Delineating Neighborhoods Using Location Choices

Rolando Campusano
University of Toronto

January 31, 2021

Abstract

Research on neighborhoods has relied on administrative definitions that do not coincide with agents' decision problems. This produces a spatial misalignment between administrative and "economic" boundaries that bias research findings and the policies designed around them. I propose a novel methodology to delineate neighborhoods using a machine learning algorithm that groups locations based on revealed preferences. I apply the methodology to Toronto's industrial and residential neighborhoods and show that they are not like each other and that they remarkably differ in size and shape from their administrative counterparts. In particular, economic neighborhoods tend to have an elliptical shape and to locate around major streets. Moreover, neighborhoods are different across industries or property types. These characteristics have implications for the study the effects of neighborhood segregation and concentration.

1 Introduction

A neighborhood is a geographically localized community where its members interact with each other in various ways. Neighborhoods have become an important focus of policy and research, as economists recognize the importance of interactions in shaping behavior and outcomes. Phenomena such as labor-market referrals, crime, disease propagation, gentrification, segregation, social mobility, agglomeration economies, among many others; are all phenomena that all arise from interactions that occur within and spill across neighborhoods.

The standard approach to neighborhoods is to use administrative or legally-defined boundaries. However, this can be a problem as administrative neighborhoods are delineated as a result of an optimization process that does not necessarily align with agents' decision problems. For example, postal codes are defined for the purpose of optimizing mail delivery and census blocks (or tracts) are defined for the purpose of optimizing the process of a census.[1] A persistent or even increasing misalignment between 'legal' and 'economic' boundaries that may lead to measurement and inference biases that compromise research findings and the policies designed around them.

For these motives, I propose a revealed preference approach to delineate economic neighborhood boundaries. In particular, I develop a machine learning algorithm that uses historical geocoded location choices of agents to identify neighborhoods as a collection of *similar-neighboring-choices.* The algorithm starts by using a dataset of historical location choices and assigns them to a geocoded hexagon grid.[2] Each location (a cell in the grid) has a series of geographical and non-geographical characteristics. Then, the algorithm incorporates revealed preferences by computing the raw propensity score associated to a location choice given their characteristics. Finally, the algorithm agglomerates adjacent cells so that so that the resulting neighborhoods minimize the within variance of the propensity score conditional to differences in propensities score among locations that are lower than a given 'threshold'.

The threshold defines where one neighborhood ends and another begins. The use of thresholds is unavoidable for any approach that seeks to discretize a continuous territory (de Bellefon et al., 2019). Intuitively, the threshold affects the size of the neighborhoods. A small threshold will deliver neighborhoods that are composed of more similar cells than a larger threshold. This is desirable as it implies more homogeneous neighborhoods. However, a threshold that is too small results in neighborhoods that are too small to make

[1] School and political districts are another potential source but their delineation is often more obscure and involve many objectives.

[2] In principle any side size is possible and is only constrained to computational power.

inferences about the unobservable differences across them as they will account for most (if not all) of the variance of the propensity score between neighborhoods.[3] Conversely, a threshold that is too big makes neighborhoods that are so big that they account for most of the within and less of the between variance of the propensity score across neighborhoods.

The spatial distribution of location choices has a direct effect on how sensitive the results are to changes in the threshold and how likely is to get to the extreme cases of neighborhoods that are too small or big. Cities with an uneven distribution of economic activity will have a distribution of bilateral differences that is highly skewed, whereas more even cities have a distribution of bilateral differences that is more uniform. This implies that applying the algorithm to uneven cities will result in neighborhoods that are less sensitive to the threshold. Moreover, this observation also implies that even under large thresholds, those neighborhoods are less likely to be 'too' big.

In order to understand better the choice of a threshold, I carry out a set of simulations. I simulate the location choice of 100,000 agents in two cities for 100 thresholds. The first city is uneven, with agents that cluster around certain locations. The second city is even, with agents that locate randomly across space. The results of the simulation confirm the intuition. Depending on the threshold, neighborhoods can be too big at very large thresholds or too small at tiny thresholds. Moreover, the boundaries of the neighborhoods are more stable in the uneven city. This implies that in order to obtain neighborhoods with boundaries that are less sensitive to the threshold, the algorithm needs a context in which location choices are clustered in the space. This stability implies that applying the algorithm to uneven cities does not require an extensive threshold sensitivity exercise, as we can be confident that the threshold will not make a huge difference in the resulting neighborhoods.

The misalignment between 'legal' and 'economic' boundaries produces three issues that the algorithm aims to solve. First, urban areas that are defined too narrowly or too broadly may present the modifiable areal unit problem by, for instance, misstating the extent of urban sprawl or by missing important positive or negative spatial spillovers of urban policy interventions. By identifying neighborhoods using economic decisions, these neighborhoods decrease the relevance of the modifiable area unit problem (MAUP) as these neighborhoods are by construction 'the' unit of analysis at which agents are making choices (Openshaw, 1983; Briant et al., 2010).[4] Moreover, by using the raw propensity score to summarize the correlation between choices and location characteristics, the al-

[3] In the limit, a threshold that is tiny makes the algorithm useless as it stops in the first iteration and only agglomerates cells that have the same propensity score.

[4] We can also think that this algorithm is in essence identifying the spatial choice set agents consider when making location choices.

gorithm incorporates the economic concept that a choice is the result of incorporating trade-offs. This implies that neighborhoods might be different across agents as each type of agent values location characteristics differently.[5]

Second, assumptions of symmetric interactions within neighborhoods are common in research at the neighborhood level (Topa and Zenou, 2015).[6] Symmetric interactions are more likely to hold at smaller geographies as a significant portion of social interactions occur at very close physical distance among agents. However, the smaller the geography the less the number of agents used for estimation, leading to a granularity problem (Dingel and Tintelnot, 2020). By agglomerating cells with similar propensity scores, this algorithm addresses both problems by delivering bigger neighborhoods composed by very similar cells, implying both less granularity and more symmetric interactions.[7]

And third, structural models that use Type 1 distributions to match the spatial distribution of economic activity tend to assume uncorrelated shocks. In the study of neighborhood effects, this assumption usually takes the form of assuming that interactions are not only symmetric within neighborhoods but are also incomplete across them (Gibbons et al., 2015). Because of the misalignment between 'legal' and 'economic' neighborhoods, this is unlikely to hold in 'legal' neighborhoods which implies that identification of the neighborhood effects fails (Gibbons et al., 2015). Without making assumptions about the network structure of interactions between agents within and across neighborhoods,[8] this algorithm delineates neighborhoods that are distinct in propensity score from their immediate neighbors. This implies that there exists zero spatial correlation between the location choices between a neighborhood and their immediate neighbors which makes them more feasible to understand research questions that focus on the (very)local aspects of the propagation of shocks and spillovers.

I present two applications of the algorithm for the Greater Toronto Area and use the

[5]For example, location amenities such as the proximity to a school or to family care are presumably valued more by households than companies, whereas proximity to delivery services or banking services are presumably valued more by companies than households.

[6]This is translated in assumptions about the way one's residential neighborhood affects one's outcomes. In particular, a common assumption takes the form of assuming that there is no geography-level correlation in unobserved attributes among residents, after taking into account the broader reference group.

[7]This is clearly under the assumption that similar propensity scores are associated to a collection of similar agents making similar location choices.

[8]The critique has also arises in the trade literature that use these distributions more frequently following the seminal work by Eaton and Kortum (2002). In particular, Lind and Ramondo (2020) points out that these research has aim to capture the Ricardian insight that different degrees of correlation in productivity lead to heterogeneity in the gains from trade by incorporating sectors, multinational production, or global value chains. However, the way the literature incorporates correlation is still restrictive, potentially removing empirically relevant sources of heterogeneity. Concretely, although multi-sector models create correlation because each sector is present in many countries, productivity is independent across sectors, implying that sectors do not share technologies.

results to compare how 'legal' (in this case postal codes) and 'economic' neighborhoods differ around these issues. The first application uses location choices of all firms with the aim to provide 'industrial' neighborhoods by industry. Location characteristics in this application are given by access to amenities (Banks, Hotels, Touristic Attractions, Parks, etc) and market access measures for upstream, competing and downstream firms. The second application uses real estate transactions with the aim to provide 'residential' neighborhoods by property type. Location characteristics in this application are given by the same measures of access to amenities as in the case of industrial neighborhoods and measures of the characteristics of real estate being sold in a given period of time.

Results show that neighborhoods do not look like postal codes. Moreover, they show that industrial and residential neighborhoods do not look alike either. Industrial neighborhoods tend to be long with a centroid located around main streets, whereas residential neighborhoods are shorter with a centroid located around secondary streets. Separating neighborhoods by industry leads to industry-specific neighborhoods, and shows that neighborhoods also differ across industries. In particular, manufacturing and entertainment neighborhoods are larger on average than retail and services neighborhoods, which can be related to the degree of tradability of the goods and services provided by each of those industries.

Separating residential neighborhoods by property type also lead to neighborhoods that are different. Condo neighborhoods tend to be larger and longer than House neighborhoods, which makes sense in a city like Toronto where condos tend to locate around main streets and houses tend to locate along secondary streets. Results also show that visual inspection and inside knowledge of the city at study is important when evaluating the quality of the results. In particular, visual inspection shows that the results for industrial neighborhoods are better than the ones for housing neighborhoods. The main differences between both applications that might lead to these results is that the distribution of activity for the 'residential' neighborhoods is not as skewed as the case for 'commercial' neighborhoods, which makes sense as real-estate tends to be equally spread across the city.

Turning into the comparisons between 'legal' and 'economic' neighborhoods. First, in order to analyze how these neighborhoods might lead into different conclusions, I analyze the prevalence of a Zipf's law for neighborhoods. In geography, Zipf's is an empirical law about the size distribution of spatial units, in particular cities. Taking it to neighborhoods, if the law holds perfectly, then the neighborhood size distribution can be approximated with a Pareto distribution with shape parameter equal to one. If the shape parameter is equal to one then the power law implies that, within a city, the largest

neighborhood is roughly twice the size of the second largest neighborhood, about three times the size of the third largest neighborhood, and so on. If the shape parameter is greater than one indicates that the size is more evenly distributed across neighborhoods than what the Zipf's law predicts. I perform this exercise for counts (number of firms or number of real estate transactions), and area sizes and find that the distribution of size is different between postal codes and economic neighborhoods signaling that if the method proposed in this paper correctly identifies the choice set then this neighborhoods decreases the relevance of the modifiable area unit problem.

Second, in order to analyze the degree to which the zero-spatial correlation assumption is satisfied, I compute the Moran (1950)'s I index of spatial correlation. This index indicates the correlation between a variable and its surrounding values.[9] A positive Moran's I indicates that similar observations are closer to each other, while a negative value indicates that dissimilar values are closer to each other, and values around zero would indicate that there is no spatial correlation, and instead, observed values are randomly distributed. I perform this exercise for both counts (number of firms or number of real estate transactions) and propensity scores, and find that the algorithm delivers neighborhoods that present significantly lower spatial correlation than that of postal codes for all measures across industries and real estate property types. In particular, and consistent with the intuition of the algorithm, neighborhoods present zero or negative spatial correlation between the number of firms located (or transactions performed) in a neighborhood and their neighbors. In terms of propensity scores, even though neighborhoods present positive spatial correlation this arises mechanically and is still substantially lower than that of postal codes.[10]

This paper is contributing to the intersection of two different literatures: the literature in urban economics that seeks to define spatial units and the literature that brings machine learning methods to economics. The most direct contribution is to the literature that seeks to define spatial units. A long-standing focus in this literature has been to provide a rigorous definition of cities. In fact, this literature has received a renewed interest the last decade due to urbanization concerns about urbanization concerns in developing countries and the availability of new sources of detailed geographical data. The latter allow researchers to not only develop algorithms that enable a better delineation of

[9]This index can be computed globally and locally. The global index summarizes how the whole dataset, or in this case the city of Toronto, is spatially correlated. The local index decomposes the index by neighborhood and summarizes how each neighborhood is spatially correlated to their neighbors.

[10]This result arises mechanically. Propensity scores are calculated using market access measures calculated by aggregating the underlying measure within with a distance decay function. These market access measures smooths out the distribution of the underlying variables which implies a propensity score that is also smoothed across locations.

cities in general, but that also enable countries that do not have detailed data on commuting patterns (to give an example of data commonly used by some developed countries, see Duranton (2015), for a review) to provide consistent delineation of cities within and across countries.[11] These newly available datasets include lights (night and/or day-time) data from satellite data (Baragwanath et al., 2019; Ch et al., 2020), gridded population data (Henderson et al., 2019; Moreno-Monroy et al., 2020) and map buildings (Arribas-Bel et al., 2019; de Bellefon et al., 2019).

This paper contributes to this literature in three fronts. Firstly, this paper focuses on delineating neighborhoods within cities which has been absent in the delineation literature even when considering the increasing impact of neighborhoods in research and policy. The second contribution is introducing a machine learning algorithm that uses geocoded location choices, as indication of revealed preferences for a certain location, to delineate neighborhoods within a city. This algorithm is flexible enough to be adapted to different scenarios and research questions. The third contribution is the enablement of the algorithm flexibility. It introduces the use of points of interest, or real estate transactions data, to understand the conditions -and data requirements- under which this algorithm provides neighborhoods that can be useful for inference about the differences within a city.

More generally, this paper contributes to the increasing literature that brings data-based algorithms and machine learning methods to economics. Though the adoption of these methods in economics has been slower, their use in empirical work has gained momentum during the last decade. The observation made by Mullainathan and Spiess (2017) and Athey and Imbens (2019) that machine learning not only provides new tools, but also solves different problems than current empirical methods, has increased the appeal to the community. This is particularly relevant in this case, as the objective of this paper is not to estimate a parameter for the relationship between location choices and their characteristics, but to find generalizable patterns in the characteristics of location choices that help us to predict the neighborhood at which these choices are located. This observation is what makes these methods great to be applied to new kind of data to attempt to answer traditional questions such as measuring economic activity using light satellite data, and to answer new questions like the one this paper attempts to answer: delineating neighborhoods using geocoded location choices.

These benefits have also been taken to the literature delineating geographical units.

[11] Commuting patterns are indicators of journey-to-work relationship between two areas which generally allows researchers to determine whether workers belong to the same local labor market and, hence, if they can be considered to form part of the same urban area.

Two examples of this recent research are Galdo et al. (2019) and Arribas-Bel et al. (2019). Galdo et al. (2019) use a variety of data sources combined with human judgement to train a small subsample of locations in India, then the trained algorithm is used to predict all the cities in the country. Arribas-Bel et al. (2019) delineate all cities in Spain using a detailed map of all buildings for Spain that are then use to feed a clustering algorithm that draws boundaries around clusters of buildings that reach a minimum density threshold. This paper contributes to this literature by introducing an algorithm that not only uses highly detailed data to make predictions but also incorporates economic theory in its core elements by using location choices as a signal of revealed preferences.

The rest of the paper is organized as follows: Section 2 briefly describes the approaches the literature has used to delineate urban areas, it introduces the algorithm, and discusses its potential benefits relative to other approaches. In an attempt to better understand to which contexts this algorithm can be applied, Section 3 introduces a simulation exercise that studies the sensitivity and stability of the results of the algorithm under two extreme distributions of economic activity. Section 4 provides two applications of the algorithm for the Great Toronto Area and discusses the resulting economic neighborhoods. Section 5 establishes basic facts about the differences between 'economic' and 'legal' neighborhoods, and Section 6 concludes.

2 Methodology

Thinking about neighborhoods as a collection of very similar cells provides a clear reason to why machine learning clustering methods are a good way to identify them. Clustering methods have the goal of grouping a collection of objects into subsets or "clusters," such that those within each cluster are more closely related to each other than to objects assigned to other clusters. These methods group observations based on a definition of similarity (or dissimilarity) provided by the researcher. The most popular clustering method in economics is the K-means algorithm (Hartigan and Wong, 1979) that provides an efficient and computationally light way to find a previously defined number of clusters "K".[12] As pointed out by Athey and Imbens (2019), choosing the number of clusters K is difficult because there is no direct cross-validation method to assess the performance of one value versus the other. Moreover, this method, although modifiable to include

[12]The algorithms defines K centroids elements that are sufficiently spread out over the characteristics space. Given this set of centroids, it then assign each observation to the cluster so that it minimizes the dissimilarity between the unit and the centroid of the cluster. It then updates the centroid and repeats the process until all observations to a cluster.

adjacency constraints might still provide spatially non-compact clusters, which is an important aspect of the economic definition of neighborhood this paper is focusing on.

Instead, motivated by Rozenfeld et al. (2011) bottom-up algorithm to delineate cities, I extend Ward (1963)'s unsupervised agglomerative hierarchical algorithm to include economic theory. As is, the original algorithm has a very important feature for the delineation of economic neighborhoods within a city. This algorithm does not require defining a priori the size or the number of neighborhoods as most clustering methods do (including k-means).[13] In fact, the algorithm builds up all possible neighborhoods based on the dissimilarity in characteristics between groups of location. Once all possible neighborhoods are defined, the algorithm requires the researcher to set a maximum threshold for which a neighborhood is set to be different from another. This threshold is the one that ends up defining the number of neighborhoods that the algorithm delivers.

Even though the algorithm is more flexible than other clustering methods, it comes with an important computational burden. First, the algorithm requires a full bilateral dissimilarity matrix between all geographical cells. The focus of the algorithm is to build-up neighborhoods starting from very small geographies, quickly incresing the size of this matrix as the size of this small geographies decreases. For example, the applications in section (4) are for the Greater Toronto Area that has a total area of about $7,124.15\,km^2$. The whole area is divided by 75-meter hexagon grid cells whose area is about $0.015\,km^2$ each, which implies about $474,943$ cells to cover the entire city.[14] This would imply a dissimilarity matrix of size $474,943 \times 474,943 = 225 \times 10^9$ which would roughly require about $14TB$ of RAM to run if each element of the matrix is a real number.[15] This extensive memory requirement is clearly infeasible even with the current computational power. Second, if we want to apply the algorithm using more than one characteristic, then the memory requirement increases as many times as characteristics we want to cluster with.

To reduce this burden and make the application of this algorithm feasible, I take two features from the conceptual definition of a neighborhood and implement them in the algorithm. First, neighborhoods are spatially compact. Including adjacency constraints reduces the set of cells the algorithm compares with from all possible cells to only those

[13] Please refer to the clustering documentation of the python machine learning library 'scikit-learn' out of which this algorithm is based on. This library has been in continues development since (Pedregosa et al., 2011) was published.

[14] The actual application has even more cells $(740,286)$ as it also includes parts of Lake Ontario and a buffer area so that all points included in the border of the GTA are also included in the exercise.

[15] Actually, the matrix is composed by floats which is a number representation used in computing for real numbers. There are different types that differ by precision (or number of decimals). I chose to use 64bytes floats (double) that use 64 bytes each and provide precision up to 15 decimals.

that are immediately adjacent.[16] This constraint not only ensures spatially compact neighborhoods but it also importantly decreases the memory requirement of running the algorithm. In particular, this modification implies that we can use a sparse square matrix instead of the full square matrix. This matrix is populated with the bilateral dissimilarity between a cell and their immediate neighbors (12 in the case of an hexagon grid) and an arbitrarily large constant number for the rest of the elements.[17] This implies that the number of elements of this dissimilarity matrix becomes $474,943 \times 12 = 5 \times 10^6$ plus the arbitrarily large constant number, implying a memory requirement of about $400MB$ RAM which is manageable in most modern computers.[18]

Second, neighborhoods are the result of agents making location choices. Assuming that all the information about all potential alternative neighborhoods is included in the probability that an agent chooses a give location, the propensity score summarizes how those choices are correlate with location characteristics and decreases the number of characteristics the algorithm is using to define the neighborhoods to just one while at the same time considering how some agents might have different preferences for location. Applying the algorithm to the propensity score is very similar in spirit to taking Rosenbaum and Rubin (1984)'s propensity score stratification to an spatial setting.[19] Within each propensity score stratum (neighborhood in this case), treated and untreated subjects have roughly similar values of the propensity score. Therefore, when the propensity score has been correctly specified, the distribution of measured baseline covariates will be approximately similar between treated and untreated subjects within the same stratum. Stratas are formed by defining a threshold (or caliper) that determines how different propensity scores have to be in order to be assigned to different groups.

After implementing these two features, neighborhoods are constructed by grouping together contiguous cells based on the bilateral differences in the propensity of observing an agent. After computing the raw propensity score $\mathrm{P}_t^b = Pr\left(d_t = b\right)$ that is cell-specific, I compute a dissimilarity matrix by calculating the bilateral euclidean distance between

[16]The idea of incorporating these type of constraints to Ward (1963)'s algorithm mainly started in with Michel et al. (2012)'s machine learning application to medical imaging that cluster functional Magnetic Resonance Imaging (fMRI) data into contiguous (three-dimensional) brain regions.

[17]This number has to be at least larger than the threshold.

[18]This is consistent **?** results for genome studies that shows that the implementation of adjacency constraints to Ward (1963)'s agglomerative algorithm reduces the space requirements from $O(mn^2)$ to to $O(mnh)$ where m is the number of characteristics being considered, n is the number of observations (in this case cells) being clustered, and h is number of neighbors to which each element is allowed to be linked to. Moreover, is also reduces the time required to run (time complexity) from $O(mn^2)$ to $O(n(h + log(n)))$.

[19]This method has been proved to significantly reduce the bias due to measured confounders when estimating linear treatment effects. Moreover, increasing the number of strata used should result in improved bias reduction, although the marginal reduction in bias decreases as the number of strata increases (Cochran, 1968).

adjacent cells. This creates a sparse matrix populated by non-zero elements around the diagonal and an arbitrarily large number for the rest of the elements. The algorithm starts with each cell as a neighborhood by itself. Further iterations merge one neighborhood at a time so that the bilateral distance between the neighborhood and the new elements are below a given threshold and minimizes the within cluster variance in the propensity score.

Specifically the algorithm works as follows. Let $\mathcal{P} = \{\mathrm{P}_t^i\}_{i=1}^{\mathbf{B}}$ the set of all cell-level probabilities to be clustered. A neighborhood is a subset of $\mathcal{P}$. The loss of information when grouping objects into a neighborhood $N \subset \mathcal{P}$ is quantified by the *Error Sum of Squares*, ESS:

$$I(N) = \sum_{\mathrm{P}_t^i} \parallel \mathrm{P}_t^i - \overline{\mathrm{P}}_N \parallel^2$$

where $\overline{\mathrm{P}}_N = n^{-1} \sum_{i=1}^{n} \mathrm{P}_t^i$ is the *centre of gravity* of N and n is the number of cells in the neighborhood. Starting from a partition $\{N_1, ..., N_l\}$ of $\mathcal{P}$, the loss of information when merging two neighborhoods N_u and N_v is quantified by:

$$\delta\left(N_u, N_v\right) = I(N_u \cup N_v) - I(N_u) - I(N_v).$$

This is what is known as the Ward's linkage and it is equal to the variation of *within-neighborhood sum of squares* after merging two neighborhoods. The algorithm starts from the trivial partition $\{N_1, ..., N_{\mathbf{B}}\}$with $\mathbf{B}$ singletons neighborhoods $N_i = \{\mathrm{P}_t^i\}$. The algorithm then creates a sequence of partitions by successively merging the two adjacent neighborhoods whose linkage δ is the smallest. The algorithm finalizes when all objects have been merged into a single cluster. Every step in this agglomeration is recorded and the final set of neighborhoods $\{N\}$ is identified at the point in which the within neighborhood variance is minimized given that the distance between neighborhoods is below a predetermined threshold $\overline{\delta}$. The algorithm is summarized in Algorithm (1).

3 Simulations

Besides geo-coded data of location choices, the algorithm requires two inputs from the researcher. First, it requires an specification for the propensity score that leads to cell specific probabilities. This specification will vary depending on the application and the nature of the underlying economic process that drives agents' location decisions. Second, it requires a predetermined threshold $\overline{\delta}$ that provides a limit to how big can the loss of information δ can be when merging two neighborhoods. This threshold is directly related

Algorithm 1 Propensity Score Agglomerative Spatial Strata-Clusters

For a given data set $\mathcal{P} = \{\mathrm{P}_t^i\}_{i=1}^{\mathbf{B}}$, the algorithm is as follow:

1. Initialize the set of neighborhoods to be $\{N_1, ..., N_B\}$ where, for $i = 1, ..., \mathbf{B}$, the i^{th} neighborhood is the i^{th} datum, $N_i = \{\mathrm{P}_t^i\}$
2. Compute the dissimilarity between all pairs of neighborhoods, that is, compute $\delta(N_u, N_v)$ for all $u < v \in \{adjacent_u\}$
3. While there is more than one neighborhood in the set of neighborhoods:
   (a) Merge a pair of neighborhoods which have minimal dissimilarity. Given indices $u' < v'$ such that $\delta(N_u, N_v) = \min_{i<j} \delta(N_u, N_v)$ and $\delta(N_u, N_v) \leq \overline{\delta}$,set $N_{i'} = N_{i'} \cup N_{j'}$ and remove $N_{j'}$ from the set of neighborhoods
   (b) Compute the dissimilarity between the new neighborhood $N_{i'}$ and all other neighborhoods in the set of neighborhoods
4. The final set of neighborhoods $\{N\}$ is defined at the point in which the within neighborhood variance is minimized given that the dissimilarity between neighborhoods is below a given threshold $\overline{\delta}$.

to the size of the resulting neighborhoods, as having a threshold too small will result in neighborhoods that are composed of only one cell, whereas having a threshold too big could presumably result in just one big neighborhood composed by all cells.

Provided there is a good specification for the propensity score available, this sub section analyzes the effect that different applications (represented by two spatial distributions of geocoded location data) and different thresholds might have in the resulting neighborhoods. This results in the simulation of 100,000 agents location choices in two cities, the first where agents locate randomly in the space, and the second where agents cluster next to each other around 50 sub centers. These two simulations can be interpreted in the context of firms making location choices under the presence or absence of agglomeration economies. In the presence of agglomeration economies, firms incorporate the benefits of locating next to each other when making location choices and that results in firms clustering whereas in the absence of agglomeration economies, firms would spread out across the city.[20]

These simulations are performed following an Homogeneous Poisson Point Process

[20] Another way to interpret this exercise is in the context of household location choices. When household preferences for location include other household choices (such as when households choice location based on school quality or based on access to certain rival amenities) it could result into more clustering. However, then when they do not include other household choices or the distribution of amenities mimics the preference of households for location then which clustering of house location choices might be less relevant.

and a Clustered Poisson Point Process and then aggregating these points into hexagon grid cells of 1km length.[21] Figure 1 shows how the simulated location choices result in different distributions of economic activity in the space after aggregating them into their respective grid cells. It clearly shows higher concentration in some areas for the clustered scenario and no clear concentration in the random scenario. This is consistent with the density as it shows that the distribution of probabilities is centered at zero for the clustered point process with a long right tail (which is consistent with a high standard deviation (the red line) of the propensity score) showing that even though most cells have zero agents, there are some few hexagon cells that concentrate higher most of them. In comparison, the random process implies a probability centered in a positive number with a support entirely in positive numbers but an order of magnitude smaller than the clustered case showing that even though there is no high concentration, most hexagon cells have a positive probability.

Given these two environments, the performance of the algorithm can be measured by the degree in which the results mimic what we observe in their distributions. In particular, at a given threshold, we would expect to detect fewer and clearer neighborhoods for the case of the clustered point process relative to the case of the random cluster point process. Moreover, we should expect that increasing the threshold would quickly lead towards having all cells merged together into one neighborhood in the random case, whereas it would lead to a few very stable neighborhoods surrounded by zero-probability cells in the clustered case. In fact, this provides us with the main data requirement of this methodology. In order to identify neighborhoods that are meaningful we need an uneven distribution of economic activity with clear patterns of concentration. Hence, we need a context in which economic agents preferences for location are skewed towards certain parts of the city.

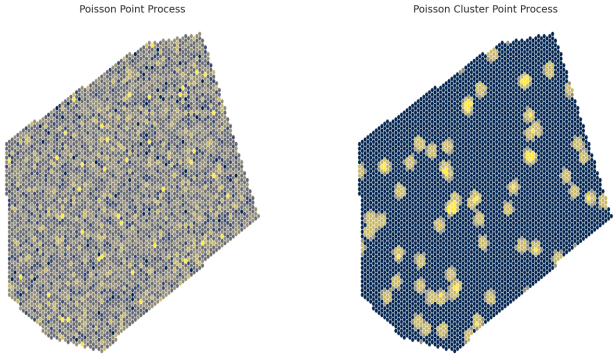
Following the propensity score literature, I set the threshold (or caliper) as a function of the standard deviation of the propensity score and run the algorithm in both environments for 100 different values starting from 0.01 standard deviations and ending in 15 standard deviations.[22] Figure 2 shows the resulting neighborhoods for 6 evenly spaced thresholds from 0.01 to 15 for a total of 100 thresholds. Visually, a neighborhood is composed by all the cells that contiguously share the same color, so even though there are

[21] For more details on how this simulation is performed please refer to Lomax et al. (2019) and the documentation of the `python` package `pointpats` from the Python Spatial Analysis Library (PySAL) ongoing project that is based on Rey and Anselin (2007).

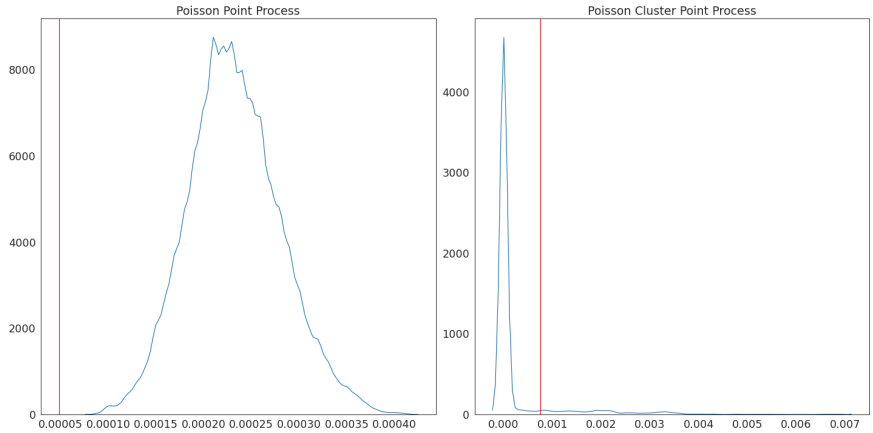
[22] **TODO: Explain why too many standard deviations still leave singletons (because the threshold is in propensity score terms and $I(N_u \cup N_x)$ is in ESS which is equal to the propensity score when is a singleton). Maybe think a way to re-define the threshold in those terms.**

Figure 1: Propensity score distribution: Random vs Cluster Processes

(a) Spatial Distribution



(b) Density



some repeated colors, if they are not attached that means they do not belong to the same neighborhood.[23]

The algorithm correctly identifies locations with high concentration as neighborhoods in the clustered simulation while it incorrectly identifies locations as neighborhoods in the random simulation.[24] This means that even though the algorithm is capturing some neighborhoods in the random case, these are meaningless as, by construction, there are no big differences between cells in this scenario.[25] In contrast, the neighborhoods identified in the clustered case have meaning as they are composed by very similar cells with non-

[23]This is thanks to the adjacency constraint in the algorithm which forces compactness in the resulting neighborhoods.

[24]By construction, the random environment is composed by many cells with very similar (close to zero) probabilities, while the clustered environment is composed by cells that have similar probabilities only around 50 randomly located sub centers and zero probability elsewhere.

[25]The algorithm joins neighborhoods based on the the minimization of the loss of information δ. At the starting point, the loss of information between joining one cell with their immediate neighbors is given by the bilateral difference between cell-level propensity scores.

zero probabilities that are around the sub centers.

In addition to the degree in which neighborhoods are correctly identified or not, another aspect that is of interest is how the results vary while increasing the threshold. Figure 2 shows that the random case starts with a higher number of neighborhoods than in the clustered case and quickly decreases to around 5 big neighborhoods when setting the threshold to 15 standard deviations. In contrast, even though the clustered case starts with a few hundred neighborhoods, they are always around their respective sub centers and they stay that way when increasing the threshold, so that even though the composition of the neighborhood is less homogeneous in that case, the neighborhoods that the algorithm identifies are still composed by cells that are very similar to each other and different from the ones outside of those neighborhoods.

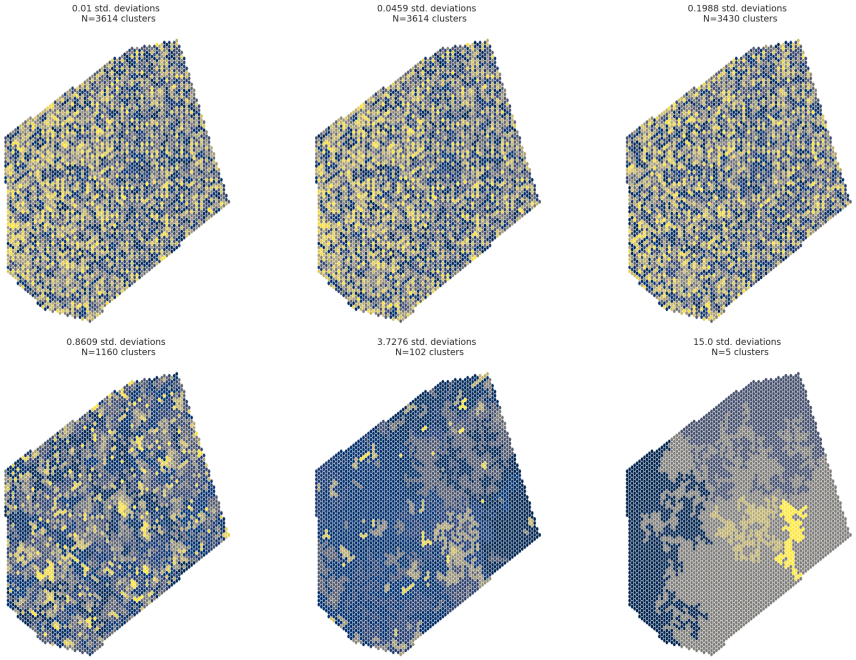
As noted before, the threshold in this algorithm mimics the use of a caliper in the propensity score literature. This implies a trade-off between the homogeneity of cells within a neighborhood and the number of neighborhoods. A small threshold will lead to neighborhoods that are composed by cells that share a very similar propensity score at the expense of having too many neighborhoods or no neighborhoods, leaving no use for the algorithm. A large threshold will lead to neighborhood that are composed by cells that have heterogeneous propensity scores with the benefit of delivering a few neighborhoods substantially decreasing the spatial dimensionality of the data.

Depending on the application and identification strategy the researcher is trying to use, the trade-off between within and between neighborhoods homogeneity might be more or less relevant. For example, if we are using fixed effects to control for the variation between neighborhoods then we need to be enough variance left within neighborhoods so that the within coefficients are identified. In contrast, if for instance, we want to understand how demographic characteristics vary across neighborhoods we might want to be sure that there is enough variance between neighborhoods so that the comparison is meaningful.

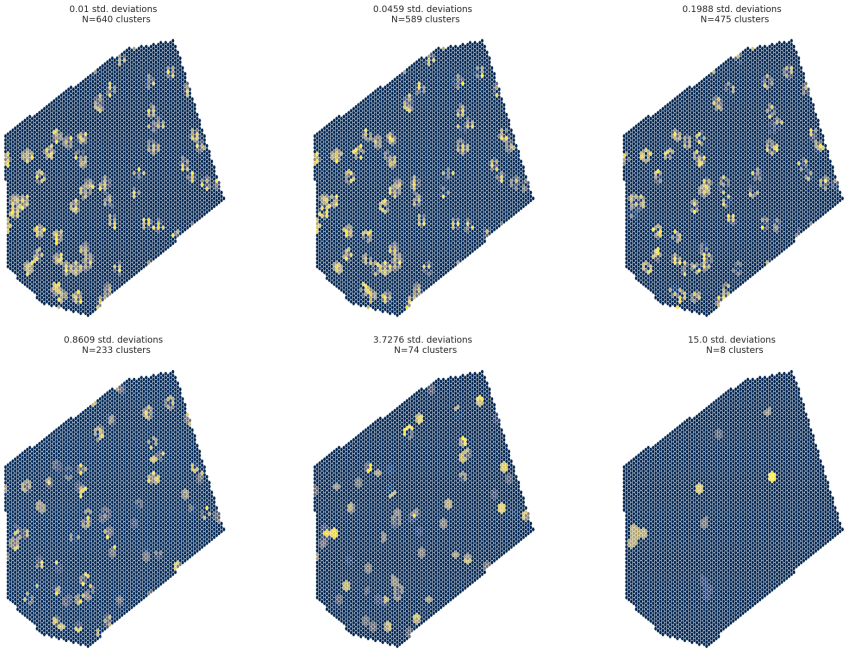
By doing a variance decomposition, Figure 3 shows how this trade-off evolves while the threshold increases. The two lines represent the percentage of the total variance that is explained by the between (or MSS) versus the within (RSS) variation in propensity score. As expected, while the threshold increases, the variance explained by the within variation increases as neighborhoods are bigger and are composed by more heterogenous cells. In contrast, the variance explained by the between variation decreases as there are less number of neighborhoods and those neighborhoods are more similar to each other.

Figure 2: Results

(a) Poisson Random Process



(b) Poisson Cluster Process

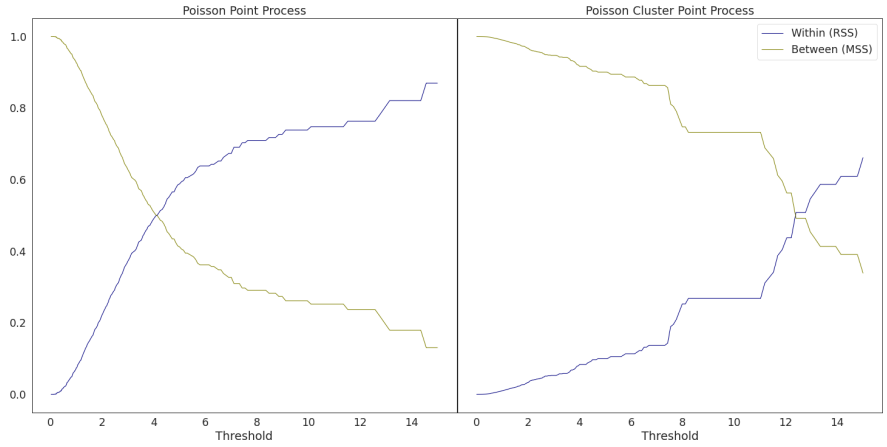


An interesting result from this exercise is the difference in the speed at which the within variation increases or the between variation decreases. For the random case, the within variation quickly overtakes the between variation, whereas in the clustered case

this is never the case. In fact, up until around 7 standard deviations, the within variation remained more or less constrained to represent less than 20% of the total variation with no big changes between different thresholds. This implies that under a clustered scenario, neighborhoods are more stable and are subject to less changes than under the random scenario.

To summarize, the simulation exercise implies that in order to obtain meaningful and stable neighborhoods an uneven distribution of location choices is needed. This will not only make the algorithm obtain clear differences between neighborhoods but it will also make the algorithm less sensitive to the definition of the threshold, and hence, produce neighborhoods that are more stable across thresholds. These results provide good news in contexts where there exist high concentration of economic activity, as given the computational burden associated with taking this to smaller grid cells, we need to be confident that the threshold will not make a huge difference in the resulting neighborhoods.

Figure 3: ANOVA across Thresholds



4 Economic Neighborhoods

One of the main advantages of this algorithm is that is flexible enough to be applied to different contexts. However, the previous section showed that there are certain data requirements that are needed in order to obtain sensible neighborhoods. In this section, I present two applications for the Great Toronto Area to illustrate these findings and to further analyze the properties of the algorithm. The applications use geographical data on land use alongside location choice data for two different type of agents: firms and households.

The first application aims to identify industrial neighborhoods using points of interest data that includes all firms and their establishments in the city.[26] The second application aims to identify residential neighborhoods using the universe of real estate transactions captured by Multiple Listing Services (MLS). The key feature that the algorithm exploits from the datasets for both applications is that they are datasets that captures revealed preferences. This means that each observation represents the optimal choice of an agent given the characteristics of the chosen option compared to all possible alternatives. The whole data used is shown in Figure 4.

It is important to note that these applications do not attempt to answer the question of what determines location choices as there is no identification strategy to deal with the identification problems associated with the estimation of the coefficients associated with that decision. The aim of this exercise is to show the capabilities of the algorithm in delivering neighborhoods that are meaningful and that will help us understand the similarity of location choices over the space and across agents. In this setting, we can think of the propensity score as playing a mechanical role in balancing the covariates associated with the location choice. This is in fact how propensity scores have started to be seen in the matching literature, in which the role of the specification is to lead to an accurate approximation of the conditional expectation of choosing a location given the covariates (Imbens, 2015).

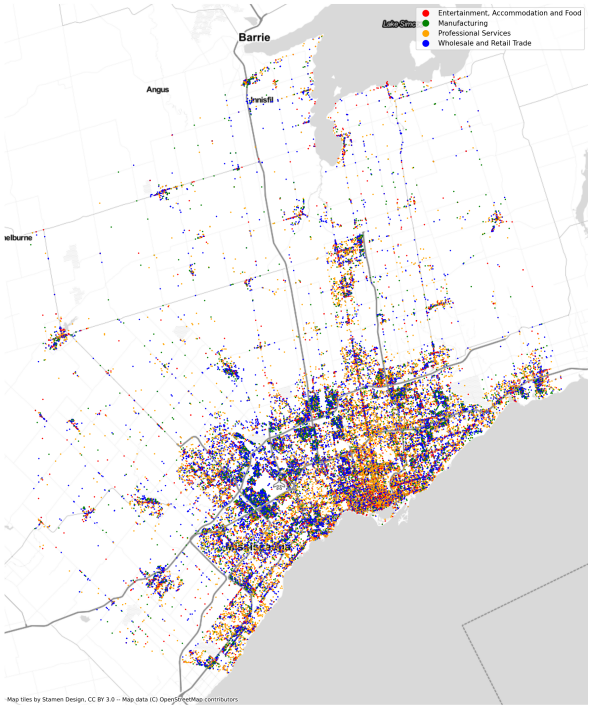
As mentioned in section (4), the Great Toronto Area has a total usable area of about $7,124.15\,km^2$ divided into 75-meter hexagon grid cells that comprise $474,943$ ($740,286$ when including water bodies and a buffer zone) grid cells of about $0.015\,km^2$ each. For each application, the process consists of assigning each choice observation to one of these grid cells and geographically matching them with the location characteristics. Location characteristics are derived from the same datasets of location choices and from an additional dataset that contain land uses for the whole city. Then, a propensity score is calculated based on a combination of the location characteristics that presumably affect the underlying decision process of the agents. Finally, the algorithm is applied.

The next subsections explain the details about the data and the propensity score specification used in each application.

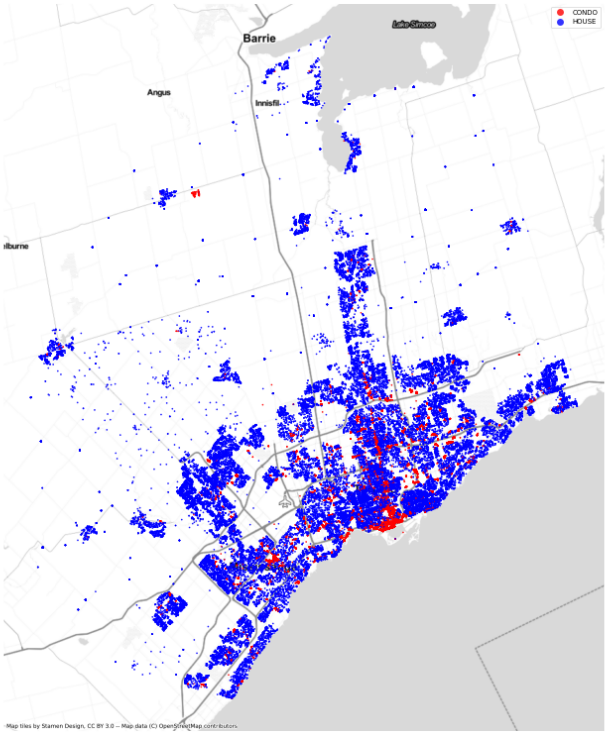
[26] A point of interest, or POI, is a specific point location that someone may find useful or interesting. When thinking about a POI, most consumers would refer to hotels, campsites, fuel stations or any other categories used in modern automotive GPS systems. However, in general, POI data also includes locations for all kind of firms as is not only used by end consumers traveling within or across cities but also by logistic and shipping companies.

Figure 4: The Data

Firms



Real Estate Transactions



4.1 Industrial Neighborhoods

The underlying location choice problem for firms is one of profit maximization. Under profit maximization, the observed location is the one that maximizes the expected profits relative to all other possible locations. This implies that the probability that a firm locates in a particular place depends on how the characteristics of that location affect profits relative to those of all other locations. Therefore, the explanatory variables to be considered are those specific to location believed to have an impact on the expected profits of the firm. In this application, the variables I consider will measure agglomeration economies, input prices, and demand conditions. In the following paragraphs, I explain the data used to measure these variables and the methodology to do so.

The main dataset used to delineate the industrial neighborhoods is therefore a dataset of firm location choices. These location choices comes from the 2012 Enhanced Points of Interest (EPOI) files produced by DMTI Spatial Inc. that is a national database of over 1 million Canadian businesses and recreational points of interest at the establishment level, and arguably comprises the universe of businesses and recreational points of interest in major metropolitan areas, such as the GTA. For each EPOI, the dataset includes the North American Industry Classification System (NAICS) industry classification code, name, and address. I filter this dataset to only include firms that are located in the GTA and belong to the manufacturing (NAICS 3), retail (NAICS 4), service (NAICS 5), or food and lodging (NAICS 7) sectors. This leaves us with a total of 125,435 firms/establishments in the area. A snapshot of the data can be found in Figure 4.

In general, agglomeration economies consist of those externalities resulting from the spatial concentration of economic activity. The benefits of this agglomeration can come from the concentration of firms in the same industry as well as from concentration in suppliers or buyers. For estimation purposes that will be clear in the next few paragraphs, this paper assumes that all firms in the same industry are equal and face the same conditions. That means that to measure the level of agglomeration, I only need to focus on the measurement of upstream and downstream firms. To do this, I combine the location choices data with the 2012 Input-Output Tables produced by Statistics Canada. These tables allow me to obtain the relative importance of firms of a certain industry for the production of a firm from another industry. This allows me to measure the upstream and downstream relationship in a given place. In particular, I compute the weighted sum of firms upstream and downstream where the weights are given by the input-output relationships found in the input-output tables. Moreover, in order to account for the centrality of a given location, I compute market access measures that include all the upstream or downstream firms within one kilometer of the location.

Other supply conditions, in particular input prices, are also considered in the location choice process. Obtaining said prices is a hard task as they are often private information. However, given the fact that this paper uses small geographical areas (the grid cells) as locations, it is safe to assume that input prices are constant within each cell. Therefore, there is no need to measure them.

On the demand side, depending on the industry, firms might want to stay close to the public. To measure that, I use the EPOI data to capture locations that attract people. In particular, I consider access to parks and waterbodies, banks, hotels, tourist attractions and postal services. Moreover, in order to further measure access to consumers, I consider access to real estate housing using the 2013 Address Points files also from DMTI that contain all the residential and commercial addresses for Canada.

This is translated into the following specification for the decision rule of a given firm i from industry j that is deciding to locate in cell-grid l at time t:

$$d_{ijlt} = \beta + f_{POI}\left(\#\text{POI}_{lt}\right) + f_{LAND}\left(\text{LAND}_{lt}\right) + f_{HOUSE}\left(\#Houses_{lt}\right) + f_{UP}\left(\#Up_{jlt}\right) + f_{DOWN}\left(\#Down_{jlt}\right) + \epsilon_{ijlt} \quad (1)$$

where d_{ijlt} is one if the firm decides to locate in cell-grid l and zero otherwise, $\#POI_{lt}$ are the number of points of interest (banks, hotels, tourist attractions or postal services) in cell-grid l, $LAND_{lt}$ is the percentage of land of cell i that correspond to different land uses (parks and waterbodies, commercial or industrial areas), ${}_i\#Houses_{lt}$ is the number of houses, $\#Up_{jlt}$ is the weighted sum of firms upstream, $\#Down_{jlt}$ is the weighted sum of firms downstream and ϵ_i is the error term. The function $f_X(X_i)$ is a function that represent the access to X in location i, and is given by

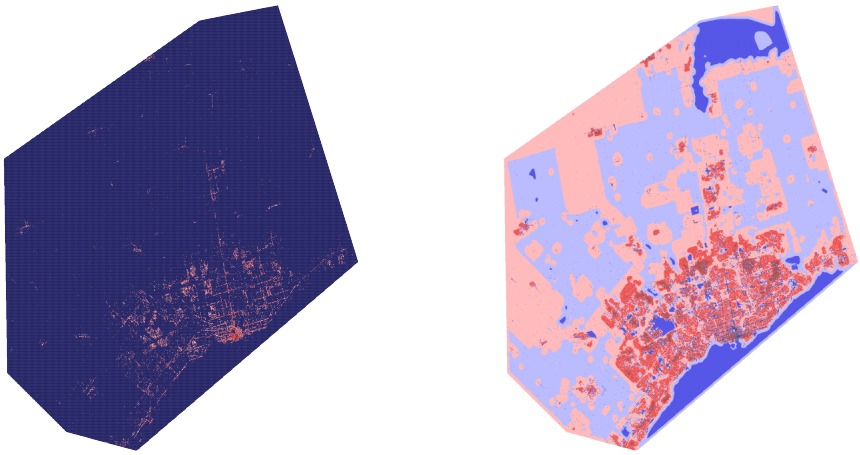
$$f_X\left(X\right) = \beta_X X + \beta_{\text{MA\_}X}\text{MA\_}X \quad (2)$$

where $\text{MA\_}X_i$ is a measure of centrality of a given location relative to X_i. The measure is computed as weighted sum of X of all cells within one kilometer of the origin cell where the weights are given by the inverse of distance with an exponential decay and $\rho = 1$.

This decision rule is estimated following McFadden (1973)'s conditional logit model in which location choices reveal the preferences associated with the agent's utility/profit maximization problem. In practice, the application of this approach to any location choice poses a problem related to the definition of the spatial choice set. Ideally, small areas should be used, because factors usually identified as relevant for location decisions (such as agglomeration economies, labor market conditions, or the cost of inputs) apply at the local level and consequently cannot be adequately taken into account when the model

considers large areas in the spatial choice set. However, the use of small areas poses a difficulty for estimation, as the conditional logit model does not handle large choice sets very well. Guimarães et al. (2003) provides a solution to this problem by demonstrating that, under the assumption that individual decisions are based exclusively in a vector of choice-specific attribute variables common to all decision-makers, the coefficients of the conditional logit model can be equivalently estimated using a Poisson regression, which is what I use instead.

Figure 5: Distribution of Raw vs Propensity Score Probabilities: Industrial Neighborhoods



Different industrial sectors might have different valuation for location attributes implying different neighborhoods across sectors. To account for that, the decision rule is estimated for all firms as well as for firms in the manufacturing, retail, service, and food and lodging sectors. The estimation of the decision rule leads to a propensity score whose distribution for all firms can be found in Figure 5.[27][28] This figure presents the comparison between the raw probability and the propensity score associated to a given cell. The color code of the figure is from blue (lower probability / propensity score) to red (higher probability / propensity score). The first thing to note is that both distributions share similar patterns signaling that the propensity score is accurately approximating the conditional expectation of finding a firm in a given grid-cell. The patterns show that the spatial distribution of economic activity is highly skewed. In particular, there is a higher concentration of economic activity in the downtown areas, along major streets and specially around the Toronto Financial District. Moreover, it shows that moving away from the city quickly

[27]Even though interesting, the results for the coefficients of the propensity score are not the main objective of this paper and hence are left in Table A.3 in the Appendix.

[28]The distribution of the propensity scores for sectors can be found in Figure A.2 in the Appendix.

decreases the probability of finding a firm keeping them relegated to only around major streets. The main difference between the two spatial distributions is the sharpness. The propensity score is more blurred than the raw probability and this is because the propensity score is effectively smoothing out the probability as it considers market access measures that account for the access within 1km from the grid-cell.

The algorithm is applied after the estimation of the propensity score. Following the discussion in Sections 2 and 3, I use six thresholds and define them as a function of the standard deviation of the propensity score. The thresholds are set at $4, 2, 1, 0.5, 0.1$ and 0.01 standard deviations of the propensity score. Neighborhoods are defined as an area that has been the result of the agglomeration process of the algorithm. This means that a neighborhood is every area that is composed by at least two grid-cells. Table 1 shows the characteristics of the resulting neighborhoods for the minimum threshold level (one standard deviation) that leads to neighborhoods that account for at least 90% of the firms in the whole sample.[29]

Table 1: Neighborhood Characteristics: Industrial Neighborhoods

	Number of Firms	Firms in Neighborhoods	Number of Neighborhoods	Area (sq km)	Length (km)	Width (km)	Length/Width
All Firms	125,435	91.03	13,947	0.713	0.838	0.431	1.764
				(20.945)	(2.411)	(1.003)	(0.891)
Manufacturing	17,661	82.35	6,130	1.467	1.099	0.548	1.807
				(29.642)	(3.967)	(1.726)	(0.954)
Wholesale and Retail Trade	44,846	84.58	10,187	1.055	0.966	0.505	1.813
				(24.591)	(2.661)	(1.255)	(0.808)
Professional Services	47,464	76.62	7,085	0.996	0.867	0.467	1.62
				(23.151)	(3.155)	(1.265)	(2.978)
Entertainment, Accommodation and Food	15,464	79.91	5,079	1.818	0.964	0.526	1.695
				(58.8)	(3.499)	(1.733)	(1.018)

Note: Results correspond to running the algorithm for each group of firms with a threshold set to the minimum threshold that leads to neighborhoods accounting for 90% of the firms in the whole sample. This threshold is one standard deviations in the propensity score. Firms in neighborhoods correspond to the percentage of firms that belong to neighborhoods that have at least two cells. Length (and width) correspond to the longest (and shortest) side of the minimum bounding rectangle that contains the neighborhood. Standard deviations are in parenthesis.

The main result from the exercise is that neighborhoods are different from each other across industries.[30] There are three main statistics reported in the table. The first one is the area in square kilometers that each neighborhood has. On average industries that have less reach due to window shopping or less within city tradability, such as professional services and retail, have smaller neighborhoods than industries that are more tradable within a city, like manufacturing and entertainment services.[31]

[29] All results are available in Table A.1 in the Appendix.

[30] Another interesting result is that even though the propensity score specification includes interactions across firms through input-ouput connections, these connections are not enough to capture the relationships between them. This is clear because splitting by industries decreases the percentage of firms within the industry that belong to a resulting neighborhoods, which means that the algorithm learns about interactions across industries more than what input-output relationships capture.

[31] Even though there might be some specialization within industries, it is not captured in the current

The second and third statistics provide information about the shape of the neighborhoods by showing the average length (longest side of the minimum bounding rectangle that contains a neighborhood) and width (the shortest side of the minimum bounding rectangle that contains a neighborhood). An interesting result arises here. On average, neighborhoods can be represented by a shape longer than it is wide, like a rectangle or an ellipse. Further visual inspection shows that these neighborhoods tend to locate around major streets. This is true for all industries, but more so for the manufacturing and trade industries. Moreover, the table shows that the smaller area of professional services and retail is given more by differences in length than in width, with neighborhoods that are between 42% to 59% shorter than the manufacturing and entertainment services.

4.2 Residential Neighborhoods

The underlying location choice problem for firms is one of utility maximization. Under utility maximization, the observed location is the one that maximize the expected utility of the household relative to all other possible locations. This implies that the probability that a household locates in a particular place depends on how characteristics of that location affect the household utility relative to the characteristics of all other locations. Therefore, the explanatory variables to be considered are those specific to a location that are believed to have an impact on a household's expected utility. In this application, the variables I consider measure the location's access to goods, services and amenities together with the average characteristics of a housing unit in that location. In the following paragraphs, I describe the data I use to measure these variables and the methodology to do so.

The main data comes with a wide range of transaction and property attributes: the transaction date, the postal code, the asking and sold price, the type of property, the number of floors, the number of rooms, bedrooms, kitchens, washrooms, family rooms, fireplaces, the size of the lot, and parking space. The dataset contains 68,184 transactions for 2013 in the GTA. To geolocate properties, I assign to each transaction the coordinates of the centroid of the postal code at which the property is located.[32] A snapshot of the data can be found in Figure 4.

To measure access to goods, services and amenities, I combine this data with the EPOI data from the industrial application and calculate the same market access measures. To

specification as the data is only split by the macro industries listed above. Further division of the data into more specific industries is possible if the number of observations increases. This could be accomplished by introducing a time-component in the location choice.

[32] Postal codes in Canada are small and generally represent a block face in large urban areas, like the GTA.

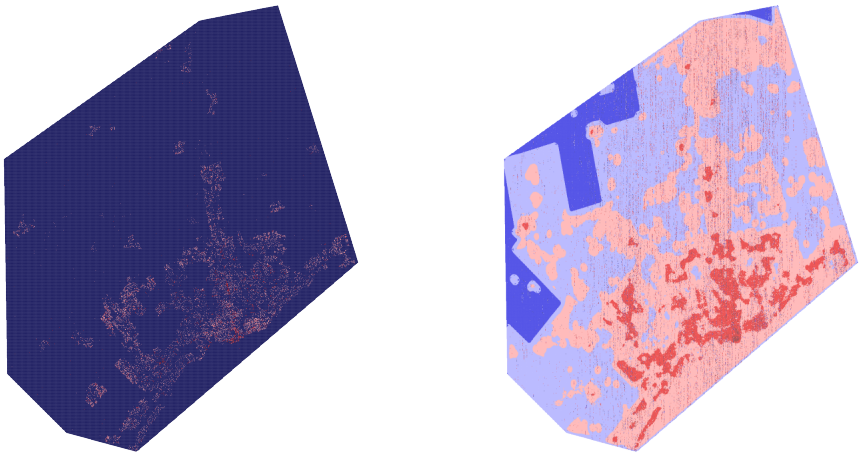
measure characteristics of the average housing unit in a given cell, I take the average number of rooms, the total number of rooms and the average lot size from the transaction data. Finally, each observation represents a transaction and hence it does not represent the housing stock at a given period in time. For this reason, I also use the 2013 Address Points files to obtain a measure of housing supply by calculating the number of residential addresses in a given cell.

This is translated into the following specification for the decision rule of a given household i that is deciding to locate in cell-grid l at time t:

$$d_{ilt} = \beta + f_{POI}\left(\#\text{POI}_{lt}\right) + f_{LAND}\left(\text{LAND}_{lt}\right) + f_{HOUSE}\left(\#Houses_{lt}\right) + \beta_{rooms} AvgRooms_{lt} + \beta_{allrooms} SumRooms_{lt} + \beta_{lot} AvgLotSize_{lt} + \epsilon_{ilt} \quad (3)$$

where d_{ilt} is one if the household decides to locate in cell-grid l and zero otherwise, $\#POI_{lt}$ are the number of points of interest (banks, hotels, tourist attractions or postal services) in cell-grid l, $LAND_{lt}$ is the percentage of land of cell l that correspond to different land uses (parks and waterbodies, commercial, government or industrial areas), $\#Houses_{lt}$ is the number of houses, $AvgRooms_{lt}$ is the average number, $SumRooms_{lt}$ is the total number of rooms available for sale, $AvgLotSize_{lt}$ is the average lot size of the location (zero in the case of condos), ϵ_{ilt} is the error term. As in the case of industrial neighborhoods, the function $f_X(X_i)$ is a function that represent the access to X in location i, and is given by equation 2.[33]

Figure 6: Distribution of Raw vs Propensity Score Probabilities: Residential Neighborhoods



[33]The coefficients of the propensity score can be found in Table A.4 in the Appendix. The distribution of the resulting score by property type can be found in Figure A.4 in the Appendix.

Different households might have different valuation for location attributes implying different neighborhoods across household types. Unfortunately, I cannot measure that but I proxy it by separating the estimation by property types (condos or houses) and assuming that households that buys condos have different preferences than those that buy houses. The estimation of the decision rule leads to a propensity score whose distribution for all household types can be found in Figure 6. As with industrial neighborhoods, this figure presents the comparison between the raw probability (left) and the propensity score (right) associated to a given cell. The first thing to note is that both distributions share similar patterns signaling that the propensity score is approximating the conditional expectation of finding a property in a given grid-cell very well. Although the raw probability is very low, it still present some concentration in certain parts of the city, especially in downtown areas. Moreover, it shows that moving away from the city quickly decreases the probability of finding a transaction keeping them relegated to only around major streets. Replicating the case of industrial neighborhoods, the main difference between the two spatial distributions is their sharpness. The propensity score is more blurred than the raw probability and this is because the propensity score is effectively smoothing out the probability as it considers market access measures that account for the access within 1km from the grid-cell.

The algorithm is applied after the estimation of the propensity scores. As in industrial neighborhoods, I use six thresholds and define them as a function of the standard deviation of the propensity score. The thresholds are set at $4, 2, 1, 0.5, 0.1$ and 0.01 standard deviations of the propensity score. Table 2 shows the characteristics of the resulting neighborhoods for the minimum threshold ("two" standard deviations) level that leads to neighborhoods that account for at least 90% of the transactions in the whole sample.

Table 2: Neighborhood Characteristics: Residential Neighborhoods

	Number of Transactions	Transactions in Neighborhoods	Number of Neighborhoods	Area	Length	Width	Length/Width
All Transactions	68,184	89.57	6,921	0.685	0.528	0.284	1.633
				(19.66)	(2.572)	(1.164)	(1.081)
House Transactions	48,510	92.84	6,834	0.671	0.495	0.274	1.613
				(19.535)	(2.413)	(1.119)	(0.947)
Condo Transactions	19,674	77.73	5,146	1.442	1.062	0.497	2.049
				(27.742)	(3.694)	(1.608)	(1.845)

Note: Results correspond to running the algorithm for each type of transaction with a threshold set to "two" standard deviations in the propensity score. Transactions in neighborhoods correspond to the percentage of transactions that belong to neighborhoods that have at least two cells. Length (and width) correspond to the longest (and shortest) side of the minimum bounding rectangle that contains the neighborhood. Standard deviations are in parenthesis.

4.3 Neighborhood Similarity

One of the benefits of this algorithm is the ability to obtain neighborhoods that are presumably different across types of agents. In particular, in the applications describe above,

the algorithm was run for different industries and different real estate property types. In order to measure the degree of agreement between these neighborhoods, I compute the adjusted Rand index (Hubert and Arabie, 1985), which is a measure of similarity between two groups or classifications (boundaries of neighborhoods in this case) that is widely used in machine learning. The unadjusted index is given by

$$\text{RI} = \frac{\text{\# of agreements}}{\text{\#agreements} + \text{\#disagreements}}$$

where the number of agreements is given by number of grid-cell pairs that are assigned to the same (and different) neighborhoods in both boundaries, and the number of disagreements is given by the number of pairs of grid-cells that are assigned to same neighborhoods in one boundary but different neighborhoods in the other boundary. The adjusted version of this index corrects by chance by ensuring to have a value close to 0.0 for random classification independently of the number of neighborhoods and exactly 1.0 when the classifications are identical.[34] As a result, the Rand index measures the ratio of agreements between the two boundaries over the total number of grid-cell pairs, and its values range between 0, dissimilarity, and 1, maximum similarity.

Table 3: Agreement Between Industrial Neighborhoods (ARI)

	All Firms	3	4	5	7
All Firms	1.00				
3 - Manufacturing	0.77	1.00			
4 - Wholesale and Retail Trade	0.76	0.84	1.00		
5 - Professional Services	0.76	0.80	0.81	1.00	
7 - Entertainment, Accommodation and Food	0.37	0.36	0.35	0.34	1.00

Based on the findings of the previous subsections, I perform three exercises to assess the degree of agreement between and across types of neighborhoods. Table 3 shows that, even though there is a high degree of agreement between industrial neighborhoods is still far from 1 showing. Moreover, it shows that the entertainment industries have neighborhoods that are very different from the ones from other industries which signals that the

[34]The unadjusted rand index suffers from one drawback; it yields a high value for pairs of random partitions of a given set of examples. To understand this drawback, think about randomly grouping a number of examples. When the number of partitions in each grouping, that is when the number of clusters, is increased, more and more example pairs are going to be in agreement because they are more likely to be not grouped together. This will result in a high RI value. Thus, RI is not able to take into consideration effects of random groupings. To counter this drawback, an adjustment is made to the calculations by taking into consideration grouping by chance. This is done by using a specialized distribution, the generalized hypergeometric distribution, for modeling the randomness. The resulting measure is known as the adjusted Rand index (ARI).

location choice process that these industries face is fundamentally different. Table 4 replicates the exercise for residential neighborhoods and shows a similar degree of agreement between condo and house neighborhoods. Finally, Table 5 performs the exercise between industrial and residential neighborhoods and shows Rand indices close to zero which support the intuition that that industrial and residential neighborhoods are fundamentally different.

Table 4: Agreement Between Residential Neighborhoods (ARI)

	all	houses	condos
All Transactions	1.00		
Houses	0.83	1.00	
Condos	0.79	0.79	1.00

Additional results come from analyzing how residential (relative to industrial) neighborhoods change across thresholds. Table A.2, in the appendix, gathers all the results for the exercise and shows that residential neighborhoods are less clear than industrial neighborhoods and the differences between them less informative. This is shown by the requirement that the threshold be higher ("two" in this case vs "one" in the industrial neighborhoods) in order to account for 90% of the data. Moreover, this is further demonstrated by the sensitivity of residential neighborhoods to changes in the threshold. This can be attributed to the fact that the distribution of real estate sales (houses in particular) is widely spread across the city which is consistent with the results from the previous section signaling that in order for the algorithm to work properly we need a highly skewed distribution of economic activity. In this case, splitting the data helps, as the distribution of condo sales is more concentrated in certain locations of the city, which in fact delivers condo neighborhoods that are more stable across thresholds.

Table 5: Agreement Between Industrial and Residential Neighborhoods (ARI)

	All	Houses	Condos
All Firms	0.084	0.085	0.082
3 - Manufacturing	0.083	0.083	0.081
4 - Wholesale and Retail Trade	0.081	0.081	0.080
5 - Professional Services	0.080	0.080	0.079
7 - Entertainment, Accommodation and Food	0.047	0.046	0.045

This section has shown that different agents making location choices lead to different economic neighborhoods across agents. These economic neighborhoods differ in size and shape and are distributed differently across the space. These differences are a signal of

the potential issues that a misalignment between economic and legal neighborhoods may produce. The next section focuses on that by studying how the economic neighborhoods are different from their legal counterpart and how their differences relate to the potential misalignment issues.

5 Economic vs. Legal Neighborhoods

The misalignment between 'legal' and 'economic' neighborhood boundaries may produce three issues that the algorithm aims to avoid. First, legal neighborhoods may present the modifiable areal unit problem as they are not the unit of analysis at which agents make location choices but they are areas that are "arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did" them (Openshaw, 1983). Second, assumptions of symmetric interactions within neighborhoods are common in research but symmetric interactions are more-or-less likely to hold depending on the size of the neighborhoods. And third, structural models that use Type 1 distributions to match the spatial distribution of economic activity tend to assume uncorrelated shocks.

In this section, I study how 'legal' neighborhoods differ from 'economic' neighborhoods besides their descriptive statistics and through the lens of the issues mentioned above. Addressing the first and second issue, I first study the size distribution of neighborhoods by analyzing the extent to which neighborhoods can be described by Zipf's law. In geography, Zipf's is an empirical law about the size distribution of spatial units. If the law holds perfectly, then the size distribution can be approximated with a Pareto distribution with shape parameter equal to one. If the shape parameter is equal to one then the power law implies that, within a city, the largest neighborhood is roughly twice the size of the second largest neighborhood, about three times the size of the third largest neighborhood, and so on. If the shape parameter is greater than one, it indicates that the size is more evenly distributed across neighborhoods than what the Zipf's law predicts.

Since the seminal work of Gabaix (1999), an enormous amount work on the size distribution of cities has been published (Nitsch, 2005). However, the evidence reported by this literature is inconclusive and debatable. The difference among results seems to be attributable to the city definition employed suggesting that the analysis of size ranking distribution may be affected by a modifiable areal unit problem (MAUP). But, as noted by Duranton and Puga (2014), even though it's empirical validity is debatable it is still useful to study the differences in size distribution through the lens of Zipf's law because both the regularities and the observed empirical deviations from it can be used to guide the modeling of economic processes underlying neighborhood size distributions.

To address the third issue, I then compute the Global Moran (1950)'s I index of spatial correlation to test the intuition that the algorithm delineates neighborhoods that have zero spatial correlation between the location choices between a neighborhood and their immediate neighbors. This index represents the correlation between a variable and its surrounding values. A positive Moran's I indicates that similar observations are closer to each other, while a negative value indicates that dissimilar values are closer to each other, and values around zero would indicate that there is no spatial correlation, and instead, observed values are randomly distributed.

5.1 Neighborhood Size Distribution

Following Gabaix and Ibragimov (2011), this subsection tests the Zipf's law for neighborhoods by estimating the relationship between the log of size and the log of size rank. For each threshold and firm/household type I estimate:

$$\log\left(\text{Rank}_i - \frac{1}{2}\right) = \alpha + \beta \log\left(\text{Size}_i\right) + \epsilon_i \tag{4}$$

where i is a neighborhood indicator. Size_i can be two measures of the relative importance of a neighborhood. The first measure is the number of agents choosing a given neighborhood, this is translated to the number of firms for industrial neighborhoods and the number of real estate transactions for residential neighborhoods. The second measure is the resulting area of the neighborhood. Rank_i computes the rank across neighborhoods of Size_i.

As mentioned before, if Zipf's law holds then $\widehat{\beta} = 1$. A coefficient greater than 1 suggests that within a city the size is more evenly distributed across neighborhoods than what the rank size rule predicts.[35] The results from estimating the coefficients of equation (4) for industrial and residential neighborhoods are reported on Tables (6) and (7) respectively. Following the previous definition, neighborhoods are those areas that are composed of at least two grid-cells cells (i.e. neighborhoods delivered by the algorithm). Moreover, I further restrict the sample to those neighborhoods and postal codes that have at least one firm (real estate transaction) for the case of industrial (residential) neighborhoods. This means that the set of postal codes being used in industrial vs residential neighborhoods might be different and hence leads to different results.

Results show that the area of industrial neighborhoods is more or less similarly dis-

[35]The $\widehat{\beta}$ coefficient can also be interpreted as an indirect indicator of a neighborhood primacy, as a coefficient smaller than one suggests that one or a few large neighborhoods dominate the urban landscape of the city (Rosen and Resnick, 1980).

tributed to the area of postal codes. This is not the case when measuring size as the number of firms in an area. In that case, industrial neighborhoods are more evenly distributed than postal codes. An explanation for this comes straight from the algorithm. The objective of the algorithm is to put very similar areas together. For example, if an area in the middle of downtown with a high number of firms has a certain level of attractiveness, and is adjacent to areas with similar attractiveness but less number firms, then the algorithm will put them together whereas in the postal code scenario they would be separate. This is what makes the number of firms more evenly distributed. This argument, however, does no necessarily hold for the area of the neighborhoods, which is presumably why we observe a more even distribution.

Table 6: Zipf's Law: Industrial Neighborhoods

	Area		Number of Firms	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Firms	-0.736	-0.774	-0.722	-0.37
	(0.001)	(0.001)	(0.001)	(0.001)
Manufacturing	-0.528	-0.711	-0.894	-0.362
	(0.002)	(0.002)	(0.004)	(0.002)
Wholesale and Retail Trade	-0.639	-0.734	-0.771	-0.363
	(0.001)	(0.001)	(0.002)	(0.002)
Professional Services	-1.561	-0.771	-0.822	-0.37
	(0.006)	(0.001)	(0.001)	(0.001)
Entertainment, Accommodation and Food	-0.494	-0.694	-0.875	-0.338
	(0.002)	(0.002)	(0.005)	(0.002)

Note: Results correspond to running the algorithm for each group of firms with a threshold set to number of standard deviations in propensity score that one standard deviations in the propensity score. The reported values correspond to the β coefficient that results from running the following regression $\log(Rank(var)) = \alpha + \beta \log(var) + \epsilon$ where var is the area or the number of firms at the neighborhood or postal code level and $Rank(var)$ correspond to the rank of said variable across geographies. Standard errors are reported in parenthesis.

Residential neighborhoods do not only differ to industrial neighborhoods in terms of shape and size, but also in how the size is spatially distributed. I find evidence that the area of residential neighborhoods is less evenly distributed than postal codes, which is the opposite to the case of industrial neighborhoods. Moreover, the distribution number of transaction of residential neighborhoods is similarly distributed to the number of transactions in postal codes, which is again the opposite to the case of industrial neighborhoods.

Table 7: Zipf's Law: Residential Neighborhoods

	Area		Number of Transactions	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Transactions	-0.441	-0.704	-0.380	-0.386
	(0.002)	(0.001)	(0.002)	(0.003)
House Transactions	-0.409	-0.707	-0.373	-0.379
	(0.003)	(0.003)	(0.002)	(0.004)
Condos Transactions	-0.366	-0.505	-0.319	-0.322
	(0.004)	(0.006)	(0.004)	(0.004)

Note: Results correspond to running the algorithm for each group of transactions with a threshold set to one standard deviations in the propensity score. The reported values correspond to the β coefficient that results from running the following regression $\log(Rank(var)) = \alpha + \beta \log(var) + \epsilon$ where var is the area or the number of transactions at the neighborhood or postal code level and $Rank(var)$ correspond to the rank of said variable across geographies. Standard errors are reported in parenthesis.

5.2 Spatial Correlation

This subsection computes the Moran (1950)'s I index of spatial correlation to test the intuition that the algorithm delineates neighborhoods that present zero spatial correlation between the location choices in a neighborhood and their immediate neighbors. More specifically, the Global Moran's I is a statistic of the degree of overall spatial autocorrelation in a spatial data set. It measures the covariance between neighborhoods and their immediate neighbors (or the "spatial lag"), relative to a random distribution. The statistic can be used as an index of the global degree of spatial clustering or dispersion in data, and is given by:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

where z_i is the deviation of the variable of interest with respect to the mean. w_{ij} is the matrix of weights that determines the neighbors j for each neighborhood i, and N is the number of neighborhoods. The matrix of weights w_{ij} is given by the inverse distance between i and all the neighborhoods. The numerator is a measure of covariance between observations and their neighbors, summed across the whole data set, and the denominator is a measure of global variance, producing a normalized index that can be tested against the null hypothesis that the variable of interest is randomly distributed across neighborhoods.[36]

Table 8: Spatial Autocorrelation (Moran's I): Industrial Neighborhoods

	Number of Firms		Average Propensity Score	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Firms	0.002	0.273	0.071	0.731
	(0.008)	(0.001)	(0.001)	(0.001)
Manufacturing	0.002	0.208	0.085	0.862
	(0.016)	(0.001)	(0.001)	(0.001)
Wholesale and Retail Trade	0.002	0.25	0.186	0.615
	(0.009)	(0.001)	(0.001)	(0.001)
Professional Services	0.003	0.335	0.279	0.655
	(0.004)	(0.001)	(0.001)	(0.001)
Entertainment, Accommodation and Food	0.007	0.469	0.265	0.817
	(0.009)	(0.001)	(0.001)	(0.001)

Note: Results correspond to running the algorithm for each group of firms with a threshold set to one standard deviations in the propensity score. The reported values correspond to the Moran's I statistic after performing an spatial union of all the cells that belong to the neighborhood and defining weights based on the inverse distance to the rest of neighborhoods. Pseudo p-value is reported in parenthesis.

[36] The index can be either positive (in the case that the standardized value Z_i of the observation is positive and the sum of its standardized neighbors Z_j is positive, or in the case that Z_i is negative and the sum of its standardized neighbors Z_j is negative—considered positive spatial association) or negative (in the case that Z_i is positive and the sum of its standardized neighbors is negative, or in the case that Z_i is negative and the sum of its standardized neighbors is positive—considered negative spatial association). The expected value is $-\frac{-1}{N-1}$ under complete spatial randomness (no positive or negative spatial autocorrelation/association).

Results support the intuition. Tables 8 and 9 shows the estimated indices and their pseudo p-value for the number of firms (or transactions) in a neighborhood and the propensity score. The algorithm is directly linked to the degree in which the propensity score is spatially correlated. A very small threshold decreases the correlation as it would create neighborhoods that are more dissimilar from each other. The opposite is also true if a large threshold is chosen. Depending on the research question, while we might choose one over the other, one thing is clear. Regardless of the threshold, the degree of spatial correlation between neighborhoods decreases substantially relative to the case of postal codes. Now, the results are even more promising when we take this exercise to the number of firms or transactions as this variable is not directly part of the algorithm process but still shows a substantial decrease in spatial correlation even reaching to a precisely estimated zero in the case of industrial neighborhoods.

Table 9: Spatial Autocorrelation (Moran's I): Residential Neighborhoods

	Number of Firms		Average Propensity Score	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Transactions	0.000	0.46	0.074	0.571
	(0.207)	(0.001)	(0.001)	(0.001)
House Transactions	-0.001	0.352	0.108	0.453
	(0.001)	(0.001)	(0.001)	(0.001)
Condo Transactions	0.000	0.473	0.158	0.766
	(0.352)	(0.001)	(0.001)	(0.001)

Note: Results correspond to running the algorithm for each group of transactions with a threshold set to one standard deviations in the propensity score. The reported values correspond to the Moran's I statistic after performing an spatial union of all the cells that belong to the neighborhood and defining weights based on the inverse distance to the rest of neighborhoods. Pseudo p-value is reported in parenthesis.

6 Conclusions

This paper is motivated by the potential biases that can come from the misalignment between economic and legal neighborhoods. Starting from that motivation, this paper proposes a revealed preference approach to delineate 'economic' neighborhood boundaries. In particular, the paper develops a machine learning algorithm that uses historical geocoded location choices of agents to identify neighborhoods as a collection of *similar-neighboring-choices*.

Part of the misalignments arises from the fact that economic neighborhoods are not necessarily constant across agents. This is in part why the algorithm is flexible enough to be adapted to different economic agents making different location choices. The paper addresses this by applying the algorithm to two different datasets and showing that in fact economic neighborhoods are different from each other and do not look like postal codes.

Finally, the paper goes back to the initial misalignment discussion and analyzes the

differences between economic and legal neighborhoods from the perspective of the biases that may arise when using one over the other. By analyzing the spatial size distribution of neighborhoods, this paper shows that economic neighborhoods are in general equal or more evenly distributed than their legal counterparts, meaning that studying things like segregation or economic concentration using legal neighborhoods may lead to higher estimates than what we observe in real life. Moreover, by analyzing the degree of spatial correlation between neighborhoods, this paper shows that economic neighborhoods are in fact more suitable for modern models of economic geography that typically assume uncorrelated shocks, and that using legally defined areas violates that assumption.

Given the recent attention to neighborhoods in the economic literature, empirical research in Urban Economics has to address the challenge of identifying the best geographical unit of analysis for measuring what constitutes a neighborhood. Such a quest in now possible due to the increasing availability of highly detailed spatial data sources that are at the choice level. However, the data is still scarce and this algorithm is not necessarily feasible for all potential research questions. This leaves the space open for further research in this area, so that we can minimize the misalignment bias and hence better guide policy makers when making decisions that need to take into account the precise scope of neighborhoods.

References

Arribas-Bel, D., Garcia-López, M. À., and Viladecans-Marsal, E. (2019). Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, page 103217.

Athey, S. and Imbens, G. (2019). Machine Learning Methods Economists Should Know About. *arXiv:1903.10075 [econ, stat]*.

Baragwanath, K., Goldblatt, R., Hanson, G., and Khandelwal, A. K. (2019). Detecting urban markets with satellite imagery: An application to India. *Journal of Urban Economics*, page 103173.

Briant, A., Combes, P. P., and Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, 67(3):287–302.

Ch, R., Martin, D. A., and Vargas, J. F. (2020). Measuring the size and growth of cities using nighttime light. *Journal of Urban Economics*, page 103254.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24(2):295–313.

de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillon, L., and Gorin, C. (2019). Delineating urban areas using building density. *Journal of Urban Economics*, page 103226.

Dingel, J. I. and Tintelnot, F. (2020). Spatial Economics for Granular Settings. Working Paper 27287, National Bureau of Economic Research.

Duranton, G. (2015). Delineating Metropolitan Areas: Measuring Spatial Labour Market Networks Through Commuting Patterns. In Watanabe, T., Uesugi, I., and Ono, A., editors, *The Economics of Interfirm Networks*, Advances in Japanese Business and Economics, pages 107–133. Springer Japan, Tokyo.

Duranton, G. and Puga, D. (2014). Chapter 5 - The Growth of Cities. In Aghion, P. and Durlauf, S. N., editors, *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, pages 781–853. Elsevier.

Eaton, J. and Kortum, S. (2002). Technology, Geography, and Trade. *Econometrica*, 70(5):1741–1779.

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics*, 114(3):739–767.

Gabaix, X. and Ibragimov, R. (2011). Rank - 1 / 2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.

Galdo, V., Li, Y., and Rama, M. (2019). Identifying urban areas by combining human judgment and machine learning: An application to India. *Journal of Urban Economics*, page 103229.

Gibbons, S., Overman, H. G., and Patacchini, E. (2015). Spatial Methods. In Duranton, G., Henderson, J. V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics*, volume 5, pages 115–168. Elsevier.

Guimarães, P., Figueirdo, O., and Woodward, D. (2003). A Tractable Approach to the Firm Location Decision Problem. *Review of Economics and Statistics*, 85(1):201–204.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Henderson, J. V., Nigmatulina, D., and Kriticos, S. (2019). Measuring urban economic density. *Journal of Urban Economics*, page 103188.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Imbens, G. W. (2015). Matching Methods in Practice: Three Examples. *Journal of Human Resources*, 50(2):373–419.

Lind, N. and Ramondo, N. (2020). Trade with Correlation. Working Paper 24380, National Bureau of Economic Research.

Lomax, N., Malleson, N., and Kieu, L.-M. (2019). Point Pattern Processes and Models. *arXiv:1910.00282 [stat]*.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., and Thirion, B. (2012). A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45(6):2041–2049.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Moreno-Monroy, A. I., Schiavina, M., and Veneri, P. (2020). Metropolitan areas in the world. Delineation and population trends. *Journal of Urban Economics*, page 103242.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, 57(1):86–100.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Geo Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Rey, S. J. and Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1):5–27.

Rosen, K. T. and Resnick, M. (1980). The size distribution of cities: An examination of the Pareto law and primacy. *Journal of Urban Economics*, 8(2):165–186.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524.

Rozenfeld, H. D., Rybski, D., Gabaix, X., and Makse, H. A. (2011). The Area and Population of Cities: New Insights from a Different Perspective on Cities. *American Economic Review*, 101(5):2205–2225.

Topa, G. and Zenou, Y. (2015). Neighborhood and Network Effects. In Duranton, G., Henderson, J. V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics*, volume 5, pages 561–624. Elsevier.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

A Appendix

Figure A.1: Neighborhoods when the clustered case results in 50 neighborhoods

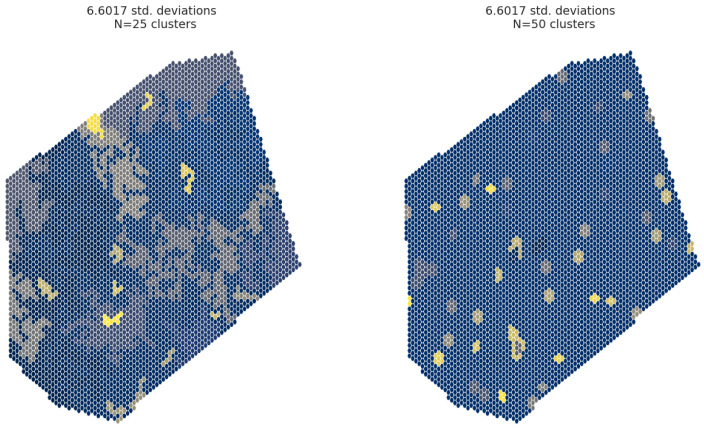
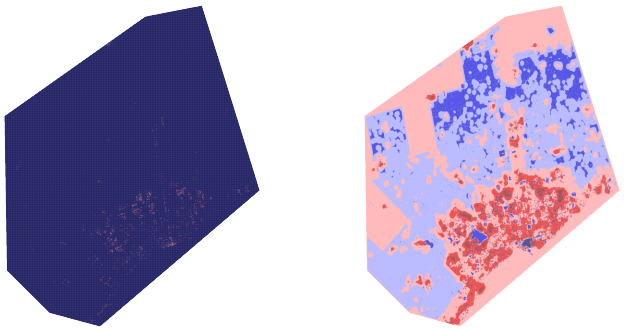
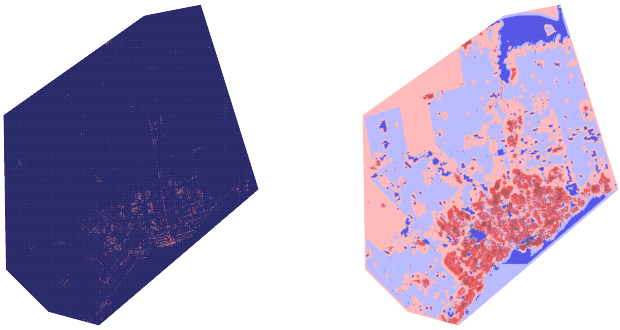


Figure A.2: Distribution of Raw vs Propensity Score Probabilities: Industrial Neighborhoods

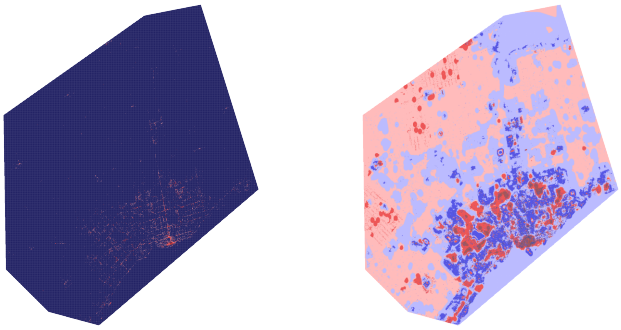
(a) Manufacturing



(b) Wholesale and Retail Trade



(c) Professional Services



(d) Entertainment, Accomodation and Food Services

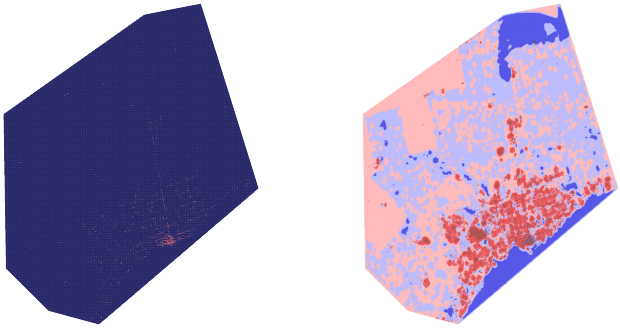
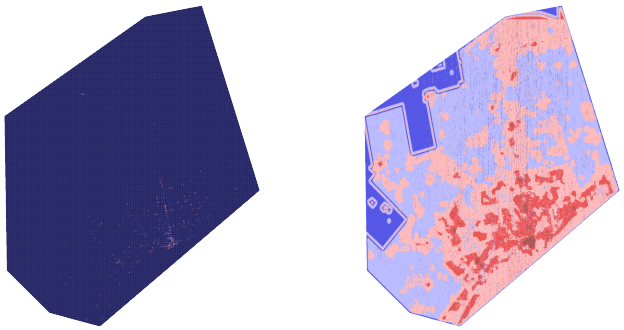


Figure A.4: Distribution of Raw vs Propensity Score Probabilities: Residential Neighborhoods

(a) Condos



(b) Houses

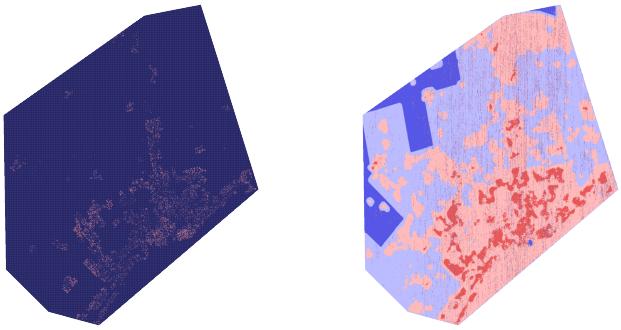


Figure A.3: Distribution of Propensity Score At Neighborhood Level: Industrial Neighborhoods

Figure A.5: Distribution of Propensity Score At Neighborhood Level: Residential Neighborhoods

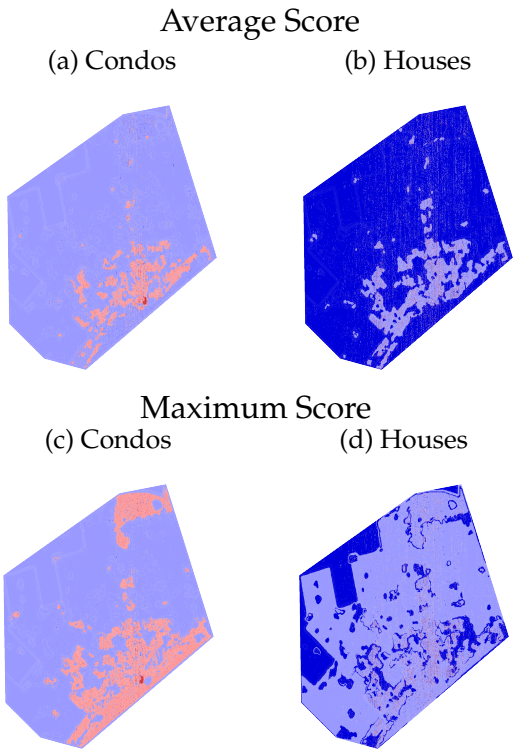


Table A.1: All Results Firm Locations

	Threshold (std.dev. score)	Number of Firms	Number of Neighbor-hoods	Number of Singletons	% Firms in Neighbor-hoods	Area (sq. km)	Length (km)	Width (km)	Length/Width	Zipf's Beta # Firms	Zipf's Beta Area	Moran's I Avg. Score	Moran's I # Firms
All Firms	four	125,435	3,389	455	98.65	3.919	1.945	1.060	1.803	-0.640	-0.485	0.060	0.001
						(73.468)	(4.875)	(2.552)	(0.704)	(0.002)	(0.003)	(0.001)	(0.122)
	two	125,435	7,288	3,228	95.07	1.433	1.133	0.603	1.719	-0.650	-0.599	0.077	0.003
						(41.243)	(3.242)	(1.524)	(0.746)	(0.001)	(0.001)	(0.001)	(0.015)
	one	125,435	13,947	7,198	91.03	0.713	0.838	0.431	1.764	-0.722	-0.736	0.071	0.002
						(20.945)	(2.411)	(1.003)	(0.891)	(0.001)	(0.001)	(0.001)	(0.008)
	half	125,435	22,590	16,015	84.57	0.390	0.638	0.325	1.748	-0.767	-1.152	0.121	0.002
						(14.051)	(1.941)	(0.707)	(0.993)	(0.001)	(0.002)	(0.001)	(0.004)
	tenth	125,435	42,865	61,154	72.5	0.145	0.417	0.220	1.656	-0.814	-1.773	0.185	0.003
						(5.443)	(1.230)	(0.407)	(1.168)	(0.001)	(0.005)	(0.001)	(0.002)
	centile	125,435	50,424	190,647	61.11	0.063	0.283	0.176	1.483	-0.841	-2.304	0.699	0.042
						(2.355)	(0.549)	(0.219)	(0.929)	(0.001)	(0.011)	(0.001)	(0.001)
	four	17,661	1,844	799	95.44	5.700	1.765	0.953	1.736	-0.695	-0.411	0.091	0.005
						(81.158)	(6.622)	(3.443)	(0.741)	(0.006)	(0.003)	(0.001)	(0.037)
	two	17,661	3,435	2,022	90.23	2.761	1.391	0.715	1.776	-0.787	-0.445	0.084	0.003
						(43.991)	(5.213)	(2.536)	(0.881)	(0.005)	(0.002)	(0.001)	(0.037)
	one	17,661	6,130	4,138	82.35	1.467	1.099	0.548	1.807	-0.894	-0.528	0.085	0.002
						(29.642)	(3.967)	(1.726)	(0.954)	(0.004)	(0.002)	(0.001)	(0.016)
Manufacturing	half	17,661	10,152	7,543	66.83	0.851	0.894	0.440	1.804	-0.947	-0.739	0.114	0.002
						(20.781)	(3.048)	(1.270)	(0.969)	(0.003)	(0.003)	(0.001)	(0.010)
	tenth	17,661	23,809	28,521	51.24	0.288	0.592	0.290	1.729	-1.055	-1.802	0.169	0.002
						(9.529)	(1.855)	(0.646)	(1.381)	(0.002)	(0.009)	(0.001)	(0.001)
	centile	17,661	51,037	114,701	33.12	0.091	0.344	0.195	1.604	-1.177	-2.585	0.835	0.005
						(3.449)	(0.689)	(0.284)	(1.283)	(0.002)	(0.018)	(0.001)	(0.003)
	four	44,846	2,464	432	96.49	5.202	2.008	1.099	1.816	-0.663	-0.455	0.127	0.000
						(96.300)	(5.541)	(2.783)	(0.730)	(0.004)	(0.004)	(0.001)	(0.353)
	two	44,846	5,381	1,641	91.1	2.146	1.313	0.699	1.809	-0.725	-0.526	0.136	0.001
						(52.892)	(3.795)	(1.842)	(0.775)	(0.003)	(0.002)	(0.001)	(0.047)
	one	44,846	10,187	4,096	84.58	1.055	0.966	0.505	1.813	-0.771	-0.639	0.186	0.002
						(24.591)	(2.661)	(1.255)	(0.808)	(0.002)	(0.001)	(0.001)	(0.009)
Wholesale and Retail Trade	half	44,846	17,695	9,334	77.69	0.557	0.717	0.369	1.797	-0.835	-0.790	0.210	0.002
						(16.995)	(2.017)	(0.861)	(0.891)	(0.002)	(0.002)	(0.001)	(0.005)
	tenth	44,846	36,456	48,323	64.64	0.178	0.427	0.231	1.629	-0.910	-1.583	0.255	0.003
						(8.099)	(1.238)	(0.440)	(0.961)	(0.001)	(0.006)	(0.001)	(0.003)
	centile	44,846	61,968	175,275	46.85	0.064	0.299	0.181	1.531	-0.951	-2.269	0.589	0.005
						(2.130)	(0.526)	(0.212)	(1.084)	(0.001)	(0.011)	(0.001)	(0.002)
	four	47,464	3,029	1,237	93.88	3.532	1.333	0.741	1.687	-0.721	-0.526	0.339	0.002
						(52.780)	(4.901)	(2.245)	(1.009)	(0.004)	(0.002)	(0.001)	(0.054)
	two	47,464	4,879	6,702	77.09	1.301	0.800	0.460	1.513	-0.823	-1.400	0.296	0.003
						(28.233)	(3.126)	(1.425)	(1.747)	(0.001)	(0.005)	(0.001)	(0.009)
	one	47,464	7,085	8,042	76.62	0.996	0.867	0.467	1.620	-0.822	-1.561	0.279	0.003
						(23.151)	(3.155)	(1.265)	(2.978)	(0.001)	(0.006)	(0.001)	(0.004)
Professional Services	half	47,464	11,032	10,538	75.7	0.698	0.855	0.442	1.707	-0.823	-1.590	0.236	0.002
						(17.154)	(2.742)	(1.029)	(2.671)	(0.001)	(0.006)	(0.001)	(0.006)
	tenth	47,464	36,114	28,110	60.99	0.235	0.588	0.292	1.834	-0.875	-1.843	0.238	0.003
						(6.079)	(1.528)	(0.542)	(2.318)	(0.001)	(0.007)	(0.001)	(0.001)
	centile	47,464	62,130	193,323	47.97	0.059	0.289	0.177	1.525	-0.920	-2.462	0.526	0.012
						(1.764)	(0.457)	(0.201)	(0.932)	(0.001)	(0.013)	(0.001)	(0.002)
	four	15,464	1,255	430	96.57	8.941	1.722	0.983	1.730	-0.578	-0.349	0.214	0.000
						(224.206)	(6.507)	(3.611)	(0.734)	(0.010)	(0.006)	(0.001)	(0.251)
	two	15,464	2,608	1,162	89.12	3.996	1.287	0.707	1.729	-0.710	-0.394	0.239	0.002
						(136.396)	(4.849)	(2.435)	(0.770)	(0.008)	(0.004)	(0.001)	(0.100)
	one	15,464	5,079	3,207	79.91	1.818	0.964	0.526	1.695	-0.875	-0.494	0.265	0.007
						(58.800)	(3.499)	(1.733)	(1.018)	(0.005)	(0.002)	(0.001)	(0.009)
Entertainment, Accommodation and Food	half	15,464	8,315	7,081	65.31	0.979	0.760	0.414	1.656	-0.936	-0.882	0.247	0.008
						(37.507)	(2.652)	(1.308)	(0.933)	(0.003)	(0.004)	(0.001)	(0.001)
	tenth	15,464	22,403	25,191	57.51	0.317	0.536	0.284	1.658	-0.958	-1.737	0.289	0.007
						(11.169)	(1.765)	(0.662)	(1.274)	(0.003)	(0.010)	(0.001)	(0.001)
	centile	15,464	60,020	92,234	44.1	0.099	0.362	0.201	1.662	-1.051	-2.173	0.665	0.023
						(3.562)	(0.653)	(0.302)	(1.064)	(0.003)	(0.015)	(0.001)	(0.002)

Note: Results correspond to running the algorithm for each group of firms across different threshold of standard deviations in the propensity score. % Firms in neighborhoods correspond to the percentage of firms that belong to neighborhoods that have at least two cells. Length (and width) correspond to the longest (and shortest) side of the minimum bounding rectangle that contains the neighborhood. Morgan's I statistics are obtained after performing an spatial union of all the cells that belong to the neighborhood and defining weights based on immediate contiguity. Zipf's β correspond to the coefficient that results from running the following regression $\log(Rank(var)) = \alpha + \beta \log(var) + \epsilon$ where var is the average propensity score or the number of transactions at the neighborhood or postal code level and $Rank(var)$ correspond to the rank of said variable across geographies. Standard errors are reported in parenthesis of all values but the Morgan's I where pseudo p-value are reported in parenthesis instead.

Table A.2: All Results Location of Housing Transactions

	Threshold (std.dev. score)	Number of Transac-tions	Number of Neighbor-hoods	Number of Singletons	% Transactions in Neighbor-hoods	Area (sq. km)	Length (km)	Width (km)	Length/Width	Zipf's Beta # Firms	Zipf's Beta Area	Moran's I Avg. Score	Moran's I # Firms
All Transactions	four	68,184	5,121	8,045	97.820	1.144	0.568	0.296	1.748	-0.254	-0.304	0.217	-0.001
						(42.199)	(3.377)	(1.533)	(1.174)	(0.008)	(0.006)	(0.001)	(0.012)
	two	68,184	6,921	15,068	89.570	0.685	0.528	0.284	1.633	-0.169	-0.441	0.074	0.000
						(19.660)	(2.572)	(1.164)	(1.081)	(0.003)	(0.002)	(0.001)	(0.207)
	one	68,184	9,233	25,317	57.110	0.436	0.493	0.265	1.573	-0.102	-0.573	0.094	0.002
						(14.384)	(1.994)	(0.891)	(1.222)	(0.001)	(0.002)	(0.001)	(0.018)
	half	68,184	10,523	34,258	47.860	0.336	0.503	0.267	1.552	-0.081	-0.619	0.098	0.003
						(11.418)	(1.693)	(0.736)	(1.307)	(0.001)	(0.002)	(0.001)	(0.007)
	tenth	68,184	25,429	52,476	28.810	0.193	0.528	0.260	1.752	-0.078	-0.982	0.104	0.016
						(5.031)	(1.099)	(0.493)	(1.493)	(0.001)	(0.003)	(0.001)	(0.001)
	centile	68,184	68,703	171,551	16.360	0.063	0.301	0.181	1.543	-0.080	-1.779	0.191	0.028
						(1.734)	(0.462)	(0.206)	(0.843)	(0.001)	(0.012)	(0.001)	(0.001)
House Transactions	four	48,510	5,224	8,308	98.810	1.113	0.542	0.285	1.744	-0.284	-0.284	0.193	-0.001
						(27.809)	(3.278)	(1.415)	(1.055)	(0.011)	(0.008)	(0.001)	(0.039)
	two	48,510	6,834	15,623	92.840	0.671	0.495	0.274	1.613	-0.192	-0.409	0.108	-0.001
						(19.535)	(2.413)	(1.119)	(0.947)	(0.004)	(0.003)	(0.001)	(0.001)
	one	48,510	9,052	23,715	72.690	0.460	0.490	0.267	1.572	-0.107	-0.548	0.080	-0.001
						(15.072)	(2.034)	(0.891)	(0.973)	(0.001)	(0.002)	(0.001)	(0.001)
	half	48,510	10,277	33,643	49.540	0.343	0.493	0.268	1.519	-0.082	-0.648	0.095	0.001
						(11.883)	(1.719)	(0.749)	(1.023)	(0.001)	(0.002)	(0.001)	(0.105)
	tenth	48,510	25,482	49,009	26.590	0.202	0.529	0.262	1.721	-0.082	-1.095	0.066	0.020
						(6.535)	(1.185)	(0.508)	(1.475)	(0.001)	(0.004)	(0.001)	(0.001)
	centile	48,510	68,777	159,268	11.770	0.066	0.310	0.182	1.580	-0.078	-1.966	0.117	0.043
						(1.962)	(0.488)	(0.224)	(0.921)	(0.001)	(0.014)	(0.001)	(0.001)
Condo Transactions	four	19,674	2,038	1,657	93.570	4.077	1.766	0.848	2.239	-0.314	-0.298	0.126	-0.002
						(52.017)	(6.326)	(2.919)	(1.927)	(0.013)	(0.008)	(0.001)	(0.043)
	two	19,674	5,146	5,305	77.730	1.442	1.062	0.497	2.049	-0.237	-0.366	0.158	0.000
						(27.742)	(3.694)	(1.608)	(1.845)	(0.008)	(0.004)	(0.001)	(0.352)
	one	19,674	9,536	10,691	69.220	0.745	0.858	0.394	2.019	-0.207	-0.494	0.183	0.007
						(17.698)	(2.525)	(1.105)	(1.910)	(0.007)	(0.004)	(0.001)	(0.002)
	half	19,674	15,337	20,539	45.850	0.420	0.694	0.326	1.910	-0.184	-0.936	0.226	0.022
						(12.054)	(1.796)	(0.794)	(1.772)	(0.005)	(0.011)	(0.001)	(0.001)
	tenth	19,674	40,195	51,668	36.220	0.164	0.508	0.248	1.835	-0.182	-1.919	0.259	0.017
						(4.326)	(0.935)	(0.415)	(1.401)	(0.005)	(0.027)	(0.001)	(0.001)
	centile	19,674	66,870	223,926	29.420	0.052	0.269	0.173	1.461	-0.183	-3.033	0.366	0.013
						(1.496)	(0.402)	(0.178)	(0.715)	(0.005)	(0.047)	(0.001)	(0.001)

Note: Results correspond to running the algorithm for each group of type of transactions across different threshold of standard deviations in the propensity score. % Transactions in neighborhoods corresponds to the percentage of transactions that belong to neighborhoods that have at least two cells. Length (and width) correspond to the longest (and shortest) side of the minimum bounding rectangle that contains the neighborhood. Morgan's I statistics are obtained after performing an spatial union of all the cells that belong to the neighborhood and defining weights based on immediate contiguity. Zipf's β correspond to the coefficient that results from running the following regression $\log(\textit{Rank}(\textit{var})) = \alpha + \beta \log(\textit{var}) + \epsilon$ where *var* is the average propensity score or the number of transactions at the neighborhood or postal code level and $\textit{Rank}(\textit{var})$ correspond to the rank of said variable across geographies. Standard errors are reported in parenthesis of all values but the Morgan's I where pseudo p-value are reported in parenthesis instead.

Table A.3: Industrial Propensity Score

	n_firms	NAICS_3	NAICS_4	NAICS_5	NAICS_7
area_Parks	-0.17***	0.02	-0.28***	-0.03**	-0.26***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)
area_Residential	0.20***	-0.04***	0.22***	-0.06***	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
area_Industrial	0.18***	0.04***	0.11***	-0.00	0.06***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
area_Commercial	-0.08***	0.17***	-0.00	-0.01**	0.11***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
area_Government	-0.24***	-0.16***	-0.26***	0.08***	0.10***
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
poi_POST	0.05***	0.09***	0.09***	0.02***	-0.16***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
poi_TOUR	-0.02***	-0.10***	-0.05***	-0.02***	0.15***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
poi_BANK	0.08***	0.06***	0.04***	0.09***	0.10***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
poi_RESTA	0.01***	-0.06***	-0.00	0.01***	0.04***
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
poi_HOTEL	-0.11***	-0.01**	-0.06***	-0.10***	-0.11***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
MA_1km_area_Parks	-0.26***	-0.06***	-0.03**	-0.07***	-0.10***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)
MA_1km_area_Open	-0.23***	-0.36***	-0.27***	-0.03***	-0.09***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_area_Residential	0.38***	0.19***	0.12***	-0.07***	-0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
MA_1km_area_Industrial	0.23***	0.43***	0.32***	0.06***	-0.13***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_area_Commercial	-0.10***	-0.09***	0.23***	0.03***	0.11***
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
MA_1km_area_Government	-0.02***	-0.01	-0.05***	-0.09***	0.13***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_poi_POST	0.10***	-0.07***	-0.12***	0.07***	-0.04***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_poi_TOUR	0.09***	0.08***	0.46***	-0.02*	-0.08***
	(0.00)	(0.02)	(0.01)	(0.01)	(0.02)
MA_1km_poi_BANK	-0.12***	-0.01	-0.35***	-0.09***	0.01
	(0.00)	(0.02)	(0.01)	(0.01)	(0.01)
MA_1km_poi_RESTA	0.06***	-0.26***	-0.18***	0.09***	0.06***
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
MA_1km_poi_HOTEL	-0.04***	0.20***	-0.09***	0.06***	-0.12***
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
stock_houses_DMTI	-0.06***	0.04***	-0.02***	0.05***	0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Down		0.01	-0.02	-0.02	0.13
		(0.12)	(0.09)	(0.10)	(0.09)
Up		-0.01	0.26***	0.04	0.12
		(0.14)	(0.09)	(0.10)	(0.10)
const	-2.46***	-4.34***	-3.52***	-2.96***	-4.24***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	740,286	740,286	740,286	740,286	740,286
Pseudo R-squared	0.33	0.28	0.31	0.25	0.24

Table A.4: Residential Propensity Score

	n_sales_MLS	n_houses_MLS	n_condos_MLS
area_Parks	-0.21***	-0.01	-0.94***
	(0.04)	(0.03)	(0.07)
area_Open	-0.28***	-0.01	-1.24***
	(0.05)	(0.05)	(0.09)
area_Residential	-0.18***	0.04	-0.84***
	(0.03)	(0.03)	(0.06)
area_Industrial	-0.12***	0.01	-0.59***
	(0.02)	(0.02)	(0.04)
area_Commercial	-0.04***	0.00	-0.18***
	(0.01)	(0.01)	(0.01)
area_Government	-0.08***	-0.01	-0.28***
	(0.01)	(0.01)	(0.02)
poi_POST	0.01	0.01	-0.00
	(0.01)	(0.01)	(0.01)
poi_TOUR	-0.07***	0.01	0.03
	(0.01)	(0.01)	(0.02)
poi_BANK	-0.04***	-0.01	-0.08***
	(0.01)	(0.01)	(0.01)
poi_RESTA	-0.01***	-0.03***	0.00
	(0.00)	(0.01)	(0.00)
poi_HOTEL	0.00	-0.01	0.00
	(0.00)	(0.01)	(0.00)
MA_1km_area_Parks	0.78***	0.34***	1.32***
	(0.03)	(0.03)	(0.06)
MA_1km_area_Open	0.93***	0.31***	1.38***
	(0.04)	(0.04)	(0.09)
MA_1km_area_Residential	0.74***	0.43***	1.14***
	(0.02)	(0.02)	(0.05)
MA_1km_area_Industrial	0.29***	0.12***	0.53***
	(0.01)	(0.01)	(0.03)
MA_1km_area_Commercial	0.08***	0.03***	0.13***
	(0.00)	(0.01)	(0.01)
MA_1km_area_Government	0.17***	0.06***	0.29***
	(0.01)	(0.01)	(0.01)
MA_1km_poi_POST	-0.07***	0.02**	-0.05***
	(0.01)	(0.01)	(0.01)
MA_1km_poi_TOUR	-0.05***	0.00	-0.09***
	(0.01)	(0.01)	(0.01)
MA_1km_poi_BANK	0.10***	-0.08***	0.12***
	(0.01)	(0.01)	(0.01)
MA_1km_poi_RESTA	0.05***	0.02***	0.04***
	(0.00)	(0.01)	(0.00)
MA_1km_poi_HOTEL	0.02***	-0.01	0.05***
	(0.00)	(0.01)	(0.01)
mean_rooms_MLS	0.40***	0.41***	0.53***
	(0.00)	(0.00)	(0.00)
sum_rooms_MLS	0.07***	0.06***	0.11***
	(0.00)	(0.00)	(0.00)
avg_lotsize_MLS	-0.15***	0.00***	-98.21***
	(0.02)	(0.00)	(1.13)
stock_houses_DMTI	0.02***	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
const	-2.92***	-3.17***	-5.23***
	(0.01)	(0.01)	(0.02)
Observations	740286	740286	740286
Pseudo R-squared	0.37	0.34	0.49