Development Engineering 6 (2021) 100064

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

The last mile in school access: Mapping education deserts in developing countries

Daniel Rodriguez-Segura [*], Brian Heseung Kim

University of Virginia – School of Education and Human Development, PO Box 400879, Charlottesville, VA, 22904, USA

ARTICLE INFO

Keywords:
Access to education
Education deserts
School placement
Education in developing countries

ABSTRACT

With recent advances in high-resolution satellite imagery and machine vision algorithms, fine-grain geospatial data on population are now widely available: kilometer-by-kilometer, worldwide. In this paper, we showcase how researchers and policymakers in developing countries can leverage these novel data to precisely identify "education deserts" – localized areas where families lack physical access to education – at unprecedented scale, detail, and cost-effectiveness. We demonstrate how these analyses could valuably inform educational access initiatives like school construction and transportation investments, and outline a variety of analytic extensions to gain deeper insight into the state of school access across a given country. We conduct a proof-of-concept analysis in the context of Guatemala, which has historically struggled with educational access, as a demonstration of the utility, viability, and flexibility of our proposed approach. We find that the vast majority of Guatemalan population lives within 3 km of a public primary school, indicating a generally low incidence of distance as a barrier to education in that context. However, we still identify concentrated pockets of population for whom the distance to school remains prohibitive, revealing important geographic variation within the strong country-wide average. Finally, we show how even a small number of optimally-placed schools in these areas, using a simple algorithm we develop, could substantially reduce the incidence of education deserts in this context. We make our entire codebase available to the public – fully free, open-source, heavily documented, and designed for broad use – allowing analysts across contexts to easily replicate our proposed analyses for other countries, educational levels, and public goods more generally.

1. Introduction

Developing countries have recently made significant strides in improving fundamental educational outcomes like literacy rates and primary school enrollment. For instance, net enrollment in primary school worldwide went from 72% in 1970 to 89% in 2018, thanks to widespread efforts and strategic investments from governments and international agencies (World Bank, 2017a). These encouraging advances have motivated a corresponding change in the policy priorities of development organizations and policy institutions from getting students *into* school, to improving the learning outcomes of students while attending school (World Bank, 2017a). However, despite this meaningful progress in terms of enrollment, much of the developing world is *still* far from achieving universal education. For instance, 1 of every 6 age-appropriate children for primary and secondary school in low-income countries remained out of school by 2018 – a total of 258 million children around the world (UNESCO, 2019).

While the particular reasons students remain unenrolled in school varies by context and individual, available evidence shows that actually having a school physically nearby is *the* first-order necessity for attending school and improving human capital. As Evans and Mendez-Acosta put it, "ultimately, construction is likely a necessary condition for other interventions to work when there are insufficient schools" (Evans and Mendez-Acosta, 2021). As such, ensuring that the full population of a region has reasonable physical access to a school is a critical first step in this pursuit of universal school enrollment. Adequately addressing this need requires that policymakers and researchers identify highly localized areas in which populations lack physical access to school. Yet to date, fine-grain analyses of this kind for developing countries have been logistically and financially prohibitive due to the costs of conducting local surveys and standing up the extensive analytic infrastructure required.

In this paper, we develop an open-source analytic framework to precisely identify areas of lower physical access to schools (i.e.,

[*] Corresponding author.
E-mail addresses: dan.rodriguez@virginia.edu (D. Rodriguez-Segura), brian.kim@virginia.edu (B.H. Kim).

https://doi.org/10.1016/j.deveng.2021.100064
Received 19 February 2021; Received in revised form 30 April 2021; Accepted 30 April 2021
Available online 25 May 2021

"education deserts", per Hillman, 2016) using recently available estimates of the distribution of population across nearly every square kilometer on the planet (WorldPop, 2018). By cross-referencing these publicly-accessible data with administrative records on school locations within a given country – data that are also broadly available and accessible to the public across many contexts – we can empirically quantify the extent to which distance to school is a problem within a given country, and further identify the exact areas, if any, where people do not have access to schools nearby. Prior analyses of educational access, particularly in developing countries, were typically limited to characterizing broad regional tracts, such as counties or departments (e.g., Lehman et al., 2013), or local areas with extensive data collection resources, such as larger urban centers. By comparison, our framework can identify education deserts across nearly every country in the world down to the 1 km^2 level – a resolution substantially more amenable to targeted policy interventions like school construction when paired with the contextual expertise of local policymakers. To provide a demonstration of this analytic framework in the present paper, we exemplify our approach in the Guatemalan context, a country which has historically struggled with educational access and equity.

Ultimately, our analytic framework offers a multitude of actionable insights for policymakers and researchers. First, it allows us to estimate how far individuals in every square kilometer of a country must travel to reach a school – analyzable separately by primary/secondary/post-secondary schools, public/private, or other categories of interest. We further visualize these results using a variety of figures and maps to make the wealth of output easily parsed by policymakers. Second, we can re-contextualize these results by setting a baseline "threshold" norm of what should constitute reasonable physical access to school and thus identify education deserts. For example, if policymakers wish to ensure that every child lives within 3 km of a school (a commonly used international benchmark),[1] our framework can quickly identify what proportion of the population lacks this access, and precisely where those populations are located. Such insights allow for a more nuanced understanding of regional-level enrollment rates and potential barriers to greater enrollment, as well as changes in physical access over time. Third, using this same threshold definition for an education desert, our framework can algorithmically identify school construction sites that would most reduce the share of population living in an education desert and thus maximize the efficiency of school construction as a lever for improving educational access. To illustrate the potential value of this algorithmic optimization, we conduct a simulation analysis in Guatemala and find that building a mere 350 optimally-placed schools based on the algorithm's recommendations from 2008 data would have had the same impact on the share of population living in a public primary school desert as the 7000 schools that were actually opened in the ensuing decade. Finally, we provide guidance for analysts who wish to further refine these analyses to account for geographic factors like elevation, impassable terrain, and similar considerations.

Most importantly, we deliver all of these analytic components in an extensively documented open-source codebase alongside this manuscript designed around the goal of "plug-and-play" utility; assuming an analyst can obtain, at minimum, school location data for a given country context, the entirety of our main analysis can be replicated with minimal effort, zero cost (all requisite software and packages used in our analysis are also free and open-source), and only modest computational resources.[2] This code base is publicly available at: https://github.com/brhkim/mapping-education-deserts, from which the code can be downloaded, and adapted by other analysts. Indeed, while we focus on Guatemala for the body of this manuscript, we include in the appendix parallel analyses for Peru, Costa Rica, Tanzania, Kenya, Rwanda, and South Africa, as a testament to the portable nature of our analysis. Aligning our codebase to analyze each additional country takes as little as 10 min, excluding time for the computation itself. And while our analysis is geared towards assessing the accessibility of schools, our codebase requires only clerical adjustments to instead analyze the physical accessibility of any other statically-located public good (e.g., vaccination sites, water sources, libraries, hospitals, etc.).

While our findings ultimately show that physical access to public primary schools is not a prominent barrier to universal school enrollment in Guatemala, we observe meaningful variation in the extent to which this is true across the country. Moreover, this analysis then offers empirical evidence to suggest that low regional enrollment rates in Guatemala are more likely the result of barriers besides physical access – insights that could prove invaluable for policymakers moving forward. In sum, we argue that as policymakers seek to traverse the last mile in school access and enrollment, fine-grain geolocated data infrastructure and identification algorithms like the one we propose here can offer enormous utility by ensuring that school investments are made in areas where they would have the highest returns in terms of educational access.

The rest of the paper is structured as follows. Section 2 describes the background and conceptual framework for this paper. Section 3 describes the data sources and Guatemalan context we focus on to demonstrate our analysis. Section 4 describes our main methodology, Section 5 reviews our main results for Guatemala, and Section 6 describes how the main methodology can be expanded and adapted to produce additional analytic insights. Finally, Section 7 explores the implications and possible applications of this analysis.

2. Background

2.1. School proximity and educational outcomes

Previous research is clear in highlighting the educational benefits of policies that target school construction in areas which are underserved by educational institutions. In a meta-analysis of the effect of physical inputs on educational outcomes from 1990 to 2010, Glewwe et al. (2014) find that there are five high-quality studies on building new schools in developing countries, which all find consistently positive effects on enrollment and the time the students spend in school. More recently, Evans and Mendez-Acosta (2021) review 6 new studies on school construction in Africa since 2014, finding general increases in enrollment and learning across contexts, and highlighting that these programs seemed most effective when physical access to schools was indeed the binding constraint to school enrollment (e.g., in rural areas with few or no schools nearby). Similarly, in experimental work in Afghanistan, Burde and Linden (2013) find that the construction of community schools that decreased students' physical distance to school increased enrollment by 47 p.p., raised test scores by 0.59 standard deviations, and helped girls more than boys, nearly eliminating the gender gap in enrollment. Duflo (2001) also shows that school construction in places in Indonesia where there were no or few schools led to returns to education of 6.8–10.6 percent in Indonesia – which in turn translated into long-run and intergenerational effects (Akresh et al., 2021) – and Koppensteiner and Matheson (2019) demonstrate that secondary school construction in Brazilian regions previously without schools led to a substantial decrease in teen pregnancy.

Not only is there evidence for the benefits of school construction on educational outcomes, but parents themselves seem to also favor school proximity. For instance, Solomon and Zeitlin (2019) run a discrete-choice experiment with Tanzanian parents, in which they find that parents indeed value outcomes (i.e., school test scores) and school

[1] Theunynck (2009) notes that this norm is in line with the recommendations of the International Institute for Education Planning (IIEP) in Paris and the World Bank (Gould, 1978).

[2] We were able to replicate our analysis on a single consumer-grade laptop, which took approximately 1 h to complete our main analytic components. Analytic extensions will take substantially longer depending on country size but should impose no additional hardware constraints.

proximity more than other inputs such as pupil-teacher ratios and desk availability. They find that the average travel distance to school in Tanzania is about 5 km, but that parents are willing to trade off more positive reported outcomes for proximity. For instance, parents are willing to send their children an additional 1.16 km for a school that scores about 8% higher over the mean on average on a primary exit exam. Conversely, similar work in Kenya by Ngware and Mutisya (2021) found that poor households often sent children to low-fee private schools because of physical convenience, as opposed to other factors like educational quality.

In all, these studies elucidate the idea that enrollment in these contexts is often negatively related to distance to education (i.e. that the distance elasticity of enrollment demand is negative) due to the logistical constraints and costs that greater distance imposes, a dynamic also well-studied in the contexts of U.S. higher education (see Alm and Winters, 2009, for a helpful review) and K-12 school choice markets (He and Giuliano, 2018). Thus, targeted school construction in areas where there are few or no schools seems to be, perhaps expectedly, a powerful way to improve school enrollment, as well as other important indicators along the lines of learning, gender parity, and equality of opportunities more broadly.

2.2. *School proximity as one barrier to access of many*

In spite of the strong evidence in favor of building schools in remote areas with low physical access to schools, little is known about how researchers and policymakers can best understand the extent to which distance, specifically, may be a barrier to enrollment for certain subpopulations and geographic areas on a comprehensive scale. For example, while local school enrollment rates are often referenced as a primary metric of school accessibility, these measures could be driven by a variety of context-specific issues ranging from family finance, motivation, cultural priorities, as well as *physical access* – each of which require drastically different policy interventions in circumstances where resources for such interventions are scarce. Relying on enrollment rates to guide intervention in this manner then masks to a large degree the potential heterogeneity in physical access to schools by region, locality, or settlement pattern. In order to maximize the effectiveness and impact of any investments made in educational access across developing nations, policymakers would ideally be able to differentiate between the previously described scenarios using a data-driven, empirical approach.

As an illustration of this quandary, the World Bank reported in 2016 that 84% of all age-appropriate children in Tanzania were enrolled in primary schools (World Bank, 2016). It is nevertheless unclear what the barriers to access look like for the remaining 16%. One can imagine a scenario where these students *would* attend school if one were available, but currently lack access; conversely, it could be that they currently have physical access, but choose not to enroll for other reasons like fees or high opportunity costs. Both stories would be consistent with the overall aggregate statistic, but they would require drastically different policy recommendations. In the case of the first scenario, policymakers might consider policies like investment in school construction and infrastructure, whereas investment in outreach campaigns or scholarships could likely be a higher priority in the second scenario. In short, without more fine-grain data than aggregate enrollment statistics, it is infeasible to systematically assess the varying educational needs in terms of increasing access to and enrollment in school.

In order to conceptualize the policy issue described here, we borrow the term "education deserts" in the spirit of Hillman (2016). Hillman's study uses data on the location of higher education institutions within commuting zones in the United States – defined by the U.S. Department of Agriculture as clusters of counties that form discrete labor market regions using detailed journey-to-work data (USDA ERS, 2019) – to identify communities that do not have reasonable access to higher education. While we focus on calculating actual distance to primary education in developing countries in the present analysis, the core of Hillman's analysis is the same as ours: the systematic identification of areas without physical access to education given a particular definition for access. More broadly, the international education literature refers to this type of rule regarding optimal school construction and placement as a "norm" (Theunynck, 2009; Lehman et al., 2013), and categorizes distance under the norm of "accessibility and efficiency." Previous policy and research efforts to establish these accessibility and efficiency norms have generally focused on selecting a maximum acceptable distance that children would be expected to travel to school, thus defining the "catchment area" for schools. For example, a commonly applied distance norm is to locate schools within a radius of 3 km from students' homes (Gould, 1978; Theunynck, 2009), though these numbers are often context-specific and can be sensitive to factors like mountainous areas where the effort of traveling such distances can vary greatly. Another example is Lehman et al. (2013), who report that in rural Mali, the distance norm in 2004 was set at 5 km.[3]

While these norms have been pervasive in the theory underpinning school construction, it has long been difficult to actually implement them at scale into decision-making frameworks given the costly and time-consuming nature of collecting such data for any given locality. For instance, Lehman et al. (2013) set out to do this in Mali, across 12 of the country's 70 educational administration districts. Ultimately, only 8 of these 12 intended districts were successfully georeferenced by surveyors, identifying all the schools, villages, and hamlets within them. While the Lehman et al. (2013) report is an extensive and valuable effort to quantify physical access to schools, the dependence on in-person surveying of schools, villages, and population makes the marginal costs of including new areas using this methodology prohibitively high for many. This is true in terms of financial costs, as well as logistical difficulty for areas that may be too remote or afflicted by conflict.

3. Data and study context

3.1. *Data specifications*

Our main methodology, by contrast, requires only two critical data components: the locations of schools across a country (through pairs of latitude and longitude coordinates), and the geographic distribution of population across a country. For the methodological extensions that we articulate in this paper, we further incorporate data on elevation geography to examine the repercussions of alternate "pathing" algorithms to school, a second wave of historical schools and population data to examine trends over time, and regional enrollment rates to facilitate comparisons across traditional and geographic measures of access.

School location data is perhaps the least standardized across contexts of our data requirements in terms of how countries report it, and stands as the primary barrier to replicating our analysis broadly. Still, this information is commonly obtainable through administrative records in many countries, either as latitude-longitude coordinates, or as physical addresses that are easily translated into coordinates through "geocoding." Recent grassroots efforts using commonly available modern technology have also shown that school locations can be "crowdsourced" in contexts where the government has not actively located where all the educational institutions are. For instance, Mulaku and Nyadimo (2011) describe the "Kenyan School Mapping Project," where the researchers identified and geolocated over 70,000 institutions across the Kenyan territory.

As is the case with any secondary data analyses, the exact process and scope of data collection for these administrative datasets will have meaningful repercussions for the robustness and interpretation of applications of our geospatial analysis. Therefore, researchers should be

[3] If a reasonable estimate for the average walking speed of a 12-year-old is 5 km/h (which is faster than for younger children), this would imply a 2-h, daily journey to school (Cavagna et al., 1983).

careful to interrogate these data accordingly before applying the algorithm we propose. For example, what are the formal conditions for a school to be included in the data? Are there relevant institutions likely to be excluded, such as private or parochial schools?[4] And how might such details affect specific areas, contexts, or populations differentially? Moreover, the concept of *location* should itself be interrogated. For example, if studying a context in which schools commonly have several linked campuses, or typically large campuses relative to the resolution of population data used for analysis, using a singular set of coordinates per school could understate access or imply unwarranted precision.[5]

For the purposes of this paper, we use government administrative data that focus exclusively on locating publicly-run primary schools in Guatemala in 2017 (Ministerio de Educación, 2020) and 2008 (SEGEPLAN, n.d.). We expect that other types of schooling in this context are valuable to consider when characterizing the broader landscape of education, but these publicly run schools as tracked by the government are likely the most policy-relevant sample to consider when analyzing, and intervening upon, the public's broad access to educational services. This is particularly true in the context of Guatemala, as primary enrollment in private schools was only 13% of the total primary enrollment in the country (World Bank, 2019).

Our geolocated, fine-grain population data come from the freely available "Global High-Resolution Population Denominators Project" datasets (WorldPop, 2018).[6] These layers provide estimates of human population distribution at a resolution of approximately 100 or 1000 m^2 for all years between 2000-2020.[7] The unusually fine-grain data comes from a combination of census and satellite imagery data, as well as careful application of machine learning algorithms (Stevens et al., 2015), developed through a partnership between School of Geography and Environmental Science at University of Southampton; the Department of Geography and Geosciences, at the University of Louisville; the Departement de Geographie, Universite de Namur, and the Center for International Earth Science Information Network (CIESIN), Columbia University. Discussion of their exact methodology is outside the scope of this paper, but the end result is that these data are highly standardized and available for nearly every country in the world at time of writing. In other words, the need to obtain these fine-grain population data to implement our proposed methodology should not pose a constraint for nearly any application.

One noteworthy feature of the Global High-Resolution Population Denominators Project is that they estimate both *overall* population within each gridded square, as well as *disaggregated* age-sex groupings, for each country. For our present analysis, this means that we are also able to isolate the population estimates to children of school-going age in this context, potentially avoiding some mismatch if relevant children are distributed distinctly from the overall population estimates. This feature will also be of use to researchers interested in other age demographics for certain school contexts (e.g., university-going age) or sex-specific policy margins (e.g., access to school specifically for female students).

That said, we still opt in the main body of this analysis to focus only on overall population estimates. This is because the methodology used to estimate these disaggregated figures impose substantially more functional form assumptions with respect to population growth and change over time (see Pezzulo et al., 2017). For example, if *migration* into and out of the various geographic units is heterogeneous with respect to age groups, or if such patterns are heterogeneous over time (as they use a singular base year to extrapolate population age pyramid ratios over time), it will be more difficult to ascertain how consistently accurate those population estimates are across a geographic context. For simplicity, and to make more transparent the limitations of the present analysis, we focus on the overall population estimates in the main body.[8] We conduct a sensitivity analysis in the Appendix to examine whether our estimates for Guatemala meaningfully change in response to using the age-specific data (children ages 5–14), finding that this distinction is completely immaterial for this particular context. We still urge analysts to consider and weigh this decision carefully for their own use-cases, however.

3.2. *The Guatemalan context*

While our main methodology should be broadly applicable given these relatively modest data requirements, we focus the current paper on Guatemala to showcase our approach for two primary reasons. First, Guatemala is a country which has historically struggled with an array of social challenges, and educational outcomes in Guatemala are particularly weak. For example, in terms of net school enrollment, 86% of school-age children were enrolled in primary school as of 2017 (compared to 94% in Latin America in 2017), and down from 94% in 2008 (World Bank, 2008, 2017b). In terms of learning, the World Bank estimates that 2 in 3 Guatemala children experience "learning poverty", meaning that they are not proficient in reading, even by the time they get to grade 6 (World Bank, 2019a). These challenges are typically worsened by the large inequities along ethnic and geographic lines within Guatemala (McEwan, 2007), given a very diverse geographic landscape with mountain ranges, lakes, and volcanos throughout the southern regions, and deep tropical jungle in more northern areas. Taken together, these challenges in terms of educational inequalities and physical characteristics make Guatemala an appropriate case study to pilot our methodology.

The second reason why we chose Guatemala is because of the public availability of all the needed data sets required for our main analysis and extensions. While our main analysis requires only a single year's worth of school and population data, additional data (such as multi-year school data) offer a useful opportunity to test the methodology's robustness and to assess the extent to which it offers new insight versus traditional measures. As such, this paper is best served by selecting a context that

[4] Note that enrollment in private schools can vary widely by context. As an example, private school enrollment amounted to 82% of all primary school students in Belize (World Bank, 2019b). In such contexts, policymakers are faced with the additional choice of first reducing the number of people without access to *any* school, or prioritizing potentially more populated areas with access to only private schools where parents are burdened by higher private school fees.

[5] Our provided code can account for multiple campuses so long as each are recorded as a separate observation in the school data. Importantly, though, note that recording data in this way assumes access to any one campus is equivalent to having access to any other campus (which would not be the case if a school has geographically separated academic and athletic facilities, for example). As accounting for large campuses would require a substantially different approach to our calculations, and we leave this task to future work where these features are a more critical factor in analysis. Anecdotally, such abnormalities are nearly unheard of in the context of Guatemala.

[6] The specific version of the data used for this analysis is known as the "Top-Down Unconstrained Individual Countries 2000–2020 (1 km^2 Resolution)" dataset. No changes to the algorithm would be required if the data used was the version with resolution at the 100 m resolution. However, this does increase computational time substantially. Analysts focusing on only one country context at a single point in time may opt to use the "Bottom-Up" datasets instead; we encourage all those interested to examine the trade-offs of these datasets closely before use.

[7] WorldPop has not released a schedule of data releases for additional years going forward, but our best understanding is that these data are intended to be maintained over time.

[8] That said, our codebase can accommodate analysts interested in utilizing these disaggregated data simply by pointing the scripts to the disaggregated population dataset, instead. Note that the disaggregated data may require additional preprocessing if multiple demographics are desired (i.e. by adding the raster files together) and a resolution other than 100 m^2 is desired (as the disaggregated datasets are not provided "off-the-shelf" at 1 km^2).

facilitates these valuable comparisons, as these additional data requirements do impose meaningful constraints to the exclusion of many otherwise viable contexts. Finally, note that we further test the "portability" of our method by conducting our main analyses in the contexts of six other developing countries in Sub-Saharan Africa and Latin America for which we could easily find data. We include this analysis in the appendix and in an additional online appendix (https://doi.org/10.1016/j.deveng.2021.100064), and remark on individual data sources there.

Given our present focus on the Guatemalan context, we now move to describe the existing policies that relate to school construction norms to better understand the current business-as-usual. Unlike the distance norms we describe in section 2.2 above, the current Guatemalan policy legislating school construction instead mandates where schools *can* be built, not where they *must* be built (Ministerio de Educación, 2016; Acuerdo Ministerial, 2012). This policy imposes a dual norm: that schools cannot be built within 2 km of one another, and they must serve a minimum number of potential students within their catchment area, which varies by educational level. For our case, primary schools must serve on average 25 potential students per grade in schools with separated grade levels, or 30 potential students per grade in schools with mixed grade levels. The policy moreover allows for a "deficit" of up to 5 potential students in total within a potential area for school construction.

The framing for Guatemala's school construction policy thus does not impose an automatic trigger policy on school construction, and instead places the burden of starting the process for construction on local governments and communities. Underserved communites must compile and submit comprehensive requests to the Ministry of Education with technical details on why the school is needed and how it meets the requirements set out in the aforementioned policy (see for instance, Municipalidad de San José, Chacayá, n.d., or Municipalidad de San José, Pinula, n.d.). We were unable to ascertain the exact process by which communities are mobilized from the ground up to submit these proposals, and by which these requests are ultimately approved in any public sources, academic literature, "grey" literature, news articles, or even anecdotal evidence. It may be the case that these processes are purposefully informal so as to provide the most flexibility for local policymakers to exercise their judgment and contextual knowledge. More cynically, we have evidence in the context of other developing countries that government inefficiencies (Batabyal and Nijikamp, 2004), lack of political representation, ethnic favoritism (Ejdemyr et al., 2018; Burgess et al., 2015), information asymmetries, and coordination problems may each ultimately play a role in the provision of public goods. In either case, it remains likely that our proposed approaches for assessing and meeting school access needs in a data-driven manner provide novel insight against the current counterfactual in the Guatemalan context.[9]

4. Main methodology

The goal of our framework is to systematically identify areas of low physical access to educational facilities in a scalable and reproducible way. Our main methodology consists of a conceptually-straightforward algorithm which estimates the nearest distance from each population pocket to a public primary school, and then analyzes these distances in different ways to compute interpretable statistics and output. Specifically, the method follows these basic steps:

1. Load the fine-grain population raster data from the "Global High-Resolution Population Denominators Project," publicly available for all countries, discretized at either the 100 × 100m or 1 × 1 km plot level. Each discrete geographic unit will be treated as the basic unit of analysis, and each such observation contains an estimate of the number of people that live inside this unit.
2. Load the school location data describing the latitude and longitude of each school.
3. Estimate the straight-line distance ("as the crow flies") between the center of each population unit and its nearest public school.

The output we obtain is a geolocated set of land plots with two key attributes: a) the estimated population living in each plot area, and b) the minimum distance from that plot to a public primary school. From this dataset, we can create several outputs to understand where the areas of low physical access, or "education deserts," are. Since these high-resolution population grids are much more disaggregated than even localized aggregate statistics on school access, we can pinpoint the specific areas where the distance to schools is prohibitively far. For our geospatial analysis, we use the excellent open-source R packages "sf" (Pebesma et al., 2021) and "raster" (Hijmans et al., 2020).

Our approach has three key advantages. First, it is very straightforward to implement and to understand conceptually, facilitating its broad use and easy interpretation by analysts and policymakers. Second, and relatedly, this analysis requires nothing more than a consumer-grade laptop and access to the internet, as all software involved (at least in the implementation we provide alongside this paper) are free and open-source. Third, the data it requires are readily available for many contexts. The fine-grain population data we use is available for virtually all countries in the world, at a resolution of 100 m^2, or 1 km^2 for faster computation. There are moreover other sources that take a different approach to estimating overall and subgroup population data for which our algorithm is also compatible.[10] And as mentioned earlier, many governments already maintain administrative databases tracking the location of schools (such as Education Management Information Systems, or "EMIS"), which are often publicly available, either by default or on request.

The simplicity of our proposed methodology is an intentional decision to offer greater flexibility, allowing it to be adapted and responsive to specific contexts as necessary, but it also makes three important methodological choices that should be stated explicitly. First, the choice

[9] To provide a rough illustration, we conducted a supplementary analysis related to section 6.3 below and examined empirically how many Guatemalan schools could be built that meet the stated policy requirements. In this exercise, we require that schools be built at least 2 km away from one another and serve an age-relevant population of 175 (taking 30 students per grade, minus the allowed deficit of 5 students per grade, times 6 grades). In brief, we find that there are currently 1087 potential areas, with no overlap among them, where a school could be built while abiding by stated requirements. We estimate that if schools were constructed at all 1087 sites, these schools would reach 376,316 age-appropriate students total, or an average of 346 students each. Remarkably, we estimate from administrative data that the average *existing* public primary school in Guatemala in 2017 had 124 students, so many of these potential new schools would not be considered "small" in this context. Note that we opt to use the age-specific population datasets from WorldPop, ages 5–14, for only this analysis. While an imperfect alignment with the true primary age demographic, this seemed the most appropriate data to use for the exercise.

[10] For example, the High-Resolution Settlement Layer (HRSL) datasets, which are the product of a long-term collaboration between Columbia University and the Facebook Connectivity Lab (CIESIN, 2016; Tiecke et al., 2017). Their approach combines intensive survey work with advanced machine learning to estimate the population of every 30 × 30m block in a country, for almost every country worldwide. The disadvantage of this, admittedly more disaggregated dataset, is that the current data for most countries is for a singular year, meaning that if the school data does not match this year, there might be some meaningful mismatch in the analysis. In addition, the WorldPop has open-sourced all of their estimation procedure, code, and underlying data, making their population estimates imminently replicable. This exceptional transparency felt important to privilege and endorse given the nature of our work here, and the likely desire for future users of our code to conduct more rigorous population data diagnostics depending on their specific use-case.

of population pockets at the 1 km^2 resolution clearly defines how granular and precise our analysis is. Although the population data that we use is also available at the level of 100 m^2 resolution, we observe similar results when this population layer is used, but with the important drawback of much higher computational times and memory limits that could put the analysis beyond the computational resources of many users. Ultimately, this decision should be for the user of the algorithm to determine given their context-specific knowledge and the policy action being considered.

Second, and relatedly, we assume that population is dispersed evenly within each geographic unit of 1 km^2 when we calculate distance from the center of each plot to each school. This is because if population is distributed evenly across a 1 km^2 plot, their average distance to school will be equivalent to the distance from the center of that plot, which is what we seek to estimate. That said, this assumption is obviously untenable and may serve to cause some measurement error in our process, but is done so for conceptual and computational ease as before Importantly, this issue becomes negligible when the resolution is sufficiently small (as with the 100 m^2 resolution), and it is actually possible to use the finer-grain population data to “weigh” population within coarser-grain population data. Given what we observed when running our analysis at the 100 m^2 resolution, this assumption is unlikely to be consequential except in very specific cases.

Third, we choose to calculate distance using an “as-the-crow-flies” approach (i.e., a straight line connecting each population pocket to the nearest school). We recognize that this approach is most certainly an under-estimate as it may ignore geographic constraints such as swift elevation changes or lack of a clearly marked path or road. We discuss how to incorporate some of these features into our methodology in the extensions later. However, we decide to use to “as-the-crow-flies” as our baseline measure for several reasons. Much like in the discussion about resolution of the population data, computation time increases substantially by including these factors. Moreover, as we show in the extension later, we find that at least in the case of Guatemala, including elevation changes as a factor does not significantly change the results. Lastly, we believe that the inclusion of other constraints in the landscape should be context-dependent, as a mountainous country with a relatively low number of roads such as Bhutan may need different adjustments compared to a flat country composed of many islands such as the Maldives. As such, we default to the as-the-crow-flies approach and leave it to users to modify this base-level algorithm to their specific needs.

5. Main results

We begin our proof-of-concept analysis by running our main algorithm using the Guatemalan population and primary schools data from 2017. Using the resulting data set, we create several outputs to better understand the nature of physical access to primary schools throughout the country. First, we examine the distribution of distances to school across the whole Guatemalan population. We display this distribution in Fig. 1 (Panel A). The median Guatemalan person lives 0.8 km from a public primary school, and the person at the 95th percentile lives 2.9 km from the nearest school. For comparison, this is lower than the median distance of 2.2 km in Tanzania, the same as in Kenya, and higher than the median distance of 0.5 km in Costa Rica (see the appendix for more details and contexts). This continuous measure can be dichomitized into the share of the population that lives further than a specific distance away from a school, and those that do not, to define the population living in an “education desert.” This threshold distance for living in an education desert, effectively a distance norm, can be varied to explore the sensitivity of the dichotomous measure to different definitions/norms. We show this in Fig. 1 (Panel B), where we calculate the proportion of Guatemalan population living in an education desert on the y-axis, at varying distance thresholds along the x-axis. For example, at a distance threshold of 1 km, 36% of the population lives in a primary school desert. Conversely, at a distance threshold of 5 km, only 1% lives in a primary school desert. For the most commonly used international distance norm of 3 km, only 5% of the population lives in a public primary school desert. Broadly speaking, Fig. 1 suggests that prohibitive physical distances to school in Guatemala only affect a small share of the population, and that a relatively small but targeted school construction initiative might be effective at closing these access gaps.

Beyond quantifying the distribution of physical access to schools as an aggregated metric, our algorithm can also map out these distances to the nearest school for every square kilometer in the country. This type of figure serves as a visual primer on areas with greater and lesser physical access to school across the country, providing valuable insight on geographic heterogeneity in the aggregated measures we described above. In our map of Guatemala in Fig. 2, we see that areas of low physical access (i.e., long distances to school) are concentrated mostly in the northern region (Petén region), and in the southwestern region (around the Escuintla and Santa Rosa departments). We argue that such visualizations allow for far more contextual interpretation of these distance-to-school measures.

6. Extensions to the methodology

As mentioned earlier, the main algorithm we propose in the previous section is relatively straightforward by design to allow enough flexibility in its adaptation across contexts and educational levels. In other words, it could be extended in several ways to yield a more nuanced and tailored analysis for different policy questions in other contexts. In this section, we demonstrate four ways in which our methodology could be modified or refined accordingly. The replication files for all four extensions are likewise publicly available in our included codebase.

6.1. *Before and after comparisons*

One of the simplest extensions that can be made in our framework is the analysis of physical access trends over time, a task we facilitate in our codebase and demonstrate here. In the Guatemalan context, we were able to obtain paired schools and population data for 2008 and 2017, allowing us to compare how physical access in the country has changed over the course of about a decade. Our data shows that between 2008 and 2017, the net number of public primary schools in Guatemala increased by 2,077, or approximately 15%. However, the Guatemalan population between the same period grew from 13.7 to 16.1 million people (18%). Therefore, at its face, the effect of the increase in the number of schools is ambiguous in terms of changes to the aggregate level of physical access to schools. Our methodology can be used to compare two points in time, as we show in Fig. 3 below. Fig. 3 shows that even though population growth outpaced school construction, the distribution of peoples’ distance to their nearest school shifted leftward, i.e., that physical access to school improved over time. That said, this fact should not necessarily be taken as a straightforward endorsement of school placement policy in that period, given that many factors may be contributing to this shift besides targeted school construction. For instance, in the extreme case where population growth was exclusively concentrated in high-density areas with existing schools nearby, the share of the population living far from schools would fall *mechanically* given that the relative share of people living near schools is rising relative to the pre-existing share of people living far from schools, even if no schools were constructed at all. Therefore, instead of being a standalone evaluation of the optimality of school placement over time, this method simply provides one measure for how physical access changed over time in aggregate.

6.2. *Choosing a distance norm*

Policymakers have typically relied on fixed distance norms or thresholds to determine whether a certain population pocket is within a

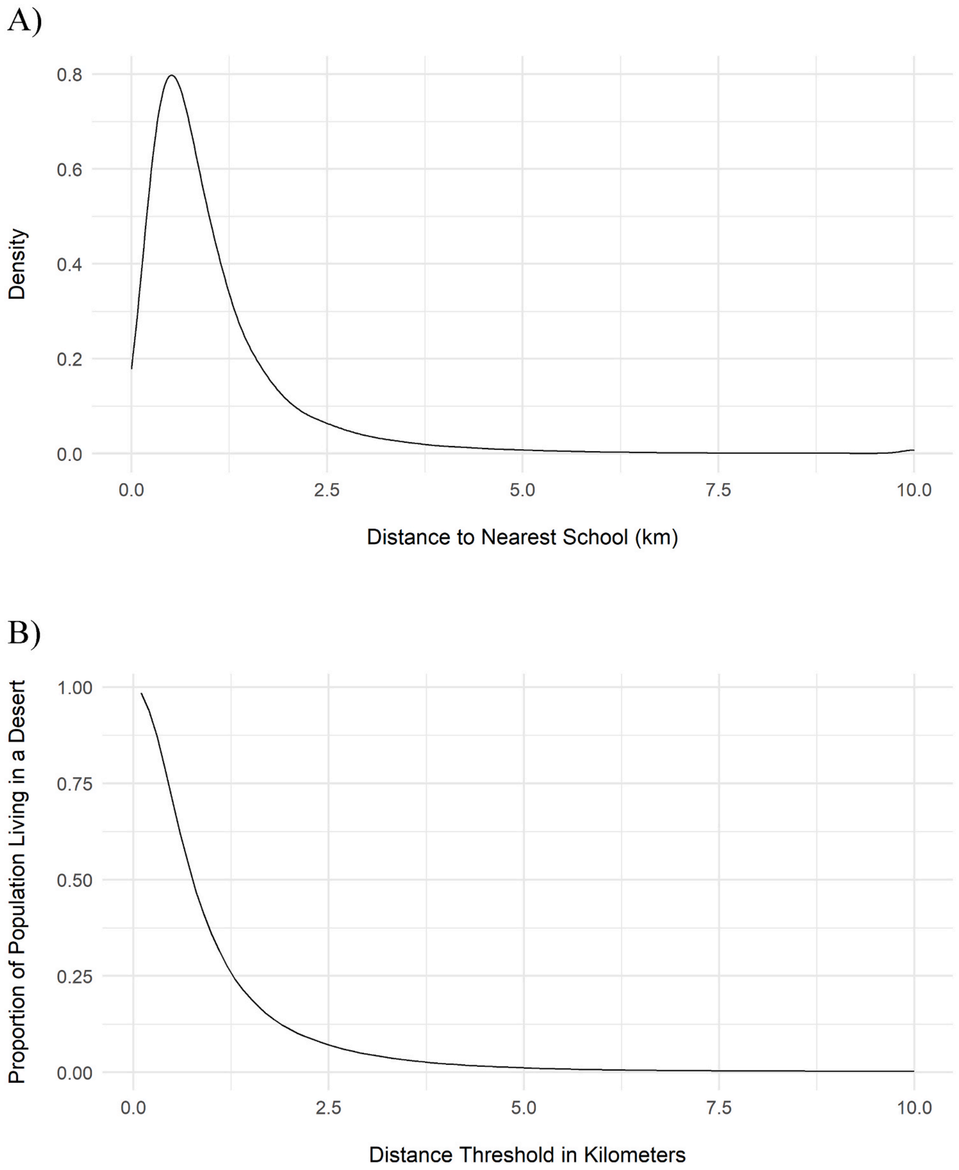


Fig. 1. (Panel B): Proportion of Guatemalan population living in education desert at varying distance norms, (Panel A): Distribution of distance to nearest school across Guatemalan population, Note: Sample subsets to only public primary schools in 2017 in Guatemala.

school's catchment area (Theunynck, 2009; Lehman et al., 2013). This threshold is highly context-dependent, and should be chosen, if at all, by agents with rich knowledge of the specific geographical, infrastructural, social, and budgetary landscape. As such, our main algorithm does not take an ex-ante stance on what this threshold should be, or what constitutes an "education desert." However, the algorithm can be easily modified to accommodate a given distance norm for more in-depth analysis. This dichotomization has two main advantages. First, it most closely resembles the previous work on identifying areas as "education deserts," with the added advantage that this task can now be done at scale in many contexts with minimal data and no surveying costs using our algorithmic approach. Second, it allows for quick identification of the most problematic areas given a certain threshold, offering a clear and interpretable "target" for policy intervention. For example, policymakers and their constituents may find it meaningful to ensure that all students in a given context live no further than X km from school.[11]

To showcase this extension to our main methodology, we choose a tentative threshold of 3 km in the Guatemalan context. Besides this being a common international distance norm, we estimate that just the cost of gas to cover even 3 km to school every day back and forth would lead to an expenditure of 4.4% (USD 7.40) of the average individual income per month in rural Guatemala, not taking into account school fees, books, bike maintenance, or other materials.[12] If instead students take the bus, the monthly transportation cost could be USD 5.20 or 3% of the monthly rural income.[13] These household expenses can start to look prohibitively high, especially for disadvantaged populations, further supporting the use of 3 km as a distance norm. This choice mirrors the spirit of Hillman (2016), where the author examines the distribution of postsecondary institutions across commuting zones in the United States as a proxy for access within a reasonable commuting distance.

[11] For example, the Virginia Community College System advertises that, "If you are in Virginia, you are 30 miles from a community college" (Rorem, 2015).

[12] Assuming an efficiency of 45 km per gallon, an average cost of 2.75 USD per gallon, and an average income in rural Guatemala of 168 USD per month (Voorend et al., 2018).

[13] Assuming a cost of 2 quetzales (0.13 USD) per ride (Cueva, 2020).

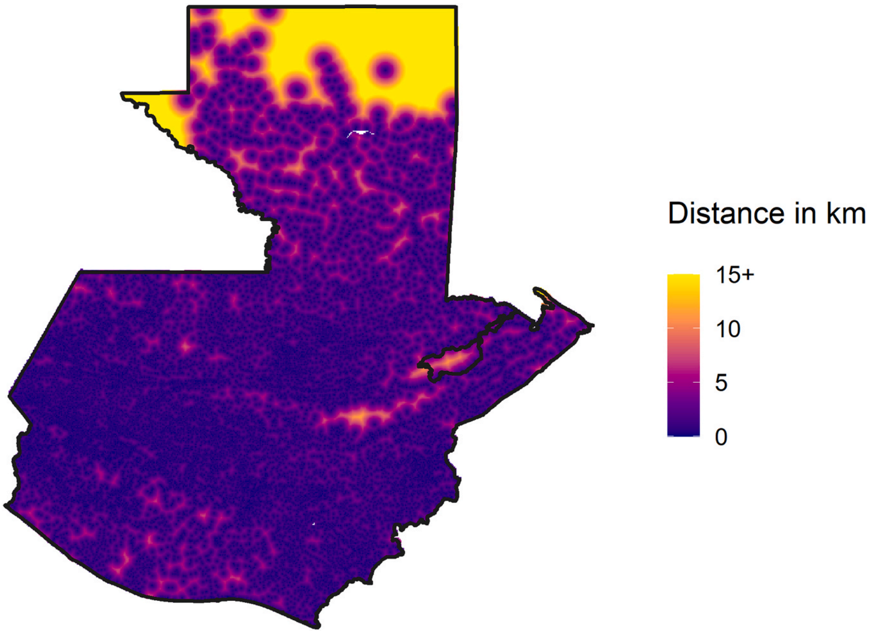


Fig. 2. Heatmap of distance to nearest public primary school by population pocket, Note: Primary school and population data from 2017. Distance is measured as-the-crow-flies from the center of each population plot to the nearest primary school.

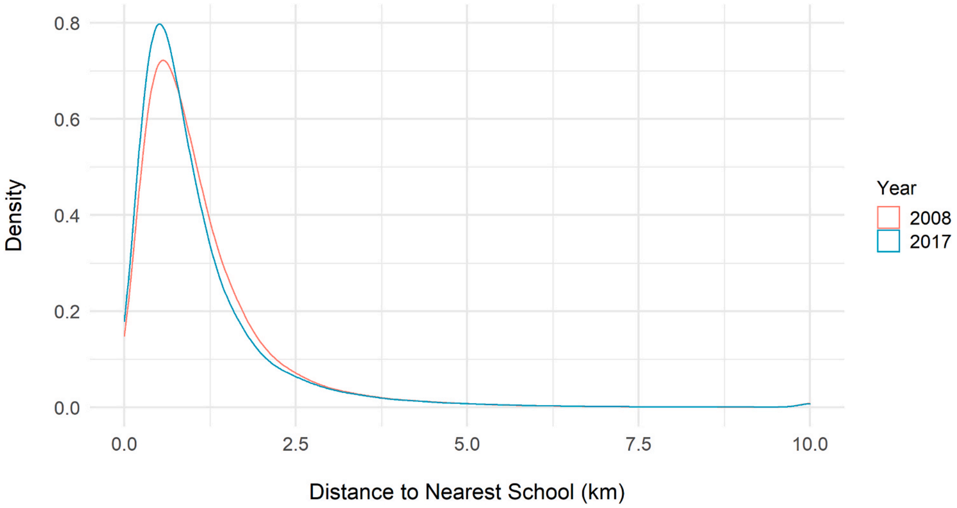


Fig. 3. Comparison of the distributions of distance to nearest school across Guatemalan population in 2008 and 2017, Note: Analysis limited to public primary schools in Guatemala for the years shown.

Fig. 4 below shows the resulting geolocated "education deserts," as defined by a distance norm of 3 km, by plotting only those population points further than 3 km from the nearest primary school. The first panel pinpoints these areas on the map of Guatemala using color to moreover represent the density of population in each of these areas (white representing areas not in an education desert), while the second panel uses the additional dimension of height to more clearly display the relative populations of these deserts. These two panels taken together highlight an important distinction: while most of the land that constitutes "education deserts" is located in the northern regions (Panel A), the real concentration of the population in education deserts is generally localized in the southern regions (Panel B). These figures, much like Fig. 2, can provide an important perspective for policymakers to decide where to strategically locate schools to increase physical access to education.

We can moreover examine the geographic distribution of population in a 3 km education desert and compare these insights against the information provided in traditional regional enrollment rates (defined in this case as the percent of age-appropriate students enrolled in primary school). Fig. 5 (Panel A) displays the same information as Fig. 4 (Panel A) except with regional enrollment rates underlaid in blue. What we observe immediately is that while some regions have high regional enrollment rates, they nonetheless contain several areas, of non-trivial population size, in education deserts. For example, the southern department of Escuintla (annotated with a red "A") has a fairly high enrollment rate relative to other departments, yet still has many pockets of education deserts. Conversely, Totonicapán (annotated with a red "B") has some of the lowest enrollment rates in the country, yet has no incidence of education deserts by our measure. We examine this relationship more explicitly using the basic scatterplot in Fig. 5 (Panel B), where each region is plotted as a single point according to its population, proportion of population in an education desert, and proportion of age-appropriate children enrolled in primary school. If enrollment rates were solely driven by whether people lived in an education desert, we would expect a perfectly negative relationship between regional

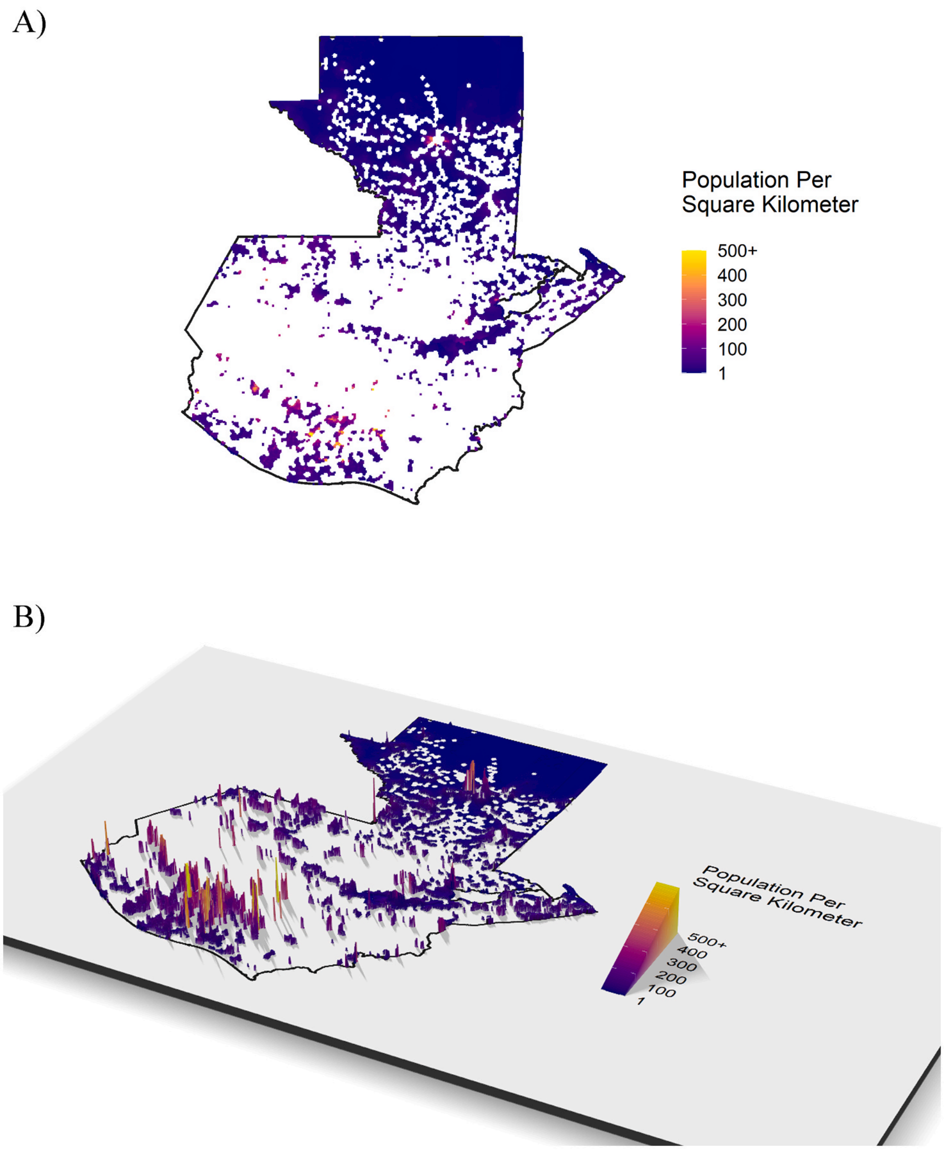


Fig. 4. (Panel B): 3-dimensional geographic distribution of Guatemalan population at least 3 km away from a school, (Panel A): Geographic distribution of Guatemalan population, including only those population points at least 3 km away from a school, Note: Sample subsets to only public primary schools in 2017 in Guatemala. Panel B represents the exact same information as Panel A, but plots both color and height to population count to better visualize differences in population than color alone. 3D visualization made possible by the "rayshader" package from Morgan-Wall (2021).

enrollment rate and their share of population living in a 3 km education desert. Yet what we observe is only a weak relationship; running a simple population-weighted regression of the proportion of population in a desert on the proportion of age-appropriate population enrolled at the department-level, we estimate a coefficient on proportion enrolled of −0.32 (p-value of 0.04 and R-squared of 0.15).[14] This indicates to us, at least on a conceptual level, that our measure of physical access is providing novel information compared with enrollment rates alone, and that the picture remains complex and multi-faceted even after analyzing physical access as we do here.

6.3. Prioritization of school construction sites based on population

A natural extension of the identification of education deserts for a given distance norm is determining how to prioritize these areas given their relative population sizes. In other words, if policymakers were to invest in school construction, what construction locations would most reduce the share of population in an education desert? To do so, we propose an additional algorithm that extends our main methodology. After the main algorithm is applied, we use the previously discussed extension to identify the areas that fall outside of a given distance norm (i.e., the "education deserts"). Then, the new algorithm examines where a school could be constructed (within a 1 square kilometer area) to maximize new population reached given the distance norm. It is able to do this iteratively for any set number of schools to be constructed (i.e., it can produce any number of optimally-placed schools, always taking into account any previously placed schools for the next school). This process can be reiterated until the desired number of schools is reached (e.g., as determined by some budget constraint), or a minimum target of population reached by schools is reached (e.g., "for a school to be built, it needs to have at least X population within its catchment area"). Therefore, this approach is especially helpful to policymakers under constraint conditions: if the budget constraint only allows the government to build a given number of schools, and the goal is to maximize the number of people reached, then this approach can ensure a more efficient placement of schools. Similarly, this approach could be helpful if governments have tiered proposals to address issues of physical access to education. In other words, a government might require a minimum number of people served for a school to be built, and locations that fall below this minimum might be prescribed other policies like remote

[14] Interestingly, this again implies a negative distance elasticity of enrollment demand per our literature review – albeit calculated using less direct proxy measures for both distance and demand. That said, an unweighted regression produces a non-significant coefficient of −0.24 instead.

instruction (such as "telesecundarias" in Mexico).[15] In this case, this extension could help to quickly categorize localities at a large scale.

We test the efficiency of this algorithm at minimizing the share of population in a 3 km education desert by leveraging Guatemala data from 2008 to 2017 to conduct a simple simulation exercise: how different would the share of population in education deserts in 2017 look if Guatemala had used our algorithm in 2008 to determine new school placements instead of its business-as-usual procedure? To begin, we first conduct our main and distance norm analysis on Guatemala using population and primary school data from 2008, and a distance threshold of 3 km. Then, we run our school placement algorithm as described above given these data.

Once that analysis is complete, we determine how many schools Guatemala would have constructed in the time period between 2008 and 2017. Our dataset shows that Guatemala had a total of 14,033 public primary schools in 2008, but of these, only 9040 remained open by 2017. Given that 16,110 schools were on record by 2017, we infer approximately 7070 new schools were constructed by 2017.[16] To be realistic, we assume that policymakers in this exercise would not have known which schools in 2008 were going to close over the next decade, nor how the distribution of population would change by 2017. In other words, they choose to construct and place new schools based only on the "snapshot" of population in an education desert using 2008 data.

We find that if policymakers had placed *all* 7070 new schools using our school placement algorithm and given these parameters, there would not be a single person living in an education desert by 2017; indeed, this feat would have been accomplished after constructing only 3167 optimally-placed schools. That said, we recognize that there exist many other factors determining how new schools are placed, making this scenario fairly unrealistic. For example, Panel A of Fig. 6 shows the cumulative new population reached per new school constructed, demonstrating the quickly diminishing returns to each additional optimally-placed school. This panel also highlights the important caveat that each additional new school would likely lack the requisite student body to justify new school construction well before this benchmark was reached (because building a school to serve a single person would not actually happen).

To explore a more realistic scenario, we proceed to ask the following question: given that the proportion of Guatemalan population in an education desert actually did decline from 2008 to 2017 after the 7070 schools were constructed (see Section 6.1 above), how few optimally-placed schools would it take to produce this same reduction? Panel B of Fig. 6 below displays the results of this thought experiment. The blue line shows the share of Guatemalan population in an education desert across varying distance thresholds, for the actual schools that existed in Guatemala in 2017 – essentially, our target to meet. The red line shows this same dynamic, but under the hypothetical circumstance that Guatemala had constructed *no* new schools at all between 2008 and 2017 – serving as our reference baseline. We find that it would take only 350 new optimally-placed schools to match the actual reduction of population living in a 3 km education desert by 2017, the hypothetical circumstance represented by the green line. Put another way: *350 optimally-placed schools had the same impact on the share of population in an education desert as the 7070 schools actually built between 2008 and 2017*. We take this finding as especially hopeful and actionable for policymakers because it roughly indicates that – at least in the Guatemalan context – substantial strides in physical access can be made even if only *one in 20* schools are constructed with physical access in mind. Conversely, it also makes clear that even a large amount of school construction may not necessarily increase physical access to school across the country by default (e.g., new schools are built in locations already being served by other schools). Policymakers are the best suited to determining when and to what extent physical access should be a consideration for new school construction, but so long as it remains even a minute priority, progress can be made with the help of these proposed algorithms.

6.4. Elevation and geographic features

Our main algorithm relies on estimating distance "as-the-crow-flies", or a completely linear trajectory between the population pockets and school locations. This approach has three key advantages. First, it is a simple and straightforward measurement choice that allows for easy conceptualization of the way in which distance was measured and minimizes the number of contextually-dependent assumptions made about travel patterns, infrastructure, etc. Second, it makes computation vastly faster than other approaches (like the extension we will discuss here). Third, it does not require additional data layers besides what we have described before: solely population data and school locations. Still, all of these advantages come at the expense of ignoring potential barriers like geographic features or lack of roads connecting two places in a fairly linear fashion.[17]

Therefore, we showcase an extension of our main algorithm where we consider elevation changes and compute the "path of least resistance" between a population pocket and a school.[18] Put simply, we first obtain elevation data across Guatemala from ArcGIS's online servers (gspeedAIST, 2019) – though note that robust elevation data are universally available for all regions of the world from a variety of sources. Using these data, we can then calculate how elevation changes when moving from each geographic cell to each adjacent cell.[19] As in our main algorithm, we calculate distances between each population point and each nearby school; however, we instead calculate the distance of the route that minimizes *walking time* after accounting for the fact that speed is inversely related to the steepness of the terrain's gradient (per Tobler's Hiking Function; Tobler, 1993). Any sufficiently steep gradient is considered impassable and avoided for any routing entirely. In practice, this might take different forms. If there is a very large mountain

[15] Telesecundarias are "are a type of junior secondary school that delivers all lessons through television broadcasts in a classroom setting, with a single support teacher per grade" (Navarro-Sola, 2019). Although these schools were initially introduced to deal with issues of delivering education in remote areas, they are also used now in urban areas to deal with issues of poor teacher quality. While these schools do require certain personnel and a physical building, these requirements are less stringent in terms of teacher training and building size. For instance, Navarro-Sola (2019) mentions that the administrative cost per student of telesecundarias is half the cost of brick-and-mortar schools. As such, our algorithm can support the identification of areas that may be best served with a full-fledged school (with the logistical, staffing, and administrative requirements this might pose) versus a lighter investment like a telesecundaria.

[16] Note that these numbers come from the presence of schools by their unique administrative ID in either data set (2008 or 2017). However, if schools simply had their unique IDs changed over this period (e.g., if they merged with another school, took on an additional level, etc.), we would still consider this as a school closing, and another one opening by this tallying method. That said, the precise number of new schools we estimate here is not hugely consequential, given the nature of the results we describe later.

[17] While roadways are an attractive feature to consider, geographically heterogeneous data availability and reliability, as well as computational complexity and costs, make such analysis infeasible and potentially biased for certain contexts (e.g., if roadway data is more complete and accurate in regions of higher income). Given our intention to provide a broadly applicable and easily accessible toolset in this paper, as well as the methodological concerns such analyses present, we opt not to explore this style of analysis ourselves.

[18] We leverage the implementation offered by van Etten and Sousa (2020) in the R package "gdistance."

[19] For computational tractability, we use elevation data at a resolution of 500 m^2. Finer-grain data allow for more nuanced pathing, but also drastically increase computational time and the likelihood of hitting software memory storage constraints.

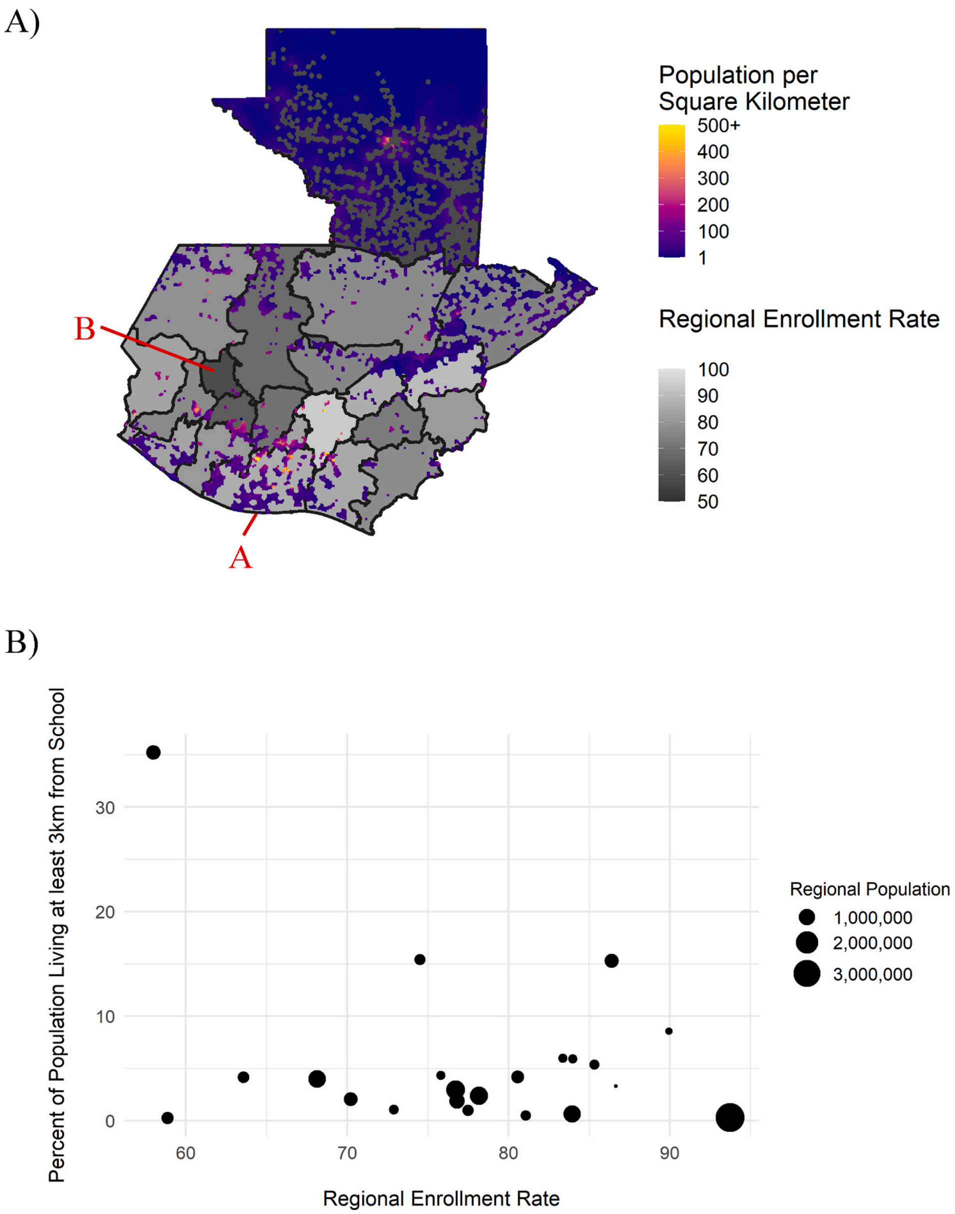


Fig. 5. (Panel B): Scatterplot of regional enrollment rates against percent of regional population living at least 3 km away from a school, (Panel A): Geographic distribution of Guatemalan population at least 3 km away from a school against regional enrollment rates, Note: Sample focuses on only public primary schools in 2017 in Guatemala. Enrollment data were collected in 2016. Panel B displays each region of Guatemala as a point, plotting its population (in point size), proportion of population in desert (on the y-axis), and proportion of age-appropriate children enrolled in primary school (on the x-axis). When running a simple population-weighted regression of the proportion of population in a desert on the proportion of age-appropriate population enrolled at the department-level, we estimate a coefficient on proportion enrolled of −0.32 (p-value of 0.04 and R-squared of 0.15).

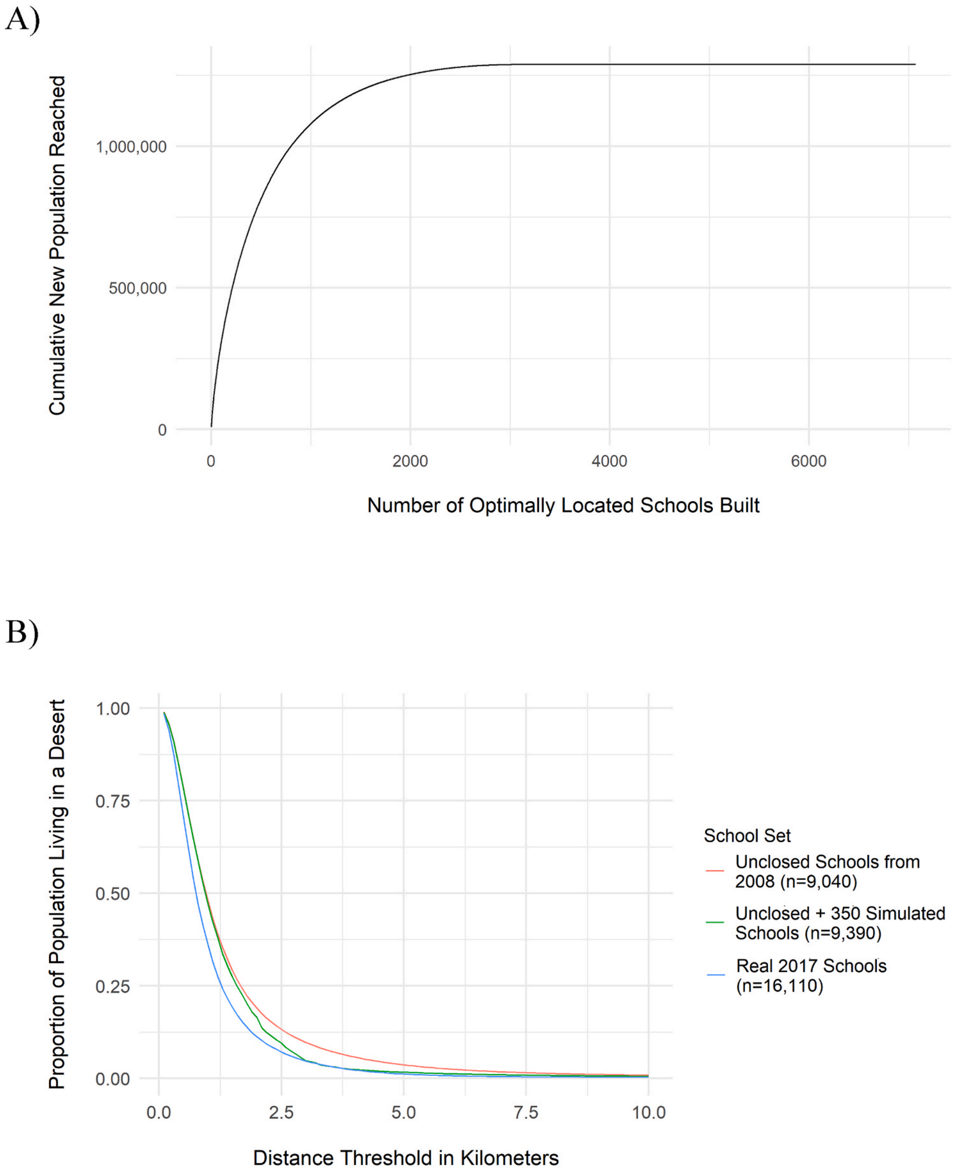


Fig. 6. (Panel A): New population reached per optimally located school. Note: For simulated public primary schools in 2017. (Panel B): Comparison of the distribution of the Guatemalan population living in an education desert in 2017, across several real and simulated school construction scenarios. Note: Population data used are from 2017 regardless of school construction scenario.

between a school and a population pocket, the "path of least resistance" is likely around the mountain. If instead there is a very small hill between these two areas, the path of least resistance might still be a straight line over the hill (depending on the elevation of the hill and its circumference), instead of going all the way around it.

After incorporating this extension to our algorithm, we compare the results to our main results using the as-the-crow-flies methodology for Guatemala in 2017. Fig. 7 (Panel A) plots, for each population pocket, the estimated distance to school using the as-the-crow-flies methodology (x-axis) against the estimated distance to school consider the path of least resistance (y-axis). For visual clarity, we bin observations and scale color according to the sum of population in that bin. The vast majority of population indeed cluster close to the 45-degree line in red, meaning that for nearly all cases, the difference in distance between the two methodologies is small.[20] In fact, Fig. 7 (Panel B) below displays the distribution of the difference in estimated distances between the two methodologies. The vast majority of the observations fall below a 20% difference between the two methodologies. Therefore, in the case of Guatemala, accounting for elevation does not make much of a difference in the identification of where education deserts are, and may come at the expense of increased barriers to analysis (e.g., data requirements, computational costs). However, this extension might be particularly valuable for other hilly or rugged contexts like Rwanda. Importantly, the estimation of the path of least resistance can also accommodate further geographical barriers such as accounting for internal bodies of water or

[20] Note that in all cases, the distance for the algorithm that takes into account elevation is equal or larger than for the main algorithm, since the main algorithm computes a straight line connecting two points.

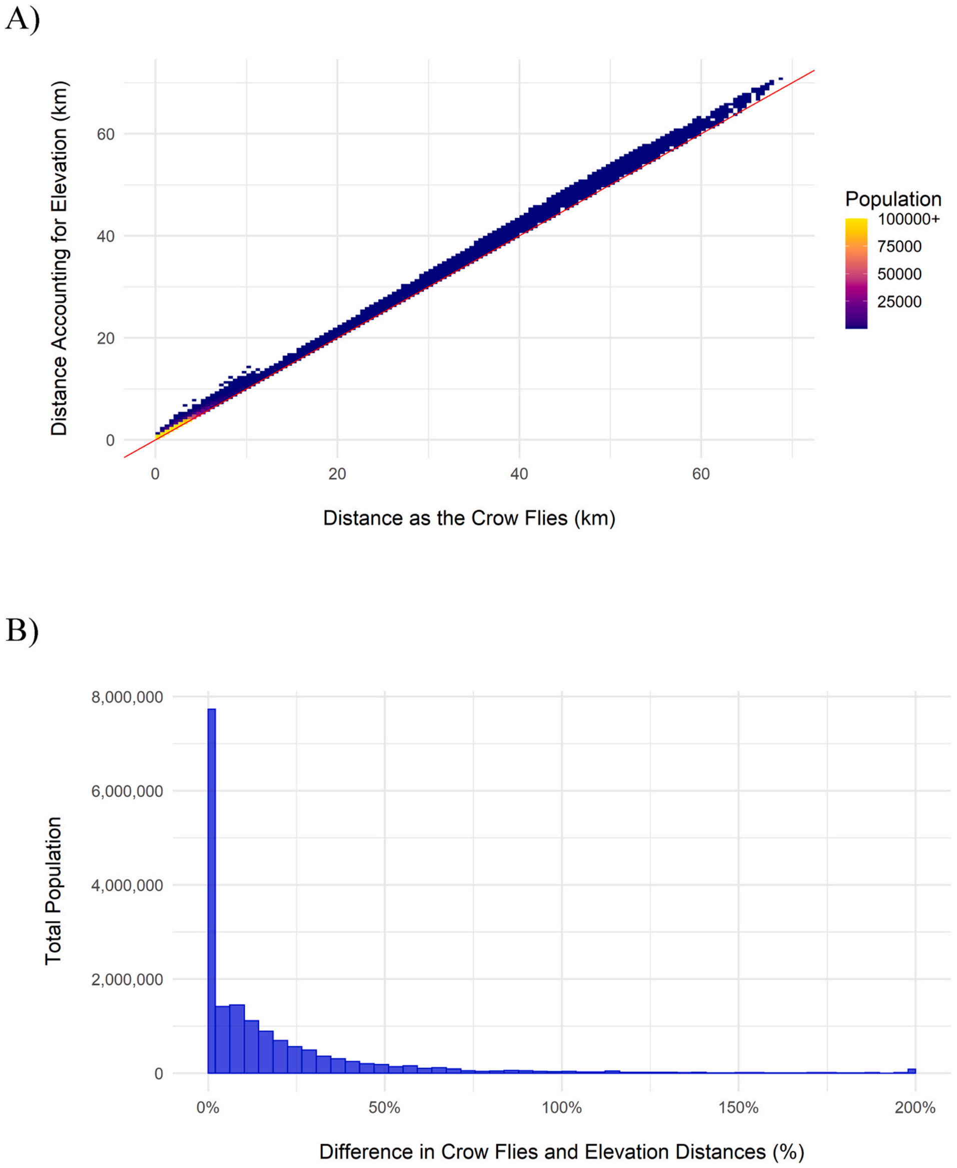


Fig. 7. (Panel B): Histogram displaying the distribution in the difference between "as-the-crow-flies" distances with distances calculated using the "path of least resistance" through elevation changes, (Panel A): Comparison of "as-the-crow-flies" distances with distances calculated using the "path of least resistance" through elevation changes, Note: Sample subsets to only public primary schools in 2017 in Guatemala.

impassable national parks.[21] In this sense, this extension provides the most flexibility to further adapt our main algorithm to local conditions, at admittedly much longer computation times.[22]

7. Discussion

In this paper, we propose a framework to identify populated areas that are not served by public primary schools in developing countries, where surveying costs may be prohibitively high and other types of administrative data may be lacking. We use Guatemalan data as a proof-of-concept to identify geographic areas within the country where individuals lack physical access to primary schooling, as well as to showcase some of the useful extensions we propose to our main methodology. We find that education deserts, defined as pockets of population outside of a school's catchment area, are somewhat rare in Guatemala, and that a relatively few but strategically placed schools could significantly universalize physical access to education.

This type of disaggregated, fine-grain analyses can be especially valuable as policymakers and investors around the world attempt to guarantee universal access to education. If indeed a country has pockets

[21] These could be incorporated in two ways. The first option would be to clip "holes" in the population and elevation raster data files using layers that signal where the national park or water bodies are. The second option would be to change the elevation of these impassable areas to an unrealistically high number. This way, the algorithm will never consider these as viable routes while searching for the path of least resistance.

[22] Conducting this analysis for Guatemala took our workstation computer approximately 9 h, compared with only 30 min for the main analysis. Moreover, we expect the computational time of this extension to increase exponentially with country area.

of population in remote areas where there are no schools, and information is not readily available on where new schools could be more impactful, then it is not clear how to make these investments in a way that creates as much social welfare as possible. Unfortunately, the regions where it is most important to identify education deserts are often the same regions where traditional, aggregate administrative data is typically most lacking. In such circumstances, policymakers would need to resort to either costly surveying endeavors, or fall back on analyses aggregated in larger regions that could critically mask meaningful heterogeneity within those aggregations. By strategically locating educational institutions using these finer-grain analyses and their own contextual expertise, policymakers can indeed ensure that all populations are served by such reforms, at least in terms of physical access to a school.

That said, primary school access is far from the only frontier in which physical access is a relevant consideration for equity and social welfare, and most of the data required to replicate this style of analysis in similar circumstances is publicly available or is of easy access to researchers and policymakers. We thus create and make available highly documented and portable code as a public good for others to recreate and extend our analysis to other contexts. Applying our codebase to analyzing primary school access in additional countries (as we show in our Appendix) can take as little as 10 min, excluding time for data acquisition and computational processing. Similarly, applying our codebase to analyzing the parallel issues of secondary school access – an increasingly prominent goal for many development organizations and governments (Cosentino, 2017) – or postsecondary institution access should be equally straightforward. While outside of our expertise, we also ensured that the codebase should be fully capable of applications to other statically located public goods, for example libraries, health institutions, vaccination facilities, water wells, and so on. In short, if a good can be meaningfully characterized by a coordinate, one can apply our code to better understand a population's physical access to it.

But in closing, we will caution that while applying the code to said contexts should be nearly costless from a logistical perspective, any such analyses should still attend to the many important contextual and data quality considerations we have outlined in this article. For instance, it remains an important critique of our approach that we assume the costs associated with traveling a kilometer in one geographic area is equal to the costs of traveling a kilometer in another geographic area. On its face, this assumption can be entirely untenable – whether comparing within the same country, same region, same city, or even same neighborhood – even after accounting for elevation as we do in the extension analysis above. Analysts must then be cognizant of how their own context and data constraints relate to the value of such analysis in spite of this assumption. To put it concisely, we subscribe to an adapted version of the old adage: if all you have is an education desert mapping tool, everything may look like an education desert problem. We thus ultimately hope that analyses stemming from our methodology provide an *additional* source of insight for researchers and policymakers, to be understood and contextualized in concert with many other sources of evidence, to better serve the public and their well-being more broadly.

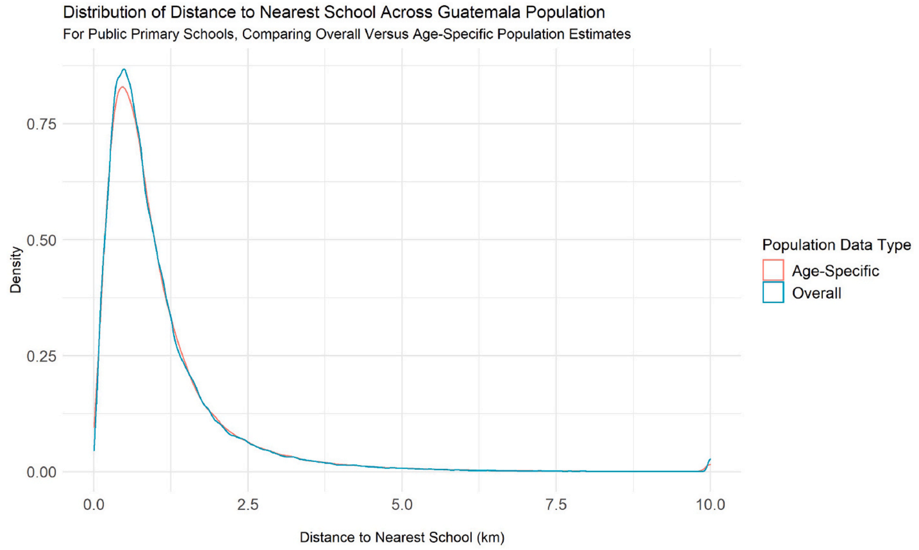
Declaration of competing interest

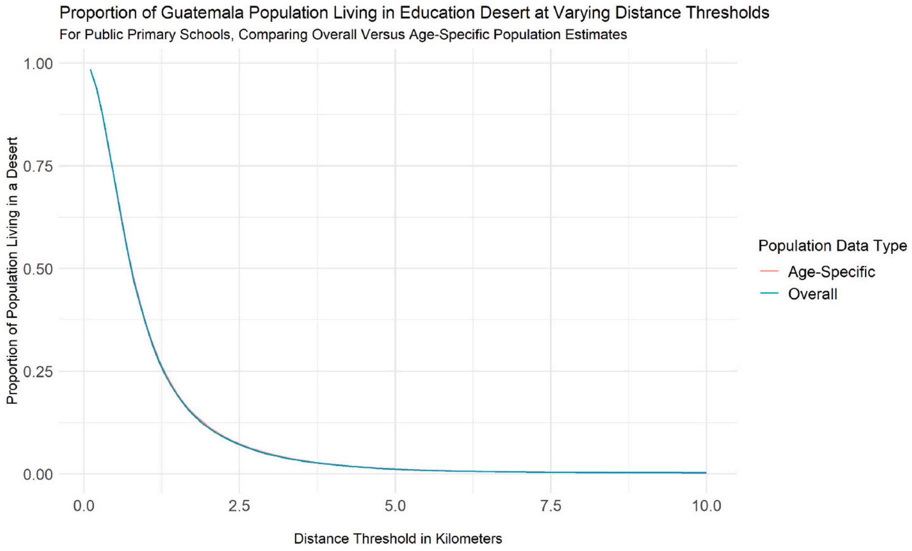
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful for the support and feedback from Benjamin Castleman, Daphna Bassok, James Wyckoff, Erich Purpur, two very helpful referee reports, and our colleagues in the Center on Education Policy and Workforce Competitiveness (EdPolicyWorks) and the Nudge[4] Solutions Lab. This work has been supported by the Institute of Education Sciences, grant #R305B140026, through the Virginia Education Science Training Pre-Doctoral Fellowship program.

Appendix: Comparisons of Results Using Overall and Age-Specific Population Data





Appendix: Other Contexts and Cross-National Comparisons

Our main methodology is primarily designed to identify areas *within* a given region where physical access to education is limited. However, we use this appendix to (1) demonstrate the portability of our analysis, and (2) illustrate some of the considerations when extending the analysis to example multiple countries by adding similar analyses for four Sub-Saharan African countries and two Latin American countries: Tanzania (in 2016), Rwanda (2012), South Africa (2020), Kenya (2018), Peru (2020), and Costa Rica (2020), as we display in this appendix and in an additional online appendix (https://doi.org/10.1016/j.deveng.2021.100064).

We observe two main benefits to cross-country analyses. First, applying this methodology to other contexts allows analysts to create potentially informative benchmarks for a given region of interest. For example, we report in the main narrative that 95% of the population in Guatemala lives within 3 km of a public primary school. In a vacuum, this number is not too informative. But when coupled with distance norms, policy goals, and statistics from peer countries, this can serve as a meaningful data point of comparison. In the case of this metric, Guatemala performs better than all other countries analyzed in Appendix Table 1 except for Costa Rica. Second, this type of comparison can moreover facilitate a rough classification for countries in terms of the issues they face with enrollment. In an ideal world, countries would have high enrollment rates and a low prevalence of education deserts, like Peru and Costa Rica in the table below. Deviations from this categorization can offer a useful shorthand for thinking about extant enrollment barriers. For instance, Guatemala and South Africa can be thought of as having relatively low desert prevalence and low enrollment, while Rwanda can be thought of as having relatively high enrollment *in spite of* high desert prevalence– indicating countries where distance may not be the primary issue for enrollment. We can moreover examine countries where desert prevalence is high while enrollment is low – perhaps contexts where deserts are more impactful – like Tanzania, with 4 in 10 people living further than 3 km from a public primary school and deserts pervasive throughout the country.

We also want to highlight that there are clear challenges in the cross-country comparison of our analyses. First, while the data-generating process for the *population* data is fairly uniform across countries, the data-generating process for *school* data can vary meaningfully by country. As we allude to in section 3 above, what qualifies as a "public" school may vary across contexts (e.g., is it only schools run by governments, or does it also include privately-run government schools?), as well as what qualifies as a "primary" school (e.g., if the grades covered in primary schools differ by location). Similarly, the data collection capabilities of governments may vary, and the degree of missingness for geo-locations can differ as well.

Finally, differences in the actual geographic distribution of a country's population can also affect the usefulness of cross-country comparisons. Costa Rica, where ∼45% of the overall population lives in an extended capital area of only about 2000 km^2 (*Gran Área Metropolitana*), is arguably incomparable to a largely rural context like Tanzania, where the most populous metropolitan area (Dar es Salaam) houses only 11% of its population, and the next-largest city only has about a fifth of this number (Mwanza). This non-exhaustive list of contextual factors can lead to shortcomings in cross-country comparisons in results derived from the methodology we proposed, and as such, these comparisons should be made carefully and sparingly, if at all.

Appendix Table 1
Comparison of "education desert" analyses across countries

Country (year of analysis)	Median distance to a public primary school (km)	Mean distance to a public primary school (km)	Share of the population that lives further than 3 km from a public primary school	Net primary enrollment rate, according to World Bank Development Indicators (latest year available)	Classification
Guatemala (2017)	***0.8***	***1.1***	***4.7%***	***85.6% (2017)***	***Low desert prevalence, low enrollment***
Tanzania (2016)	2.2	5.9	40.6%	83.5% (2016)	High desert prevalence, low enrollment
Peru (2020)	0.6	1.4	11.5%	95.7% (2018)	Low desert prevalence, high enrollment
Costa Rica (2020)	0.5	0.6	3.0%	97.3% (2018)	Low desert prevalence, high enrollment
Kenya (2018)	0.8	2.0	12.9%	80.0% (2012)	High desert prevalence, low enrollment
Rwanda (2012)	1.4	1.7	11.9%	98.8% (2016)	High desert prevalence, high enrollment
South Africa (2020)	0.7	1.1	5.1%	87.0% (2017)	Low desert prevalence, low enrollment

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.deveng.2021.100064.

References

No. 4025-2012 Acuerdo Ministerial, 2012. https://leyes.infile.com/index.php?id=182&id_publicacion=67124.

Akresh, R., Halim, D., Kleemans, M., 2021. Long-Term and Intergenerational Effects of Education : Evidence from School Construction in Indonesia. Policy Research Working Paper; No. 9559. World Bank. © World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/35208 License: CC BY 3.0 IGO.

Alm, J., Winters, J.V., 2009. Distance and intrastate college student migration. Econ. Educ. Rev. 28 (6), 728–738. https://doi.org/10.1016/j.econedurev.2009.06.008.

Batabyal, A.A., Nijkamp, P., 2004. Favoritism in the public provision of goods in developing countries. Econ. Bull. 15 (1), 1–9.

Burde, D., Linden, L.L., 2013. Bringing education to Afghan girls: a Randomized Controlled trial of village-based schools. Am. Econ. J. Appl. Econ. 5 (3), 27–40 (JSTOR).

Burgess, R., Jedwab, R., Miguel, E., Morjaria, A., Padró i Miquel, G., 2015. The value of democracy: evidence from road building in Kenya. Am. Econ. Rev. 105 (6), 1817–1851. https://doi.org/10.1257/aer.20131031.

Cavagna, G.A., Franzetti, P., Fuchimoto, T., 1983. The mechanics of walking in children. J. Physiol. 343, 323–339.

CIESIN: Facebook Connectivity Lab and Center for International Earth Science Information Network, 2016. High Resolution Settlement Layer (HRSL). Columbia University and DigitalGlobe. https://www.ciesin.columbia.edu/data/hrsl/.

Cosentino, C., 2017, September 18. Supporting Secondary Education in Developing Nations. Mathematica. https://www.mathematica.org/commentary/supporting-secondary-education-in-developing-nations.

Cueva, D., 2020, July 29. Transportistas anuncian un incremento de hasta el triple del costo del pasaje. Prensa Libre. www.prensalibre.com.

Duflo, E., 2001. Schooling and labor market Consequences of school construction in Indonesia: evidence from an unusual policy experiment. Am. Econ. Rev. 91 (4), 795–813 (JSTOR).

Ejdemyr, S., Kramon, E., Robinson, A.L., 2018. Segregation, ethnic favoritism, and the strategic targeting of local public goods. Comp. Polit. Stud. 51 (9), 1111–1143. https://doi.org/10.1177/0010414017730079.

Etten, J. van, Sousa, K. de, 2020. Gdistance: Distances and Routes on Geographical Grids (1.3-6) [Computer Software]. https://CRAN.R-project.org/package=gdistance.

Evans, D.K., Mendez Acosta, A., 2021. Education in Africa: what are we learning? J. Afr. Econ. 30 (1), 13–54. https://doi.org/10.1093/jae/ejaa009.

Glewwe, P., Hanushek, E.A., Humpage Liuzzi, S., Ravina, R., 2014. School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010. In: Glewwe, P. (Ed.), Education Policy in Developing Countries. The University of Chicago Press, pp. 13–64. https://doi.org/10.1086/680396.

Gould, William T.S., 1978. "Guidelines for School Location Planning." Staff Working, Paper 308. World Bank, Washington, DC.

gspeedAIST, 2019, March 11. *Guatemala DEM and Hillshade*. ArcGIS. https://www.arcgis.com/home/item.html?id=c9f9b32c4221455ca3600d28c961c642.

He, S.Y., Giuliano, G., 2018. School choice: understanding the trade-off between travel distance and school quality. Transportation 45 (5), 1475–1498. https://doi.org/10.1007/s11116-017-9773-3.

Hijmans, R.J., Etten, J. van, Sumner, M., Cheng, J., Baston, D., Bevan, A., Bivand, R., Busetto, L., Canty, M., Fasoli, B., Forrest, D., Ghosh, A., Golicher, D., Gray, J., Greenberg, J.A., Hiemstra, P., Hingee, K., Geosciences, I., for, M.A., Karney, C., Wueest, R., 2020. Raster: Geographic Data Analysis and Modeling (3.4-5) [Computer Software]. https://CRAN.R-project.org/package=raster.

Hillman, N.W., 2016. Geography of college opportunity: the case of education deserts. Am. Educ. Res. J. 53 (4), 987–1021. https://doi.org/10.3102/0002831216653204.

Koppensteiner, M., & Matheson, J. (n.d.). Secondary Schools and Teenage Childbearing: Evidence from the School Expansion in Brazilian Municipalities. Policy Research Working Paper. No. 9420.

Lehman, D., Buys, P., Atchina, F., Laroche, L., Prouty, B., 2013. The Rural Access Initiative: Shortening the Distance to Education for All in the African Sahel. World Bank, Washington, DC.

Ministerio de Educación, 2016. Manual de Criterios Normativos para el Diseño Arquitectónico de Centros Educativos Oficiales. Policy report, ISBN 978-9929-688-70-4.

Ministerio de Educación, 2020. Establecimiento Educativos. https://datosabiertos.mineduc.gob.gt/dataset/establecimiento-educativos.

Morgan-Wall, T., 2021. Rayshader: Create Maps and Visualize Data in 2D and 3D (0.24.5) [Computer software]. https://CRAN.R-project.org/package=rayshader.

Mulaku, G.C., Nyadimo, E., 2011. GIS in education Planning: the Kenyan school mapping Project. Surv. Rev. 43 (323), 567–578.

n.d Municipalidad de San José, Chacayá, Sololá. Construcción Escuela Primaria Colonia Romec, san José Chacayá, Sololá. http://snip.segeplan.gob.gt/share/SCHE$SNIP/PLANOS_DISENOS/186461-TQBEVQZYMJ.pdf.

n.d Municipalidad de San José, Pinula. Construcción Escuela Primaria y pre-Primaria Colonia Santa Sofía, Municipio de san José, Pinula, Departamento de Guatemala. https://www.guatecompras.gt/concursos/files/1241/6202896%40PERFIL%20DE%20ESCUELA%20SANTA%20SOFIA.pdf.

Navarro-Sola, L., 2019. Secondary School Expansion through Televised Lessons: the Labor Market Returns of the Mexican Telesecundaria. Working Paper. https://laianaso.github.io/laianavarrosola.com/Navarro-Sola_JMP.pdf.

Ngware, M.W., Mutisya, M., 2021. Demystifying Privatization of education in sub-Saharan Africa: do poor households utilize private schooling because of Perceived quality, distance to school, or low fees? Comp. Educ. Rev. 65 (1), 124–146. https://doi.org/10.1086/712090.

Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Müller, K., Pedersen, T.L., Baston, D., 2021. Sf: Simple Features for R (0.9-8) [Computer software]. https://CRAN.R-project.org/package=sf.

Pezzulo, C., Hornby, G.M., Sorichetta, A., Gaughan, A.E., Linard, C., Bird, T.J., Kerr, D., Lloyd, C.T., Tatem, A.J., 2017. Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. Scientific Data 4 (1), 170089. https://doi.org/10.1038/sdata.2017.89.

Rorem, A., 2015, January 20. America's College Promise in Virginia. Stat Chat - A Web Series from the University of Virginia Weldon Cooper Center for Public Service Demographics Research Group. http://statchatva.org/2015/01/20/americas-college-promise-in-virginia/.

n.d. SEGEPLAN. Descargas SINIT: Escuelas de Guatemala (MINEDUC) http://ide.segeplan.gob.gt/descargas.php.

Solomon, S., Zeitlin, A., 2019. What do Tanzanian parents want from primary schools—and what can be done about it? RISE Insight. https://doi.org/10.35489/BSG-RISE-RI_2019/009.

Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., 2015. Disaggregating census data for population mapping using random forests with Remotely-Sensed and ancillary data. PloS One 10 (2), e0107042. https://doi.org/10.1371/journal.pone.0107042.

Theunynck, S., 2009. School Construction Strategies for Universal Primary Education in Africa: Should Communities Be Empowered to Build Their Schools? the World Bank. https://doi.org/10.1596/973-0-8213-7720-8.

Tiecke, T.G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E.B., Dang, H.-A.H., 2017. Mapping the World Population One Building at a Time. ArXiv:1712.05839 [Cs]. http://arxiv.org/abs/1712.05839.

Tobler, W., 1993. Three Presentations on Geographical Analysis and Modeling. http://www.ncgia.ucsb.edu/Publications/Tech_Reports/93/93-1.PDF.

United Nations Educational, Scientific and Cultural Organization Institute for Statistics, 2019. New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School (UIS/2019/ED/FS/56; Fact Sheet). http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf.

USDA ERS - Commuting Zones and Labor Market Areas, 2019, March 26. USDA Economic Research Service. https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/documentation/.

Voorend, K., Anker, R., Anker, M., 2018. *Living Wage Report: Guatemala* (Series 1, Report 16). Global Living Wage Coalition.

World Bank, 2008. World Development Indicators. School enrollment, primary (% net) [Data file]. Retrieved from. https://data.worldbank.org/indicator/SE.PRM.NENR.

World Bank, 2016. World Development Indicators. School enrollment, primary (% net) [Data file]. Retrieved from. https://data.worldbank.org/indicator/SE.PRM.NENR.

World Bank, 2017a. World Development Report 2018: Learning to Realize Education's Promise. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1.

World Bank, 2017b. World Development Indicators. School enrollment, primary (% net) [Data file]. Retrieved from. https://data.worldbank.org/indicator/SE.PRM.NENR.

World Bank, 2019a. Guatemala: Learning Poverty Brief. http://pubdocs.worldbank.org/en/640231571223409894/LAC-LCC2C-GTM-LPBRIEF.pdf.

World Bank, 2019b. World Development Indicators. School enrollment, primary, private (% of total primary) [Data file]. Retrieved from. https://data.worldbank.org/indicator/SE.PRM.PRIV.ZS.

WorldPop, 2018. 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. Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://doi.org/10.5258/SOTON/WP00670. www.worldpop.org.