Delineating Neighborhoods using Location Choices

Rolando Campusano

October 9, 2020

Motivation

- Neighborhoods have become an important focus of economic policy and research *(e.g. mobility - competition - gentrification - sprawl - segregation - agglomeration)*

Motivation

- Neighborhoods have become an important focus of economic policy and research *(e.g. mobility - competition - gentrification - sprawl - segregation - agglomeration)*
- Conceptually, a neighborhood is a geographically localized community where members engage in face-to-face social interactions

Motivation

- Neighborhoods have become an important focus of economic policy and research *(e.g. mobility - competition - gentrification - sprawl - segregation - agglomeration)*
- Conceptually, a neighborhood is a geographically localized community where members engage in face-to-face social interactions
- However, neighborhoods have been typically defined using 'official' boundaries that do not necessarily represent the level at which those interactions occur
  - Census blocks (or tracts)
  - Postal (ZIP) codes
  - School/political districts
- Which means that there might be miss-alignment and/or overlapping between 'official' and 'economic' boundaries

Motivation

- These miss-alignment and/or overlapping between 'official' and 'economic' boundaries may lead to a variety of biases
  - Measurement error
  - Modifiable areal unit problem (Briant et al., 2010)
  - Spatial correlation across neighborhoods (Lind and Ramondo, 2018)
- Do not have clear solutions without imposing more structure and assumptions (Gibbons et al., 2015; Lind and Ramondo, 2018)
- This paper provides an alternative solution
  - use historical geocoded location choices of agents to
  - identify neighborhoods as a collection of *similar-neighboring-choices*
  - provide a machine-learning algorithm to obtain their boundaries

This paper: algorithm

- Data is composed by agents location choices at very-fine geographies that can (ideally) be allocated to an arbitrarily defined grid using geocodes
- Each grid is associated with a series of choices that can be summarized in a *propensity score* as a function of agents and grid characteristics
- Boundaries are delineated using a bottom-up machine learning agglomerative algorithm with *adjacency constraints* in the spirit of Ward (1963) and Rozenfeld et al. (2011)
- Neighborhoods are composed by cells with very similar propensity score (like Rosenbaum and Rubin (1984)'s propensity score strata)

This paper: economic neighborhoods

- Using economic decisions to delineate neighborhoods => 'economic' neighborhoods
- Adjacency constraints => compact and statistically distinct neighborhoods but not necessarily statistically different from neighborhoods in other parts of town
- Computing economic neighborhoods helps address the main issues with current definitions
  1. decreases the relevance of the modifiable area unit problem by identifying the spatial choice set based in economic decisions (Briant et al., 2010).
  2. less granularity and more symmetric interactions by agglomerating cells with very similar propensities (Topa and Zenou, 2015; Dingel and Tintelnot, 2020).
  3. zero spatial correlation between neighborhoods and their immediate neighbors by creating neighborhoods that are distinct from their immediate neighbor (Gibbons et al., 2015; Lind and Ramondo, 2018)

This paper: applications

- This algorithm can be applied to identify neighborhoods in any application that involves an underlying location choice
- The main requirement of this algorithm is a dataset with geocoded location choices
- Fortunately, the availability of these datasets has been increasing over time
- Two example applications are analyzed for the city of Toronto
  - Industrial neighborhoods using a points of interest dataset
  - Residential neighborhoods using a dataset of real-estate transactions

This paper: Preview of Results

- Simulation exercises show that
  - meaningful neighborhoods need highly skewed distributions of economic activity
  - neighborhoods can be too big (very large threshold) or too small (very small threshold)
  - however, under highly skewed distributions the algorithm delivers more stable sizes
  - visual inspection is crucial to asses the quality of the neighborhoods

This paper: Preview of Results

- Simulation exercises show that
  - meaningful neighborhoods need highly skewed distributions of economic activity
  - neighborhoods can be too big (very large threshold) or too small (very small threshold)
  - however, under highly skewed distributions the algorithm delivers more stable sizes
  - visual inspection is crucial to asses the quality of the neighborhoods
- Applying the method to two datasets for the GTA shows that industrial and housing neighborhoods
  - do not look like postal codes, specially industrial neighborhoods
  - are larger (even with small thresholds) and mostly located around big streets
  - are different in shape and characteristics across industries and housing types
  - compared to postal codes, they
    - have (way) less spatial correlation
    - follow a power law (Zipf's law)

Outline

Methodology

- The main idea is very similar in spirit to Rosenbaum and Rubin (1984)'s propensity score stratification
- Machine learning clustering methods provide a way to incorporate the spatial aspect
  - As they have the goal of grouping a collection of objects into subsets or "clusters"
  - Using some sort of definition of similarity (or dissimilarity) provided by the researcher
  - Similarity in attributes and similarity in location
- What algorithm?
  - k-means algorithms is fast and known but requires to know the number of clusters a-priori and does not produce compact clusters (Athey and Imbens, 2019)
  - hierarchical algorithms creates all 'possible' clusters and produces compact clusters but requires bilateral similarity matrices and a pre-defined threshold

Methodology

- Hierarchical algorithms are bottom-up (Rozenfeld et al., 2011; de Bellefon et al., 2019; Arribas-Bel et al., 2019)
- But it requires a similarity matrix to compute all possible clusters => unfeasible as is
- Two important modifications makes it suitable for this setting:
  - adjacency constraints and the use of sparse similarity matrices reduce computational burden (assumption on network structure)
  - use propensity score as a clustering feature (assumption of statistical sufficiency)
- Caveat: we need to define a threshold (and in our case also an specification for the propensity score)

Methodology - Algorithm

What is $\delta(N_u, N_v)$??

For a given $\mathcal{P} = \{\mathrm{P}_t^u\}_{u=1}^{\mathbf{B}}$ set of all block-level probabilities to be clustered.

1. Initialize with set of neighborhoods to be $\{N_1, ..., N_B\}$ where $N_u = \{\mathrm{P}_t^u\}$ for all $u = 1, ..., B$

Methodology - Algorithm

What is $\delta(N_u, N_v)$??

For a given $\mathcal{P} = \{\mathrm{P}_t^u\}_{u=1}^{\mathbf{B}}$ set of all block-level probabilities to be clustered.

1. Initialize with set of neighborhoods to be $\{N_1, ..., N_B\}$ where $N_u = \{\mathrm{P}_t^u\}$ for all $u = 1, ..., B$
2. Compute the dissimilarity between all pairs, that is, compute $\delta(N_u, N_v)$ for all $u < v \in \{adjacent_u\}$

For a given $\mathcal{P} = \{\mathrm{P}_t^u\}_{u=1}^{\mathbf{B}}$ set of all block-level probabilities to be clustered.

1. Initialize with set of neighborhoods to be $\{N_1, \ldots, N_B\}$ where $N_u = \{\mathrm{P}_t^u\}$ for all $u = 1, \ldots, B$
2. Compute the dissimilarity between all pairs, that is, compute $\delta(N_u, N_v)$ for all $u < v \in \{adjacent_u\}$
3. While there is more than one neighborhoods in the original set:
    1. Merge a pair which have minimal dissimilarity

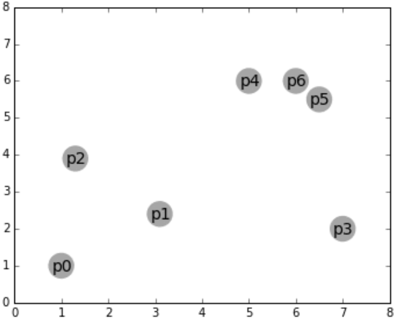
    $$\delta(N_{u'}, N_{v'}) = \min_{u'<v'} \delta(N_u, N_v)$$

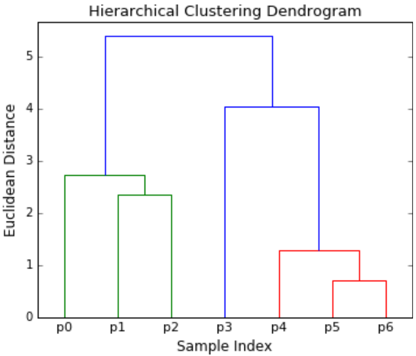
    set $N_{u'} = N_{u'} \cup N_{v'}$ and remove $N_{v'}$ from the set of neighborhoods
    2. Compute dissimilarity between $N_{u'}$ and the remaining neighborhoods in original set

For a given $\mathcal{P} = \{\mathrm{P}_t^u\}_{u=1}^{\mathbf{B}}$ set of all block-level probabilities to be clustered.

1. Initialize with set of neighborhoods to be $\{N_1, ..., N_B\}$ where $N_u = \{\mathrm{P}_t^u\}$ for all $u = 1, ..., B$
2. Compute the dissimilarity between all pairs, that is, compute $\delta(N_u, N_v)$ for all $u < v \in \{adjacent_u\}$
3. While there is more than one neighborhoods in the original set:
   1. Merge a pair which have minimal dissimilarity
   $$\delta(N_{u'}, N_{v'}) = \min_{u' < v'} \delta(N_u, N_v)$$
   set $N_{u'} = N_{u'} \cup N_{v'}$ and remove $N_{v'}$ from the set of neighborhoods
   2. Compute dissimilarity between $N_{u'}$ and the remaining neighborhoods in original set
4. The final set of neighborhoods $\{N\}$ is defined as the subset of $\mathcal{P}$ in which the variance is minimized given that the dissimilarity between **all** neighborhoods is below a given threshold $\bar{\delta}$.

Methodology - Illustration





Outline

Simulations

- Besides geo-coded data of location choices, the algorithm requires two inputs from the researcher
  - it requires an specification for the propensity score that leads to cell specific probabilities
  - it requires a predetermined threshold $\overline{\delta}$
- Now, even though this algorithm is feasible, it is still computationally intensive specially in very local contexts
- This means that even though cross-validation of the threshold is possible, it is also very limited
- I simulate two cities, each with a different distribution of economic activity, to study how the algorithm behaves under different thresholds

Simulations - Two Cities

- 100,000 agents location choices under two different distribution of economic activity
  1. a poisson random point process
  2. a poisson cluster point process with points clustered around 50 sub centers
- points are then aggregated at the grid cell level Kernel Density

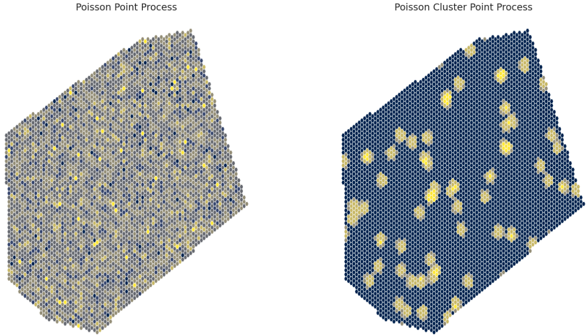


Figure: Two Cities Simulated Data

Simulations - Two Cities Neighborhoods

And run the algorithm for a given threshold obtaining this

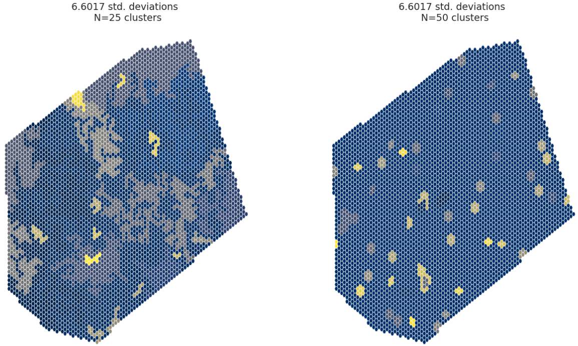
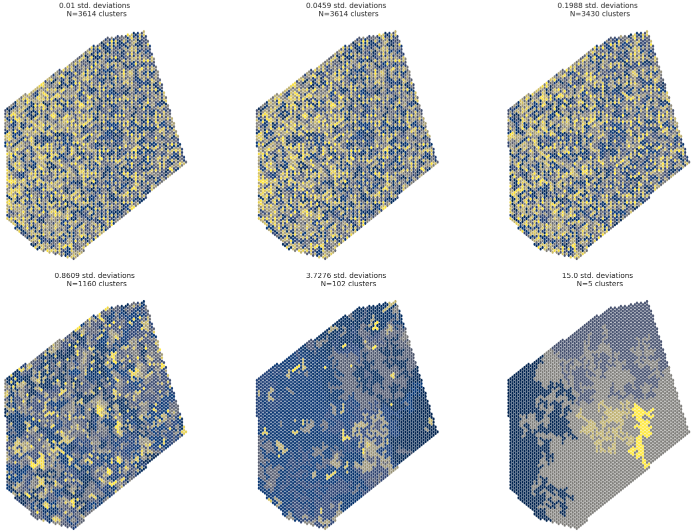
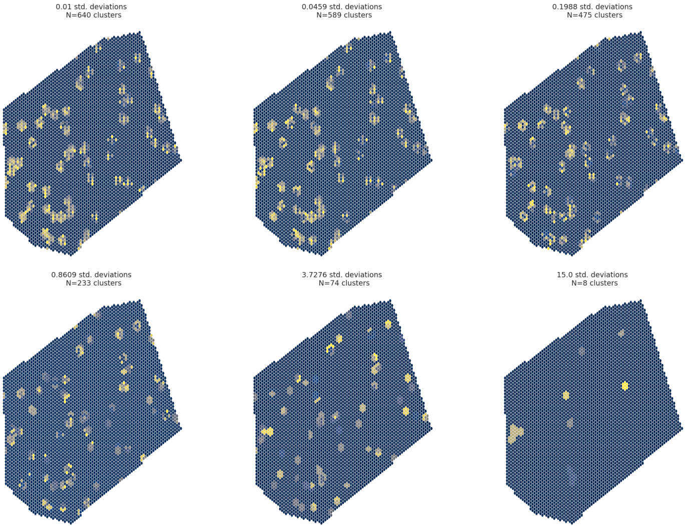


Figure: Two Cities Neighborhoods

Simulations - Results for Random City



Simulations - Results for Cluster City



Simulations - Threshold

- A threshold too small produces too many neighborhoods => no between variance
- A threshold too large produces too few neighborhoods => no within variance
- Skewed distributions are less sensitive to this

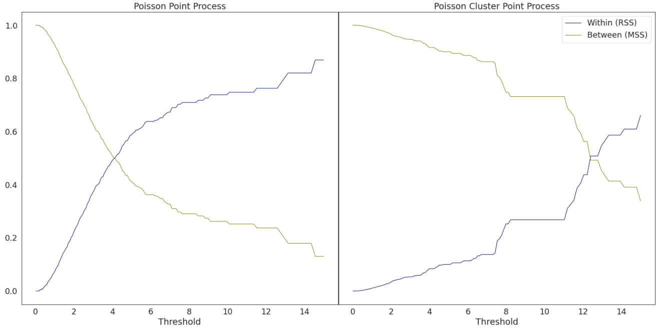


Figure: Variance Decomposition Between/Within Neighborhoods

Simulations - Conclusions

- In order to obtain meaningful and stable neighborhoods we need an uneven distribution of location choices
- This will not only make the algorithm obtain clearer differences between neighborhoods
- But it will also make the algorithm less sensitive to the definition of the threshold and hence producing neighborhoods that are more stable across thresholds

Outline

Applications

- Two applications using location choice data and then compare the results with postal codes
  - Firm Location Choices
  - Housing Transactions
- Choices are assigned to an hexagon grid cell with 75m side
- The algorithm is ran for 6 different thresholds defined as $(0.01, 0.1, 0.5, 1, 2, 4) \times Std.Dev.$ of the propensity score
- All results showing today are with the threshold set at the smallest value that produces neighborhoods grouping at least 90% of the choices

Outline

Industrial Neighborhoods - Setup

- The propensity score captures the probability that a given firm chooses grid cell i based on its characteristics

$$\#\text{Firms}_i = \beta + \sum_{POI} \beta_{POI}\text{POI}_i + \sum_{POI} \beta_{\text{MA}POI}\text{MA\_POI}_i + \sum_{POI} \beta_{LAND}\text{LAND}_i + \sum_{LAND} \beta_{\text{MA}LAND}\text{MA\_LAND}_i +$$
$$\beta_{UP}\#Up + \beta_{\text{MA}UP}\text{MA\_}\#Up + \beta_{DOWN}\#Down + \beta_{\text{MA}Down}\text{MA\_}\#Down + \epsilon_i$$

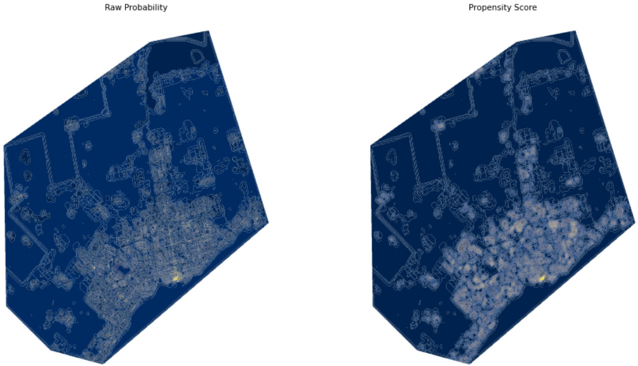
- Dataset: DMTI Spatial Inc. CanMap POI dataset (at street address level)
  - Amenities (Hotels, Schools, Banks, Hospitals, Attractions, and Police/Fire Stations)
  - Land use data (Parks/Waterbodies, Commercial, Residential, Industrial)

POI are points of interest, *LAND* are land uses and $MA(1KM)\_$ are measures within 1 km with distance exp decay (rho=1)

Regression Results | Score and Raw Probability Distribution

Industrial Neighborhoods - Visual Inspection

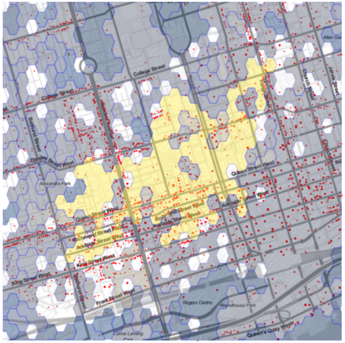
- Industrial neighborhoods do not look like postal codes
- They are larger and around main streets



Industrial Neighborhoods - Visual Inspection

- Industrial neighborhoods do not look like postal codes
- They are larger and around main streets

all
centroid: [-79.392,43.652]
96.33 % of firms in square clusters
91.03 % of firms in all clusters



Note: **The graph shows the surroundings of the neighborhood with the highest average score without counting singletons**

Industrial Neighborhoods - Characteristics

- And there are some differences across industries
- And across the city Size Spatial Distribution

	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	80.95	5,869	1.847	1.208	0.603	1.915
				(49.891)	(4.106)	(1.819)	(0.975)
Wholesale and Retail Trade	44,846	77.94	10,316	0.867	0.852	0.459	1.674
				(20.681)	(2.771)	(1.143)	(2.093)
Professional Services	47,464	79.34	10,808	0.684	0.77	0.401	1.664
				(17.779)	(2.774)	(1.051)	(3.064)
Entertainment, Accommodation and Food	15,464	81.65	3,998	2.051	1.225	0.612	1.662
				(47.342)	(4.893)	(1.988)	(1.087)

Note: Results correspond to running the algorithm for each group of firms with a threshold set to 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.

Outline

Housing Neighborhoods - Setup

- The propensity score captures the probability that a given firm chooses grid cell i based on its characteristics

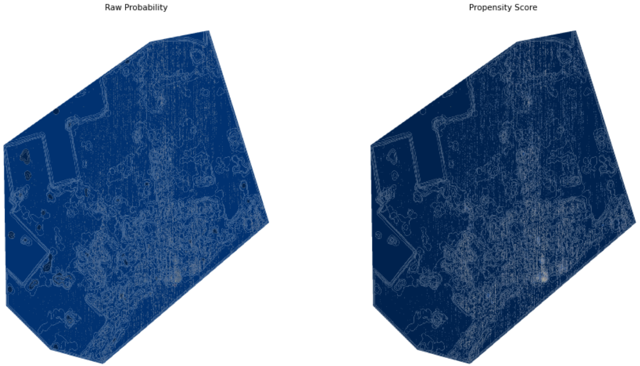
$$\#\text{Sales}_i = \beta + \sum_{POI} \beta_{POI}\text{POI}_i + \sum_{POI} \beta_{\text{MA}POI}\text{MA\_POI}_i + \sum_{POI} \beta_{LAND}\text{LAND}_i + \sum_{LAND} \beta_{\text{MA}LAND}\text{MA\_LAND}_i + \beta_{rooms} AvgRooms + \beta_{allrooms} SumRooms + \beta_{lot} AvgLotSize + \beta_{stock} Stock + \epsilon_i$$

- Dataset (in addition to previous data)
  - 2012 MLS Housing Transactions
  - Stock of Housing Units Based on DMTI List of Addresses

Regression Results
Score and Raw Probability Distribution

Housing Neighborhoods - Visual Inspection

- Housing neighborhoods are smaller and look more similar to postal codes around big streets
- and show less variance in the score than their industrial neighborhoods which is an indication that this method (or the score specification) might be not suitable for this type of data

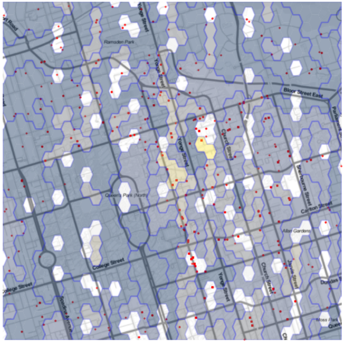


Note: **The graph shows the surroundings of the neighborhood with the highest average score without counting singletons**

Housing Neighborhoods - Visual Inspection

- Housing neighborhoods are smaller and look more similar to postal codes around big streets
- and show less variance in the score that their industrial neighborhoods which is an indication that this method (or the score specification) might be not suitable for this type of data

all
centroid: [-79.387,43.667]
72.11 % of transactions in square clusters
89.57 % of transactions in all clusters



Neighborhoods by Type of Housing

Note: **The graph shows the surroundings of the neighborhood with highest average score without counting singletons**

Figure: Housing Transaction Neighborhoods

	Number of Transactions	Transactions in Neighborhoods	Number of Neighborhoods	Area	Length	Width	Length/Width
All Transactions	68,184	57.11	9,233	0.436	0.493	0.265	1.573
				(14.384)	(1.994)	(0.891)	(1.222)
House Transactions	48,510	72.69	9,052	0.46	0.49	0.267	1.572
				(15.072)	(2.034)	(0.891)	(0.973)
Condo Transactions	19,674	69.22	9,536	0.745	0.858	0.394	2.019
				(17.698)	(2.525)	(1.105)	(1.91)

Note: Results correspond to running the algorithm for each type of transaction with a threshold set to one 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.

Size Spatial Distribution

Outline

Neighborhoods vs Postal Codes

- The miss-alignment between 'official' and 'economic' boundaries may lead to a variety of biases
  - Measurement error
  - Spatial correlation across neighborhoods (Lind and Ramondo, 2018)
  - Modifiable areal unit problem (Briant et al., 2010)
- Here we analyze the last two by looking at
  - Global Moran's I measure of spatial correlation
  - The relationship between the rank of a neighborhood in a given characteristic and the characteristic (Zipf's law)

Neighborhoods vs Postal Codes - Spatial Correlation (Moran's I)

Industrial Neighborhoods

	Average Propensity Score		Number of Firms	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Firms	-0.022	0.731	0.002	0.273
	(0.001)	(0.001)	(0.064)	(0.001)
Manufacturing	0.495	0.862	0.000	0.208
	(0.001)	(0.001)	(0.357)	(0.001)
Wholesale and Retail Trade	0.365	0.615	0.040	0.25
	(0.001)	(0.001)	(0.001)	(0.001)
Professional Services	0.268	0.655	-0.005	0.335
	(0.001)	(0.001)	(0.001)	(0.001)
Entertainment, Accommodation and Food	0.461	0.817	-0.03	0.469
	(0.001)	(0.001)	(0.001)	(0.001)

Housing Neighborhoods

	Average Propensity Score		Number of Transactions	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Transactions	0.033	0.571	-0.252	0.46
	(0.001)	(0.001)	(0.001)	(0.000)
House Transactions	0.015	0.453	-0.622	0.352
	(0.004)	(0.001)	(0.001)	(0.000)
Condo Transactions	0.231	0.766	0.204	0.473
	(0.001)	(0.001)	(0.001)	(0.000)

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 immediate contiguity. Pseudo p-value is reported in parenthesis.

Neighborhoods vs Postal Codes - Zipf's Law

Industrial Neighborhoods

	Average Propensity Score		Number of Firms	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Firms	-0.722	0.14	-0.407	0.271
	(0.001)	(0.002)	(0.007)	(0.164)
Manufacturing	-0.861	0.152	-0.516	0.034
	(0.004)	(0.003)	(0.015)	(0.062)
Wholesale and Retail Trade	-0.762	0.156	-0.391	0.084
	(0.002)	(0.002)	(0.01)	(0.067)
Professional Services	-0.803	0.165	-0.533	0.07
	(0.001)	(0.002)	(0.009)	(0.061)
Entertainment, Accommodation and Food	-0.811	0.198	-0.426	0.03
	(0.005)	(0.003)	(0.017)	(0.066)

Housing Neighborhoods

	Average Propensity Score		Number of Firms	
	Neighborhoods	Postal Codes	Neighborhoods	Postal Codes
All Transactions	-1.13	-0.38	-0.436	0.124
	(0.002)	(0.002)	(0.01)	(0.003)
House Transactions	-1.278	-0.373	-0.61	0.128
	(0.004)	(0.002)	(0.012)	(0.003)
Condos Transactions	-0.644	-0.319	-0.099	0.169
	(0.009)	(0.004)	(0.021)	(0.006)

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 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.

Outline

Conclusions

- This paper provides an data-based solution for the miss-alignment between 'official' and 'economic' neighborhoods
- It shows that the 'economic' neighborhoods
  - do not look like postal codes and are different across types
  - show less spatial correlation and follow a power law which is not the case for 'official' neighborhoods (postal codes)
- Simulation exercises show that there are some distribution requirements in order to provide good and stable neighborhoods
  - this is corroborated by the differences between industrial and housing neighborhoods
- Visual inspection is crucial

Thanks!

Outline

Daniel Arribas-Bel, M. À. Garcia-López, and Elisabet Viladecans-Marsal. Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, page 103217, November 2019. ISSN 0094-1190. doi: 10.1016/j.jue.2019.103217.

Susan Athey and Guido Imbens. Machine Learning Methods Economists Should Know About. *arXiv:1903.10075 [econ, stat]*, March 2019.

A. Briant, P. P. Combes, and M. Lafourcade. Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, 67(3):287–302, May 2010. ISSN 0094-1190. doi: 10.1016/j.jue.2009.09.014.

Marie-Pierre de Bellefon, Pierre-Philippe Combes, Gilles Duranton, Laurent Gobillon, and Clément Gorin. Delineating urban areas using building density. *Journal of Urban Economics*, page 103226, December 2019. ISSN 0094-1190. doi: 10.1016/j.jue.2019.103226.

Jonathan I Dingel and Felix Tintelnot. Spatial Economics for Granular Settings. Working Paper 27287, National Bureau of Economic Research, May 2020.

References II

Steve Gibbons, Henry G. Overman, and Eleonora Patacchini. Spatial Methods. In Gilles Duranton, J. Vernon Henderson, and William C. Strange, editors, *Handbook of Regional and Urban Economics*, volume 5, pages 115–168. Elsevier, 2015.

Nelson Lind and Natalia Ramondo. Trade with Correlation. Working Paper 24380, National Bureau of Economic Research, March 2018.

Paul R. Rosenbaum and Donald B. Rubin. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387): 516–524, 1984. ISSN 0162-1459. doi: 10.2307/2288398.

Hernán D. Rozenfeld, Diego Rybski, Xavier Gabaix, and Hernán A. Makse. The Area and Population of Cities: New Insights from a Different Perspective on Cities. *American Economic Review*, 101(5): 2205–2225, August 2011. ISSN 0002-8282. doi: 10.1257/aer.101.5.2205.

Giorgio Topa and Yves Zenou. Neighborhood and Network Effects. In Gilles Duranton, J. Vernon Henderson, and William C. Strange, editors, *Handbook of Regional and Urban Economics*, volume 5, pages 561–624. Elsevier, 2015.

References III

Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963. ISSN 0162-1459. doi: 10.1080/01621459.1963.10500845.

Appendix

Methodology - Merging Criteria Go back

Let $\mathcal{P} = \{\mathrm{P}_t^i\}_{i=1}^{\mathbf{B}}$ the set of all cell-level probabilities to be clustered

The loss of information when grouping blocks into a neighborhoods $N \subset \mathcal{P}$

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

where $\overline{\mathrm{P}}_N = n^{-1} \sum_{u=1}^{n} \mathrm{P}_t^u$ is the *centre of gravity* of N and n is the number of blocks in the neighborhoods.

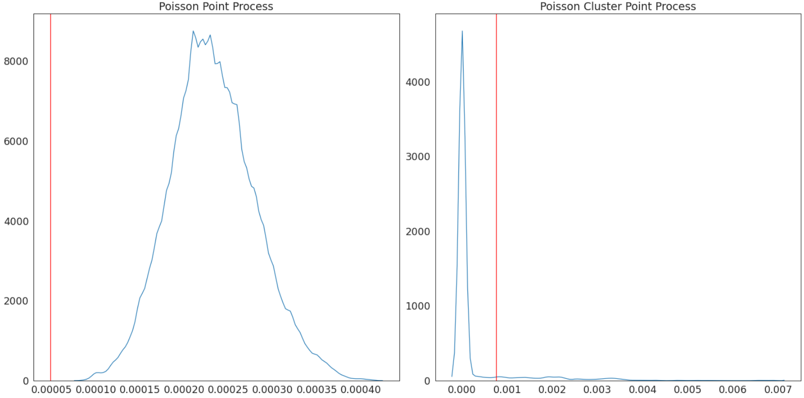
Starting from a partition $\{N_1, ..., N_l\}$ of $\mathcal{P}$, the loss of information when merging two neighborhoods N_u and N_v is quantified by:

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

That, when minimized, it is equal to minimizing the variation of *within-cluster sum of squares* after merging two clusters (Ward, 1963)

Simulations: Two Cities - Kernel Density

Go back

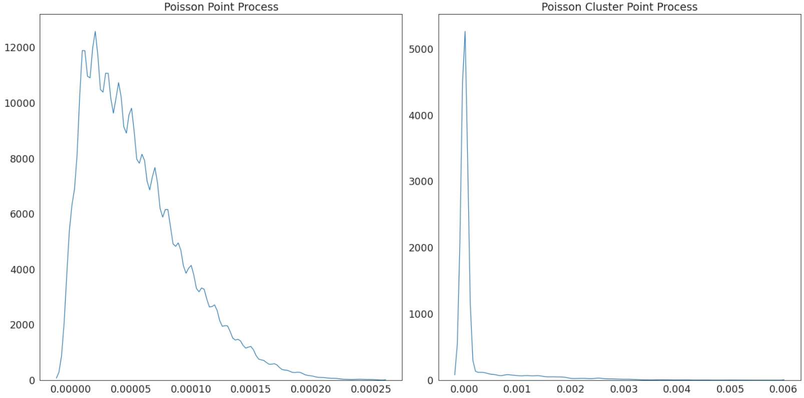


red lines are the standard deviations

Simulations: Distribution of Bilateral Distances

Go back

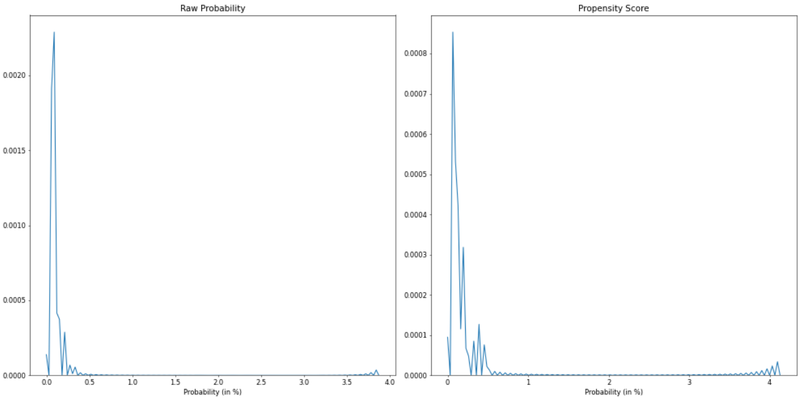
- At the starting point, the dissimilarity between neighborhoods is given by the euclidean distance between scores

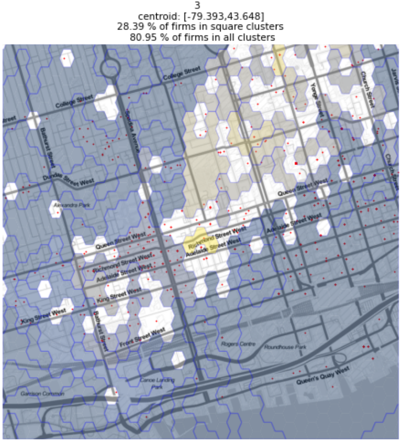


Firms: Propensity Score Regressions Go back

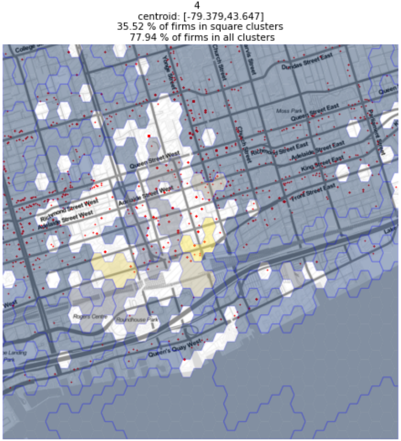
	All Firms	Manufacturing	Trade	Services	Ent/Food/Lodge
MA_1km_area_Parks	-0.26***	-0.25***	0.12***	0.07***	-0.20***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
MA_1km_area_Open	-0.23***	-0.28***	-0.19***	-0.13***	-0.34***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
MA_1km_area_Residential	0.38***	0.16***	0.15***	0.22***	-0.06***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
MA_1km_area_Industrial	0.23***	0.43***	-0.09***	-0.27***	-0.26***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_area_Commercial	-0.10***	-0.23***	-0.02***	0.11***	-0.01
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
MA_1km_area_Government	-0.02***	0.16***	-0.08***	0.03***	0.12***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_poi_POST	0.10***	0.21***	0.02***	-0.01*	-0.04***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
MA_1km_poi_TOUR	0.09***	0.24***	0.14***	-0.11***	-0.17***
	(0.00)	(0.02)	(0.01)	(0.01)	(0.01)
MA_1km_poi_BANK	-0.12***	-0.11***	0.12***	0.08***	0.17***
	(0.00)	(0.02)	(0.01)	(0.01)	(0.01)
MA_1km_poi_RESTA	0.06***	-0.04***	-0.01***	0.09***	0.12***
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
MA_1km_poi_HOTEL	-0.04***	0.06***	-0.24***	-0.08***	0.10***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)
Stock_Houses_DMTI	-0.06***	0.01**	-0.09***	-0.00**	0.03***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Downstream		0.06	0.06	0.25**	-0.03
		(0.12)	(0.09)	(0.11)	(0.07)
Upstream		-0.38***	-0.09	-0.12	-0.01
		(0.13)	(0.10)	(0.11)	(0.07)
const	-2.46***	-4.39***	-3.21***	-3.07***	-4.29***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	740286	740286	740286	740286	740286
Pseudo R-squared	0.33	0.29	0.28	0.30	0.27

Firms: Propensity Score Distribution Go back





(a) Manufacturing

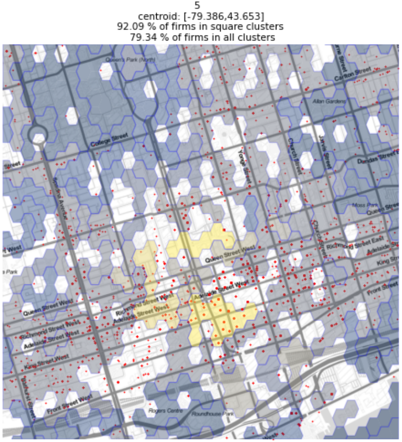


(b) Wholesale and Retail Trade

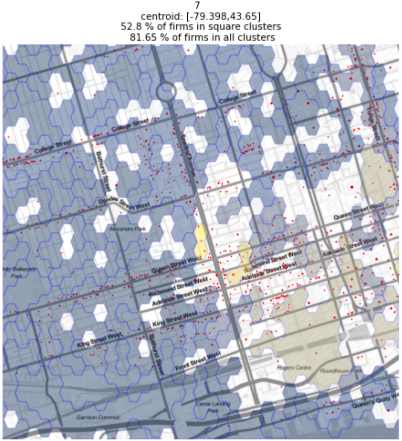
Figure: Neighborhoods by Industry (I)

Firms: Results by Industry

Go back



(a) Professional Services

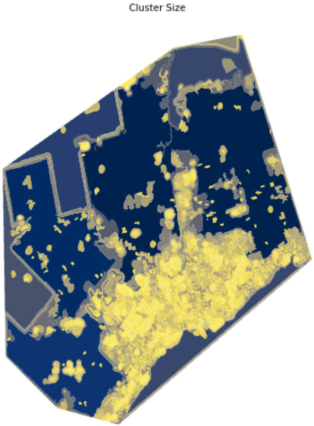


(b) Entertainment, Accommodation and Food

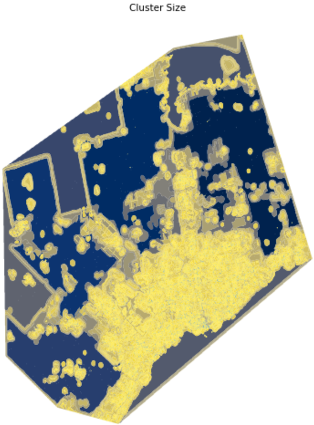
Figure: Neighborhoods by Industry (II)

Firms: Spatial Size Distribution

Go back



(a) Manufacturing



(b) Professional Services

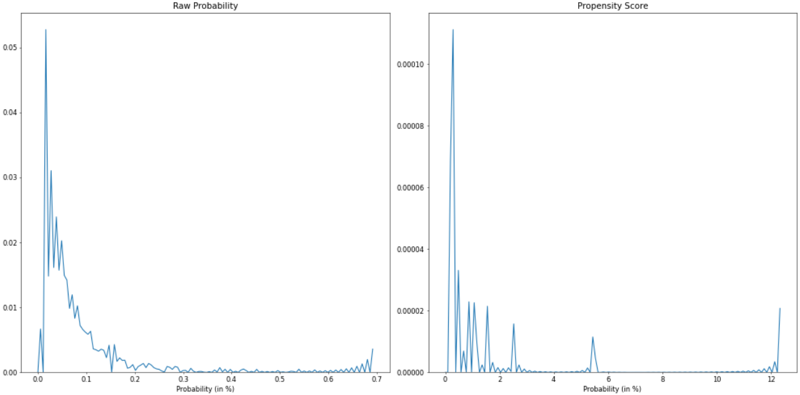
Figure: Neighborhood Size Distribution

Housing: Propensity Score Regressions Go back

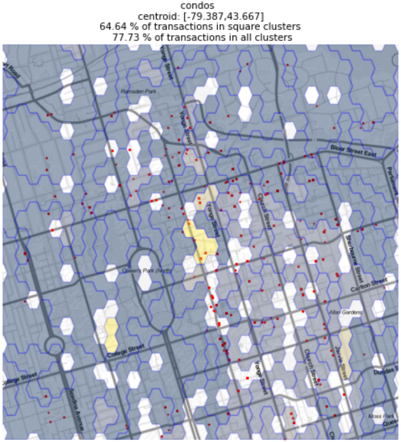
	#Sales	#House Sales	#Condo Sales
MA_1km_area_Parks	0.78***	0.34***	1.32***
	(0.03)	(0.03)	(0.06)
MA_1km_area_Open	0.93***	0.31***	1.37***
	(0.04)	(0.04)	(0.09)
MA_1km_area_Residential	0.74***	0.43***	1.13***
	(0.02)	(0.02)	(0.05)
MA_1km_area_Industrial	0.29***	0.12***	0.53***
	(0.01)	(0.01)	(0.03)
MA_1km_area_Commercial	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***	-98.25***
	(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

Housing: Propensity Score Distribution

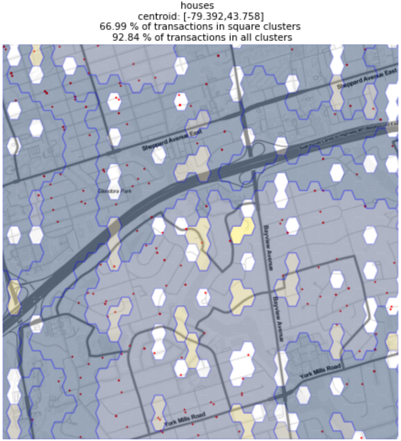
Go back



 Go back



(a) Condos

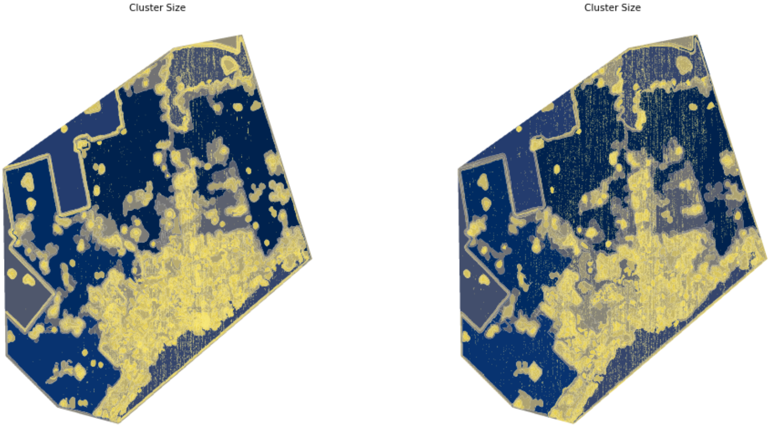


(b) Wholesale and Retail Trade

Figure: Neighborhoods by Type

Housing: Spatial Size Distribution

Go back



(a) Condos (b) Houses

Figure: Neighborhood Size Distribution