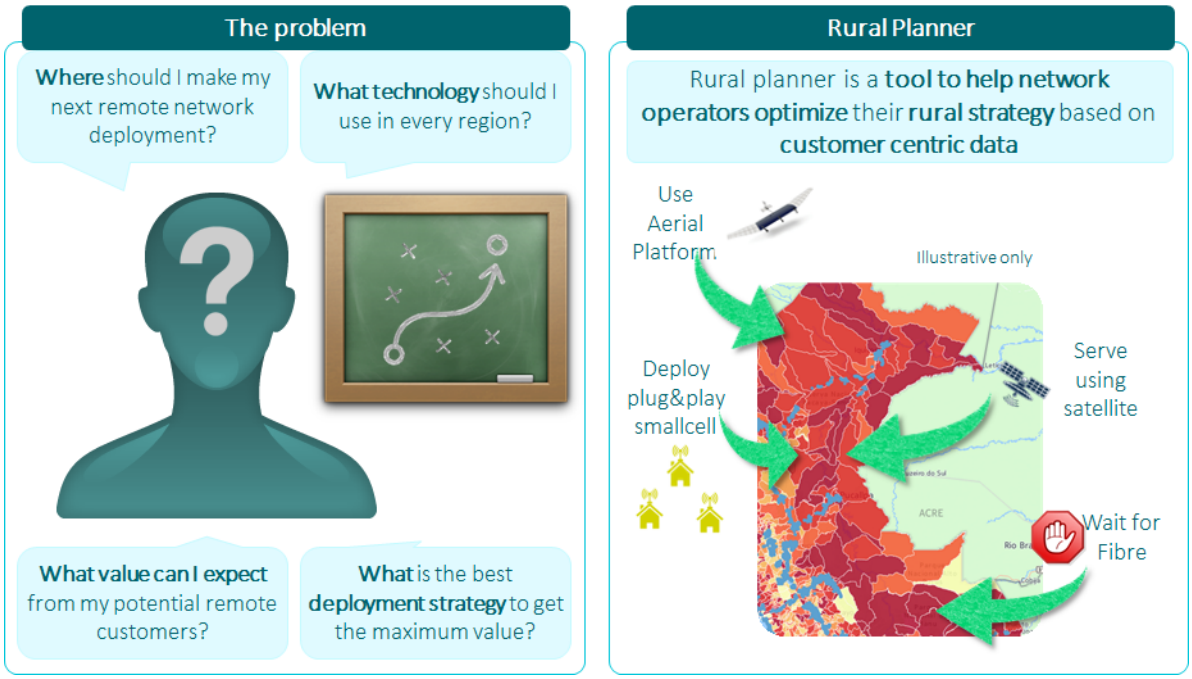
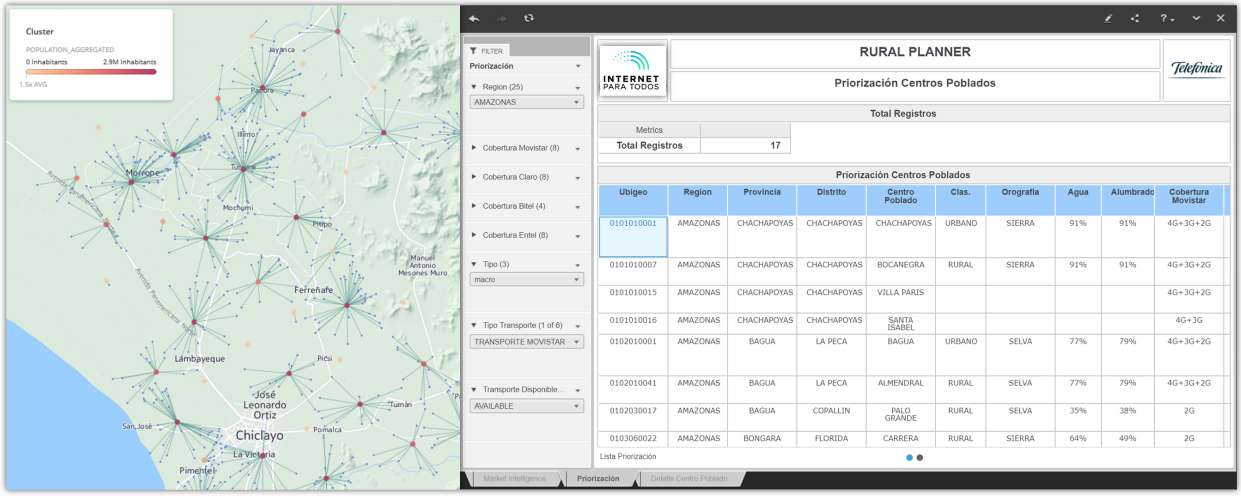


Image 2: Rural planner, to help business and network managers design their remote deployments

With all this information, Rural Planner aggregates the features at a "settlement" level and enables the end user to prioritize network deployments and assess the business potential of any type of deployment, e.g., competitor coverage hedge, low hanging fruit (minimum CAPEX) strategy etc. For instance, a decision maker could use this tool to answer queries such as, *"Where are the clusters with population of more than 2000 people that only have 2G coverage, with ARPU > $10 per month where fiber backhaul is available?"*

The picture below shows some snapshots of the interfaces used within Rural Planner.



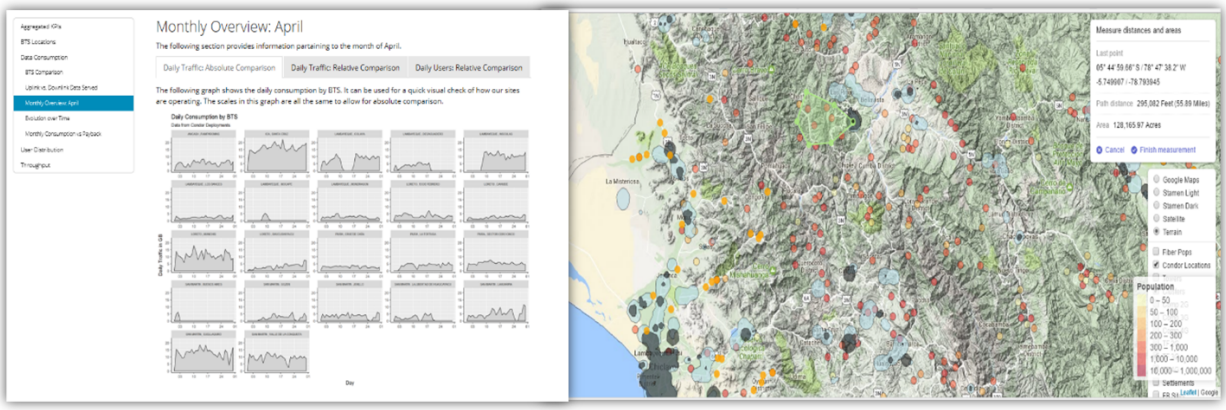


Image 3: Rural Planner's database sample interface

Key Areas of Collaboration Between Facebook and Telefónica

There are two key areas where Rural Planner can help improve connectivity, including:

- **Population Density**: Currently, it is difficult to accurately assess where people live with enough granularity. This is crucial when it comes to evaluating the priority level of any given deployment, or designing the right network solution.
- **Real Coverage:** To get a complete picture of the actual coverage people experience, we can use the information obtained from geolocated mobile sessions to get a more granular and neutral (user-based) view of the coverage footprint in any given country. This information would complement existing coverage information from operators, regulatory bodies and other public sources.

Both of these data points are critical inputs for any prioritization algorithm that aims to help make network deployment decisions to address rural connectivity. **Mapping population and coverage is the first challenge that Facebook and Telefónica have worked together to tackle**. To do so, we have used Peru as the benchmark country to create, calibrate and train our population and coverage models, so that they can then be rapidly adapted and used in any other country.

The **rationale for selecting Peru as a benchmark country for this study were**:

- Existence of granular population and coverage data to train our algorithms
- Peru has highly diverse topography, making it a good benchmark for several terrain scenarios (jungles, high mountains, deserts, coasts & islands).

Population Density

DATASETS

To improve the accuracy and reliability of Rural Planner's deployment recommendations, it is important to have the most granular population density information possible. The issue with this is that the census data that Governments offer often lack accuracy, they are not granular enough (one settlement refers to a very wide area) and they are outdated very quickly.

One of the reasons why this project started in Peru is the quality of its census data. The government publishes this information every 10 years, with an intermediate update every 5 to 7 years. The level of granularity of these datasets is high: Peru's most granular level of settlement can have a 1-5 people population and they can be 100 meters from one another. As a result, Peru has an information grid of 100,000 settlements, giving a very detailed view of how population is distributed across the country.

However, this is not the case for most countries, where census data is often not granular enough and is outdated. In order to solve this challenge, Facebook and Telefónica have come up with a methodology to provide this information without the limitations of public datasets.

How is isolated population detected without reliable census data?

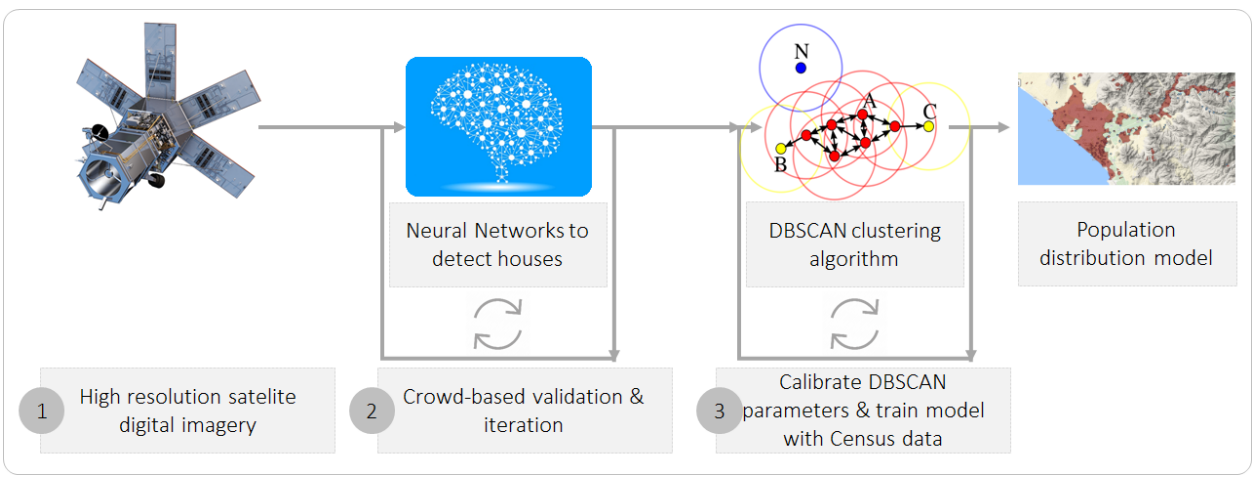


Image 4: Population distribution - FB & TEF methodology in a nutshell

Taking advantage of high resolution satellite imagery data, we enhance the current resolution of population to higher levels of granularity. We use 50 cm x 50 cm resolution satellite imagery to detect houses in areas of size 30m x 30m squares. We use a manually curated data set to train a convolutional neural network to identify whether a 30m x 30m region has a house or not.

We collaborate with the Center for International Earth Science Information Network (CIESIN)

to augment their census tracks with the predicted location of houses. Assuming that the number of houses is directly proportional to the population of a region, we generate population data at a resolution of 30m x 30m. This methodology gives us the opportunity to estimate the population of rural areas at a higher granularity as well.

Using the predicted locations of houses, we use a two-stage clustering technique to identify and create settlement clusters that best correlate with the available settlement data from the Peruvian Census.

The clustering is done in two phases. First, we use a standard clustering methodology called DBSCAN [2] to cluster houses using location coordinates. DBSCAN is a density-based clustering method which tends to cluster points which are packed together and marks points which are farther away in lower density regions as outliers. We do a second phase of clustering for the outliers. We have tried alternative parameter settings for the clustering algorithms to generate a set of reasonable settlements. Since we use settlements to make connectivity related decisions, a reasonable settlement is defined as one which can be correlated with an actual region in the country and also can be connected efficiently. If a settlement is too big we try to change the parameters to reduce its size. The tuning of the settlements was done via multiple feedback loops between Telefónica and Facebook.

In the figure below, we see three different algorithms used for clustering a single region. The left most figure uses a Louvain-based method while the rightmost figure shows a two-phase DBSCAN methodology to cluster the detected houses. The middle figure shows a DBSCAN clustering methodology followed by a CURE methodology to cluster. We can see the change in density of clusters and also the shape in the three different methodologies. For our purposes , we choose the two-phase DBSCAN clustering method since the clusters are of more reasonable size and align well within the Peru settlements.

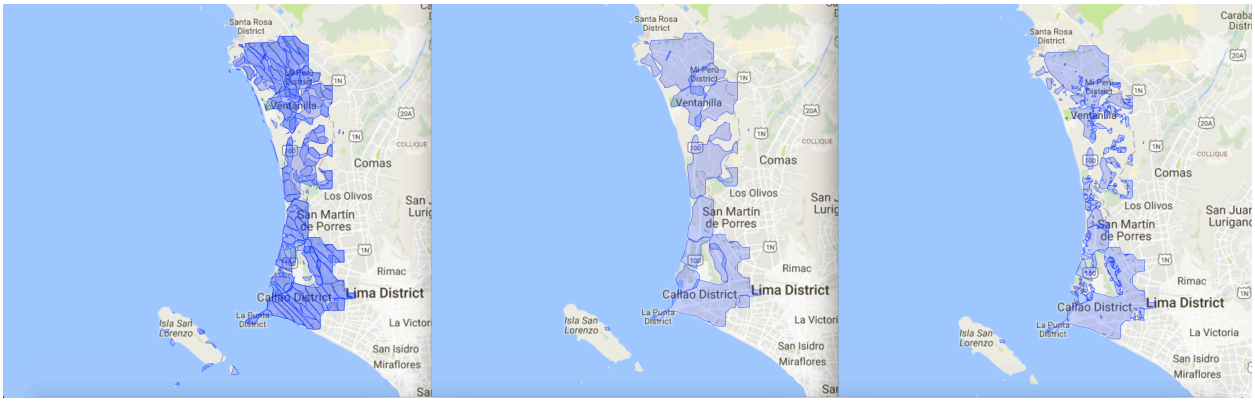


Image 5 - Alternative clustering methodologies tested for grouping houses into settlement clusters

VALIDATION AND CALIBRATION

In this section, we describe the methodology used to validate the choice of parameters in the clustering algorithms. The parameters determine settlement attributes, such as the size

of the clusters and the number of clusters in a region. To perform this validation, we define a population metric to compare the clusters with Peru's census data.

Peru's census data is more granular and detailed compared to other countries. Where the population is given for ~100k settlements that are identified by their UBIGEO codes, which is a coding system that INEI uses to uniquely identify each settlement that was counted in the latest census [5]. Roughly 40K of these settlements have a population of less than 10 people according to the census, of which ~15K has no population. This census data and the number of houses for a given settlement are used to calculate an estimate for the population. This data is used to validate and calibrate the population density obtained from satellite imagery. In 2018, there will be another census, at which point we will be able to re-calibrate the whole model.

We use the following KPIs to validate the model:

- % of population inside generated clusters (urban and rural)
  - Strictly inside
  - Within a given threshold from the closest cluster
- Characterization of the settlements not included inside a cluster
- Main variables that affect the detection's performance

The results are shown in the following tables and charts. These results correspond to the version of the algorithm that we considered the best. We iterated with different algorithms (e.g. Louvain, Cure and DBSCAN) and with different parameters for each algorithm.

	Population inside clusters - Global				
% of ubigeos inside polygons	% of ubigeos in a 3 km range to their closest polygon	% of ubigeos in a 5 km range to their closest polygon	% of population inside polygons	% of population in a 3 km range to their closest polygon	% of population in a 5 km range to their closest polygon
26.02 %	77.1 %	91.13 %	84.41 %	96.92 %	98.25 %

Population inside clusters - Rural	
% of ubigeos in a 5 km range to their closest polygon	% of population in a 5 km range to their closest polygon
87.09 %	95.08 %

Image 6 - Comparison of UBIGEO settlements with settlements detected via Facebook's algorithm

Around 75% of all the settlements from the Peru census tracks are within 3 km of a cluster detected within our predicted polygons, while around 90% of these settlements are within

5 km. 85% of the population is within the clusters, while 97% of the population is within 3 km of the clusters.

For rural areas, 87% of the settlements are within 5 km of the identified clusters, and 95% of the population lies within 5 km of the identified clusters. Since rural areas are sparsely populated and spread out, we use a higher cut-off for comparison.

Based on the above numbers, we see that the clusters generated closely mimic what we obtain from census tracts. Specifically, these datasets can be generated more easily and on a regular cadence compared to a census track, which takes a longer period of time to update.

False positives

Another important element to analyze is the number of false positives that the detection algorithm generates. We identify false positives when the algorithm identifies clusters where there are actually no houses.

To do so, the first step is to check how many 'empty' clusters we find. To make sure that the clusters are not only empty, but far away from any settlement defined in the census, we set a threshold of 5 km. This means that we are looking for clusters whose closest settlement (as defined by Peru's census) is at least 5 km away.

With this definition, 3.95% of the clusters are classified as 'empty'. We are working under the assumption that the census data describes the population distribution accurately. However, through visual inspection we identify houses/buildings in clusters which are not included in the census tracks. An example of this is shown in the figure below.

Image 7: Polygon with no census point nearby (5 km)

Image 8: Zoom into the polygon area with some houses.

Therefore, the clusters obtained from satellite imagery might capture a more updated scenario and the number of false positives might be lower than the estimated number. This is why we consider that the number of false positives will vary from 1-3% based on the visual analysis.

Characterization of the undetected settlements

We defined the settlements that are farther away than 5 km to the nearest settlement in Facebook's settlement dataset as undetected. Most of the undetected settlements are very sparsely populated, around 70% of the these settlements have a population less than 25 people. The plot below shows the cumulative population as a function of distance, note that ~28M of the 31M people living in Peru, according to the census, are within 5 km of the centroid of a settlement Facebook's algorithm has detected.

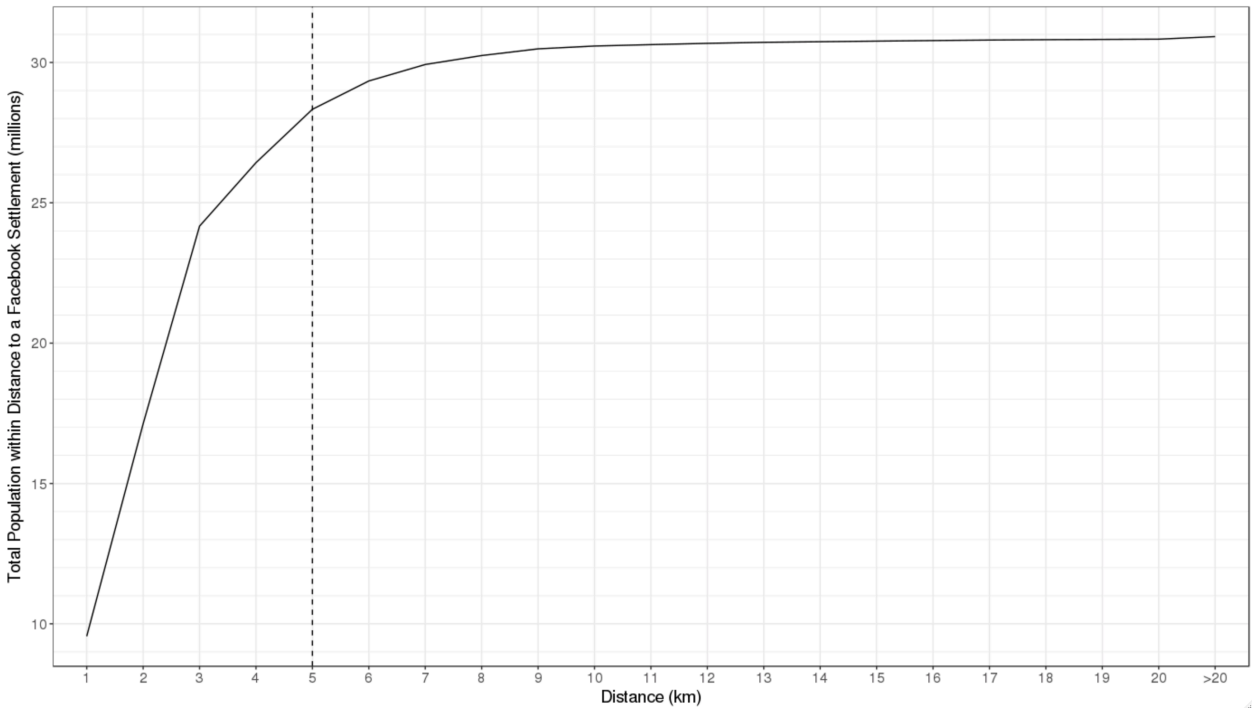


Image 9: Cumulative population of UBIGEOs as a function of distance from the nearest cluster detected via Facebook's population density analysis.

It is important to note that the performance of the methodology is dependent on the terrain morphology. Utilized satellite imagery quality varies with the orography and weather (e.g., cloud cover). Jungles and mountains are one of the most challenging terrains in terms of the percentage of undetected settlements, compared to coasts and desert. We have identified that most of the false positives appear in mountainous areas.

The following figures show the locations of the settlements from the Peruvian census vs the estimated clusters from Facebook. A good fit example of our jointly calibrated model.

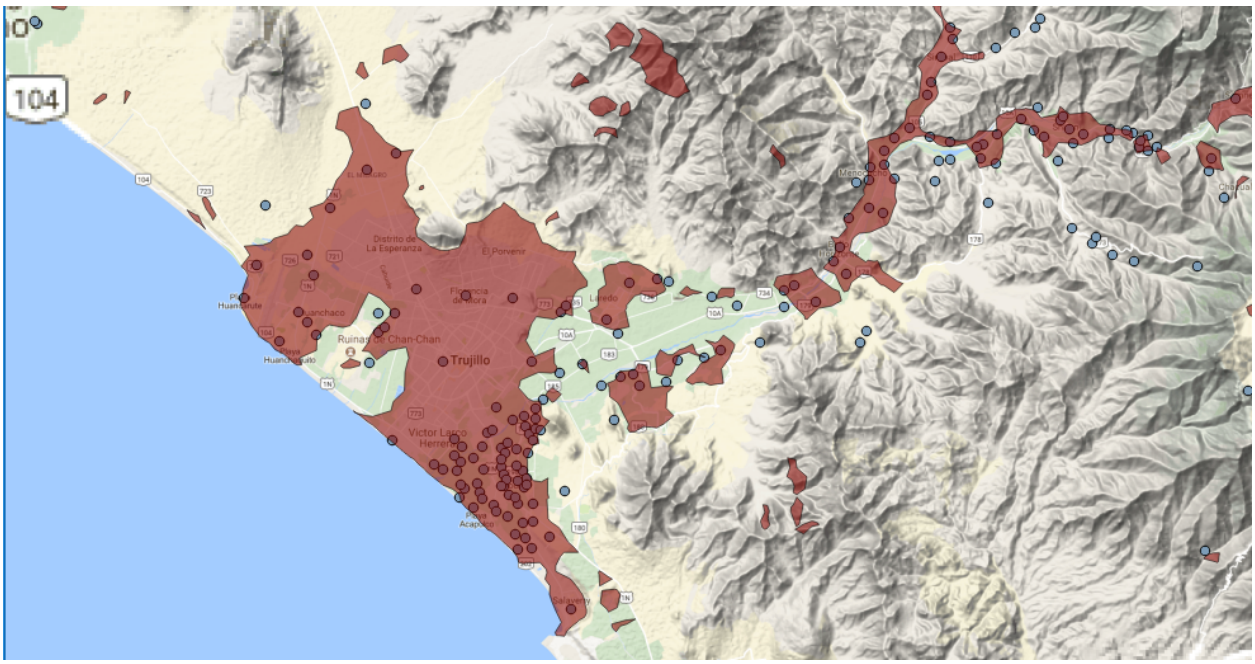


Image 10: Census settlements (dots) and the clusters

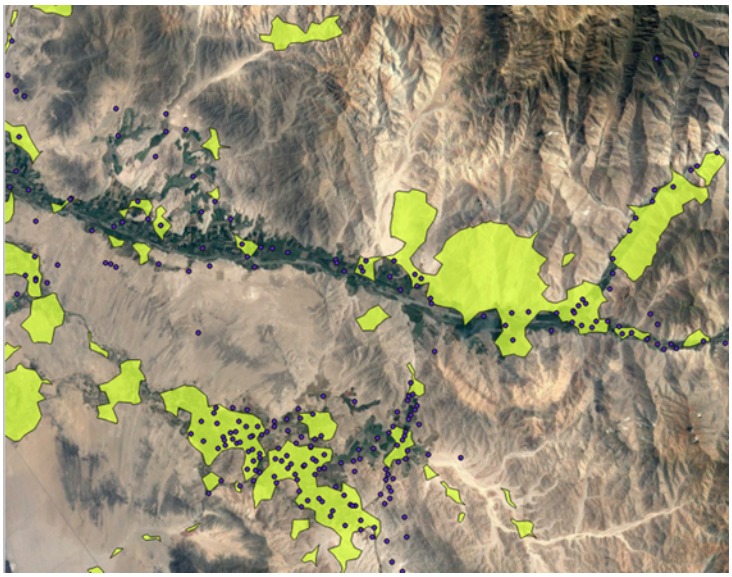


Image 11: Detail on spatial distribution for a given area. Many settlements are close to the edge of a cluster

Conclusions and limitations:

We have found a methodology to distribute population (detect houses) using high resolution satellite imagery data. Using available census data from Peru, we have calibrated the results of the clustering algorithm with the available census data.

We believe that, for Peru, the correlation we found between the population distribution data set based on census and the resulting data from our developed algorithms is sufficiently ac-

curate, since it can explain:

- 88% of the settlement sample (88% of the settlements provided by census within a 5 km range from its closest polygon).
- 95% of the population sample (95% of the Peruvian population provided by census within a 5 km range from its closest polygon).
- With less than 3% of false positives found. The algorithm doesn´t create more clusters than it should.

However, the methodology has some limitations, including:

- **Rural vs. urban**: It is hard to calibrate a unique clustering algorithm that suits all scenarios equally. A rural vs. urban clustering based on house proximity could be a good compromise to improve the fit between the predictive model and reality (training sample).
- **Imagery data**: The high resolution satellite imagery data is not always available with sufficient resolution. In jungle orographies, our population distribution predictive model works worst simply because there are not good enough images (e.g., clouds, foggy, rain fades). This is why we have identified that most of the worst predictions fit within jungle territories.
- **Census mis-geolocation**: Our training data sample for Peru is not perfect. In some cases the settlement coordinates are not 100% accurate, which is the reason why we consider a result where the settlements are within a 5 km radius from the predicted cluster a "good enough fit." Also, since the last Peruvian census was performed in 2007, it might be that some settlements have moved or disappeared (this is a common thing in some orographies, such as the jungle, where changes in river paths make people move their houses often).

Coverage

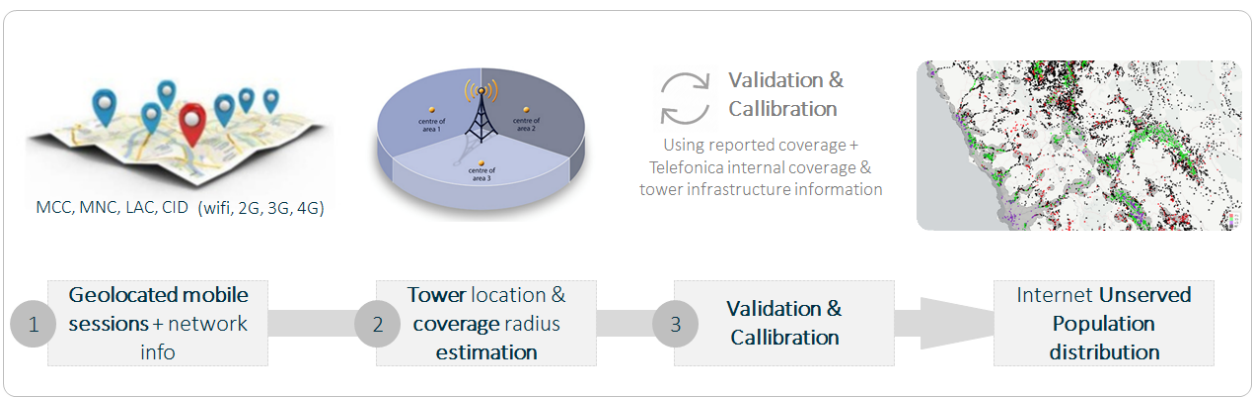


Image 12: Workflow diagram depicting how the estimated coverage was generated and compared to reported coverage

DATASETS

Availability is one of the main barriers to connectivity, together with affordability, relevance, and readiness [3]. National carriers in most countries have been evaluating the footprints of their own cellular networks (as well as those of their competitors) on a country level without knowing exactly how much of the population is left uncovered and where the unconnected are located. In other words, they do not have the sufficient spatial resolution to make their knowledge actionable. As discussed in the previous section, the population density dataset that Facebook has been creating and making available to the public for select countries [4], has been one of the most important advancements in understanding the dynamics of how people are living in relation to connectivity with a high spatial resolution. This insight is relevant and crucial to many industry participants, including internet companies, Original Equipment Manufacturers (OEMs) and Mobile Network Operators (MNOs).

Given that telcos are in a good position to enable technologies and business models that will address the problem of rural connectivity, it is worth looking at how they have approached this problem. Telcos have, historically, determined the places where there is coverage through reports published by the regulators. In most markets, all connectivity providers are obligated to report the places where they have coverage. Moreover, once they do this, they have to agree to a certain quality of service (i.e. SLA – Service Level Agreement). To make sure all telcos meet their SLAs and are true to their reported coverage, the regulator performs a series of tests in the reported locations. These tests consist of accessing a select set of mobile services (instant messaging, web browsing, etc.) and measuring signal strength in several different spots within the reported area.

The regulator publishes the information gathered with a certain cadence. The coverage report contains all the places with coverage for each telco operating in the country. In this way, telecommunications companies have a full picture of their footprint in the country and that of their competitors.

There are some notable issues with this way of gauging the coverage footprint of telcos in a given country. Not all telcos have the same policies when it comes to reporting their coverage. Some are very strict and only report coverage when they are certain that they meet all SLAs. Whereas others, may emphasize coverage footprint and report coverage in areas where quality of service might not meet their SLAs . This heterogeneity of what is implied by “coverage” makes it difficult to interpret the reported coverage.

Facebook and Telefónica have been working on finding a new way to estimate coverage footprint of telcos based on geolocated mobile sessions information from multiple data sources. The output of this exercise is an accurate coverage footprint for all network generations and carriers combined, objective, and closer to the real user perceived coverage.

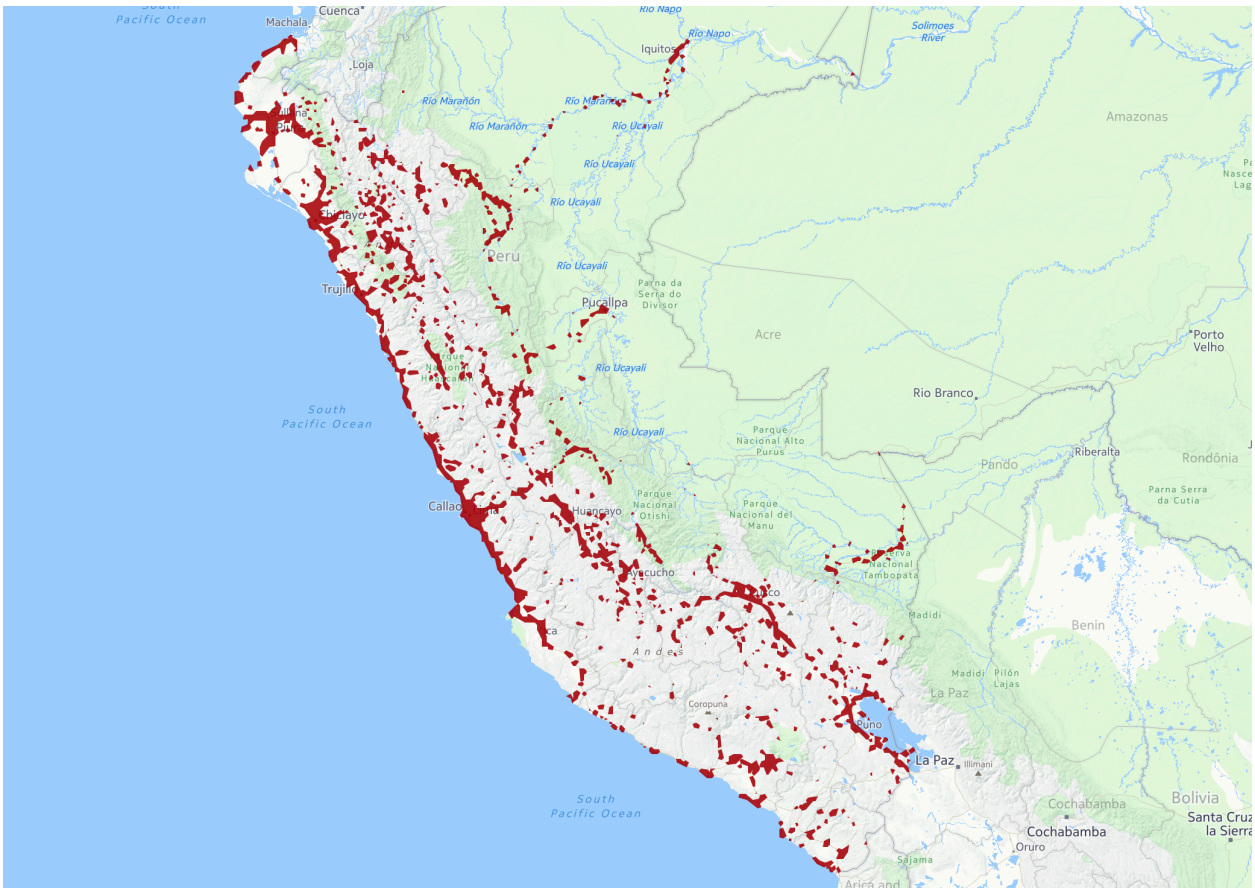


Image 13: The coverage footprint for all carriers and all network generations estimated for Peru

This methodology isn't without its drawbacks. For example, the coverage footprint at a micro scale (block level) will over-estimate coverage in certain directions and under-estimate in others. This over/under-estimate is a function of the network generation the coverage is estimated for, and is on the order of few kilometers for 2G to hundreds of meters for 4G. This skew in the coverage footprint is not consequential as we go to a macro scale.

Given that in rural areas the main form of access to internet is through cellular networks, the cellular network footprint has become more and more relevant to rural connectivity programs. Through this coverage footprint data we are able to get a more detailed and accurate view of both urban and rural coverage. Combining this data with settlements provides a way to segment the population that takes into account the financial aspects of connectivity, which helps in choosing which business model to use for a given segment.

VALIDATION AND CALIBRATION

In addition to the dataset described in the previous section for this work, we had access to the reported coverage for Peru for each UBIGEO. The table below compares this dataset to the dataset Facebook procured by combining population and coverage data (as described in this work):

	A	B	C	D	E	F	G	H	I

	A	B	C	D	E	F	G	H	I
1	2G Covered	3G Covered	4G Covered	# of Settlements (UBIGEO)	# of Settlements (Facebook)	Population by Reported Coverage	Population by Estimated Coverage [6]	% of Population by Reported Coverage	% of Population by Estimated Coverage
2	FALSE	FALSE	FALSE	66,009	9,420	3,370,239	7,055,316	11%	22%
3	FALSE	FALSE	TRUE	24	59	5,515	41,549	0%	0%
4	FALSE	TRUE	FALSE	7,210	1,210	631,530	1,947,919	2%	6%
5	FALSE	TRUE	TRUE	41	309	13,219	463,662	0%	1%
6	TRUE	FALSE	FALSE	13,129	832	2,883,566	608,628	9%	2%
7	TRUE	FALSE	TRUE	28	32	12,163	18,723	0%	0%
8	TRUE	TRUE	FALSE	10,984	911	3,486,621	1,640,858	11%	5%
9	TRUE	TRUE	TRUE	2,133	2,663	20,634,047	19,598,884	66%	62%
10	**Total**			**99,558**	**15,436**	**31,036,900**	**31,375,537**		

Image 14 - Comparison of coverage numbers for the two datasets

The new approach indicates that there are more unconnected people than the previous methods suggested. Here we define a settlement connected if >80% of the settlement is covered by a given network technology. Even in the most conservative case (e.g., defining a settlement as connected if >0% of it is covered), the unconnected population is slightly larger than the previous method.

The below snapshot from the north-eastern region of Lima, shows that the estimated coverage footprint described above agree well with the reported coverage except in some areas where the reported 2G coverage is not detected. In the map below, the grey area is the estimated coverage footprint and the black dots are the UBIGEO settlements that are reported to have no coverage. We have removed all settlements with populations of less than 10 people. The colors denote the highest network technology reported in a particular settlement.

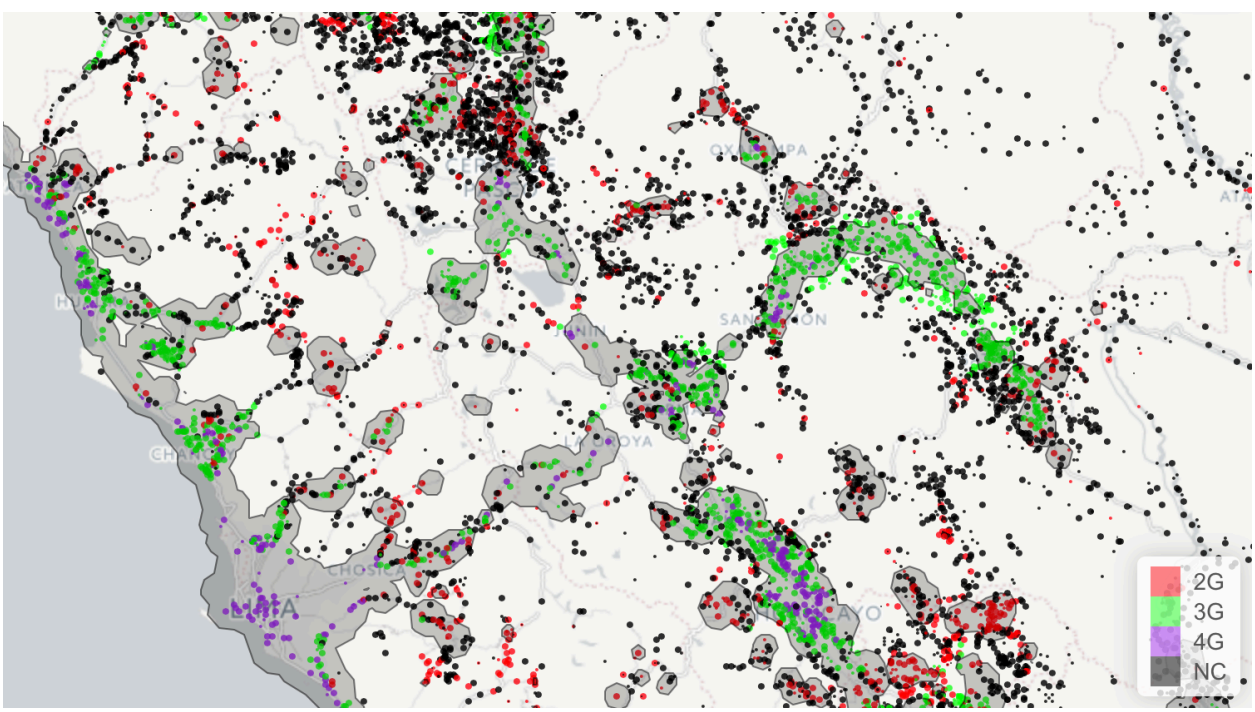


Image 15 - Comparison of coverage - UBIGEO settlements and the derived coverage footprint

Initial results

The methodology explained in this paper is already being used in the wider scope of the collaboration between Facebook and Telefónica in Peru, providing useful and accurate information to guide better (faster and cheaper) decision making.

The first use of the data set was to estimate the mobile internet unserved (no coverage at all) population distribution in Peru. The number of people that do not have access to the internet based on Telefónica's previous estimations, was around 6.3M out of 31M population (20% of total) whereas this number is around 7.6M people (25% of total) based on the work outlined in this paper.

Another joint use case has been finding internet access within mobile internet unserved areas as a proxy to find and prioritize pilots with WISPs (wireless internet service providers) to test new network technologies & operating models.

Moreover, the resulted population distribution, is now more granular and accurate. This is proving extremely helpful to:

- Estimate project ROIs with less uncertainty,
- Pick the right locations for specific pilots and real network deployments, and
- Design the right network architecture solution for the selected location, avoiding designing errors which might cause bad performance while saving in the required number of onsite field visits.

Conclusions and next steps

In order to bring mobile internet connectivity to rural and remote areas in Latin America, we need to use data science to deploy the right technologies in the right areas, with the best operating models for that area. The two main elements needed to build this intelligence are **population distribution/density** (where do people live?) and **coverage information** (where can we find mobile internet coverage?). When these two datasets are accurate and reliable enough, it makes direct impact in peoples lives by introducing in most cases for the first time internet connectivity to their neighborhoods.

The use and validation of this methodology in Peru allowed Telefónica to map population and coverage status of remote rural areas much faster and more precise, compared to other parts of Latin America. This has a major impact helping Telefónica lower the risk for future deployments that will extend the existing coverage by providing a standard and robust methodology to estimate population and prioritize deployments; and accurately design the adequate network architecture.

The next steps will include:

- Validate the methodology for other countries and improve/adapt where needed
- Start using it in all the countries where Telefónica will continue to extend coverage
- Add new layers, such as quality of the service (in order to know not only where people are connected but also how good their experience is): Unserved vs Underserved
- Keep exploring new use cases for sharing data and knowledge

References

[1] UN Sustainable Development Goals, Goal 9 (8 of 8): "Significantly increase access to information and communications technology and strive to provide universal and affordable access to the Internet in least developed countries by 2020"

[2] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231

[3] State of Connectivity 2015, A Report on Global Internet Access, internet.org by Facebook

[4] https://ciesin.columbia.edu/data/hrsl/

[5] https://en.wikipedia.org/wiki/UBIGEO

[6] The population given here is based on the GPW v4 (http://sedac.ciesin.columbia.edu/data/collection/gpw-v4) and is the 2015 UN adjusted population estimate.