Using Mobile Phone Data to Make Policy Decisions

A study in how new data sources optimized health facility placement in Malawi

COOPER/SMITH

dial Digital Impact Alliance

ACKNOWLEDGMENTS

This technical report was prepared as part of a partnership between the Digital Impact Alliance (DIAL) and Cooper/Smith. Primary contributing authors include Erwin Knippenberg, Rachel Sibande, Emily Chirwa, Tyler Smith, Senthil Kumar Subramanian, Arbind Kumar Sinha & Syed Raza.

ABOUT DIAL

DIAL aims to realize a more inclusive digital society in emerging markets, in which all women, men and children benefit from life-enhancing, mobile-based digital services. A partnership among USAID, the Bill & Melinda Gates Foundation, the Swedish government, and the United Nations Foundation, DIAL helps accelerate the collective efforts of government, industry and NGOs to realize this vision.

DIAL is staffed by a global team and is guided by a board of leading emerging market entrepreneurs, technologists and development experts. With this leadership, DIAL is uniquely positioned to serve as a neutral broker, bringing together government, industry and other development stakeholders to promote new solutions to old problems. For more information about the Digital Impact Alliance or this technical report, please visit our website: www.digitalimpactalliance.org

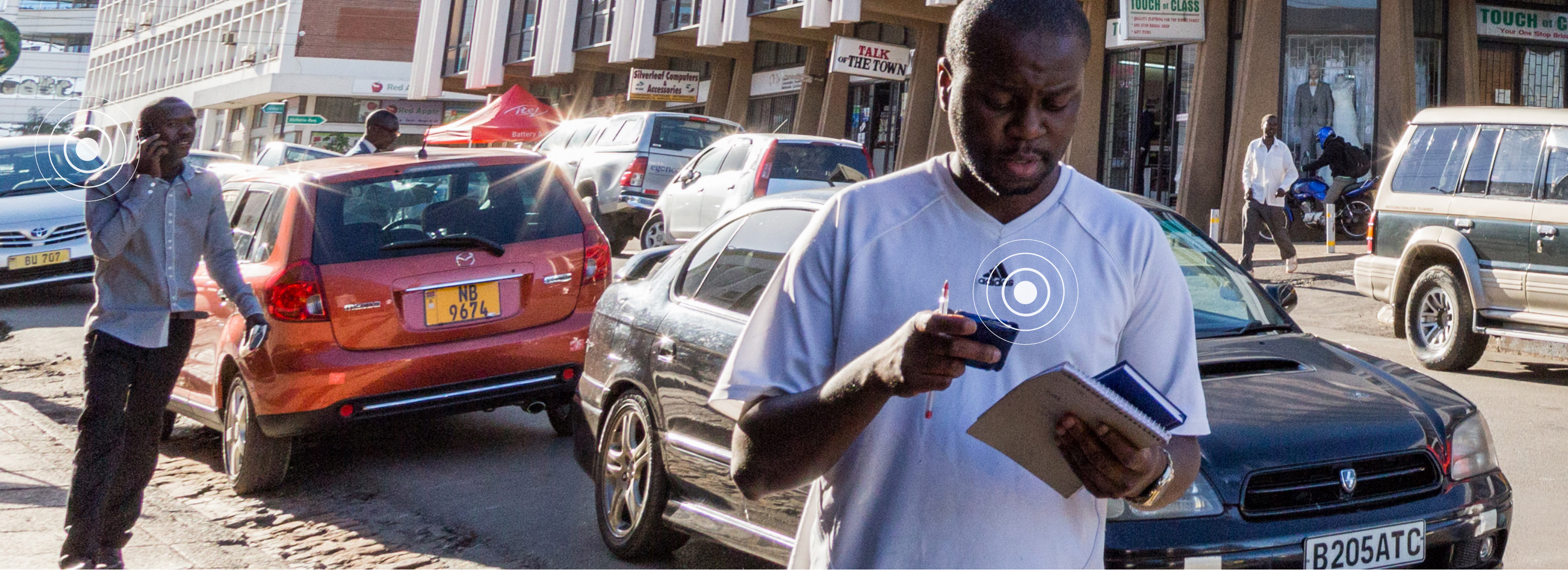
ABOUT COOPER/SMITH

Cooper/Smith is a technical assistance organization that uses hard data to increase the effectiveness and efficiency of development programs worldwide. The organization brings deep knowledge and expertise in strategic planning; health financing and evidence-based resource allocation; digital health systems implementation; operations, behavioral, and clinical research; as well as application of advanced analytic methods, including econometrics, modeling, and machine learning.

We support large-scale data systems and data use projects in Malawi, funded by the Bill & Melinda Gates Foundation and the Digital Impact Alliance. We support other programs globally, and locally within Burkina Faso, Cameroon, Kenya, Liberia, Madagascar, Nigeria, Senegal, Thailand, and Zambia. Former and current partners include Georgetown University; the London School of Hygiene and Tropical Medicine; the Boston Consulting Group; Catholic Relief Services; Canada's Minister of International Development; the Center for Strategic and International Studies; the International Fund for Agricultural Development; the Global Fund to Fight AIDS, Tuberculosis, and Malaria; the World Bank, and the Global Financing Facility in Support of Every Woman, Every Child. We participate in numerous technical and advisory groups and are committed to the democratization of data access and use worldwide.

For more information, please visit our website at www.coopersmith.org

TABLE OF CONTENTS



EXECUTIVE SUMMARY

> **This technical report demonstrates how MNO data can be used as a public good to inform policy and decision-making by governments and their agencies.**

This technical report demonstrates how Mobile Network Operator (MNO) data can be used as a public good to inform policy and decision-making by governments and their agencies across various development sectors. In this use case, we demonstrate the use of MNO data to understand population density and migration patterns of people and how such insights can inform governments where to deploy services such as water, schools, health posts and other service points.

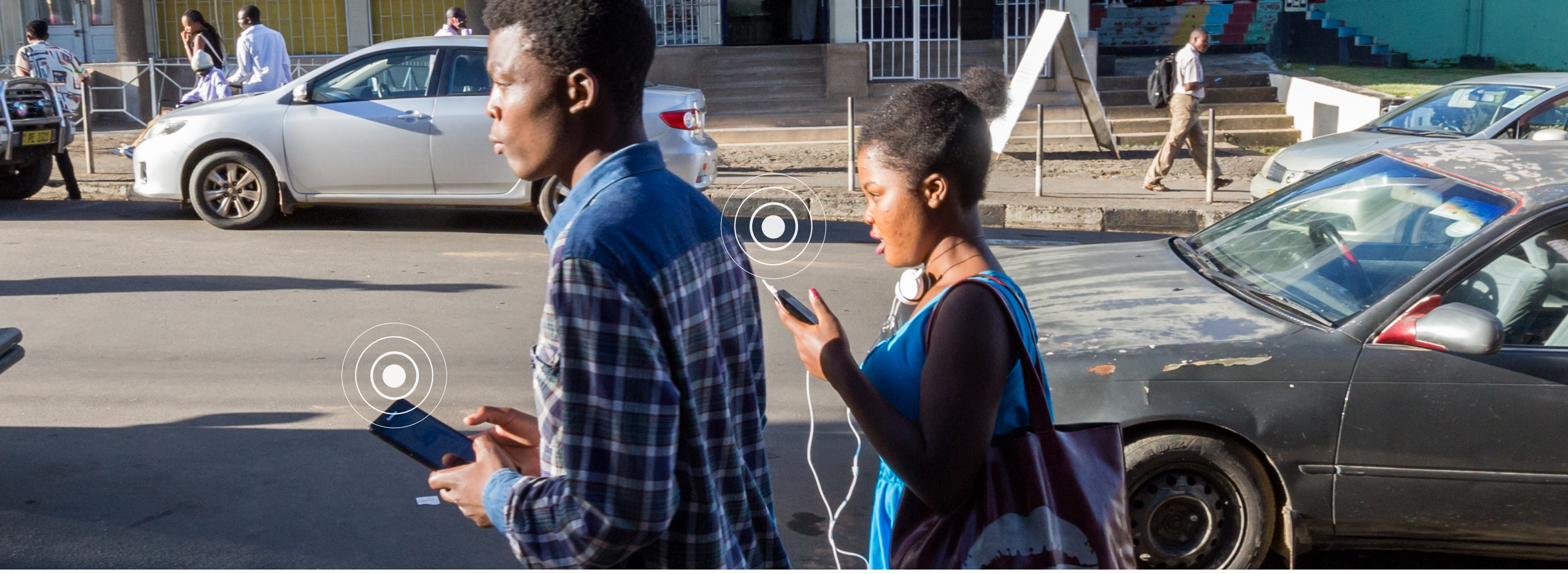
In this technical report, we explore the use of MNO data to deliver an optimized allocation model for 900 new health posts using MNO data. The model that will ultimately reduce the proportion of Malawians without access to health care from 44.7 percent to just 5 percent by 2023. By using MNO data rather than relying solely on census data, government officials can better predict population movements in certain districts after potential flooding, and thus make more informed decisions about where to place future health posts.

Targeted placement of health posts is key because providing widespread access to primary health care remains a challenge in many developing countries. In Malawi, an estimated 7.73 million people, or 44.7 percent of the population, live more than 5 km from a health facility, which severely hampers their access to essential health care services. Malawi's population is projected to grow from 17.6 million in 2018 to 21.6 million in 2023. Without action, the number of Malawians without access to a health facility would grow to 9.7 million by 2023.

To remedy this, the Malawi Ministry of Health created a Capital Investment Plan (CIP) that proposes to build 900 new health posts between 2020 and 2023. The goal is to locate these new health posts strategically to ensure that 95 percent of the population lives within 5 to 6 km of a health facility by 2023. Additionally, the plan aims to increase resilience during the rainy season by allocating more health posts to areas where flooding could cut off access and where large numbers of people migrate to during these periods.

In order to get a clearer picture of population density and movement patterns throughout the country to properly place the health posts and maximize the efficiency of allocated resources, the Ministry of Health worked with development partners to integrate mobile network operator (MNO) data into the CIP it was developing. The proposition was that anonymized call data records (CDRs) and unique call density could be used as a complement to national census statistics to better reflect both seasonal migration patterns and long-term urbanization. The optimized analysis would then be integrated into the CIP as a technical appendix, informing the allocation of new health posts. This offered a concrete, policy-relevant use-case for the use of MNO data to inform policymaking.

To ensure sustainability and replicability, the authors engaged with government partners to integrate the analysis into country systems and provide a ranked list of priority locations for new health posts. The authors will also work with government partners in Malawi and across the globe to illustrate how such data for development models can be made replicable across development sectors and geographies. Insights on population density and movement patterns can also inform locations of water points, schools, agricultural cooperatives and improve people's access to other critical programs and services.



BACKGROUND

New data sources such as mobile phones and geographic information systems (GIS) are being widely used in the developed world for commercial and public service purposes. And while they have been experimented with in the developing world, they have not yet been incorporated as routine practice in development situations where they could provide the most benefit. The Digital Impact Alliance (DIAL) and its partners are working to increase the use of these new types of data in the developing world for humanitarian purposes by identifying use cases across health, agriculture, education and other areas.

In 2017, DIAL, Cooper/Smith and Infosys, in collaboration with the Malawi Ministry of Health (MoH) and Malawi Communications Regulatory Authority (MACRA), embarked on a collaborative effort to demonstrate the value of mobile network operator (MNO) data project to inform governmental policy and decision-making. The Ministry its of Health solicited input from the DIAL project consortium for Plan its Capital Investment Plan (CIP), which included a proposal to included build 900 new health posts between 2020 and 2023 as part of its emphasis on expanding access to primary health care. A health-post is staffed by a clinician or community health worker and stocked with essential medicines in order to provide frontline care.

The Ministry of Health solicited input from the DIAL project consortium for its Capital Investment Plan (CIP), which included a proposal to build 900 new health posts between 2020 and 2023 as part of its emphasis on expanding access to primary health care.

While the CIP includes investments in secondary and tertiary care, the government emphasized that expanding access to primary health care facilities such as health posts is its priority. At that time, 44.7 percent of Malawians lived more than 5 km from any health facility, leaving 7.73 million people without access to primary health care services. Unless the government acted, that number would grow to 9.7 million. The Ministry's goal was to optimize the allocation of these health posts so that 95 percent of the population lives within 5 to 6 km from a health facility by 2023.

The problem statement was presented as follows: To infer population movement patterns using anonymized, aggregated MNO data as a proxy to understand population densities, migration and urbanization patterns. Such data complements a national census in that it allows for dynamic tracking of population movement in a timely and cost-effective manner, which does not require expensive on-the-ground surveys. As a policy-relevant use case, the analysis develops a model forecasting migration patterns and optimizing the placement of health facilities in line with national priorities. For the purposes of this exercise access to any health facility is considered sufficient, with health-posts as the most cost-effective means to expand access to primary health facilities.

LITERATURE REVIEW

The background literature draws from three types of research:

With the increased availability of high-resolution satellite data combined with powerful algorithms, there have been sustained improvements in forecasting population density (Stevens et al., 2015). The WorldPop project, an open source collaboration between researchers, trains its algorithms on historical census data and uses it to project annual population density at 100-meter resolution (WorldPop 2018).

This technical report intends to bring together insights from the literature to inform a policy-relevant use case that harnesses MNO data to identify current and future gaps in the availability of health services, drawing up recommendations to remedy those gaps.

Complementary to this approach, researchers have demonstrated the potential to harness mobile phone data to map population movements dynamically. Deville et al. (2014) showed that the density of unique users in a cell tower's catchment area corresponded closely with population density. Therefore, researchers can extrapolate to predict shifts in population densities between day and night, weekdays and weekends, and across seasons. Using Portugal and France as case studies, they showcase significant shifts in population density, enough to impact service provision. Erbach-Schoenberg et al. (2016) show how similar methods can be used in the context of developing countries to account for seasonal fluctuations in calculating disease incidence. Working with mobile phone data in Namibia, they highlight how seasonal mobility affects estimates of malaria incidence, leading to differences of up to 30 percent compared to estimates created using static population maps.

Finally, research focusing on service provision has sought to link mobile phone data to development outcomes. Blumenstock et al. (2015) use mobile network data and machine learning algorithms to predict poverty outcomes in Rwanda, identifying hotspots with a high degree of precision. In the domain of health, Wesolowski et al. (2015) demonstrate the relationship between travel distance, as calculated using a household's radius of gyration,[1] and the percentage of households not receiving antenatal care. Travel distance remains a salient issue when it comes to the provision of basic health services.

This technical report intends to bring together insights from the literature to inform a policy relevant use case that harnesses MNO data to identify current and future gaps in the availability of health services, drawing up recommendations to remedy those gaps.

[1] The radius of gyration is the distance from a central point defined as the center of inertia. In this context, it is the assumed distance a household can travel on average given its pattern of locations identified using MNO data.
[2] WorldPop uses geospatial data calibrated on census data to create high-resolution (30 m by 30 m) maps of population distribution. See http://www.worldpop.org.uk/.

DATA COLLECTION

Identifying Data Sources

The authors (DIAL, Infosys and Cooper/Smith) drew from existing literature to develop an analytical plan that sought to combine three streams of data:

1) Anonymized MNO data provided by an MNO partner, including:
   a. CDR data (2016-2017)
   b. Geo-tagged location of each cell phone tower
2) High-resolution population density data for 2015 compiled by WorldPop[2]
3) Location and catchment area of existing health facilities, provided by MoH in collaboration with UNICEF

The analysis proposed to map the MNO data to administrative units, calculating the density of unique callers, then calibrate this against population density using WorldPop data. Fluctuations in the density of callers can then be extrapolated to infer population migration trends.

Establishing Data Pipeline

In parallel, the authors worked to establish a viable pipeline to import MNO data to a secure in-country server. To ensure proper anonymization of the MNO data, Infosys provided training to MNO technical staff.[3] A total of 26 months of data representing every call and SMS sent and received in Malawi between January 2016 and May 2018 were transferred to a secure in-country server.[4] Access to the anonymized raw data was restricted to the server system administrator, with a small group of additional technical users given access only to the processed data.

Once stored on the secure server, the data was cleaned for analysis. In the process of cleaning, the authors flagged several issues concerning data quality, including missing observations. Addressing these concerns around data quality and preparing the data for analysis proved to take more effort and time than initially anticipated. The authors had to iterate several times with MNO counterparts to ensure the proper transfer and formatting of complete datasets, leading to delays. This was a useful lesson learned, highlighting important logistical constraints that must be tackled to implement such a project.[5]

---

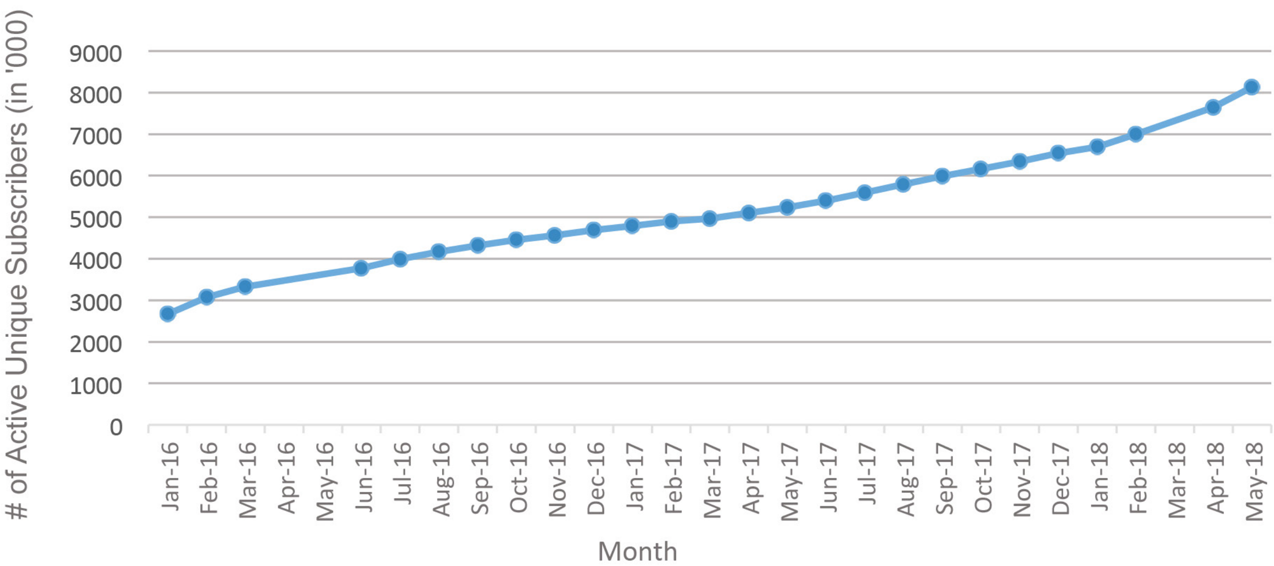
[3] The phone numbers were anonymized using a cryptographic hash function, SHA 256 algorithm. These functions are collision resistant, in that they generate a unique output, and one-way, so they cannot be decrypted back.
[4] The data is over the span of 29 months (Jan 2016-May 2018) but three months are missing (April 2016, May 2016 and March 2018)
[5] This depends on a number of factors, as has been highlighted in a number of earlier DIAL publications, including here.

Data Compiling and Cleaning

When reviewing the data transferred from our MNO partner, several issues were identified requiring further cleaning and wrangling. Three months of data were originally missing. In addition, two months out of the 26 only had SMS data and were missing observations for calls sent and received. Our analysis trains on the whole dataset, on the assumption that these missing months are not anomalies. Trends in the growth of unique users suggest the authors can safely extrapolate from observed months. See Figure 1.

FIGURE 1

Rapid Increase in the Number of Unique Active MNO Subscribers



The data contain 12.9 billion records, including 23.4 million unique receiving numbers. Active unique users were defined as users having used their phones at least once in the past three months. The ratio between unique originating numbers and unique receiving numbers is high for two reasons: the originating numbers are from one MNO, while the receiving numbers could be from either of the MNOs active in Malawi.[6] Furthermore, in order to minimize information loss, there were no filters applied to account for application to person (A2P) calls and messages or for the practice of "buzzing," which is a brief phone call alerting the other person of the need to call back.[7] The authors have not applied any filters to the records, such as A2P calls and SMS. Therefore, there is a high chance that the ratio between the amount of unique receiving numbers to the amount of unique originating numbers will be high, especially for A2P calls. The data also contains unique location area codes (LAC IDs), and tower-level cell IDs, allowing the unique originating numbers to be matched to the nearest cell tower.[8] Furthermore, the data accounted for the number of calls and SMS separately, as well as call forwarding and roaming. See Figure 2.

[6] The two MNOs exist as a duopoly, with the one under consideration holding more than half of market share.
[7] The ratio between receiving and originating drops between 2016 and 2017, which may be tied to a government push to register all SIM cards, rendering inactive SIM cards defunct. This does not affect the analysis, which is built on the number of unique originating numbers.
[8] There were 7,025 unique cell tower IDs in the master list of towers and 6,743 unique cell IDs in the CDR data, including towers that entered and exited the dataset. There were 6,049 matching IDs across both datasets. These towers constituted the sample size in terms of coverage.

FIGURE 2

Rapid Increase in the Number of Unique Active MNO Subscribers

Raw MNO Data	2016	2017	2018 (Jan-May)	Total
# Records	5,004,750,666	5,960,543,159	1,934,707,383	12,900,001,208
# Unique originating numbers	12,261,549	16,762,626	8,821,575	23,400,290
# Unique receiving numbers	50,092,409	31,288,109	18,993,028	78,600,039
# Unique LAC IDs	39	54	54	66
# Unique cell IDs	3,926	5,562	6,648	6,743
# Voice calls	3,073,839,525	3,375,532,407	592,711,535	7,042,083,467
# SMS	1,927,785,317	2,582,155,890	1,341,477,160	5,851,418,367
# Call forwarding	1,278,911	1,996,888	429,349	3,705,148
# Roaming call forwarding	1,846,913	857,974	89,339	2,794,226

The unique originating numbers were matched to the nearest tower's latitude and longitude using the cell ID. Of the 6,743 unique tower IDs in the MNO data, 6,049 were matched to the roster of cell towers with geospatial coordinates. Catchment area was calculated using Voronoi polygons, which delineate the area closest to every cell tower, with a maximum range of 20 km in rural areas.[9] Accounting for missing observations, the pre-analysis concluded that signal from the observed towers could reach a land surface where an estimated 95 percent of Malawi's population lives. This was considered sufficient to proceed with the analysis. The density of unique users was calculated by dividing the number of observed users by the cell tower's catchment area.

To conduct the analysis, the above MNO data was combined with several Malawi-specific data sources. This includes 2015 population density at 100-meter resolution from WorldPop, calculated using satellite imagery trained on Malawi's previous census, collected in 2009 (WorldPop 2018).

UNICEF conducted an extensive survey of every operating health facility in the country and calculated the relevant catchment area. This catchment area constitutes the distance such that a patient would have to walk no more than 5 km to reach a health facility. In its analysis, UNICEF accounted for the road networks, topography and potential for flooding in calculating the 5 km catchment area for each health facility. It calculated a "best case," where there were no impediments to travel, and a "worst case," where flooding made certain roads and health facilities inaccessible.[10] The report used the best-case model unless otherwise noted.

In collaboration with the Malawi Ministry of Health, the analysis also incorporates facility-level disease burden data, as reported on a monthly basis to the ministry.

---

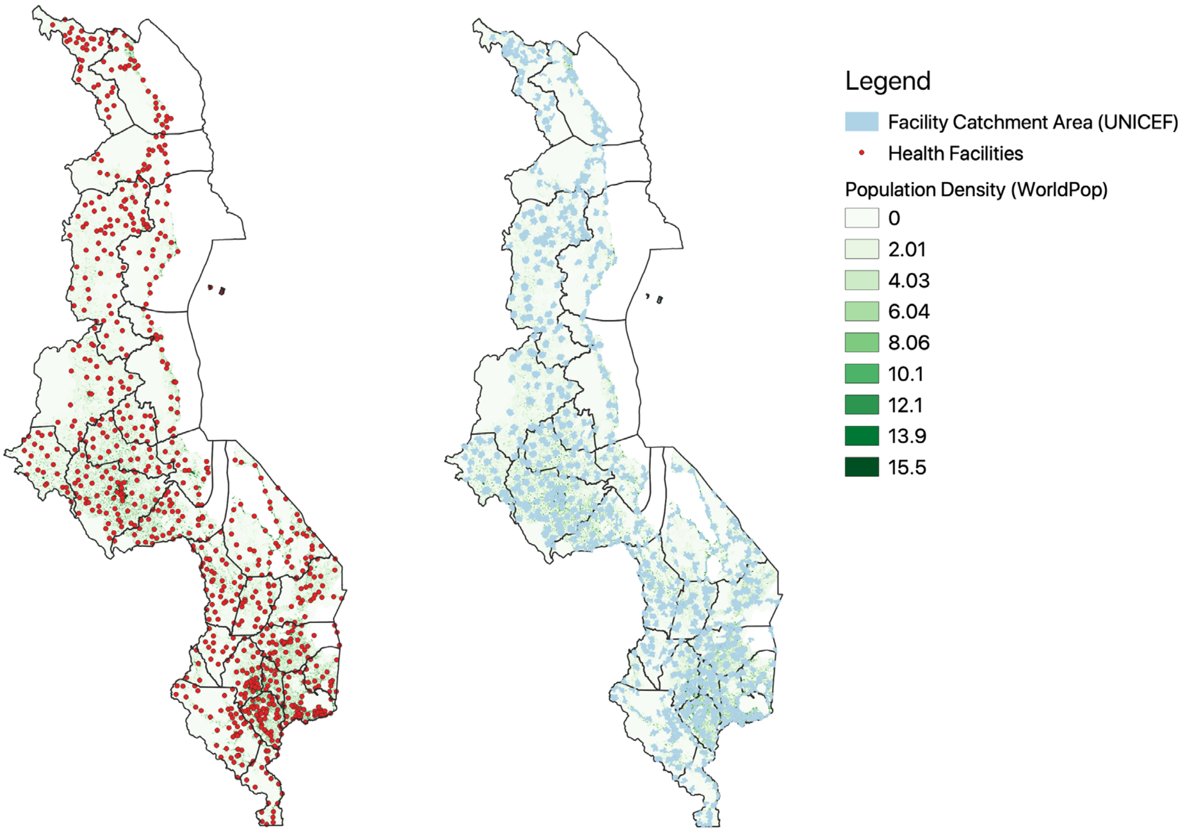
[9] Though the MNO provider shared the location of its towers, it did not provide a map of its countrywide coverage, requiring the authors to estimate this coverage instead.
[10] Special thanks to our colleagues at UNICEF for sharing these catchment area calculations with us.

GAP ANALYSIS

A first step was to establish an estimate of coverage gap in terms of health services at the district and Traditional Authority (TA)[11] levels based on facility catchment areas calculated by UNICEF combined with 2015 WorldPop population data. For the purposes of this exercise, access to any facility, whether primary, secondary or tertiary, was considered sufficient. While access to surgical and reference facilities are critical elements of a well-functioning health system, at the government's directive, the study focused on identifying and expanding access to primary care. See Figure 3.

FIGURE 3

44.7% of Malawians Live Outside the Current Catchment Areas



Overlaying these calculated catchment areas on WorldPop population distribution makes it possible to calculate the population in each district that is not within 5 km of a health facility and, therefore, lacks readily available access to primary health-care.

The following table presents the populations with and without access in each of Malawi's districts, as both a total and a percentage of each district's population. Estimates suggest more than 450,000 people in each of five districts are currently underserved. These include Lilongwe, Mangochi, Dowa, Kasungu and Mzimba.

[11] In Malawi, the traditional authority is the administrative unit below the district level.

Mapping the coverage gap of the traditional authority level reveals similar patterns of coverage. Certain smaller districts have fewer people without access in total but are underserved as a percentage of the district population. For example, in Balaka only 54 percent of the population has access to health facilities. This suggests a trade-off between ensuring access to the maximum number of people and access to the highest percentage of people per district. See Figure 4.

FIGURE 4

Population with Access to Health Facilities Within 5 km

District		UNICEF Health Facility Catchment Areas Coverage		
Name	**Population**	**Total with Access**	**Total Without Access**	**% Population Without Access**
Total	**17,303,307**	**9,567,641**	**7,735,666**	**44.71%**
Balaka	421,134	227,263	193,871	46.04%
Blantyre	1,300,397	1,032,835	267,562	20.58%
Chikwawa	577,274	356,047	221,227	38.32%
Chiradzulu	381,370	273,895	107,475	28.18%
Chitipa	236,697	126,245	110,452	46.66%
Dedza	830,299	403,326	426,973	51.42%
Dowa	739,222	285,761	453,461	61.34%
Karonga	358,380	233,863	124,517	34.74%
Kasungu	829,530	267,784	561,746	67.72%
Likoma	11,962	11,962	0	0.00%
Lilongwe	2,526,221	1,565,465	960,756	38.03%
Machinga	648,531	323,784	324,747	50.07%
Mangochi	1,058,506	523,622	534,884	50.53%
Mchinji	605,201	224,441	380,760	62.91%
Mulanje	689,479	468,309	221,170	32.08%
Mwanza	122,127	50,493	71,634	58.66%
Mzimba	1,137,498	544,283	593,215	52.15%
Neno	141,353	65,379	75,974	53.75%
Nkhata Bay	286,593	126,249	160,344	55.95%
Nkhotakota	404,014	207,532	196,482	48.63%
Nsanje	314,478	211,262	103,216	32.82%
Ntcheu	623,126	317,567	305,559	49.04%
Ntchisi	298,223	126,582	171,641	57.55%
Phalombe	416,471	212,139	204,332	49.06%
Rumphi	232,241	126,774	105,467	45.41%
Salima	448,545	219,501	229,044	51.06%
Thyolo	777,455	473,007	304,448	39.16%
Zomba	886,982	562,268	324,714	36.61%

POPULATION MOVEMENT DYNAMICS

Estimating Population Density

Having identified a gap in coverage based on static population patterns, the next step was to incorporate MNO data to account for population growth and migration patterns.

For each cell tower, the authors identified the number of unique users within the cell tower's catchment area.[12] Dividing the number of unique users by the tower's catchment area gives the density per tower. This density was mapped to TAs by calculating the weighted average of overlapping polygons, providing the density of unique originating IDs in each traditional authority.

The analysis sought to determine the relationship between the density of unique originating IDs σc and population density P^c. The training data for population density came from the 2015 WorldPop data, extrapolated to 2016 and 2017, calculated for each traditional authority to ensure one-to-one correspondence between σ^c and P^c.

The relationship was estimated using a linear regression formula estimated using ordinary least squares (OLS):

$$logP^c = \alpha + \beta log\sigma^c + \mu_k$$

Where *α* is the regression constant, *β* is the coefficient of interest and μ_k is a regional fixed effect that allows for inter-regional variation. The regression was calculated separately for 2016 and 2017. To allow for potential spatial correlation, the specification was tested for the significance of Moran's I, a matrix weighted for spatial adjacency (Odland 1988). If the null of no spatial correlation was rejected, the model was estimated to account for spatial co-variance.

The central region of the country has a higher beta coefficient than the south or the north, likely because it has a much higher mobile penetration, as evidenced by the ratio of population density to unique call density. This validates the need to separate coefficients across regions. The coefficients were consistent within regions between years. See Figure 5.

[12] Unique users were defined as those active within the last three months (at least one SMS or call). As in Deville et al. (2014), the number of unique users per tower was calculated as the sum of unique users active within the cell tower's polygon between 8:00 pm and 7:00 am, when users are assumed to be at home. Their location was determined as the tower they use most frequently, the model tower. The number of unique users per night was then averaged over the entire year.

FIGURE 5

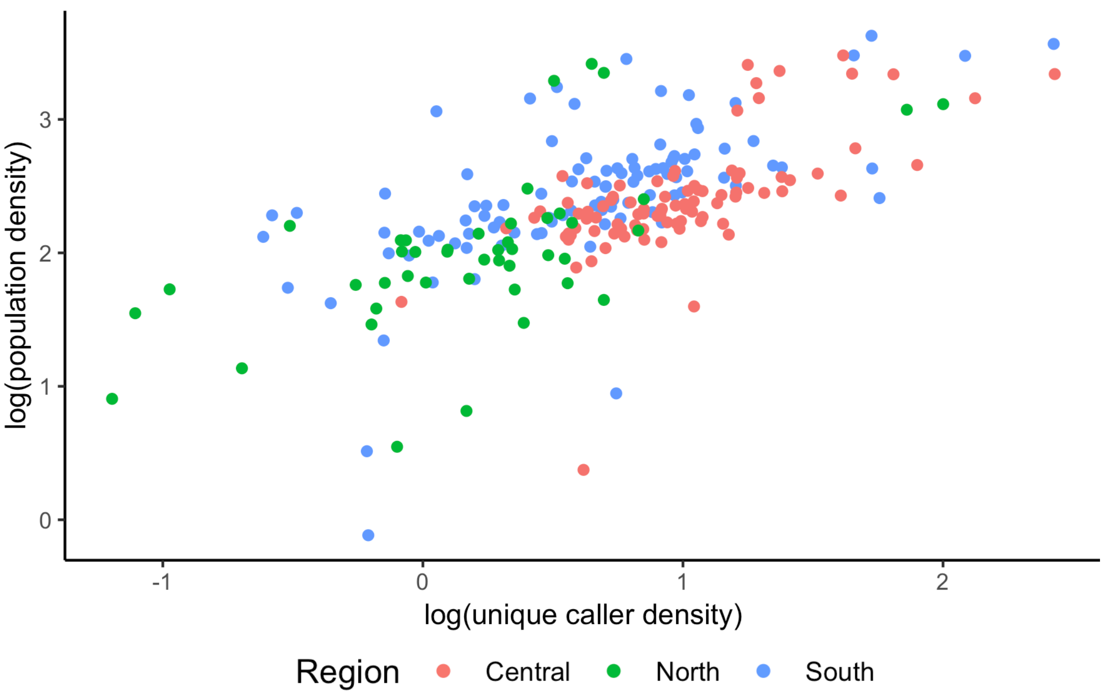
Results of Regressing Population Density on Density of Unique Callers, OLS

	2016			2017		
	South	**Central**	**North**	**South**	**Central**	**North**
α	4.52	3.43	4.1	4.51	3.29	4.01
β	0.624	0.86	0.595	0.597	0.89	0.597
P-Value	(6.815*10^-9)	(0.00478)	(0.0002662)	(3.388*10^-8)	(0.0349)	(0.0002321)
R^2	0.2748	0.3701	0.2533	0.2529	0.3707	0.2575
Moran's I	0.108	0.021	0.344	0.197	0.026	0.36822
Account for Spatial Correllation	No	Yes	No	No	Yes	No
(P^c/σ^c)	60.83	19.81	59.64	51.56	17.41	49.92
N	107	95	48	107	95	48

Spatial k-fold cross-validation was used to validate the above results (Pohjankukka 2017). The model was re-run k=8 times, each time omitting 1/8 of the sample. The estimated coefficient $\hat{\beta}$ was then used to calculate the predicted population density $logP^c$, and compared to the actual population density. The R^2 values suggest that variation in unique caller density accounts for between a quarter and a third of the variation in population density. See Figure 6.

FIGURE 6

Correspondence Between Log Call Density and Log Population Density



As evidenced from the figure above, the correspondence between log call density and log population density is largely linear, though there are some outliers at the low-end of population density, sparsely inhabited areas where the number of unique callers is an imperfect proxy for population.

Unlike the static population estimates from the WorldPop data, the number of unique user IDs from the MNO data varies from day to day and month to month. As we see above, the regional coefficients are consistent between 2016 and 2017, tentative evidence that the relationship holds over time. The estimated coefficients can therefore be used to calculate predicted population densities across time, uncovering the dynamics of population movement.

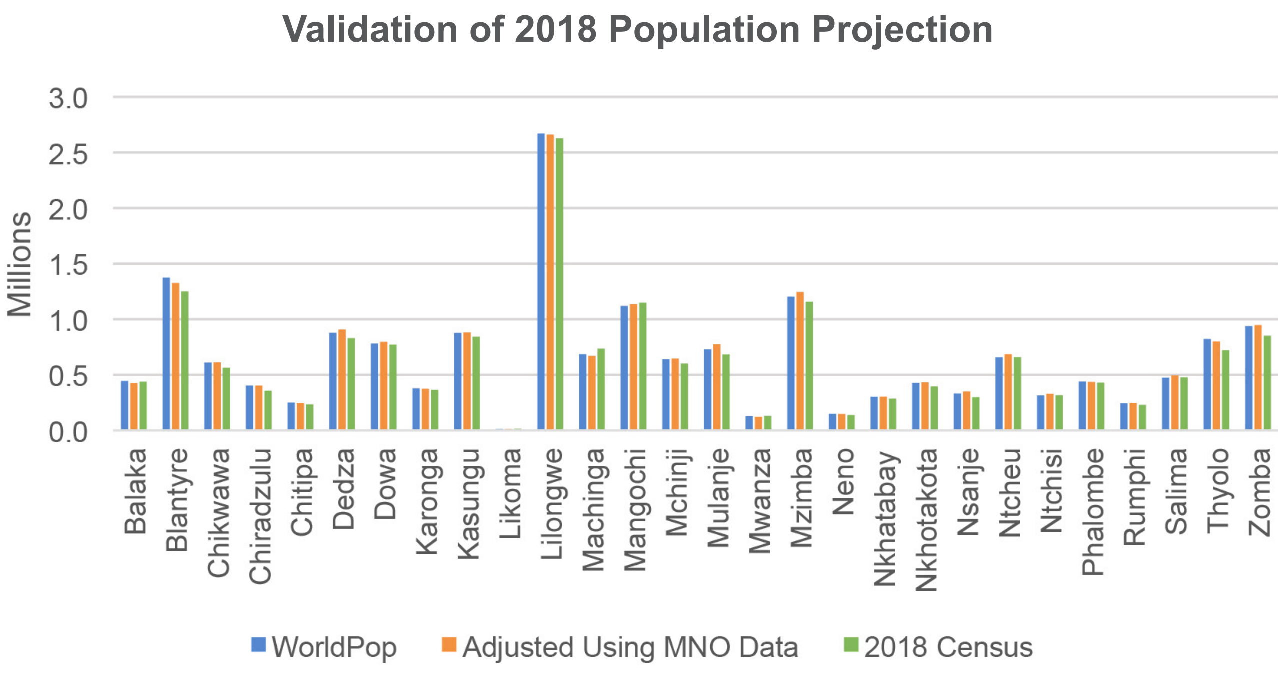
Validation Using Census Data

With the beta coefficients estimated, the next step was to project population growth patterns forward for every TA. The static WorldPop data uses a uniform growth rate across all TAs to project population growth, assuming that it will grow at a rate of 2.785% per year across all administrative areas.[13] Instead, the authors calculated the change in the density of unique users across time in each TA. In order to account for subscriber growth, a unique set of active users were identified at the beginning of January 2016 and their movements tracked throughout the two years to capture shifts in population.

Using the coefficients estimated above, these shifts in unique users were used to estimate the percentage change in population density for each TA. While the overall country's population growth rate was held constant each TA grew at a different rate. Certain TA's, largely in urban areas, were allowed to grow faster, while other TAs, notably in rural areas, grew slower or even shrunk. The country's population was therefore redistributed for allow for observed migratory patterns.

In order to validate this methodology, the predicted population levels were compared with the results from the 2018 census. The census provides a detailed headcount of population based on on-the-ground surveys at the national, district and traditional authority levels in 2018. This is considered the authoritative, ground-truth data. This was compared with predicted population levels based on a) the WorldPop data projected uniformly forward to 2018[14] and b) the TA-specific growth rates inferred from shifts in a fixed number of unique users. So, while the totals for a) and b) are identical, the district level population levels will differ. See Figure 7.

FIGURE 7



There is less than 5 percent discrepancy between the predicted and actual population levels for each district. Using satellite data and MNO data are both good proxies for actual population levels.[15]

---

[13] This number was derived by comparing TA level population levels as projected by WorldPop in 2015, 2016 and 2017.

[14] Recall that WorldPop based its projections on the 2009 census, training satellite data on it and projecting forward using a uniform growth rate.

[15] While this validates long-term growth patterns, as the authors do not know the composition of cell phone users relative to those who tend to migrate, short- and medium-term population mobility may be overstated or understated.

Short- and Medium-Term Population Movements

Once benchmarked to population density, the authors calculated population movement from month to month and across times of day. This allowed for the following three types of analysis:

1) Commuting

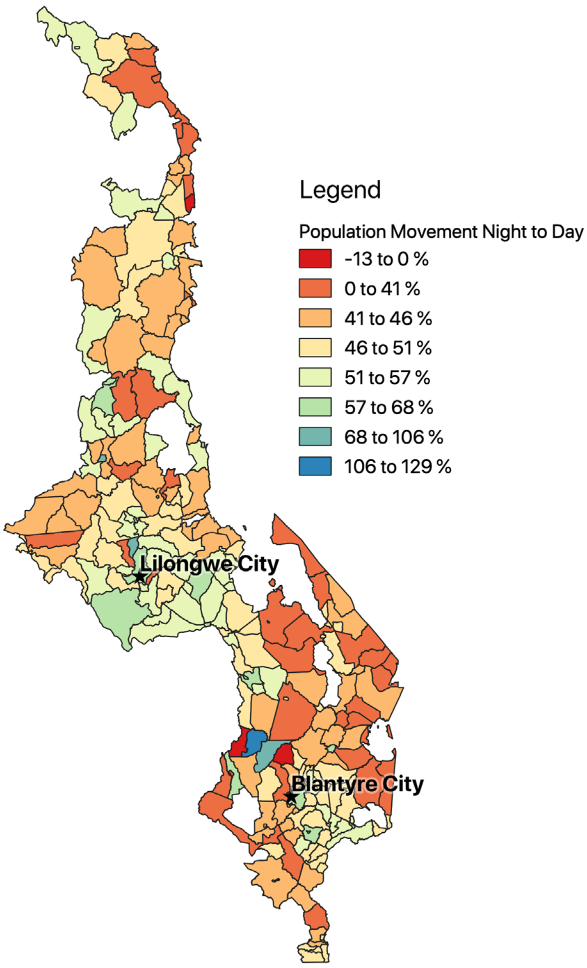
To understand commuting patterns, the authors compared the population as inferred by cell phone activity during the nighttime (8:00 pm – 7:00 am) and daytime (7:00 am – 8:00 pm).[16] The map below presents the percentage change in estimated population density during the day relative to the night. The percentage changes are mostly >0, reflecting a higher level of cell activity during the day relative to the night.

Nighttime location is assumed to indicate where a person lives. Shifts in their movement during the daytime offer evidence of commuting behavior. This may be relevant for patients seeking care when going to or returning from work.

Comparing nighttime and daytime activity at the TA level, there is a general surge during the daytime and some evidence of commuting into the larger cities. The report accounts for these commuting patterns in recommending the location of health posts by emphasizing areas where people spend the night, between 8pm and 7am. See Figure 8.

FIGURE 8

Evidence of Commuting from Periphery to City Centers



[16] Each unique number's daytime and nighttime location was inferred using their modal location during the daytime and nighttime.

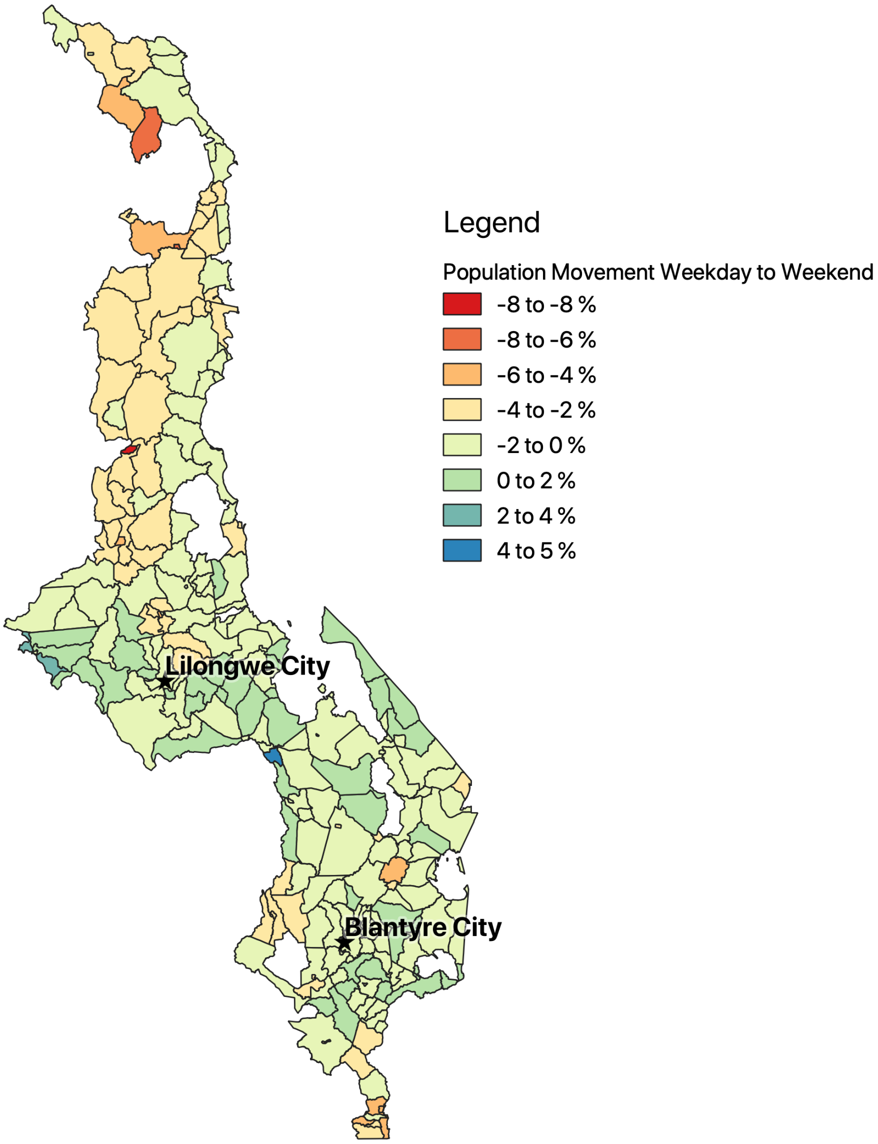
2) Weekend vs. Weekday

A second analysis observed population movement across the weekday and weekend to infer whether there were significant population movements to be accounted for when providing health services. This was done by comparing the average number of nighttime unique users per TA during the week to the average nighttime unique users per TA during the weekend and extrapolating to population using the estimated coefficients. The map below displays the percentage change in estimated population density on weekends relative to weekdays.

The most significant shift seems to be in the north, when on weekends a large number of people move from west to east, potentially to the lake. In the center, there is some evidence that individuals leave Lilongwe on weekends. High numbers of people along the Mozambique border over the weekend suggest the presence of market places. See Figure 9.

FIGURE 9

Population Shifts to Coast and Markets on Weekends



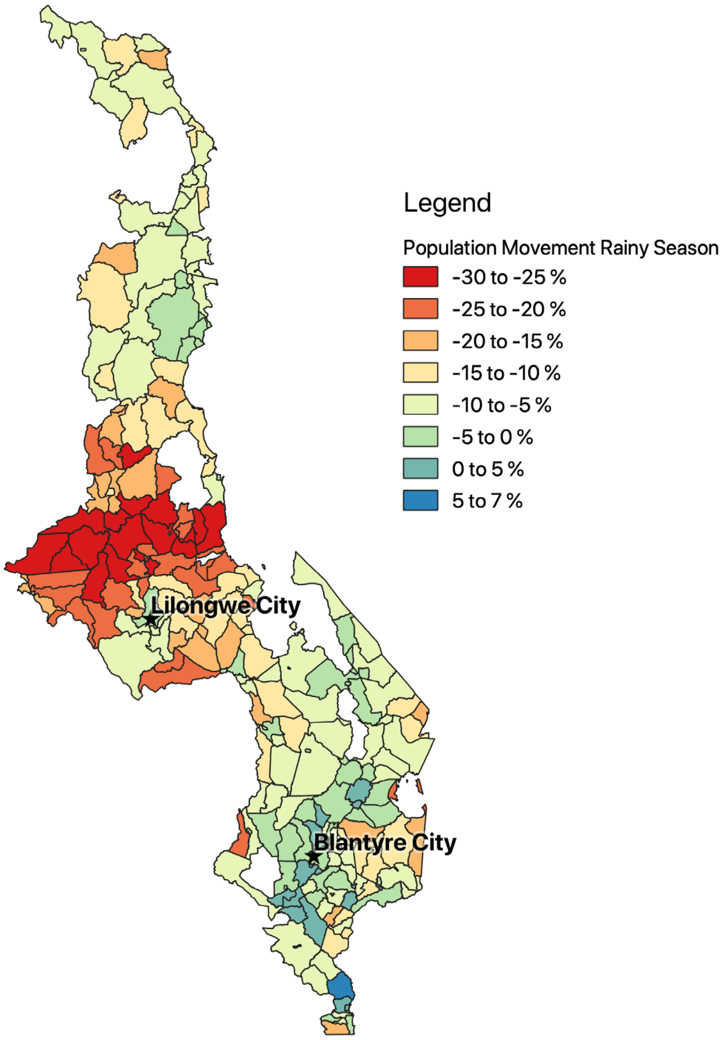
3) Migration

A third area of analysis was seasonal migration. Since an estimated 80 percent of Malawi's population engages in agricultural activities, many of them are subject to seasonal migration, particularly seasonal laborers. The authors therefore compared population density in the rainy season (November - April) to the non-rainy season (May - October). This meant comparing the average number of active unique users per month[17] in the rainy and dry seasons and extrapolated to population using the coefficients outlined above.[18]

The map below presents the change in population density, illustrating population movement during the rainy season as a percent change. See Figure 10.

FIGURE 10

Population Shifts from the Central Region to the Southern Region During the Rainy Season



During the rainy season there is a large-scale migration from the center of Malawi, especially around Lilongwe, towards the south. In particular, populations shift towards the Shire River Basin, an agriculturally fertile land where intensive agriculture is practiced. This offers evidence of seasonal migration, likely driven by the need for labor for agricultural activities.

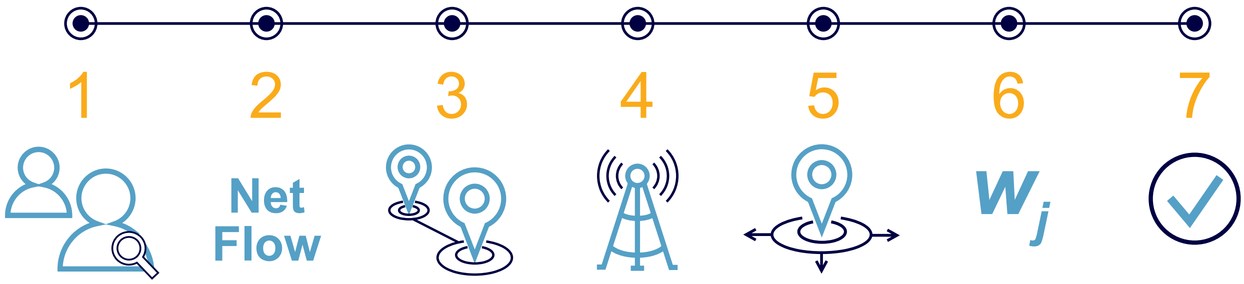
---

[17] This includes all users who sent an SMS or made a call in the past three months.

[18] As the demographic characteristics of cell phone users are not available, these estimates may either be an upper bound, if those who migrate are more likely to own cell phones, or a lower bound, if those who migrate are poorer and thus less likely to own cell phones. The magnitude of the shift suggests an upper bound estimate.

OPTIMIZATION OF FACILITY ALLOCATION

Given the observed gap in service provision, an optimal allocation of health posts can be calculated using projected population growth based on MNO data. This optimization seeks to maximize coverage of currently unserved populations, accounting for population growth and migration patterns.



The analysis proceeded in the following sequence:

STEP 1

Identified a consistent set of unique users in 2016 and 2017 and **calculated the net-flows in or out of each Traditional Authority** for each year. The net-flow is the difference between number of unique users in the TA at the beginning of the year and number of unique users in the TA at the end of the year. Restricting the analysis to a consistent sub-sample ensures that additional subscribers do not bias the sample. The TA-level 2016 and 2017 net-flows were not statistically different, so the average was taken.

STEP 2

Used NetFlow to adjust projected population growth for each TA. As described above, the national-level projected growth rate was pegged to the WorldPop growth rate, but TA level growth was allowed to vary. This method was validated using the 2018 census. Some TAs grew faster, especially in urban areas, and some TAs grew slower or even shrank. Population growth was then projected forward to 2019 to 2023, allowing differential growth rates that more accurately reflect migration and urbanization patterns.

STEP 3

Used the catchment areas of existing health facilities defined by UNICEF, along with **population density** and **long-term population movement**, to identify the availability of health facilities within each TA relative to the population. The total estimated uncovered population at time t is U_t.

STEP 4

Calculated spatial clusters containing a given uncovered population based on the WorldPop data adjusted for population growth and migration. To obtain such clusters, population density was calculated for each cell tower catchment area with the highest resolution available, using the estimated coefficients and adjusting for TA-level population growth. Short-term movements were used to identify highest-used locations within each TA. These clusters were restricted to be outside of the existing catchment areas, have a maximum radius of 6 km and contain no more than 12,500 people. If a given area contained more than 12,500 people, it was split, and two new clusters were calculated. The centroid for these spatial clusters was identified. These demand points are the candidate sites for new facilities, denoted P_{ij} for facility i in TA j.

STEP 5

Selected a distance beyond which a patient cannot travel to the health post: D called Impedance cutoff. This defines the catchment area. To maximize coverage, the cutoff was allowed to vary between 5 and 6 km. The population within this catchment is denoted C_i.

STEP 6

At the TA level, the **importance of population can be mediated by weight** w_j capturing TA-level disease burdens.

STEP 7

Optimization was conducted iteratively for every year, accounting for health posts allocated to date. Given the demand points, the number of facilities allocated in a given year is optimized to minimize the number of under-served U_t given the constraint $\overline{P}$, the maximum number of health centers that can be built within a year. Once allocated, in the subsequent year the demand point P_{ij} is removed from consideration, since it's already allocated. The model seeks to optimize the following objective function:

$$\text{Min } U_t - \sum\sum w_j * \sum P_{ij} * C_i$$

$$s.t \sum P_{ij} \leq \overline{P}$$

Based on the above, three different models were run.

MODEL 1

Accounts for underserved population only, giving all TAs an equal weight w_j of 1.

MODEL 2

Accounts for underserved population and TA-level disease burden but doesn't rank the severity of disease. w_j is the number of patients at the TA level, regardless of disease.

MODEL 3

Accounts for underserved population, TA-level disease burden and the severity of disease, ranking these by severity in terms of disability adjusted life years (DALYs) from the Global Burden of Disease (GBD) compiled by the Institute for Health Metrics and Evaluation (IHME).[19] w_j is the number of patients at the TA level, weighted by disease.

Model 1: Underserved Population Only

The results find that by strategically placing 900 new health posts to account for existing gaps, 95 percent of the population will be within 5 to 6 km of a health post by 2023. Adjusting for disease burden does not significantly alter these results. If those health posts are not built, 9.7 million Malawians, 44.85 percent of the population, would still be uncovered by 2023. See table below.

Results from Model 1, 2 and 3 in Terms of Coverage

Descriptions	Model 1	Model 2	Model 3
Year 4 (2023) forecasted population	21,621,892	21,621,892	21,621,892
Year 4 (2023) estimated uncovered population before 900 new health posts	9,696,818	9,696,818	9,696,818
Year 4 (2023) estimated uncovered % population before 900 new health posts	44.85%	44.85%	44.85%
Year 4 (2023) estimated uncovered population after 900 new health posts	1,122,720	1,228,322	1,242,047
900 new health posts	5.19%	5.68%	5.74%

These health posts are each expected to serve a maximum of between 12,000 and 12,500 people within a 5 to 6 km radius. The proposed schedule of construction is drawn from the CIP, with the assumption that the one health post scheduled for 2019 will be built in 2020 instead, for a total of 198 health posts in 2020. In each subsequent year, 234 health posts are to be built in order to reach the goal of 900 health posts by 2023. See table below.

Year-on-Year Distribution of Health Posts by Catchment Population (Model 1)

Year	Up to 12,000	12,000 to 12,500	Total
1	108	90	198
2	181	53	234
3	209	25	234
4	233	1	234
Total	**731**	**169**	**900**

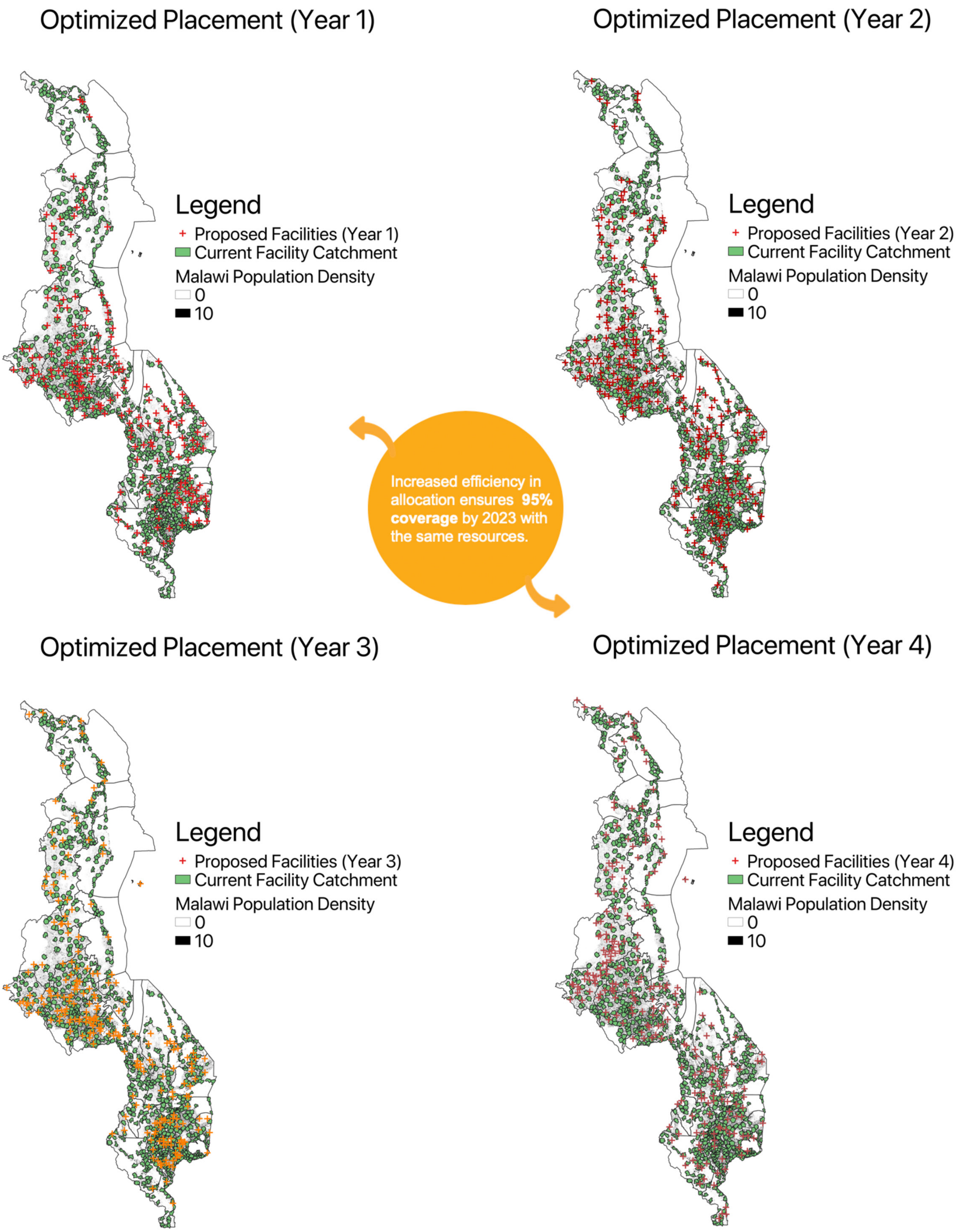
The allocation of these new health posts reflects population growth patterns, seeking to fill gaps in coverage in both rural areas and rapidly expanding urban areas. The initial facilities are expected to serve upwards of 12,000 people each, reflecting the pent-up demand for services. As new facilities are built, the additional number of people served by each health post gradually goes down. The 900th health post is expected to serve fewer than 3,000 people living within 6 km.

This model is currently informed by historical MNO data but can be updated using periodic or close to real-time MNO data as it becomes available. See Figure 11 on page 21.

[19] Available at http://www.healthdata.org/malawi.

FIGURE 11

Proposed Allocation of New Health Posts



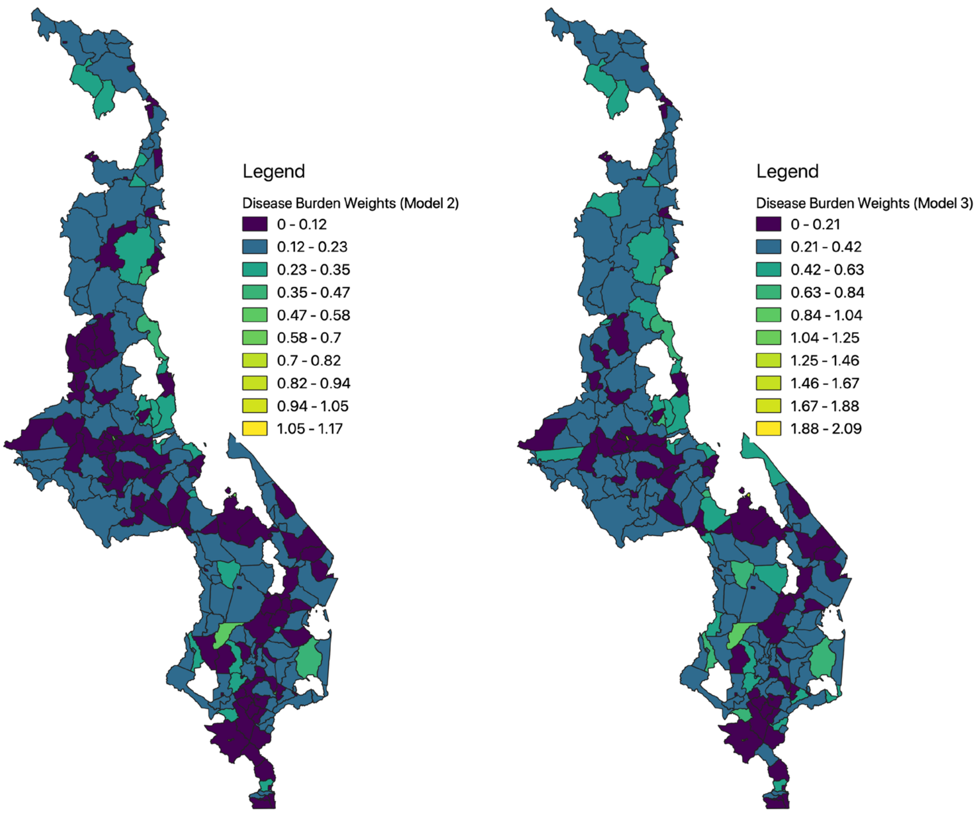
Model 2: Underserved Population and Disease Burden

Model 2 incorporates TA-level disease burden as reported by the Ministry of Health through the Health Management Information System. These are reported as 34 indicators, reflecting the number of patients diagnosed and receiving treatments every month in each reporting facility. The researchers chose to use administrative data because of both its timeliness and granularity relative to survey data, while allowing that cases reported may not necessarily reflect the actual disease burden in all areas due to suppressed demand.

In order to create weights from this HMIS data, each indicator was first summed across the year to reflect the total annual disease burden. Indicators from each facility were then summed up to the TA level. To accurately compare disease burdens across TAs, they were normalized by dividing the sum or reported cases by the TA population. An exception was made for district and central hospitals, which were assumed to service the entire district rather than a single TA. Their disease burden was therefore allocated to each TA in the district in proportion to the population, rather than to a single TA. See Figure 12.

FIGURE 12

Disease Burden Weights Based on HMIS Data



Finally, to calculate the weights, the 34 indicators were summed into a single index to reflect overall disease burden. This raw sum included the number of patients, cases and deaths across all categories. As the administrative data does not differentiate by patient, some of these indicators, such as OPD visits, included potential repeat visits by the same patient for the same condition. While an imperfect approximation for disease burden, it does reflect the workload health facilities currently face.

In terms of results, TA-level disease burdens were incorporated as weights to the objective function outlined at the top of page 18, and therefore skewed the allocation of facilities towards TAs that had a higher burden of disease. These weights adjusted the prioritization of allocation slightly, though the overall coverage after four years was near identical at an estimated 94.32 percent of the population.

Model 3: Underserved Population, Disease Burden and DALY

Model 3 differs from model 2 in that, in the final step, the sum of HMIS indicators incorporated the severity of the disease in terms of disability adjusted life years (DALYs), as determined by the Global Burden of Disease compiled by the Institute for Health Metrics and Evaluation (IHME). The weights were on a scale of 1 to 5, with 5 being the most severe. Adjusting for DALYs therefore put more weight on the indicators that have the greatest impact on health. See table below.

Disease	Weight Without DALY	Weight With DALY
# of 15 - 49 age group tested HIV positive	1	5
# of HIV positive persons receiving ARV treatment	1	1
# of HIV positive women treated for PMTCT	1	5
acute respiratory infections - new cases (u5)	1	5
HIV confirmed positive (15-19 years) new case	1	5
diarrhea non - bloody -new cases (under5)	1	4
malaria - inpatient deaths under 5	1	4
malaria -inpatient deaths (5 & over)	1	4
malaria new case (under 5)	1	4
malaria- new cases (5 & above)	1	4
malnutrition -inpatient deaths (under 5)	1	4
malnutrition new case (under 5)	1	4
# of deliveries attended by skilled health personnel	1	3
# of direct obstetric deaths in facility	1	3
# of postpartum care within 2 weeks of delivery	1	3
# of pregnant women starting antenatal care	1	3
# of road accidents - inpatient deaths	1	3
cholera - inpatients deaths	1	3
dysentery- inpatients deaths	1	3
total # of live births	1	3
# of fully immunized under 1 child	1	2
# of persons receiving Depo-Provera	1	2
# of persons receiving IUCD	1	2
# of persons receiving Norplant	1	2
# of persons receiving 3 months' supply of condoms	1	2
# of under 1 children given BCG	1	2
# of under 1 children given pentavalent	1	2
cholera - confirmed new cases	1	2
dysentery - new cases	1	2
ear infections - new cases	1	2
measles - confirmed new cases	1	2
# of OPD attendance	1	1
total # of discharges	1	1

Results in terms of allocation were very similar to model 2, with an ANOVA test rejecting significant differences between all three models at the district and TA levels. Model 2 and Model 3 tended to produce more balanced allocations in terms of catchment populations.

ALIGNMENT WITH THE MOH CAPITAL INVESTMENT PLAN

Overview of CIP

The Malawi Ministry of Health is planning to roll out 900 health posts over the next five years across all 28 districts of Malawi. These include both upgrades to existing facilities and the construction of new buildings to expand access, particularly in rural and remote areas, with an emphasis on the provision of primary health care.

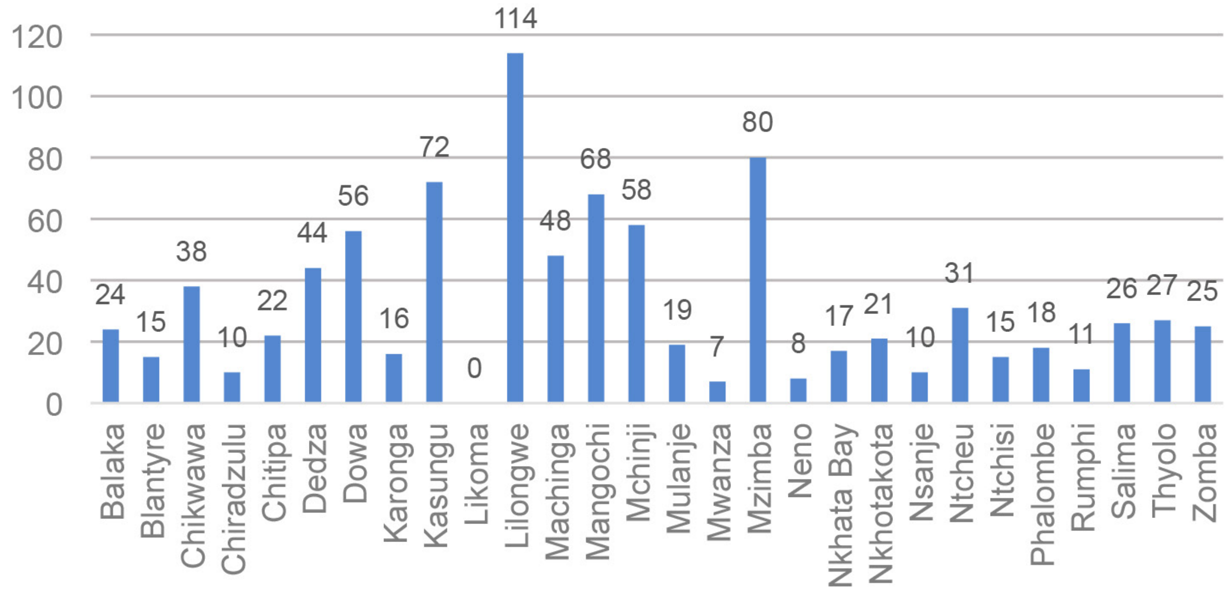
The Malawi Ministry of Health is planning to roll out 900 health posts over the next five years across all 28 districts of Malawi.

As part of its draft Capital Investment Plan, the government of Malawi developed proposed allocations of new facilities using the following four criteria:

1) Catchment population
2) Distance to nearest existing health facilities
3) Facility accessibility (high, medium, low)
4) Preferred year for work to take place as expressed by the District Health Monitoring Teams

The CIP projects that the cost of equipping and building these health posts will be **$41,954,407** over five years. Based on a review of the CIP appendix, the plan recommends the following allocation of health posts per district:

Proposed New Health Posts in Capital Investment Plan

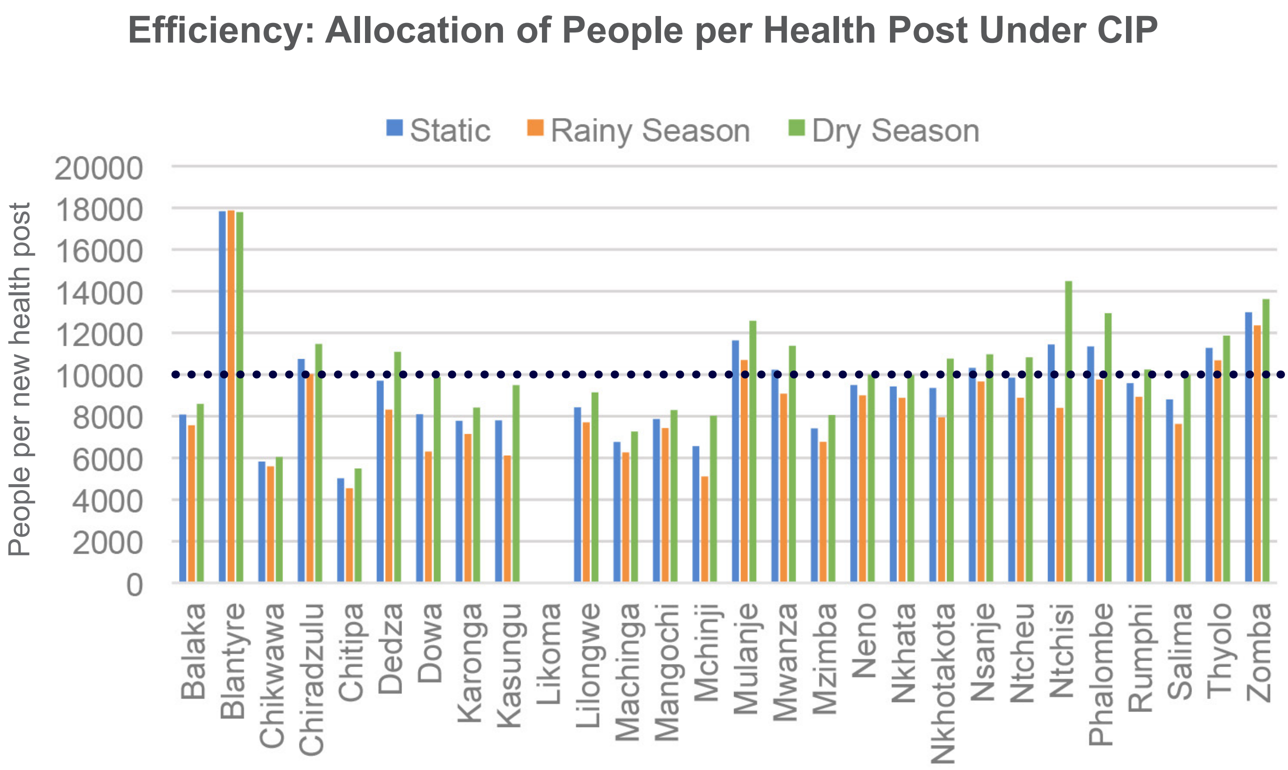


Allocative Efficiency

Since the exact location of these proposed health posts is yet to be determined, the authors could not conduct an analysis of their proposed catchment area. Instead, by incorporating this data into the analysis presented above, it can calculate the "people per health post." That is, based on the current number of people without access per district, how many people each new health post will service assuming they service all the people without current access. People per health post therefore measures allocative efficiency.

$$\frac{\text{(\# of People Without Access in District)}}{\text{(\# of New Health Posts)}} = \text{People per Health Post}$$

Incorporating MNO data, the metric changes when accounting for population shifts in the rainy and dry seasons. The people per health post for each scenario is presented in the figure below:



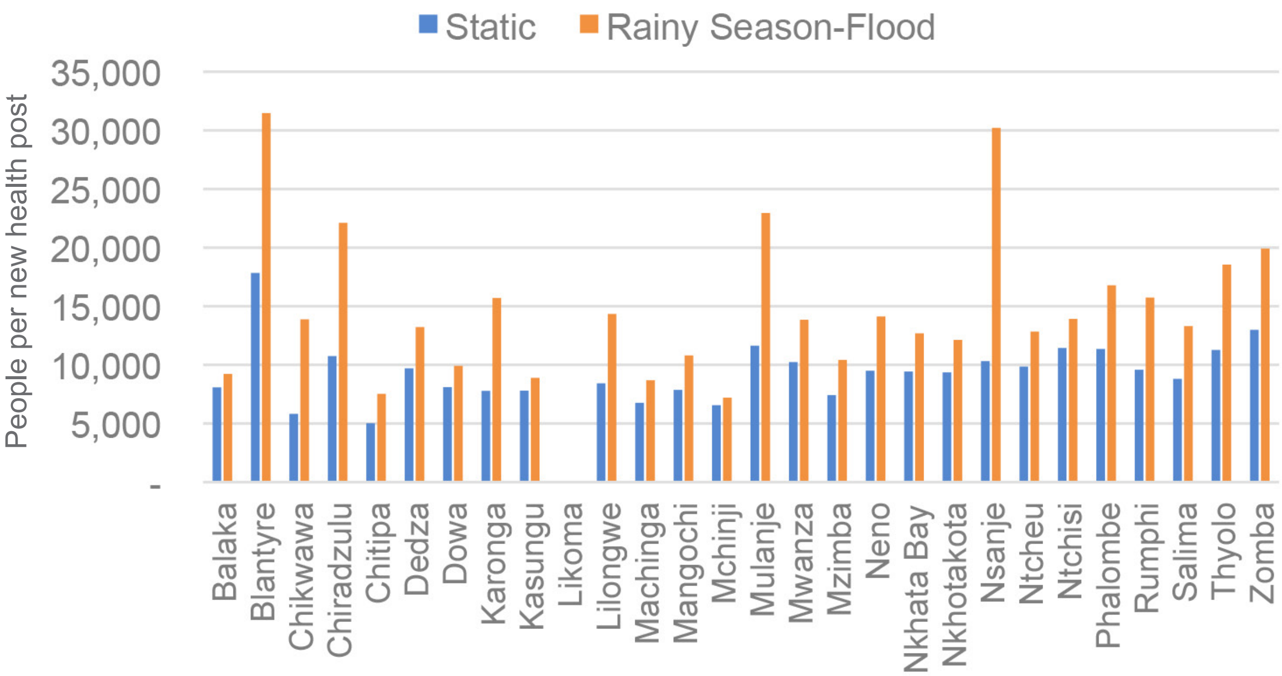
The dotted line represents an illustrative threshold of 10,000 per new health post, showing how certain districts are below that threshold and others above it.[20] Blantyre has a particularly high people per health post value, suggesting there are insufficient proposed posts to provide services to the population currently not covered.

[20] This threshold is illustrative only, as the CIP does not specify the estimated maximum capacity of a health post. The authors recommend consulting with MoH to establish a recommended threshold.

Allocative efficiency can also be adjusted to allow for catastrophic shocks, such as floods cutting off access to health facilities. By combining the UNICEF worst case scenario for flooding with the population movement analytics, the people per new health post ratio can be calculated in the event of catastrophic floods, like those in 2015, and thereby identify the most vulnerable districts. See Figure 13.

FIGURE 13

Resilience: People per Health Post, Accounting for Flooding



From the above, Blantyre, Nsanje, Chiradzulu and Mulanje are particularly vulnerable in the event of a flood. In Blantyre and Nsanje, a flood may mean that each new health post must service up to 31,500 people, straining its capacity.

COMPARISON WITH OPTIMIZED ALLOCATION

Comparison between the CIP and optimal allocation models could only be made at the district level, as the CIP does not recommend TA-level allocations. The optimization model therefore adds value in providing additional resolution, recommending both the Traditional Authority and the community where new health posts should be built.

Based on the model, if 900 health posts are built in optimal locations, 95 percent of Malawians will live within walking distance of a health post by 2023.

At the district level, though the allocations were broadly aligned, there were significant differences between the allocation of new health posts as outlined in the CIP and those determined by the allocation. These differed most in Blantyre, Zomba and Chikwawa, reflecting the difference in terms of allocative efficiency. In Blantyre alone, the re-allocation would reduce the people per health post from 17,800 to 9,200.

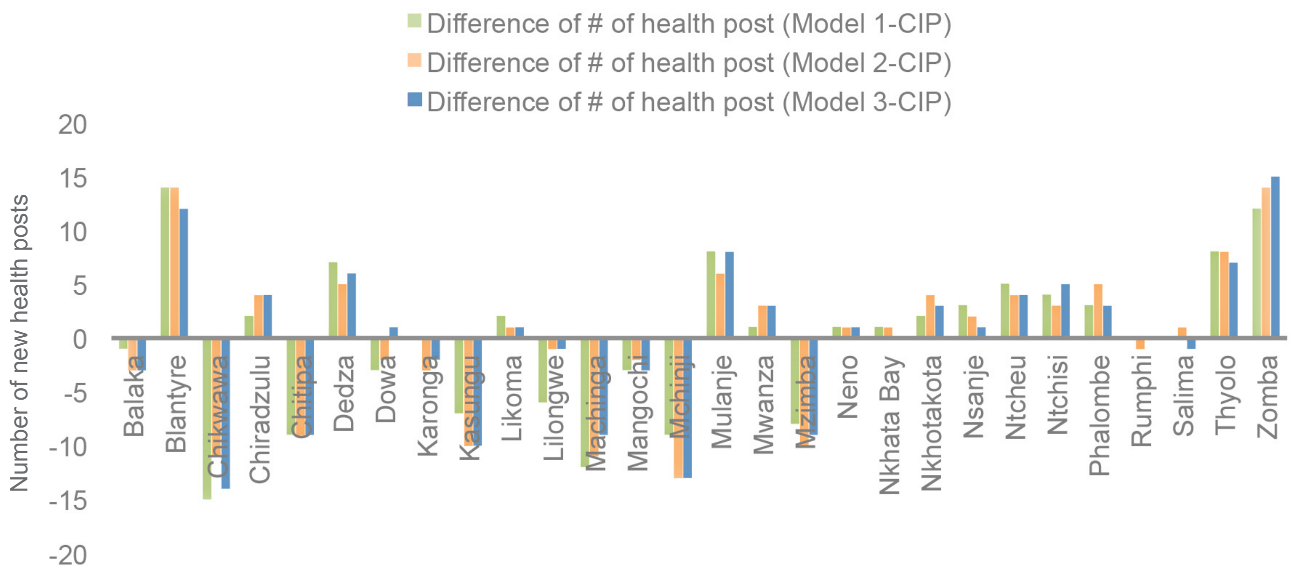
As a counterfactual, if each district built the number of health posts recommended in the optimized model and each health post could serve no more than 12,500 people, an additional 226,000 Malawians would have access to health services, relative to the allocation under the CIP.[21]

Given that these regions are most vulnerable to having health facilities cut off in times of flood, building additional facilities may also increase their resilience, ensuring that Malawians have access to health facilities when they need them most. In the worst-hit districts, the maximum number of people per health post drops by between one-third and one-half, and the highest burden drops from 31,500 to 23,200, a more manageable number in times of crisis.

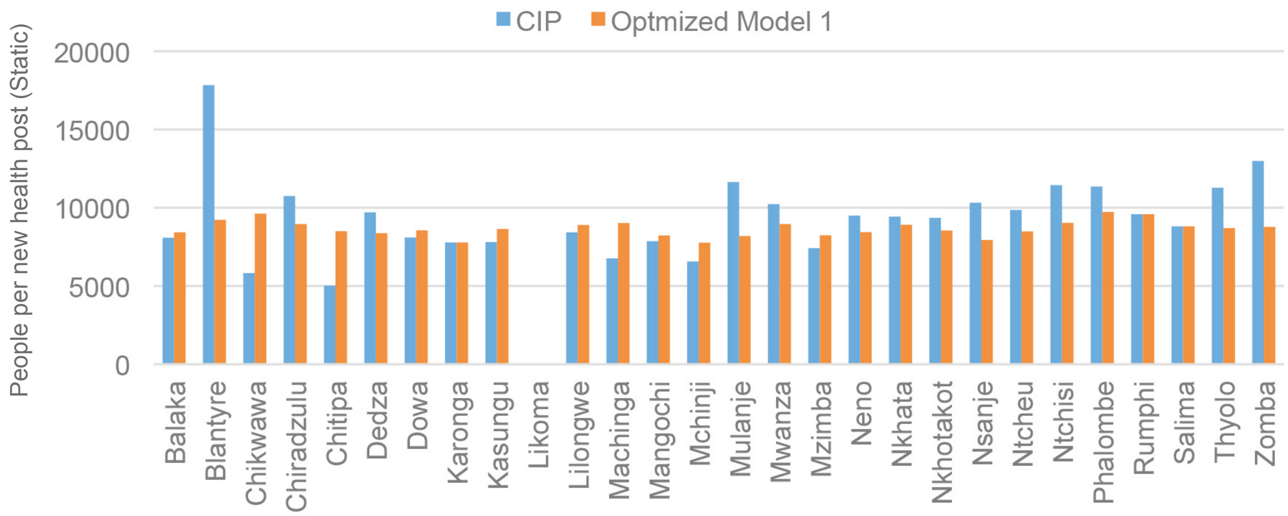
Based on the model, if 900 health posts are built in optimal locations, 95 percent of Malawians will live within walking distance of a health post by 2023.

[21] This was calculated as $\sum[(U_k - P_k^{CIP} * 12{,}500) * 1(U_k - P_k^{CIP} * 12{,}500 > 0)]$ the sum of the difference between U_k, the unserved population per district adjusted for population growth, and P_k^{CIP} the number of health posts proposed by the CIP multiplied by 12,500, but only if the difference is positive.

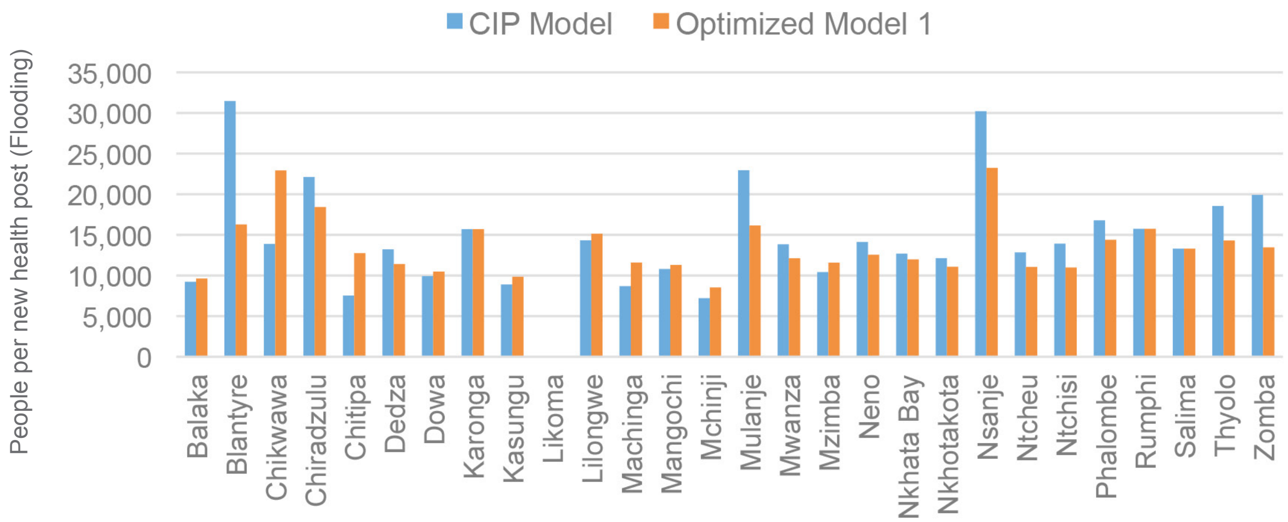
Change in Allocation Relative to the CIP



Counterfactuals: Optimized Model Improves the Efficiency of Allocation…



… and Its Resilience to Flooding



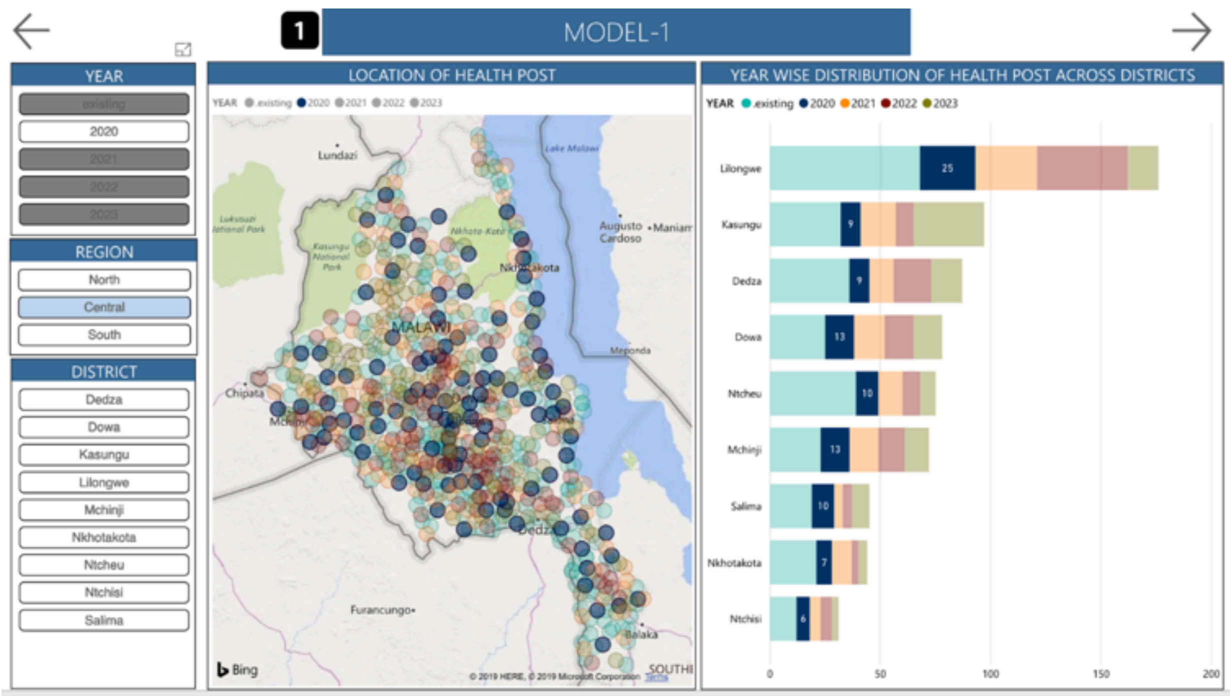
SUSTAINABILITY AND SUPPORT TO GOVERNMENT PARTNERS

Dashboard

To provide further insights, the authors developed an interactive dashboard using Power BI. This dashboard provides an overview of:

- Estimated population density
- Health post coverage
- Cell phone usage patterns
- Long-term population movements
- Short-term population movements

Snapshot of User Dashboard



The authors intend to integrate data on patients and disease burden into the dashboard, combining these with the above in a user-friendly format based on feedback from MoH.

Engagement with Ministry of Health

The key to ensuring the relevance of this use case is engaging with government counterparts at both the policy and technical levels. Input was solicited throughout the process and incorporated into the model. Direct engagement around the analytical products allowed technical counterparts to provide feedback. These inputs guided the authors in designing analytical products around the requirements and capacities of the ministry.

In order to ensure sustainability, DIAL and Cooper/Smith organized a series of dissemination exercises. These included one-on-one meetings with government counterparts, deep-dive presentations where input was solicited and a dissemination workshop with a detailed walk-through of the analytical products. Participants came away with an understanding of how the products inform their activities and provide concrete feedback the authors could incorporate to ensure the products fit into existing country systems. This process also initiated the conversation around sustainability, in order to integrate the use case into the ministry's decision-making.

These conversations around sustainability emphasized tapping into a broader set of use cases to answer other questions regarding the provision of health services. This included conversations around additional datasets that could be combined with MNO data. The conversations also emphasized the importance of using appropriate tools that ministry counterparts were familiar with. In order to turn this analytical exercise into a viable long-term solution, the one-time transfer of MNO data needs to evolve into a data pipeline, providing updates on population movement dynamics on a monthly or bi-monthly basis.

LESSONS LEARNED AND RECOMMENDATIONS

The process of acquiring the data, analyzing it and integrating it into country systems generated a number of lessons learned and recommendations, notably:

1. **Define a specific, demand-driven use case** to structure the analytical process within a realistic time horizon that allows for unforeseen delays.
2. **Emphasize country-level buy-in from the very beginning of the process** to ensure that the use case is policy-relevant and that the research is in full regulatory compliance with regards to data encryption and user confidentiality.
3. **Engage with private-sector partners** on the potential value-add from both a CSR and business development perspective. Mobile network operators are looking to engage with development partners to expand the use of their data products, but expectations must be managed to allow for differing perspectives.
4. **Bring together a broad-based analytical team of researchers** to tackle the many technical challenges inherent in preparing, cleaning and analyzing multiple datasets and bringing them together to deliver relevant insights.
5. **Engage continuously with technical counterparts** to ensure the relevance of analytical products, laying the groundwork for integrating these products into country systems.

Limitations of the model include:

1. **The model is gender blind by construct**, since the data was stripped of identifying characteristics. Research by the Global System for Mobile Communications Association (GSMA) has shown that in Sub-Saharan Africa there is a 15% gender gap in mobile ownership, so this model likely over-represents the movements of men compared to women, who are more likely to visit health facilities, particularly for pre-natal care.
2. **The model is based on data from one of two principle telecom providers** in the country. Our validation exercise has found no evidence that this systematically biases the data.
3. **By construct, there is no data for the approximately 5 percent of zones outside of mobile coverage** in Malawi. Population movement and growth in these zones has been inferred based on observed patterns in adjacent areas with mobile coverage, but since these are the most remote areas, these inferences might not provide a complete picture.

With these recommendations in mind, this use case demonstrates both the feasibility and potential impact of combining new data streams with rigorous analytics to improve service delivery.

REFERENCES

Blumenstock, J., Cadamuro, G., and On, R. “Predicting Poverty and Wealth From Mobile Phone Metadata,” Science, 350(6264), 1073-1076 (2015).

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., and Tatem, A. J. “Dynamic Population Mapping Using Mobile Phone Data,” *Proceedings of the National Academy of Sciences*, 111(45), 15888-15893 (2014).

Erbach-Schoenberg, E., Alegana, V. A., Sorichetta, A., Linard, C., Lourenço, C., Ruktanonchai, N. W., and Tatem, A. J. “Dynamic Denominators: The Impact of Seasonally Varying Population Numbers on Disease Incidence Estimates,” *Population Health Metrics*, 14(1), 35 (2016).

Odland, J. *Spatial Autocorrelation* (Vol. 9). Newbury Park, CA: Sage Publications (1988).

Pohjankukka, Jonne and Pahikkala; Tapio and Nevalainen; Paavo and Heikkonen; Jukka. “Estimating the Prediction Performance of Spatial Models via Spatial K-fold Cross Validation. *International Journal of Geographical Information Science*, 1-19. 10.1080/13658816.2017.1346255 (2017).

Stevens, F.R., Gaughan, A.E., Linard C., Tatem A.J. “Disaggregating Census Data for Population Mapping Using Random Forests With Remotely-Sensed and Ancillary Data,” *PLoS ONE*, 10(2): e0107042 (2015).

Wesolowski, A., O’meara, W. P., Tatem, A. J., Ndege, S., Eagle, N., and Buckee, C.O. “Quantifying the Impact of Accessibility on Preventive Healthcare in Sub-Saharan Africa Using Mobile Phone Data,” *Epidemiology* (Cambridge, Mass.), 26(2), 223 (2015).

WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High-Resolution Population Denominators Project - Funded by the Bill & Melinda Gates Foundation (OPP1134076).


GOGO CHATINKHA BANDA
MATERNITY WING
BUILT IN 1980
sobo
ONCOLOGY WARD
JANE
CLINIC

Digital Impact Alliance

digitalimpactalliance.org