How Densely Do Manufacturing Establishments Occupy Land?

Kristian Behrens* Florian Mayneris† Théophile Ndjanmou Biéda‡

August 22, 2024

Abstract

We construct a new dataset containing parcel sizes and building footprints of Canadian manufacturing plants. Using these data, we decompose industrial density (parcel size per worker) into: crowding (floorspace per worker); building height (floorspace to building footprint); and parcel coverage (building footprint to parcel size). We find that establishments occupy parcels more densely in big cities and central locations, and that larger establishments use less land per worker. Floorspace per worker is, however, unrelated to distance from the city centre. These facts cannot be replicated by a Cobb-Douglas production function but by one with fixed land requirements and a CES aggregator for land- and non-land inputs. We estimate the elasticity of substitution between land- and non-land factors at between 0.14 and 0.42.

Keywords: Land use; Industrial density; Geo-referenced data; Building and parcel polygons; Production function.
JEL Classification: R32; R14; L60.

*Université du Québec à Montréal, Canada; and CEPR. Email: behrens.kristian@uqam.ca
†Université du Québec à Montréal, Canada; and CEPR. Email: mayneris.florian@uqam.ca
‡Caisses Desjardins, Canada. Email: ndjanmou@yahoo.fr.

Acknowledgements. We thank the editor, Daniel Xu, three anonymous reviewers, our discussants Matt Freedman and Miquel-Àngel Garcia-López, as well as Gabriel Ahlfeldt, Théophile Bougna, Julien Martin, Marlon Seror, Nate Baum-Snow, and participants at many seminars and conferences for helpful comments and suggestions. We gratefully acknowledge financial support from SSHRC (Insight Development Grant #430-2018-00959).

1 Introduction

Competition for land has intensified substantially over the past 30 to 40 years. This is due to a combination of increased demand, on the one hand, as the population of large cities continues to grow; and limited supply, on the other hand. Although the deindustrialization process has freed up some space in cities—via the reconversion of former industrial buildings into offices and apartments—a dearth of developable vacant land,[1] administrative red-tape and restrictive land-use and building regulations,[2] as well as the pressing need to reduce land consumption to meet climate goals and preserve ecosystems,[3] substantially constrain land supply. Against this background, we need to better understand the human footprint on land. How much land do we use? And how do we use it?

Inspired by the monocentric city model and its refinements, many studies on *residential land use* address different topics such as city shape and growth, the extent and causes of urban sprawl, the determinants of housing supply, the political economy of zoning, or the determinants of housing- and land prices (see Duranton and Puga, 2015, for an extensive review of the literature). Given a lack of data, we know much less about *industrial land use* for production purposes, outside of the residential construction sector.[4] This is problematic as the share of land used for industrial

[1] There exist few data on vacant land in cities. Using survey information, Pagano and Bowman (2000, p.7) document that there is on average 15% of vacant land in US cities. However, only a fraction of that land is developable: *"three characteristics [not large enough, odd-shaped, or in the "wrong" location] individually, but especially in concert, seemed to limit the development potential of vacant land."*

[2] There has been little change in the restrictiveness of land-use regulations. If anything, *"to the extent there is change [between 2006–2016], it is to strengthen the control regime."* (Gyourko et al., 2021, p.3)

[3] See the 'zero net artificialization' law passed in France in 2021, which aims to reduce by 50% the consumption of natural, agricultural, and forest areas between 2021 and 2031, and which ultimately strives for no net new land consumption by 2050. The United Nations' Sustainable Development Goals 11 (Indicator 11.3.1) is about reducing the "Ratio of land consumption rate to population growth rate", i.e., to strive to make cities more compact.

[4] When available at all, data on land and buildings are usually lumped together with capital in balance sheets. Datasets sometimes provide values for land and buildings, but no comprehensive quantity data on firms' land or floorspace consumption exist to our knowledge.

and commercial purposes within cities is sizeable, generally above 20%.[5] Moreover, although much attention has been given to the analysis of residential land use and sprawl, industrial and commercial land use are major drivers of low-density suburban development, thus potentially generating large environmental costs.[6]

In this paper, we make progress in understanding land use for manufacturing production, both from empirical and theoretical perspectives. To this end we first build a new *quantity dataset* containing the parcel sizes, the footprints of the buildings, and the number of workers for a large sample of Canadian manufacturing plants. Inspired by the literature on urban residential density (see, e.g., Angel et al., 2021), we then decompose manufacturing establishments' *industrial density* (number of workers per unit of land) into three components: *crowding* (number of workers per unit of floorspace), *building height* (the ratio of floorspace to building footprint), and *parcel coverage* (the ratio of building footprint to parcel size). For a large sample of Canadian manufacturing establishments, we then document how *industrial density* and *parcel coverage* vary with establishment characteristics and attributes of their environment, in particular establishment size (as measured by employment), city size, and distance to the closest city centre. Using the subset of establishments located in Montréal, for which we have more detailed data, we do the same with *crowding* and *building height*. We finally use our empirical results to revisit the modeling of land use for production and to infer a range for the elasticity of substitution between land and non-land factors.

Previewing our key results, firms located in large cities and closer to city centres occupy their parcels more densely, both in terms of employment and building footprint. Put differently, the ratio of employment to parcel size (industrial density) and the ratio of building footprint to parcel size (parcel coverage) increase with city size and decrease with distance to the closest city centre. The number of workers per unit

[5]For example, around 30% of the built-up area in the Paris metropolitan region was for commercial use in 2012 (Duranton and Puga, 2015). Figures for Canadian cities are slightly lower but in the same ballpark, with 22.86% in Calgary, 22.52% in Montréal, and 17.94% in Vancouver (authors' computations).

[6]As documented by Burchfield et al. (2006, p.589), contrary to residential development, commercial land development became substantially more sprawled between 1976 and 1992 in the US. The same holds true for Canada: "*The main contributors to low-density suburban development are not residential uses but non-residential activities (commercial, industrial, distributional uses) [...] Surprisingly, no one seems to notice.*" (Bourne, 2001, p.27)

of floorspace (crowding), instead, is essentially flat within cities: it does not vary with distance to the nearest city centre. We further find that large establishments, in terms of their number of employees, have higher industrial density and crowding. Based on a theoretical production function framework, we then derive three main implications for the way we should model land as a production factor from our empirical results: (i) the small values of the (semi-)elasticities of industrial density to city size and to distance to the city centre suggest that the elasticity of substitution between land and non-land production factors is far below the canonical Cobb-Douglas value of 1; (ii) the essentially zero semi-elasticity of crowding with respect to distance to the city centre suggests that floorspace and other production factors are complements, related in a Leontief way; and (iii) industrial density and crowding increase with establishment size, which suggests that land and floorspace have a fixed-cost component.

We believe that these results are important. First, they suggest that the canonical Cobb-Douglas case may not be a good theoretical approximation of how land and floorspace enter the production function of manufacturing establishments. Actually, while capital and labor may be substitutes in a way compatible with the Cobb-Douglas case—an iPad can take an order instead of a human—this seems more difficult to rationalize for land (or floorspace) and non-land factors. How should we exactly think about labor-land isoquants? A framework combining a CES production function with minimum land requirements seems empirically more plausible. The way we model land may not be innocuous, especially given the increasing use of quantitative spatial models to quantify the effects of changes in infrastructure and land-use regulations (see Redding and Rossi-Hansberg, 2017 for a survey). Second, the fact that industrial density and parcel coverage, but not crowding, change with land prices suggests that most of the adjustment of manufacturing establishments to land prices happens through outdoor space. This is reminiscent of the floorspace-vs-yardspace trade-off in residential urban models (see, e.g., Brueckner, 1983), although outdoor space for industrial use—mainly parking and storage—differs from yardspace in that it can have larger environmental costs (see, e.g., Davis et al., 2010, for parking). Last, although we focus on manufacturing only, we believe that our results are relevant. Manufacturing is often a land-intensive activity, and considering the recent signs of reindustrialization in

several developed countries such as the US, Germany, or France, understanding better the needs of manufacturing establishments in terms of land is important given these (sometimes gigantic) new manufacturing projects.[7]

Our work contributes to two main strands of literatures. First, it relates to the literature on the role of land in the production process of firms. The few empirical studies on that matter relate either to the effects of land-use regulations on productivity in the retail sector (e.g., Haskel and Sadun, 2012; Cheshire et al., 2014) or to the determinants of commercial real estate prices (e.g., Drennan and Kelly, 2010; Ahlfeldt and McMillen, 2018; Liu et al., 2018). Except for some work on the patterns and determinants of the floor-to-area ratio (FAR; see, e.g., Barr and Cohen, 2014; Brueckner et al., 2017), we are not aware of studies on the quantity of land used by firms. This is likely due to a lack of data, and our work partially fills this gap by generating a new dataset. On the theoretical front, canonical urban models generally assume that production is concentrated in dimensionless 'business districts', i.e., production requires no land. Notable exceptions—where firms and residents compete for land—include Ogawa and Fujita (1980), Fujita and Ogawa (1982), Lucas and Rossi Hansberg (2002), Pflüger and Tabuchi (2010), and Wrede (2013), but the way land enters the production function varies greatly across existing models—sometimes it is a pure fixed cost or Leontief, but more often it is via a Cobb-Douglas production function. To the best of our knowledge, we are the first to empirically analyze the ingredients that should be used to model land in the production function of manufacturing firms.[8]

Second, our work speaks to the literature on the determinants of urban density. As

[7]See, e.g., `https://home.treasury.gov/news/featured-stories/unpacking-the-boom-in-us-construction-of-manufacturing-facilities` for recent US trends, the Tesla gigafactory in Berlin-Brandenburg in Germany, or the construction of several gigafactories in the north of France.

[8]Several recent studies find that the production function of many manufacturing sectors is not Cobb-Douglas (see, e.g., Oberfield and Raval, 2021 for the US; Imbert et al., 2022 for China; and Mayneris, 2022 for France). However, these studies analyze the substitution between labor and capital without accounting explicitly for land. The construction sector is the only one for which we are aware of studies analyzing the role of land in the production process. Epple et al. (2010) and Combes et al. (2019), for example, estimate production functions for housing where land and non-land inputs (called "capital") are the two factors. The latter show that for newly-built single-family homes, the *production function for housing* is well, though not perfectly, approximated by a Cobb-Douglas function.

extensively discussed in Duranton and Puga (2015), this literature mainly focuses on population density. In particular, in recent years several works have proposed margin decompositions of population density (population over urban footprint) to analyze how a given density level is compatible with different types of the urban environment (see, e.g. Angel et al., 2021; Narro et al., 2020; Seidel and Krause, 2024). We adapt this type of decomposition to analyze the density with which manufacturing establishments concentrate employment in space. Instead of using the margin decomposition to compare neighborhoods or cities, we use it to compare manufacturing establishments within cities to better understand how land consumption varies with their individual characteristics and their urban environment.

The remainder of the paper is organized as follows. Section 2 presents our decomposition framework, explains the construction of our dataset, and provides sectoral descriptive statistics on land consumption by manufacturing establishments. Section 3 estimates the (semi-)elasticities of plant-level land consumption to city- and plant-level characteristics. We propose a conceptual framework in Section 4 that puts some structure on our regression results and guides us in the quantification of the elasticity of substitution between land and non-land inputs. Section 5 concludes. Details on our data and robustness checks are relegated to the online appendix.

2 A quantity-based dataset: Decomposing density

The objective of our analysis is to describe how densely manufacturing establishments occupy land, and from there to better understand how they use land in their production function.

It has been increasingly recognized that the same level of overall density can be achieved in different ways, i.e., it can arise from different combinations of underlying factors. An emerging literature on 'density accounting' thus decomposes density along various margins. For example, in the residential context, Angel et al. (2021, pp.268–269)

suggest to decompose *urban density*—population over urban area—as follows:[9]

$$\underbrace{\frac{\text{Total population}}{\text{Total area of the urban footprint}}}_{\text{Urban density}} = \underbrace{\frac{\text{Total population}}{\text{Total residential floor area}}}_{\text{Crowding}} \underbrace{\frac{\text{Total residential floor area}}{\text{Total area of residential plots}}}_{\text{Floor area ratio, height}} \underbrace{\frac{\text{Total area of residential plots}}{\text{Total area of the urban footprint}}}_{\text{Residential share}} \quad (1)$$

Hence, urban density depends on how crowded and how tall buildings are, and on how much of the city's surface area is devoted to residential use.

We use a framework analogous to (1) to decompose, at the parcel level, *industrial density*—number of workers over parcel size—as follows:

$$\underbrace{\frac{\text{Establishment employment}}{\text{Establishment parcel size}}}_{\text{Industrial density}} = \underbrace{\frac{\text{Establishment employment}}{\text{Establishment floorspace}}}_{\text{Crowding}} \underbrace{\frac{\text{Establishment floorspace}}{\text{Establishment building footprint}}}_{\text{Building height}} \underbrace{\frac{\text{Establishment building footprint}}{\text{Establishment parcel size}}}_{\text{Parcel coverage}} \quad (2)$$

$$= \underbrace{\frac{\text{Establishment employment}}{\text{Establishment floorspace}}}_{\text{Crowding}} \underbrace{\frac{\text{Establishment number of floors}}{\text{Building height}}}_{\text{Building height}} \underbrace{\frac{\text{Establishment building footprint}}{\text{Establishment parcel size}}}_{\text{Parcel coverage}} \quad (3)$$

where (3) approximates floorspace by number of floors times building footprint. Hence, industrial density depends on how crowded and how tall buildings are, and on how much of the parcels' surface area is covered by buildings. The product of the last two terms, $\frac{\text{Establishment floorspace}}{\text{Establishment parcel size}}$ is sometimes called 'structural density'. Provided that the establishment's number of floors equals one, which we later show to be a good approximation, it is also given by $\frac{\text{Establishment building footprint}}{\text{Establishment parcel size}}$.

We now explain how we build a quantity-based dataset that allows us to measure and to decompose industrial density as in (2) and (3).

2.1 Data sources and dataset construction

We collect information on the amount of land occupied by manufacturing establishments.[10] Ideally, our data would allow us to construct industrial density and its three components: crowding, building height, and parcel coverage. Unfortunately, as explained below, we do not have floorspace and height information for all buildings in our data and cannot easily infer it. We hence focus mostly on two key measures: *industrial density* and *parcel coverage*. Computing the first requires the surface area of

[9] Several other recent contributions have decomposed the average residential density of cities along various margins (see, e.g., Narro et al., 2020 for Japan; and Seidel and Krause, 2024 for Norway).

[10] In what follows, we interchangeably use the terms 'establishment', 'plant', and 'firm'.

the polygon of the parcel where the plant is located, whereas computing the second requires information on the building polygons. Note that for establishments located in single storey buildings, building footprint closely matches floorspace. We exploit this idea later in the paper when we push further the analysis on crowding thanks to more detailed data on building height and floorspace for the city of Montréal.

We first describe the methodology used to construct our dataset. Details for each step and an extensive discussion of the quality of the final dataset are relegated to online appendices A and B.

2.1.1 Data collection and processing

Establishment-level data. We use the proprietary Scott's National All Business Directories, a dataset of geo-referenced manufacturing plants operating in Canada that draws information from Business Register records and telephone surveys. It provides a fairly exhaustive coverage of the manufacturing sector (which is the key reason why we focus on that sector). We use cross-sectional data for 2017, the closest year to the reference year for the polygon datasets we use. This choice reduces potential measurement error due to changes in the delineation of buildings and parcels. It also allows for more precise geocoding as street names and configurations may change over time. The variables of interest for our analysis include the plant's address and industry (North American Industry Classification System, NAICS 6-digit level), an estimate of the number of onsite workers, and dummy variables for whether it is a headquarter (HQ) or has export activities. The dataset also contains information on the products manufactured by the plants (up to ten) and its broad type of activity (manufacturing, wholesale, professional, scientific and technical services etc.). We geocode the plants using the procedure explained in online appendix A.

Polygon datasets. We collect parcel- and building polygons from numerous provincial and metropolitan sources. The full list of sources is given in online appendix A (see Table A1). For parcels, we collect more than 4.5 million polygons covering the entire provinces of British Columbia (BC), Quebec (QC), and New Brunswick (NB). For the other provinces, we obtain data for Toronto, Oshawa, Windsor, and York in Ontario

(ON); Banff in Alberta (AB); Winnipeg in Manitoba (MB); and Regina and Saskatoon in Saskatchewan (SK). We did not obtain data for Nova Scotia (NS), Newfoundland and Labrador (NL), Prince Edward Island (PE), and the three Territories. For buildings, we collect open source data on building polygons released by Microsoft, which contain 12,663,475 building footprints covering all provinces and territories.[11]

Other datasets. To leverage spatially fine-grained population census data, we combine data from the 2016 population census with boundary shapefiles of dissemination areas (the smallest geographic units at which census data are publicly released in Canada), census metropolitan areas (CMA), census agglomerations (CA), economic regions, and provinces and territories. Last, we collect information on major infrastructure such as highway junctions (from the Canadian road network files), as well as rail freight stations, major airports, and major seaports (from the Open Government geographic data portal).

We complement our data with more detailed information that we can access for the city of Montréal. We have the 2016 assessment roll data that allows us to see tax lots and the associated parcel and building polygons. Crucially, this dataset contains direct measures of floorspace and information on the number of floors for most observations, which will allow us to compute the full decomposition that appears in equations (2) and (3). We also will make use of the Québec business register which contains information on all establishments registered in the province. This will allow us to construct finer controls for the number of other establishments that are located on the same parcels as the manufacturing plants in the Scott's database.

2.1.2 Construction of the surface measures

We use GIS tools to link plants in the establishment dataset to parcel- and building polygons (see online appendix A for details). This mapping allows us to construct the measures of land occupied by a manufacturing establishment, namely its parcel size and building footprint. The parcel size is the surface of the parcel polygon that

[11] For additional information, see `https://blogs.bing.com/maps/2019-03/microsoft-releases-12-million-canadian-building-footprints-as-open-data`.

contains the establishment, while the building footprint is the ground floor area of all building polygons on the establishment's parcel. We must deal with three potential problems while building the dataset.[12]

First, there is unfortunately no one-to-one mapping between establishments and parcel- and building polygons. Sometimes, several establishments fall on the same parcel or building, i.e., there is some 'sharing' by neighboring establishments. In that case, it is unclear how to divide space between establishments. In our sample used for the analysis of parcel size, the average number of neighbors—identified from the Scott's data—is 1.3, whereas the median is 0. Hence, sharing does not seem to be a major problem for our analysis. We still control in our regressions for the number of neighbors sharing parcels or buildings using a flexible fourth-order polynomial. We will also show that our main results are unchanged when focusing on establishments with no identified neighbors.[13]

Second, the locations occupied by manufacturing establishments may consist of several contiguous parcels and not just the ones on which the establishments fall during the geocoding process. Following discussions we had with employees of the Québec Land Registry, we think that this situation rarely occurs. This is consistent with Brooks and Lutz (2016), who show that assembled parcels have a much higher value than the sum of the individual parcels, so that owners of contiguous parcels have an incentive to assemble them. Moreover, using the Montréal assessment roll data we can compute the number of parcels associated with the establishments' tax lots. We find that 85% of the manufacturing establishments in Montréal are on tax lots composed of a single parcel, confirming our discussions with the Land Registry. Hence—provided that Montréal is nationally representative—measurement error seems limited in that specific matter.

Last, the surface area of the parcel may be smaller than the building footprint for

[12] Assigning geocoded data to polygons delineated from satellite imagery raises several issues regarding quality and accuracy. We relegate a detailed discussion of these issues to online appendix B

[13] Although the Scott's database provides a near exhaustive coverage of the manufacturing sector, it is not the universe of plants in Canada, especially for services which are more sparsely present in that database. There is hence potential measurement error in the count of neighbors. We provide robustness checks using the universe of establishments in Québec to show that our results change little when using more precise data to control for the number of neighbors.

some plants. This may be due to the separate assignments of establishments to parcels and buildings: an establishment can be assigned to a building and to a parcel that do not correspond to the same lot. Furthermore, there is likely some measurement error in the original building polygons data, which may be mis-identified by Microsoft's automated recognition procedure. In particular, adjacent buildings may get amalgamated—thus being visible as a single polygon in the data—thus yielding larger polygons that straddle several parcels. We will show that our results are unchanged when using only the sample of establishments for which the parcel size exceeds the building footprint.

2.1.3 Building height

As explained before, we do not have direct measures of floorspace for most of our establishments since the building polygons do not report height or volume information. Unfortunately, we have not been able to find comprehensive satellite data on building height (LIDAR data), as some cities do not have those data or are only partially covered. We can nevertheless provide an analysis of floorspace consumption as follows.

First, we have detailed assessment roll data for Montréal, which reports direct measures of floorspace and number of floors. Hence, for this subsample, we can make use of high quality data to investigate establishments' floorspace consumption.

Second, we can use those data to investigate the number of floors of the buildings to which our manufacturing establishments are assigned. We find that 59% of the manufacturing establishments in Montréal are located in single-floor buildings, 18.1% in two-floor buildings, and 5.3% in three-floor buildings. Hence, less than 18% of establishments are located in taller buildings, and those are most probably headquarters or sales offices (which we control for in the empirical analysis). In a nutshell, the Montréal data reveal that manufacturing is a relatively 'low-rise activity', so that $\frac{\text{Establishment floorspace}}{\text{Establishment building footprint}} = \text{Number of floors} \quad 1$ is not too bad an approximation.[14] However, we will further use the assessment roll data to compute sectoral shares of establishments that operate in single-floor buildings in Montréal. We consider that

[14]Montréal is an 'older' city, and manufacturing in earlier years more often occurred in multistorey buildings in dense city centres (e.g., 'light' manufacturing such as apparel or some textiles). Modern manufacturing operations in Montréal and elsewhere probably occur mainly on a single floor.

all industries where the median number of floors of the establishments in Montréal is below 1.6 are 'low-rise industries'.[15] We will then provide results on crowding for the whole of Canada using only those industries and taking building footprint as a proxy for floorspace. The underlying idea is that being able to operate in multi-floor buildings, as compared to single-floor buildings, is likely a technological characteristic and can thus be extrapolated from one city to the rest of the country.

2.2 Quality assessment and representativeness of the final dataset

To assess the quality of our data obtained from the geo-coding and assignment procedure, we make use of a subset for the province of Québec (QC). The reason is that the polygon identifiers in the QC dataset are the same as the official identifiers recorded on the government website of the Land Registry "Infolot". We can thus randomly draw a subset of plants in QC from our dataset and compare their parcel identifier from "Infolot" (obtained from the address of the establishment) to the one obtained from our assignment procedure. Doing so allows us to build a measure of quality—which we construct for the whole dataset, not just QC—with three categories: excellent, good, and poor (see online appendix B for details).

In the remainder of the paper, we only keep observations of excellent quality (79.1% of the observations for which we have a measure of parcel size; see table B1 in the online appendix). Thus, our final dataset contains the plants that: (i) are precisely geo-coded; and (ii) have an excellent assignment quality to parcel- or building polygons. We verify later that our results hold when we: (i) include observations of lower quality; and (ii) restrict our analysis to Montréal for which we have administrative data from the property assessment roll.

Out of 32,829 manufacturing plants recorded in the Scotts database for 2017, we assign parcel sizes of excellent quality to 10,428 (31.8%) of them, and building foot-

[15]The industries include: Non-metallic mineral product manufacturing (NAICS 327); Fabricated metal product manufacturing (NAICS 332); Computer and electronic product manufacturing (NAICS 334); and Electrical equipment, appliance and component manufacturing (NAICS 335). We exclude Textile Mills (NAICS 313); and Petroleum and coal product manufacturing (NAICS 324) because these two sectors have very few establishments in our Montréal sample.

prints of excellent quality to 25,289 (77%) of them. Focusing on observations for which we have excellent quality data for both parcel size and building footprint, and further removing observations that are located outside of urban areas, or for which some values of the covariates used in the regression analysis are missing, leaves us with a regression sample of 8,704 observations representing 30% of the manufacturing employment present in the original data.[16] Note that the loss of data is mainly due to the absence of parcel polygons for some provinces and, to a lesser extent, the accuracy of the geocoding and polygon assignments (see Table B1 for more details).

Turning to the sectoral representativeness of our data, Table B2 in online appendix B shows that the distribution of the 3-digit industries is broadly similar to that in the raw Scotts database. The correlation between these distributions exceeds 0.97. Turning to geography, Table B3 shows the distribution of plants across provinces. As explained before, we lack parcel polygons for entire provinces which are thus missing from our final dataset. Yet, the correlation of the distribution of establishments in our sample across provinces with the raw Scott's data remains reasonably high at 0.62. Last, although our final dataset is sectorally and geographically representative of manufacturing in Canada, there could still be selection on observables. We hence use a probit model to assess the extent to which the establishments in the final sample exhibit specific observable characteristics compared to the others. Table B4 shows that, beyond the geographic fixed effects, the only characteristic that is significantly related to the probability to be included in the regression sample is the size of the CMA, with a slight over-representation of establishments from large urban areas. Moreover, the pseudo R^2 of the regression is small, and only 5 percentage points are related to establishment characteristics. Put differently, there is little selection on observables in the sample used for our analysis.

[16]Statistics Canada reports 1,767,500 manufacturing jobs in 2017. The sum of employment across the 32,829 establishments in the raw Scott's data is 1,531,087, which hence covers close to 87% of national employment.

2.3 Some descriptive statistics

We now briefly present some descriptive statistics for our main variables of interest, i.e., industrial density (number of workers over parcel size) and parcel coverage. These statistics are based on our final regression samples.

Industrial density. Panel (a) of Figure C1 in the online appendix depicts the distribution of the ratio of establishment employment over parcel size (see Table C1 in the online appendix for the exact figures), which we take as the measure of how densely land is occupied by manufacturing establishments in terms of employees. The patterns in Figure C1 reveal substantial heterogeneity between and within sectors: the coefficient of variation is 917% for the whole sample, ranging from 110% to 1589% across industries. While some industries such as NAICS 314 ('Textile product mills') have on average lower density than others such as NAICS 311 ('Food mfg'), i.e., fewer employees per unit of parcel size, differences between industries are dwarfed by within-industry variation.

Parcel coverage. Turning to parcel coverage, panel (b) of Figure C1 in the online appendix shows again substantial heterogeneity in terms of the ratio of building footprint over parcel size, both between and within sectors (see Table C2 in the online appendix for the exact figures). Yet, this heterogeneity is less pronounced than that for parcels (the coefficient of variation equals 61% for the whole sample, and ranges from 35% to 81%). Some sectors, such as NAICS 324 ('Petroleum and Coal Products'), or 321 ('Wood Products') have small building-to-parcel ratios, while 315 ('Clothing and Textile') and 323 ('Printing and Support Activities') exhibit high ratios. This sectoral heterogeneity likely reflects different needs for outdoor space in terms of parking, storage, or loading and unloading. The sectors with the highest coverage belong to what could be called 'light manufacturing' (clothing, printing, textiles) and are found in denser areas.

The substantial amount of variation in industrial density and parcel coverage highlight that a detailed analysis controlling for plant-level characteristics and locational characteristics is important to better understand the patterns. This is what we do next in our multivariate econometric analysis.

3 Empirical analysis

Since the within-industry variation in industrial density and parcel coverage is huge, a more detailed analysis at the establishment level is required to understand the drivers of industrial density. How does land consumption depend on establishment-level individual or location-specific environmental characteristics, and how does it vary across and within urban areas?

3.1 Methodology

Let i index establishments, s index 4-digit industries, and z index economic regions. When relevant, we will write $i(s,z)$ to make clear that establishment i operates in sector s and is located in zone z. Assume that each individual density component $y_{i(s,z)}$ in the decomposition (2) is a function of environmental characteristics, $\text{Env}_{i(z)}$, and individual characteristics, Estab_i, as follows:

$$y_{i(s,z)} = \alpha \text{Env}_{i(z)} + \beta \text{Estab}_i + \theta_s + \eta_z + \epsilon_{i(s,z)}, \tag{4}$$

where $y_{i(s,z)} \in \{$industrial density, crowding, building height, parcel coverage$\}$ denotes one component of our decomposition.

The first term in equation (4), $\text{Env}_{i(z)}$, is a vector of characteristics related to establishment i's environment in zone z. We include the (log) size of the urban area it is located in, both in terms of population and surface (as per the 2016 Census); the distance of the establishment to the nearest city centre;[17] fixed effects identifying the type of zoning (commercial/industrial, residential, or recreational) in use at its location; a polynomial of degree four in the number of neighbors on the same parcel to flexibly control for 'sharing'; and measures of distance to specific types of transportation infrastructure (distance to the closest major airport, major seaport, rail freight station, and highway junction).[18] In some specifications, we also control for the local density around the plant, measured by the (log) population density in the dissemination areas

[17] A city corresponds either to a Census Metropolitan Area or a Census Agglomeration. To identify city centres, we use a routine that detects clusters of densely populated dissemination areas. The details of the procedure are presented in Appendix C.

[18] The proximity to these different types of infrastructure might influence industrial density and parcel

within a 500 metre radius. The latter will show how important local density is compared to city-level density in determining how manufacturing land use changes. As will become clear later, among these various local characteristics, we consider distance to the closest city centre and population density as proxies for local land prices.

The second term in equation (4), Estab_i, is a vector of plant-level characteristics related to the size and type of activities carried out by the establishment. We control, in particular, for the (log) number of employees of the plant, dummies identifying headquarters and exporters, as well as counts of the number of 4-digit industries, of products, and of broad types of activity the plant is involved in (e.g., if the establishment produces or is a wholesale outlet). These covariates aim to capture in the best possible way factors related to establishment-level specific needs in terms of land that may drive the large within-sector variations highlighted in section 2.3.

Last, θ_s and η_z stand for sector and economic region fixed effects.[19] They capture technological parameters and regional determinants beyond those already included that may drive how densely manufacturing establishments occupy land.

To account for auto-correlation between observations within urban areas, we cluster all standard errors at the CMA/CA level (Moulton, 1990). As mentioned in Section 2.2, we restrict the sample to observations for which the data on parcel size and/or building footprint are of the highest quality.

Taking logs and the derivative with respect to distance from the closest city centre, equation (3) yields

$$\begin{aligned}\frac{\partial \ln \frac{\text{Establishment employment}}{\text{Establishment parcel size}}}{\partial \text{distance}} = & \frac{\partial \ln \frac{\text{Establishment employment}}{\text{Establishment floorspace}}}{\partial \text{distance}} \\ & + \frac{\partial \ln \text{Number of floors}}{\partial \text{distance}} + \frac{\partial \ln \frac{\text{Establishment building footprint}}{\text{Establishment parcel size}}}{\partial \text{distance}},\end{aligned} \tag{5}$$

which is the distance-gradient version of decomposition (2). Letting $\widehat{\alpha}_{dist}$ denote the estimate of the component of $\boldsymbol{\alpha}$ with respect to distance from the closest city centre, we thus have: $\widehat{\alpha}_{dist}^{\text{Industrial density}} = \widehat{\alpha}_{dist}^{\text{Crowding}} + \widehat{\alpha}_{dist}^{\text{Building height}} + \widehat{\alpha}_{dist}^{\text{Parcel coverage}}$.

coverage, either because of the size of the parcels available close to this infrastructure or because of how 'packed' establishments accept to be in order to enjoy proximity to this infrastructure.

[19] There are 76 economic regions in Canada. They correspond to census division aggregates and are a standard geographic unit for analysis of regional economic activity.

3.2 Baseline results

3.2.1 Industrial density: Establishment employment over parcel size

Table 1 summarizes results for six regressions of industrial density (establishment employment over parcel size) on environmental and individual establishment characteristics. Column (1) includes only the main covariates of interest for the geographic environment of the establishment (CMA population size and surface area, and distance to the nearest city centre). Column (2) adds the log employment of the establishment to the set of regressors. Column (3) adds the other individual characteristics of the establishment (especially HQ and exporter dummies), and column (4)—our preferred specification that we will use later for the decomposition and quantification exercises—adds controls for the distance to transport infrastructure. Column (5) takes a more local perspective on density and adds the log population density in a 500 metre radius around the establishment to the set of environmental characteristics. Last, column (6) presents standardized (beta) coefficients for our benchmark specification (4) to convey an idea of the economic significance of the correlations we highlight.

The results in Table 1 exhibit several clear patterns. First, plants in larger urban areas (as measured by CMA population) and plants closer to city centres (as measured by straight-line distance) occupy their parcels more densely, i.e., use less land per worker.[20] This certainly reflects the fact that land prices are higher in big cities and closer to city centres. In our preferred specification (4), the elasticity of industrial density to city population equals 0.135 and the semi-elasticity of parcel size to the distance to the closest city centre equals 0.027. Both coefficients decrease in absolute value when we control for the population density within 500 metres of the establishment in column (5), with CMA size becoming insignificant. This reflects variation in land prices within cities, with dense areas being on average closer to city centres and more expensive. Note that the effect of distance from the centre, though weaker, remains statistically significant, even when controlling for local density.

Insert Table 1 and Figure 1 about here.

Turning to establishment characteristics, headquarters occupy their parcels more

[20] Although the point estimates for CMA surface are negative, they are imprecisely measured.

densely in terms of employment, while the opposite is true for exporters: 'office' functions require less space than functions related to production and exports for which factory space and warehousing are more important. As shown in columns (2)–(6), the by far most important individual characteristic is establishment size (in terms of employees), with an elasticity between 0.62 and 0.65 depending on the specification: larger establishments have a higher industrial density. We see three possible explanations for this finding.

First, as already mentioned, the Scott's data are exhaustive for manufacturing but not for services. Hence, we possibly mismeasure the number of neighbors on the parcel. This problem is potentially more severe for small firms that are more likely to share their location with other businesses, thus leading us to underestimate their industrial density. As shown in Figure 1, when we run our benchmark regression separately by establishment-size bins, the correlation between parcel size per worker and establishment size is close to 1 for establishments with 1–5 employees, and close to 0.65, the coefficient found for the whole sample, for establishments with 5–15 and 15–50 employees. Larger establishments (50+ employees) have a coefficient of 0.55. This pattern is inconsistent with the idea that the large positive correlation estimated for the whole sample stems from an underestimation of industrial density for smaller establishments.[21]

Second, some manufacturing establishments may occupy buildings with several floors. It is likely that larger establishments occupy taller buildings but not necessarily much bigger parcels, which would then show up in a higher establishment employment over parcel size ratio. Using detailed data for Montréal, we find in unreported regressions that controlling for the number of floors does not affect the coefficient on

[21] Figure 1 suggests that there may be large adjustment costs for changing land consumption. Moving, opening, or closing an establishment is indeed costly so that firms adjust their land use sluggishly; only when shocks are large enough do firms adjust land use, most likely by moving or by opening and closing establishments (Bergeaud and Ray, 2021). When firms grow or shrink following transitory shocks, they do so by first adjusting their number of employees and, eventually, later their land use. If large firms have grown more relative to their initial size, the positive correlation between the establishment employment over parcel size and establishment size could be related to the existence of adjustment costs. We ran regressions controlling for plant-level employment growth between 2013 and 2017, and the coefficient on establishment size is unaffected. These results are available upon request.

the number of workers, which remains positive, large, and highly significant. Hence, differences in building height (and thus floorspace) do not explain that result.

Last, land obviously has some characteristics of a fixed cost. It is costly to negotiate the lease or purchase, to assemble parcels, to make changes to the land, and to prepare it for use (decontamination, teardown of existing structures). Furthermore, concerning the buildings, some parts such as corridors, bathrooms, and meeting rooms have a size that is partly independent of the number of workers using them. If land were a pure variable cost, under usual functional forms of the production function such as Cobb-Douglas or CES, land per worker would be independent of firm size, which runs counter to our results in Table 1. Thus, our findings have implications for how to model land use for production. We come back to this point in Section 4.

Note finally that, from a quantitative perspective, the R^2s of our regressions are fairly large, above 0.55 in our preferred specification.[22] Thus, although we work with micro data at the establishment level, the empirical model explains a substantial part of the variation in establishment-level industrial density. The standardized coefficients in column (6) show that establishment size, the population of the CMA, and the distance to the closest city centre have first-order effects on plants' land use per worker.

3.2.2 Parcel coverage: Building footprint-to-parcel size

We now turn to parcel coverage to understand how the built density of industrial land changes with environmental and individual establishment characteristics. Table 2 summarizes the results of six regressions of building footprint-to-parcel size on those different characteristics. As shown, two of the four main determinants for industrial density are also important determinants of parcel coverage: CMA population and the distance to the nearest city centre. Building footprint-to-parcel size increases with city size and decreases with the distance to the nearest centre. In a nutshell, establishments in larger cities and closer to the centre occupy buildings that cover a larger share of the parcels they are built on. However, and contrary to industrial density, all else equal parcel coverage is only modestly correlated with establishment size.

Insert Table 2 about here.

[22] We checked that these R^2s are not solely driven by the industry- and economic region fixed effects.

Bearing in mind that land is more expensive in large cities and closer to city centres, the observation that parcel coverage decreases with distance to the centre and in smaller cities suggests that establishments use more outdoor space compared to indoor space when land prices are lower. Outdoor space may include, e.g., parking lots, open-air storage, or green space. It is important to recognize that outdoor space has some contribution to the establishment's production. However, being probably less central to production than indoor space, firms can restrict their use of outdoor space when land prices are high, especially when other arrangements can provide the services otherwise provided by outdoor space. We know, for example, that the cost of surface parking increases with the value of land, which implies that firms and households save on land by investing in underground or structural parking when being closer to the city centre (Brueckner and Franco, 2017). Furthermore, city centres are better served by public transit, which allows firms to reduce parking space for their employees and cover their parcels more extensively with buildings. Outdoor space thus seems fairly reactive to changes in land prices, whereas indoor space may exhibit a stronger complementarity with the other production factors and may less easily be compressed (which will be corroborated in section 3.4 by the results on Montréal).

3.3 Robustness checks

Table 3 summarizes a battery of robustness checks based on specification (4) in Tables 1 and 2. To save space, we only present the coefficients for the three main variables of interest—CMA population, distance to the nearest city centre, and establishment size—but all covariates of the benchmark specification are included in the different regressions. We report nine different robustness checks. Column (1) shows the benchmark results as a point of comparison. In column (2), we expand the sample to include all establishments for which we have information on the dependent variable, irrespective of the quality of the geocoding and the polygon assignment. In column (3), we trim the bottom and top 1% of the distribution of the dependent variable. In column (4), we restrict the sample to observations with excellent quality for both the parcel size and building footprint measures. In column (5), we eliminate observations for which the parcel size is smaller than the building footprint. In column (6), we restrict the

sample to manufacturing establishments with less than 50 employees to mitigate the fact that large establishments may occupy several adjacent parcels, in which case we would underestimate the amount of space they use. In column (7), we restrict the sample to those establishments that have no identified neighbors on the same parcel or in the same building. Despite the fourth-order polynomial control for the number of neighbors, we may still mismeasure the actual amount of land occupied by establishments when several manufacturing firms occupy the same parcel. In the same vein, in column (8), when both parcel size and building footprints are available, we replace the number of neighbors by the average of the number of neighbors on the parcel and in the building. In column (9), we restrict the sample to establishments located farther than 5 kilometres from the nearest city centre to ensure that the patterns we uncover are not driven by what happens in very central locations. Finally, in column (10) we include CMA fixed effects to control for other systematic differences across metro areas. In that case, CMA size is of course no longer separately identified.

Insert Tables 3 and 4 about here.

For both industrial density and parcel coverage, Table 3 shows that the correlations with CMA population, distance to the nearest city centre, and establishment size in terms of employment are remarkably stable, both qualitatively and quantitatively. The only exception is the correlation between building footprint-to-parcel size and establishment size, which is less stable but generally close to zero. We are thus confident that our key results are robust: controlling for establishment size, manufacturing establishments have higher industrial density and parcel coverage in big cities and in central locations within cities. Moreover, controlling for distance to the centre and CMA size, larger establishments use less land per worker.

A second set of robustness checks is provided in Table 4. There, we use property assessment roll data for the city of Montréal to verify if our results are sensitive to the use of province-level parcel datasets and open source building polygons.[23] Using the geographic coordinates of the properties in the assessment roll, we merge this information with the data used for the core analysis for 1,115 establishments. Reassur-

[23]Property assessment rolls are decentralized in Canada and mostly available at the municipal level. Unfortunately, we could not collect similar data for the other provinces or cities.

ingly, when both variables are available, the correlation between the surface area of the parcels we have assigned to establishments in our analysis so far and the ones filled in the property assessment roll equals 0.89. This confirms that the spatial join procedure we have implemented generally allows us to recover reliable parcel size information. Table 4 shows that the main results are unchanged when using the Montréal sample. If anything, the distance gradients are a bit steeper and more precisely estimated, and remain significant even when local density is included as a covariate. To summarize, Montréal is representative of the whole Canadian sample. This observation is important for the following analysis focusing on crowding, i.e., the number of workers per unit of floorspace, as it draws heavily on this sample.

3.4 Crowding: workers per unit of floorspace

Until now, we have provided estimates for industrial density and parcel coverage. We now leverage the Montréal assessment roll data to provide estimates for crowding (establishment employment over floorspace). Indeed, the assessment roll data contain information about the floorspace—and the number of floors—of the properties located on the parcels and allow thus for a direct measure of floorspace.

The results in column (1) of Table 5 show that, as for parcels, bigger establishments occupy floorspace more densely as the number of workers per unit of floorspace increases with establishment size. However, conditional on establishment size, floorspace is not significantly related to the distance to the closest city centre. This means that establishments do not reduce the amount of floorspace they use per worker when locating closer to city centres: they use less land (as measured by parcel size) but a similar amount of floorspace by occupying their parcels more densely, i.e., by using less outdoor space and/or increasing building height.

Insert Table 5 about here.

We next make use of the detailed floor information for establishments in Montréal to classify sectors into those that are 'low rise' and those that are 'high rise'. More precisely, we look at the distribution of the number of floors of the buildings occupied by different manufacturing industries and consider that all industries where the median number of floors is less than 1.6 are low rise industries. We then re-estimate our

model in column (2) using only the low rise industries for which the building height component in (3) is approximately one. We still find an insignificant effect of distance on floorspace per worker, though the standard error increases substantially due to the smaller sample size. Column (3) extends the analysis to Canada, using only the low-rise industries as identified from the Montréal data. As in column (1), the coefficient on distance from the closest city centre in column (3) is very close to zero. Thus, we conclude that crowding in Canadian manufacturing establishments, as measured by the ratio of workers to floorspace, is independent of distance from the city centre.

3.5 Quantifying the decomposition

We now check whether the decomposition (5) holds exactly or approximately in the data. To provide clean results, we use the data for Montréal and a consistent sample of the same establishments in the four separate regressions that we run. We separately estimate the four terms of the gradient decomposition of (5),

$$\widehat{\alpha}_{\text{dist}}^{\text{Ind. density}} = \frac{\partial \ln \frac{\text{employment}}{\text{parcel size}}}{\partial \text{dist}}, \qquad \widehat{\alpha}_{\text{dist}}^{\text{Crowding}} = \frac{\partial \ln \frac{\text{employment}}{\text{floorspace}}}{\partial \text{dist}},$$

$$\widehat{\alpha}_{\text{dist}}^{\text{Building height}} = \frac{\partial \ln \text{number of floors}}{\partial \text{dist}} \quad \text{and} \quad \widehat{\alpha}_{\text{dist}}^{\text{Parcel coverage}} = \frac{\partial \ln \frac{\text{building footprint}}{\text{parcel size}}}{\partial \text{dist}}$$

noting that:

$$\alpha_{\text{dist}}^{\text{Crowding}} = \widehat{\alpha}_{\text{dist}}^{\text{Ind. density}} \quad \widehat{\alpha}_{\text{dist}}^{\text{Building height}} \quad \widehat{\alpha}_{\text{dist}}^{\text{Parcel coverage}} \tag{6}$$

Table 6 summarizes the results. Columns (1), (2) and (4) mirror results previously presented, while column (3) shows that the number of floors is a mildly decreasing function of distance from the closest centre, as predicted by standard urban models.

Insert Table 6 about here.

Using the results in Table 6, we see that the decomposition (5) holds approximately in our data. Indeed, comparing columns (1) and (5) in Table 6, which report separate regressions of the left-hand side and the right-hand side of equation (6) on the same set of covariates, we see that the estimated coefficients are both close to zero and statistically insignificant. Columns (2)–(4) further show that the insignificant coefficient

on floorspace with respect to distance stems from the offsetting effects of three significant coefficients. Firstly, industrial density, as measured by employment over parcel size, decreases with distance to the centre (column (2)). Secondly, the number of floors decreases slightly as we move away from the centre (column (3)). Finally, parcel coverage decreases with distance to the centre, showing that parcel size increases faster than building footprints (column (4)). Put differently, all else equal (in particular, conditional on establishment size), floorspace per worker remains constant with respect to distance to the closest city centre because buildings are increasingly flatter with smaller groundfloor footprints as we move away from the centre, leaving workers with the same per capita floorspace but more outdoor space.

4 Some implications for theory

We now discuss in more details some implications of our empirical findings for modeling land for production, and then back out—through the lens our our theory—the value of the elasticity of substitution between land and labor that is compatible with our empirical results.

4.1 The canonical Cobb-Douglas production function

Consider the canonical setting with perfectly competitive factor markets.[24] Let w_z, r_z, and p_z denote the price of labor (L), capital (K), and land (H) in zone z (we do not specify here if land should be interpreted as parcel size or floorspace, but we come back to this distinction when we discuss the quantitative implications of the theoretical framework). Assuming a Cobb-Douglas production function, the output $Y_{i(s,z)}$ of firm i—operating in sector s and located in zone z—is given by:

$$Y_{i(s,z)} = A_{i(s,z)} L_i^{\ s} K_i^{\ s} H_i^{1 \quad s \quad s} \quad \text{so that} \quad \frac{L_i}{H_i} = \frac{\alpha_s}{1 \quad \alpha_s \quad \beta_s} \frac{p_z}{w_z}, \tag{7}$$

where $A_{i(s,z)}$ denotes TFP and α_s and β_s are cost shares.

Expression (7) has two testable implications. First, land per worker is independent of establishment size L_i. This property, however, contradicts our empirical finding that

[24] We remain agnostic as to competition on the product markets, which we do not model here.

the number of workers per unit of land significantly increases with establishment size.

Second, (7) predicts that $\frac{@\ln(L_i/H_i)}{@\text{Dist}_i}$ equals $\frac{@\ln(p_z/w_z)}{@\text{Dist}_i}$, i.e., the semi-elasticity of workers per unit of land with respect to distance from the centre equals that of the rent-wage ratio. As predicted by urban models and vindicated by substantial empirical analyses, the price of land is a decreasing function of distance from the centre, whereas wages for a given job seem to vary little across local labor markets.[25] Hence, $\frac{@\ln(p_z/w_z)}{@\text{Dist}_i} \quad \frac{@\ln p_z}{@\text{Dist}_i}$ seems to be a reasonable approximation. We do not have land price data to estimate land price gradients in Canadian cities. However, Albouy et al. (2018) provide estimates of the ratio of land values per acre in the city centre (0.5 miles from downtown) and 10 miles away from it for more than 300 urban areas in the US. The weighted average ratio equals 6.5 (using urban area population as weights), which corresponds to a semi-log gradient of $\frac{@\ln(p_z/w_z)}{@\text{Dist}_i} = \quad 0.197$.[26] This value is a magnitude larger than our estimates for $\frac{@\ln(L_i/H_i)}{@\text{Dist}_i}$, which range from $\quad 0.04$ to $\quad 0.02$ (see Table 1 and Table 3, panel (a)). Although Canadian and US cities are unlikely to have exactly the same land price gradients, the difference between our estimates and those in Albouy et al. (2018) is so large that it suggests that the elasticity of substitution between labor and land is (much) smaller than the one predicted by the Cobb-Douglas specification.

To summarize, neither the increase in the number of workers per unit of land with respect to firm size nor its very small changes with respect to distance from the centre seem to fit well with the standard Cobb-Douglas production function. Based on that observation, we now revisit the specification of the production function including land and then discuss its quantitative implications.

[25] Mulalic et al. (2014) use quasi-random variation in commuting distance due to firm relocations and find very little effects on wages (0.15% per kilometre after three years).

[26] Assuming that the log of land price linearly depends on the distance to the city centre, and since Albouy et al. (2018) estimate the ratio of land values at 0.5 and 10 mile from downtown to equal 6.5 on average, the gradient is given by $\quad \ln(6.5)/9.5 = \quad 0.197$.

4.2 Augmenting the production function

Assume that output is given by

$$Y_{i(s,z)} = A_i \left\{ \xi_s \left[\frac{H_i \quad \overline{H}_i}{\kappa_{i(s,z)}} \right]^{\frac{\sigma_s \quad 1}{\sigma_s}} + (1 \quad \xi_s) \left(L_i{}^s K_i^{1 \quad s} \right)^{\frac{\sigma_s \quad 1}{\sigma_s}} \right\}^{\frac{\sigma_s}{\sigma_s \quad 1}} \tag{8}$$

where capital and labor are aggregated using a Cobb-Douglas technology, and land and the capital-labor bundle are aggregated using a CES technology.[27] The elasticity of substitution (σ_s) between land and the other factors, as well as the technological land-intensity (ξ_s), are specific to industry s, whereas the land-augmenting productivity ($\kappa_{i(s,z)}$) is specific to firm i. There is a fixed land requirement ($\overline{H}_i$) for any level of output, reflecting the fact that some space must be used irrespective of the amount of output produced.

Taking the ratio of the first-order conditions with respect to land and labor yields:

$$\frac{L_i}{H_i \quad \overline{H}_i} = \left[\frac{\xi_s}{\alpha_s(1 \quad \xi_s)} \right]^{s} [\kappa_{i(s,z)}]^{s \quad 1} \left(\frac{p_z}{w_z} \right)^{s} \left(\frac{L_i}{K_i} \right)^{(1 \quad s)(1 \quad s)}.$$

Similarly, the ratio of the first-order conditions with respect to labor and capital yields:

$$\frac{L_i}{K_i} = \frac{\alpha_s}{1 \quad \alpha_s} \frac{r_z}{w_z}.$$

Log-linearizing the former relationship and substituting the latter, we obtain:

$$\begin{aligned} \ln\left(\frac{L_i}{H_i \quad \overline{H}_i}\right) = \quad & \sigma_s \ln\left[\frac{\xi_s}{\alpha_s(1 \quad \xi_s)}\right] + (\sigma_s \quad 1) \ln \kappa_{i(s,z)} \\ & + \sigma_s \ln\left(\frac{p_z}{w_z}\right) + (1 \quad \sigma_s)(1 \quad \alpha_s)\left[\ln\left(\frac{\alpha_s}{1 \quad \alpha_s}\right) + \ln\left(\frac{r_z}{w_z}\right)\right] \end{aligned} \tag{9}$$

When fixed costs $\overline{H}_i$ are not too large, the term $\ln(H_i \quad \overline{H}_i)$ can be approximated by $\ln(H_i \quad \overline{H}_i) \quad \ln H_i \quad \frac{\overline{H}_i}{H_i}$, so that we finally obtain:

$$\begin{aligned} \ln\left(\frac{L_i}{H_i}\right) = \quad & \underbrace{\sigma_s \ln\left[\frac{\xi_s}{\alpha_s(1 \quad \xi_s)}\right] + (1 \quad \sigma_s)(1 \quad \alpha_s)\left[\ln\left(\frac{\alpha_s}{1 \quad \alpha_s}\right)\right]}_{0s} \\ & + \sigma_s \ln\left(\frac{p_z}{w_z}\right) \quad \frac{\overline{H}_i}{H_i} + \underbrace{(\sigma_s \quad 1) \ln \kappa_{i(s,z)} + (1 \quad \sigma_s)(1 \quad \alpha_s) \ln\left(\frac{r_z}{w_z}\right) + \varepsilon_i}_{i} \end{aligned} \tag{10}$$

[27] When $\sigma_s \; ! \; 1$, (8) reduces to $Y_{i(s,z)} = A_i \kappa_{i(s,z)}^{\xi_s} H_i^{\xi_s} L_i^{\alpha_s(1 \quad \xi_s)} K_i^{(1 \quad \alpha_s)(1 \quad \xi_s)}$, which is isomorphic to the baseline case (7).

where ϵ_i is the structural error term (and ε_i a reduced-form error term, including the error from the approximation).

The main purpose of the econometric results in Section 3 was to provide a careful description of the correlations between land consumption and various characteristics of manufacturing establishments and their environment. We now discuss how our reduced-form empirical results relate to the theoretically grounded equation (10), and use the structure of the equation to discuss the biases our correlations may have if one were to interpret these correlations more structurally.

The industry-specific term β_{0s}, which implies that low-α_s and high-ξ_s sectors sort into places where land is relatively less expensive as they use relatively more land, is captured in our regressions by industry fixed effects. The term $\ln\left(\frac{p_z}{w_z}\right)$ varies both across and within metro areas, and in the absence of price and wage data for Canadian cities, two of the covariates in the previous regression analysis proxy for it: city population and the distance to the nearest city centre. Land prices increase faster than wages with city size (see, for example, in the French case, Combes et al., 2008, 2019), and they decrease with distance to the centre (wages varying little within metro-areas). Through the lens of equation (10), as long as $\sigma_s > 0$, we thus have:

$$\frac{\partial \ln\left(L_i/H_i\right)}{\partial \ln \text{Pop}_z} = \sigma_s \frac{\partial \ln\left(p_z/w_z\right)}{\partial \ln \text{Pop}_z} > 0 \quad \text{and} \quad \frac{\partial \ln\left(L_i/H_i\right)}{\partial \text{Dist}_i} = \sigma_s \frac{\partial \ln\left(p_z/w_z\right)}{\partial \text{Dist}_i} < 0,$$

which is exactly what we found in Section 3.

Finally, we do not have a direct measure of the fixed land requirement, but since the share of the fixed requirement in overall land consumption mechanically decreases with establishment size, the log size of the establishment included in the regression analysis is a proxy for $\frac{\overline{H}_i}{H_i}$. Put differently, $\frac{\overline{H}_i}{H_i}$ increases with establishment size, so that the correlation between the number of workers per unit of land and establishment size should be positive, which is also exactly what we found in Section 3.

How do the two remaining components in the structural error term of equation (10) affect the correlations we estimated in our regressions? First, the presence of the plant-specific land-augmenting productivity parameter, $\kappa_{i(s,z)}$, implies that high $\kappa_{i(s,z)}$-firms (i.e., firms with a low land-augmenting productivity) use relatively more land when $\sigma_s < 1$ (see our discussion in Section 4.1), and thus sort into places where land is relatively cheap. The battery of firm-level controls we introduced in the regression

analysis (dummies for headquarters and exporters, number of production lines and functions, proximity to different types of infrastructure) are one way to control as well as possible for this unobserved parameter. However, since firms that use relatively more land sort into places where land is relatively cheap (i.e., into small cities and far from city centres within cities), the possible remaining bias is such that (in absolute value) the estimated correlation between $\ln\left(\frac{L_i}{H_i}\right)$ on the one hand, and city size or distance to the centre on the other, are upper bounds of the 'real' correlations.

The structural error term also contains a function of the relative price of capital with respect to labor, $(1 \quad \sigma_S)(1 \quad \alpha_S)\ln\left(\frac{r_z}{w_z}\right)$. Assuming that capital markets are integrated in Canada, we can consider that the price of capital does not vary substantially across and within cities. On the other hand, wages increase with city size but do not vary within cities. Overall, it is thus reasonable to think that $\ln\left(\frac{r_z}{w_z}\right)$ decreases with city size but is unrelated to the distance from the city centre. Hence, the presence of $(1 \quad \sigma_S)(1 \quad \alpha_S)\ln\left(\frac{r_z}{w_z}\right)$ in the error term should not affect the estimated correlation between $\ln\left(\frac{L_i}{H_i}\right)$ and distance from the city centre. Table 3 shows that including city fixed effects does not affect the estimated correlation between industrial density and distance to the city centre, in line with the foregoing discussion. Under the plausible assumption that both $\sigma_S < 1$ and $\alpha_S < 1$, and ignoring possible interactions with $\frac{\overline{H}_i}{H_i}$, it will however lead to under-estimating the correlation with city size.

To conclude this discussion, the correlation we estimate between the number of workers per unit of land and the distance from the city centre is an upper bound in absolute value. However, the sign of the bias is ambiguous for the correlation with city size.

4.3 Quantitative implications

We now provide some tentative estimates for the value of σ_S. As already mentioned, through the lens of equation (10), we have

$$\frac{\partial \ln(L_i/H_i)}{\partial \ln \mathrm{Pop}_z} = \sigma_S \frac{\partial \ln(p_z/w_z)}{\partial \ln \mathrm{Pop}_z} \quad \text{and} \quad \frac{\partial \ln(L_i/H_i)}{\partial \mathrm{Dist}_i} = \sigma_S \frac{\partial \ln(p_z/w_z)}{\partial \mathrm{Dist}_i}.$$

Hence, we can construct an estimate $\widehat{\sigma}_S$ as follows:

$$\widehat{\sigma}_S = \frac{\partial \ln(L_i/H_i)}{\partial \ln \mathrm{Pop}_z} \Big/ \frac{\partial \ln(p_z/w_z)}{\partial \ln \mathrm{Pop}_z} = \frac{\partial \ln(L_i/H_i)}{\partial \mathrm{Dist}_i} \Big/ \frac{\partial \ln(p_z/w_z)}{\partial \mathrm{Dist}_i}. \tag{11}$$

Since we estimated a zero semi-elasticity of floorspace with respect to distance from the city centre, it follows that the elasticity of substitution between floorspace and other production factors equals 0, which amounts to a Leontief production function relating floorspace and the labor-capital bundle.

To operationalize (11), when land is interpreted as parcel size, we need estimates for the (semi-)elasticity of p_z/w_z with respect to distance from the centre and metropolitan population size. Lacking estimates for Canadian cities, we make use of the results and data from Albouy et al. (2018) for US metropolitan areas. As explained before, since wages do not vary much within metro areas with distance from the centre, we can take the estimate of 0.197 from Albouy et al. (2018) as the reference value for $\partial \ln(p_z/w_z)/\partial \text{Dist}_i$ for Canada. Using equation (11) and our estimates, it follows that $\widehat{\sigma}_s = 0.027/0.197 = 0.137$ (we use our preferred estimate $\frac{@\ln(H_i/L_i)}{@\text{Dist}_i} = 0.027$ from column (4) of Table 1).

Albouy et al. (2018) do not provide estimates of the elasticity of land values per acre with respect to metropolitan population size. We can, however, combine their data with population sizes to estimate it. We obtain an estimate of $\frac{@\ln p_z}{@\ln \text{Pop}_z} = 0.399$ using land values for the overall MSA and of $\frac{@\ln p_z}{@\ln \text{Pop}_z} = 0.856$ using land values for the MSA central cities. Since we need an estimate for $\frac{@\ln(p_z/w_z)}{@\ln \text{Pop}_z}$, we further require the elasticity of wages with respect to CMA population for US cities. Behrens and Robert-Nicoud (2015) estimate 0.081 (unconditional) or 0.042 (controlling for MSA education) for the population-wage elasticities across US cities. Combining this with the above estimates, we obtain a range from $\frac{@\ln(p_z/w_z)}{@\ln \text{Pop}_z} = 0.399 \quad 0.081 = 0.318$ to $\frac{@\ln(p_z/w_z)}{@\ln \text{Pop}_z} = 0.856 \quad 0.042 = 0.814$. Using our preferred estimate $\frac{@\ln(L_i/H_i)}{@\ln \text{Pop}_z} = 0.135$ from column (4) of Table 1 we obtain the following range: $\widehat{\sigma}_s \in [0.166, 0.425]$.[28]

[28] As an additional check, based on French data, Combes et al. (2019) find that the elasticity of the price of parcels (per square metre) to city population is roughly equal to 0.6, while using French data too, Combes et al. (2008) find an elasticity of individual wages to population density of 0.03. These two elasticities are not estimated for the same period and at the exact same spatial scale, but they are cleanly estimated with very detailed data. Hence, the elasticity of relative land prices to city size/city population density in France equals 0.57. Taking this as a reference value for Canada, equation (11) implies a value of $\widehat{\sigma}_s = 0.135/0.57 = 0.237$, which falls in the range of the values obtained using the estimates of Albouy et al. (2018) with US data.

The range of σ_S implied by our quantification exercise is 0.14 to 0.42, with three out of four values being below 0.25. These values are far from 1 that obtains under the ubiquitous Cobb-Douglas specification used in much of the existing literature. They suggest that labor and land, as measured by parcel size, are little substitutable in the production function of manufacturing establishments.[29]

5 Conclusions

We have constructed a new quantity dataset for Canadian manufacturing establishments containing measures of the parcel sizes they occupy and the floorspace of their buildings. Using those data, we have decomposed industrial density (parcel size per worker) into three components: crowding (floorspace per worker); building height (floorspace to building footprint); and parcel coverage (building footprint to parcel size). Using these components, we have estimated their elasticity with respect to city size and distance to the city centre to see how density changes and along which margins manufacturing firms modulate their land consumption.

Our results show that, controlling for establishment size, manufacturing establishments occupy parcels more densely in big cities and in central locations within cities. This is the industrial analogue of the results on residential density substantiated by the urban economics literature. This result holds both in terms of industrial density (employment over parcel size) and parcel coverage (building footprint over parcel size). We further find that, controlling for location, larger establishments use parcels more densely: employment-to-parcel size increases strongly with establishment size (but parcel coverage does not depend much on establishment size). Last, our results show that floorspace per worker appears unrelated to distance from the centre. This latter point suggests that much of the adjustment in terms of land consumption is related to outdoor space such as parking, storage, or green space. To our knowledge this aspect,

[29] In the conceptual framework, we assumed σ_s is sector specific since there is no reason a priori to believe that land and labor are equally substitutable in all sectors. To see whether the average σ_s masks heterogeneity, we investigate the cross-sectoral heterogeneity in the two elasticities we can estimate from our data. Table C2 in the online appendix shows that the sectoral variation is limited, which implies that the sectoral variation in σ_s is limited too.

has not been investigated in more detail until now.

Our results have implications for modeling firms' land consumption. Since floorspace does not change with distance from the centre, hence with land prices, this suggests that floorspace and labor are used in (industry-specific) fixed proportions, as in a Leontief production function. Furthermore, there seems to be a strong fixed costs component for land. Our key stylized facts can be replicated with a production function that has fixed land requirements and that aggregates land- and non-land factors with a CES specification. Using that framework, we have finally quantified the implied elasticity of substitution between land- and non-land factors. As expected, this elasticity is low at 0.14 to 0.42, way below the value of 1 implied by the Cobb-Douglas specification.

We hope that our results will be useful to other researchers interested in understanding how industrial activity uses land and along which margins firms can adjust their consumption of land and floorspace. This seems especially important for the recent quantitative spatial models that largely use Cobb-Douglas production functions, as the results may be sensitive to that choice. We view our analysis as a first step in the direction of better understanding land use by firms. Our results may be specific to manufacturing and not transpose to other industries such as services and retail. It would thus be important to extend our analysis to economic activity more broadly defined. Although the data are increasingly available, this is likely a challenging task.

References

Ahlfeldt, Gabriel M. and Daniel P. McMillen, "Tall Buildings and Land Values: Height and Construction Cost Elasticities in Chicago, 1870-2010," *The Review of Economics and Statistics*, 2018, *100* (5), 861–875.

Albouy, David, Gabriel Ehrlich, and Minchul Shin, "Metropolitan Land Values," *The Review of Economics and Statistics*, July 2018, *100* (3), 454–466.

Angel, Shlomo, Patrick Lamson-Hall, and Zeltia Gonzalez Blanco, "Anatomy of density: measurable factors that constitute urban density," *Buildings & Cities*, 2021, *2* (1), 264–282.

Barr, Jason and Jeffrey P. Cohen, "The floor area ratio gradient: New York City, 1890–2009," *Regional Science and Urban Economics*, 2014, *48*, 110–119.

Behrens, Kristian and Frédéric Robert-Nicoud, "Agglomeration theory with heterogeneous agents," *Handbook of regional and urban economics*, 2015, *5*, 171–245.

Bergeaud, A. and S. Ray, "Adjustment Costs and Factor Demand: New Evidence From Firms' Real Estate," *Economic Journal*, 2021, *131* (633), 70–100.

Brooks, Leah and Byron Lutz, "From Today's City to Tomorrow's City: An Empirical Investigation of Urban Land Assembly," *American Economic Journal: Economic Policy*, 2016, *8* (3), 69–105.

Brueckner, Jan K, "The economics of urban yard space: An "implicit-market" model for housing attributes," *Journal of Urban Economics*, 1983, *13* (2), 216–234.

Brueckner, Jan K. and Sofia F. Franco, "Parking and Urban Form," *Journal of Economic Geography*, 2017, *17* (1), 95–127.

_ , Shihe Fu, Yizhen Gu, and Junfu Zhang, "Measuring the Stringency of Land Use Regulation: The Case of China's Building Height Limits," *The Review of Economics and Statistics*, July 2017, *99* (4), 663–677.

Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew A. Turner, "Causes of Sprawl: A Portrait from Space," *The Quarterly Journal of Economics*, 05 2006, *121* (2), 587–633.

Cheshire, Paul C., Christian AL Hilber, and Ioannis Kaplanis, "Land use regulation and productivity–land matters: evidence from a UK supermarket chain," *Journal of Economic Geography*, 2014, *15* (1), 43–73.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon, "Spatial wage disparities: Sorting matters!," *Journal of urban economics*, 2008, *63* (2), 723–742. Publisher: Elsevier.

_ , _ , and _ , "The Costs of Agglomeration: House and Land Prices in French Cities," *The Review of Economic Studies*, July 2019, *86* (4), 1556–1589.

Davis, Amélie Y., Bryan C. Pijanowski, Kimberly Robinson, and Bernard Engel, "The environmental and economic costs of sprawling parking lots in the United States," *Land Use Policy*, 2010, *27* (2), 255–261.

Drennan, Matthew P. and Hugh F. Kelly, "Measuring urban agglomeration economies with office rents," *Journal of Economic Geography*, 2010, *11* (3), 481–507.

Duranton, Gilles and Diego Puga, "Urban land use," in "Handbook of Regional and Urban Economics (Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.)," Vol. 5, Elsevier, 2015, pp. 467–560.

Epple, Dennis, Brett Gordon, and Holger Sieg, "A new approach to estimating the production function for housing," *American Economic Review*, 2010, *100* (3), 905–24.

Fujita, Masahisa and Hideaki Ogawa, "Multiple equilibria and structural transition of non-monocentric urban configurations," *Regional science and urban economics*, 1982, *12* (2), 161–196.

Gyourko, Joseph, Jonathan S Hartley, and Jacob Krimmel, "The local residential land use regulatory environment across US housing markets: Evidence from a new Wharton index," *Journal of Urban Economics*, 2021, *124*, 103337.

Haskel, Jonathan and Raffaella Sadun, "Regulation and UK retailing productivity: evidence from microdata," *Economica*, 2012, *79* (315), 425–448.

Imbert, Clément, Marlon Seror, Yifan Zhang, and Yanos Zylberberg, "Migrants and Firms: Evidence from China," *American Economic Review*, 2022, *112* (6), 1885–1914.

Liu, Crocker H., Stuart Rosenthal, and William Strange, "The vertical city: Rent gradients, spatial structure, and agglomeration economies," *Journal of Urban Economics*, 2018, *106* (C), 101–122.

Lucas, Robert E. and Esteban Rossi Hansberg, "On the internal structure of cities," *Econometrica*, 2002, *70* (4), 1445–1476.

Mayneris, Florian, "Does the urban wage premium imply higher firm-level labor shares in cities?," Technical Report 2022.

Moulton, Brent R, "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," *The Review of Economics and Statistics*, May 1990, *72* (2), 334–338.

Mulalic, Ismir, Jos N Van Ommeren, and Ninette Pilegaard, "Wages and commuting: Quasi-natural experiments' evidence from firms that relocate," *The Economic Journal*, 2014, *124* (579), 1086–1105.

Narro, Delgado, Augusto Ricardo, and Yuya Katafuchi, "Decomposition of density into their components: Analysis for the case of Japan," Technical Report 2020.

Oberfield, Ezra and Devesh Raval, "Micro Data and Macro Technology," *Econometrica*, 2021, *89* (2), 703–732.

Ogawa, Hideaki and Masahisa Fujita, "Equilibrium land use patterns in a nonmonocentric city," *Journal of regional science*, 1980, *20* (4), 455–475.

Pagano, Michael A and Ann O'M Bowman, *Vacant land in cities: An urban resource*, Brookings Institution, Center on Urban and Metropolitan Policy Washington, Survey Series, 2000.

Pflüger, Michael and Takatoshi Tabuchi, "The size of regions with land use for production," *Regional Science and Urban Economics*, 2010, *40* (6), 481–489.

Redding, Stephen J and Esteban Rossi-Hansberg, "Quantitative spatial economics," *Annual Review of Economics*, 2017, *9* (1), 21–58.

Seidel, Andre and Melanie Krause, "Unlocking Neighborhood Density," *Journal of Urban Economics, forthcoming*, 2024.

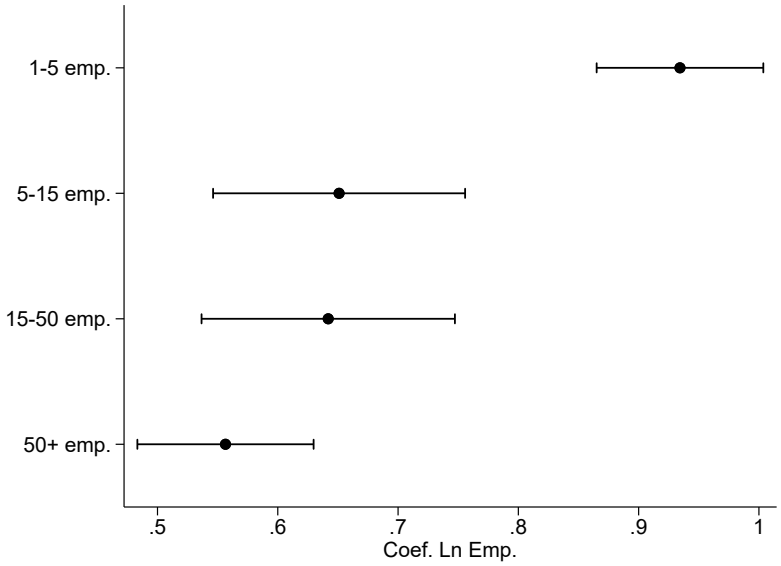
Wrede, Matthias, "Heterogeneous skills and homogeneous land: segmentation and agglomeration," *Journal of Economic Geography*, 2013, *13* (5), 767–798.

Table 1: Determinants of industrial density (establishment employment over parcel size).

Dependent variable	Ln establishment employment over parcel size					
	(1)	(2)	(3)	(4)	(5)	(6)
Environmental characteristics						
Ln CMA population	0.217[a]	0.193[a]	0.193[a]	0.135[a]	0.051	0.143
	(0.055)	(0.042)	(0.042)	(0.046)	(0.039)	
Ln CMA surface area	-0.131	-0.108	-0.111[c]	-0.111	-0.030	-0.061
	(0.087)	(0.065)	(0.064)	(0.078)	(0.060)	
Distance to the closest city centre	-0.039[a]	-0.037[a]	-0.037[a]	-0.027[a]	-0.020[a]	-0.128
	(0.005)	(0.006)	(0.006)	(0.008)	(0.005)	
Ln Population density within 500m					0.150[a]	
					(0.013)	
Individual characteristics						
Ln employment		0.622[a]	0.627[a]	0.629[a]	0.649[a]	0.601
		(0.016)	(0.015)	(0.015)	(0.017)	
1 is HQ			0.077[a]	0.079[a]	0.057[a]	0.017
			(0.021)	(0.021)	(0.021)	
1 is exporter			-0.116[a]	-0.111[a]	-0.089[a]	-0.038
			(0.023)	(0.022)	(0.020)	
Observations	8,704	8,704	8,704	8,704	8,704	8,705
R^2	0.268	0.555	0.557	0.562	0.585	0.562

Notes: OLS regressions. 1 denotes $f0, 1g$ dummy variables. All regressions include industry (4-digit) fixed effects, economic region fixed effects, dummies for local land-use category, and a forth-order polynomial in the number of neighbors of the establishment on its parcel. Columns (3)–(6) include, but do not report, the following individual controls (# functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment). Columns (4)–(6) include, but do not report, the infrastructure distance controls (Ln distance to major airport, Ln distance to major seaport, Ln distance to freight station, and Ln distance to highway junction). Column (6) reports beta coefficients and therefore omits standard errors. We include only observations of excellent quality for parcel size. Standard errors in parentheses are clustered at the CMA/CA level. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Figure 1: Industrial density (establishment employment over parcel size) and establishment size across size bins.



Notes: This figure shows the coefficient and the 95% confidence interval on establishment size for the benchmark regression in column (4) of Table 1 run separately for each employment-size bin.

Table 2: Determinants of parcel coverage (establishment building footprint over parcel size).

Dependent variable	Ln establishment building footprint over parcel size					
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of the local environment						
Ln CMA population	0.313^a	0.313^a	0.315^a	0.287^a	0.226^a	0.355
	(0.063)	(0.063)	(0.063)	(0.078)	(0.074)	
Ln CMA surface area	-0.203^c	-0.203^c	-0.206^c	-0.242^c	-0.183	-0.155
	(0.107)	(0.107)	(0.107)	(0.124)	(0.114)	
Distance to the closest city centre	-0.044^a	-0.044^a	-0.044^a	-0.033^a	-0.028^a	-0.184
	(0.006)	(0.006)	(0.006)	(0.010)	(0.008)	
Ln population density 500m					0.108^a	
					(0.020)	
Characteristics of the establishment						
Ln employment		-0.024^a	-0.022^a	-0.024^a	-0.009	-0.027
		(0.008)	(0.007)	(0.007)	(0.008)	
			(0.036)	(0.034)	(0.034)	
1 is exporter			-0.019	-0.018	-0.003	-0.007
			(0.019)	(0.020)	(0.018)	
Observations	8,510	8,510	8,510	8,510	8,510	8,511
R^2	0.271	0.272	0.272	0.283	0.300	0.283

Notes: OLS regressions. 1 denotes $f0, 1g$ dummy variables. All regressions include industry (4-digit) fixed effects, economic region fixed effects, dummies for local land-use category, and a forth-order polynomial in the number of neighbors of the establishment on its parcel. Columns (3)–(6) include, but do not report, the following individual controls (# functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment). Columns (4)–(6) include, but do not report, the infrastructure distance controls (Ln distance to major airport, Ln distance to major seaport, Ln distance to freight station, and Ln distance to highway junction). Column (6) reports beta coefficients and therefore omits standard errors. We include only observations of excellent quality for parcel size. Standard errors in parentheses are clustered at the CMA/CA level. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table 3: Robustness checks for sample selection, data quality, and CMA fixed effects.

(a) Industrial density:	Ln establishment employment over parcel size									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Ln CMA population	0.135^a	0.125^a	0.138^a	0.135^a	0.089^b	0.131^a	0.119^b	0.150^a	0.104^c	
	(0.046)	(0.036)	(0.043)	(0.050)	(0.039)	(0.048)	(0.049)	(0.048)	(0.058)	
Distance to the closest city centre	-0.027^a	-0.021^a	-0.024^a	-0.025^a	-0.019^a	-0.028^a	-0.025^a	-0.028^a	-0.009^b	-0.026^a
	(0.008)	(0.004)	(0.007)	(0.008)	(0.006)	(0.008)	(0.008)	(0.009)	(0.004)	(0.010)
Ln employment	0.629^a	0.648^a	0.569^a	0.627^a	0.628^a	0.707^a	0.546^a	0.634^a	0.628^a	0.628^a
	(0.015)	(0.016)	(0.014)	(0.014)	(0.018)	(0.017)	(0.014)	(0.018)	(0.020)	(0.014)
Observations	8,704	12,143	8,528	8,045	7,423	6,944	5,331	8,704	5,040	8,701
R^2	0.562	0.507	0.543	0.565	0.588	0.537	0.450	0.537	0.574	0.567

(b) Parcel coverage:	Ln establishment building footprint over parcel size									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Ln CMA population	0.287^a	0.276^a	0.222^a	0.275^a	0.220^a	0.291^a	0.232^a	0.279^a	0.233^b	
	(0.078)	(0.061)	(0.065)	(0.075)	(0.045)	(0.082)	(0.087)	(0.073)	(0.115)	
Distance to the closest city centre	-0.033^a	-0.028^a	-0.029^a	-0.031^a	-0.022^a	-0.036^a	-0.031^a	-0.033^a	-0.007^c	-0.032^a
	(0.010)	(0.007)	(0.008)	(0.010)	(0.005)	(0.011)	(0.010)	(0.009)	(0.004)	(0.012)
Ln employment	-0.024^a	-0.019^c	-0.004	-0.028^a	0.016^c	0.009	-0.023^b	-0.012	0.011	-0.026^a
	(0.007)	(0.011)	(0.011)	(0.006)	(0.010)	(0.017)	(0.011)	(0.008)	(0.015)	(0.007)
Observations	8,510	11,797	8,334	8,045	7,423	6,827	5,204	8,510	4,913	8,506
R^2	0.283	0.256	0.273	0.278	0.225	0.296	0.303	0.283	0.216	0.296

Notes: OLS regressions. 1 denotes $f0, 1g$ dummy variables. (1) Benchmark; (2) no quality restrictions; (3) 1% trimming; (4) both excellent; (5) coverage < 1; (6) less than 50 employees; (7) no neighbors; (8) parcel and building neighbors; (9) more than 5km from centre; (10) include CMA fixed effects. All regressions include industry (4-digit) fixed effects, economic region fixed effects, and the following controls (not shown): dummies for local land-use category; # functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment; Ln distance to major airport, Ln distance to major seaport, Ln distance to freight station, and Ln distance to highway junction. All regressions include a polynomial of degree 4 in the number of neighbors of the establishment on its parcel. We include only observations with the most reliable information on parcel size. Column (6) reports beta coefficients and omits standard errors. Standard errors in parentheses are clustered at the CMA/CA level. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table 4: Robustness checks using assessment roll data for Montréal.

Dependent variable	Ln establishment employment over parcel size			Ln establishment building footprint over parcel size		
	Original data	Roll data	Roll data	Original data	Roll data	Roll data
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of the local environment						
Distance to the closest city centre	-0.037^a	-0.056^a	-0.039^a	-0.039^a	-0.057^a	-0.049^a
	(0.009)	(0.010)	(0.011)	(0.009)	(0.011)	(0.011)
Ln population density 500m			0.198^a			0.087^b
			(0.035)			(0.035)
Characteristics of the establishment						
Ln employment	0.637^a	0.611^a	0.643^a	-0.087^a	-0.114^a	-0.099^a
	(0.028)	(0.033)	(0.032)	(0.026)	(0.034)	(0.034)
1 is HQ	0.241^b	0.344^a	0.321^a	0.071	0.183	0.172
	(0.098)	(0.116)	(0.112)	(0.100)	(0.128)	(0.127)
1 is exporter	-0.076	-0.144^c	-0.140^c	-0.050	-0.128	-0.126
	(0.068)	(0.083)	(0.082)	(0.063)	(0.090)	(0.090)
Observations	1,041	1,041	1,041	1,028	1,028	1,028
R^2	0.636	0.546	0.570	0.341	0.278	0.284

Notes: The regression sample only includes establishments in Montreal. OLS regressions. 1 denotes $f0, 1g$ dummy variables. All regressions include industry (4-digit) fixed effects, economic region fixed effects, and the following controls (not shown): dummies for local land-use category; # functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment; Ln distance to major airport, Ln distance to major seaport, Ln distance to freight station, and Ln distance to highway junction. All regressions include a polynomial of degree 4 in the number of neighbors of the establishment on its parcel. We include only observations with the most reliable information on parcel size. Column (6) reports beta coefficients and omits standard errors. Standard errors in parentheses are clustered at the CMA/CA level. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table 5: Crowding with Montréal assessment roll data and our Canada-wide sample of 'low-rise sectors'.

Dependent variable:	Ln Establishment employment over floorspace		
	Roll data MTL	Roll data MTL	Canada-wide sample
	(1)	(2)	(3)
Characteristics of the local environment			
Ln CMA population			-0.080^{c}
			(0.044)
Ln CMA surface area			0.022
			(0.065)
Distance to the closest city centre	-0.009	-0.040	0.006
	(0.013)	(0.026)	(0.006)
Characteristics of the establishment			
Ln employment	0.686^{a}	0.696^{a}	0.658^{a}
	(0.034)	(0.064)	(0.022)
1 is HQ	0.055	-0.100	0.050
	(0.115)	(0.270)	(0.064)
1 is exporter	-0.142^{c}	-0.126	-0.160^{a}
	(0.080)	(0.153)	(0.037)
Observations	873	188	2,191
R^2	0.719	0.683	0.537

Notes: (1) Sample includes establishments in Montreal with floorspace information; (2) sample includes establishments in Montreal associated with 'single floor (flat)' industries; (3) sample includes establishments in Canada associated with 'single floor (flat)' industries. OLS regressions. 1 denotes $\{0,1\}$ dummy variables. All regressions include industry (4-digit) fixed effects, economic region fixed effects, and the following controls (not shown): dummies for local land-use category; # functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment; Ln distance to major airport, Ln distance to major seaport, Ln distance to freight station, and Ln distance to highway junction. All regressions include a polynomial of degree 4 in the number of neighbors of the establishment on its parcel. We include only observations with the most reliable information on parcel size. Standard errors in parentheses are clustered at the CMA/CA level. $^{a}p < 0.01$, $^{b}p < 0.05$, $^{c}p < 0.1$.

Table 6: Decomposing the density distance gradient.

Dependent variable	Ln Establishment employment over floorspace	Ln Establishment employment parcel size	Ln number of floors	Ln building footprint over parcel size	(2)-(3)-(4)
	(1)	(2)	(3)	(4)	(5)
Characteristics of the local environment					
Distance to the closest city centre	-0.001	-0.052^{a}	-0.014^{b}	-0.052^{a}	0.014
	(0.014)	(0.012)	(0.006)	(0.013)	(0.014)
Characteristics of the establishment					
Ln employment	0.688^{a}	0.657^{a}	-0.027	-0.071^{c}	0.755^{a}
	(0.036)	(0.036)	(0.019)	(0.040)	(0.043)
1 is HQ	-0.007	0.101	0.117^{c}	0.119	-0.136
	(0.125)	(0.124)	(0.061)	(0.141)	(0.158)
1 is exporter	-0.172^{b}	-0.168^{c}	0.008	-0.148	-0.029
	(0.085)	(0.096)	(0.042)	(0.097)	(0.089)
Observations	723	723	723	723	723
R^2	0.758	0.631	0.328	0.725	0.771

Notes: The regression sample only includes establishments in Montreal. To provide the decomposition, we restrict ourselves to the largest sample that allows for estimation of each component. OLS regressions. 1 denotes $f0,1g$ dummy variables. All regressions include industry (4-digit) fixed effects, economic region fixed effects, and the following controls (not shown): dummies for local land-use category; # functions in the establishment, # of 4-digit NAICS codes of the establishment, and # of products produced by the establishment; Ln distance to major airport, Ln distance to freight station, and Ln distance to highway junction. All regressions include a polynomial of degree 4 in the number of neighbors of the establishment on its parcel. We include only observations with the most reliable information on parcel size. The dependent variable in column (5) is the difference between the dependent variable in column (2) and those in columns (3) and (4). Standard errors in parentheses are clustered at the CMA/CA level. $^{a}p < 0.01$, $^{b}p < 0.05$, $^{c}p < 0.1$.

Online Appendix

A Data Appendix

Census geography. Since we largely draw on the standard Canadian census geography, which may be little known to many readers, we provide here some information on its fundamental structure and concepts. A census metropolitan area (CMA) or a census agglomeration (CA) is formed by one or more adjacent municipalities centered on a population center (known as the core). A CMA must have a total population of at least 100,000 of which 50,000 or more must live in the core based on adjusted data from the previous Census of Population Program. A CA must have a core population of at least 10,000 also based on data from the previous Census of Population Program. To be included in the CMA or CA, other adjacent municipalities must have a high degree of integration with the core, as measured by commuting flows derived from data on place of work from the previous Census Program. An economic region (ER) is a grouping of complete census divisions (CDs), created as a standard geographic unit for analysis of regional economic activity. There are 76 economic regions in Canada that constitute a partition of the country. They are much smaller than provinces but, except for the very largest metropolitan areas, much bigger than cities. Finally, the 10 provinces and 3 territories are the federal political units in Canada.

Census data is published at different geographic levels. For our work, we rely on the most disaggregated public data, which is published at the dissemination area (DA) level. A DA is a small area that is composed of one or more neighbouring dissemination blocks. It has about 400 to 700 residents.

Geocoding. The recent years of the Scott's dataset already reports geographic coordinates for each plant. However, some of these are based on postal code centroids, which are necessarily less accurate than coordinates obtained from rooftop geocoding and do not permit to precisely associate plants with parcel- or building polygons. We hence geocode all plants based on their address information.

Geocoding consists in providing an address to a geocoder—a particular Application

Programming Interface (API) used to recover geographic coordinates of addresses– which returns the latitude and longitude of the corresponding address. The geocoder also provides the address related to the coordinates of the points it returns so that we can verify if the input address and the return address match.

For the sake of precision, we use three different options to perform the geocoding. The first option uses the commercial API of the Google Map server to geocode each plant based on the address recorded in the Scotts database. The second option uses the same API but combines the company's name with the address as the input for the geocoder. In doing so, small errors in the address reported in the Scotts data can be corrected and the accuracy of the geocoding improved. The third option uses the point coordinates provided in the DMTI database, which is an extensive database containing more than 15 million feature points representing Canadian addresses and their related geographic coordinates with 'rooftop' precision.[30] We merge the Scotts addresses with the DMTI address using the API of ArcGIS, a commercial Geographic Information Systems (GIS) software.

Once we have geocoded the addresses, we compare the coordinates (latitude, longitude) returned by the three options and assign to each plant the coordinates that are most likely the accurate ones. Accuracy is based on two criteria: (i) the distances between the point coordinates yielded by the three options (so as to identify probable errors, i.e., points that are very far away from the other return values); and (ii) the match between the postal codes recorded in the Scotts database and the postal codes returned by the geocoder for each option (so as to keep only the points for which the postal code corresponds to the one recorded in the Scotts database). If several different points are returned for the same establishment, the coordinates retrieved from Google Maps based on the company name and the address are preferred to the coordinates obtained via Google Map using the address only, which are themselves preferred to the DMTI coordinates.

Finally, we construct a variable with three categories to grade the accuracy of the geocoding process for each plant based on how convergent the three options are in terms of establishment location. We retain only observations that are either 'rooftop'

[30] See `https://www.dmtispatial.com` for more information.

(i.e., exactly coded) or 'range interpolated' (i.e., interpolated based on a range of address numbers); we do not consider the rest (e.g., postal-code level) as being accurate enough to assign plants to polygons.

Data sources. We extensively explored existing open-access data sources on various websites and got in touch with several institutions to obtain information on parcel- and buildings polygons and footprints in Canada. The main relevant data sources for our work are the following:

- Statistics Canada, via the official website of the Canadian Government, provides several datasets including data on buildings that are open for public use.
- Some Assessment Rolls of different municipalities—which are in charge of computing the value of the tenure taxes based on the nature, the location, and the scope of the properties—provide open-access data.
- Cadastral information: Some provinces and cities in Canada do have information on the parcels where buildings are located.
- GIS databases of cities: The websites of some cities provide GIS data which record parcels polygons and/or footprints of buildings of their localities.
- Open source data on building footprints in Canada released by Microsoft: These datasets contain 12,663,475 building footprints covering all provinces and territories.[31]

Table A1 provides the complete list of polygon datasets that we collected along with the links where they can be accessed.

Polygon dataset quality. We collected polygon datasets from the above sources. These datasets come in different data formats (KML, shapefile, geodataset, etc) and are for different reference years. During their processing, we identified and solved the following challenges linked to the quality of the data:

[31] For additional information, see `https://blogs.bing.com/maps/2019-03/microsoft-releases-12-million-canadian-building-footprints-as-open-data`.

Table A1: Datasources for the parcel- and building polygons.

Locality	Coverage	Last update		Polygon type	Licence	Links
Locality	**Coverage**	**Last update**		**Polygon type**	**Licence**	**Links**
Alberta	AB province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Alberta	Banff	2017	Banff	Parcels	open data	http://banffmaps.ca/opendata/
Alberta	Winnipeg	2017	Winnipeg	Parcels	open data	https://data.winnipeg.ca/Assessment-Taxation-Corporate/Map-of-Assessment-Parcels/rt7t-3m4m
British Columbia	CB province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
British Columbia	CB province	2016	province	Parcels	Open data	https://catalogue.data.gov.bc.ca/dataset/parcelmap-bc-parcel-fabric
Manitoba	MB province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Manitoba	Brandon	2017	Brandon	Parcels	open data	http://opengov.brandon.ca/OpenDataService/opendata.html
New Brunswick	province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
New Brunswick	province	2019	province	Parcels	open data	https://gnb.socrata.com/api/geospatial/rzzg-85tb?method=export\
Newfoundland and Labrador	NL province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Newfoundland and Labrador	St john	2019	St John	Parcels	open data	http://catalogue-saintjohn.opendata.arcgis.com/
North-West Territories	NT territories	2019	province	Building footprints	opendata	http://opendata.yellowknife.ca
Nova Scotia	NS province	2019	province	Building footprints	open data	https://github.com/Microsoft/CanadianBuildingFootprints
Nunavut	NU territories	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Ontario	ON province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Ontario	Oshawa	2017	Oshawa	Parcels	open data	https://city-oshawa.opendata.arcgis.com/datasets?t=Durham\%20Housing
Ontario	York	2019	York	Parcels	open data	https://insights-york.opendata.arcgis.com/datasets/parcel
Ontario	Toronto	2017	Toronto	Parcels	open data	https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/
Ontario	Windsor	2017	Windsor	Parcels	open data	www.citywindsor.ca/opendata/Lists/OpenData/Attachments/20/Land\%20Parcels.kmz
Prince-Edward Island	PE province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Quebec	QC province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Quebec	QC province	2018	province	Parcels	InfoLot	UQAM data warehouse (https://appli.mern.gouv.qc.ca/infolot
Saskatchewan	Regina	2017	Regina	Parcels	open data	http://open.regina.ca/
Saskatchewan	SK province	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints
Yukon Territories	YT territories	2019	province	Building footprints	OSM/Statcan	https://github.com/Microsoft/CanadianBuildingFootprints

Notes: Listing of all the datasources that we use in the paper.

- Quality of the collected files: The polygon datasets we collected are not homogeneous. The formats of the files are not always the same and the reference units of the polygon datasets are different in some cases (feet, meters, etc.) and sometimes not indicated at all in the files. To solve this problem we converted all the files into shapefile format (.shp), harmonized the units to meters, and projected each dataset into a suitable coordinate system according to the position of the locality it refers to. We consider as a suitable coordinate system one which does not alter distances. In most cases, the 'Albers conic conformal system' is used, as generally recommended for Canada. We also construct for each polygon dataset the following key variables: a unique identifier, the surface area, and the number of neighbors of each polygon recorded in the dataset. The latter variable is useful to check for the quality of the area assignation process for each plant. The LIDAR dataset source gives the building footprints along with an estimation of the height of the building. These files report the minimum and the maximum height detected by the signal used to scan the space.

- Matching buildings to parcels: The polygon datasets we collected have two different features. The first one is the parcel-polygon that represents the amount of land used by a plant to host its main building and possibly some other spaces (auxiliary buildings, parking, storage, etc.). The second polygon type is the building-polygon that represents only the building of the plant. Theoretically, the building footprint should be included in the parcel outline. Yet, in some cases the building overlaps with more than one parcel. As a result, the surface of the building footprint is greater than the surface of the parcel to which its is related. We solve this issue by aggregating up all the parcels that overlap with the building.

Assignment to polygons. We have, on the one hand, a geocoded establishment-level dataset and, on the other hand, different polygon datasets. To merge them, we use the spatial join tools available in the open-source software Quantum GIS (QGIS) to map each plant to a polygon. More precisely, we overlay the polygon datasets (parcels and buildings) with the coordinate point layers representing the geocoded establishments.

Figure A1 shows an example of how the geocoded Scott's plants are overlaid on the building polygon layer for the spatial join process. Figure A2 shows the same for the building- and parcel polygons using the Montréal assessment roll data.

Figure A1: Polygon layer with overlaid geocoded establishments.

As is well known, spatial join can be a somewhat noisy process. Hence, not all plants fall exactly onto a polygon (neither parcels nor buildings). For each plant, we thus perform three assignment options. The first option relates each plant to the polygon onto which it falls; in that case, the distance between the plant and the polygon is assumed to be 0. If the plant does not fall exactly onto a polygon, it has no associated polygon. The second option then relates each plant to the polygon whose centroid is the closest, and we compute the distance between the plant and that centroid. Finally, the third option relates each plant to the polygon whose border is the closest; we again compute the distance between the plant and that border. We then compare the three (or two) distances obtained in the three option and we take as the final assignment the polygon corresponding to the shortest distance. Obviously, when the plant falls onto a polygon, it is that polygon which is assigned to the plant since the distance is zero. When the shortest distance is greater than 75 meters we consider that the process is too

Figure A2: Establishments (yellow), parcels (green), and buildings (red) from the roll data.

noisy and we do not assign that polygon to the plant. In addition, to avoid assigning the surface of corridors to plants, we compute for each polygon its number of neighbors. If an assigned polygon has more than 10 neighors, we consider that the polygon is a corridor or a common space and we do not assign that polygon to the plant.

We then construct a variable corresponding to the combination of assignments pointing in the direction of the polygon the establishment is assigned to. For example if the options "Border" and "Center" assign the plant to the same polygon whereas the "Within" option points to a different polygon for the same plant, then assignment variable for that plant will be "Center-Border". Thus, the assignment variable has the following 7 categories : (1) "Within-Center-Border"; (2) "Within-Center"; (3) "Within-Border"; (4) "Center-Border" (5) "Within"; (6) "Center"; (7) "Border".

Based on this assignment variable, we construct a quality variable as follows: i) we cross-tabulate the assignment variable with the dummy we could build for the observations from Quebec and that identifies those establishments which are assigned to the right polygon (described in section 2.2); ii) for all of our observations, we define as "Excellent" those observations whose assignment category has a high probability of being located on their actual polygon as measured based on observations from Quebec; "Good" is for observations whose assignment category has an intermediate probability of being located on their actual polygon; and "Poor" is for all the categories with a low

probability of being located on their actual polygon. Doing so, we implicitly assume that the mapping between the assignment variable and the dummy identifying correct observations in Quebec is representative of the entire country.

For the Parcel-based measure, the process leads to grade as "Excellent" the plants whose assignment category is "Within-Border-Center" or "Within". These plants with an "Excellent" parcel-based measure have a 89% probability of being positioned on their actual polygon. Plants graded as "Good" are those whose assignment category is "Within-Border" or "Within-Center". The plants of "Good" quality have a 60% probability of being positioned on their actual polygon. Finally, "Poor" is the grade for observations whose assignment category is "Border"; "Center-Border" or "Center"; these observations have a 16% probability of being located on their actual polygons.

For the Building-based measure, the category "Excellent" comprises the plants whose assignment category is "Within-Border-Center", "Within-Center", "Within-Border" and "Border-Center". The observations rated as "Excellent" for the Building-based measure have a 78% probability of being positioned on their actual polygon. The quality "Good" is for observations whose assignment category is "Within"; for them, the probability of being positioned on their actual polygon is equal to 66%. The grade "Poor" encompasses observation whose assignment category is "Border" or "Center". These plants have a 36% probability of being positioned on their actual polygon.

Summary: step-by-step data-construction procedure. Below is a summary of the main workflow to construct our dataset.

Step *1. Creating a unique addresses*. From the Scotts dataset, unique addresses are identified since several plants can share the same location. We create a unique identifier for each address. This step prepares the geocoding process, and will avoid to geocode several times the same address. A dataset of unique addresses is then generated with variables, the detailed address, and the address identifier.

Step *2. Geocoding unique addresses*. We use the dataset of unique addresses as input for the geocoding process described above. The output file contains geocoded addresses, in addition to the inputs variables, the geographic coordinates of each address, the detailed address as recorded in the database of the geocoder (Google or

DMTI) as well as a quality variable indicating the degree of accuracy of the returned coordinates.

Step *3. Extracting polygon surfaces*. Using a Geographic Information System, the geocoded addresses are overlaid on the polygons featuring parcel or building footprints. Then spatial join techniques are used to associate parcel and/or building polygons to addresses. Three different spatial join approaches are used to associate polygon areas to addresses. The output contains for each address, the associated polygon area from each of the three spatial join approach, as well as the distance between each associated polygon and the geographic coordinates of the address.

Step *4. Extracting location characteristics*. Using a Geographic Information System, the geocoded addresses are overlaid on shapefiles of dissemination areas, Census Metropolitan Areas (CMAs), zoning restrictions, highways, seaports and airports to compute various location variables : population and surface area of dissemination areas and CMA, distance to closest seaport, freight station, airport and highway junction as well as dummies for zoning categories.

Step *5. Creating a raw land variable*. This process compares the results of the three different spatial join approaches and finally assign to each address the 'best' result. Quality variables are constructed.

Step *6. Creating the final dataset*. The Scotts dataset is merged with location characteristics and land measures to obtain the final dataset used in the paper.

B Quality assessment and representativeness.

Beyond the measurement challenges mentioned in the previous subsection, geocoding data and assigning them to polygons retrieved from satellite data inherently bring issues regarding the quality of the data and the methodology employed to assign plants to polygons.

First, there can be errors in the polygon datasets. Representing a parcel or a building by a polygon is subject to minor errors. For example, the algorithm used to convert satellite building images into polygon building outlines may fail in some cases to fit exactly the building into its representative polygon. The level of such errors—known as

the matching precision—is estimated at 1.3% by the data provider.[32] This type of error only affects the building polygons. Parcel polygons are derived from administrative data and should, therefore, not be subject to measurement error of the type inherent to satellite data.

Second, there can be errors in the plant-to-polygon assignments. Geocoding microdata is an inherently noisy process. Even minor errors in the geocoding of plants can lead to their mis-assignment to polygons. To gauge the scope of false assignments in our dataset, we make use of the subset of data for the province of Quebec (QC). The reason is that the polygon identifiers in the QC dataset are the same as the official identifiers of the polygons as recorded on the governmental website of the land register "Infolot".[33] We can, therefore, randomly draw a set of addresses of plants in QC from our dataset and compare the parcel identifiers from "Infolot" to those obtained by our assignment procedure. Using a sample of 1,667 addresses, we find 1,320 correct assignments. Put differently, the probability for a plant in QC to be located exactly on its actual polygon is 79.16%.

As explained in the part "Assignment to polygons" of the Appendix, the assignment of plants to polygons is based on three options that can potentially point to different polygons. Among the 1,667 addresses that we use for validation, if we restrict ourselves to the subset of observations for which the three options in the assignment procedure point to the same polygon, the share of correct assignments increases to 91.3%. In other words, plants for which the three assignment options point to the same polygons are very likely to be correctly assigned. Making use of that observation, we finally construct a 'quality' variable based on: (i) how accurate the geocoding of the establishment is; and (ii) how likely a correct assignment to a polygon is. This quality variable—which we construct for the whole dataset, not just Quebec—has three categories: excellent, good, and poor (see the part "Assignment to polygons" for more details). Table B1 summarizes the distribution of observations across data-quality categories for parcels and building footprint. In the remainder of the paper, unless noted

[32] See `https://github.com/Microsoft/CanadianBuildingFootprints` on the GitHub website where the data are released.

[33] On that website, it is possible to recover the identifier of a parcel by providing the address of a location. See `https://appli.mern.gouv.qc.ca/infolot/`.

Table B1: Assignment quality.

Assignment quality	Parcel (PB)		Building (BB)	
	N	%	N	%
Excellent	10,428	76.98	25,289	95.45
Good	843	6.22	775	2.93
Poor	2,275	16.79	431	1.63
Total	13,546	100.0	26,495	100.0

Notes: Distribution of geocoded establishments in 2017 by quality categories. The classification includes the quality of the geocoding and the quality of the polygon assignment process. Concerning the geocoding quality, all observations with a less than excellent quality are removed, and the remaining observations are used to construct the three groups: excellent, good, and poor. The final sample we use includes observations of excellent quality located in CMAs and CAs and for which the values of the covariates used in the regression analysis are not missing. We have a final regression sample of 8,704 parcels.

otherwise, we only keep observations of 'excellent' quality.

C Detecting city centres.

For each establishment, we compute its distance to the nearest city centre. We use an algorithm that identifies clusters of densely populated dissemination areas. The latter correspond to zones that dominate—in a statistical sense—the population distribution in the metro area. The clustering algorithm works as follows:

1. For each metro area, determine all dissemination areas (DAs) that have a population density in the top quartile of the metro area density distribution. Flag those with 1, and the remaining DAs with 0.

2. For each metro area, compute for each DA i the number of DAs flagged with 1 and the number of DAs flagged with 0 within a 750 metres radius around the centroid of the DA. Assume that there are N_{i0} DA's flagged with 0 and N_{i1} DA's flagged with 1 within 750 metres around DA i

3. Compute the total population within the dissemination areas N_{i0} and N_{i1} in the 750 metres radius.

4. Use the hypergeometric probability distribution to compute the probability to see N_{i1} ones among the draw of N_{i0} + N_{i1} DAs within 750 metres, given the total number of 1's and of 0's in the city-wide distribution.

5. Flag all DA's that: (i) have population density in the top quartile; (ii) have a p-value from the hypergeometric distribution below 1%; and (iii) that have a population share in the top 5% of all the DA's in the top quartile of the population density distribution.

6. Put the remaining DA's on the map. Draw buffers with 750 metres radius round those DAs and merge all contiguous buffers. These correspond to the identified centres.

7. Take the centroid of that merged buffer as the location of the centre.

Note that there is no a priori restriction on the number of centres in a metropolitan area. We detect 166 centres in our 152 metro areas. We finally associate each firm with the nearest centre and compute the great circle distance between the firm's centroid and the city centre's centroid. We only consider firms that are less than 30 kilometres from an urban centre, all firms that are more than 30 kilometres from an urban centre are not considered as urban.

D Additional tables and results.

Table B2: Distribution of plants across industries.

	Regression sample (Table 1)		Scotts Data	
	N	%	*N*	%
311 Food mfg	745	8.6	2,929	8.9
312 Beverage and tobacco product mfg	77	0.9	340	1.0
313 Textile mills	32	0.4	97	0.3
314 Textile product mills	227	2.6	752	2.3
315 Clothing mfg	307	3.5	724	2.2
316 Leather, allied product mfg	40	0.5	131	0.4
321 Wood product mfg	353	4.1	1,933	5.9
322 Paper mfg	149	1.7	509	1.6
323 Printing, support activities	726	8.3	2,274	6.9
324 Petrol, coal product mfg	20	0.2	135	0.4
325 Chemical mfg	433	5.0	1,578	4.8
326 Plastics, rubber products mfg	545	6.3	1,915	5.8
327 Non-metallic mineral product mfg	387	4.4	1,985	6.1
331 Primary metal mfg	125	1.4	543	1.7
332 Fabricated metal product mfg	1,300	14.9	5,267	16.0
333 Machinery mfg	1,061	12.2	4,614	14.1
334 Computer, electronic product mfg	299	3.4	1,032	3.2
335 Electrical, appliance mfg	256	2.9	787	2.4
336 Transportation equipment mfg	295	3.4	1,117	3.4
337 Furniture, related product mfg	421	4.8	1,405	4.3
339 Miscellaneous mfg	906	10.4	2,753	8.4
Total	8,704	100.0	32,829	100.0

Notes: This table reports the distributions of the Scott's database along with the regression sample in Table 1 across the different industries at the NAICS 3-digit level.

Table B3: Distribution of plants across provinces.

	Regression sample (Table 1)		Scotts Data	
	N	%	*N*	%
Alberta	0	0.0	2,818	8.6
British Columbia	2,208	25.4	3,865	11.8
Manitoba	419	4.8	1,028	3.1
New Brunswick	214	2.5	713	2.2
Newfoundland	0	0.0	298	0.9
Nova Scotia	0	0.0	781	2.4
Ontario	1,881	21.6	13,850	42.2
Prince Edward Island	0	0.0	148	0.5
Quebec	3,708	42.6	8,456	25.8
Saskatchewan	274	3.1	972	2.7
Total	8,704	100.0	32,829	100.0

Notes: This table reports the distributions of the Scott's database along with the regression sample in Table 1 across the Canadian provinces. We remove the three territories as they have very few establishments in the Scotts' data.

Table B4: Sample selection on observables.

Dependent variable	1 is in sample			
	(1)	(2)	(3)	(4)
Ln Employment			-0.010	0.004
			(0.015)	(0.019)
# functions in the estab.			0.018	-0.006
			(0.018)	(0.017)
# 4-digit NAICS			-0.002	0.004
			(0.010)	(0.008)
# products			0.006	0.006
			(0.007)	(0.007)
1 is HQ			0.042	-0.017
			(0.030)	(0.031)
1 is exporter			-0.001	-0.017
			(0.051)	(0.039)
Ln CMA population				0.254[a]
				(0.058)
Ln population density 500m				0.085
				(0.053)
Ln distance to major airport				0.025
				(0.076)
Ln distance to major seaport				-0.043
				(0.093)
Ln distance to freight station				0.072
				(0.058)
Ln distance to junction				-0.024
				(0.021)
Fixed effects:				
4-digit industry	Yes	Yes	Yes	Yes
Province	No	No	No	No
Observations	24,457	24,457	24,457	24,457
R^2	0.0157	0.271	0.272	0.326

Notes: Probit regressions. 1 denotes $\{0, 1\}$ dummy variables. Standard errors in parentheses are clustered at the CMA/CA level. $^a p < 0.01$, $^b p < 0.05$, $^c p < 0.1$.

Table C1: Employment over parcel size by NAICS 3-digit industry.

	Employment over parcel size			
	N	Mean	Median	CV
311 Food mfg	745	13.8	5.62	2.24
312 Beverage and tobacco product mfg	77	34.7	3.47	5.13
313 Textile mills	32	5.95	3.26	1.1
314 Textile product mills	227	5.03	2.49	1.69
315 Clothing manufacturing	307	12.28	4.02	2.43
316 Leather, allied product manuf.	40	8.07	3.61	1.25
321 Wood product manufacturing	353	7.05	2.49	4.15
322 Paper manufacturing	149	6.48	2.39	2.61
323 Printing, support activities	726	5.91	2.63	1.94
324 Petrol, coal product manuf.	20	4.68	1.44	2.14
325 Chemical manufacturing	433	15.26	2.59	7.57
326 Plastics, rubber products manuf.	545	6.17	3.3	1.8
327 Non-metallic mineral product manuf.	387	5.48	2.39	2.03
331 Primary metal manufacturing	125	6.96	3.35	1.99
332 Fabricated metal product manuf.	1300	7.21	3.06	7.23
333 Machinery manufacturing	1061	6.65	2.95	3.14
334 Computer, electronic product manuf.	299	9.21	3.35	3.16
335 Electrical, appliance manuf.	256	7.21	3.32	2.51
336 Transportation equipment manuf.	295	5.73	3.22	1.57
337 Furniture, related product manuf.	421	5.88	3.06	1.97
339 Miscellaneous manufacturing	906	13.64	2.58	15.89
Total	8704	8.82	3.05	9.17

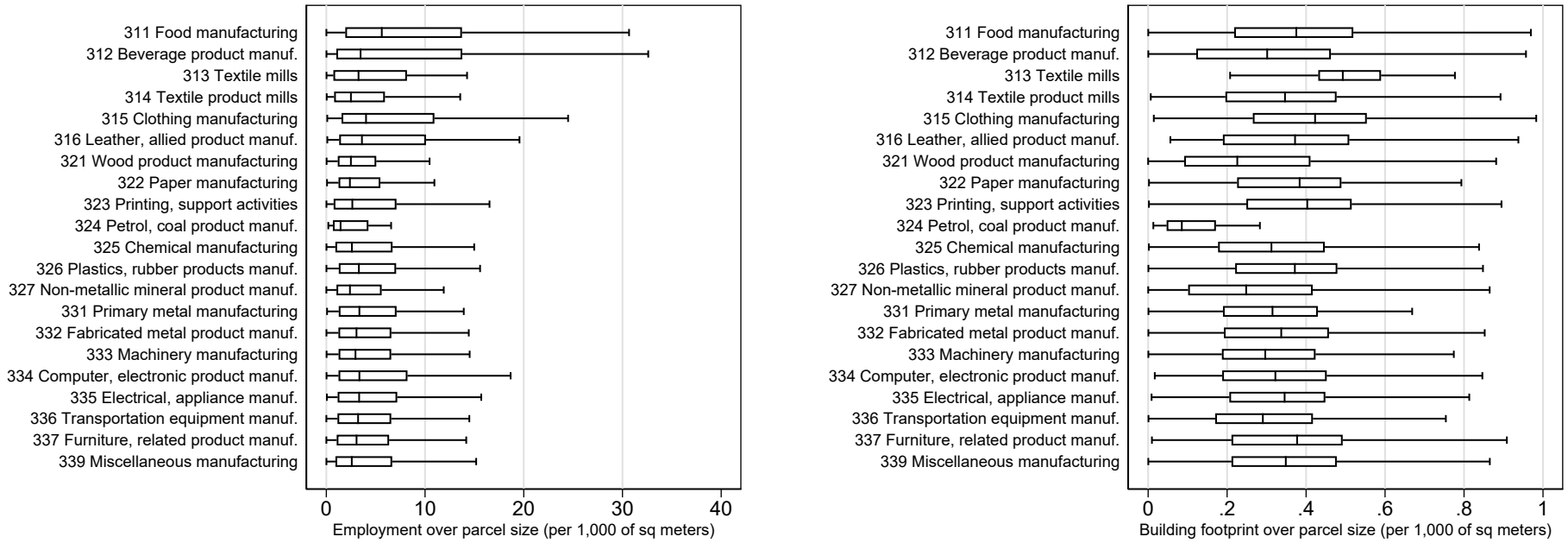
Notes: This table reports descriptive statistics for the number of workers pr unit of parcel size (in number of workers per 1,000 of square meters) across 3-digit industries. The sample is the regressions sample in Table 1.

Table C2: Building footprint over parcel size by NAICS 3-digit industry.

	Building footprint over parcel size			
	N	Mean	Median	CV
311 Food manufacturing	615	.39	.38	.6
312 Beverage product manuf.	55	.31	.3	.76
313 Textile mills	24	.51 .49	.35	
314 Textile product mills	183	.35	.35	.6
315 Clothing manufacturing	216	.43	.42	.51
316 Leather, allied product manuf.	35	.37	.37	.59
321 Wood product manufacturing	321	.27	.23	.81
322 Paper manufacturing	129	.39	.38	.52
323 Printing, support activities	572	.41	.4	.53
324 Petrol, coal product manuf.	15	.11	.09	.78
325 Chemical manufacturing	388	.33	.31	.64
326 Plastics, rubber products manuf.	498	.37	.37	.56
327 Non-metallic mineral product manuf.	349	.28	.25	.76
331 Primary metal manufacturing	106	.32	.31	.61
332 Fabricated metal product manuf.	1153	.35	.34	.6
333 Machinery manufacturing	968	.32	.3	.59
334 Computer, electronic product manuf.	265	.34	.32	.6
335 Electrical, appliance manuf.	232	.35	.35	.53
336 Transportation equipment manuf.	255	.31	.29	.66
337 Furniture, related product manuf.	358	.38	.38	.55
339 Miscellaneous manufacturing	687	.37	.35	.59
Total	7424	.35	.34	.61

Notes: This table reports descriptive statistics for the ratio of building footprint over parcel size. The sample corresponds to observations in the regression sample of Table 2 for which parcel size exceeds the building footprint.

Figure C1: Employment over parcel size and building footprint over parcel size by NAICS 3-digit industry.

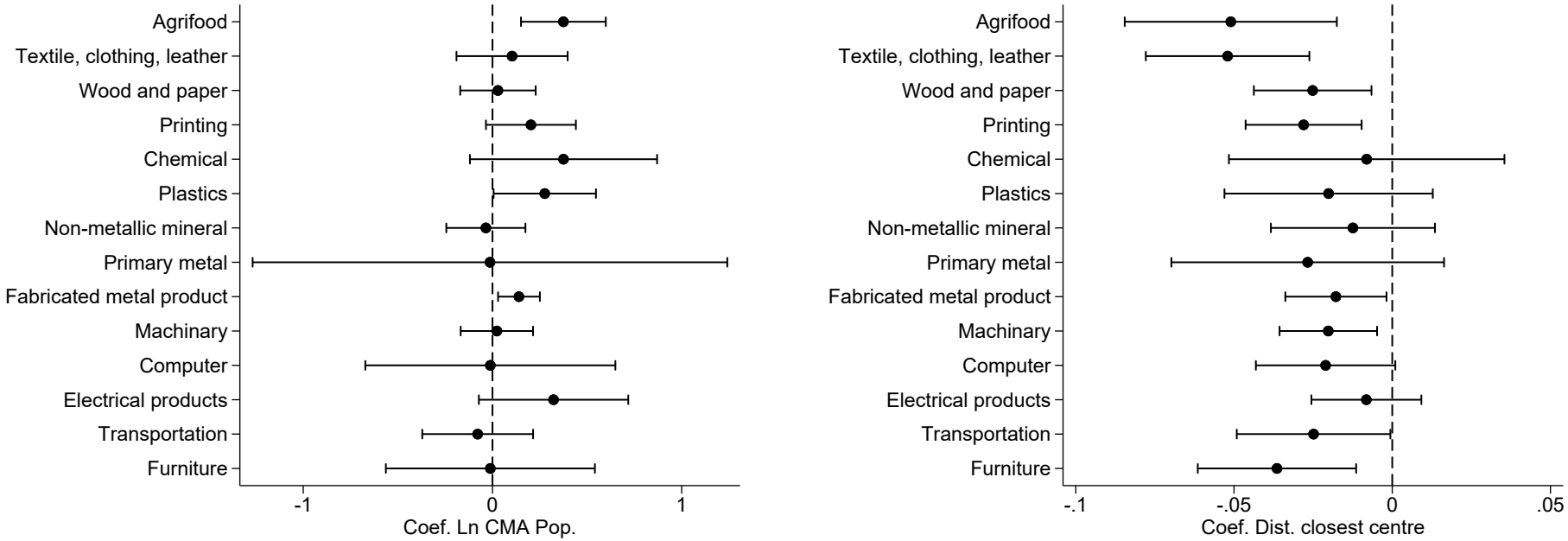


(a) Employment over parcel size.

(b) Building footprint over parcel size.

Notes: Panel (a) shows the distribution of employment over parcel size across industries. The sample contains observations from the regression sample in Table 1. Panel (b) shows the distribution of building footprint over parcel size across industries. The sample contains observations from the regression sample in Table 2 for which the parcel size exceeds the building footprint.

Figure C2: Elasticity of employment over parcel size to CMA size and distance from the closest centre by NAICS 3-digit industry.



(a) CMA size.

(b) Distance from the closest center.

Notes: The industries NAICS 312 ('Beverages and Tobacco') and NAICS 313 ('Textile mills') which have very few observations have been ignored. Panel (a) shows the distribution of the elasticity of employment over parcel size to CMA population size. Panel (b) shows the distribution of the semi-elasticity of employment over parcel size to distance from the nearest city centre.