Do we need another coffee house?

The amenity space and the evolution of neighborhoods

César A. Hidalgo[1,2] and Elisa E. Castañer[1],

[1] Macro Connections, The MIT Media Lab, Massachusetts Institute of Technology

[2] to whom correspondence should be addressed: hidalgo@mit.edu

Abstract:
Neighborhoods populated by amenities—such as restaurants, cafes, and libraries—are considered to be a key property of desirable cities. Yet, despite the global enthusiasm for amenity-rich neighborhoods, little is known about the empirical laws governing the colocation of amenities at the neighborhood scale. Here, we contribute to our understanding of the naturally occurring neighborhood-scale agglomerations of amenities observed in cities by using a dataset summarizing the precise location of millions of amenities. We use this dataset to build the network of co-location of amenities, or Amenity Space, by first introducing a clustering algorithm to identify neighborhoods, and then using the identified neighborhoods to map the probability that two amenities will be co-located in one of them. Finally, we use the Amenity Space to build a recommender system that identifies the amenities that are missing in a neighborhood given its current pattern of specialization. This opens the door for the construction of amenity recommendation algorithms that can be used to evaluate neighborhoods and inform their improvement and development.

Significance Statement:

Mapping the amenities that are likely to locate in the same neighborhood has not been possible because of the scarcity of data on the precise location of amenities and because of the lack of an algorithmically implementable definition of a neighborhood. Here we use a dataset on more than 1.2 million amenities to introduce an algorithm to identify neighborhoods, and use the neighborhoods identified by this algorithm to map the network of amenities that are likely to co-locate in the same neighborhood. We validate the utility of this network of amenities by using it to create a recommendation algorithm that helps identify the amenities that are missing from a neighborhood.

Introduction

How do businesses choose where to locate? For decades, scholars have been studying where businesses locate and why businesses agglomerate. But while the theoretical literature explaining the location and agglomeration of businesses is long and vast[1-10], the empirical literature documenting the location of business, especially at the intra-city scale, is much shorter and more recent.

The theoretical efforts explaining agglomerations in cities go back to Alfred Marshall's industrial districts[1], and to the mathematical models advanced by Johann Von Thünen[2], Harold Hotelling[3], Walter Christaller, and August Lösch[4-7]. In Von Thünen's model, differences in land use are explained as consequences of a location's distance from the market, which is a center of agglomeration[2]. In Hotelling's model, businesses agglomerate to maximize their catchment area—that is, to be the closest business to the largest number of potential customers[3]. In Christaller and Lösch's central place theory, a hierarchy of central places emerges when goods differ on how far a person would be willing to travel to purchase them[4-7].

In recent decades, however, these seminal models were expanded to endow them with economic micro foundations[8-10] and to include new mechanisms that could help explain agglomerations that were left out of the seminal models. Among these additional mechanisms we have demand externalities[11-12], which predict agglomerations when planned purchases trigger unplanned purchases; search costs, which predict agglomerations when customers like to compare prices[12]; and transportation effects, which predict agglomerations because transportation technologies reduce the cost of carrying goods and create an incentive to bundle purchasing trips[12].

The richness of this theoretical literature, however, has not been matched by equivalent empirical work, especially at the neighborhood scale. The relative scarcity of empirical work stems in part from the lack of data on the precise location of businesses and amenities, and in part because working at the neighborhood scale requires an implementable definition of what a neighborhood is. Not surprisingly, these limitations have pushed empirical work on agglomerations to focus on coarser scales, such as countries, cities, and regions, where data is readily available and the spatial units of analysis are exogenously defined[13-16].

Yet, scholars have still made important methodological progress at these coarser scales, by advancing network techniques to map products that are co-exported[13-14], or industries that hire similar workers[15-16]. The advantage of these network techniques is that they help preserve the identity of the elements involved in a dataset, and as such, are useful to study the effect of product and industry relatedness on the process of industrial diversification.

An example of this work in the context of economic development is the network connecting products that are likely to be co-exported, or product space[13-14], which has been used to anticipate the evolution of a country's export structure and the constraints that different productive structures impose on the ability of countries to generate income. Similarly, at the regional scale, scholars have used networks of similarity between industries and occupations to study the importance of skill relatedness in the evolution of regional economies[15-16].

Here we extend the use of these network methods to neighborhood scale agglomerations by using data on the precise location of more than 1.2 million amenities and by introducing a spatial clustering method that we use to identify neighborhoods. These data and methods allow us to solve the technical problems that have limited mapping the network of amenities that are likely to co-locate in the same neighborhood. We then validate the utility of this network of amenities, or Amenity Space, to build a recommender system that exploits a neighborhood's pattern of specialization to estimate the number of amenities of each type that should locate in it. This recommender system allows us to detect amenities that are potentially missing from a location, and also, represents an extension of the network methods used at the international and regional scale, to the neighborhood scale.

Data

We use data from the Google Places API containing the latitude, longitude, and type of amenity (i.e. cafe, restaurant, library, etc.), for more than 1.26 million amenities across 47 US cities (see SM for details). Certainly the data from Google's Places API is not free of biases. The data on amenities registered in Google Places focuses on customer facing businesses and places of interests (from hair salons and bakeries to airports and cemeteries), and hence, fails to include information on other forms of economic activity, such as manufacturing. Also, the data might have coding issues, such as having a restaurant registered as a bar. Yet, despite these limitations, the Google Places API is accurate enough to be the backbone of the world's most popular mapping service (Google Maps) and is used daily by millions of individuals to find the location of businesses. This makes the Google Places API data an imperfect, yet attractive dataset to study the spatial organization of amenities at the intra-city scale.

Finally, we remind the reader that our results should be interpreted in the narrow context of the data from which these results were derived. This is data from an online mapping service and for U.S. cities only. The question of whether the results presented below can be generalized to other locations, and also, of whether these results hold for other datasets, is beyond the scope of this paper.

Results

To identify what amenities co-locate in each neighborhood we begin by introducing a method to identify neighborhood scale agglomerations and the amenities that are present in each of them.

Our clustering procedure begins by calculating the *effective number of amenities* A_i that are present in each location i. We define A_i as the number of amenities that can be reached by walking from location i. Formally, the effective number of amenities in location i is the scalar function A_i:

$$A_i = \sum_{j=1}^{N} e^{-\gamma d_{ij}}, \tag{1}$$

where d_{ij} is the distance between amenity i and amenity j, γ is a decay parameter that discounts amenities based on their distance to location i, and N is the total number of amenities in city c. To interpret A it is useful to note that an amenity located where the measurement is taking place (i.e. with d_{ii}=0) contributes one to the effective number of amenities in that location, whereas an amenity at distance $d_{ij}=1/\gamma$—which would imply walking $1/\gamma$ kilometers from amenity i to j—will contribute only $1/e$ to that location's effective number of amenities (A_i). We find that our algorithm finds meaningful neighborhoods when we set γ=16, which implies that the contribution of an amenity to the effective number of amenities of a location roughly halves every 62.5 meters and becomes negligible at about 500 meters. This is consistent with research showing that the volume of pedestrian traffic becomes negligible after a ten-minute walk.[12,17]

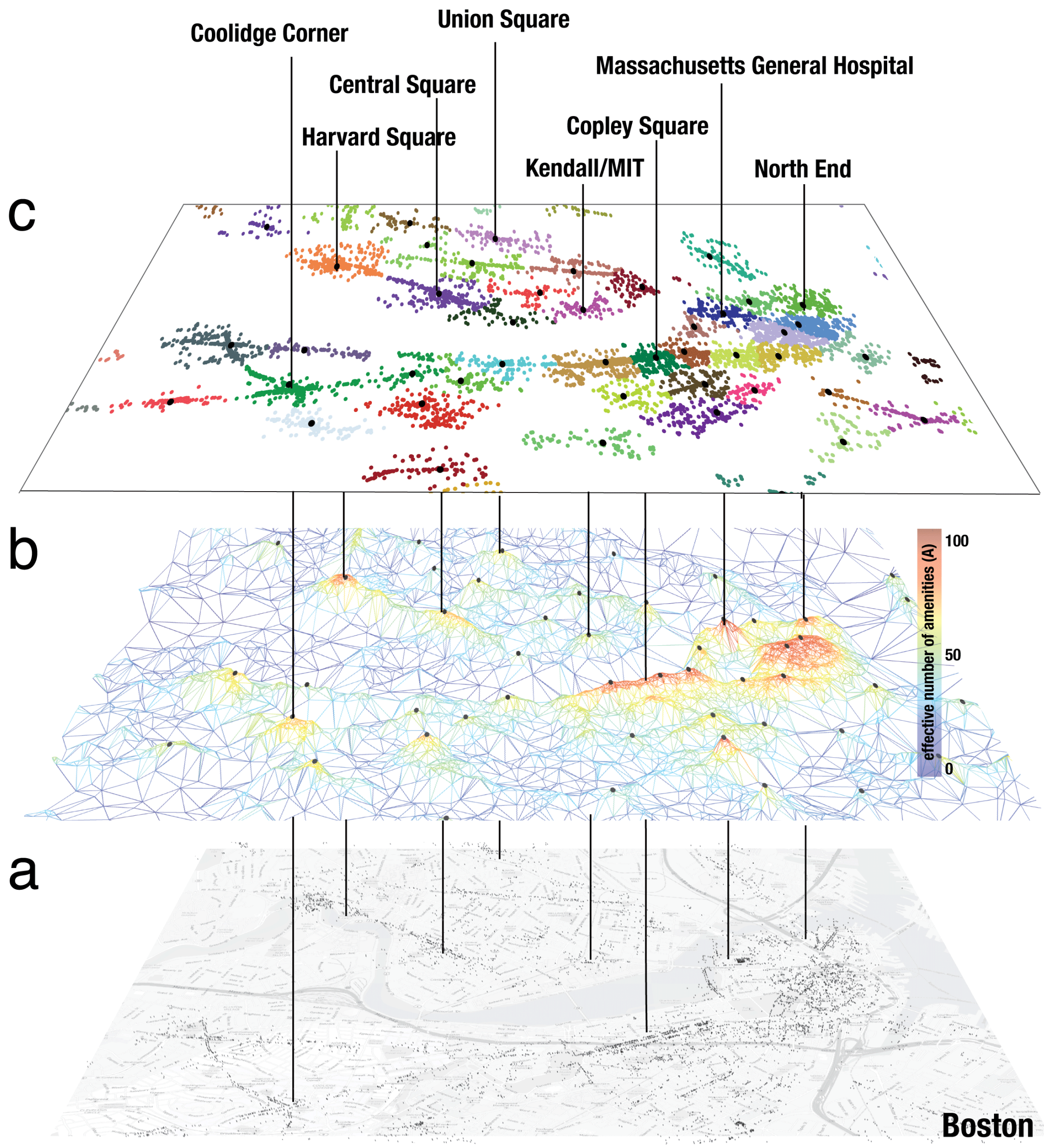

Figure 1: Clustering algorithm. a Map of Boston **b** The number of effective amenities (A) at each location where an amenity is present in Boston. Peaks represent locations with a high number of effective amenities and valleys represent locations with a low number of effective amenities. The black dots in the peak of the hills represent local maxima identified by our clustering algorithm, and are the center of neighborhoods. **C** Neighborhoods identified after the 90% of points with highest A has been assigned to a location using our clustering algorithm. Neighborhoods are shown as sets of dots of the same color. Neighborhood centers are also marked by black dots.

We then use A to identify the amenities belonging to each neighborhood using the following algorithm. First, we remove the 10% of amenities that have the lowest value of A, which represent amenities not located in an agglomeration. For the remaining 90% of amenities we identify local peaks on the landscape defined by A (Fig 1b) by searching for locations that have an effective number of amenities that is larger than their n nearest neighbors using the functional heuristic ($n_i = 3A_i + 50$). This heuristic helps avoid identifying multiple peaks in locations with a large concentration of amenities. Then, we assign amenities to each of the identified peaks using the following greedy algorithm: (i) we initialize each neighborhood by assigning to each peak all amenities that are in close proximity to it (less than 500 meters). Then, (ii) we calculate the distance between each amenity that has not been assigned to a neighborhood and all amenities that have been assigned to a neighborhood. Then, (iii) we assign to a neighborhood only the amenity that is closest to an amenity that has already been assigned to a neighborhood. Finally (iv), we recalculate the distance between assigned and unassigned amenities (repeat step (ii)) and assign one new amenity to a neighborhood by repeating step (iii). We continue until all amenities have been assigned to a neighborhood. An example of the neighborhoods found for the city of Boston is shown in Figure 1c (see SM for New York and San Francisco).

The neighborhoods identified using our algorithm (Fig 1c) correspond to well-known centers of urban activity. In the case of Boston these neighborhoods include Harvard Square and Central Square in Cambridge, and The North End and Coolidge Corner in Boston, among others.

We also note that the distribution of the effective number of amenities in a city is characterized by some universal properties. Figure 2a shows the distribution of the effective number of amenities (A) for every city in our dataset while Figure 2b shows the same distribution after normalizing the effective number of amenities in a city by that city's average effective number of amenities ($\langle A \rangle = \sum_i A_i / N$). For comparison, we show the same distributions but for an ensemble of cities where the location of each amenity has been randomized. These randomized cities are characterized by a narrow distribution, meaning that these random cities lack the high concentrations of amenities that signal the presence of neighborhood scale agglomerations in real cities. More importantly, figure 2b shows that once we normalize the effective number of amenities in a city by that city's average all cities follow the same lognormal distribution

$$P(\frac{A_i}{\langle A \rangle} = x) = \ln N\ (\mu, \sigma), \tag{2}$$

with μ =-0.404 and σ=0.89. The existence of a universal distribution for the effective number of amenities across all cities in our sample means that all of these cities have an equal number of peaks and valleys of a given magnitude when the magnitude of these peaks and valleys is measured in units of that city's average.

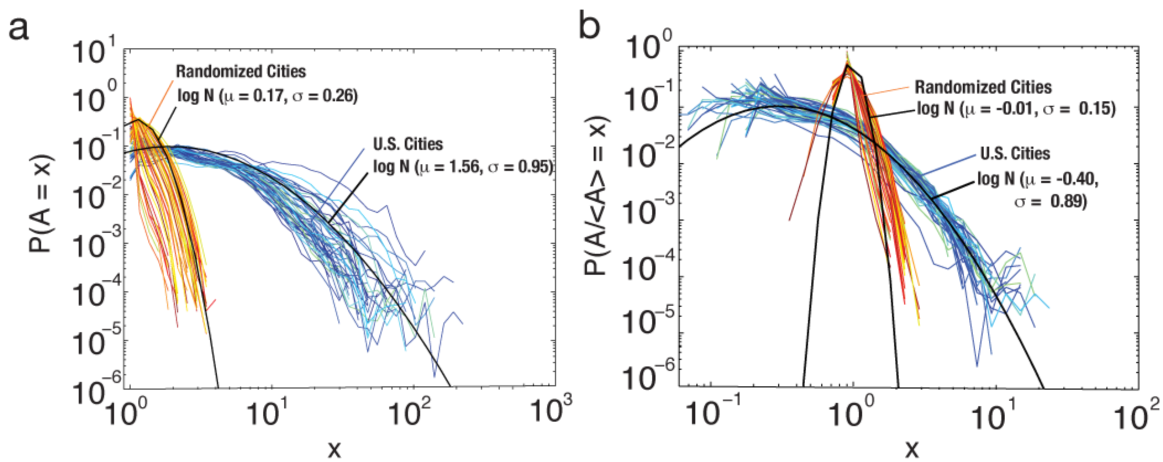

Figure 2: City micro-agglomerations. a The distribution of the effective number of amenities (A) in each US city. Blue lines show the distribution observed in our urban amenities data and orange lines show the distribution observed after randomizing the location of amenities for each city. **b** The distribution of the effective number of amenities (A_i) in each US city normalized by the average effective number of amenities in that city. Blue lines show the distribution observed in the cities data and orange lines show the distribution observed in the same cities but after randomizing the location of amenities

The Amenity Space

After having identified neighborhoods for the 47 cities in our data we map the network connecting pairs of amenities that are likely to co-locate in the same neighborhood. We construct this network of amenities, or Amenity Space, by using spearman's rank correlation to identify pairs of amenities that are likely to be present in the same neighborhood. Figure 3a shows a visualization of this network containing the network's Maximum Spanning Tree[13] and the links that have a pairwise correlation equal or larger than 0.3 (see SM for the full correlations matrix). This subset of links provides a visualization that avoids visual clutter and that reveals what amenities tend to collocate with others. For example, car repair shops collocate with car dealers (Spearman's ρ=0.45), just like religious centers collocate with schools (Spearman's ρ=0.46). What is more important, however, is that this network tell us what combinations of amenities should predict the presence of others, providing a mean to recommend the amenities that should locate in the same neighborhood based on that neighborhood's current pattern of specialization. For instance, the network tells us that neighborhoods that specialize in beauty salons, accountants, and dentist, should also specialize in real state agents, but not in convenience stores or car rentals.

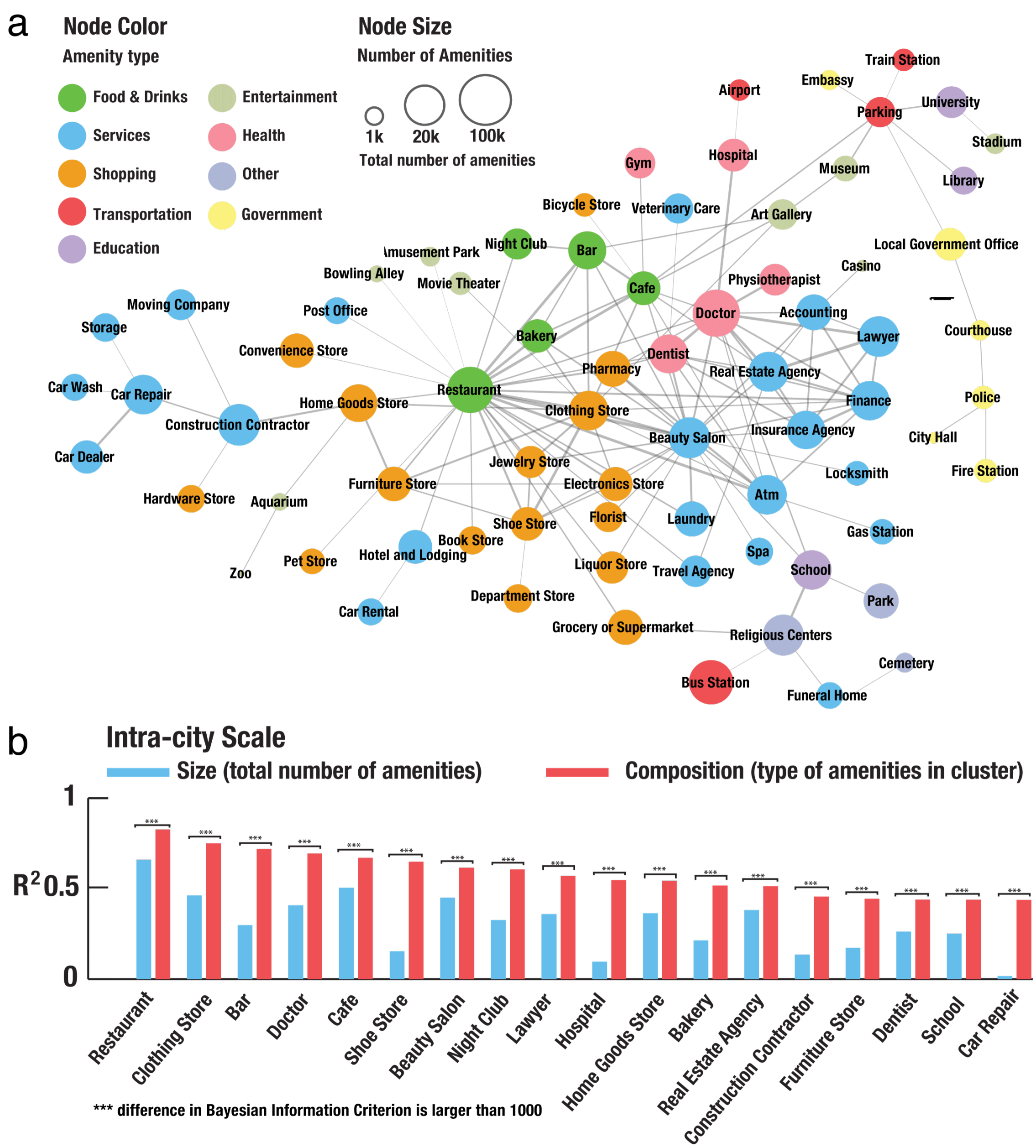

Figure 3: a Network of amenity co-locations. The nodes in the network represent different types of amenities and the edges connect amenities that are likely to collocate in the same neighborhood (see SM). The width of the edges connecting a pair of nodes is proportional to the spearman correlation obtained from the collocation of the two types of amenities across all neighborhoods. The size of a node is proportional to the number of times that an amenity is present in our data set. The color of each node represents the category that the amenity belongs to. **b** Comparison of the accuracy of two models used to predict the total number of amenities of each type on a neighborhood. The light-blue bars show the R^2 of a model predicting the number of amenities of each type in a neighborhood using only the total number of amenities in that neighborhood. The red bars show the R^2 of a model using information on the number of amenities of other types that are present in a neighborhood.

We then use the Amenity Space to build a parsimonious recommendation algorithm[18-19] for each type of amenity. We build this recommendation algorithm using multivariate regression and a forward selection algorithm that iteratively includes new types of amenities to the regression until the contribution of a new type of amenity is statistically insignificant (characterized by a *p*-value of more than 0.001 (see SM)). In addition, we control for over-fitting by using both Akaike's Information Criterion (AIC) and Bayes's Information Criterion (BIC). To provide a benchmark for the accuracy of the model we also predict the number of amenities of each type in a neighborhood using only the total number of amenities in that neighborhood (as a measure of the size of the economy of that neighborhood).

Figure 3b compares the R^2 of the models constructed using the patterns of specialization of neighborhoods with the models using only their size (the neighborhood's total number of amenities). In most cases (66/74=89%), the BIC test chooses the regression using the model based on the pattern of specialization over the regression using the neighborhood size (the exceptions are airports, aquariums, bus stations, car rentals, casinos, convenience stores, gas stations, and zoos), indicating that the model based on a neighborhood's pattern of specialization and the amenity space is better at predicting which amenity should locate in each neighborhood. Also, we note that the differences between the two models are not just statistically significant, but characterized by strong size effects. On average, for the 66 amenity types in which the amenity space model works better, the R^2 of the amenity space model is twice that of the model using size only (R^2=17% on average using size vs. R^2=35% on average using composition). This means that the increase in predictive power obtained by considering the types of amenities that locate in a neighborhood is not only statistically significant, but also substantial.

Finally, we illustrate our predictive algorithm by showing the recommendations it makes for specific amenities and neighborhoods for the city of Boston. Here, we recommend amenities by looking at the difference between the number of amenities observed in a neighborhood and those predicted by the model.

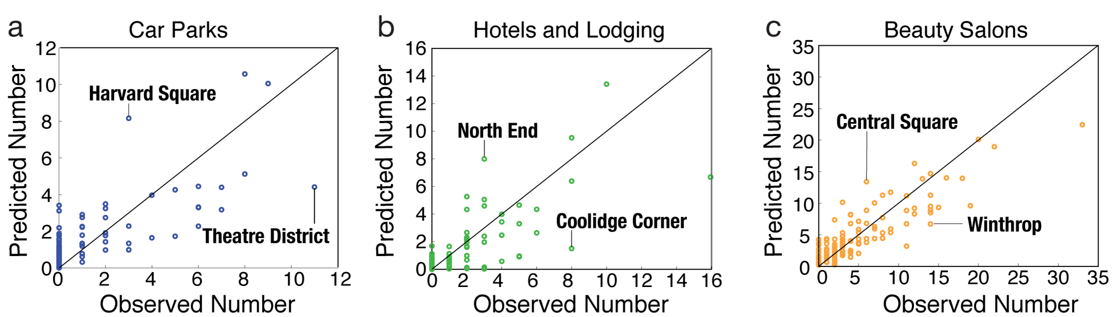

Figure 5: Prediction of amenities in Boston's neighborhoods. a Observed vs. predicted number of car parks, **b** hotels, and **c**, beauty salons for each neighborhood in Boston. Points above the lines represent neighborhoods where the predicted number of amenities is higher than the observed, suggesting instances of under supply. Points below the lines represent neighborhood where the predicted number of amenities is lower than the observed, suggesting instances of excess supply.

Figure 5 a-c compare the number of car parks, hotels, and beauty salons, observed and predicted, for each neighborhood in Boston. Points above the line, such as

Harvard Square in the case of car parks (Figure 5a), the North End for hotels (Figure 5b), and Central Square for Beauty Salons (Figure 5c), indicate amenities that are under expressed in a location (given that location's current pattern of specialization). Points below the lines such as Boston's Theatre District in car parks, Coolidge Corner in hotels, and Winthrop in beauty salons, suggest instances of excess supply. Of course, the recommendations of the model should be taken with care. For instance, our model recommends more car parks in Harvard square, but of course, a decision to build new parking there should consider other aspects of Harvard Square that are not included in our model, such as the aesthetics of its architecture[20-21], or the externalities caused by cars. The lesson here is that the model successfully detects a known reality of Harvard square, which is that there is limited parking. Figure 5b shows another example in which our model detects a lack of hotels in the North End, a well-known tourist spot in Boston where only a handful of hotels are present. This could mean that there is a great potential for new hotels to locate in Boston's North End, but once again, that's a decision that would require additional considerations.

Discussion

In this paper we introduced the Amenity Space, a network summarizing the patterns of co-locations characterizing neighborhood scale agglomerations, and used it to create a recommendation algorithm that we can use to evaluate the composition of amenities in a neighborhood.

Of course, our results and models are not free of biases and limitations. Beyond the data biases described above, our model is limited by its simplicity, which bounds the total amount of variance in the presence of amenities that we can explain. Our statistical model is based on linear regression techniques that could be potentially improved by using more complex functional forms, but also, by adding to them information that is not expressed in the presence of amenities, such as information on population density, the aesthetic appeal of a neighborhood's architecture[20-21], its foot traffic, and the daily and seasonal variations in traffic captured by mobile phone data[22]. Also, our model does not take into account zoning laws that can restrict the type of amenities that locate in each neighborhood.

Still, the results and methods presented here point to interesting new avenues of research. For example, time resolved data sources for both amenities and streetscapes could be used to explore the interaction between the dynamics of the amenities that locate in a neighborhood and the aesthetic of the buildings being constructed in it. Also, these results could help inform what types of business permits or incentives need to be given out to help balance a city's neighborhoods. On the computational side, the information uncovered here could also be used to create interactive online resources that would deliver the recommendations uncovered by our algorithm or similar algorithms. Together, our results, and the new avenue of research they open, should help stimulate the quantitative study of cities at the intra-city scale.

Acknowledgements:

We acknowledge the support of The MIT Media Lab consortia and of Google's Living Labs award. We also acknowledge comments from Luis Bettencourt, Andres Sevtsuk, Scott Kominers, Mia Petkova, Luis Valenzuela, Matias Garretón, and Mark Hubberty.

References

[1] Alfred, Marshall. "Principles of economics." (1895).

[2] von Thünen, Johann Heinrich. *Isolated state: an English edition of Der isolierte Staat*. Pergamon Press, 1966.

[3] Hotelling, Harold. "Stability in Competition." *The Economic Journal* 39.153:41-57 (1929).

[4] Christaller, Walter. *Central places in southern Germany*. Prentice-Hall, (1966).

[5] Lösch, August. "The Economics of location." New Haven, CT: Yale University Press (1954).

[6] Eaton, B. Curtis, and Richard G. Lipsey. "An economic theory of central places." *The Economic Journal* 56-72 (1982).

[7] Mulligan, Gordon F. "Agglomeration and central place theory: a review of the literature." *International Regional Science Review* 9.1:1-42. (1984)

[8] Krugman, Paul. "On the number and location of cities." *European Economic Review* 37.2: 293-298. (1993)

[9] Krugman, Paul. "What's new about the new economic geography?." *Oxford review of economic policy* 14.2: 7-17 (1998).

[10] Fujita, Masahisa, and Paul Krugman. "The new economic geography: Past, present and the future." Papers in regional science 83.1 (2004): 139-164.

[11] Nelson, Richard Lawrence, and Alan Curtis. *The selection of retail locations*. New York: FW Dodge Corporation, (1958).

[12] Sevtsuk, Andres. "Location and Agglomeration The Distribution of Retail and Food Businesses in Dense Urban Environments." *Journal of Planning Education and Research* 34.4: 374-393. (2014)

[13] Hidalgo, César A., et al. "The product space conditions the development of nations." *Science* 317.5837 (2007): 482-487.

[14] Hausmann, Ricardo, and César A. Hidalgo. *The atlas of economic complexity: Mapping paths to prosperity*. MIT Press, (2014).

[15] Boschma, Ron, and Koen Frenken. "The emerging empirics of evolutionary economic geography." *Journal of Economic Geography* 11.2 (2011): 295-307.

[16] Neffke, Frank, Martin Henning, and Ron Boschma. "How do regions diversify over time? Industry relatedness and the development of new growth paths in regions." *Economic Geography* 87.3 (2011): 237-265.

[17] Handy, Susan L., and Debbie A. Niemeier. "Measuring accessibility: an exploration of issues and alternatives." *Environment and planning A* 29.7 (1997): 1175-1194.

[18] Maes, Pattie. "Agents that reduce work and information overload." Communications of the ACM 37.7 (1994): 30-40.

[19] Resnick, Paul, and Hal R. Varian. "Recommender systems." Communications of the ACM 40.3 (1997): 56-58.

[20] Salesses, Philip, Katja Schechtner, and César A. Hidalgo. "The collaborative image of the city: mapping the inequality of urban perception." PloS one 8.7 (2013): e68400.

[21] Naik, Nikhil, et al. "Streetscore--Predicting the Perceived Safety of One Million Streetscapes." Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, (2014).

[22] Gonzalez, Marta C., César A. Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns." Nature 453.7196 (2008): 779-782.

SUPPLEMENTARY MATERIAL:
The Amenity Space and the development of neighborhoods

Elisa E. Castañer[1], César A. Hidalgo[1,2]

I. DATA

A. Amenities Data: We collect the data from the Google Places API containing the latitude, longitude and type (cafe, restaurant, library, etc) of the urban amenities in 47 US cities. The original data set contains 95 different types of amenities. However, we merge amenities that have the same functionality (Table 1) and exclude amenities that are scarce or ambiguous to obtain a data set composed of 74 different types of amenities. The amenities we exclude are: taxi stand, campground, store, subway station, RV park, movie rental, and shopping mall. The resulting amenities are shown in Table 2.

Original Amenities	New Amenities
Hindu temple Mosque Place of worship Synagogue Church	Religious center
Meal delivery Meal takeaway Food Restaurant	Restaurant
Health Doctor	Doctor
Finance Bank	Finance
Roofing Contractor Electrician Plumber	Construction contractor

Painter General Contractor	

Table 1: The left column shows the amenities that were merged into a new amenity type, shown in the right column.

Amenity	Points	Amenity	Points	Amenity	Points
Accounting	17280	Dentist	26071	Movie Theater	1232
Airport	1535	Department store	3515	Moving Company	12744
Amusement park	1017	Doctor	153772	Museum	2161
Aquarium	492	Electronics store	11876	Night Club	5675
Art gallery	5358	Embassy	688	Park	25723
ATM	30753	Finance	32221	Parking	5527
Bakery	9255	Fire station	2050	Pet Store	2270
Bar	21506	Florist	5102	Pharmacy	15204
Beauty salon	41851	Funeral home	2761	Physiotherapist	7929
Bicycle store	1409	Furniture store	12379	Police	1613
Book store	3417	Gas station	2552	Post Office	2723
Bowling alley	366	Grocery or supermarket	15206	Real Estate Agency	39484
Bus station	110642	Gym	5934	Religious Centers	58468
Cafe	9485	Hardware store	4595	Restaurant	112430
Car dealer	11603	Home goods store	29537	School	46516
Car rental	2968	Hospital	7942	Shoe Store	8612
Car repair	40215	Hotel and lodging	11452	Spa	2843
Car wash	3202	Insurance agency	27866	Stadium	1245
Casino	172	Jewelry store	6751	Storage	5849
Cemetery	2386	Laundry	14391	Train Station	1262
City hall	140	Lawyer	37611	Travel Agency	7394
Clothing store	29806	Library	3466	University	6597
Construction contractor	86044	Liquor store	7948	Veterinary Care	5373
Convenience store	13818	Local Government Office	10081	Zoo	114
Courthouse	717	Locksmith	2182	**Total**	1262374

Table 2: Total number of each type of amenity in the Google Places data set in the 47 US cities in our study.

II. CLUSTERING

A. Effective number of amenities

We begin our clustering procedure by calculating the effective number of amenities at each location. The effective number of amenities, A_i, in a location i represents the number of amenities that can be reached by walking from that location. We define A_i as:

$$A_i = \sum_{j=1}^{N_c} e^{-\gamma d_{ij}} = \sum_{j=1}^{k} e^{-\gamma d_{ij}} + \sum_{j=k+1}^{N_c} e^{-\gamma d_{ij}} \qquad (1)$$

$$= \sum_{j=1}^{k} e^{-\gamma d_{ij}} + \varepsilon$$

where d_{ij} is the distance (in km) between amenity i and amenity j, and N_c is the total number of amenities in a city c. γ is a decay parameter that discounts amenities based on their distance to location i,. We set γ=16, meaning that the contribution of an amenity to the effective number of amenities at a location roughly halves every 62.5 meters and becomes negligible at about 500 meters. Moreover, k determines the number of amenities that we use, instead of N_c, to calculate the effective number of amenities, I, at a location. Theoretically all of the amenities in a city should contribute to the effective number of amenities at a location in the city. However, because amenities that are far from a location i have an insignificant contribution to the effective number of amenities at that location i, we only calculate the contribution of the k closest amenities to each location. This yields an error ε in the effective number of amenities. We set k=2000, which is a large enough so that the effective number of amenities at a location always converges before summing the k^{th} amenity.

B. Identifying cluster centers

We continue our clustering procedure by identifying the center of each neighborhood as the local peaks on the landscape defined by A. We identify local peaks by searching for locations that have an effective number of amenities, A_i, larger than their n_i nearest neighbors. We define n_i as: $n_i = 3I_i + 50$, i.e. a function of the effective number of amenities at location i, so that the centers of very dense neighborhoods are required to have a large A_i to be considered a peak. By setting n_i proportional to I_i we avoid assigning multiple neighborhood centers to areas with high density of amenities, and we avoid not assigning any neighborhood center to areas with a low density of amenities.

C. Assigning points to clusters

Finally, we assign points to a neighborhood using the peaks we obtained. First, we remove the 10% of the points in each city with the lowest effective number of amenities, to eliminate isolated amenities that are not part of an agglomeration. After that, we assign all amenities that are within a distance of 0.5km of a neighborhood center to that neighborhood. Then, we calculate the distance from each unassigned point to each assigned point using the following algorithm:

1. Choose the unassigned point, u, which is closest to an assigned point, a.
2. Assign point u to the neighborhood point a belongs to.
3. Calculate the distance from each unassigned point to the newly assigned point u.

The algorithm finalizes once all points have been assigned to a neighborhood. Figure 2 shows the effective number of amenities in the cities of Boston, San Francisco, and

New York (left figures), and the corresponding assignments of amenities to neighborhoods (right figures).

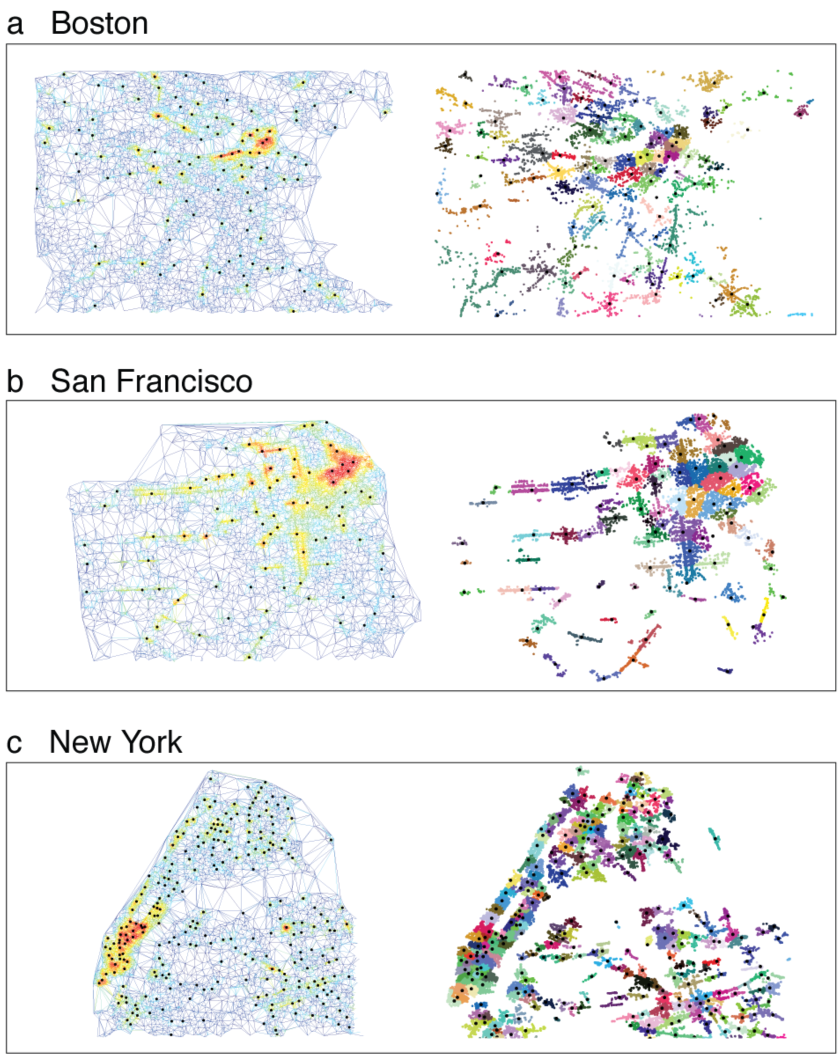

Figure 1: Clustering algorithm. The figures on the right show the effective number of amenities at each location in the cities of **a** Boston, **b** San Francisco, and **c** New York. Red lines correspond to areas with a high effective number of amenities and blue lines correspond to areas with a low effective number of amenities. The black dots represent the locations we assign as neighborhood centers. The figures on the left show the corresponding assignment of amenities to neighborhoods. Each dot represents an amenity, and sets of dots of the same color constitute a neighborhood.

IV. COLLOCATION OF AMENITIES

To study the collocation patterns of amenities, we calculate the spearman correlation between all pairs of amenities across neighborhoods. We show the resulting correlations in the form of a network, where nodes represent amenity types and edges connect amenities that are highly correlated across neighborhoods. To construct this network we first create a Maximum Spanning Tree (MST) of the network and then add edges only between amenities that have a pairwise correlation equal or larger than 0.3.

Here, we show the values of all spearman correlations between amenities across neighborhoods in the form of a matrix (Figure 2). We cluster amenities using Ward linkages.

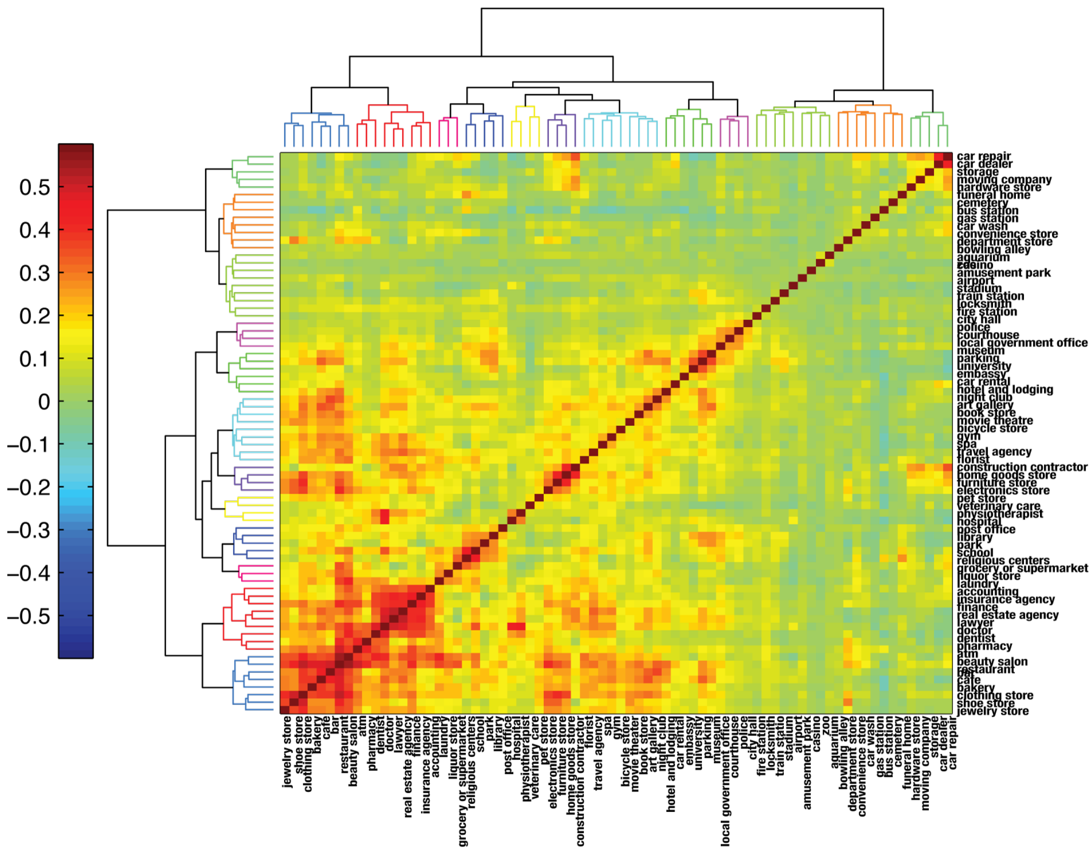

Figure 2: Amenities correlations. Matrix showing the Spearman correlation between each pair of amenities. Amenities are clustered using Ward linkages.

V. PREDICTIONS

We develop four models to predict each type of amenity in the inter city and intra-city scale using two different metrics. In the inter city scale, we create a model (a linear regression) that uses the total number of amenities in a city to predict the number of each type of amenity in that city, and a model (regression using forward selection with p-enter value of 0.001), that uses the composition of amenities in a city to predict the number of each type of each amenity in that city. In the intra-city scale, we create a model (linear regression) that uses the total number of amenities in a micro-cluster to predict the number of each type of amenity in that micro-cluster, and a model (regression using forward selection with p-enter value of 0.001), that uses the composition of amenities in a micro-cluster to predict the number of each amenity in that micro-cluster. Table 5 shows the R^2 obtained for each of these models.

	Inter-City Model		**Intra-City Model**	
	Size	**Composition**	**Size**	**Composition**
Accounting	0.946	0.985	0.291	0.448
Airport	0.575	0.816	0.016	0.114
Amusement Park	0.382	0.724	0.002	0.005
Aquarium	0.709	0.880	0.014	0.028
Art Gallery	0.603	0.930	0.114	0.271
ATM	0.911	0.967	0.320	0.465
Bakery	0.777	0.980	0.364	0.543
Bar	0.649	0.966	0.462	0.750
Beauty Salon	0.952	0.989	0.449	0.615
Bicycle Store	0.594	0.919	0.080	0.183
Book Store	0.878	0.980	0.245	0.344
Bowling Alley	0.478	0.702	0.004	0.014
Bus Station	0.242	0.431	0.023	0.237
Cafe	0.649	0.956	0.505	0.670
Car Dealer	0.608	0.850	0.003	0.231
Car Rental	0.831	0.942	0.042	0.118
Car Repair	0.867	0.976	0.016	0.437
Car Wash	0.828	0.970	0.005	0.071
Casino	0.016	0.000	0.002	0.008
Cemetery	0.126	0.585	0.001	0.015
City Hall	0.379	0.449	0.031	0.151
Clothing Store	0.884	0.993	0.298	0.718
Construction Contractor	0.824	0.978	0.135	0.456
Convenience Store	0.629	0.928	0.042	0.134
Courthouse	0.676	0.738	0.088	0.446
Dentist	0.954	0.974	0.262	0.439
Department Store	0.673	0.945	0.016	0.200
Doctor	0.957	0.986	0.408	0.694
Electronics Store	0.924	0.966	0.224	0.355
Embassy	0.102	0.419	0.046	0.114
Finance	0.953	0.983	0.424	0.610
Fire Station	0.490	0.632	0.018	0.058
Florist	0.889	0.981	0.207	0.259
Funeral Home	0.476	0.787	0.018	0.146
Furniture Store	0.912	0.980	0.173	0.444
Gas Station	0.443	0.777	0.000	0.028
Grocery or Supermarket	0.791	0.955	0.116	0.377
Gym	0.911	0.984	0.229	0.339

Hardware Store	0.896	0.953	0.020	0.194
Home Goods Store	0.908	0.986	0.213	0.517
Hospital	0.958	0.979	0.096	0.546
Hotel and Lodging	0.795	0.824	0.250	0.435
Insurance_agency	0.825	0.981	0.234	0.433
Jewelry Store	0.902	0.978	0.208	0.352
Laundry	0.933	0.984	0.180	0.354
Lawyer	0.871	0.894	0.359	0.570
Library	0.610	0.937	0.180	0.416
Liquor Store	0.753	0.815	0.175	0.301
Local Government Office	0.901	0.937	0.181	0.567
Locksmith	0.671	0.752	0.033	0.053
Movie_theater	0.780	0.952	0.125	0.190
Moving Company	0.721	0.931	0.012	0.131
Museum	0.499	0.951	0.221	0.412
Night Club	0.735	0.957	0.326	0.606
Park	0.669	0.745	0.149	0.320
Parking	0.666	0.938	0.374	0.610
Pet Store	0.812	0.943	0.077	0.192
Pharmacy	0.878	0.949	0.169	0.371
Physiotherapist	0.863	0.931	0.081	0.260
Police	0.681	0.866	0.052	0.201
Post Office	0.859	0.964	0.090	0.130
Real Estate Agency	0.835	0.952	0.381	0.513
Religious Centers	0.744	0.868	0.171	0.430
Restaurant	0.921	0.995	0.659	0.826
School	0.948	0.976	0.251	0.438
Shoe Store	0.916	0.966	0.153	0.648
Spa	0.784	0.940	0.182	0.297
Stadium	0.613	0.749	0.010	0.107
Storage	0.632	0.912	0.010	0.123
Train Station	0.099	0.414	0.047	0.087
Travel Agency	0.813	0.931	0.292	0.402
University	0.238	0.351	0.020	0.328
Veterinary Care	0.814	0.966	0.020	0.115
Zoo	0.343	0.680	0.001	0.011

Table 5: R^2 of each of the models for each amenity.

Given that these four models use a different number of samples and parameters, we calculate the Akaike Information Cirterion (AIC) and Bayesian Information Criterion (BIC) of each of the models. These criteria allow us to differentiate the models: the lower the AIC and BIC values, the more desirable the model (better fit and less overfitted). The AIC and BIC values obtained for each model are summarized in Table 6.

	Inter-City Scale	**Intra-City Scale**

	Size		Comp.		Size		Comp.	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
Accounting	387.1	389.0	610.2	615.7	7233.6	7240.7	5467.9	5630.0
Airport	283.6	285.4	534.2	536.0	-14564.4	-14557.4	-14565.1	-14480.5
Amusement Park	252.4	254.3	492.8	496.5	-6923.4	-6916.4	-6940.3	-6933.2
Aquarium	160.8	162.6	395.3	399.0	-24141.2	-24134.2	-23744.2	-23709.0
Art Gallery	404.3	406.2	600.4	605.9	15448.2	15455.3	13780.7	13893.4
Atm	458.2	460.1	614.1	617.8	14260.7	14267.7	12446.0	12664.5
Bakery	437.6	439.5	479.8	483.5	2507.6	2514.7	-349.5	-208.5
Bar	508.4	510.2	724.1	731.5	20416.0	20423.0	14125.5	14358.1
Beauty Salon	471.4	473.3	625.0	628.7	19820.0	19827.0	16895.3	17113.8
Bicycle Store	268.8	270.6	283.5	285.4	-16203.0	-16196.0	-17083.1	-16970.3
Book Store	278.1	280.0	510.0	517.4	-7584.9	-7577.8	-8461.1	-8313.1
Bowling Alley	132.9	134.8	414.8	418.5	-28639.3	-28632.3	-28784.8	-28756.6
Bus Station	667.5	669.3	735.0	736.9	34335.6	34342.7	34766.9	34936.1
Cafe	443.1	445.0	461.9	465.6	4533.6	4540.6	1134.4	1338.8
Car Dealer	461.3	463.1	714.4	716.2	10190.0	10197.1	8802.5	8873.0
Car Rental	290.1	291.9	443.9	449.4	-3146.7	-3139.7	-3181.3	-3110.8
Car Repair	521.0	522.8	731.4	738.8	22908.0	22915.1	18230.6	18371.6
Car Wash	293.8	295.6	460.0	467.4	-11654.7	-11647.6	-11747.7	-11663.2
Casino	216.3	218.2	443.5	443.5	-35421.5	-35414.5	-35127.2	-35113.1
Cemetery	341.0	342.9	586.3	590.0	-21285.1	-21278.0	-21423.3	-21402.2
City Hall	76.2	78.0	293.2	295.0	-37437.7	-37430.7	-38618.9	-38562.5
Clothing Store	500.4	502.2	592.9	600.3	31184.7	31191.8	23911.6	24024.4
Construction Contractor	591.7	593.6	574.4	578.1	23556.0	23563.1	20067.1	20243.3
Convenience Store	455.2	457.1	570.9	572.8	1726.9	1733.9	2246.0	2394.0
Courthouse	160.9	162.8	404.9	406.8	-13810.3	-13803.3	-17901.4	-17788.6
Dentist	436.7	438.6	714.8	718.5	19519.7	19526.7	17163.9	17311.9
Department Store	327.4	329.2	510.3	514.0	-6448.4	-6441.4	-7219.3	-7064.3
Doctor	578.7	580.5	607.2	614.6	42180.3	42187.3	36517.8	36672.9
Electronics Store	386.1	388.0	587.0	590.7	3688.0	3695.0	2189.0	2322.9
Embassy	329.3	331.1	632.8	634.7	-2578.3	-2571.2	-3268.8	-3205.4
Finance	440.6	442.5	615.9	623.3	20231.0	20238.1	16860.1	17036.3
Fire Station	293.9	295.8	600.6	602.4	-17404.5	-17397.5	-17771.8	-17722.5
Florist	332.2	334.1	524.5	530.0	-4705.5	-4698.5	-5303.1	-5197.4
Funeral Home	336.3	338.1	568.3	570.1	-10028.7	-10021.7	-10844.2	-10703.3
Furniture Store	389.5	391.4	441.7	447.2	10673.1	10680.1	7599.0	7697.7
Gas Station	333.2	335.0	543.6	545.5	-13926.6	-13919.5	-11923.4	-11867.0
Grocery or Supermarket	467.7	469.6	592.6	596.3	9280.9	9288.0	6500.8	6677.1
Gym	321.2	323.1	463.7	469.2	-2721.9	-2714.9	-4013.8	-3837.6
Hardware Store	299.0	300.9	516.4	522.0	-8239.8	-8232.8	-9960.1	-9854.4
Home Goods Store	469.2	471.0	645.5	651.0	17169.9	17177.0	13170.9	13290.7
Hospital	310.3	312.1	428.4	433.9	11386.1	11393.2	5907.4	6027.2
Hotel and Lodging	413.6	415.5	734.4	736.3	12585.9	12592.9	10292.7	10483.0
Insurance Agencyncy	496.8	498.7	692.3	697.9	14861.7	14868.7	12397.0	12538.0
Jewelry Store	353.3	355.1	444.2	447.9	14860.2	14867.3	13143.0	13269.9
Laundry	400.6	402.4	566.7	570.4	4144.1	4151.2	2146.9	2316.0

Lawyer	479.9	481.8	728.9	730.8	38846.6	38853.6	35662.3	35831.4
Library	343.4	345.3	429.1	432.8	-5993.1	-5986.1	-8949.8	-8808.9
Liquor Store	405.4	407.2	632.6	634.4	-1736.3	-1729.2	-2355.6	-2242.8
Local Government Office	331.1	332.9	481.5	487.0	12849.6	12856.6	7505.0	7638.9
Locksmith	309.2	311.0	542.0	543.9	-14495.9	-14488.8	-14640.9	-14591.5
Movie_theater	223.5	225.4	352.4	356.1	-16822.2	-16815.1	-17422.3	-17337.7
Moving Company	443.6	445.5	628.4	632.1	300.1	307.2	-457.0	-372.4
Museum	318.3	320.1	385.0	392.4	-4793.4	-4786.3	-6985.0	-6872.2
Night Club	377.0	378.9	545.3	550.8	6321.7	6328.7	1774.4	1922.5
Park	504.8	506.7	683.4	685.3	11027.2	11034.2	10194.1	10363.3
Parking	372.0	373.9	596.1	599.8	5373.6	5380.6	1963.1	2153.4
Pet Store	285.8	287.6	452.0	455.7	-13001.7	-12994.6	-14005.4	-13885.6
Pharmacy	429.4	431.2	643.7	647.4	7366.7	7373.7	5035.9	5176.9
Physiotherapist	353.1	355.0	667.6	673.1	791.0	798.0	-544.4	-459.9
Police	252.8	254.6	349.2	352.9	-15255.4	-15248.4	-16701.6	-16602.9
Post Office	269.4	271.3	504.4	508.1	-12492.7	-12485.7	-12755.2	-12670.6
Real Estate Agency	515.7	517.6	668.7	672.4	20820.6	20827.7	19273.1	19449.4
Religious Centers	565.0	566.9	729.1	732.8	24793.2	24800.3	21728.3	21883.4
Restaurant	602.3	604.2	745.4	752.8	32182.1	32189.1	26651.6	26912.4
School	488.4	490.3	741.3	745.0	15330.2	15337.2	13283.1	13445.2
Shoe Store	363.4	365.2	607.3	611.0	18001.5	18008.6	11416.1	11528.9
Spa	307.5	309.4	456.4	460.1	-8683.6	-8676.6	-9852.6	-9732.7
Stadium	225.0	226.8	553.0	554.9	-13695.6	-13688.5	-13931.8	-13875.4
Storage	394.6	396.4	582.3	586.0	-6999.5	-6992.4	-7697.7	-7606.1
Train Station	334.7	336.6	545.1	547.0	-10105.8	-10098.8	-10424.3	-10389.0
Travel Agency	398.7	400.6	545.0	548.7	3926.0	3933.1	2523.5	2671.5
University	403.4	405.3	557.5	559.3	24047.2	24054.3	21500.1	21627.0
Veterinary Care	336.8	338.6	619.2	624.8	-3348.5	-3341.5	-3679.7	-3538.8
Zoo	70.8	72.7	144.4	148.1	-43402.7	-43395.6	-42907.2	-42879.0

Table 6: Akaike Information Criterion (AIC) values and Bayesian information Criterion (BIC) values of each model. The models with smaller AIC and BIC values are preferred.