





ARTICLE



<https://doi.org/10.1057/s41599-021-00892-w>

OPEN

Mapping out-of-school adolescents and youths in low- and middle-income countries

V. A. Alegana^{1,2}[✉], C. Pezzulo², A. J. Tatem², B. Omar³ & A. Christensen⁴

Education is a human right and a driver of development, but, is still not accessible for a vast number of adolescents and school-age-youths. Out-of-school adolescents and youth rates (SDG 4.3.1) in lower and middle-income countries have been at a virtual halt for almost a decade. Thus, there is an increasing need to understand geographic variation on accessibility and school attendance to aid in reducing inequalities in education. Here, the aim was to estimate physical accessibility and secondary school non-attendance amongst adolescents and school-age youths in Tanzania, Cambodia, and the Dominican Republic. Community cluster survey data were triangulated with the spatial location of secondary schools, non-proprietary geospatial data and fine-scale population maps to estimate accessibility to all levels of secondary school education and the number of out-of-school. School attendance rates for the three countries were derived from nationally representative household survey data, and a Bayesian model-based geostatistical framework was used to estimate school attendance at high resolution. Results show a sub-national variation in accessibility and secondary school attendance rates for the three countries considered. Attendance was associated with distance to the nearest school ($R^2 > 70\%$). These findings suggest increasing the number of secondary schools could reduce the long-distance commuted to school in low-income and middle-income countries. Future work could extend these findings to fine-scale optimisation models for school location, intervention planning, and understanding barriers associated with secondary school non-attendance at the household level.

¹ Population Health Unit, Kenya Medical Research Institute - Wellcome Trust Research Programme, Nairobi, Kenya. ² WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, UK. ³ Tanzania Data Lab (dLab), Dar es Salaam, Kinondoni District, Tanzania. ⁴ Plan International (PlanBørnefonden), Copenhagen, Denmark. ✉email: valegana@kemri-wellcome.org

Introduction

Despite efforts to improve access to education, many adolescents and youths of school-age remain marginalised, disproportionately by geography, social-economic status, cultural norms and gender (Graetz et al., 2020; UNESCO, 2020a). The 2030 Sustainable Development Goal (SDG) agenda will not be achieved without substantial investment in education and associated inequalities (Friedman et al., 2020) for pre-adult age groups (UNESCO, 2015a, 2015b). This age group constitutes ~18% of the global population and represents the future well-being of society and its socio-economic development potential. The lower secondary school age (12–14 years) and youths of upper secondary school age (15–17 years) is also a period when higher risk (or protective) behaviours start or become entrenched, having a major impact on their health and development as adults (UNFPA, 2007). While there has been increased investment and initiatives in many countries since 2010 to improve access to secondary school education (Morgan et al., 2014; Koski et al., 2018), the proportion of out of school adolescents of lower secondary age and youth of school age remains unacceptably high—progress has been at a near standstill for 8 years (UNESCO, 2018) with sub-national heterogeneities and their drivers, such as geographic distance, poorly understood in many low-income and middle-income countries where policies are targeted. Without education access and protection, the immediate economic and social well-being of any country is at risk.

Distance to school is a recognised barrier to education access alongside socio-economic and demographic household characteristics such as parents level of educations, wealth, and early marriage (Buchmann, 1999; Yu, 2007; ILO, 2010). For example, according to the World Inequality Database on Education (WIDE) for 2020 (UNESCO, 2020b) only 8% of youths completed secondary school in Tanzania, 57% in the Dominican Republic and 21% in Cambodia. To aid in understanding inequalities related to physical or geographic accessibility, information on the location of populations, schools, and socio-demographic characteristics are increasingly available to develop fine spatial resolution maps of geographic accessibility. Further, triangulating data on the spatial location of school with the household-level data from nationally representative household surveys undertaken every 3–5 years (Anderson and Cleland, 1984; Ayad et al., 1997; Burgert-Brucker et al., 2015) can be useful in estimating school attendance at a community level. Finally, the improved mapping of the age-structured global population (Stevens et al., 2015; Wardrop et al., 2018; Worldpop, 2018) provides opportunities for understanding the location of services within populations and improves the estimation of those marginalised from schools. There have been no previous attempts that triangulate available community survey data, with the spatial databases of schools and fine-scale age-structured population maps at a sub-national level to estimate access and attendance amongst adolescents and school-age-youths at a fine geographic scale.

Here we employ a geospatial approach using an example of three countries classified by the World Bank as low- and middle-income (United Republic of Tanzania—Tanzania mainland, Cambodia and the Dominican Republic). The aim was to examine geographic accessibility to secondary schools and associate this with the predicted out-of-secondary school rates at a fine spatial resolution (1 by 1 km). The approach integrates locations of secondary schools with fine-scale geospatial covariates to estimate geographic accessibility in a Geographic Information System (GIS).

Methods

Spatial database for schools. Geographic location data for schools were assembled from governmental sources for the three

countries. These countries were selected based on the geographic differences and heterogeneities in the distribution of secondary schools in Africa (Tanzania), Southeast Asia (Cambodia), and Latin America (Dominican Republic). For Tanzania 3258 secondary schools location data were obtained from the United Republic of Tanzania data portal (The United Republic of Tanzania—Government Basic Statistics Portal, 2015). The designated age range for a secondary school in Tanzania is 14–19 years old. For Cambodia, these were obtained from the Ministry of Education, Adolescent and Sport (MoEYS) consisting of 1615 schools classified as College, LyceeG10-12, and LyceeG7-12 with an age range of 13–18 years. In the Dominican Republic, data were obtained from the Ministerio de Educación de la República Dominicana ($n = 4618$) (Ministerio De Educación, 2018) and the corresponding age-range was 14–17 years.

Cluster-level data on school attendance. The Demographic and Health Surveys (DHS) for Tanzania (2015–16 DHS, $n = 595$ clusters) (Ministry of Health et al., 2016), Cambodia (2014 DHS, $n = 611$ clusters) (National Institute of Statistics/Cambodia et al., 2015), and Dominican Republic (2013 DHS, $n = 476$ clusters) (Centro De Estudios Sociales Y Demográficos—Cesdem/República Dominicana and ICF International, 2015) were first used to derive the rates of attendance adjusted for DHS sampling and stratification. The DHS survey sampling in each country was based on a two-stage stratified sampling design using the national census sampling frame. During the first stage, enumeration areas (EAs), also known as clusters, were selected by using a probability proportional-to-population size. During the second stage, households were sampled from a complete household listing in the selected EAs. Specific details on the sampling procedures for the three countries of interest for this work can be found in the DHS final reports, and in the DHS Sampling Manual (ICF International, 2012). DHS clusters were defined as a group of households in the same area or a block (if in urban areas) selected for the interview within the complex survey design used by the DHS, and usually cluster level spatial coordinates (latitude and longitude) are also provided in the surveys.

Ancillary covariates and population data. Additional covariate data were assembled to aid in the estimation of geographic access and interpolation of cluster-level data. Land use and land cover maps for the three countries were obtained from MERIS GlobCover (Arino et al., 2007). Globcover classification uses 22 classes defined based on the United Nation's Land Cover Classification System (UN-LCCS) (Fao, 2000). The current GlobCover V.2.3 was derived from a time-series of medium resolution imaging spectrometer (MERIS) satellite imagery acquired from December 2004 to June 2006 at a spatial resolution of 300 m. An improved gap-filled digital elevation model (DEM) data was obtained from the HydroSHEDS dataset based primarily on NASA's Shuttle Radar Topography Mission (SRTM) (Lehner et al., 2008). Roads data were assembled from Open Street Maps (OSM) and online resources such as the National Geospatial-Intelligence Agency (NGA) (NGA, 2015) and independent data from MapCruzin (an independent open-source data repository <http://www.mapcruzin.com/>). A gridded night-time light dataset based on low-light imaging of earth at night was downloaded, obtained from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) [<https://www.ngdc.noaa.gov/eog/viirs.html>] operated jointly by NASA and NOAA (Elvidge et al., 2017). This gridded image of lights at night has been shown to correlate highly with the urban population (Small et al., 2005; Shi et al.,

2014). Lastly, a 1×1 km population maps were downloaded from WorldPop (Worldpop, 2018). These represented disaggregated census-based maps using a combination of standardised dasy-metric mapping approaches informed by population density weights calculated using random forest (RF) methodology (Stevens et al., 2015).

Geospatial modelling of travel time to secondary schools. A gridded layer of travel times to secondary schools was estimated using land use, elevation (the DEM), and roads in AccessMod (version 5.0) (Ray and Ebener, 2008). The other assembled covariates were used in interpolating attendance rates in space. In deriving the travel time grid, each GIS layer was converted into a raster surface 1×1 km and each pixel was assigned an impedance value representing the speed of traversing a grid pixel based on land use type (Table 2). The resulting rasters were then combined into a gridded friction layer. Travel speeds were assigned to different land use classes, roads and slope by assuming multiple modes of transport within a single journey. For instance, for primary roads, motorised transport was assumed with a maximum speed of 80 km h^{-1} . On other tertiary roads, a 5 km h^{-1} walking speed was adopted with a correction for non-motorised transport (Cycling) at 10 km h^{-1} applied on residential roads. Details of travel speed and mode of travel were selected based on recommendations from previous studies (Noor et al., 2006; Tanser et al., 2006). The DEM was used to derive slope and different speeds calculated for each degree rise based on Tobler's equation:

$$V = 6 * \exp\pi(-3.5 \text{ abs}[\text{Tan}(\text{slope in degrees}/57.296) + 0.05])$$

where V is the calculated speed (Tobler, 1993). Travel times to each school were computed separately.

Estimating net school attendance rates at cluster level. The calculation of the adjusted net school attendance rate followed the guidelines and coding proposed by the MEASURE DHS programme (Croft et al., 2018). The methodology adopted Stata (Statacorp., 2017) (software for statistics and data science) code for estimating rates adjusting for survey weighting. The adjusted net attendance rate estimated the total number of students of the official secondary school age group who attended secondary education (or primary, or higher education) at any time during the reference academic year. The numerator was the de facto total population of secondary school-age attending primary or secondary or higher school, while the denominator was the total number of de facto secondary school-age adolescents or school-age-youths. It, therefore, included students of official school age who accessed school earlier or later than the normal enrolment age and was expressed as a percentage of the corresponding population (UNESCO, 2019), giving a more precise picture of school participation. Age ranges were established based on guidelines from the National Ministry of Education and the UNESCO Institute for Statistics database. The age at the start of the academic year was used to determine the eligible secondary school-age population used in the numerators and denominators for the net attendance rate. To establish these age ranges, full information on the date of birth of the child in question was triangulated with the start of the academic year, to account for the temporal gap between the interviews and the start of the academic year. For geospatial mapping purposes, these rates were aggregated at cluster level and the proportion of secondary school age attending (or out-of-school) in each cluster was computed. The computed proportions at each georeferenced cluster were then interpolated through modelling at the second stage.

Geostatistical modelling of secondary school attendance. A model-based geostatistical method (Diggle et al., 1998) was used to spatially interpolate cluster-level estimates of attendance with gridded covariates to define mean attendance at 1×1 km. The renaissance of model-based geostatistical (MBG) approaches has occurred in other fields (Banerjee et al. 2004; Lindgren, 2013), with the added advantage of estimating uncertainty associated with the estimation of school attendance. At the first stage, covariates were selected using a statistical procedure. Covariates considered included the modelled travel time to the nearest school, the enhanced vegetation index, night-time light, minimum and maximum temperature in all three countries. A *bestglm* (Mcleod and Xu, 2008) procedure was then implemented for each country separately resulting in a parsimonious set for modelling.

The main objective in modelling was to predict net attendance at fine-scale for all locations nationally using a parsimonious set of covariates that were statistically important in explaining variation in observed attendance rates. For this purpose, a Bayesian hierarchical spatial model was implemented in the Integrated Nested Laplace Approximation in R software (R-INLA) (Rue et al., 2009; Cameletti et al., 2012; Martins et al., 2013) to estimate a continuous map of the proportion attending secondary school-level education at 1×1 km spatial resolution. A stochastic partial differential equation (SPDE) approach was adopted using R-INLA, and computation performed via Gaussian Markov random function (GMRF). A stationary model was implemented using Matérn covariance with the smoothness of process ν and variance σ^2 given by

$$\alpha = \nu + d/2$$

where d is spatial dimension and marginal variance $\sigma^2 = 1/(\Gamma(\nu)\Gamma(\alpha)(4\pi)^{d/2}k^{2\nu}\tau^2)$. A linear model was implemented using a Gaussian likelihood for the proportion attending school adjusted for sampling and strata. Thus,

$$z(s) = x(s)\beta + w(s) + \varepsilon(s)$$

where $z(s)$ are realisations of the underlying attendance process linked to a spatial structured predictor in an additive way, $x(s)$ denotes set of covariates with β coefficients and $\varepsilon(s)$ is the measurement error. $w(s)$ represents the spatial process associated with the spatial association between clusters. The Bayesian specification was completed by assigning non-informative priors to hyper-parameters to the fixed effects (covariates) and the random parameters (spatial and the measurement error). For SPDE parameters, a penalised complexity (PC) priors framework was used for the model range and the marginal variance (Fuglstad et al., 2019, 2020).

Model calibration (statistical consistency) and sharpness (concentration) were assessed using the probability integral transform (PIT) and the conditional predictive ordinate (CPO), a leave-one-out cross-validation approach in which an estimate was validated based on the fitted model and the remaining data only (Spiegelhalter et al., 2002; Czado et al., 2009). A 20% subset of data selected randomly was used in the computation of the mean prediction error (MPE), the root mean square error (RMSE), and a Pearson's product-moment correlation coefficient that quantified the association between observed and predicted values. Figure 1 shows the overall methodology for geostatistical prediction of out-of-school rates (Breiman and Spector, 1992). Methodology for school attendance

Results

Summary of data and distance to school. There were 3258 secondary schools in the Tanzanian mainland, 1615 in Cambodia and 4618 in the Dominican Republic. The average straight-line

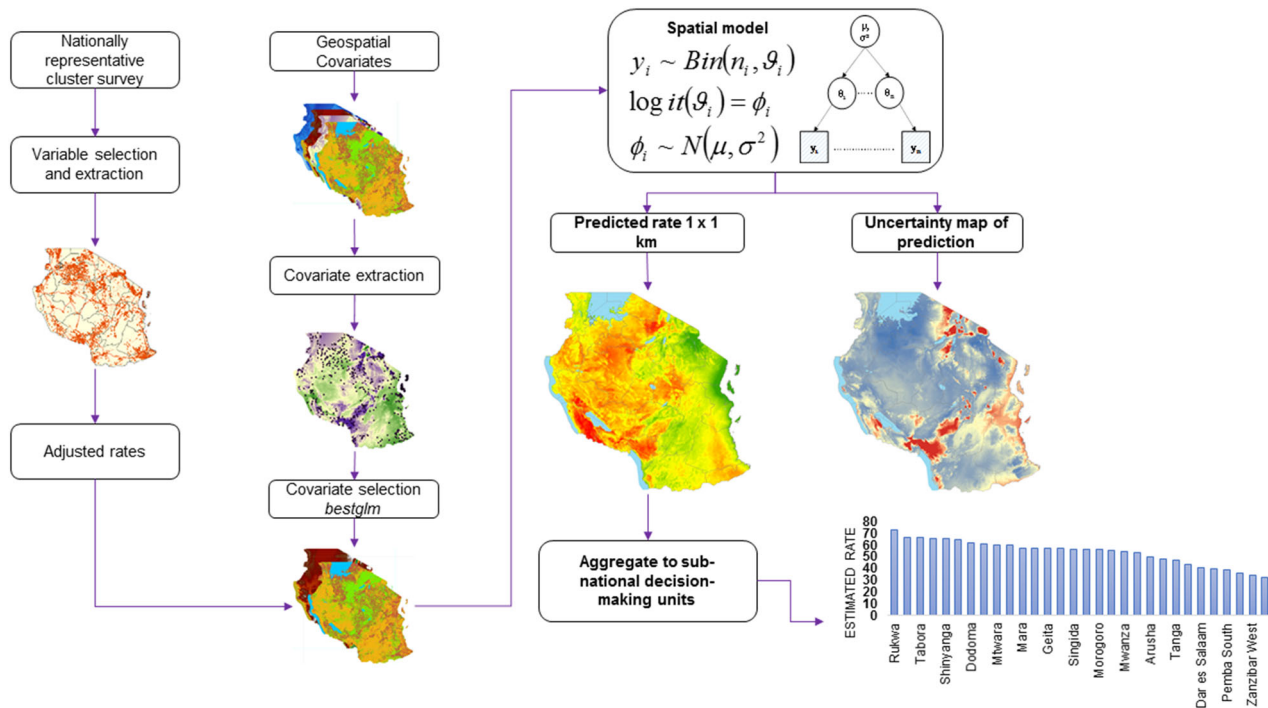


Fig. 1 Methodology for school attendance modelling. Overall schematic flow of the geospatial analysis of out-of-school rates.

Table 1 Model goodness of fit and prediction performance for Tanzania, Cambodia and the Dominican Republic.

	Deviance information criterion	Number of effective parameters	Mean absolute error	Root mean square error	Correlation
Tanzania mainland	911.88	19.32	0.29	0.39	0.62
Cambodia	435.93	45.74	0.11	0.14	0.78
Dominican Republic	409.41	28.04	0.20	0.27	0.87

The goodness of fit was assessed using DIC. The model prediction performance was assessed using the mean absolute error (MAE), the root mean square error (RMSE) and Pearson correlation between the predicted and a 20% validation set.

distance from any population centre to the nearest school was estimated as 6.6 km in Tanzania (mainland), 3.3 km in Cambodia and 1.3 km in the Dominican Republic. This suggested that schools were geographically located at a further straight-line distance in Tanzania compared to the other two countries. This aspect was also reflected in travel time with the mean estimated travel time to the nearest school of 0.8 h in Tanzania (~50 min), 0.4 h in Cambodia (~25 min) and only 0.1 h (~10 min) in the Dominican Republic (Fig. 2).

Covariate selection and model validation. From the covariate selection procedure across the three countries, only temperature variables and night-time light (a proxy for urbanisation) were important statistically in explaining variation in school attendance rates. Travel time to the nearest secondary school (an indicator of geographic accessibility) was not selected for predictive modelling. Therefore, this covariate was used in associating geographic accessibility with predicted estimates of secondary school attendance at sub-national levels (Administrative level 1).

Table 1 lists model prediction performance for each country. For the three models, the Pearson correlation between the predicted estimate and the out-of-sample validation set (20% of clusters) was >60% in all countries. This suggested a good association of the prediction when compared to the observed

data. The mean absolute error was calculated based on residuals between observed and predicted estimates and was relatively small at 0.29 (Tanzania), 0.11 (Cambodia), and 0.20 (Dominican Republic).

The predicted rate of secondary school non-attendance. Fig. 3 shows predictions of the percentage of adolescents and school-age-youths out-of-school in Tanzania, Cambodia and the Dominican Republic at a 1 km spatial resolution. The green areas are those with a low percentage of adolescents and school-age youths not attending secondary school. The second panel shows the difference between the upper and lower 95% Bayesian credible interval as a measure of uncertainty in estimates. Uncertainty is contributed by several factors including survey sampling of the clusters, few data points and the goodness-of-fit of the model. Fig. 4 shows a quadrant level analysis of the percentage out of secondary school and the estimate of adolescents and school-age-youths based on population distribution. Fig. 5 shows scatter plots between travel time and out-of-secondary school rates in the three countries with a fitted non-linear model via generalised additive models (GAM) regression. The corresponding R^2 from GAM regression was 73.3% in Tanzania, 68.8% in Cambodia, and 87.5% in the Dominican Republic.

Table 2 shows that, on average, approximately 57.3 (54.5–58.3) of secondary school age adolescents and school-age-youths were

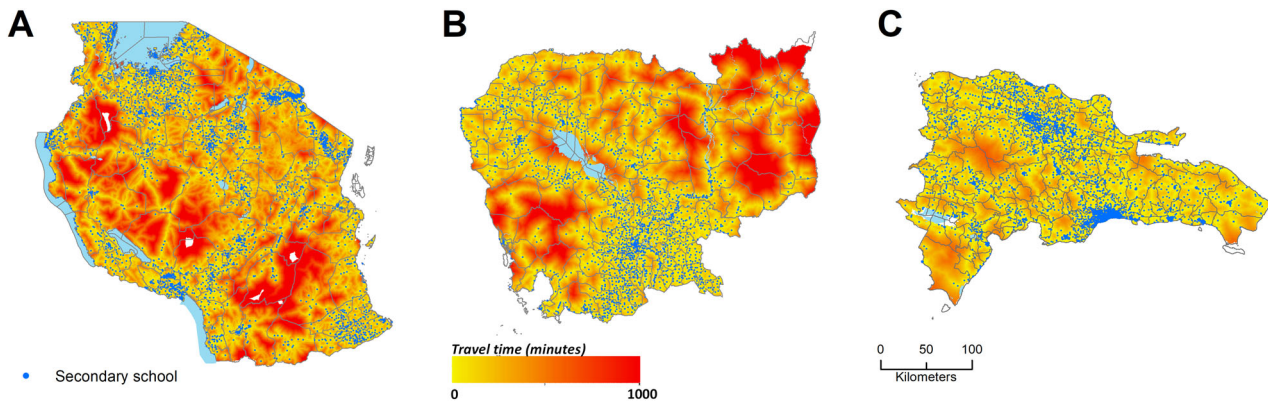


Fig. 2 Secondary school accessibility maps. Estimated travel time (minutes) to the nearest secondary school in the three countries (A) Tanzania, (B) Cambodia, and (C) the Dominican Republic. The blue dots represent the spatial distribution of school secondary schools in the three countries, respectively.

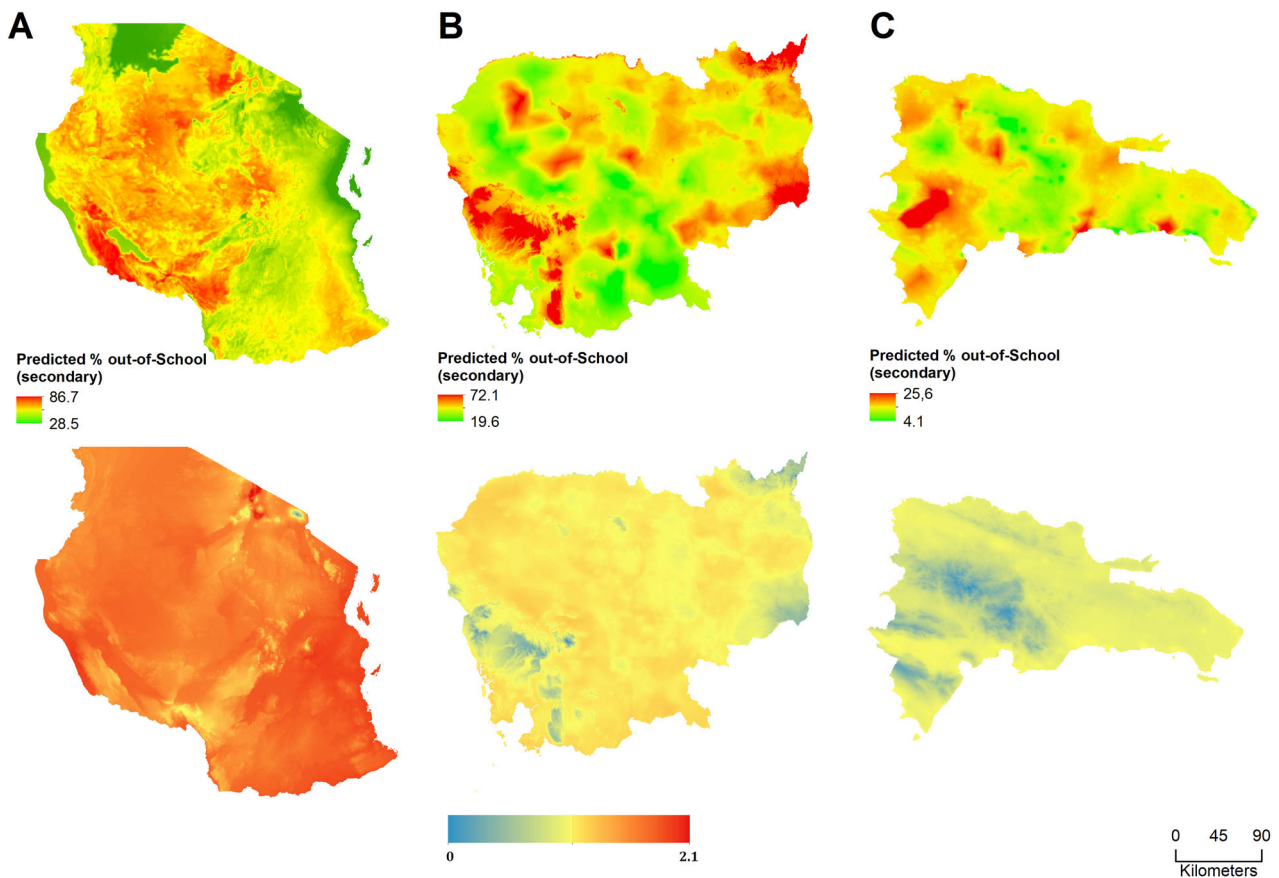


Fig. 3 Out-of-secondary school rates. Maps at 1×1 km spatial resolution of the predicted (mean) percentage of secondary school age adolescents and school-age-youths who were out-of-school in (A) Tanzania, (B) Cambodia and (C) the Dominican Republic. The lower panel maps show the difference between the upper and lower 95% Bayesian credible interval.

estimated to be out of school in the Tanzanian mainland. This translated to approximately 2.8 million adolescents and school-age youths out of school in 2016. The regions with the lowest attendance rates were associated with longer travel times e.g. Tabora, Mbeya and Njombe. There were 8 regions in the Tanzanian mainland with >60% out-of-school rates as classified in the first quadrant of Fig. 4. These were in Dodoma, Katavi,

Mbeya, Mtwara, Njombe, Rukwa, Shinyanga, Simiyu and Tabora. The total number of out-of-school adolescents and school-age-youths in these 8 regions was ~1.01 million, representing more than a third of the 2.8 million out-of-school.

In Cambodia, ~40.0% (37.4–42.3%) were estimated to be out-of-secondary school representing ~0.59 million (annexe Table A2). The Mòndól Kiri region had the largest population, with an

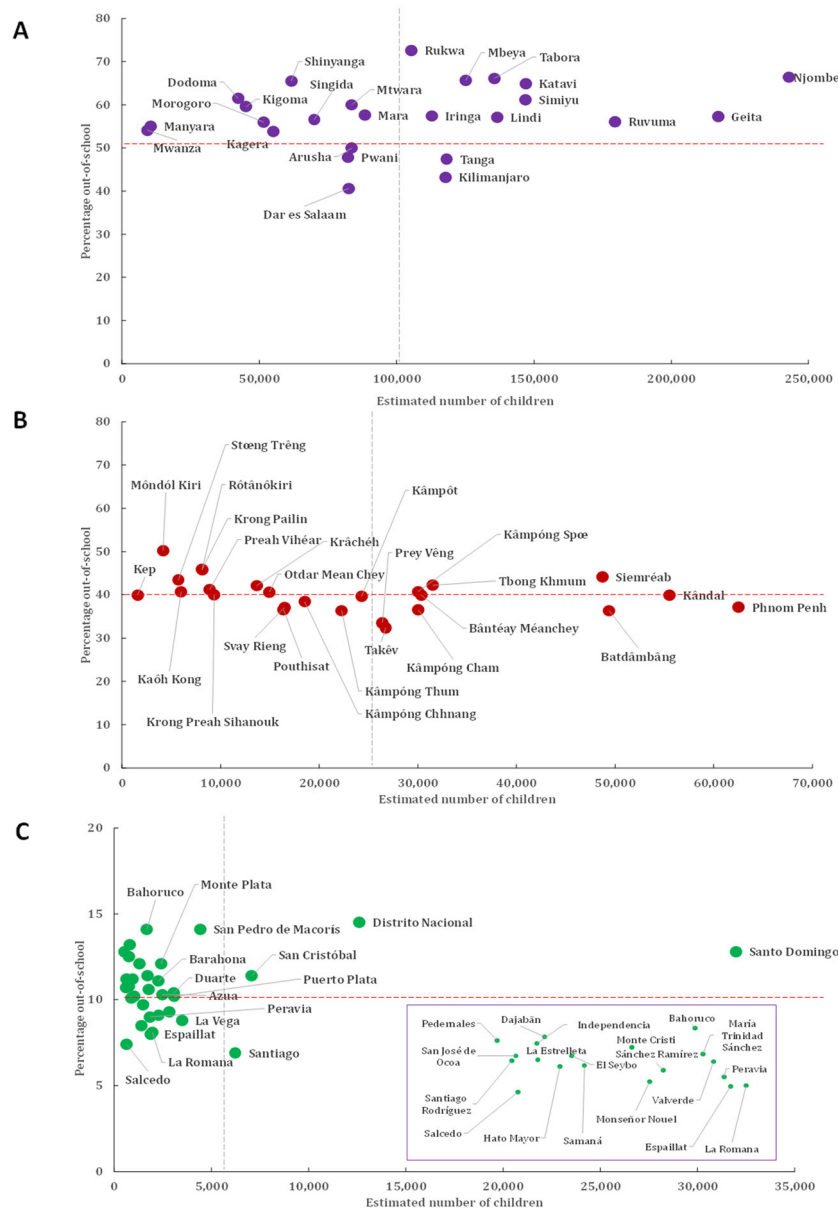


Fig. 4 Quadrant analytical plots. Scatter plots showing the variation of attendance rates by region and by estimated number out of school in **A** Tanzania, **B** Cambodia and **C** the Dominican Republic. The red and grey line show national averages for percentage (y-axis) and number (x-axis) of adolescents and school-age youths out of secondary school.

estimated 50.2% (44.4–58.1%) of adolescents and school-age youths out-of-secondary school. In total, 11 out of 25 regions in Cambodia exceeded the national average of adolescents and school-age youths out-of-secondary school (17%; $n = 170,079$). For the Dominican Republic, the percentage of adolescents and school-age youths out-of-secondary school was lower at 10.7% (9.7–11.7%) representing ~0.1 million adolescents and school-age-youths. However, half of the regions ($n = 17$) in the Dominican Republic exceeded the national average with a population of ~68.2% ($n = 70,398$) of adolescents and school-age youths out of school.

Discussion

This study focused on secondary school attendance for adolescents and school-age-youths in Tanzania, Cambodia and Dominican Republic. In Tanzania, more than 50% of this age group (14–19

years) were estimated to be out of secondary school education (mean 53.8% IQR 51.4–60.2%). Based on estimated distance (Table 2 and Fig. 5), secondary schools were twice the distance (6.6 km, IQR 2.2–19.6 km) and at a greater travel time (0.8 h, IQR 0.2–3.0 h) from the population in Tanzania compared to Cambodia and the Dominican Republic. In Cambodia, the estimated percentage of 13–18 years adolescents and school-age youths out-of-secondary schools was 40.0% (IQR 37.4–42.3%). While in the Dominican Republic only 10.77% (IQR 9.7–11.7%) amongst adolescence and school-age-youths between 14 years to 17 years adolescents and school-age youths were estimated to be out of secondary school. This represented ~2.8 million out-of-secondary schools in Tanzania in 2016, 0.6 million in Cambodia in 2014 and 0.1 million in the Dominican Republic in 2013. Maps of school attendance and geographic access (Figs. 2 and 3) are important in characterising heterogeneities at a fine geographic scale and can be particularly

Table 2 Estimates of geographic accessibility to secondary school, and number of adolescents and school-age youths not attending secondary school in the Tanzania mainland by region.

Region	Distance in km: mean (IQR)	Travel time (h): mean (IQR)	Mean to ratio travel time (h): mean (IQR)	Out-of-school: mean (IQR)	Male 15-20 years	Female 15-20 years	Total 15-20 years	Estimated male 15-20 years out of school	Estimated female 15-20 years out of school	Total 15-20 years out of school
Arusha	6.1 (1-24)	1.1 (0.1-6.3)	0.3 (0-2)	50 (45.8-58.8)	102,413	108,330	210,743	51,186	54,143	105,328
Dar es Salaam	12.9 (9.2-20.6)	0.9 (0.2-1.3)	0.3 (0.2-0.4)	40.6 (38.4-43.1)	263,624	334,498	598,122	107,031	135,806	242,837
Dodoma	11.7 (1.2-30.4)	1 (0.2-2.9)	0.3 (0.1-0.9)	61.5 (57.9-67.7)	116,245	104,203	220,448	71,522	64,113	135,635
Geita	5.6 (2.2-14)	0.8 (0.2-2.5)	0.2 (0.1-0.8)	57.3 (55.4-62.5)	111,489	106,834	218,323	63,259	61,255	125,179
Iringa	4.9 (2.2-16.3)	0.7 (0.2-2.9)	0.2 (0.1-0.9)	57.4 (52.4-67.6)	57,015	50,359	107,374	37,751	28,928	61,680
Kagera	5 (2.1-12.3)	0.9 (0.2-4.6)	0.3 (0.1-0.8)	53.8 (51.6-60.4)	140,680	132,433	273,113	75,722	71,283	147,005
Katavi	8.7 (1.4-28.4)	1.3 (0.1-4.9)	0.4 (0-1.5)	64.9 (63.5-72.1)	32,065	33,167	65,232	20,798	21,506	42,304
Kigoma	5.5 (2.2-13.3)	0.8 (0.2-2.7)	0.2 (0.1-0.8)	59.6 (57.5-63.9)	122,571	123,927	246,498	73,073	73,882	146,955
Kilimanjaro	1.5 (1-4.5)	0.3 (0-1.5)	0.1 (0-0.5)	43.2 (40.4-51.1)	98,695	95,132	193,827	42,591	41,053	83,645
Lindi	4.7 (2.1-11.7)	0.5 (0.2-1.2)	0.2 (0.1-0.4)	57.1 (52.7-63.9)	41,695	37,319	79,013	23,826	21,325	45,151
Manyara	9.3 (2.8-26)	1.4 (0.5-4.2)	0.4 (0.2-1.3)	55 (51.9-67.2)	84,753	76,084	160,837	46,627	41,858	88,485
Mara	5.3 (1.4-15)	0.6 (0.1-1.3)	0.2 (0-0.4)	57.6 (56.3-63.9)	100,319	95,502	195,821	57,823	55,047	112,870
Mbeya	5.6 (1-26.9)	0.5 (0-1.8)	0.1 (0-0.6)	65.7 (61.2-80)	162,524	167,776	330,300	106,810	110,262	217,072
Morogoro	7.1 (1.4-48.1)	1 (0.2-3.3)	0.3 (0-1)	56 (54.5-63.4)	125,573	118,571	244,144	70,289	66,370	136,659
Mtwara	3.7 (1-9.4)	0.4 (0.1-1.4)	0.1 (0-0.4)	60 (58.4-65.3)	59,337	57,378	116,714	35,593	34,420	70,014
Mwanza	3.4 (1-16.4)	1.2 (0.1-1.4)	0.4 (0-0.4)	54.1 (50-65.4)	166,442	165,541	331,983	89,991	89,504	179,495
Niamba	4.8 (2.2-11.4)	0.4 (0.1-1.4)	0.1 (0-0.4)	66.4 (62.1-72.9)	40,245	37,295	77,541	26,739	24,779	51,518
Njombe	4.8 (2.2-35.1)	0.4 (0.1-1.4)	0.1 (0-0.4)	47.8 (43.3-60.3)	58,814	56,371	115,185	28,139	26,923	55,062
Pwani	12.9 (4.2-25.1)	0.7 (0.2-2.3)	0.2 (0.1-0.7)	72.6 (71.9-80.4)	56,065	59,138	115,204	40,702	42,933	83,635
Rukwa	5.4 (3-12.1)	0.7 (0.2-2.3)	0.2 (0.1-0.7)	56.1 (54-61.6)	74,899	71,906	146,806	42,003	42,328	84,331
Ruvuma	6.3 (2.2-15.8)	0.7 (0.1-2.2)	0.2 (0-0.7)	56.1 (54-61.6)	90,314	90,252	180,566	42,003	42,003	84,006
Shinyanga	3.8 (2.2-9.6)	0.5 (0.2-1.7)	0.2 (0.1-0.5)	65.5 (64.2-69.9)	90,314	90,252	180,566	42,003	42,003	84,006
Simiyu	4.7 (2.6-9.6)	0.8 (0.2-1.8)	0.2 (0-0.6)	61.2 (60.2-64.9)	94,853	97,677	192,530	58,035	59,769	117,798
Singida	3.8 (1.4-8.9)	0.6 (0.2-1.7)	0.2 (0.1-0.5)	56.6 (54.2-61.5)	80,398	65,447	145,845	45,516	37,052	82,568
Tabora	15.8 (3.8-39)	1.5 (0.2-4.6)	0.5 (0.1-1.4)	66.1 (64.1-71.2)	138,293	134,137	272,430	91,479	88,730	180,210
Tanga	5.5 (1.4-30.3)	0.4 (0-1.1)	0.1 (0-0.3)	47.4 (41.2-59.2)	116,734	107,711	224,446	55,384	51,103	106,487
Total	6.4 (2.2-16.4)	0.8 (0.2-2.0)	0.2 (0.1-0.9)	57.3 (54.5-58.3)	2,536,054	2,526,880	5,062,934	1,416,676	1,401,441	2,818,117

The number out of school is estimated by gender. For purpose of space, the tables for the Dominican Republic and Cambodia are included as supplementary tables (Supplementary Tables S1 and S2).

important when targeting education interventions. For countries such as Tanzania and Cambodia, a possible geographic-related intervention could be to increase school availability and reduce travel time to secondary schools in regions with poor access (Table 2).

The secondary school non-attendance rate in mainland Tanzania estimated here corroborates previous education research and enrolment data for Tanzania (The United Republic of Tanzania—Government Basic Statistics Portal, 2016; Human Rights Watch, 2017). It is worth noting an average distance of 5 km is commuted twice daily for secondary schools without boarding facilities. The long journey to secondary school contributes to the overall out-of-secondary school numbers estimated to be 2.8 million here, alongside other factors not explored here, e.g. socioeconomic status, individual characteristics (e.g. attitude towards school), cultural factors, home environment, and lack of teachers (Sabates et al., 2010; Inoue et al., 2015; Gubbels et al., 2019). The modelled predictions of the out-of-secondary school population in mainland Tanzania are consistent with findings from the education policy brief report in 2014 of 3 million (Tanzania Education Network, 2018). However, it is not clear, at sub-national levels, that those who enrol at the lower secondary complete (forms 1–4) and progress to the advanced level (forms 5 and 6) (Mashala, 2019). More research is required on secondary school enrolment, drop-out or the likelihood of secondary school completion and retention in low- and middle-income countries.

The average travel time to school in the Dominican Republic was approximately 10 minutes (0.1 h, IQR 0.0–0.2) with ~0.1 million estimated to be out-of-secondary school. The Dominican Republic is smaller in terms of land surface area but had a larger number of secondary schools ($n = 4618$) compared to the mainland Tanzanian ($n = 3258$) or Cambodia ($n = 1615$). The Tanzanian mainland is 19 times larger by geographic size, while Cambodia is 4 times larger than the Dominican Republic. This suggests that physical access to school in the Dominican Republic is boosted by school availability (short distance) relative to population distribution.

There were some limitations in the analyses undertaken. Firstly, the assumptions underlying the analysis of travel times assigned to different land surfaces include a degree of subjectivity, although all the assumptions made during this analysis are based on values derived from previously published geographic accessibility studies. This includes the measurement error in covariates and depending on the country context, these assumptions may lead to over or under-estimation of the actual rates. Secondly, variations in the sizes of schools were not explicitly modelled. The size of the school could be driven by the school location (urban or rural) as well as the size of underlying populations influencing use. The use of night-time light as a proxy for urbanisation adjusted for differences between urban and rural rates of attendance. However, other barriers such as household socioeconomic status, parents level of education, decision-making at a household level, and gender differences (Huisman and Smits, 2009) were not explored further. The focus here was on predictive modelling of non-attendance rates and the aspect of explanatory modelling should be explored by future studies. Lastly, the geographical displacement of the DHS clusters could influence fine-scale estimates of school attendance. A random displacement is applied to urban clusters (by maximum 2 km) and rural clusters (by maximum 5 km), and; an additional 1% of clusters are displaced by a maximum of 10 km. Given the interaction between cluster values and covariates used in the statistical models, the state of error introduced by the displacement was not explored analytically, but has been found to

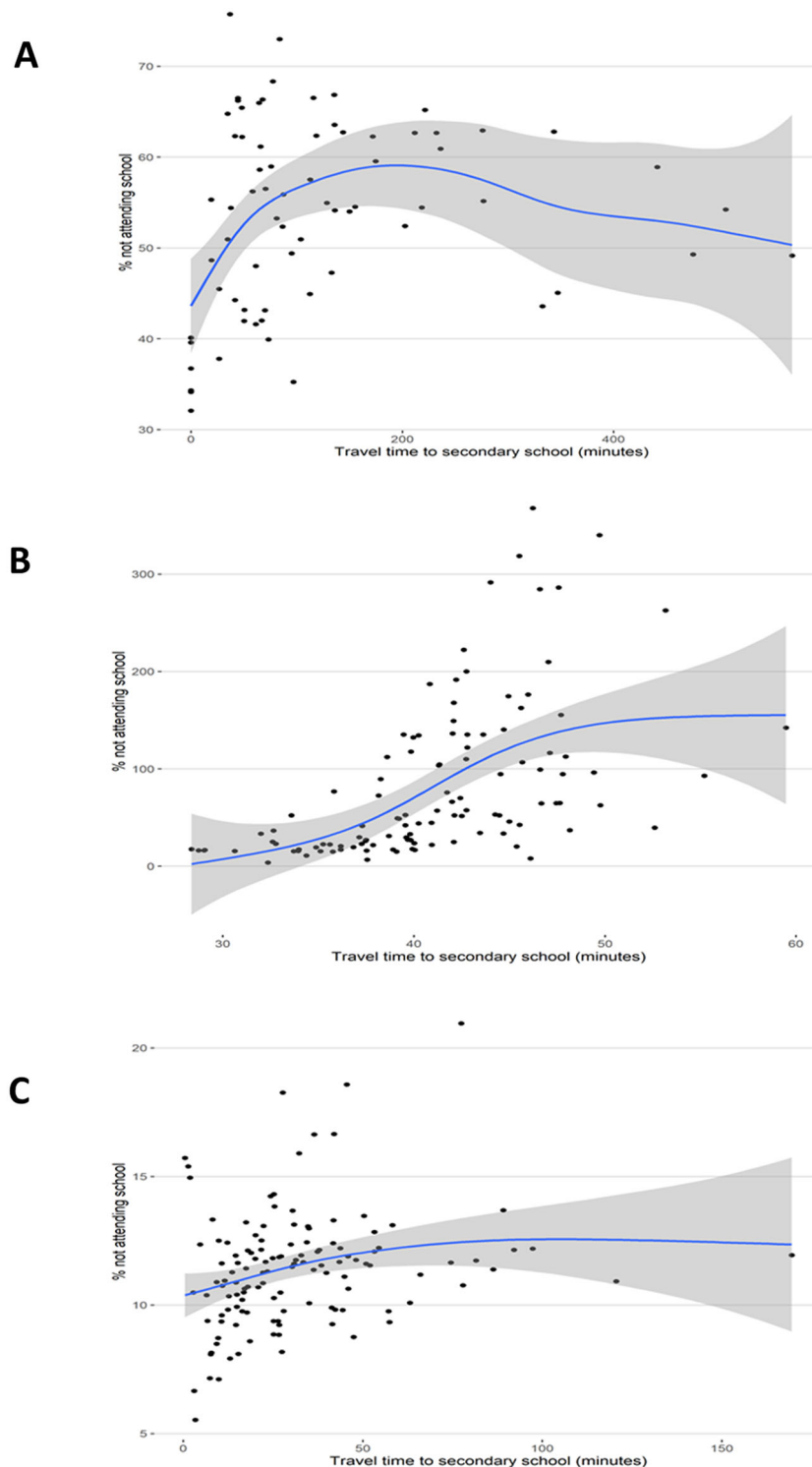


Fig. 5 Travel time and attendance. Scatter plots at the sub-national level (Administrative level 1) showing the association between mean travel time (x-axis) and the modelled posterior mean secondary school non-attendance (y-axis) for **(A)** Tanzania, **(B)** Cambodia and **(C)** the Dominican Republic. The fitted blue line is the non-linear fit via GAM regression with corresponding 95% CI (grey ribbon).

impact such modelling in a minimal way (Burgert, 2014; Gething et al., 2015).

Conclusion

The 2030 Agenda for Sustainable Development may not be achieved without investment in the education of the adolescent and

youths. Thus, unearthing the subnational variation in secondary school geographic accessibility and attendance is important in estimating the fine-scale variation of those physically marginalised from education and provides indicators for monitoring SDG 4.3.1 (UNDP, 2019). Using an example of three low-income and middle-income countries (Tanzania mainland, Cambodia, and the Dominican Republic), the number of adolescents and youth of

school-age out-of-school vary within and between countries and many are physically marginalised. In general, inequalities in access to secondary education has a future impact on national economies and require national investment to remove disparities and ensure no adolescents and youths are left behind. Alongside improving physical access and inequality in these countries, it would be beneficial to investigate at a micro (household-level) and macro-level the role of other factors such as direct and indirect costs, and the quality of provision on out-of-school rates.

Data availability

School data is publically available for all three countries as referenced in the main article and the URLs have been provided. The DHS data is publicly available online through data request <https://www.dhsprogram.com/data/available-datasets.cfm>.

Received: 4 October 2020; Accepted: 27 August 2021;

Published online: 15 September 2021

References

- Anderson JE, Cleland JG (1984) The world fertility survey and contraceptive prevalence surveys: a comparison of substantive results. *Stud Fam Plann* 15:1–13
- Arino O, Gross D, Ranera F, Bourg L, Leroy M, Bicheron P, Latham J, Di Gregorio A, Brockman C, Witt R et al. (2007) GlobCover: ESA service for global land cover from MERIS. In: *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS) 2007*. IEEE International, Barcelona
- Ayad M, Barrere B, Otto J (1997) Demographic and socioeconomic characteristics of households. DHS comparative studies no. 26. Macro International, Calverton
- Banerjee S, Carling PB, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, London
- Breiman L, Spector P (1992) Submodel selection and evaluation in regression. The X-random case. *Int Stat Rev/Rev Int Stat* 60:291–319
- Buchmann C (1999) Poverty and educational inequality in sub-Saharan Africa. *Prospects* 29:503–515
- Burgert-Brucker CR, Yourkavitch J, Assaf S, Delgado S (2015) Geographic variation in key indicators of maternal and child health across 27 countries in sub-Saharan Africa. DHS spatial analysis reports no. 12. ICF International, Rockville
- Burgert CR (2014) Spatial interpolation with Demographic and Health Survey data: key considerations. DHS Spatial Analysis Reports No. 9. ICF International, Rockville
- Cameletti M, Lindgren F, Simpson D, Rue H (2012) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv Stat Anal* 1–23
- Centro De Estudios Sociales Y Demográficos - Cesdem/República Dominicana & ICF International (2015) República Dominicana Encuesta Sociodemográfica y sobre VIH/SIDA en los Bateyes Estatales 2013. CESDEM/República Dominicana and ICF International, Santo Domingo, República Dominicana
- Croft TN, Aileen MJM, Courtney KA et al. (2018) Guide to DHS statistics. ICF, Rockville
- Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. *Biometrics* 65:1254–1261
- Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *J Royal Stat Soc: Ser C (Appl Stat)* 47:299–350
- Elvidge CD, Baugh K, Zhizhin M, Hsu FC, Ghosh T (2017) VIIRS night-time lights. *Int J Remote Sens* 38:5860–5879
- Fao (2000) Land Cover Classification System (LCCS): classification concepts and user manual [online]. Natural Resources Management and Environment Department, Rome, Italy
- Friedman J, York H, Graetz N, Woyczynski L, Whisnant J, Hay SI, Gakidou E (2020) Measuring and forecasting progress towards the education-related SDG targets. *Nature* 580:636–639
- Fuglstad G-A, Hem IG, Knight A, Rue H, Riebler A (2020) Intuitive joint priors for variance parameters. *Bayesian Anal* 15:1109–1137
- Fuglstad G-A, Simpson D, Lindgren F, Rue H (2019) Constructing priors that penalize the complexity of Gaussian random fields. *J Am Stat Assoc* 114:445–452
- Gething P, Tatem A, Bird T, Burgert-Brucker CR (2015) Creating spatial interpolation surfaces with DHS data. DHS spatial analysis reports no. 11. ICF International, Rockville
- Graetz N, Woyczynski L, Wilson KF, Hall JB, Abate KH, Abd-Allah F, Adebayo OM, Adekanmbi V, Afshari M, Ajumobi O et al. (2020) Mapping disparities in education across low- and middle-income countries. *Nature* 577:235–238
- Gubbels J, Van Der Put CE, Assink M (2019) Risk factors for school absenteeism and dropout: a meta-analytic review. *J Youth Adolesc* 48:1637–1667
- Huisman J, Smits J (2009) Effects of household- and district-level factors on primary school enrollment in 30 developing countries. *World Dev* 37:179–193
- Human Rights Watch (2017) I had a dream to finish school. Barriers to secondary education in Tanzania. United States of America.
- ICF International (2012) Demographic and health survey sampling and household listing manual. MEASURE DHS, ICF International, Calverton
- ILO (2010) Micro factors inhibiting education access, retention and completion by children from vulnerable communities in Kenya. ILO, Nairobi
- Inoue K, Di Gropello E, Taylor YS, Gresham J (2015) Out-of-school youth in Sub-Saharan Africa: a policy perspective. *Directions in development—human development*. World Bank, Washington
- Koski A, Strumpf EC, Kaufman JS, Frank J, Heymann J, Nandi A (2018) The impact of eliminating primary school tuition fees on child marriage in sub-Saharan Africa: a quasi-experimental evaluation of policy changes in 8 countries. *PLoS ONE* 13:e0197928–e0197928
- Lehner B, Verdin K, Jarvis A (2008) New global hydrography derived from spaceborne elevation data. *EOS Trans Am Geophys Union* 89:93–94
- Lindgren F (2013) Continuous domain spatial models in R-INLA. *ISBA Bull* 19:14–20
- Martins T, Simpson D, Lindgren F, Rue H (2013) Bayesian computing with INLA: new features. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
- Mashala Y (2019) The impact of the implementation of free education policy on secondary education in Tanzania. *Int J Acad Multidiscipl Res* 3:6–14
- Mcleod AI, Xu C (2008) bestglm: best subset GLM. University of Western Ontario
- Ministerio De Educación (2018) Ministerio de Educación de la República Dominicana. Accessed December 2017. Available: <https://www.ministeriodeeducacion.gov.do/transparencia/conjunto-de-datos-abiertos/1-centros-educativos/2018/listados>
- Ministry of Health CD, Gender, Elderly, Children—Mhcdgce/Tanzania Mainland, Ministry of Health—MoH/Zanzibar, National Bureau of Statistics—NBS/Tanzania, Office of Chief Government Statistician—OCGS/Zanzibar & ICF (2016) Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015–2016. MoHCDGCE, MoH, NBS, OCGS, and ICF, Dar es Salaam, Tanzania
- Morgan C, Petrosino A, Fronius T (2014) Eliminating school fees in low-income countries: a systematic review. *J MultiDiscipl Eval* 10:26–43
- National Institute of Statistics/Cambodia, Directorate General for Health/Cambodia & ICF International (2015) Cambodia Demographic and Health Survey 2014. National Institute of Statistics/Cambodia, Directorate General for Health/Cambodia, and ICF International, Phnom Penh, Cambodia
- NGA (2015) NGA GEONet Names Server (GNS) [Online]. <http://geonames.nga.mil/gns/html/>. Accessed November 2018
- Noor AM, Amin AA, Gething PW, Atkinson PM, Hay SI, Snow RW (2006) Modelling distances travelled to government health services in Kenya. *Trop Med Int Health* 11:188–96
- Ray N, Ebener S (2008) AccessMod 3.0: computing geographic coverage and accessibility to health care services using anisotropic movement of patients. *Int J Health Geogr* 7:63
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc: Ser B (Stat Methodol)* 71:319–392
- Sabates R, Akyeamong K, Westbrook J, Hunt F (2010) School drop out: patterns, causes, changes and policies: Background paper for the Education for All Global Monitoring Report 2011- The hidden crisis: armed conflict and education. UNESCO, Paris
- Shi K, Huang C, Yu B, Yin B, Huang Y, Wu J (2014) Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas. *Remote Sens Lett* 5:358–366
- Small C, Pozzi F, Elvidge CD (2005) Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sens Environ* 96:277–291
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J R Stat Soc: Ser B (Stat Methodol)* 64:583–639
- Statacorp (2017) Stata statistical software: release 15. StataCorp LLC, College Station
- Stevens FR, Gaughan AE, Linard C, Tatem AJ (2015) Disaggregating census data for population mapping using Random Forests with remotely-sensed and ancillary data. *PLoS ONE* 10:e0107042
- Tanser F, Gijbsbertsen B, Herbst K (2006) Modelling and understanding primary health care accessibility and utilization in rural South Africa: an exploration using a geographical information system. *Soc Sci Med* 63:691–705
- Tanzania Education Network (2018) National education policy brief. TenMet, Mikocheni, Dar es Salaam, Tanzania

- The United Republic of Tanzania—Government Basic Statistics Portal (2015) Map of secondary schools location. http://opendata.go.tz/dataset/map-of-secondary-schools-locations/resource/563c0ddc-7781-459e-9b71-37cf3a15425d?view_id=cc889472-4eb0-44cd-b016-acf9a7bfeb19
- The United Republic of Tanzania—Government Basic Statistics Portal (2016) Enrolment in secondary schools by gender and age. President Office—Regional Administration and Local Government (PORALG), Tanzania
- Tobler W (1993) Three presentations on geographical analysis and modeling: National Center for Geographic Information and Analysis. Santa Barbara, CA93106-4060. University of California, Santa Barbara
- UNDP (2019) Sustainable Development Goals [Online]. SDGF, New York., Accessed January <https://www.undp.org/content/undp/en/home/sustainable-development-goals.html>
- UNESCO (2015a) Education for all 2000–2015: achievements and challenges: education for all global monitoring report. UNESCO, Paris
- UNESCO (2015b) How long will it take to achieve universal primary and secondary education?: technical background note for the Framework for Action on the post-2015 education agenda. UNESCO, Paris
- UNESCO (2018) One in five children, adolescents and youth is out of school. Fact sheet no. 48. UIS/FS/2018/ED/48
- UNESCO (2019) Glossary of term: adjusted net attendance rate. <http://uis.unesco.org/en/glossary-term/adjusted-net-attendance-rate>. [Online]. Accessed January 2019
- UNESCO (2020a) Global education monitoring report, 2020: inclusion and education: all means all. UNESCO, Paris
- UNESCO (2020b) World inequality database on education. UNESCO, Paris
- UNFPA (2007) UNFPA framework for action on adolescents and youth. Opening doors with young people: 4 keys. UNFPA, New York
- Wardrop NA, Jochem WC, Bird TJ, Chamberlain HR, Clarke D, Kerr D, Bengtsson L, Juran S, Seaman V, Tatem AJ (2018) Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc Natl Acad Sci USA* 115:3529–3537
- Worldpop (2018) What is worldpop? [Online]. <http://www.worldpop.org.uk/>. Accessed November 2018
- Yu G (2007) School effectiveness and education quality in southern and east Africa. University of Bristol, Bristol

Acknowledgements

The authors would like to thank Dlab, the ministry of education and finance of Tanzania, and the Tanzania National Bureau of Statistics for their insight into the education system

in Tanzania during finding presentation visits in 2019. The authors would also like to thank the staff at the DHS programme, ICF, in particular, Tom Pullum and Courtney Allen, for their support with the DHS data and advice on the construction of indicators, and staff at UNESCO for their support with the UIS Data Centre. The authors also wish to thank Amanda Roden, Alexandra Frosch, Polly Marshall, Sada Saxton and all staff at WorldPop for their invaluable assistance and support in managing this project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-021-00892-w>.

Correspondence and requests for materials should be addressed to V. A. Alegana.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021