Energy Policy 37 (2009) 4066–4078

Contents lists available at ScienceDirect

Energy Policy

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

The impact of geography on energy infrastructure costs

Alex Zvoleff [a,*], Ayse Selin Kocaman [b], Woonghee Tim Huh [b], Vijay Modi [c]

[a] Department of Earth and Environmental Sciences, Columbia University, USA
[b] Department of Industrial Engineering and Operations Research, Columbia University, 332 Mudd, 500 West 120th Street, New York, NY 10027, USA
[c] Department of Mechanical Engineering and Earth Institute, Columbia University, 220 Mudd, 500 West 120th Street, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 21 November 2008
Accepted 4 May 2009
Available online 11 June 2009

Keywords:
Energy planning
Energy economics
Electrification

ABSTRACT

Infrastructure planning for networked infrastructure such as grid electrification (or piped supply of water) has historically been a process of outward network expansion, either by utilities in response to immediate economic opportunity, or in response to a government mandate or subsidy intended to catalyze economic growth. While significant progress has been made in access to grid electricity in Asia, where population densities are greater and rural areas tend to have nucleated settlements, access to grid electricity in Sub-Saharan Africa remains low; a problem generally ascribed to differences in settlement patterns. The discussion, however, has remained qualitative, and hence it has been difficult for planners to understand the differing costs of carrying out grid expansion in one region as opposed to another. This paper describes a methodology to estimate the cost of local-level distribution systems for a least-cost network, and to compute additional information of interest to policymakers, such as the marginal cost of connecting additional households to a grid as a function of the penetration rate. We present several large datasets of household locations developed from satellite imagery, and examine them with our methodology, providing insight into the relationship between settlement pattern and the cost of rural electrification.

1. Introduction

Progress in providing electricity coverage has been particularly slow in tropical Africa, i.e. Sub-Saharan Africa with the exception of South Africa (Bekker et al., 2008). In addition to economic factors, a reason frequently cited for this slow progress is that many of the populations in this region live in dispersed rural settlements, making the costs of networked infrastructure intrinsically higher (Haanyika, 2006). The objective of this paper is to directly address the question of how these population settlement patterns influence the cost of electrification. Given limited funding, understanding the impact of the spatial structure of the population on infrastructure costs is critical in the planning phase of a project; however, existing proxies such as population density prove inadequate as predictors of costs in a rural setting.

Given the possibility of acquiring detailed local-level information from remote-sensing data, we develop a new method for estimating a more detailed, per-unit cost of infrastructure, while taking into account population structure. Since settlement patterns influence the cost of local distribution networks while not directly impacting the costs of transmission and generation, we focus on these local networks, and on the problem of how to optimize their per-unit cost. For a broader-scale look at rural energy development, at the national scale, see Parshall et al. (2009).

We first provide a background on the motivating factors for our analysis (Section 2) and then go into our approach (Section 3), testing our methods on a new dataset (Section 4). A discussion of our results is then presented (Section 5).

2. Background

2.1. Rural energy background

The United Nations reaffirmed the "Millennium Development Goals" (MDGs), at the 2002 World Summit on Sustainable Development. While there is no explicit mention of electricity or roads as specific "goals", there is substantial evidence (Modi et al., 2006; Saghir, 2004) that achievement of the MDGs will not be possible without commensurate investments in electricity and transport infrastructure. At present, the International Energy Agency estimates that nearly 1.6 billion people worldwide do not have access to electricity. Governments of countries where a dominant fraction of the population does not have access to grid electricity are now emphasizing the critical role that electricity services play in promoting human development.

* Corresponding author. current address: Department of Geography, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182-4493, USA. Tel: 619-594-8030; fax: 619-594-4938. Tel.: +1 619 594 8030; fax: +1 619 594 4938.
E-mail address: azvoleff@mail.sdsu.edu (A. Zvoleff).

0301-4215/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.enpol.2009.05.006

As they attempt to find the best way to increase coverage, governments also are seeking an understanding of the various modalities of public and private sector contributions. These modalities require a detailed understanding of the cost structure of electrification in rural settings. The broad cost structure consists of generation costs, transmission costs (everything from the power plant up to 33 kV lines, or in some countries 11 or 66 kV) and distribution costs (everything downstream from 11 to 440 V). Many countries foresee a situation where the local distribution is built and managed by franchisees that manage one or more local feeders (e.g. Senegal, Mali). Knowledge of the variability in local distribution costs due to settlement patterns can allow better structuring of arrangements with franchisees and determination of the appropriate level of public sector contribution as one attempts to shift from a state-dominated power system to one where private ownership and market forces play a larger role (Victor and Heller, eds., 2006; World Bank, 2002; Rufin et al., 2003).

Rural energy planning has, thus far, focused primarily on the national to regional scale. Experience has been gained in design and implementation of rural electrification projects at the national level, with reviews such as that by Bekker et al. (2008) in South Africa, and Haanyika (2008) in Zambia providing a basis for future policy and systems design. However, much work remains to be done, particularly on the local level. Large parts of Sub-Saharan Africa remain without access to electricity, and the lack of cost-effective, reliable energy services hampers development (World Bank, 2008).

Interconnection of regional grids and new technologies for distributed power generation and storage offer the potential for providing energy services at reduced cost. Effective implementation of these technologies; however, requires a new planning approach that can consider information across spatial scales. Energy storage, for instance, is a critical component of any distributed energy system; when effectively utilized, storage allows improved efficiency and lower cost generation, as peak demand can be met with less generation capacity (see, for example, ESMAP, 2007). To be effective; however, storage systems must be located appropriately. Transmission losses between generation and storage systems as well as from storage systems to demand centers must be minimized. Optimizing the location of these facilities requires considering the trade-offs involved on both the local and regional level. A better understanding of local-level dynamics is a first step towards this goal.

Agricultural development also requires a better understanding of small-scale energy economics. African agriculture, already vulnerable to seasonal and inter-annual climate variability, will be increasingly tested with climate change. According to the Intergovernmental Panel on Climate Change, by 2050, up to 600 million people are likely to experience increased water stress (Boko et al., 2007). Increased usage of groundwater or water storage for irrigation and domestic supply is one option for increasing the resilience of African agriculture and rural populations (Peacock et al., 2008). However, any substantial new developments will require energy for construction as well as for long-term maintenance and pumping. In addition, water planners must overcome optimization challenges on both the micro- and macro-scale similar to those encountered in energy planning (the algorithm we develop here for energy infrastructure may be relevant for planning water networks), as topography is critically important in the hydrological context. New advances in high-resolution elevation measurement (such as LIDAR) provide sufficient resolution for small-scale study of local topography (Rayburg et al., 2009). Assimilation of this data is essential for irrigation design as well as for accurate estimation of likely energy demand due to water infrastructure.

Although significant progress has been made in large-scale, national, and regional planning and optimization of energy networks supporting rural electrification, new technologies require integration of data across spatial scales to achieve the most cost-effective, efficient solutions. The development of design strategies that manage the unique difficulties encountered in rural areas and that consider local-level data in conjunction with the regional scale picture is important for planning cost-effective rural energy networks. New methods for considering data on the local scale are a first step towards bridging the gap between models. A focus on local-level planning, using household-level location data, offers the potential to optimize energy systems, reduce the high costs associated with dispersed population structure, and maximize the potential of a region for further economic development, while simultaneously increasing the capacity of developing regions to cope with environmental challenges, such as climate variability and change (Yohe et al., 2007).

2.2. Modeling local-level infrastructure costs

Household and hence demand location data is not widely used in large-scale energy models. As local arms of the utility generally have good knowledge of the demand in the immediate vicinity of their network, from a utility perspective the need for such data has not been of urgency. In most areas, location data is either unavailable, or of insufficient resolution to provide useful information for demand modeling. The general practice is to avoid direct consideration of individual users and to instead aggregate demand by assuming full connectivity of the population in a given area. A factor can then be used to estimate the likely demand of the aggregated populace (as in Kaijuka, 2007). Sparsely inhabited areas; however, are expensive to connect, and connection of the entire population of an area may be sub-optimal. The sparse nature of rural populations generally increases connection costs per household compared to urban areas due to longer average distances between households, and often extreme topography. Additionally, recovery of operating and installation costs from consumers is often difficult in rural areas due both to individuals limited ability to pay for service and to the lack of large commercial and industrial users to bear the brunt of the costs. Geographic and other factors might, therefore, make connection of some portion of the population in a given area prohibitively expensive. Current models do not consider this possibility, and do not allow determination of the optimal degree of electrification, or "penetration" in an area.

2.3. Optimizing the penetration rate

An important concern for utilities is the initial decision to extend a network into an otherwise unserved area. Initial expansions are generally limited to strategic demand points, as both the total and per-household (or demand point) investment costs can generally be reduced by excluding distant households from the grid, due to the resultant savings from avoiding very long cable runs and underutilized transformers. However, although costs may be reduced in the short term by such a strategy; the impact on longer term expansion costs must be considered: if such a network is expanded to full coverage, will the final cost be higher than if it had initially been built to serve the full population? To address this question, we examine the role of population structure in offering the potential for cost-savings due to partial penetration, as well as the trade-offs faced with the later expansion of these optimized networks.

To better handle the large amounts of data (individual demand points, or structure locations), as can be gathered using

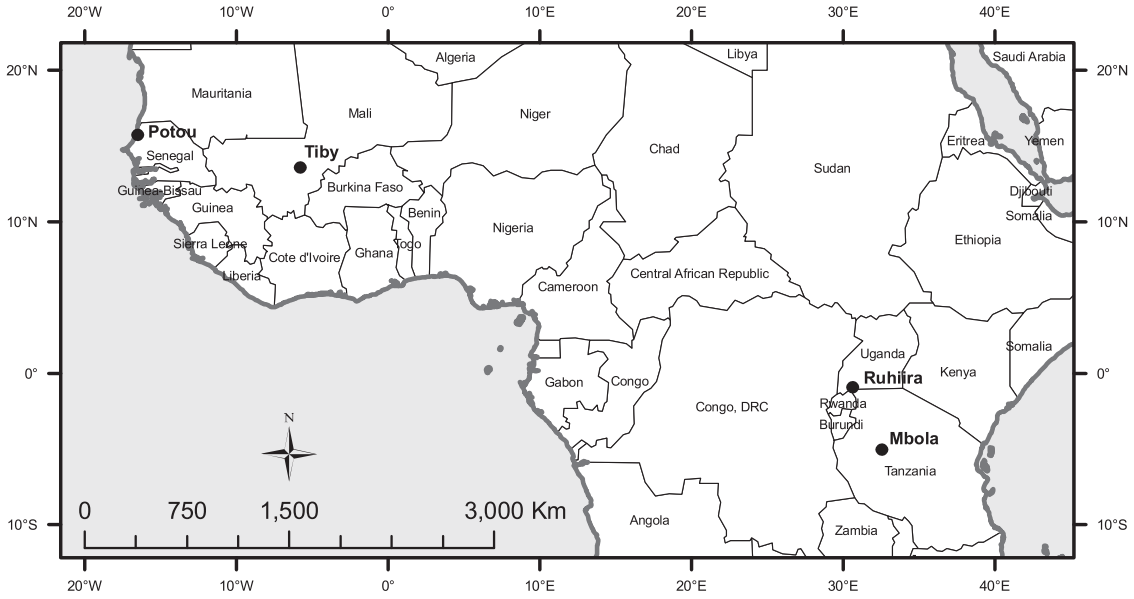

Fig. 1. The locations of the four Millennium Villages sites discussed here.

remote-sensing, while also considering the costs of network construction at a range of penetration rates, we have developed the "composite Prim's algorithm" (CPA) as a simple heuristic approach to optimizing the penetration rate.

3. Methodology

To consider the cost-effectiveness of rural electrification in a given area, two methods are developed. The first is an index to allow simple comparison between sites of the similarity of population structure. The second is an algorithm to construct a full profile of the cost of electrification with increasing penetration. While the index provides a simple metric for considering the spatial structure of a site, the algorithm we propose is useful for determining the optimal extent of electrification in a particular area, with the results used to feed regional and larger scale models, and to better inform design and policy decisions.

To test the algorithm and indices, we use a new dataset (presented in Section 3.1) of structure locations gathered at four sites in rural Africa. After clarifying our terminology for the qualitative description of settlement pattern (in Section 3.2), we present three approaches for characterizing the cost of rural energy infrastructure investment as a function of population structure (Sections 3.3–3.5).

3.1. Data

The algorithms are tested on a new dataset collected from several of the African Millennium Villages sites. The Millennium Villages project focuses on empowering villages through public investments to develop critical infrastructure and social capital needed for achieving the United Nations Millennium Development Goals. Further background, as well as preliminary project results, is discussed in Sanchez et al. (2007).

Structure-level location data has been gathered from nine of the Millennium Villages sites; four sites (Mbola, Tanzania; Potou, Senegal; Ruhiira, Uganda; and Tiby, Mali) that provide examples of commonly encountered population distributions were selected for discussion here (see the map in Fig. 1, and see Table 1 for basic statistics on the sites). The datasets collected for the Millennium Villages sites generally have around 3000 structures, with some sites having in excess of 17,000.

Table 1
QuickBird imagery collected for the four sites.

Site	Image capture date(s)	Area of image (km^2)	Number of structures
Mbola	7/3/2007	100	1175
Potou	1/26/2007	95.5	1797
Ruhiira	6/22/2003, 10/13/2006, 4/29/2007	100	6434
Tiby	12/17/2005	100	2496

The structure location data used in our model to estimate network costs were digitized from QuickBird satellite imagery of each site. The images were obtained in two forms: a four band multispectral image at 2.4 m resolution and as a panchromatic image at 60 cm resolution. Prior to digitizing structure locations, the 2.4 m resolution imagery was pan-sharpened to 60 cm resolution, and then orthorectified using a 90 m resolution SRTM digital elevation model. An appropriate stretch (generally linear) was used on each image to improve the contrast, depending on cloud cover and on the overall brightness of each image.

Structure locations were hand-digitized from the processed imagery, generally at around 1:2000 scale. A 1 km^2 grid was used on each image to ensure the entire image was covered during heads-up digitization. On each image, all visible roofs (referred to here as "structures") were manually recognized and marked (see Fig. 2). Each structure does not necessarily correspond to an individual "household". Some households, for example, might possess more than a single structure, using individual buildings for living or storage spaces. These distinctions; however, are difficult to detect from the air. We here consider "full penetration" of an area to mean complete electrification of all structures—future work is required to better understand the relationship between structures detected from the air and actual population densities (which are dependent on the number of individuals per household).

For most sites, a single QuickBird image, generally covering a $10 \times 10\,km^2$ area, was sufficient to cover the entire site. In areas

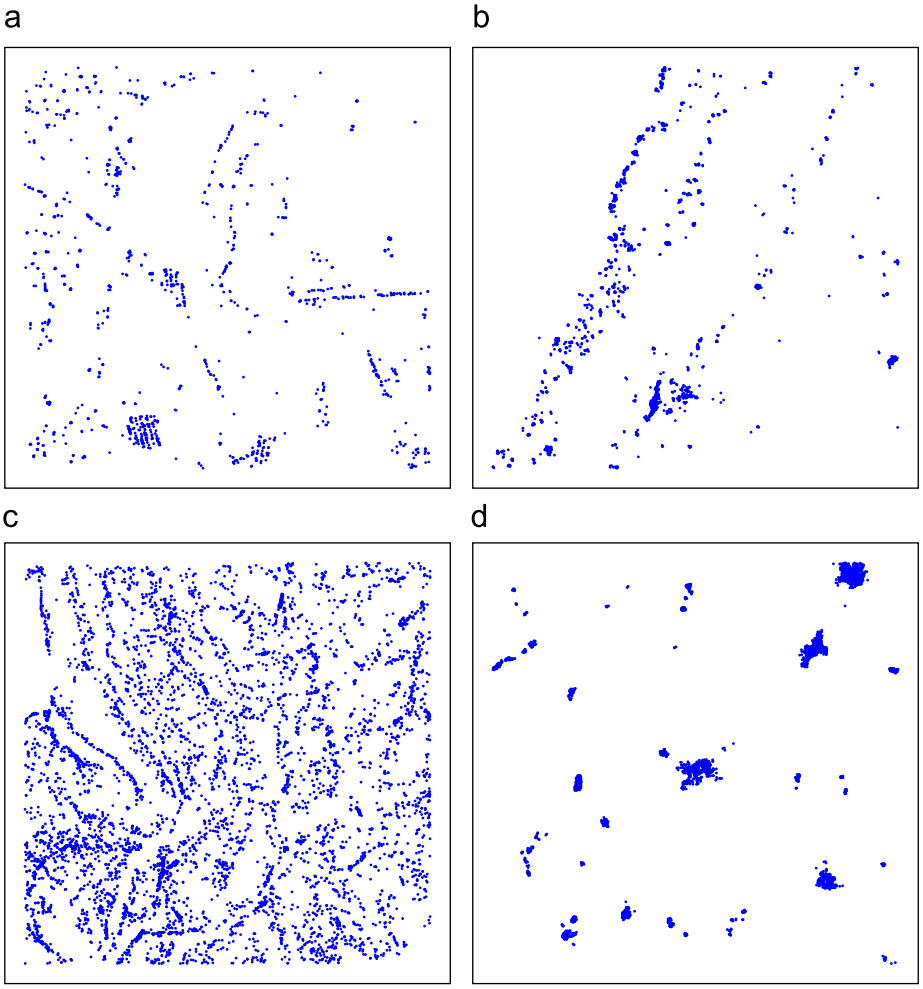

Fig. 2. Structure locations at the four sites: (a) Mbola, (b) Potou, (c) Ruhiira, and (d) Tiby. Each dot represents a single structure. Each site is approximately 10 by 10 km.

where a single image covered the entire site, all points were weighted evenly. However, several larger sites (Ruhiira, Uganda is discussed here) required the composition of multiple images to fully cover the required area. In these cases, points were digitized off each individual image in turn, and the multiple resulting datasets were combined after manually checking the data to prevent duplication of information in the final combined set of points. In Ruhiira, three images were used, two acquired in 2003, and one in 2007. A significant area of overlap between the 2003 and 2007 imagery allowed estimation of the degree of population growth that occurred in the intervening four years. To account for this growth, a correction factor was applied to points in the 2003 dataset so as to weight them more heavily in the analysis than the 2007 points.

3.2. Qualitative evaluation of settlement pattern

Although the data acquired thus far shows a broad range of spatial distributions of populations (see Fig. 2), there are common features that have been observed. Qualitatively, sites can be classified as sparse or dense, and nucleated or dispersed. A sparse site has a relatively low density as compared to a dense site. A nucleated site shows clustering of population around certain centroids, whereas a dispersed site is closer to a random distribution (a Poisson distribution) of points around the landscape.

Using these relative terms, a site can therefore be nucleated but sparse, or dense yet dispersed. While Ruhiira, Uganda is densely populated (compared to the other areas we discuss here) it is not nucleated, although topography and local road networks influence settlement locations to some extent. Tiby, Mali exemplifies the most nucleated areas, with a small number of very dense clusters of structures dotting the landscape. Potou, Senegal and Mbola, Tanzania fall in between these two extremes. Potou shows nucleation in the south and along the coast in the west, and is otherwise dispersed. Mbola features one large cluster in the south, while the remaining area is sparse.

3.3. The homogeneity index

Based on the household location data that we have collected and processed, it is evident that the distribution of households displays distinct characteristics for each of the regions under study. To facilitate the quantitative comparison of spatial structures among regions, we propose a new measure called the homogeneity index (HI) that serves as a proxy for the degree of dispersion of the households within each region. This measure is a variation of the classical nearest neighbor index due to Clark and Evans (1954), which we describe below prior to introducing the exact definition of the HI.

For a given set of n points, let R_{near} denote the average distance between each point and its nearest neighbor (where the average is

taken over the set of points). Now, as a benchmark, consider the problem of maximizing this quantity by changing the location of these n points within a given area. It turns out that one can solve this problem by arranging these points in a regular repeated hexagonal shape such that each point has six nearest neighbors (similar to a bee-hive shape). In this case, the maximum possible value of R_{near}, which we now denote by R_{max}, is given by $R_{max} = (2^{1/2})/(3^{1/4}\rho^{1/2})$ where $\rho = n/A$ denotes the density and A the area. The nearest neighbor index of Clark and Evans is defined as the ratio R_{near}/R_{max}. By definition, this index ranges between 0 and 1.

To better estimate the impact of the spatial distribution of structures on network cost, we modify the definition of the average distance by considering the minimum spanning tree (MST) (see Section 3.4 for details on calculating this network). Given a set of structures, the MST is a collection of segments (each connecting two structures) such that all the structures are connected and the total length of the network is as small as possible. For the discussion in this paper, all structures are given an equal weight except for correcting differences in the dates of image acquisition (Section 3.1), although weighting factors accounting for differential demand, topography, or other factors could easily be incorporated. Segment length is computed as two-dimensional Euclidean distances.

To compute the HI, then, let R_{MST} denote the average length of the segments in the MST, which refers to the mean interhousehold distance (MID, or average segment length) in the MST network. We define the homogeneity index by the ratio R_{MST}/R_{max}, and this index also ranges from 0, where the all the points are clustered at the same location, to 1, where all the points are in the regular repeated hexagonal shape mentioned above. We use this index as a measure of dispersion or nucleation.

3.4. Composite Prim's algorithm

The partial electrification problem, which we study in this paper, is to construct a network such that the mean interhousehold distance is minimized subject to the requirement that the network connects a pre-specified percentage of the households; we refer to this percentage as the penetration rate. We are interested in the total length of the network, since we assume that costs are linearly related to the total length as a first approximation. A further simplification is made in that the additional cost of transformers is assumed to simply be proportional to the low-voltage network length (as in rural settings the local, low-voltage line network is the dominant component of the infrastructure (ESMAP, 2007)). We, therefore, measure costs in units of length (meters), rather than in direct monetary units. When comparing different possible networks, we define the unit cost of a network as the mean segment length per connection for that network, or the mean interhousehold distance. An optimal network for connecting a given percentage p of households, is defined as MID(p%), meaning the network with the minimum MID for the penetration rate p (note that MID without any argument corresponds to MID(100%)).

While this problem is not easily solvable, the special case of the problem with the penetration rate of 100% (full penetration) is easy to solve. We review an optimal algorithm for the special case, and then adapt it for an arbitrary penetration rate. In graph theory, each point is called a "node", and each segment is referred to as an "edge". For clarity, we will refer here to the points (or nodes) as "structures" and we will call each length of wire connecting a pair of structures a "segment".

If the penetration rate is 100%, this special case of our problem is known in the literature as the minimum spanning tree problem mentioned above. Given a set of structures, where each structure represents a location requiring electricity, the objective is to find the segments (direct connections between two structures) such that all the structures are connected and the total length of the segments (or equivalently the average segment length) is as small as possible. This problem can be solved using an algorithm due to Prim (1957). The algorithm works by first choosing any structure as a starting point, and by then adding the shortest segment emanating from that structure to the network, or "tree". The process then repeats. At each iteration, the shortest segment emanating from the structures already in the tree is added. Segments that would create a cycle (connecting the existing network to itself) are avoided. When all the structures have been added to the tree, the process is completed.

We illustrate this algorithm on a set of structures shown in Fig. 3. In this four-structure example, suppose that the initial structure is D. Then the structure that is closest to D is C, and we add segment (C, D) to the network. At the next iteration, the structure that is closest to either C or D is B, and we add (B, C) to the network. Continuing with the algorithm, we obtain the spanning tree consisting of (C, D), (B, C), and (A, B), and the corresponding mean interhousehold distance is $(5+1+0.5)/3 = 2.167$. The sequence in which the segments are added to the set depends on the chosen initial structure; however, the final spanning tree does not depend on the initial structure under mild regularity assumptions (for example, that no two segments have identical lengths).

For a general penetration rate p% (that is less than 100%), there exists no computationally efficient algorithm to solve the problem of minimizing the network length subject to the penetration rate. For a fixed penetration rate, we modify Prim's algorithm as follows. For each initial structure, we run the Prim's algorithm until the required penetration rate is achieved. By running this algorithm repeatedly, using a different initial structure for each run, a series of different networks can be calculated.

Although the minimum spanning tree is identical (it is the optimal solution) regardless of the first structure chosen to start the network, the cost at a given percentage penetration is dependent on the starting point. This is intuitively clear when an example is considered: assume a given set of structures is sparsely distributed throughout a space with the exception of 10% of the structures clustered densely in one area. If a starting point for the network is picked in the middle of the dense cluster, the total cost to connect 10% of the population will be fairly low, as 10% of the population happens to be clustered densely around the starting point. If, however; a starting point were chosen far from this single dense cluster, the total cost to connect 10% of the

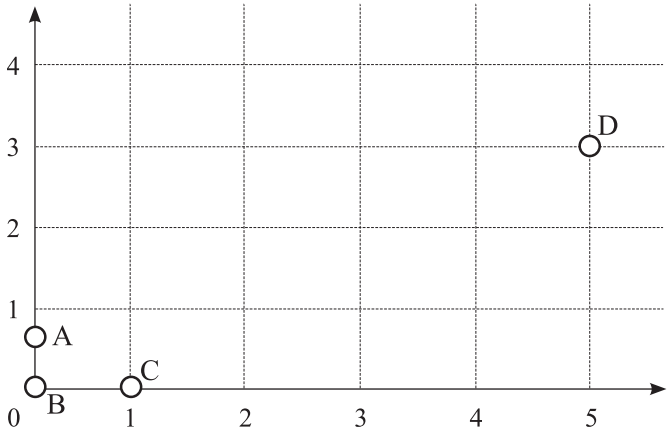

Fig. 3. A four-structure example of the Prim's algorithm.

structures would be relatively large, due to the large distance between the 90% of the structures not in the dense cluster.

From the set of networks calculated above, we take, for each penetration rate p%, the network with the minimum total length, and denote the average segment length of each of these networks by MID(p%). In the example given in Fig. 3, suppose we are interested in a penetration rate of 75%. If the initial structure is D, then we terminate Prim's algorithm after adding (C, D) and (B, C), resulting in a network with total length $5+1=6$. If the initial structure is C, then the algorithm terminates with (B, C) and (A, B) with total length $1+0.5=1.5$. If the initial structure is either A or B, we obtain the same network with the total length of 1.5. Thus, $\text{MID}(75\%)=1.5/2=0.75$. We refer to this algorithm as the composite Prim's algorithm.

One of the problems facing the planner is to decide the penetration rate in each region, and such a decision should strike a balance between the benefit of higher penetration and the cost of electrification. To this effect, we produce a curve of the cost of electrification MID(p%) with rising penetration rate p%. From this curve, we can also deduce the total interhousehold distance, as well as the marginal interhousehold distance. The marginal distance is calculated for each penetration rate as the cost of connecting the next structure to the grid, and presented here as a smoothed curve to aid in interpretation (smoothed over a window covering 10% of the segments at the site).

We make several remarks on the CPA. The MID(p%) represents the cost of the network when the network is built from scratch, and it is possible that a structure that is included in the network for a lower penetration rate becomes excluded for a higher penetration rate. Also, the total cost of the network at a given percentage penetration is dependent on the starting structure. A plot of cost versus penetration rate can be produced for each starting structure; as seen in Fig. 4, the cost curve can be radically different depending on the chosen starting structure, particularly for low penetration rates. By the definition of the algorithm, the network that the algorithm produces for any penetration rate is a truncated version of the regular Prim's algorithm, and thus it is a subset of the MST. Therefore, at full penetration, the cost computed by the algorithm is the same regardless of the starting structure. This occurs since the fully spanning network is the same in both cases; it is the MST.

This subset property has a useful implication to computation, since the algorithm can restrict its attention to consider only segments within this MST in computing the network with any penetration rate. Now, for possible choices of the initial structure, we can use all the structures if the number of structures is small (less than approximately 10,000 structures on a recent desktop computer); otherwise, we select enough initial structures such that increasing this number does not significantly affect the quality of solutions. We emphasize that the CPA is a heuristic approach to solve the partial electrification problem, and it is by no means an optimal algorithm even though it performs well in practice. We discuss another heuristic approach later (in Section 5.3).

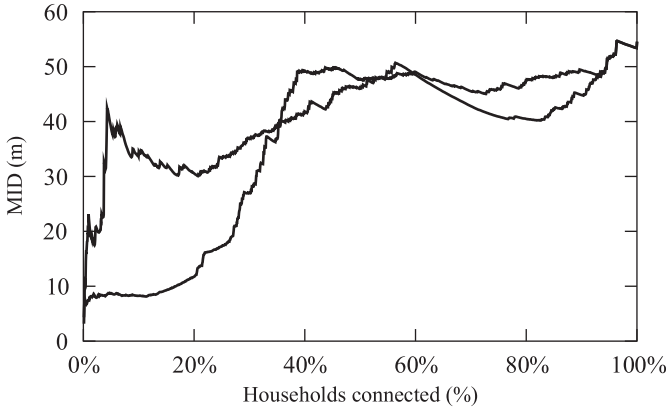

Fig. 4. The interhousehold distance versus penetration rate plotted for two of the best starting points in Potou, Senegal.

3.5. Distribution of network segment lengths

To compare the networks generated by the composite Prim's algorithm across sites, we consider each network as a collection of segments (individual lengths of wire) connecting individual structures. Different population distributions lead to different types of networks, with some networks dominated by shorter or longer segments. Networks in nucleated sites have a large amount of wire in short segments, due to the large number of closely situated structures. Networks for sites with a small number of dense clusters would be expected to also have a small number of very long cable runs in their network: those segments that make connections between clusters. A site with a random distribution of structures would be expected to show the most weight in the midrange of segment lengths, while having few very long or very short segments. The distribution of segment lengths for sites where the population distribution varies across the site would be expected to show some combination of these attributes. To make comparisons of networks at different sites, we construct a histogram for each site showing the total length of the network that is made up of a range of segments lengths. The height of each bin within the histogram is equal to the sum of the lengths of all segments that are within the range of lengths contained within the bin. For consistency of display, we choose a bin size of 25 m, with the final bin in each histogram containing all segments ranging from 425 m up to the longest segment in the network.

4. Results

The CPA was run (Fig. 5), and the homogeneity index calculated (Table 2), on four of the Millennium Villages sites. The four sites were selected so as to display a range of population distributions: from nucleated, as in Tiby, Mali, to homogeneously distributed, as in Ruhiira, Tanzania. For each site, every structure was used as a potential network starting point in the CPA. Mean and marginal network costs were calculated for each site using the CPA (Fig. 5), and the segment length distributions of the MST of each site were plotted (Fig. 6).

Below we discuss the results for each of the four sites, considering how each type of population structure determines the observed cost profiles. We then follow (in Section 5) with a more general evaluation of the algorithm, and discuss potential applications. We do not attempt to discuss in detail those political, economic, and geographic factors that have determined the population structure at these sites; our focus here is to understand how the existing patterns themselves might impact the cost of infrastructure investments.

4.1. Mbola, Tanzania

The area of Mbola (near the town of Tabora), Tanzania considered here shows a sparse pattern of population, with little clustering of households except for two areas in the southwest. These dense areas allow interconnections within a portion of the population (up to about 15%) with a relatively low MID (see Fig. 5a). Connecting the remaining population, however, is

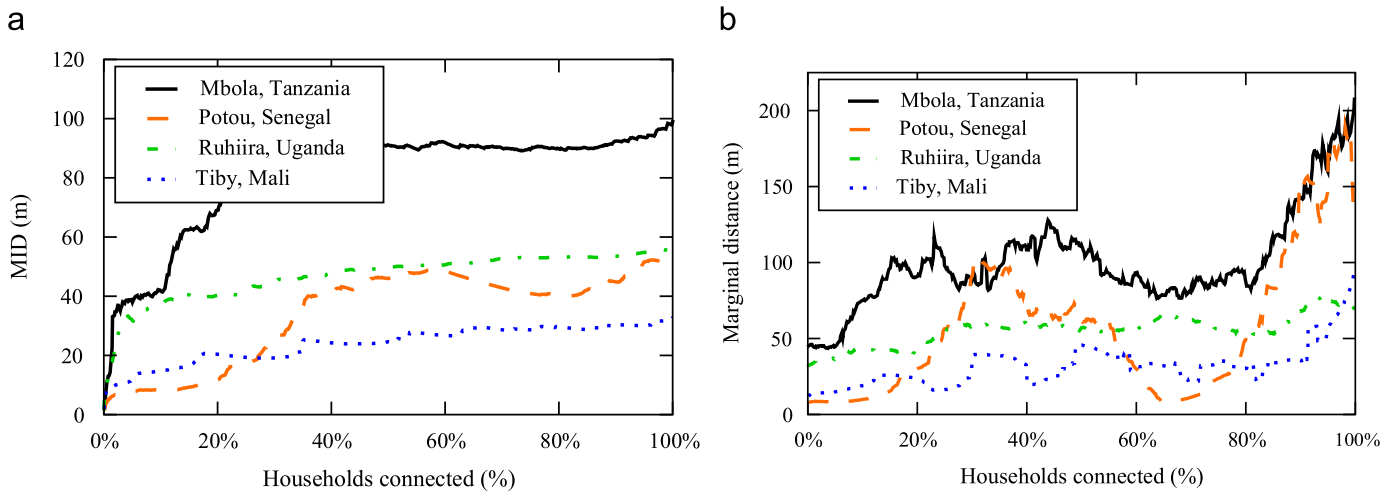

Fig. 5. Mean interhousehold distance (MID) (a) and marginal distance (b) for each of the four Millennium Villages sites.

Table 2
Comparison of statistics for the four sites.

Site	Maximum possible MID(100) (m)	Actual MID(100) (m)	Density (structures/km^2)	Homogeneity index
Mbola	313	99.7	11.7	.32
Potou	218	54.3	18.8	.22
Ruhiira	134	56.9	64.3	.42
Tiby	215	32.7	25.0	.15

The "Maximum Possible MID(100)" is the maximum possible mean interhousehold distance for a hexagonal distribution of points of the given density.

expensive. Considering the marginal distance curve, it is apparent that once these dense areas are connected (about 10% of the population), the next 70% of the population (from 10% to 80% penetration) can be connected with roughly constant marginal cost. The marginal distance does declines somewhat between 45% and 80% penetration, as the relatively dense settlements along roads in the south-east and eastern part of the image are connected to the network. However, connecting the last 20% of the population is relatively expensive (best observed in the marginal cost curves of Fig. 5b) due to the sparse arrangement of the remaining structures. This can be seen in the significant rise in the marginal distance curve for Mbola after about 80% penetration. The distribution of segments in the MST (Fig. 6a) is characteristic of a site that is sparsely and homogeneously populated with some dominant clusters. The connections between these clusters contribute to the 30 km of network length made up of segments between 450 and 1505 m.

4.2. Potou, Senegal

Potou (with the largest town nearby being Louga), Senegal is sparsely populated, with two nucleated areas; one in the south and another in the northwest along the coast. The remaining coastal area is uniformly, though sparsely, populated. Small outlying clusters line several rural roads running north-south through the area. Network costs in Potou are dominated, however, by the areas outlying the two main population centers. The MID rises slowly as the largest population center (in the south of the image) is connected to the grid (Fig. 5a). After 20% penetration, the MID begins to rise more quickly as the outskirts of the main cluster are connected to the grid. After 60% penetration, the network begins to reach nucleated areas along the coast; this is reflected in the drop in the marginal distance plot. There is a steep rise in marginal distance as the (sparsely distributed) last 20% of the structures join the network. The Potou MST segment distribution in Fig. 6b has two peaks, one for short segments and one for long segments, characteristic of a nucleated settlement pattern surrounded by sparsely populated outlying areas; however, the peaks are not as distinct at the short end as in Fig. 6d for Tiby where the nucleation is strong.

4.3. Ruhiira, Uganda

Ruhiira (with the largest city nearby being Mbarara), Uganda is far more densely populated than the other sites considered here. The road network and extreme topography in the area gives some structure to the population distribution; however, Ruhiira clearly has the least clustering of the four distributions shown here. The lack of clustering leads to a situation where after 5% penetration, the MID is essentially the same, as also seen in the flat marginal distance curve. In considering Ruhiira in comparison to the other sites, the differences in scale must be remembered; Ruhiira has 6436 structures compared to only 2496 in Tiby, the next largest site. There is no significant jump in the marginal distance even as the last 20% of the households are connected (Fig. 5b). The near uniform population distribution in Ruhiira is evident in the distribution of MST segment lengths (Fig. 6c) with nearly all of the network length composed of segments between 25 and 150 m. The majority of the segments are concentrated around the mean, with very few segments of the MST longer than 100 m. The result of this pattern is nearly flat curves for both mean and marginal cost, with the mean cost quickly reaching a stable value of around 40 m, and then slowly and monotonically rising to 60 m (Fig. 5a).

4.4. Tiby, Mali

The population of Tiby (near the city of Segou), Mali is split into several nucleated clusters, with few outliers. The ease of connecting these dense, nucleated clusters allows connection of the entire population with an MID of only 32.7 m. However, the large separation between population centers in Tiby leads to "jumps" in the marginal distance (as seen at 30%, 50%, and 75% penetration in Fig. 5b) as connections are made between clusters. Fig. 6d shows that the nearly half of the total network length consists of very short segments, with the other half made up of a few long segments, up to 2500 m in length, connecting the

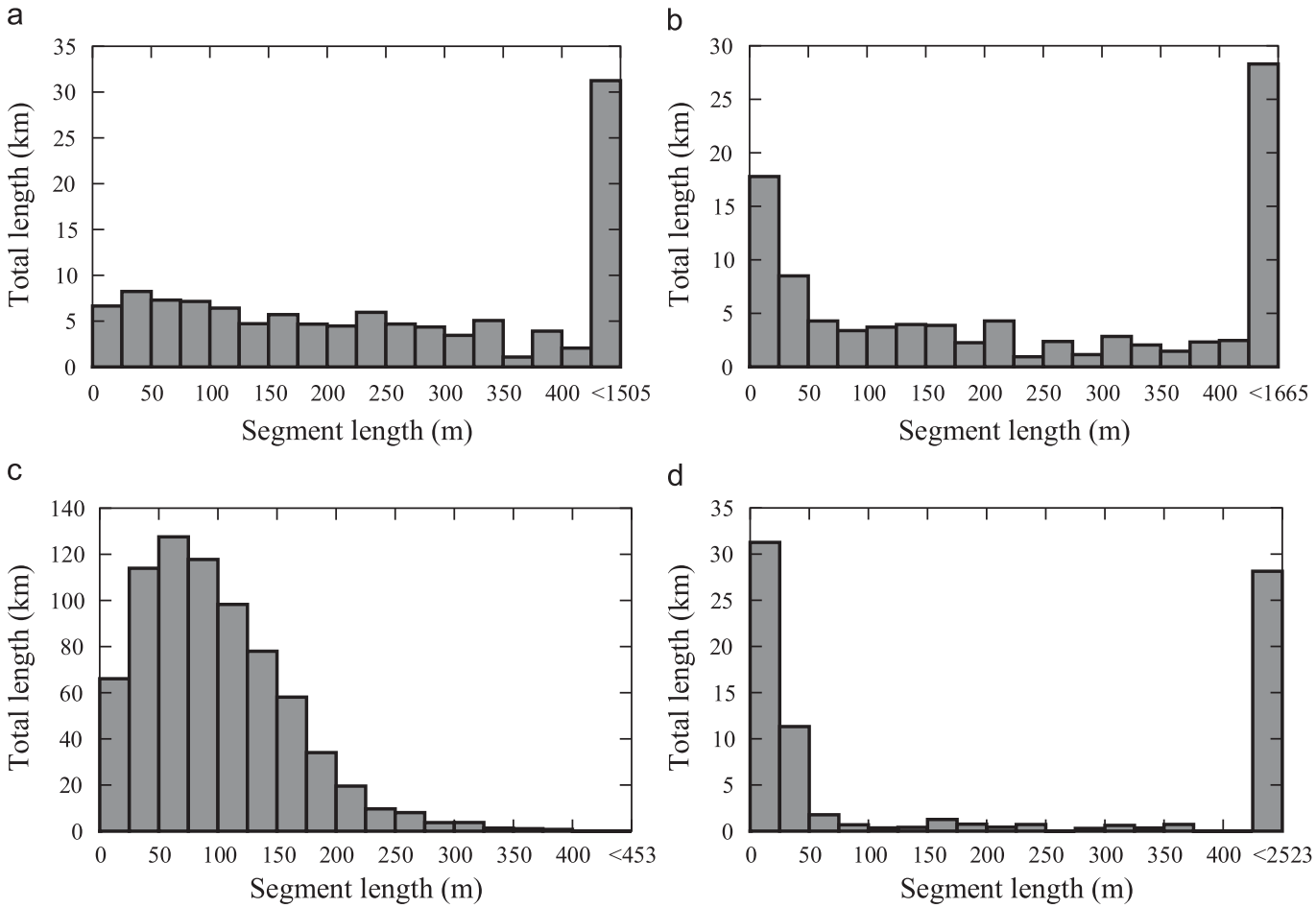

Fig. 6. Comparison of the dominant segment lengths at the four sites: (a) Mbola, (b) Potou, (c) Ruhiira, and (d) Tiby. The height of each bar is the sum of the lengths of all segments contained within that bin. The final bin includes all segments longer than 425 m, with the longest segment in the network indicated to the far right of the *x*-axis, next to the "<"sign.

nucleated settlements. For an electricity grid, such a distribution would imply that a fewer number of transformers would be needed for a given population (compared to a distribution such as that shown in Fig. 6c for Ruhiira); as each transformer can serve multiple demand points with short low-voltage segments.

5. Discussion

We first highlight the impact of several characteristic settlement patterns on infrastructure costs by considering a simplified (artificial) dataset (Section 5.1). Following, we discuss the general performance of the composite Prim's algorithm and homogeneity index, and issues of importance to planners considering the usage of these algorithms (Section 5.2). To gauge the performance of the CPA, we then compare it to an existing algorithm (Section 5.3) and present an alternative measure of dispersion, the micro density variation index (MDVI) in Section 5.4 to corroborate the results of the homogeneity index discussed previously.

5.1. Characteristic settlement patterns

Our observations from the Millennium Villages sites suggest several characteristic settlement types that can lead to particular cost curves. However, the complexity and noise in the Millennium Villages data complicates understanding the relationship between spatial pattern, and network cost. To isolate the impacts of settlement pattern on infrastructure costs, we consider here a set of artificial patterns. The hypothetical patterns in Fig. 7(a–f) all have the same total number of structures (400), and identical spatial extents, varying only in the spatial arrangement of the structures.

Pattern a is a sparse but uniform grid as might approximate a highly organized area. Pattern b shows a sparse but random distribution of points, as might be found in a flat rural area with few roads organizing the population. Patterns c and d depict a nucleated population with several clusters, with the clusters in d having a higher density than in c. Similarly, patterns e and f compare two highly nucleated settlements, in this case each with a single nucleus, of varying density.

Comparing the MIDs (Fig. 8) for patterns a and b as penetration rate varies, we see that the degree of "spatial randomness", or departure from a uniformly gridded layout, decreases the MID at all penetration rates. This makes intuitive sense because as the population gets closer to a gridded layout (pattern a), the MID(100%) approaches R_{max} (as discussed in Section 3.3). Patterns c and d isolate the impact of multiple nucleation centers on MID: the long connections between dense population nuclei lead to "jumps" in the MID (as is also seen in Tiby). The increased density of nucleation in pattern d (vs. c) amplifies this effect. Patterns e and f show the impact of nucleation at a single site, with varying density. The MID curve for these sites fairly quickly reaches a stable value, with the denser site (f) having a lower MID(100%).

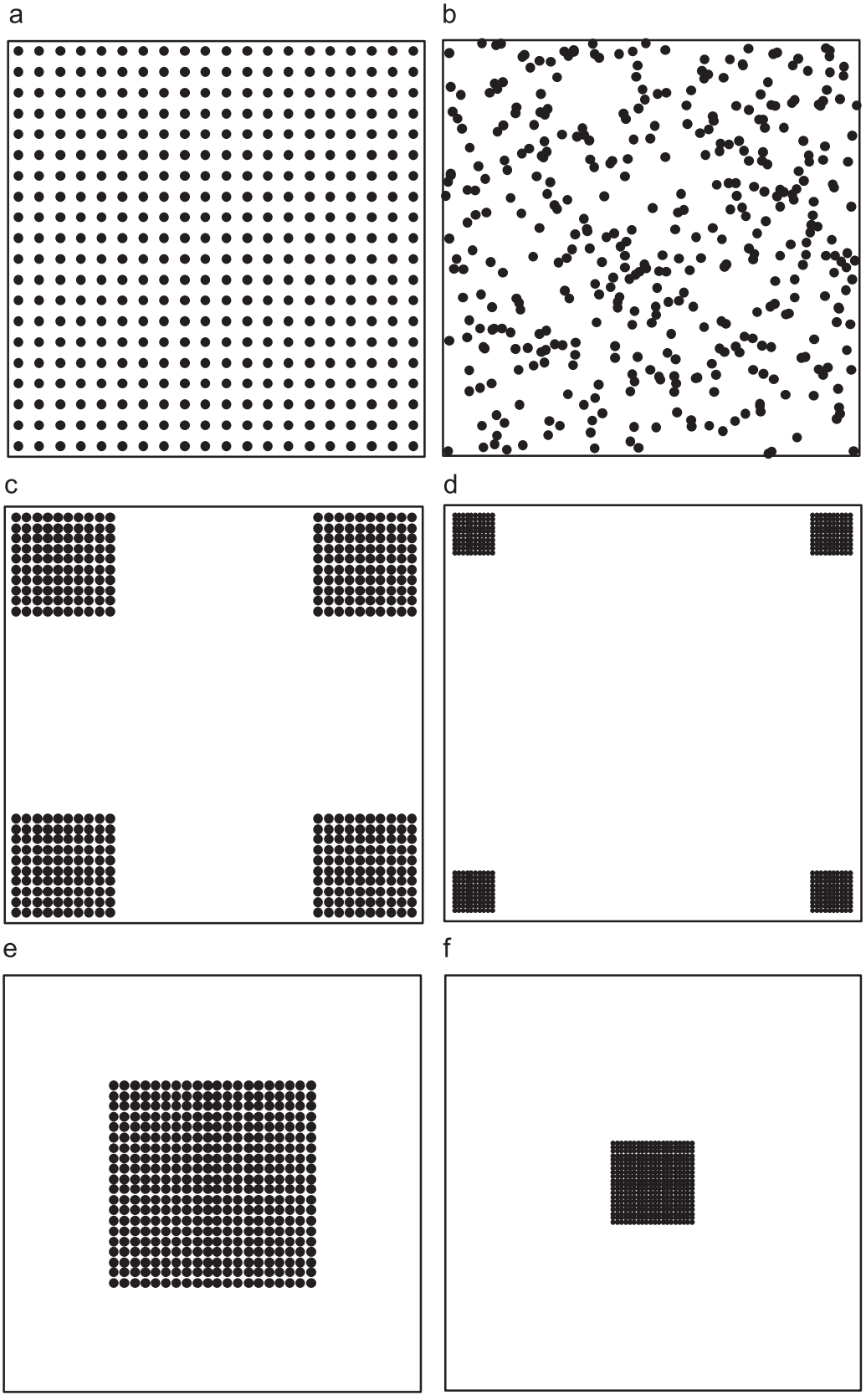

Fig. 7. A set of hypothetical structure distributions. Each dot represents a single structure. Each image has 400 structures in total.

5.2. General discussion

The most useful on-the-ground information that a network planner would need to carry out preliminary assessments is the mean interhousehold distance. The results for four different rural areas in East and West Africa show HI values ranging from .15 to .42. Settlement patterns determine these differences, with the larger HI of .42 coming from a near homogeneously distributed settlement and the lower figure of 15 from a more nucleated settlement.

As planners consider the possibilities for rural electrification, the next question is the relationship between network expansion cost and penetration rate. The usual assumption is that marginal costs will increase dramatically as the last few households are connected to the network. Indeed, this is generally observed in our datasets, with the effect being minimal when the settlement pattern is nearly homogeneous (as in Ruhiira). However, even though this effect is seen for the majority of the sites, differences in HI and in the variation of cost with penetration rate between sites are striking. Finally, there is a concern that optimization of local networks by excluding high-cost households from the grid will greatly increase the later costs of grid expansion; we provide a method of addressing this issue with the CPA.

The sites we have examined, though sparsely populated, tend towards distributions of two types: non-nucleated or uniformly populated (as in Ruhiira and Mbola) or with nucleated population centers arranged into several dense clusters with a small number of outliers (as in Potou and Tiby). In areas such as Ruhiira, Uganda, the lack of significant structure in the spatial layout of the population leads to the MID quickly approaching the MID(100%)—penetration rate essentially becomes irrelevant except from a total cost perspective. As seen in the distribution of segments, nearly all of the network length consists of segments that are 25–150 m, but with a peak for segment lengths that are between 50 and 75 m. An electricity grid in Ruhiira would need more transformers per-unit population, due to the lack of nucleation, as each transformer would serve a relatively small number of households. The high HI in Ruhiira, .42 is consistent with these observations.

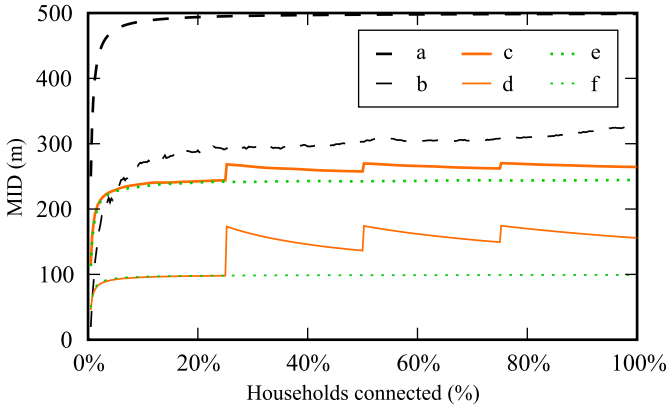

Fig. 8. Mean interhousehold distance (MID) for each of the six hypothetical settlement patterns depicted in Fig. 7.

Areas with significant clustering, as in Tiby, Mali, with an HI of .15, are a trivial case at the other end of the spectrum. Small, nucleated clusters lend themselves to be better served by local "mini-grids" or through networks where one transformer can serve a number of households. This type of distribution also shows little structure in the marginal cost profile, except as relatively long segments are added to the network to connect clusters. The distribution of MST segment lengths in a highly nucleated site like Tiby (when compared to another site) shows more weight in the shorter range, due to the short distances between clustered households, little weight in the midrange (100–450 m) and a relatively large fraction of length in longer distance cable runs connecting the individual clusters.

The MST segment distribution of Mbola, Tanzania (HI of .32) is closer to that of Ruhiira, but with long segments that separate the little clustering that exists there, and with more weight in the midrange. On the other hand Potou, Senegal (HI of .22) is more similar to Tiby, showing more nucleation than Mbola, though without as tight a clustering as in Tiby. These observations are consistent with the differences in MID at all penetration rates among Mbola and Potou, although their overall densities are similar: 11.7 and 18.8 structures/km^2, respectively.

It is important to note that the sites discussed above vary greatly in terms of the total number of structures per site. As seen in Table 3, although Potou and Ruhiira have similar MIDs for 50% penetration, 46.1 and 49.8 m, respectively, the total cost of connecting 50% of the population in Ruhiira is about five times as expensive as connecting 50% of the population in Potou, due to the larger total number of structures in Ruhiira. Depending on the objectives and economics of a project, total cost, rather than MID, could be the most important metric to consider. Total cost is also likely to be important for planning when network expansion is considered.

As the CPA is based on Prim's algorithm, it is important to recognize that it necessarily builds subgraphs of the MST. Therefore, any network produced by the CPA can be expanded to 100% penetration with no penalty. In other words, the cost of initially building a network with 100% coverage is equal to the cost of building a network (calculated with the CPA) with less than 100% coverage, plus the cost of expanding it to full coverage. The same is not necessarily true for expansion of a minimum cost network spanning a subset of the population to another network also providing less than full coverage—the optimal network for 20% penetration may not be a subgraph of the optimal network for 40% penetration, as the two may start from different root structures. For example, a planner interested in constructing a network spanning only 20% of the grid and who is planning for later expansion up to 40% connectivity could minimize their expansion costs by estimating with the CPA the optimal network for 40% connectivity, and building initially building a network that is a subgraph of this network. A planner interested in minimizing

Table 3
Comparison of MID (in meters) and total cost (in parentheses) for the four sites at several different penetration rates.

Site	10% Penetration	25% Penetration	50% Penetration	75% Penetration	100% Penetration	Total number of structures
Mbola	41.4 (5013)	74.6 (22,155)	90.9 (53,703)	90.2 (79,770)	99.7 (116,225)	1174
Potou	8.3 (1505)	17.4 (7852)	46.1 (41,604)	40.8 (55,229)	54.3 (95,925)	1796
Ruhiira	37.7 (25,068)	42.1 (69,787)	49.8 (164,268)	52.8 (260,672)	56.9 (364,859)	6560
Tiby	14.8 (3743)	19.3 (12,162)	24.6 (30,858)	28.6 (53,639)	32.7 (81,360)	2494

Differences in structure density account for the variation in total cost.

initial costs could instead estimate the best network for 20% connectivity, and choose to construct this network initially, while calculating the penalty of doing so by comparing it to the best network for 40% connectivity. Though there is no analytical guarantee for the CPA, it appears to handle these cases reasonably well, based on the observed empirical results.

The CPA performs well enough to be used on reasonably sized datasets; we have experimented with calculations on a site having in excess of 17,000 structures. However, the computational complexity of the algorithm necessitates limiting the potential starting points considered by the algorithm when the number of points in the dataset becomes large. Another possibility for preliminary assessments is to use a small sample of the dataset as an approximation to the whole population while calculating the MID. This technique can only be used with a dataset where the population is nearly Poisson distributed (or homogenous). Ruhiira is an example of one site nearing spatial homogeneity. Sites with a large degree of nucleation would be difficult to segment in this fashion.

Although a planner might give preference to the electrification of certain sites, such as health centers, markets, and educational areas, over individual households, the algorithm discussed above weights all buildings evenly. Other important factors in considering the cost-effectiveness of electrifying a site could include proximity to existing infrastructure, and the topography of the area. These factors could easily be added to the analysis by differentially weighting points according to their relative demand or perceived importance (health centers weighted more heavily than households, for example)—the algorithm itself does not need to be modified to take this additional information into account. We have chosen to consider all structures equally here as acquisition of these weighting factors for the sites we mention would be difficult, due to their derivation from remotely sensed imagery. A more detailed dataset; however, would allow prioritization of structures for electrification based on their perceived importance.

Extension of the composite Prim's approach using weighting functions to take account of existing infrastructure and population would allow more accurate cost estimates (given the availability of data), as well as consideration of the economic development potential of a region, and the expected non-linear relationship between economic development and population density. Different structures could be weighted differentially to both reflect a planners preferences for electrification of certain areas, as well as to incorporate the potential for economic development. Prior research suggests that areas with higher population density tend to develop faster than areas that are sparsely populated, due to the increased size of markets, as well as returns to scale in production (World Bank, 2008). Planners could take advantage of this tendency, with the CPA, by incorporating a non-linear factor weighting areas with higher population densities. Areas with higher population density, and therefore higher economic development potential, would be weighted more heavily than sparse areas.

Not mentioned here are the political issues potentially associated with grid planning. Political, as well as economic, considerations would likely play a large role in determining, for example, the network starting point, or a potential weighting scheme. Haanyika (2008) relate the problem in Kenya of a poorly regulated energy market leading to utilities "cherry-picking" the most attractive consumers to supply—our framework for choosing a network would clearly be difficult to implement in this case. If broader-scale electrification is a policy objective, these considerations must be taken into account, and incentive structures and pricing planned accordingly.

We also do not consider here an example of a grid expansion in an area with a pre-existing grid. The areas we have considered in rural Africa generally lack the pre-existing infrastructure to make this problem a concern; therefore, our dominant focus here is on the problem of grid construction with no pre-existing local electricity infrastructure, and on understanding, in these cases, the impacts of population structure on infrastructure investment costs. However, extension of a pre-existing grid is one area where future work could be carried out. The CPA algorithm should apply fairly simply to this special case.

Future work is also required to better relate household and structure densities. Here, we have focused on connecting "structures" to the grid: individual buildings visible from remote-sensing imagery. However, it is important not to conflate structures and households. Although households are the common unit of analysis for other areas of assessment, such as the provision of health services, we have avoided the conversion of data expressed on a per-structure basis to representation based on households due to the limited data available on household density versus structure density in the areas we consider here. Our preliminary work has noted significant differences between sites in the number of structures per household. Improving the quality of this information is important for policymakers to be able to better gauge the impact of infrastructure investments. Were this data available, and constant across a site, a simple scaling factor could be used to express mean, marginal, and total costs on a per-household basis (or on a per-person basis, if total population is also accurately known). Future work could also consider possibilities for merging structure-level point data into clusters, so as to approximate household locations using structure-level data based on remotely sensed information.

5.3. Composite Prim's algorithm vs. an existing algorithm

In this paper, we have proposed the composite Prim's algorithm for the partial electrification algorithm. In the computer science literature, an abstraction of this problem is known as the *k*-MST problem, and it has been well-studied. Chudak et al. (2001) have taken a technically rigorous approach to this problem, and proposed a heuristic algorithm that produces a network whose total length is guaranteed to be at most twice the total length of the optimum network. Their approach is based on the fact that the Lagrangian version of the *k*-MST problem is a well-studied Prize Collecting Steiner Tree problem, for which a similar guarantee has been developed using the so-called primal-dual algorithm (Goemans and Williamson, 1995). The primal-dual algorithm searches for the best possible solution by keeping track of both the feasible solution and its shadow prices related to the constraints. While the algorithm of Chudak et al. (2001) has a theoretically appealing property, its implementation is much more complex than that of the CPA. We have implemented both algorithms, and find that both have similar performances. As seen in Fig. 9, the mean cost curves obtained from these two algorithms match closely.

5.4. Homogeneity index and micro density variation index

In this paper, we have adopted the homogeneity index, defined in Section 3.3, as the measure for the dispersion of structures within each region. To assess the validity of this measure, we propose another indicator called the micro density variation index, which is shown to be consistent with the HI. MDVI is computed as follows: (i) start with a square tile that is relatively small compared to the area of the region under study, (ii) place these tiles throughout the region of interest such that the number of squares is sufficiently large, and (iii) count the number of structures in each square. Then, we compute the sample coefficient of variation on the number of structures, and denote it by the MDVI. If the structures are uniformly distributed in the

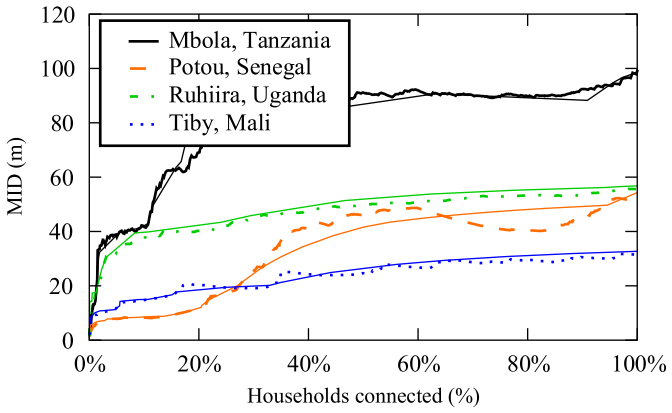

Fig. 9. Comparison of results of the composite Prim's algorithm (thick lines) with those obtained from the *k*-MST algorithm proposed by Chudak et al. (2001) (thin lines).

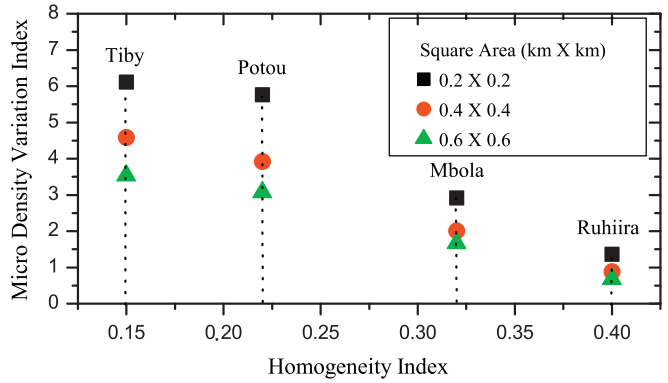

Fig. 10. Comparison of the homogeneity index (HI) with the micro density variation index (MDVI).

region, then each tile would contain exactly the same number of structures, resulting in an MDVI of zero (or very close to zero); otherwise, if the structures are highly clustered, the corresponding MDVI would be high. For the placement of tiles in the region, we use a systematic grid approach where the region is subdivided using a grid with spacing that is one-quarter of the edge length of a tile, and consider all possible tiles that are aligned with the grid and completely contained in the regions. (Due to the choice of the grid spacing, we permit the overlapping of squares.) For example, for the 10 km × 10 km region, if we use tiles of size 0.4 km × 0.4 km, then it can be shown that there are 9409 distinct tiles, from which the numbers of structures can be counted. We recommend that the size of the tiles be such that a reasonable number of tiles can cover the entire region, and a sufficiently large number of tiles can be placed without overlapping too many times.

In Fig. 10, we have computed the MDVI values using tiles of sizes 0.2 km × 0.2 km, 0.4 km × 0.4 km, and 0.6 km × 0.6 km, and plotted them against the HI. We observe that the MDVI is negatively related to HI, a property that is expected from the definition of MDVI. This observation suggests that both indices are good measures of the dispersion of structures within each region. Note that the evaluation of MDVI is computationally easier as it does not require one to compute the minimum spanning tree; hence it is plausible that the MDVI can be used to obtain an approximation of the value of the HI using a regression approach. This allows one to have a simple technique to rapidly assess the length of wire that is needed to connect all households in a region. One would first compute the MDVI for the region and then consider Fig. 10 to determine an approximation for HI. With this estimate of HI and the maximum possible MID(100%), one can infer an estimate of the MID(100%) value as without computing the minimum spanning tree.

6. Conclusions

The four sites described here are typical of the settlement patterns encountered by infrastructure planners in rural Sub-Saharan Africa. The results indicate the inadequacy of existing proxies such as population density, and of the importance of considering the cost of electrification at varying penetration rates: the presumption that cost per connection will monotonically increase with penetration rate is incorrect. Future analyses at the national and regional scale can utilize this knowledge by optimizing local penetration rates so as to ensure cost-effectiveness of the entire grid; thereby ensuring full penetration in those areas where it is optimal, while reducing unnecessary costs in areas where it is not. Although the metric we offer, the homogeneity index, cannot yet fully capture these subtleties, we have begun to develop a means to allow better comparison among sites. We also find that some population settlement patterns offer the potential for savings (on a per-unit basis) in upfront investments through an initial roll-out that covers part of the population; later expansion of these partially spanning networks can be undertaken at little additional cost in the long-term.

The model described herein need not be limited solely to the analysis of electricity infrastructure. Other problems in rural infrastructure design, such as water and communication networks could benefit from a similar modeling approach. Sitting of new health care and educational facilities could also benefit from an approach aimed at maximizing penetration and cost-effectiveness.

References

Bekker, B., Eberhard, A., Gaunt, T., Marquard, A., 2008. South Africa's rapid electrification programme: policy, institutional, planning, financing and technical innovations. Energy Policy 36 (8), 3115–3127 Aug.

Boko, M., Niang, I., Nyong, A., Vogel, C., Githeko, A., Medany, M., Osman-Elasha, B., Tabo, R., Yanda, P., 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Ch. Africa, pp. 433–467.

Chudak, F.A., Roughgarden, T., Williamson, D.P., 2001. Approximate *k*-MSTs and *k*-Steiner trees via the primal-dual method and Lagrangean relaxation. In: In Proceedings of the 8th Conference on Integer Programming and Combinatorial Optimization. Springer, Berlin, pp. 60–70.

Clark, P.J., Evans, F.C., 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35 (4), 445–453 Oct.

ESMAP, 2007. Technical and economic assessment of off-grid, mini-grid and grid electrification technologies. Tech. Rep. 121/07, Energy Sector Management Assistance Program.

Goemans, M.X., Williamson, D.P., 1995. A general approximation technique for constrained forest problems. SIAM J. Comput. 24 (2), 296–317.

Haanyika, C.M., 2006. Rural electrification policy and institutional linkages. Energy Policy 34, 2977–2993.

Haanyika, C.M., 2008. Rural electrification in Zambia: a policy and institutional analysis. Energy Policy 36 (3), 1044–1058.

Kaijuka, E., 2007. GIS and rural electricity planning in Uganda. Journal of Cleaner Production 15 (2), 203–217.

Modi, V., McDade, S., Lallement, D., Saghir, J., 2006. Energy services for the Millennium Development Goals. Tech. rep., Energy Sector Management Assistance Programme, United Nations Development Programme, UN Millennium Project, and World Bank.

Parshall, L., Pillai, D., Mohan, S., Sanoh, A., Modi, V., 2009. National electricity planning in settings with low pre-existing grid coverage: development of a spatial model and case study of Kenya. Energy Policy 37 (6), 2395–2410.

Peacock, T., Ward, C., Gambarelli, G., 2008. Investment in Agricultural Water for Poverty Reduction and Economic Growth in Sub-Saharan Africa. World Bank Tech. Rep.

Prim, R.C., 1957. Shortest connection networks and some generalizations. Bell System Technical Journal 36, 1389–1401.

Rayburg, S., Thoms, M., Neave, M., 2009. A comparison of digital elevation models generated from different data sources. Geomorphology 106 (3–4), 261–270.

Rufin, C., Rangan, U.S., Kumar, R., 2003. The changing role of the state in the electricity industry in Brazil, China and India. American Journal of Economics and Sociology 62 (4), 649–675.

Saghir, J., 2004. Energy and Poverty, Paper Prepared for the International Energy Forum. The World Bank.

Sanchez, P., Palm, C., Sachs, J., Denning, G., Flor, R., Harawa, R., Jama, B., Kiflemariam, T., Konecky, B., Kozar, R., Lelerai, E., Malik, A., Modi, V., Mutuo, P., Niang, A., Okoth, H., Place, F., Sachs, S.E., Said, A., Siriri, D., Teklehaimanot, A., Wang, K., Wangila, J., Zamba, C., 2007. Poverty and hunger special feature: the African millennium villages. Proceedings of the National Academy of Sciences 104 (43), 16775–16780.

Victor, D., Heller, T., 2006. The Political Economy of Power Sector Reform: The Experiences of Five Major Developing Countries. Cambridge University Press, Cambridge Tech. Rep.

World Bank, 2002. Private Participation in Infrastructure: Trends in Developing Countries in 1990–2001. World Bank, Washington DC Tech. rep.

World Bank, 2008. World Development Report 2009: Reshaping Economic Geography. World Bank, Washington DC Tech. rep.

Yohe, G., Lasco, R., Ahmad, Q., Arnell, N., Cohen, S., Hope, C., Janetos, A., Perez, R., 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Ch. Perspectives on climate change and sustainability, pp. 811–841.