Operationalizing Central Place Theory and Central Flow Theory with mobile phone data

Derek Doran, Andrew Fox

Central Place and Central Flow Theory are geographic principles explaining why and how cities develop across large regional spaces. Central Place Theory postulates that cities self-organize into a spatial hierarchy were small numbers of very large ‘Central Places’ support numerous surrounding and less developed ‘Low Places’, while ‘Middle Places’ develop at the periphery of where Central Places carry spatial influence. Central Flow Theory is a comple- mentary notion that explains the cooperative development of cities through joint information sharing. Both theories are often discussed, with multiple regional development and economic models built upon their tenents. However, it is very difficult to quantify the degree to which Central Place and Central Flow Theory explains the development and positions of cities in a region, particularly in developing countries where socioeconomic data is difficult to collect. To facilitate these measurements, this paper presents a way to operationalize Central Place and Central Flow Theory using mobile phone data collected across a region. It defines a set of mobile phone data attributes that are related to basic facets of the two theories, and demonstrates how their measurements speak to the degree to which the theories hold in the region the mobile phone data covers. The theory is then applied in a case study where promising locations for economic investment in a developing nation are identified.

Operationalizing Central Place and Central Flow Theory With Mobile Phone Data

Derek Doran
Dept. of Computer Science & Engineering
Kno.e.sis Research Center
Wright State University
derek.doran@wright.edu

Andrew Fox
Dept. of Industrial Engineering and
Management Science
Northwestern University
andrewfox2014@u.northwestern.edu

Abstract

Central Place and Central Flow Theory are geographic principles explaining why and how cities develop across large regional spaces. Central Place Theory postulates that cities self-organize into a spatial hierarchy were small numbers of very large 'Central Places' support numerous surrounding and less developed 'Low Places', while 'Middle Places' develop at the periphery of where Central Places carry spatial influence. Central Flow Theory is a complementary notion that explains the cooperative development of cities through joint information sharing. Both theories are often discussed, with multiple regional development and economic models built upon their tenents. However, it is very difficult to quantify the degree to which Central Place and Central Flow Theory explains the development and positions of cities in a region, particularly in developing countries where socioeconomic data is difficult to collect. To facilitate these measurements, this paper presents a way to operationalize Central Place and Central Flow Theory using mobile phone data collected across a region. It defines a set of mobile phone data attributes that are related to basic facets of the two theories, and demonstrates how their measurements speak to the degree to which the theories hold in the region the mobile phone data covers. The theory is then applied in a case study where promising locations for economic investment in a developing nation are identified.

1 Introduction and Motivation

The mechanisms behind where, why, and how cities spatially organize themselves across a large geographic area is inherently complex [2], yet an ability to summarize and model such mechanisms is important for theoretical and practical purposes. They may be theoretically related to the *importance* of a city, the kinds of services they provide, and the features that attract different types of people to settle within them [1, 52, 27]. For example, very large cities with major hospitals may have a tendency to attract medical doctors if the cities surrounding it have no hospital

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1342v1 | CC-BY 4.0 Open Access | rec: 4 Sep 2015, publ: 4 Sep 2015

or if the next city with a major hospital is very far away. It may also be used to explain why the surrounding areas of a large city are dotted with smaller towns, suburbs, or rural areas with little economic activity [15, 18]. Theoretical models may also be a useful tool to anticipate the growth of a city, region, and the spatial distribution of a population over time [41]. Practically, these models may be used by urban panning researchers to identify and control access to open space within a city [51], to forecast the production of goods and services [42], and to drive decisions about whether urban areas should be expanded [50].

Central Place Theory (CPT) is a long-standing theory that summarizes the ways city's develop and position themselves over large regions [9, 4]. It imagines a theoretical geographic area without physical barriers, uniform soil fertility, population, transportation network and access, and a free market where any good can be sold to any other city. The theory hypothesizes that these conditions lead to the mergence of 'Central Places' that carry a very high population and produce a disproportionately large number of goods. Other types of communities, namely small 'Low Places' that are represented by 'villages' and 'towns' naturally develop at different distances from Central Places depending on their reliance to its goods, people, and economy. 'Middle Places' would also develop as cities that are self-sustaining, but are less developed than central places. The more recent, complementary Central Flow Theory (CFT) postulates that cities develop in a cooperative manner by sharing information and interests using modern technology unconstrained by distance [55].

Whether CPT and CFT are valid theories that explain the development of cities across a geographic landscape is an important question. A positive answer would suggest that CPT and CFT-based models of city population growth [48], hierarchical organization [25], economic growth [27], and tourism [11] could be applied to manipulate and control the development of urban cities. It could also lead to novel types of applied analysis, such as identifying regions in a country that is apt for economic investment or has high growth potential [12]. One way to perform this assessment is to codify CPT and CFT concepts in a mathematical or statistical model, and then assess how well the model fits or explains patterns within spatial and socioeconomic data across a region. However, the numerous design decisions necessary to create such models would make them unrealistic at worst, and controversial at best. For example, models that do not adequately consider the relationship between population and economic growth and rates of communication among cities are may not reflect the basic notions of CPT and CFT. Moreover, models that could appropriately codify CPT and CFT concepts may still be questioned by how they also consider the many exogenous factors that affect how well real city development deviates from CPT and CFT. Such factors may include climate, topography, and the position of significant cultural landmarks and historical events in a region [45, 56].

Given the inherent challenge in building integrated models of CPT and CFT, an alternative approach is to search for a collection of patterns in data related to the development of cities in a geographic region. The alignment of these patterns to CPT and CFT concepts, and the interactions of the types of patterns, can help an analyst decide whether to reject the hypothesis that CPT and CFT explains urban development in the region. If the patterns do not reject the hypothesis, they may be thought of as a quantitative *operationalization* of CPT and CFT that unlocks the use of existing CPT- and CFT-based models in applied settings, and a quantitative analysis of regional development based on the two theories.

This paper illustrates how one may assess if CPT or CFT explains the development of a geographic region, and

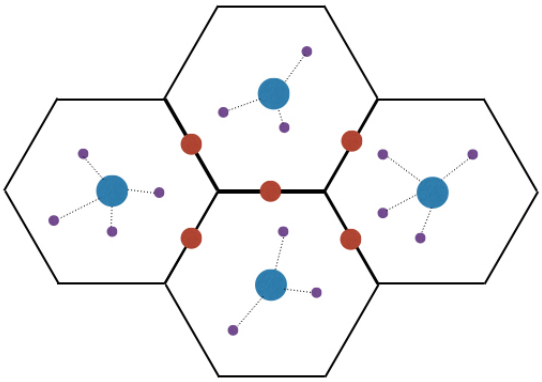


Figure 1: Idealistic spatial hierarchy of Central Places (blue), Low Places (Purple), and Middle Places (red). Hexagons correspond to the region Central Places influence by providing low- and high-level outputs. Low places rely on the Central Place to thrive. Middle places are necessarily self-sustaining due to their distance between Central Places.

how the concepts may be operationalized, using mobile phone data. Mobile phone data is a particularly useful source of data for this purpose because: (i) the mobile phone is a ubiquitous communication device that is used across countless geographic regions [13]; (ii) mobile phone data can be collected by service providers at large scale, even in developing nations where sources of urban, socioeconomic, and geographic data is difficult to collect [5]; and (iii) mobile phone data has demonstrated its use as a way to approximate transportation flows [7] and marketplace actions [26], which are basic CPT concepts. Twelve months of mobile phone calling data cross arrondissements in Senegal are used in our analysis. After establishing that CPT and CFT reasonably explains the spread of urban developments in Senegal, we offer a use case of its operationalization in a study that searches for promising regions of economic development.

The layout of this paper is as follows: Section 2 gives a finer overview of CPT and CFT and discusses the related work on building theories and applications on them. Section 3 presents the mobile phone dataset and analyze patterns supporting the hypothesis that CPT and CFT are applicable. Section 4 discusses a use-case for operationalizing CPT and CFT where promising places for economic investment are identified. Conclusions and future work are presented in Section 5.

2 Central Place and Central Flow Theory: An Overview

CPT is a leading theory for explaining the tendency of villages, towns, and cities to self-organize along a cascading hierarchy [9]. It proposes a spatial organization of cities across a region as illustrated in Figure 1 where small villages, towns, and secondary centers lie in regions where a large urban center carries local influence. The hierarchy is rooted at a *Central Place* - large population zones that are able to supply *low-level outputs* such as physical goods and services and *high-level outputs* such as knowledge and culture to the surrounding area it influences. Central places are thus expected to be positioned far away from each other, so that a local region is not exposed to the possibly conflicting influence of multiple central places and redundant outputs. CPT defines *Low Places*, typically manifested

as a town, village, or suburb, to be small population zones that live within the influence of a Central Place. Low Places strongly rely on their nearby Central Place for both low- and high-level outputs due to a small or weakly developed local economy and basic infrastructure. Finally, CPT identifies communities that live on the periphery of a Central Place's influence zone as a *Middle Place*. They are by necessity more self-reliant compared to low places because of their larger geographic distance to the Central Place. As a consequence of their self-reliance, Middle Places are able to produce some, but not all of the low-level outputs provided by Central Places and may remain dependent on them for many high-level outputs. However, Middle Places tend to be situated between a number of Central Places and thus agglomerate the collection of resources the Central Places provide into one location [40]. The ability to collect resources from many different Central Places, each of which may offer different types of low- and high-level outputs, mean Middle Places have the unique opportunity to integrate the ideas and norms of many Central Places and may generate their own culture and new knowledge [39].

Patterns where small communities are placed near large ones and moderate sized communities are greater distances from large ones is a sign that CPT explains regional development. Such a pattern has been noticed across the world [10, 28, 36, 17, 49]. With this anecdotal evidence, researchers have devised models that uses CPT tenents to explore facets of regional development. For example, Richardson *et al.* create models for predicting the size of city populations and their growth rate based on the hierarchical city size distribution that CPT asserts [47]. Hsu *et al.* use the CPT notion of spatial hierarchy to identify the aspects of an optimal distribution of cities given constraints about the scale of their local economies and total size [25]. Ikeda *et al.* build a model based on bifurcation analysis to tie CPT to economic geography theories, bridging CPT to multiple modern explanations of regional economic development [27]. Daniels integrate CPT into a model that predicts whether or not a destination will successfully profit by hosting a major sporting event [11]. Although all of these models are theoretically sound, they may not be applicable without satisfactory evidence that CPT reasonably explains the development of cities in a region. Furthermore, the models may not be practical without a way to quantify the extent to which the tenents of CPT they are built on applies.

CFT is a recent, complementary theory for explaining urban development [55]. Whereas CPT explains the effect of Central Places and their influence over local regions, CFT is based on non-local interactions among places without regard for physical distance. It emphasizes the cooperative aspects of place interactions where information, ideas, specialists and other 'foreign' commerce are exchanged for mutual economic benefit. CPT and CFT are complementary in that, while large Central Places interact with their geographic surroundings and nearby cities (CPT) [24] to provide outputs that drive their economy, their further development hinges on the free exchange of ideas and integration of 'foreign' ideas and commerce (CFT) [54, 53]. This complementary relationship has been seen in the historical development of various urban places [58, 46, 20] and can be inferred from theoretical models [31]. Like CPT, the application of CFT is limited without data that quantifies the total 'transportation of information' between places in a region. This information may be provided by traffic flows between places [29], by the rate of requests among Internet servers [21], or by telephone conversations between cities [32].

3 CPT and CFT Signals in Mobile Phone Data

In this section, we explore how the records of mobile phone calls made over a geographic region can be used to quantify aspects of CPT and CFT. For this purpose, we consider a year-long dataset of mobile phone data captured from January 1st 2013 through December 31st 2013 [5] across Senegal. The data provider filtered out records from mobile phone devices that were active less than 75% of the days during this period and from devices that made more than 1,000 calls per week. For privacy protection purposes, the data did not contain information about individual devices, rather the total number of calls placed between antenna's across the country, the total duration of these calls per day, the geo-location of the antenna, and the arrondissement (administrative area) the antenna belongs in. In total, 2.079 billion calls between 1,666 phone towers are included in the dataset. We next explore how patterns relating to different CPT and CFT concepts, namely *prominence*, *distance*, *partnership*, and *centrality*, can be measured and evaluated for each arrondissement using mobile phone data. These characteristics, while not exhaustive, represent core aspects of CPT and CFT and hence may be used to quantify and evaluate how well the theories explain urban development in a region. Since the data captures calls placed between towers in different arrondissements rather than cities, we consider the arrondissement to be the unit of "place" for CPT and CFT. For example, arrondissements that match the characteristics of a Central Place likely foster some of the most populated and important cities in a country.

3.1 Prominence

Communication is a basic necessity for the trade of goods and services, the spread of knowledge and ideas, and to disseminate influence across a broad region. The total number of calls placed within an arrondissement is thus a measure reflecting its *prominence*, which we define as a value proportionate to the social, technological, and economic progression of the cities within it. It is intuitive to believe that any Central Place should exhibit large prominence as it communicates ideas and supports the economy of the low places that depend on it. Large prominence may also be explained by the way Central Places impart commerce, knowledge, and influence onto nearby Middle Places. Middle Places should have moderate prominence because, by definition, they have a large, self-sustaining population and a need to communicate with Central Places to support their economy and exchange new ideas. Owing to their influence and size, CPT expects a small number of Central Places, more but still few Middle Places positioned far from many Central Places, and a large number of Low Places that rely on nearby Central and Middle Places.

We explore the prominence of arrondissements in Senegal by plotting the distribution of the number of calls placed from them on log-log scale in Figure 2. We observe a slow decay in the call volume of the twenty most active arrondissements before a rapid exponential drop. This behavior is a signature of a dataset that is log-normally distributed, which we confirm by fitting a log-normal distribution as the black trend in the figure. Log-normal distributions are a class of distributions that all exhibit a *long* or *heavy-tail* [35], where the probability of observing values many times the mean of the distribution is not negligibly small.

From the above discussion, the arrondissements whose call volume dwarfs all others (those in the left tail of the

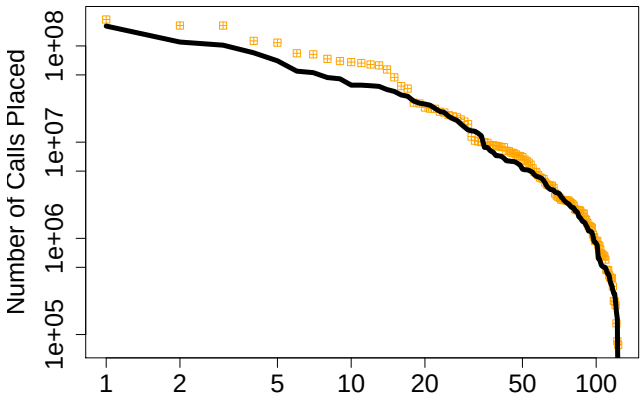
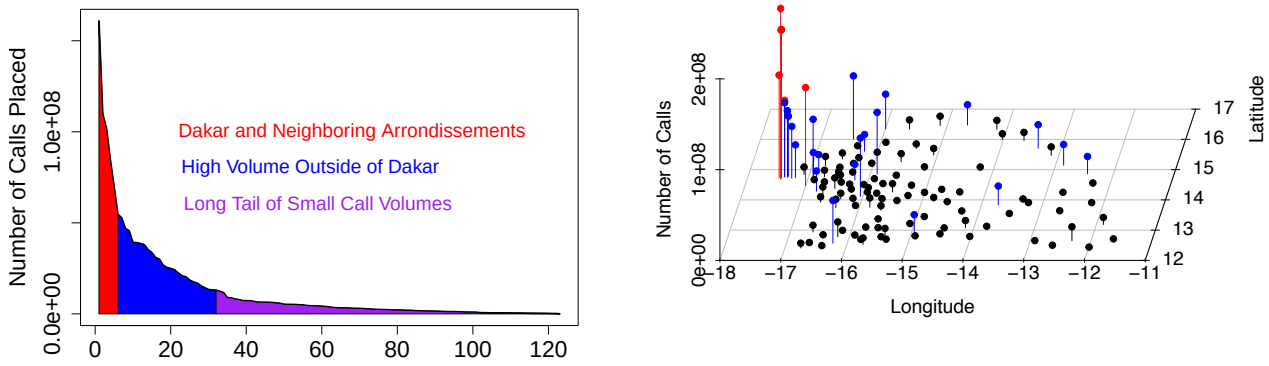


Figure 2: Rank-order distribution of calls placed per arrondissement and log-normal fit



(a) (b)

Figure 3: Positions of arrondissements in the left tail, body, and right tail of call volume distribution

distribution in Figure 2) exhibit high prominence, and hence, may be Central Places responsible for supporting a number of low places. To identify regions of 'high', 'moderate' and 'low' prominance corresponding to Central, Middle and Low Places respectively, we consider a classification of arrondissements by their call volume into three separate groups: one representing the left tail of the call volume distribution, another representing those whose calling volume is significant but lower than those seen in the left tail, and a third group that exhibits low call volumes in the long right tail of the distribution. Given the skewness of the distribution (the mean of the distribution is 16,900,544 calls, yet the median is only 3,816,449 calls), we use the following heuristic: (i) if the number of calls in an arrondissement is less than the mean of the distribution, consider it as part of the right tail; (ii) if the number of calls in an arrondissement is greater than and within two standard deviations of the mean, consider it as the body of the distribution; (iii) otherwise, consider the arrondissement as in the left tail of the distribution.

Figure 3(a) visualizes the partitioning of arrondissements into the left (red), body (blue), and right tail (purple)

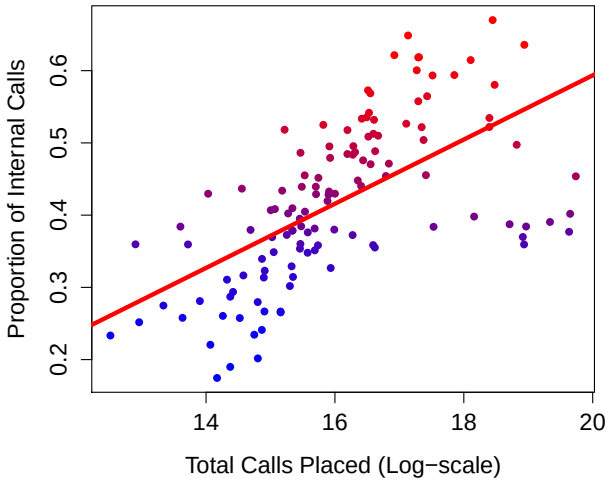
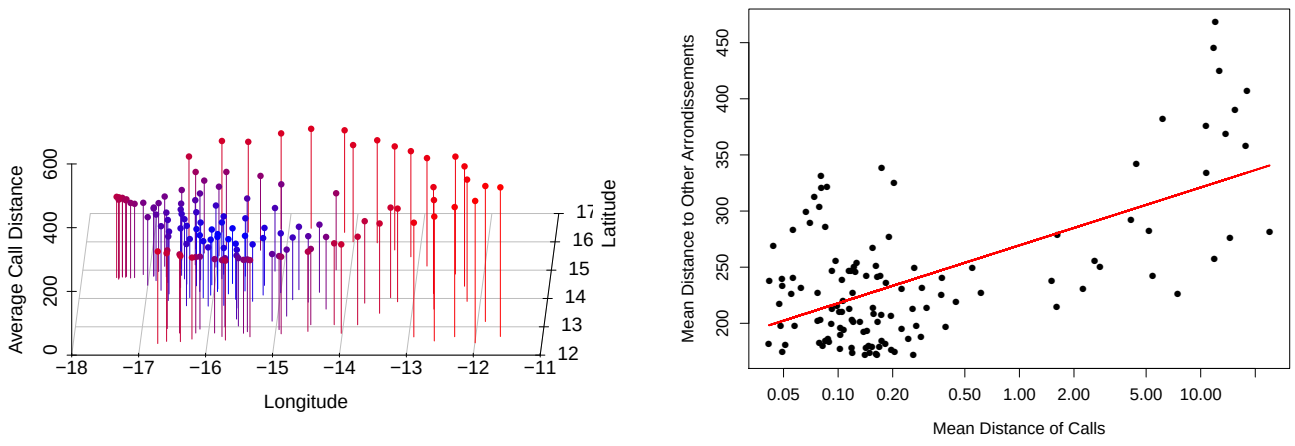


Figure 4: Relationship between total and internal calls of an arrondissement

of the call volume distribution representing 5%, 18%, and 77% of all arrondissements respectively. We observe that the proportion of arrondissements falling into the left, body, and tail reflect the proportion of Central, Middle, and Low Places expected to appear in CPT's idealistic spatial hierarchy of places in Figure 1, with a small number of Central, moderate number of Middle, and a large number of Low Places. Figure 3(b) further demonstrates links between a region's prominence and CPT by plotting the positions of arrondissements by latitude and longitude along with the color of their classification. Places in the left tail of the distribution correspond to Dakar and its suburbs, which is an area of Senegal that has an intense population density and is the economic center of the country. We also find, as anticipated by CPT, that places in the body of the call volume distribution are well dispersed through the country. The significant call volumes seen in these places suggest that they are self-sufficient with a significant population, and likely satisfy the definition of a Middle Place.

We further explore the notion of self-sufficiency in arrondissements that could be Central or Middle Places based on calling volume. According to CPT, Central and Middle Places should exhibit a high level of self-sufficiency compared to Low Places that are dependent on nearby Central Places. This self-sufficiency can be seen by considering the percentage of all mobile phone calls that are among towers *within* the same arrondissement. It reflects the "locality" of calls made within an administrative region; areas with strong internal communication suggest a weaker reliance on the information provided by people located in other arrondissements. We thus compare the total number of calls placed in an arrondissement against the number of such calls that are internal in Figure 4. The figure suggests a strong positive relationship between the level of internal calls (self-sufficiency) and total call volume (prominence). This finding is in harmony with how self-sufficiency and prominence are related in CPT, giving further support that the call volume of an arrondissement captures the notion of prominence.



(a) (b)

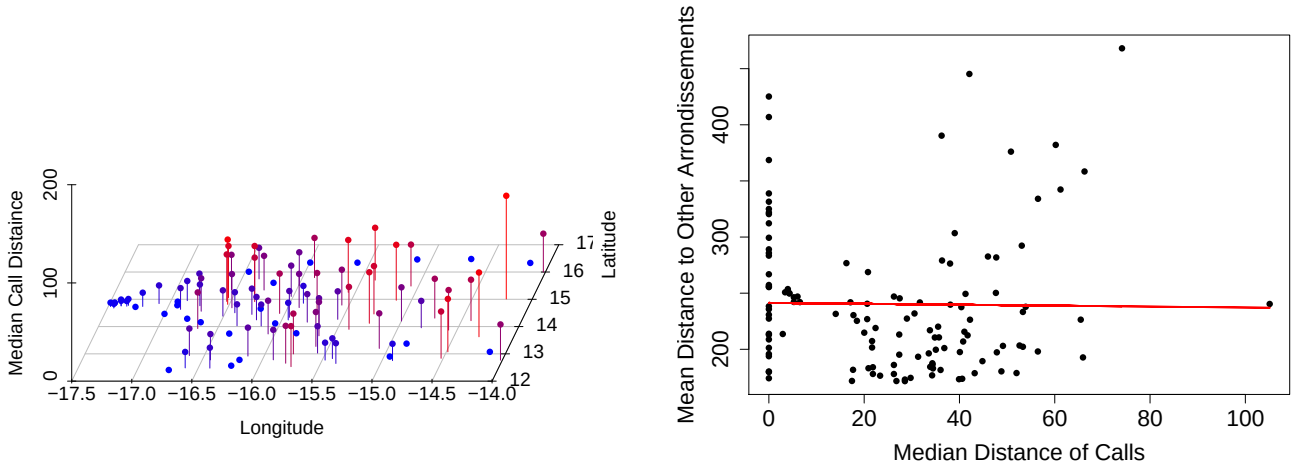
Figure 5: Average calling distances and correlation with distance from other arrondissements

3.2 Communication distance

Another aspect of CPT concerns the distance with which goods and ideas travel in a country. According to the theory, goods and ideas at Central and Low Places should only travel a small distance to the surrounding places. However, Middle Places may send goods, services, and knowledge larger distances since they are positioned far away from Central Places that are economic and cultural hubs in a country. Using mobile phone data, we may approximate the distance that goods and ideas travel by the distance calls made from an arrondissement travel. This is because a communication from one place to another signals the transmission of an idea, and because communication is a prerequisite for nearly every economic exchange.

We define the 'distance' of a communication as the physical distance between the calling towers that transmit a mobile phone call. Although the average distance of communications originating from cell towers in an arrondissement seems like a natural way to define communication distance, plotting this average against the arrondissement's position in Figure 5(a) shows a natural skew towards locations farther away from the center of the country. Figure 5(b) quantifies this skew by showing a positive linear correlation between an arrondissement's log-transformed average call distance and its average distance to all other arrondissements. This pattern is a natural byproduct of the physical geographic position arrondissements are located in. For example, it is likely that arrondissements around the border have large call distance by virtue of the fact that they are far away from any other arrondissement, not because they exhibit the tendencies of a Middle Place.

An order statistic may exhibit less sensitivity to an arrondissements position compared to the mean. For example, consider one of the border arrondissements with large mean calling distance in Figure 5(a). It may be the case that most calls are placed between other nearby arrondissements, but the distance that a smaller number of calls placed across the country (such as to Dakar) significantly influences the mean. An order statistic such as the median, however, better captures the smaller number of calls to nearby arrondissements. In fact, Figure 6(a) visualizes



(a) (b)

Figure 6: Median calling distances and correlation with distance from other arrondissements

how the median distance communication travels has little association with its physical location. Furthermore, Figure 6(b) identifies no statistically significant correlation (Pearson's $\rho = -0.1214; p = 0.8777$ and Kendall's $\tau = -0.0988; p = 0.1132$) between the location of an arrondissement and the median of its call distance distribution. The relatively high median call distances correspond to arrondissements where the majority of calls placed are to far away locations, independent of its physical location. In Figure 6(a), we observe that many of the same arrondissements whose total calling volume indicated it may be a Central or Low place according to CPT (Figure 5) exhibits a low median communication distance, as anticipated by the thoery. We also observe a reationship between places that exhibit high median calling distances and those falling into the body of the call volume distribution in Figure 3. This offers some evidence that median calling distances may be used to quantify the concept in the context of CPT.

3.3 Partnership

According to CPT, Middle Places synthesize information from a number of other places to create new products and knowledge. The degree to which this combining occurs may be quantified with mobile phone data by studying the number of "most active communication links" of a place. For example, Middle Places may establish a large number of active links between other Central and Low places to accumulate, create, and share new knowledge with others. We thus say that Middle Places exhibit a high level of *partnership*. On the other hand, Central and Low places may only create a small number of active links, most of which may be local, to give (receive) support to nearby Low (from nearby Central) Places. We therefore measure the partnership of a place by the number of ' "most active" communication links it maintains. This measure is derived for an arrondissement A by counting and ordering the number of times a tower in A connects to a tower in every other arrondissement B. Following the Pareto principle where 80% of a quantity may be attributed by the top 20% 'richest' of a population [30], we count

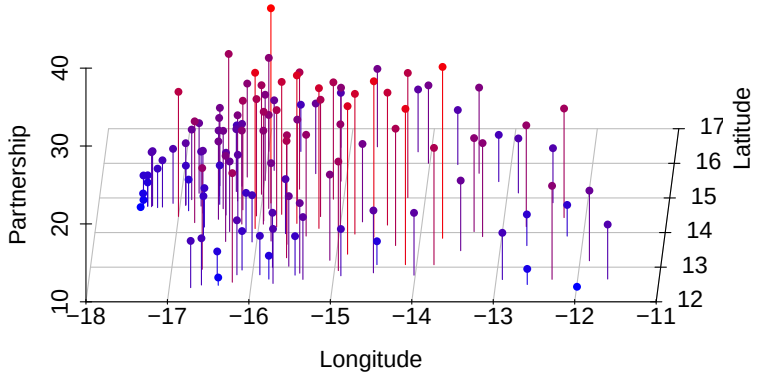
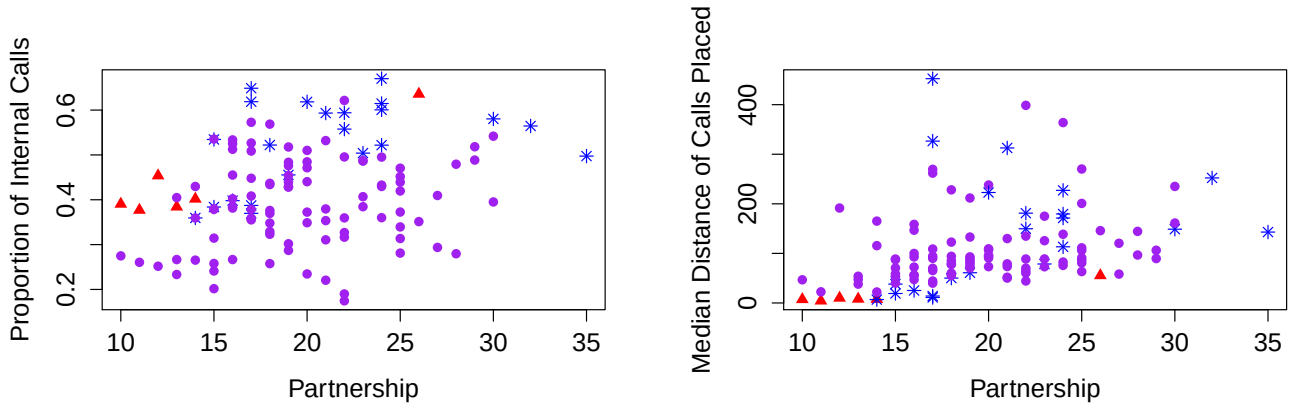


Figure 7: Spatial distribution of partnership values across Senegal

the number of arrondissements that are contacted in the top 20% of A's most often used communication links. Middle Places should have large levels of partnership as they need to maintain large volumes of communication to develop, thus a large number of number of arrondissements should be contacted in their most active communication links. Central and Low Places, however, may exhibit little partnership as the majority of their communications are devoted to a small number of regional locations.

We explore the spatial distribution of arrondissements and their partnership values in Figure 7. Arrondissements around Dakar and its suburbs have very low partnership, as anticipated by CPT since Dakar is a Central Place of the country. Furthermore, locations that would be described as a Middle Place given their prominence in Figure 3(b) tend to also exhibit high partnership in Figure 7. Further comparison of the figures, however, reveal that more arrondissements have high levels of partnership compared to moderate levels of prominance. This could indicate that places with low prominance may exhibit high partnership, which is inconsistent with CPT. We therefore check if our measure of partnership is completely incompatible with CPT by also examining its relationship with communication distance and self-sufficiency in Figure 8. Each point corresponds to an arrondissement and its shape and color represents whether its location lies in the left tail, body, or heavy right tail of the call volume distribution. It shows that arrondissements falling into the body or the right tail take on partnership values that span a similar range, but those in the body tend to have a larger proportion of calls that are internal (mean internal proportions of 52% and 38%, respectively) and tend to send external communications larger distances (average distance of 144.75km and 105.79km, respectively). Arrondissements whose prominance is moderate and communication distance is far thus exhibit the highest levels of partnership, matching the theoretical definition of a Middle Place. Thus, although partnership alone cannot disambiguate Middle Places from the remaining Low Places, but it does offer supporting evidence when considered alongside prominance, communication distance, and



(a) (b)

Figure 8: Comparing partnership against self-sufficiency (a) and median calling distance (b). Arrondissements are labeled ▲, ∗, • if they fall into the left tail, body, and right tail of the call volume distribution respectively.

self-sufficiency.

3.4 Centrality

We quantify a critical aspect of CFT, the amount of information exchanged among places, as the total duration of conversations made between cell towers across different arrondissements. If we consider CFT in conjunction with CPT, we hypothesize that Central Places will communicate with numerous other locations in a country for extended periods of time as it exposes its high-level outputs. In other words, the sum of the duration of communications from towers in a Central Place to other places may be among the longest in the country, and the Central Place should serve as a kind of "hub" for information exchanges. Note that this is different from the concept of prominance, which only measures the volume of communication without regard for the number of other places such communication travels to or from. We also expect Middle Places to participate in moderately long communications with other places as they accrue high-level outputs from many Central Places. They may also serve as a communication hub as they communicate with the Central Places that surround them, develop new ideas from these connections, and share this new knowledge with other places.

The graph theoretical concept of *pagerank centrality* may be used to quantify the "importance" of an arrodissement with respect to the number of other places it connects to, the frequency it connects to such places, and the "importance" of those connecting places [43]. Central and Middle Places should therefore exhibit high pagerank centrality because Central places are expected connect to scores of other places, and Middle Places are expected to rely on many Central places, which themselves have high pagerank centrality, to accrue their high-level outputs. To derive the pagerank centrality of arrondissements, we represent the mobile phone dataset as an undirected network where nodes represent calling towers and edges are weighted by the sum of the length of all mobile phone calls placed between them. The pagerank centrality p_i of calling tower i is defined by:

$$p_i = \alpha \sum_j \mathbf{A}_{ij} \frac{p_j}{g_j} + (1-\alpha)\frac{1}{N}$$

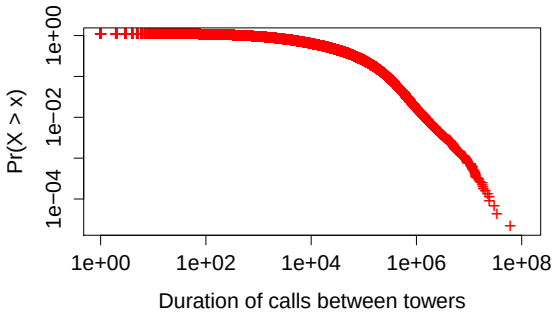


Figure 9: Distribution of total calling times across towers

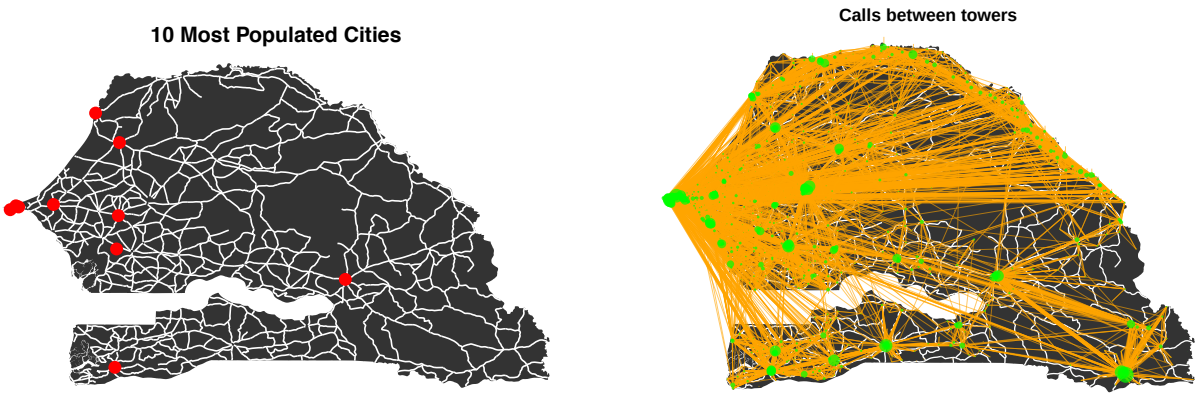


Figure 10: Spatial comparison of the most populated cities in Senegal and Pagerank centrality of calling towers

where $\mathbf{A}$ is a matrix with elements $\mathbf{A}_{ij}$ as the cumulative length of all phone calls between towers i and j, k_j is the degree of node j, $g_j = \max(1, k_j)$, N is the number of towers, and $\alpha = 0.87$ is a damping parameter set according to the recommendations of earlier work [6].

We measure the PageRank centrality of connections among the 1,666 towers in the country. Figure 9 plots the distribution of these durations and exhibits a clear power-tailed shape. The power tail in the distribution of call times are a common phenomenon across mobile phone datasets [14, 44, 8]. We therefore only consider communication between towers whose cumulative duration of all conversations fall in the top 1.5% of this distribution, where significant calling activity occurs. On average, these 38,613 remaining flows exhibit an average of 2,739 minutes of conversations per day. Figure 10 compares the location of the 10 most populated cities in Senegal from the Global Gazetteer online database[1] against the location and PageRank centrality of calling towers (larger vertices correspond to higher PageRank). It shows a strong correlation between the position of the most popular cities (located in Central Place arrondissements) and the location of call towers that exhibit the largest amount of activity. We also observe many call towers with high PageRank lying between these most populated cities. These locations likely lie in arrodissements fitting the profile of CPT Middle Places. For example, Figure 11 compares

[1] `http://www.fallingrain.com/world/index.html`

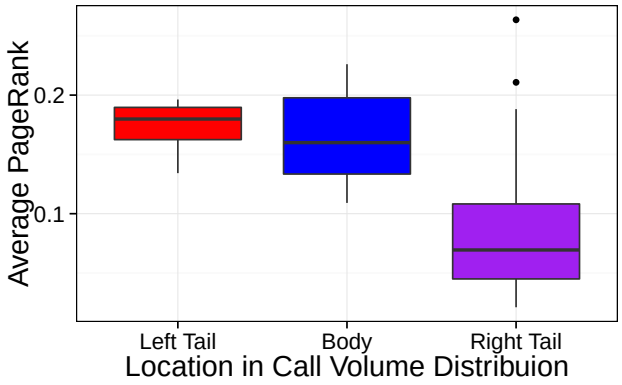


Figure 11: Average PageRank of towers in arrondissements in the left, body, and right-tail of call volume distribution

the average PageRank centrality of antennas falling in arrondissements that are in the left, body, and right tail of the call volume distribution from Figure 3. It shows that arrondissements in the left and body of the tail (CPT Central and Middle Places) have significantly larger centrality compared to those in the right tail (Low Places). This relationship exemplifies the relationship CFT and CPT have with each other; by CFT, places who communicate with other important places tend to transmit high-level outputs that cooperatively help them develop, and by CPT, Central and Middle Places are precisely the ones with such high-level outputs to share.

4 Operationalizing CPT and CFT to Address Urbanization

The previous section demonstrated that concepts of CPT and CFT may be quantified using mobile phone records captured across a country. It also demonstrated that, in the case of Senegal, measures of the prominence, communication distance, partnership, and the centrality of arrondissements are correlated in ways anticipated by the two theories. This establishes a quantitative support that CPT and CFT may explain the development and positioning of arrondissements in Senegal. With this support, we next demonstrate an a case study that operationalizes CPT and CFT concerning urbanization.

4.1 Urbanization in developing countries

A virtually universal trait across developing countries are significant rates of *urbanization*, defined as the migration of citizens from traditional, tribal, and rural regions to large city centers [19]. Ever increasing political turmoil in rural or tribal towns, ecological breakdowns, and the romantic or unrealistic notion held by citizens that great opportunity exists in urban areas all contribute to high urbanization rates [34]. Urbanization is a pressing challenge for a developing nation because it leads to poor living conditions in her most prominent cities as their population exceeds its capacity with respect to infrastructure, public works, and job availability. It also encourages an unstable bazaar economy that is impossible for the country's government to tax or regulate, high rates of crime, and pollution. The country of Senegal is no exception to the urbanization phenomenon; over 42.5% of the population lives in urban

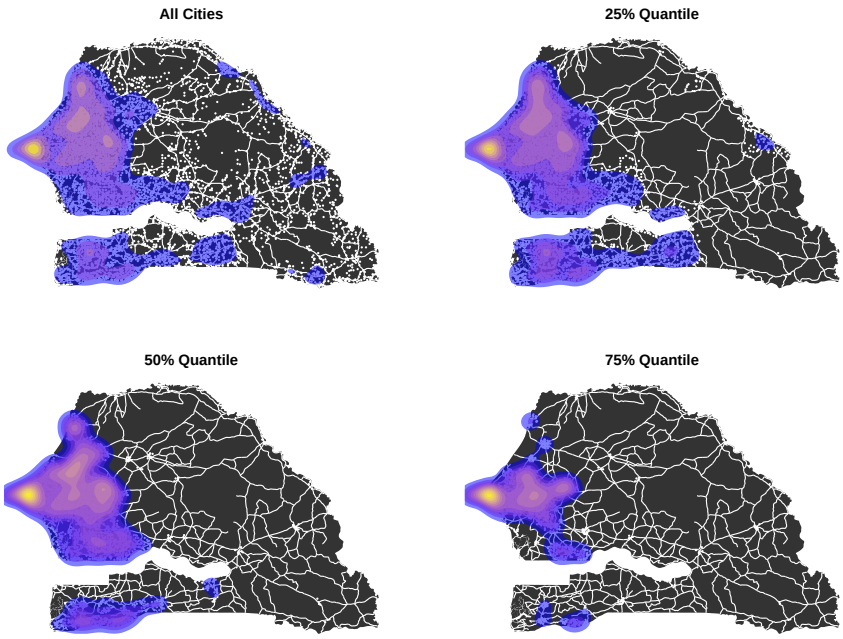


Figure 12: Evidenceof urbanization in Senegal. Each plot filters out cities whose population density falls below the stated quartile. The 25% largest cities are mostly concentrated in a band east of the capital city of Dakar.

areas and over 71.9% of citizens living in the country's 50 most popular cities reside in Dakar and Grand Dakar[2]. To demonstrate the intensity with which urbanization occurs in Senegal, Figure 12 uses data from the Global Gazetteer database to show how the majority of her population is concentrated on the west coast, and the top quarter of cities with highest population density lies only in a region to the east of Dakar and Grand Dakar.[3]

Urban planning researchers and policy makers concur that an effective way to reduce urbanization is to encourage a country's citizens to migrate out of, rather than into, overpopulated urban centers by investing in the rapid development of promising towns and cities in alternative areas of the country [23]. Doing so simultaneously relieves the pressure applied to large central cities while investing dollars into the development of new cities, adding to the country's economy. The ideal town or city for rapid investment is one that already has an established local economy, a developed infrastructure that supports its present inhabitants, and is self-sustaining; that is, it is located sufficiently far from existing large urban centers so that it does not rely on their economy, people, or services to thrive [3]. These features nearly match the description of a Middle Place defined by CPT.

An ideal way to identify locations where urban development may combat urbanization may therefore be to rank each arrondissement by the degree to which they exhibit the qualities of a Middle Place. However, such a ranking is challenging to derive quantitatively because the relative importance of the qualities of a Middle Place for measuring the positive effects of urban development is difficult to identify. We thus seek a broader, unsupervised analysis of arrondissements that classify them into 'types' based on the similarity of CPT and CFT related features. The class of arrondissements represented by feature values that best reflect Middle Places may thus be a collection of places

[2]http://www.indexmundi.com/senegal/urbanization.html
[3]http://en.wikipedia.org/wiki/Template:Largest_cities_of_Senegal

that are suitable candidates for economic investment. To perform this classification, we consider finite mixture models to search for a best fitting mixture of probabilistic models that explain the distribution of values across all features in the dataset. Since class labels are derived by the mixture component that would most likely generate the features of an arrondissement, this approach relaxes many of the constraints imposed by distance-based clustering algorithms that are typically restricted to searching for linear divisions between groups [22]. We use the `mclust` finite mixture modeling software package in R to search for clustering solutions where the mixture components are members of the exponential family of distributions. The package reports results from automatically composed probabilistic models tuned to parameter settings that encode assumptions about the shape of the distributions and the number of clusters considered [16].

4.2 Model selection

Because no observable outcome exists to compare model validity against [38], the criteria used for mixture model selection carries an inherent level of subjectivity. We thus adopted the following criteria to evaluate a potential solution. They are ordered by priority below:

1. **Multicollinearity**: High correlations among independent factors can skew a model result by counting multiple explanations of the same variation in the data twice. Thus, two variables that are highly correlated are not introduced into the models because they overstate the impact of their phenomena on the solution. Correlations larger than 0.5 are considered high, and correlations between 0.3 and 0.5 are monitored as we evaluate the solution using the remaining criteria below.

2. **Actionability**: This criterion asks if the clustering results offer actionable insights to decide where resources for urban development should be targeted. For example, if the average features of a cluster of arrondissements reflected a Middle Place but includes Dakar (a known Central Place) or contains the majority of arrondissements in the country, the ability for an analyst to take action from these results is limited. This is a logical and subjective, yet necessary, criterion.

3. **Bayesian Information Criteria (BIC)**: Finite mixture models commonly use BIC as a statistic for comparing the quality of different clustering solutions [16]. It is defined as:

$$B = 2 \log P(D|M, \Theta) - d \log n$$

   where D is the set of data vectors, M is the fitted clustering model with maximum likelihood parameters Θ, $d = |\Theta|$, and $n = |D|$. Models with larger B tend to be better models since greater log-likeilhood values are indicative of whether the data D fits the model $M(\Theta)$ well.

4. **Pseudo-F Statistic**: The Pseudo-F statistic is a measure of the efficiency of a clustering result. It is defined as the ratio of the mean sum of squares distance between vectors in different clusters to the mean sum of squares distance between vectors in the same cluster [33]. Larger Psudeo-F scores correspond to tighter clusterings where intra-cluster distances between vectors is small and inter-cluster distances are high.

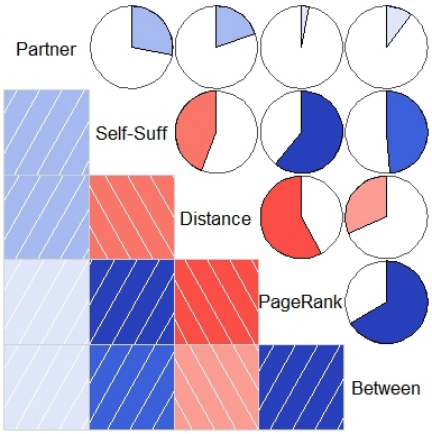


Figure 13: Correlations between CPT and CFT features. PageRank shows high ($|\rho|> 0.5$) correlations with call distance and self-sufficiency, while betweenness centrality is only moderately correlated with Self-Sufficiency.

Solution	Variables	BIC	Pseudo-F
Best	Self-Sufficiency, Partnership, Betweenness Distance ($X = 60\%$)	-1,288	41.6
Alt. A	Self-Sufficiency, Partnership, Betweenness, Distance ($X = 50\%$)	-1,297	46.2
Alt. B	Self-Sufficiency, Partnership, Betweenness, Demand-Weighted Dist.	-1,342	31.0
Alt. C	Self-Sufficiency, Partnership, PageRank Centrality, Demand-Weighted Dist.	-1,316	38.4

Table 1: Clustering solutions with different variable settings

We considered the self-sufficiency, communication distance, partnership, and PageRank centrality of each arrondissement taken from the mobile phone dataset in our clustering analysis. We intentionally do not consider the prominance of a location because the skew in the distribution of call volume overemphasizes the impact of this feature in the clustering solutions. We also found that features exhibit multicollinearity; for example Figure 13 identifies how PageRank centrality exhibits high correlation with call distance and self-sufficiency. We overcome this problem by considering the *betweenness* centrality of arrondissements instead, which is defined as the number of shortest communication paths in the country that pass through the arrondissement being measured. Like PageRank, betweenness centrality reflects the ability of cities in an arrondissement to connect to other locations in Senegal, thus acting as a hub of information and resources and as a place where ideas and knowledge across the country meet. However, betweenness centrality gives no consideration to the total length of communications between arrondissements. The choice of betweenness centrality in our analysis is thus a compromise: it only captures the degree to which an arrondissement is a hub, yet as see in Figure 13, exhibits less correlation with the other features compared to PageRank.

Table 1 enumerates through the four best clustering solutions found by the `mclust` package that were built from different subsets of CPT and CFT features. We found that the solution given in the first row identifies four clusters using the self-sufficiency, partnership, betweenness centrality, and communication distance defined by the $X = 60^{th}$ percentile. The 60^{th} percentile was used since, as seen in Table 2, it serves as an approximate elbow point that

Distance Traveled by X% of Calls	Correlation with Self-Sufficiency	Variance of Distance Traveled
50% (median)	-0.58	454
60%	-0.37	854
70%	-0.17	3,581
80%	0.10	10,872

Table 2: Correlation between distance of calls and self-sufficiency features

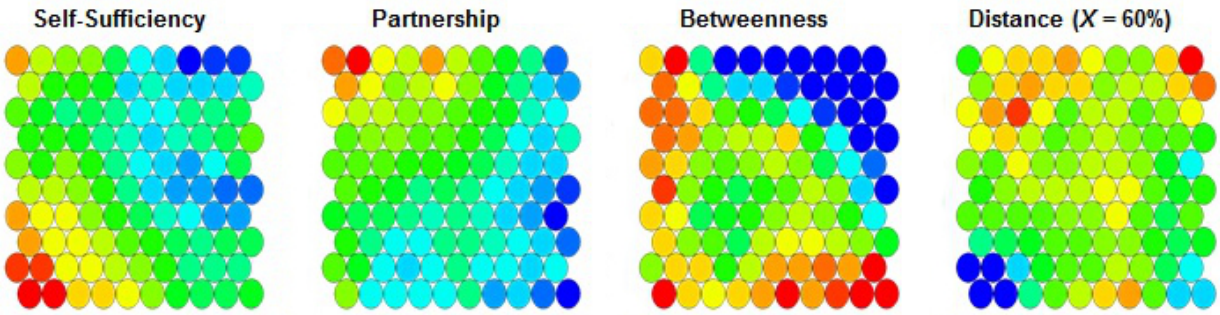


Figure 14: SOM-based visualization of groups of values for features used in the best clustering solution

reduces the correlation of communication distance with self-sufficiency while maintaining a low amount of variance in the feature's values. Because it exhibits less multicolinearity among the features used in the clustering, this solution is thus superior to the first alternative A in Table 1 even though they have very similar BIC and Psudeo-F scores.

Figure 14 uses a Self-Organizing Map (SOM) to visualize the distribution of self-sufficiency, partnership, betweenness centrality, and communication distance in the best clustering solution across the arrondissements. SOMs arrange the values of a feature across cells of a hexagonal grid, and then compares the similarity of nearby cells to decide if values should be reorganized or merged together. The result of a SOM is a two-dimensional grid where similar observations are grouped together in a neighborhood [37, 57]. They thus visualize the proportion of arrondissements that take on similarly large (red) or small (blue) values in the best clustering solution. The maps identify how the distance of calls, partnership, and self-sufficiency metric exhibit a skew towards a small number of arrondissements, which according to CPT may be those that represent Middle Places. The more even distribution of betweenness centrality is likely due to the fact that both Middle and Central Places are important brokerage locations for information and communication across the country; hence both types of places may be represented by the hotter colored nodes. The large number of cool colored betweenness centrality and partnership cells may capture the Low Places that do not serve as brokers of any kind of information nor do they communicate with a large number of external places.

Figure 15 presents centroid positions of the four clusters that emerge in the best finite mixture model as a dot plot. These values are subjectively mapped to being relatively low (□), moderate (■), or high (★) in Table 3. Based on this mapping we describe the clusters as follows:

Cluster Description	Self-Suff.	Partnership	Betweenness	Distance	Cluster Size
Dakar and Suburbs	■	□	★	□	8
Emerging Opportunities	★	■	★	□	9
Low Places	□	■	□	■	37
Common Towns	■	■	■	■	69

Table 3: Cluster labels and features (□: Low; ■: Moderate; ★: High)

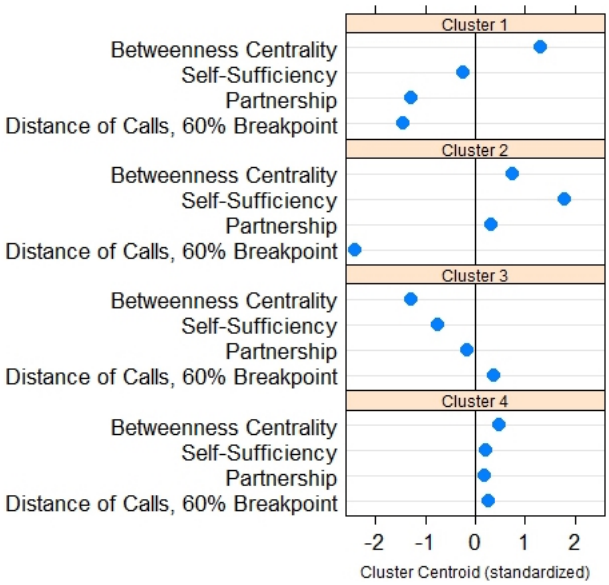


Figure 15: Dot plot of cluster centroids

- **Cluster 1: Dakar and its Suburbs**. This cluster has eight arrondissements that include Dakar and its suburbs. The arrondisements show high betweenness, meaning they are hubs for calls throughout the country. Yet their low call distance and partnership implies exclusivity: information flow passes primarily through partners within the same cluster. This high betweenness and heighted interactivity within arrondissements in the cluster paints Dakar as the principle central place of the country.

- **Cluster 2: Middle Places**. The nine arrondissements in this cluster have qualities that best support the definition of Middle Places. They have high self-sufficiency, and calls leaving these places reach out to a number of other places. Although the lower calling distance diagrees with the moderate to large calling distances expected by CPT, they are the only cluster to feature very high self-sufficiency and partnership values. Such features, along with the small size of the cluster, match the definition of a CPT Middle Place [9].

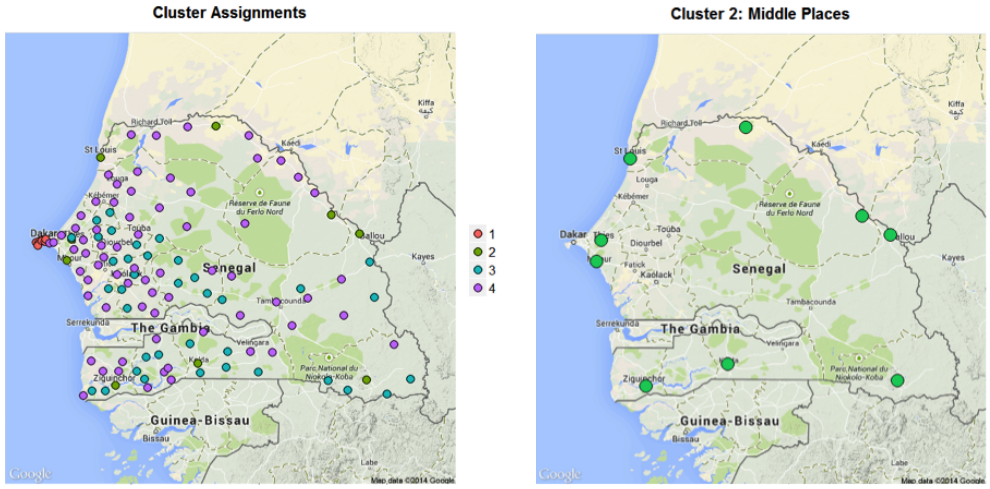
- **Cluster 3: Low Places**. The 37 arrondissements in this cluster exhibit a low degree of self-sufficiency and betweenness, and a moderate level of partnership and call distance. Low self-sufficiency is an indicator of a Low Place that relies on nearby other places for resources and information. Similarly, a low betweenness value indicates that the location is not a broker of information, and that they are not of interest to most other arrondissements. In Figure 16(a), the Low Places (blue positions) tend to be surrounded by a number of other nearby arrondissements, further emphasizing their reliance on nearby central, middle, or low-middle places.

- **Cluster 4: Low-Middle Place**. The majority of arrondissements fall into a cluster with moderate self-sufficiency, betweenness, partnership and distance, qualities of strong Low Places or a faintly emerging Middle Place. The positions of such arrondissements in Figure 16 find them to be either near Dakar and its suburbs, by the border of the country, in remote regions, or immediately surrounded by arrondissements that only support low places.

The features of Cluster 2 give the strongest evidence that its arrondissements support Middle Places. Because no reliable ground truth data about the economy or propserity of cities in these arrondissements are available, however, we cannot quantitatively assess whether locations in these arrondissements are indeed promising locations for investment. Instead, we performed a manual investigation into cities within arrondissements in this cluster, and find qualitative features that justify why they may fit the definition of a Middle Place and be promising opportunities for economic development: (i) **Thies**. Thies is one of Senegal's largest cities and sits in an area considered to be a transportation hub that services routes between St Louis, Dakar and Bamako[4]. It is also a major producer of peanuts and fertilizer that are among the country's top exports, and host reserves of important metals[5]. Further investment could therefore push Thies to becoming a strong economic hub for the country. (ii) **St Louis**. St Louis is the capital of the St Louis arrondissement and is located in the northwest of the country near the mouth of the Senegal river on the Mauritanian border. It has a heavy tourism based economy and has a high rate of sugar production, pastoral farming, and exports of peanut skins. The city was listed as a UNESCO World Heritage Site

[4] http://bit.ly/1CtZOIO

[5] http://www.britannica.com/EBchecked/topic/592085/Thies



(a) (b)

Figure 16: Cluster assignments (a) and Middle Places (b)

in 2000, making tourism is a driver of its growth and of its ability to accumulate foreign ideas[6]. The returns on its diverse economy and cultural significance may be greatly enhanced by economic investment. (iii) **Mbour**. Mbour is a city that lies 80km south of Dakar. The city's major industries are tourism, fishing and peanut processing. It is Senegal's fifth largest city and, by some indicators, is among one of the fastest growing[7]. This growth may be further accelerated with economic investment. (iv) **Ziguinchor**. Ziguinchor is a river-port town in southwestern Senegal lying along the Casamance River. It is one of the largest cities in Senegal, but is separated from the north of the country by The Gambia [8]. Ziguinchor is principally a trading port, transportation hub, and ferry terminal for the country. A major highway in the country crosses the Casamance River by Ziguinchor, which links the region with the rest of Senegal. It also has a diverse agriculure including nuts and fruits, and opened a university for agglomerating and producing knowledge in 2007[9].

5 Conclusions and Future Work

This article demonstrated the promise of mobile phone data to operationalize aspects of Central Place and Central Flow Theory over a large geographic region. It specifically examined how the CPT and CFT notions of prominance, communication distance, partnership, and centrality may be quantified using records of mobile phone calls between cell towers. With patterns that suggest the notions of CPT and CFT may hold in Senegal, a case study that uses the operationalization to study urbanization in the country was presented. Clustering arrondissements based on CPT and CFT features in the mobile phone data revealed groups that match the definition of Central and Middle Places. Qualitative analysis of the cities in Middle Place arrondissements found them to be promising locations for economic development.

[6] http://whc.unesco.org/en/list/

[7] https://www.imf.org/external/pubs/ft/dp/2013/afr1304.pdf

[8] http://www.britannica.com/EBchecked/topic/657131/Ziguinchor

[9] http://www.univ-zig.sn

Our future work will continue to examine how CPT and CFT may be operationalized and the degree to which they hold be measured from data sets that capture activities across large regions. We will also refine our case study, and search for alternative datasets and signals that establish some ground truth as to the 'best' locations for economic investment in the region. This ground truth may let us formulate a ranking of the 'best' places for economic investment by optimizing over feature valuess associated with CPT Middle Places.

References

[1] P. M. Allen and M. Sanglier. A dynamic model of growth in a central place system. *Geographical Analysis*, 11(3):256–272, 1979.

[2] B. J. Berry. Cities as systems within systems of cities. *Papers in Regional Science*, 13(1):147–163, 1964.

[3] B. J. Berry and W. L. Garrison. A note on central place theory and the range of a good. *Economic Geography*, pages 304–311, 1958.

[4] B. J. Berry and W. L. Garrison. Recent developments of central place theory. *Papers in Regional Science*, 4(1):107–120, 1958.

[5] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 4(10):36–44, 2011.

[8] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[9] W. Christaller. *Central places in southern Germany*. Prentice-Hall, 1966.

[10] D. Christian. *Maps of Time: An Introduction to Big History, With a New Preface*, volume 2. Univ of California Press, 2011.

[11] M. J. Daniels. Central place theory and sport tourism impacts. *Annals of Tourism Research*, 34(2):332–347, 2007.

[12] D. Doran, A. Fox, and V. Mendiratta. Where do we develop? discovering regions for urban investment in senegal. In *Proc. of Intl. Conference on the Analysis of Mobile Phone Datasets*, 2015.

[13] D. Doran and V. Mendiratta. *Propagation Models and Aanalysis for Mobile Phone Data Analytics*, pages 257–292. Springer, 2015.

[14] D. Doran, V. Mendiratta, C. Phadke, and H. Uzunalioglu. The importance of outlier relationships in mobile call graphs. In *Proc. of Intl. Conference on Machine Learning and Applications*, volume 2, pages 24–29. IEEE, 2012.

[15] S. N. Durlauf, L. Blume, et al. *The new Palgrave dictionary of economics*. Palgrave Macmillan Basingstoke, 2008.

[16] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

[17] J. Friedmann. The world city hypothesis. *Dev. and change*, 17(1):69–83, 1986.

[18] M. Fujita, P. R. Krugman, and A. J. Venables. *The spatial economy: Cities, regions, and international trade*. MIT press, 2001.

[19] A. Gilbert and J. Gugler. *Cities poverty and development: Urbanization in the third world*. New York NY/Oxford England Oxford University Press, 1982.

[20] A. Growe and K. Volgmann. Exploring cosmopolitanity and connectivity in the polycentric german urban system. *Tijdschrift voor economische en sociale geografie*, 2015.

[21] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of internet hosts. *Networking, IEEE/ACM Transactions on*, 14(6):1219–1232, 2006.

[22] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*. Springer, 2009.

[23] V. Henderson. Urbanization in developing countries. *The World Bank Research Observer*, 17(1):89–112, 2002.

[24] P. M. Hohenberg and L. H. Lees. *The making of urban Europe, 1000-1994*. Harvard University Press, 1995.

[25] W.-T. Hsu, T. J. Holmes, and F. Morgan. Optimal city hierarchy: A dynamic programming approach to central place theory. In *Meeting Papers from Society for Economic Dynamics*, 2009.

[26] D. Husemann, R. Hermann, M. Moser, and A. Schade. Payment for network-based commercial transactions using a mobile phone, April 2001. US Patent App. 09/843,968.

[27] K. Ikeda, K. Murota, T. Akamatsu, T. Kono, Y. Takayama, G. Sobhaninejad, and A. Shibasaki. Self-organizing hexagons in economic agglomeration: core-periphery models and central place theory. Technical report, Technical Report METR 2010–28. Department of Mathematical Informatics, University of Tokyo, 2010.

[28] J. Jacobs. *The death and life of great American cities*. Random House LLC, 1961.

[29] W.-S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

[30] L. Kaplow and S. Shavell. Any non-welfarist method of policy assessment violates the pareto principle. *Journal of Political Economy*, 109(2):281–286, 2001.

[31] D. Knitter. Central places and the environment, 2013.

[32] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[33] L. K. Lim, F. Acito, and A. Rusetski. Development of archetypes of international marketing strategy. *Journal of International business studies*, 37(4):499–524, 2006.

[34] E. Linden. The exploding cities of the developing world. *Foreign Affairs*, 1996.

[35] L. Lipsky. *Queueing Theory: A Linear Algebraic Approach*. Springer-Verlag, 2nd edition, 2009.

[36] K. Lynch. *Good city form*. MIT press, 1984.

[37] J. Malone, K. McGarry, S. Wermter, and C. Bowerman. Data mining using rule extraction from kohonen self-organising maps. *Neural Computing & Applications*, 15(1):9–17, 2006.

[38] E. Malthouse. *Segmentation and lifetime value models using SAS*. SAS Inst., 2013.

[39] P. McCann and F. van Oort. Theories of agglomeration and regional economic growth: a historical review. *Handbook of regional growth and development theories*, pages 19–32, 2009.

[40] G. F. Mulligan. Agglomeration and central place theory: a review of the literature. *International Regional Science Review*, 9(1):1–42, 1984.

[41] J. Musil. Changing urban systems in post-communist societies in central europe: analysis and prediction. *Urban studies*, 30(6):899–905, 1993.

[42] D. Nakamura. Spatial competition and consumer exclusion: social welfare perspectives in central-place system. *Letters in spatial and resource sciences*, 3(3):101–110, 2010.

[43] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[44] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States*, 104:7332–7336, 2007.

[45] S. Openshaw and Y. Veneris. Numerical experiments with central place theory and spatial interaction modelling. *Environment and Planning A*, 35(8):1389–1404, 2003.

[46] D. C. Prosperi and A. C. Oner. Spatial impacts of megaprojects on the form of metropolitan regions: a theoretical inquiry. *International Journal of Society Systems Science*, 7(1):23–46, 2015.

[47] H. W. Richardson. Theory of the distribution of city sizes: Review and prospects. *Regional Studies*, 7(3):239–251, 1973.

[48] A. I. Saichev, Y. Malevergne, and D. Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer, 2009.

[49] S. Sassen. *The global city*. Princeton University Press Princeton, NJ, 1991.

[50] V. Shuper and P. Em. Moscow city expansion: An alternative based on central place theory. *Regional Research of Russia*, 3(4):376–385, 2013.

[51] J. W. Smith and M. F. Floyd. The urban growth machine, central place theory and access to open space. *City, Culture and Society*, 4(2):87–98, 2013.

[52] M. E. Smith. The aztec marketing system and settlement pattern in the valley of mexico: A central place analysis. *American Antiquity*, pages 110–125, 1979.

[53] E. W. Soja. Cities and states in geohistory. *Theory and Society*, 39(3-4):361–376, 2010.

[54] P. J. Taylor. Extraordinary cities: Early city-ness and the origins of agriculture and states. *Intl. Journal of Urban and Regional Research*, 36(3):415–447, 2012.

[55] P. J. Taylor, M. Hoyler, and R. Verbruggen. External urban relational process: introducing central flow theory to complement central place theory. *Urban Studies*, 47(13):2803–2818, 2010.

[56] Y. Veneris. *The informational revolution, cybernetics and urban modelling*. PhD thesis, University of Newcastle upon Tyne, 1984.

[57] R. Wehrens and L. M. Buydens. Self-and super-organizing maps in r: the kohonen package. *Journal of Statistical Software*, 21(5):1–19, 2007.

[58] M. Zhao, K. Wu, X. Liu, and D. Ben. A novel method for approximating intercity networks: An empirical comparison for validating the city networks in two chinese city-regions. *Journal of Geographical Sciences*, 25(3):337–354, 2015.