UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESSAIS SUR LA CONCENTRATION SPATIALE INDUSTRIELLE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN ÉCONOMIQUE

PAR

THÉOPHILE NDJANMOU BIEDA

JUILLET 2022

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESSAYS ON SPATIAL CONCENTRATION OF INDUSTRIES

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

TO THE PH.D IN ECONOMICS

BY

THÉOPHILE NDJANMOU BIEDA

JULY 2022

REMERCIEMENTS

La réalisation de cette thèse a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais tout d'abord adresser toute ma reconnaissance à mes directeurs de thèse, les professeurs Kristian Behrens et Florian Mayneris, pour leur patience, leur disponibilité, et leur rigueur dans l'encadrement qu'ils m'ont apporté. Vous avez veillé à ce que je puisse avoir les financements nécessaires et des conditions idoines pour la réalisation de cette thèse. Vous m'avez appris à préparer un projet de recherche, à organiser et structurer un travail de recherche, à analyser les données et construire une histoire économique autour des résultats. Vous m'avez appris à rédiger un travail scientifique. Vous avez été pour moi bien plus que de simples directeurs de thèse. Merci Kristan ! Merci Florian !

Je désire aussi remercier les professeurs de l'ESG de l'Université du Québec à Montréal qui ont contribué de diverses manières à mon encadrement, par des cours magistraux, des conseils et des commentaires sur mes travaux lors des séminaires internes de l'ESG. Tout particulièrement, merci aux professeurs Julien Martin et Marlon Seror respectivement président du jury et évaluateur interne de ma thèse.

Merci également à Théophile Bougna, Economiste à la Banque mondiale qui a accepté d'être examinateur externe de ma thèse.

Merci aussi à tout le personnel administratif de l'ESG des cycles supérieurs : Merci à Martine Boisselle ! Merci à Julie Hudon ! Merci à Lorraine Brisson ! Merci à Karine Fréchette ! Merci à Natacha Bikonda !

Je voudrais également exprimer ma reconnaissance envers mes amis et collègues de doctorat qui m'ont apporté leur soutien moral et intellectuel tout

au long de ma démarche. Merci tout particulièrement à Manassé Drabo, mon fidèle compagnon de tout ce parcours. Je me souviendrai toujours de nos nombreuses heures de travail pour discipliner et géocoder les "données Scotts".

Merci à tous mes amis et frères de la famille chrétienne pour leurs prières et encouragements. Merci à ceux de la Communauté Baptiste Evangélique de Saint Léonard à Montréal ! Merci à ceux du Cercle de Prière à Yaoundé ! Merci à ceux des Groupes Bibliques Universitaires de Montréal et du Cameroun.

Je veux remercier ma famille pour le soutien, la patience et l'encouragement qu'elle m'a apporté pour que je puisse réaliser ce projet. Merci à Jean-Jacques et Lucile ! Merci à Ernest et Stella ! Merci à toute la famille Bieda et la famille Lemdjo.

Enfin, je veux dire merci à mon épouse et nos enfants d'avoir accepté et supporté mon absence et de m'avoir encouragé tout au long de ces années de recherche. Gaëlle, tu es formidable ma chérie ! Merveille, Nathan, Abigaëlle et Paul-Emmanuel, vous faites ma joie.

DEDICACE

A Dieu Tout Puissant, Créateur du ciel et de la terre ! A qui je dois la vie, le mouvement et l'être (Actes 17 :28a, Bible).

TABLE DES MATIÈRES

LISTE DES TABLEAUX

TABLE DES FIGURES

RÉSUMÉ

La concentration spatiale de l'activité économique fascine par ses extrêmes et intrigue en même temps par les paradoxes qui l'accompagnent. Pour tenter de mieux comprendre ce phénomène, cette thèse apporte une contribution essentiellement empirique à travers ses trois chapitres.

Dans le chapitre 1 intitulé, "Land for production : Evidence from Canadian Manufacturing Industries", nous combinons des microdonnées des établissements avec des données de polygons pour construire une mesure de la quantité d'espace utilisé par les établissements manufacturiers au Canada. Puis, à l'aide d'un cadre conceptuel simple, nous montrons comment le foncier par unité de travailleur varie avec les prix relatifs du foncier, le paramètre technologique et l'élasticité de substitution entre le travail et le foncier. Nos résultats suggèrent que le foncier et le travail sont des substituts imparfaits, que le foncier a une compposante fixe, et comporte des coûts d'ajustement non négligeables.

Dans le chapitre 2 intitulé "The causes of the agglomeration of innovation : Evidence from the coagglomeration patterns", nous utilisons la mesure de Duranton and Overman (2005) pour décrire la coagglomération de l'innovation au Canada. Cette description révèle que l'innovation canadienne est concentrée et même plus que la production. Ensuite, nous estimons l'effet causal de "l'échange des intrants", de "l'utilisation d'un bassin commun de travailleurs" et de la "diffusion des connaissances" sur la coagglomération de l'innovation. L'analyse montre qu'en plus de la coagglomération de la production, seule la diffusion de connaissances a un effet positif et significatif sur la coagglomération de l'innovation.

Dans le chapitre 3 intitulé "A complément to the test of localization of Duranton and Overman (2005)", nous appliquons la méthodologie de Duranton and Overman (2005), en utilisant un nouveau contrefactuel pour étudier l'agglomération des établissements manufacturiers canadiens. Le nouveau contrefactuel, qui tient compte plus précisément des choix de localisation possibles des entreprises, détecte mieux la localisation et la dispersion des industries. De plus, des différences substantielles sont observées entre la classification générée avec le nouveau contrefactuel et celle du contrefactuel classique. Nos résultats suggèrent par ailleurs que la localisation et la dispersion seraient plus fortes que ce qui est généralement admis.

ABSTRACT

The spatial concentration of economic activity fascinates by its extremes and at the same time intrigues by the paradoxes that accompany it. In an attempt to better understand this phenomenon, this thesis makes an essentially empirical contribution through its three chapters.

In Chapter 1, "Land for production: Evidence from Canadian Manufacturing Industries", we combine detailed micro-geographic establishment-level data with polygon shapefiles to construct a new dataset with information on land used by a large representative sample of Canadian manufacturing plants. Using a conceptual framework with productivity-augmenting parameters for land and labor, we show how land per worker varies with relative land prices, technological parameters, and the elasticity of substitution between labor and land. Our results suggest that land and labor are imperfect substitutes, land largely appears to have a fixed component, and it has sizable adjustment costs.

In chapter 2 entitled "The causes of the agglomeration of innovation: Evidence from the coagglomeration patterns", we use the continuous measure of Duranton and Overman (2005) to describe the coagglomeration of innovation in Canada. The observed patterns reveal that Canadian innovation is concentrated and even more than production. Then, we analyze the effects of labor pooling, input sharing, and knowledge spillover on the coagglomeration of innovation. The analysis shows that on top of the coagglomeration of the production, only the knowledge spillover unambiguously causes the coagglomeration of innovation.

In chapter 3 entitled "A complement to the test of localization of Duranton and Overman (2005)", we apply the methodology of Duranton and Overman (2005), using a new counterfactual to study the agglomeration patterns of Canadian manufacturing establishments. The new counterfactual which accounts more precisely for firms' possible location choices, better detects the departure from randomness, and generates substantial differences between the patterns that we uncovered compared to those obtained with the classical counterfactual. Our results suggest that localization and dispersion may be stronger than what is usually thought.

INTRODUCTION

L'un des faits saillants du paysage économique partout dans le monde est la concentration spatiale. Au Canada par exemple, seule 20% de la surface nationale est habitée et environ 81% de la population vit en zone urbaine. Cette extrême concentration a de tout temps fasciné et intrigué les économistes surtout à cause des apparents paradoxes rattachés à ce phénomène. Par exemple, un logement à Vancouver en Colombie-Britannique est 2.5 fois plus cher qu'un logement à Saguenay au Québec. Pourtant, la densité de la population à Vancouver a augmenté au fil du temps pour se situer à environ 5400 habitants au kilomètre carré. Au même moment la densité de la population de Saguenay n'est que de 2.9 habitants au kilomètre carré. [1]

Pour mieux comprendre la concentration spatiale, les questions qui orientent la recherche sur ce phénomène portent essentiellement sur les raisons de son existence et ses conséquences sur l'activité économique. Une autre question tout aussi importante est celle de sa mesure.

Comment la concentration affecte-t-elle l'activité économique ?

L'agglomération de l'activité économique génère aussi bien des gains que des coûts pour les acteurs économiques (Brinkman, 2016). S'agissant des gains, les externalités positives de l'agglomération sont notamment les gains de productivité (Mayer et al., 2015; Fujita and Thisse, 2002). Ces gains de productivité,

1. Voir le site https ://www.rentseeker.ca/ pour l'estimation des prix de logement, et https ://www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2015004-eng.htm et worldpopulationreview.com pour les densités de population

sont entre autres dus à la sélection spatiale des entreprises et des travailleurs les plus productifs dans les endroits les plus concentrés (Behrens et al., 2014). En termes chiffrés, il est couramment admis que doubler la taille d'une ville par exemple se traduirait par une augmentation de la productivité de l'ordre de 3 à 8% (Rosenthal and Strange, 2004), et doubler la densité de l'emploi augmenterait l'intensité d'innovation de 20% (Carlino et al., 2007).

Les coûts économiques de l'agglomération quant à eux se traduisent par une plus grande congestion spatiale (Graham, 2007), des prix plus élevés aussi bien pour les biens de consommation que pour les facteurs de production (Tabuchi and Yoshida, 2000; Combes et al., 2019). Au sujet du foncier en particulier, le modèle monocentrique utilisé comme base pour plusieurs théories en économie urbaine suggère que dans les endroits les plus denses, la quantité de logements utilisés est beaucoup plus petite que dans les endroits les moins denses. Plusieurs faits empiriques valident cette prédiction notamment pour ce qui est du foncier résidentiel (Duranton and Puga, 2015). Du côté de la production, très peu d'études empiriques se sont penchées sur la question, à cause notamment d'une absence de microdonnées appropriées pour pleinement explorer ce sujet.

Comment se forment les "clusters" économiques ?

En plus des avantages naturels que certains endroits pourraient présenter par rapport à d'autres, la théorie économique met en évidence le rôle de trois mécanismes à l'origine de l'agglomération de l'activité économique. Il s'agit notamment de «l'échange des intrants», «le partage d'un bassin commun de travailleurs", et «la diffusion de connaissance». En effet, la possibilité de partager des intrants et des installations de production, l'efficacité de l'appariement entreprise-travailleur et la facilité de créer, d'accumuler et de diffuser des connaissances au sein d'un "cluster" sont autant d'incitations à l'agglomération de l'ac-

tivité économique (Duranton and Puga, 2004). D'un point de vue empirique, de nombreuses études ont testé l'existence de ces forces et ont fourni des preuves convaincantes de leur rôle dans la génération de l'agglomération de la production (Ellison et al., 2010; Audretsch and Feldman, 1996; Rosenthal and Strange, 2001; Ellison and Glaeser, 1999). Cependant, plusieurs champs de validation de ces mécanismes restent encore inexplorés. Par exemple, comment se forme l'agglomération de l'innovation ? Cette question est d'autant plus intéressante que quelques faits notables distinguent l'agglomération de la production de celle de l'innovation. Par exemple, l'innovation est beaucoup plus concentrée que la production et les entreprises ne localisent pas leurs unités productives aux mêmes endroits que leurs unités de Recherche et Développement (Kelly and Hageman, 1999; Feldman and Kogler, 2010)

Comment l'agglomération est-elle mesurée ?

Les méthodes de mesure du degré de concentration industrielle ont connu de nombreuses avancées au cours de cette dernière décennie. En effet, mesurer adéquatement le degré de concentration de l'activité économique est une question de premier ordre, car de cet exercice dépendent plusieurs politiques en matière d'agglomération. Dans ce sillage, les divers efforts des chercheurs pour quantifier le degré de concentration ont donné lieu à des mesures intéressantes, allant du simple indice de Gini à des mesures plus élaborées telles que la fonction de Ripley (Marcon and Puech, 2003). Cependant, une caractéristique particulière de ces mesures est qu'elles n'ont pas de signification en soi. Elles doivent plutôt être comparées à un point de référence pour évaluer le degré de concentration de la réalité qu'elles sont supposées quantifier. Depuis Ellison et Glaeser (1997), il est communément admis qu'un test adéquat de concentration industrielle devrait tenir compte de la tendance naturelle de l'activité industrielle à se localiser dans l'espace. En effet, une distribution in-

égale de l'activité industrielle à travers l'espace ne traduit pas forcément de la concentration industrielle. Ce constat a porté les chercheurs à accorder un soin particulier à la définition de la référence à laquelle il conviendrait de comparer une distribution spatiale de l'activité économique. La référence la plus couramment utilisée suppose qu'en absence de toute forme d'agglomération, la distribution spatiale serait celle d'une assignation aléatoire des entreprises à travers l'espace. Cependant, il se trouve que certaines localisations ne seraient pas éligibles pour abriter certaines entreprises pour des raisons propres à la nature de l'activité de l'entreprise ou simplement pour des questions légales ou réglementaires. En clair, un site productif au Centre-ville de Montréal accueillerait difficilement une industrie de production pétrolière pour des raisons environnementales et/ou d'espace. À date, la prise en compte de ce genre de questionnement dans la définition des "références" qui servent à détecter le degré de concentration n'est pas encore courante. Pourtant, de la référence utilisée dépend entièrement la classification des industries en termes de degré d'agglomération.

Cette thèse à travers ses trois chapitres, apporte des contributions à chacune des problématiques abordées précédemment.

De manière plus précise, dans le chapitre 1 intitulé « Land for production : Evidence from Canadian Manufacturing Industries », nous discutons de l'incidence de la taille de la ville et de la distance au centre des villes (negativement correlée à la densité locale) sur l'utilisation du foncier par les industries manufacturières au Canada. En l'absence de données appropriées sur la quantité d'espace utilisée par les entreprises canadiennes, notre première tâche a consisté à construire une base de données du foncier utilisé dans la production manufacturière. Pour cela, nous exploitons les données d'accès libre sur les polygones représentant les bâtiments et parcelles du Canada. Et, à l'aide des outils

des Systèmes d'Information Géographique, nous parvenons à construire une mesure du foncier utilisé par les entreprises canadiennes. Nous utilisons ensuite cette mesure pour décrire de manière toute nouvelle l'intensité du facteur foncier dans la production. Cette description fait état entre autres d'une assez grande hétérogénéité sectorielle et spatiale dans la consommation du foncier par les entreprises manufacturières au Canada. En second lieu, éclairés par un cadre conceptuel simple, nous mettons en lumière quelques faits stylisés liés à cette variable. Nos résultats permettent de tirer trois leçons importantes sur le rôle du foncier dans la fonction de production des établissements manufacturiers : (i) l'élasticité de l'espace foncier par travailleur au niveau de l'établissement est d'environ -0,6 ce qui indique que le foncier est en partie un facteur de production fixe ; (ii) l'élasticité de la l'espace foncier par travailleur par rapport à la distance au centre des villes est positive mais faible en valeur absolue, ce qui suggère que l'espace foncier et le nombre de travailleur sont imparfaitement substituables ; (iii) enfin, notre analyse suggère également la présence d'importants coûts d'ajustement en rapport avec le facteur foncier.

Le chapitre 2 s'intitule "The causes of the agglomeration of innovation : Evidence from the coagglomeration patterns". La littérature théorique qui discute des microfondements des économies d'agglomération souligne la présence de trois causes principales liées à ce phénomène. Il s'agit notamment de "l'échange d'intrants", "le partage d'un bassin commun de travailleurs" et "la diffusion de connaissance". De nombreux faits empiriques valident l'effet positif de chacune de ces forces sur l'agglomération de la production. Ce chapitre quant à lui s'intéresse à l'effet de ces mêmes déterminants sur la colocalisation industrielle, mais plutôt du côté de l'innovation. Tout d'abord, nous nous servons d'une base de données de brevets canadiens pour construire une mesure de colocalisation de l'innovation. Ensuite, nous utilisons cette mesure pour décrire la

configuration spatiale de l'innovation au Canada. Ce premier exercice reproduit pour le cas du Canada, un fait stylisé qui met en contraste la concentration de l'innovation et celle de la production : *Au Canada également, l'innovation est plus concentrée que la production*. À la suite de cette analyse descriptive, nous estimons l'effet causal des trois déterminants microfondés des économies d'agglomération sur la colocalisation de l'innovation. Notre analyse montre qu'une part essentielle de la concentration de l'innovation est le simple reflet de la concentration de la production. De plus, outre la concentration de la production, seule «la diffusion de connaissance» a un effet positif et significatif sur la colocalisation de l'innovation.

Le chapitre 3 s'intitule "A complement to the test of localization of Duranton and Overman (2005)". Dans ce chapitre, nous proposons une amélioration des tests de détection de la localisation industrielle. Notre approche s'appuie sur la méthodologie proposée par Duranton et Overman (2005, 2008) pour identifier des industries localisées ou dispersées. Ces deux auteurs ont développé une approche de test de localisation industrielle, basée sur l'estimation du noyau de densité des distances bilatérales entre les établissements d'une industrie. Cette mesure de concentration est par la suite comparée à une mesure de référence encore appelée mesure contrefactuelle. La mesure contrefactuelle permet entre autres de distinguer la concentration industrielle effective, d'une simple inégalité spatiale naturelle à l'industrie manufacturière en général. La définition de cette référence pour une industrie donnée s'appuie sur un tirage aléatoire de sites parmi l'univers des sites manufacturiers. Ces tirages aléatoires sont ensuite attribués à cette industrie comme sites hypothétiques et utilisés pour construire la mesure de référence. Toutefois, Duranton et Overman (2005, 2008) reconnaissent que cette façon de procéder ne prend pas en compte le fait que certains sites pourraient être inappropriés, voire interdits à certaines entre-

prises pour des raisons de tailles, de procédés de production, de réglementation, etc.

Nous utilisons la même méthodologie que Duranton et Overman (2008) et proposons une nouvelle approche de construction de la mesure de référence. Notre approche utilise une méthode de classification hiérarchique pour définir des catégories de sites. Ensuite, la mesure de référence pour une industrie donnée est construite par des tirages aléatoires de sites, sous la contrainte que chaque établissement de l'industrie d'intérêt ne peut se voir attribuer qu'un site qui appartient à la même catégorie que le site sur lequel l'établissement est observé. En procédant ainsi, nous minimisons le risque qu'un établissement soit réassigné à un site "non admissible" en raison de possibles contraintes de taille, de réglementation, etc.

Nous appliquons cette nouvelle approche à des données d'entreprises canadiennes, et les résultats obtenus suggèrent que cette nouvelle approche de définition de la mesure de référence, comparée à la méthode traditionnelle est plus précise dans la détection de la localisation/dispersion des industries. De plus, des différences substantielles sont observées dans les classifications de certaines industries, telles que générées par la nouvelle approche et celle traditionnelle. En effet, près de 20% des industries passent de localisées (respectivement dispersées) avec la mesure contrefactuelle classique, à dispersées (respectivement localisées) avec notre nouvelle mesure contrefactuelle. Enfin, la localisation industrielle telle qu'identifiée par la nouvelle mesure contrefactuelle apparait à des distances plus courtes que celles identifiées par la mesure contrefactuelle classique.

CHAPTER I

LAND FOR PRODUCTION: EVIDENCE FROM CANADIAN MANUFACTURING INDUSTRIES

Abstract

We combine detailed micro-geographic establishment-level data with polygon shapefiles to construct a new dataset with information on land used by a large representative sample of Canadian manufacturing plants. Using a conceptual framework with productivity-augmenting parameters for land and labor, we show how land per worker varies with relative land prices, technological parameters, and the elasticity of substitution between labor and land. Our results suggest that land and labor are imperfect substitutes, land largely appears to have a fixed component, and it has sizable adjustment costs.

Keywords: Land use for production; geo-referenced data; building and parcel polygons; adjustment costs; manufacturing.
JEL Classification: R32; R14; L60.

1.1 Introduction

All human activity to some degree requires land. The availability of land and its allocation among competing uses are thus paramount to our understanding of the spatial structure of economic activity and society in general. Analyzing the distribution of economic activity and population across and within cities is the hallmark of urban economics and the land it requires is the focus of a large and well-established body of theoretical literature. [1] There is also a growing empirical literature that seeks to understand land supply and land-use patterns for primarily residential purposes. [2] However, surprisingly little is known—both theoretically and empirically—on the role of land in the production process of firms (Duranton and Puga, 2015). [3] How does land consumption for production vary across sectors? Across space? How substitutable are land and labor inputs for firms? How and when do firms adjust their land inputs? And is land

---

1. See, e.g., Fujita and Ogawa (1982), Lucas and Rossi Hansberg (2002) Duranton and Puga (2015), and Brinkman (2016) for land-use patterns; Cheshire and Sheppard (2004) and Hilber and Robert-Nicoud (2013) for land-use regulations; and Fujita (1989), Anas et al. (1998), and Fujita and Thisse (2013) for comprehensive general treatments.

2. See, e.g., Saiz (2010) and Carozzi (2020) for estimates of land supply elasticities; Epple et al. (2010) and Combes et al. (2019) for estimations of the production function for housing; and Glaeser et al. (2005) for the costs of land-use regulations.

3. Canonical urban models 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 Fujita and Ogawa (1982), Lucas and Rossi Hansberg (2002), Empirically, land for production has been mostly analyzed in relation to land-use regulations (e.g., Cheshire et al. 2014; Haskel and Sadun 2012). There is a literature on commercial real estate prices (Drennan and Kelly, 2010; Ahlfeldt and McMillen, 2015), and on the effects of real estate collateral on firm-level investment (Gan, 2007; Chaney et al., 2012). More closely related to our study, Barr and Cohen (2014) analyze the floor-to-area ratio gradients of commercial properties in New York.

a variable or a fixed production factor?

Providing answers to these questions is both important and difficult. It is important for urban economic theory because it can inform the way we model land as a production factor. It is also important from an applied perspective because differences in land requirements across industries and establishments of different sizes shape the patterns of within- and between-city distributions of economic activity and productivity, which are the targets of many costly economic policies. Answering these questions is, however, difficult because of the paucity of data on land consumption for production. [4] Data on land is almost always conflated with data on non-land capital inputs, and quantity measures are usually not available (price times quantity, i.e., value is generally reported). We tackle these difficulties by constructing establishment-level quantity measures of land used for production and dissect them to uncover hitherto unnoticed patterns. Previewing our key findings, we show that the elasticity of land consumption per worker to plant-level employment is around -0.6, indicating that land is partly a fixed production factor; the elasticity of land consumption per worker to distance to the city center, is positive but small in absolute value, suggesting that land and labor are poor substitutes; finally, we find that land is potentially subject to sizeable adjustment costs.

Our contribution is threefold. First, we combine detailed micro-geographic establishment-level data with polygon shapefiles to construct a new dataset with information on the amount of land used by a large representative sample of Canadian manufacturing plants. We build a measure of land consumption

4. "Mergent Intellect" and "Scotts' National All Business Directories" are two databases that report 'square footage' for establishments. However, it is mostly missing and when it is reported it is too noisy to be useful. We are not aware of the existence of usable plant-level land data in other datasets.

based on the surface of the parcels occupied by the plants. The key strength of our dataset is that it measures quantities rather than values. Combining it with plant-level employment information enables us to construct a measure of land per worker, a key theoretical quantity in the firm's production function.

Second, we propose a conceptual framework with productivity-augmenting parameters for land and labor across plants and industries, and we show how land per worker varies with relative land prices, technological parameters, and the elasticity of substitution between labor and land. Using that framework, we discuss identification problems that bias the estimates of this elasticity of substitution. We then dissect our data along various dimensions to show how land consumption varies by industry, city size, distance to the city center, and a series of plant-level characteristics such as employment size. Our empirical model explains around half of the variation in the amount of land per worker used by manufacturing establishments. We document substantial variation of land consumption within and between sectors, and land per worker varies more than total land consumption.

The key finding of our analysis is that *land and labor are imperfect substitutes and that land has some characteristics of a fixed cost*. Indeed, while city size is not significantly associated with plants' per worker land consumption, we find a significantly positive and robust effect of distance to the city center on that variable. However, this elasticity is small in absolute value, suggesting that land and labor are poor substitutes. Moreover, the elasticity of per capita land consumption for production to plant-level employment is around -0.6. Put differently, a 10% increase in employment reduces the land per worker at the plant level by about 6% on average. This result is very robust across specifications; it shows that land is partly a fixed cost for firms.

Third, we discuss how the relationship between plant-level per worker land consumption, plant size, city size, and distance to the city center varies sectors. We hardly find any significant heterogeneity across sectors. We finally, provide also suggestive evidence of sizable adjustment costs for land.[5] Plants do not change their land consumption often, and when they do it is mostly within cities. Larger plants are less likely to relocate, and employment growth over the past four years is negatively related to current land-to-labor ratios. These findings suggest that land consumption cannot be adjusted significantly in the short- or medium run. Although we find heterogeneity across sectors, our estimates of adjustment costs are too imprecise to be very revealing.

Though largely descriptive, we think our results are interesting for urban economists because there is so little we know empirically about land use for production. Taken altogether, they show that in order to quantify the contribution of land to the production process of firms, a framework different from the usual Cobb-Douglas production function approach is needed.[6] Since Canadian manufacturing is representative of manufacturing in other developed countries, our data and results are potentially interesting for other researchers, e.g., to calibrate models or for structural estimation exercises. Furthermore, the increasing availability of big open-source data on building polygons—such as Microsoft's database or the ongoing collection of building outlines by Open Street Map—

5. There is a large literature on adjustment costs for factor inputs (e.g., Hamermesh and Pfann, 1996; Hall, 2004). However, land is almost never even mentioned in that literature, in line with our observation that it is usually subsumed by 'capital'. Sood (2020) and Bergeaud and Ray (2021) are recent exceptions.

6. Duranton et al. (2015) is the only contribution we are aware of estimating production functions for manufacturing firms where land is a production factor. They find a somehow low elasticity of value added to land input (0.13) but they estimate a standard Cobb-Douglas production function.

provides ample opportunities to replicate and extend our exercise to other countries.

The remainder of the paper is organized as follows. Section 1.2 explains the construction of our dataset and shows its representativeness. Section 1.3 provides descriptive statistics and dissects keys features of the amount of land used for production by Canadian manufacturing establishments. Section 1.4 estimates the elasticity of plant-level land consumption to city-size, distnce to the city and some plant-level characteristics. Section 1.5 discusses our conceptual framework and discusses estimation issues. Last, section 1.6 concludes. Details on the construction of our database and robustness checks are relegated to the Appendix.

1.2 Data construction

We collect information on the amount of land occupied by manufacturing establishments.[7] We construct and use two main measures throughout the paper. First, the surface area of the parcels where plants are located. It is derived from the polygons of the parcels. This measure captures both the building footprint and the outdoor space used by the plant for storage, parking, or green space. The second measure is the building-to-parcel ratio, i.e. the share of the parcel covered by the footprint of the buildings where plants are located. To compute it, we need the surface area of the building footprint, which is derived from the polygons of the buildings where plants are grounded.[8] We do not observe the floor space used by manufacturing establishments and we cannot infer it from

7. In what follows, we interchangeably use the terms 'establishment', 'plant', and 'firm'.

8. The building footprint is known as Gross Area Floor in the assessment roll terminology.

the data we have. This is why our analysis is not about the amount of space used by establishments; it is about the amount of land they occupy.

1.2.1 Methodology

We first briefly present the methodology used to construct the dataset. Details for each step of the procedure and an extensive discussion of the quality of the final dataset are relegated to Appendices 1.7.2–1.7.5. The dataset we build combines proprietary data for geo-referenced Canadian manufacturing establishments with open-source data for parcel- and building polygons. We use GIS tools to associate each establishment with specific building and parcel polygons. We then compute the parcel size and the building footprint for each plant using the surface of its associated polygons.

Data collection and processing

i) Establishment-level data

We use the proprietary Scott's National All Business Directories, a dataset that draws information on plants operating in Canada from Business Register records and telephone surveys. It provides a fairly exhaustive coverage of the manufacturing sector on which we focus in this paper. Although the data span the years 2001–2019,[9] we exploit the dataset as a cross-section in 2017, the closest year to the reference year for the polygon datasets that we use. This choice reduces potential measurement error due to changes in the delineation of buildings and parcels.[10] The variables of interest for the analysis include the precise

9. Data for 2015 are missing from our dataset.

10. It also allows for more precise geocoding as street names and configurations may have

postal address information of each plant, its industrial classifications (North American Industry Classification System, NAICS 6-digit level), an estimate of the number of workers at the site, and dummy variables for whether the plant reports an export activity and whether or not it is a headquarter. The dataset also contains information on the products manufactured by the plants and the broad sectors in which it is active (manufacturing, wholesale, professional, scientific and technical services etc.). We geocode each plant in the dataset using the procedure explained in Appendix 1.7.1. [11]

ii) Polygon datasets.

We collect parcel and building polygons from numerous Canadian provincial and metropolitan sources. The full list of sources from which we collect these datasets, as well as a discussion of their quality, are relegated to Appendix 1.7.2 (see Table 1.4). Concerning parcels, we collect more than 4.5 million polygons covering the entire provinces of British Columbia (BC), Quebec (QC), and New Brunswick (NB) as well as the cities of Toronto, Oshawa, Windsor, and York in Ontario (ON). For the other provinces, we obtain data for Banff in Alberta (AB), Winnipeg in Manitoba (MB) and Regina and Saskatoon in Saskatchewan (SK). We did neither obtain data for Nova Scotia (NS), Newfoundland and Labrador (NL), and Prince Edward Island (PE), nor for the three Territories. Concerning buildings, we collect information from the Open Source data on building footprints in Canada released by Microsoft. These datasets consist of 12,663,475

---

changed significantly since 2001.

11. The dataset already records geographic coordinates for each plant but some of these coordinates are based on postal code centroids obtained from Post Canada's Postal Code Conversion files. These are necessarily less accurate than coordinates obtained from rooftop geocoding and do not permit to precisely associate plants with building- or parcel polygons.

building footprints covering all provinces and territories. [12]

iii) Other datasets.

To make use of spatially fine-grained population census data, we collect the shapefiles of the boundaries of all dissemination areas (DA; the smallest geographic unit at which public census data are released), census metropolitan areas and census agglomerations (CMA and CA), economic regions (ER, which are sub-provincial units), and provinces in Canada. Combined with data from the population census of 2016, we obtain files that contain information on the population, the surface area, and the relations between the different levels of the census geography. We also collect polygon files released by DMTI which record basic zoning restrictions in Canada, namely the main type of activity allowed in each area by local zoning policies (commercial/industrial, residential, recreational). Finally, we collect from Statistics Canada additional information on some major infrastructures such as highways junctions (from the Canadian road network files), as well as the location of airports, seaports, and train freight stations (from the Open Government geographic data portal).

Construction of the surface measures

We use GIS tools to relate each geocoded plant in the establishment dataset to parcel and building polygons (see Appendix 1.7.3 for technical details). The mapping between plants and polygons then allows us to construct the measures of land occupied by manufacturing establishments. The parcel size is the surface of the parcel polygon that contains the establishment, while the building footprint is the ground floor area of the building polygon that contains the establishment.

12. See https://blogs.bing.com/maps/2019-03/microsoft-releases-12-million-canadian-building-footprints-as

There is not a one-to-one mapping between establishments on the one hand, and parcel and building polygons on the other hand. Sometimes, several establishments fall on the same parcel. Put differently, there is some parcel- and building- sharing. This should, however, not be a major problem for our analysis: compared to services establishments, manufacturing plants are less likely to have many neighbors. In the sample used for the analysis of parcel size, the average number of neighbors identified for each establishment based on the Scotts data is 1.3 and the median is 0. Yet, this means that there are still some establishments that share the same parcel or building. Thus, we control in the regression analysis for (a polynomial function in) the number of neighbors as measured in the Scotts data. We also check that the main results remain unchanged when focusing on establishments with no identified neighbors. [13]

Also, the properties occupied by manufacturing establishments may be composed of several contiguous parcels, not only of the one on which establishments fall following the geocoding process. The discussions we had with employees from the Land Register of Quebec let us think that this situation rarely occurs. This is coherent with the fact that, as shown by Brooks and Lutz (2016), assembled parcels have a much higher value than the sum of values of the individual parcels, so that owners of contiguous parcels have an incentive to assemble them. Moreover, for the city of Montreal, we have exhaustive information from the property assessment roll that allows us to compute the number of parcels in the properties occupied by establishments. 85% of the manufacturing establishments in Montreal occupy properties composed of only one parcel,

13. Note that the Scott's dataset—though providing an extensive coverage of the manufacturing sector—is not the universe of productive plants in Canada. There may be some measurement error in the count of neighbors. We discuss the implication of this for the estimation results in section 1.4.

confirming our discussion with the Land Register of Quebec. Hence, as long as Montreal is representative of the rest of the country, measurement error remains very limited here.

Finally, for some plants, the surface area of the parcel is smaller than the building footprint. This is because the assignment of establishments to parcels on the one hand and to buildings on the other hand being done independently, an establishment can be assigned to a building and a parcel that do not correspond to the same lot. This may also come from polygons that are misidentified by the automatic recognition procedure (amalgamation of adjacent buildings). In a robustness check, we reproduce the results using only the sample of establishments for which we have both the parcel size and the building footprint, and for which the parcel size is larger than the building footprint.

1.2.2 Quality assessment

Assigning geocoded data to polygons delineated thanks to satellite images inherently brings issues regarding data quality and methodology accuracy. We relegate the detailed discussion of these issues to Appendix 1.7.4. We simply mention here that to gauge the quality of the data obtained after the whole geocoding and assignment process, 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". We can, therefore, randomly draw from the dataset we build a set of plants in QC and compare their parcel identifier from "Infolot" to the one obtained with the assignment procedure we follow. Based on this comparison we are able to build a quality variable—which we construct for the whole dataset, not just Quebec—with

three categories: Excellent, good, and acceptable (see Appendix 1.7.3 for additional details). In the remainder of the paper, we only keep observations with 'excellent quality' (77.1% of the observations for which we have a measure of parcel size). We check in a robustness test that our results hold when using observations with a lower quality.

1.2.3 The final dataset

The final sample used for the analysis contains the manufacturing plants from the Scott's database that: (i) are precisely geocoded (see Appendix 1.7.5) and (ii) have an excellent quality in terms of assignment to parcel- or building polygons (see 1.7.3). we further trimmed the 1% tails of each 3-digit industry. We now discuss its representativeness.

Out of the 32417 manufacturing plants recorded in the Scott's database for 2017, we can assign parcel size of excellent quality to 8708 (26.86%) of them and building footprint size of excellent quality to 20443 (73%) of them. The loss of data is mainly due to the absence of polygons for some cities and provinces and, to a lesser extent, to accuracy issues from the geocoding and polygon assignments (See Table 1.5 for more details). Concerning the sectoral representativeness of the estimation, Table 1.6 in Appendix 1.7.6 shows that the distribution of the (3-digit) industries is broadly similar to that in the raw Scott's database. The correlation between these distributions exceeds 0.98. Some sectors have few observations in the two samples (e.g., "313 Textile Mills", "316 Leather, allied product manufacturing" and "324 Petrol, coal product manufacturing"). We keep those sectors for the pooled analysis but we will not consider them for the sectoral analysis.

From a geographic perspective, Table 1.7 in Appendix 1.7.6 presents the distri-

bution of plants across provinces. As explained before, we lack parcel polygons for entire provinces, which are then missing in the estimation sample. Yet, the correlation of the geographic distribution with the raw Scott's data remains reasonably high (equal to 0.77). Moreover, the provinces that host the major part of manufacturing in Canada are very well represented. These provinces, namely Ontario, Quebec, and British Columbia, account for 80.1% of the overall number of manufacturing plants in Canada. They represent 87.3% of the plants in the parcel sample.

To conclude, the sectoral representativeness of the final dataset is excellent and its geographic coverage is good. There could still be some selection based on establishment characteristics. We hence use a probit model to assess the extent to which the establishments in the final sample exhibit specific characteristics compared to those for which we do not have reliable information on parcel size. Table 1.11 in Appendix 1.7.6 shows that beyond the geographic fixed effects, very few establishment characteristics are related to the probability to be in the regression sample. Moreover, the pseudo R-square of the regression is quite low, equal to 0.34. Hence, there is few selection in the sample used for the analysis, and most of it is related to selection across provinces due to the fact that we could not get parcel data for entire provinces in Canada.

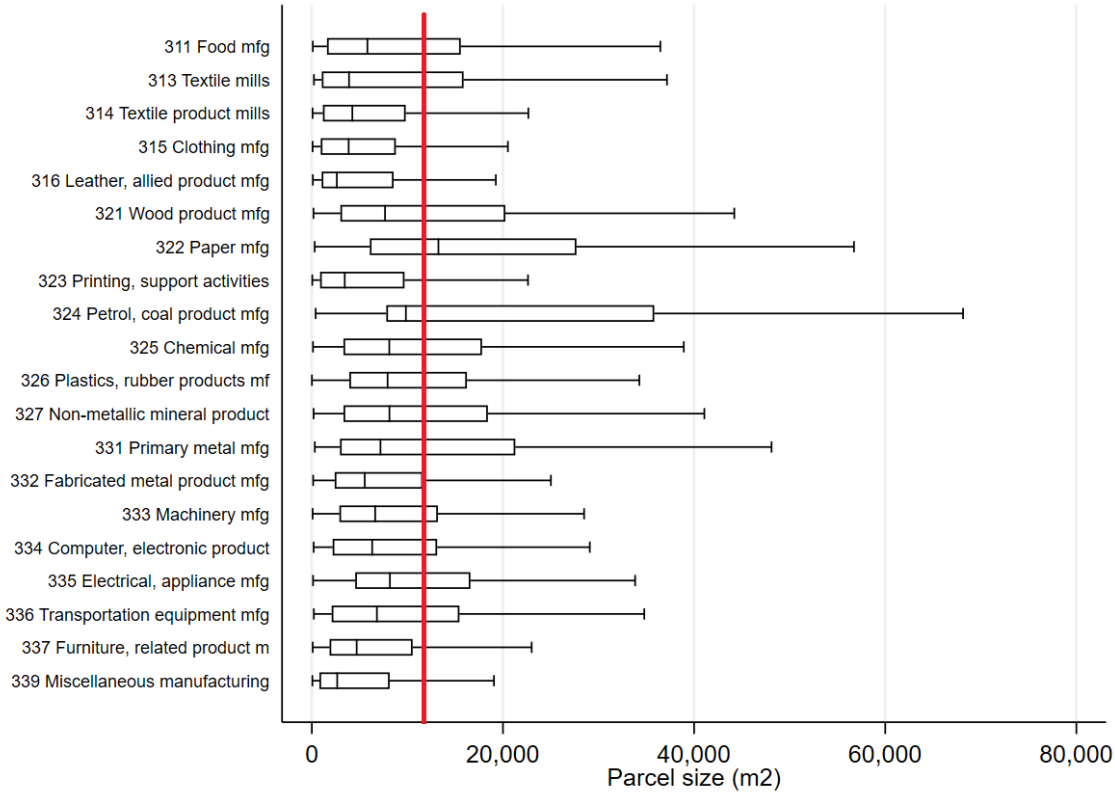
1.3 Land occupied by manufacturing establishments: Some sectoral statistics

We now present sectoral statistics on the size of the parcels occupied by manufacturing establishments in Canada, and on the share of their surface area covered by the footprint of the building on it (building-to-parcel ratio).

1.3.1 Size of parcels by sector

Figure 1.1 reveals substantial heterogeneity in the amount of land occupied by manufacturing establishments, both between and within sectors (Table 1.8 in Appendix 1.7.6 presents the figures associated to the graph).

Figure 1.1 – Plant-level parcel size by industry

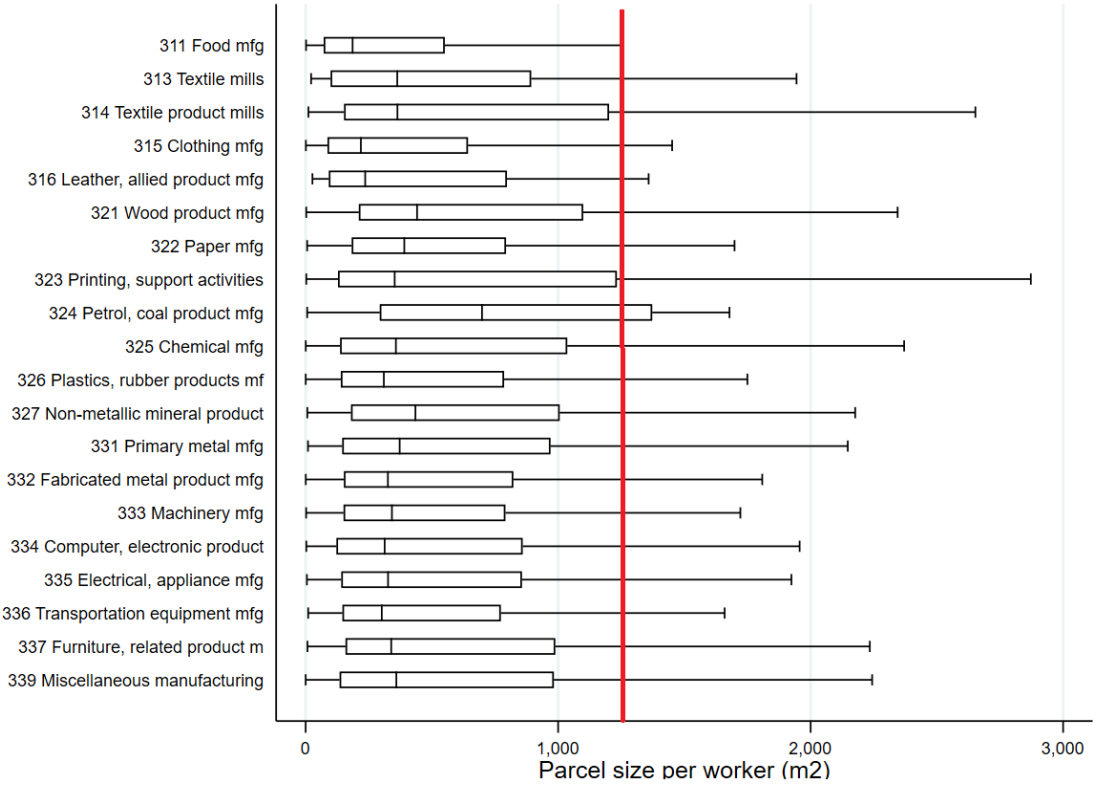


Notes: This graph represents the distribution of the parcel sizes across the industries. The industry 312 (Beverages and Tobacco) has been removed to keep the graph readable. The red line represents the mean value for the whole sample.

The average plant-level parcel size is around 13,350m^2, but the median is more than twice smaller, thus suggesting a right-skewed distribution of the parcel size. Moreover, the coefficient of variation equals 270%, revealing substantial heterogeneity in our sample. Part of that heterogeneity reflects between-sector

differences, with some sectors having on average large parcels (e.g., beverages and tobacco (not on the graph); primary metal; petroleum and coal products; and paper manufacturing) whereas others have much smaller parcels (e.g., leather; clothing and textile; printing; and miscellaneous manufacturing). The coefficient of variation is large in all sectors, ranging from 130% to 530% and showing that the size of parcels is not only heterogeneous between sectors but also within sectors.

Figure 1.2 – Plant-level parcel size per worker by industry



Notes: This graph represents the distribution of the parcel sizes per worker across the industries. The industry 312 has been removed for the sake of legibility. The red line represents the mean value for the whole sample.

Figure 1.2 focuses on parcel size per worker instead of parcel size (see Table 1.9 in Appendix 1.7.6 for the exact figures). Parcel size per worker measures how densely land is occupied by manufacturing establishments in terms of em-

ployees. The patterns on Figure 1.2 reveal even more heterogeneity between and within sectors than for parcel size (coefficient of variation of 430% for the whole sample, ranging from 130% to 880% across industries). In addition, the industry-level rank correlations between parcel size and parcel size per worker is not statistically significant: sectors with the highest plant-level parcel size are not necessarily those the most or the least densely occupied parcels in terms of workers.

Note that we checked that the ranking of industries in terms of average parcel size and parcel size per worker does not depend on the sample of establishments we use. We computed the figures for those establishments for which we observe both the parcel and the building size, for establishments with less than 50 workers, and for establishments with no identified neighbor on their parcel or in their building. The ranking of industries is not significantly different across samples, so that the patterns we uncover are not driven by specific sample-selection mechanisms (these figures are available upon request).

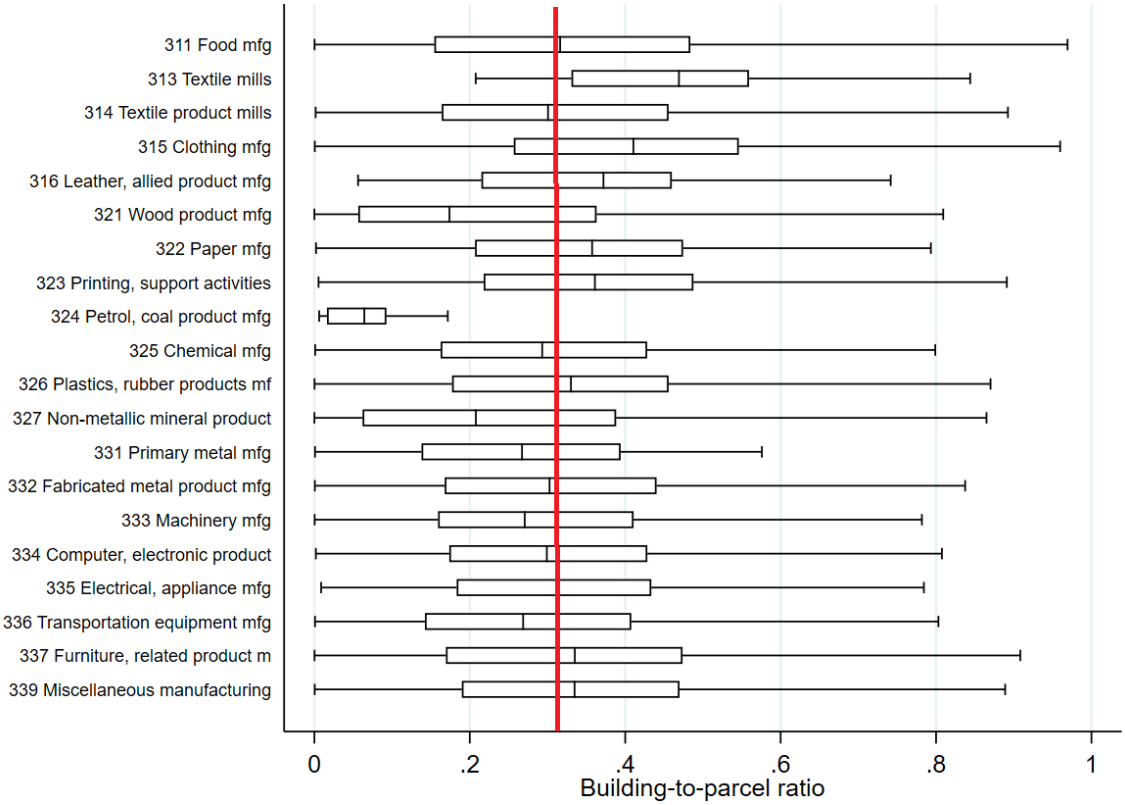
1.3.2 Building-to-parcel ratio

Another way to capture how densely occupied a parcel is is to compute the building-to-parcel ratio: the higher this ratio, the more densely occupied the parcel in terms of buildings. Figure 1.3 shows a fair amount of heterogeneity in terms of the building footprint-to-parcel ratio both between and within sectors, even though this heterogeneity is less important than the one observed for the parcel measure alone (the coefficient of variation equals 1060% for the whole sample and ranges from 220 to 1763%). Some sectors, like petroleum and coal, wood products, or primary metal products exhibit small building footprint-to-parcel ratios. On the contrary, textile mills, clothing, and printing exhibit high

ratios, thus showing that they use relatively less outdoor space.

The heterogeneity we observe across sectors certainly tells something about the different needs in terms of land across industries. The sectors with the lowest ratios definitely seem to be sectors that rely either on outdoor resources (wood, coal, non-metallic mineral products) or for which space for storage tanks (petrol, beverages) is important. By contrast, the sectors with the highest ratios are historically located in denser areas and belong to 'light manufacturing' (clothing, printing, textiles). Note that plants may react to higher land prices by reducing either their building footprint, their parcel footprint, or both; and that the ease with which either type of land ('indoors' or 'outdoors') can be adjusted may depend on the use (e.g., parking vs storage) and the industry. This might have important implications for the spatial sorting of sectors and their propensity to spatially agglomerate, a point we will return to later.

Figure 1.3 – Building-to-parcel ratios by NAICS 3-digit industry



Notes: This graph represents the distribution of the building-to-parcel ratios across the industries. The industry 312 has been removed for the sake of legibility. The red line represents the mean value for the whole sample which includes not only plants in CMA/CA but also those outside the CMA/CA. In addition we've constrained the sample such that the parcel size > building footprint size

1.4 Land occupied by manufacturing establishments: An econometric analysis

In this section, we provide a detailed analysis of the characteristics of the manufacturing establishments and their environment that determine the amount of land they occupy. In a context where the footprint of human activity becomes a first-order environmental issue, we focus on two variables: the surface area of the parcel occupied by an establishment divided by the number of workers it employs and the building-to-parcel ratio. The first variable is inversely related to the density of the parcel in terms of workers using it: the higher it is,

the lower the number of employees using the parcel. Moreover, as will become clear in Section 1.5, this variable lends itself well to a structural interpretation of the regression coefficients. The second variable is related to the density of the parcel in terms of building: the higher it is, the higher the fraction of the parcel covered by buildings. We first detail the equation to be estimated, we then present the benchmark results, and we finally propose some robustness checks.

1.4.1 Estimated equation

The equation we bring to the data is the following:

$$y_{isz} = \alpha \text{Env}_i + \beta \text{Estab}_i + \gamma \text{Infra}_i + \theta_s + \eta_z + \varepsilon_{isz} \tag{1.1}$$

where i stands for establishment, s for 4-digit NAICS industry code and z for the zone where the plant is located. y_{isz} denotes parcel size per worker or building-to-parcel ratio.

Env_i is a vector of characteristics related to the environment of establishment i. It includes the (log) size of the urban area where it is located both in terms of population and in terms of surface area (as per the 2016 Census),[14] the weighted distance of the establishment to the city centres of the urban area,[15] fixed effects identifying the type of zoning (commercial/industrial, residential,

14. The urban area corresponds to the Census Metropolitan Area (at least 100,000 of which 50,000 or more must live in the core) or the Census Agglomeration (a core population of at least 10,000).

15. To identify the centers of cities we use a routine that locates clusters of population density. The details of the procedure as well as the weighting scheme is presented in Appendix 1.7.7

recreational) that is prevalent at the location of the establishment, as well as a polynomial function of degree 4 of the number of neighbors falling on the same parcel. In some specifications, we also control for the (log) population density in the dissemination areas within a 500m radius around the establishment.

Estab_i is a vector of characteristics related to the size and the type activity run in establishment i. We control in particular for the (log) number of employees, dummies identifying headquarters and exporting plants, as well as for the NAICS 4-digit industries, products and broad sectors of activity in which the plant is involved.

Also, proximity to specific infrastructures might influence the amount of land per worker occupied by manufacturing establishments and their building-to-parcel ratio, either due to the size of the parcels available close to these infrastructures or to how "packed" establishments accept to be in order to enjoy proximity to these infrastructures. Infra_i is thus a vector containing the (log) distance between establishment i and the closest major airport, major seaport, freight station and highway junction.

Finally, θ_s and η_z stand for sector and economic regions fixed effects.[16] They account for technological parameters and regional determinants 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 1.2.2, we limit the sample to observations for which the information on parcel size or building footprint is of the highest quality.

---

16. 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.

1.4.2 Benchmark results

Table 1.1 shows estimation results when the parcel size per worker is used as the dependent variable. All regressions include industry- and economic region-fixed effects, as well as a polynomial function of degree 4 in the number of neighbors of the establishment on its parcel.

Table 1.1 – Determinants of parcel size per worker

	ln Parcel size per worker					
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of the local environment						
Ln Population CMA	-0.264^a	-0.242^a	-0.241^a	-0.227^a	-0.125^b	-0.237^a
	(0.070)	(0.056)	(0.055)	(0.065)	(0.051)	
Ln CMA surface area	0.081	0.064	0.068	0.088	0.025	0.047
	(0.091)	(0.067)	(0.067)	(0.083)	(0.069)	
Weighted Distance to city centers	0.040^a	0.038^a	0.038^a	0.033^a	0.025^a	0.220^a
	(0.003)	(0.004)	(0.004)	(0.006)	(0.004)	
$\mathbb{1}$ Residential	-0.504^a	-0.924^a	-0.918^a	-0.889^a	-0.707^a	-0.289^a
	(0.080)	(0.105)	(0.103)	(0.104)	(0.097)	
$\mathbb{1}$ Recreational	0.114	0.084	0.083	0.089	-0.000	0.022
	(0.094)	(0.107)	(0.107)	(0.105)	(0.091)	
Ln Population density 500m					-0.148^a	
					(0.012)	
Characteristics of the establishment						
Ln Employment		-0.619^a	-0.624^a	-0.626^a	-0.645^a	-0.590^a
		(0.016)	(0.016)	(0.016)	(0.017)	
$\mathbb{1}$ Headquarter			-0.070^b	-0.077^a	-0.058^b	-0.016^a
			(0.026)	(0.025)	(0.023)	
$\mathbb{1}$ Exporter			0.114^a	0.107^a	0.086^a	0.036 ^{a}
			(0.021)	(0.021)	(0.020)	
# functions in the estab.			-0.019	-0.019	-0.007	-0.007
			(0.039)	(0.037)	(0.035)	
# 4-digit NAICS in the estab.			-0.021^c	-0.020^c	-0.020^c	-0.018^c
			(0.012)	(0.012)	(0.012)	
# products produced in the estab.			-0.010^b	-0.010^b	-0.009^b	-0.017^b
			(0.004)	(0.004)	(0.004)	
Distance to transport infrastructure						
Ln Distance to major airport				-0.134^c	-0.009	-0.122^c
				(0.068)	(0.076)	
Ln Distance to major seaport				0.129^b	0.085^b	0.135^b
				(0.054)	(0.041)	
Ln Distance to freight station				0.020	0.022	0.016
				(0.052)	(0.026)	
Ln Distance to junction				-0.033	-0.007	-0.027
				(0.041)	(0.041)	
Observations	8,707	8,707	8,707	8,707	8,707	8,708
R-squared	0.287	0.561	0.564	0.568	0.588	0.568
Industry (4-digit) fixed effects	yes	yes	yes	yes	yes	yes
Economic region fixed effects	yes	yes	yes	yes	yes	yes
Controls for # neighbors	yes	yes	yes	yes	yes	yes

Notes: All regressions include a polynomial function of degree 4 in the number of neighbors of the establishment on its parcel. Only observations with the highest reliable information on parcel size are included. $\mathbb{1}$ denotes $\{0,1\}$ dummy variables. Standard errors clustered at the CMA/CA level in parentheses. ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1.

In the first column, the only other covariates are the characteristics of the geographic environment of the establishment. In column (2), the log employment of the establishment is added to the set of regressors. Column (3) accounts for other individual characteristics of the establishment, while we control in column (4) for the log distance to various transport infrastructure. We retain this regression as the benchmark specification. Indeed, through the lens of a theoretical framework proposed in Section 1.5, we can use its coefficients and others available in the literature to infer the substitution elasticity between land and labor in the production function of manufacturing plants. In column (5), we add to the set of environmental characteristics the log population density in a 500m radius around the establishment. Finally, we present in column (6) the standardized coefficients of the benchmark specification appearing in column (4). The regression results exhibit several robust patterns.

First, regarding the characteristics of the local environment, plants in populated urban areas use fewer land per worker, as well as plants that are located closer to city centres within urban areas (the coefficient on the weighted distance to city centres being positive). In our benchmark specification, the elasticity of parcel size per worker to city population is equal to -0.227 and the semi-elasticity to the weighted distance to city centres is equal to 0.033. Note that both coefficients decrease in absolute value when we control for the population density in the immediate surroundings of the establishment in column (5) (this latter variable attracting a negative and highly significant coefficient). This is coherent with the fact that population density is not homogeneous within cities and decreases on average with distance to the centre. Hence, when local population density is accounted for, the effect of the other two variables in weakened. Land prices are higher in big cities and lower at higher distances from city centres. We will thus show in the final section of this paper how we

can use these elasticities to recover values for the substitution elasticity between land and labor in the production function of establishments. Regarding zoning, not surprisingly, compared to establishments in commercial and industrial areas (the reference category), manufacturing plants in residential areas occupy fewer land per worker. This may be because the use of land is restricted or parcels are smaller in the residential parts of cities so that they attract establishments with lower land requirements.

Moving to establishment characteristics, as expected, headquarters occupy less land per worker, 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 required. The results also show that plants with a broader range of activity in terms of NAICS and products tend to occupy fewer land per worker, even though the relationship is less statistically significant. Finally, the highest correlation is found for establishment size in terms of employees, with an elasticity comprised between -0.65 and -0.60 depending on the specification: bigger establishments occupy fewer space per worker. We can see four explanations to this negative correlation. First, moving, opening, or closing a facility is costly so that firms adjust the size of the parcel they use less easily than their workforce; only when shocks are large and permanent enough do firms adjust their land consumption, most likely by moving or by opening and closing some establishments (Bergeaud and Ray, 2021). This means that when firms grow or shrink, they first do so by adjusting their number of employees only, especially if they face transitory shocks. Then, if big firms are firms that have grown a lot compared to their initial size, the negative correlation between parcel size per worker and establishment size could be related to the existence of adjustment cost. However, we ran the results controlling for plant-level employment growth between 2013 and 2017, and this leaves the

coefficient on establishment size unaffected.[17] Second, as already mentioned before, the Scotts data are exhaustive for manufacturing but not for services, so that we possibly mismeasure the number of neighbors on the parcel. If we do have a measurement error issue, it is arguably more severe for small firms: they are more likely to share their location with other businesses, and we may thus overestimate their parcel size per worker. This could also explain the negative correlation between parcel size per worker and establishment size we find. However, as shown on Figure 1.4, when we run the benchmark regression separately for bins of establishments with different sizes, the correlation between parcel size per worker and establishment size is close to -1 for establishments with 1-5 employees, and close to -0.6, the coefficient found for the whole sample, for establishments with 5-15, 15-50 and 50+ employees. This pattern of heterogeneity is inconsistent with the idea that the negative correlation estimated on the whole sample mostly reflects an overestimation of parcel size per worker for smaller plants. In the end, two explanations seem more likely to explain why parcel size per worker decreases with establishment size. First, it is suggestive of a fixed cost component in the land input. Indeed, if land is a variable cost only, under specific functional forms of the production function (CES for example, see Section 1.5 below), the quantity of land per worker is independent of the size of the firm in terms of employees. However, part of the land used by firms has the nature of a fixed cost: corridors, bathrooms, office spaces, or production spaces have a size that is partly independent of the number of workers using them. A second possible explanation is that even though this is not the most frequent situation, some manufacturing firms occupy multi-floor buildings. In Montreal for example, we know the number of

17. Results available upon request.

floors of the buildings occupied by establishments,[18] and it appears that 64% of the manufacturing establishments occupy a one-floor building, 16% a two-floor building, 5% a 3-floor building and 15% a 4^{+}-floor building. It is likely that bigger establishments occupy taller buildings but not necessarily much bigger parcels, which will show up into a lower parcel size per worker.

Finally, among the various types of transport infrastructure we consider, distance to major seaports is the only one to be robustly correlated with the amount of land per worker used by manufacturing establishments: plants that are located close to major seaports occupy fewer land per worker.

From a quantitative perspective, it is worth noting that the R^2 in our model is fairly large in all specifications, between 0.5 and 0.6, and that it is not entirely driven by the industry and economic region fixed effects. Thus, although we work with micro data at the establishment level, the model explains a substantial part of the variation in that data in terms of land per worker occupied by manufacturing plants. Among the regressors we take into account, the standardized coefficients displayed in column (6) show that four characteristics particularly stand out: the establishment size in terms of employees, being located in a residential zone, the population of the CMA and the weighted distance to city centres.

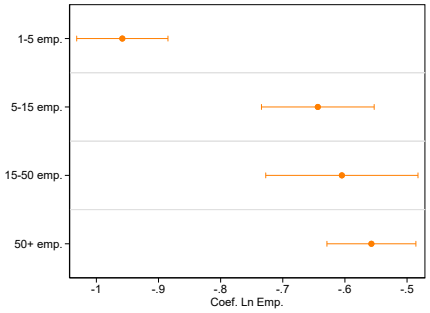
That parcels in big cities, as well as parcels in residential zones and in central locations within cities, are more densely occupied in terms of employees does not mean that the manufacturing establishments occupying these parcels use less floor space per worker: buildings on these parcels could cover a greater share of the parcel or they could have multiple floors. The data we have do not

18. This information comes from the property assessment roll we could get for Montreal. Unfortunately, we do not have it for the other cities in our sample.

allow to recover floor space, but we have the footprint of the building so that we can compute the building-to-parcel ratio: the higher this ratio, the more densely built the parcel occupied by an establishment. Note that taller buildings having a greater footprint, a higher building-to-parcel ratio is consistent with more floor space in both the "horizontal" and the "vertical" dimensions.

Table 1.2 displays the results when the building-to-parcel ratio is used as a dependent variable.

Figure 1.4 – Parcel size per worker and establishment size across size bins



Notes: This graph represents the coefficient on establishment size when the benchmark regression in column (4) of Table 1.1 is run separately for each bin.

Two of the four main determinants of the parcel size per worker appear as important determinants of the building-to-parcel ratio too: the city population and the weighted distance to city centres, with standardized coefficients well above 0.2 in absolute value. The building-to-parcel ratio increases with city size and decreases with the weighted distance to city centres. This likely reflects the fact that outdoor space is partly used for parking lots or green space, two dimensions on which firms can accept restrictions when land prices are high. We also know 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 center (Brueckner

and Franco, 2017). Indoor space probably exhibits, on the contrary, a stronger complementarity with the other production factors and can less easily be compressed.

Interestingly, the coefficient on establishment size, which is by far the main determinant of parcel size per worker, is now very close to 0. This suggests that as for parcel size per worker, building footprint per worker decreases with establishment size. This is confirmed by the results in Table 1.12 in Appendix 1.7.8 where the dependent variable is the building footprint. Building footprint per worker decreases strongly with establishment size, with an elasticity which is very close to the one obtained for parcel size. i.e. -0.65 to -0.60. Interestingly, we also find that building footprint increases with city size and decreases (even though the coefficient is not always significant) with the weighted distance to city centres. This is highly suggestive that manufacturing plants occupy taller buildings in big cities and in central locations within cities. If it were not the case, considering that land prices are higher in big cities and in central locations, we should observe that building footprint decreases with city size and increases with distance to the centre. These patterns are also consistent with the observation that multi-floor manufacturing firms are more often found in large and dense cities.

Table 1.2 – Determinants of building-to-parcel ratio

	ln Building-to-parcel ratio					
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of the local environment						
Ln Population CMA	0.366[a]	0.367[a]	0.370[a]	0.391[a]	0.321[a]	0.457[a]
	(0.074)	(0.075)	(0.074)	(0.086)	(0.079)	
Ln CMA surface area	-0.147	-0.147	-0.150	-0.209[c]	-0.164	-0.126[c]
	(0.099)	(0.100)	(0.099)	(0.123)	(0.116)	
Weighted Distance to city centers	-0.045[a]	-0.045[a]	-0.045[a]	-0.040[a]	-0.034[a]	-0.298[a]
	(0.005)	(0.005)	(0.005)	(0.007)	(0.006)	
⊮ Residential	0.432[a]	0.411[a]	0.413[a]	0.386[a]	0.261[a]	0.142[a]
	(0.073)	(0.072)	(0.072)	(0.085)	(0.074)	
⊮ Recreational	-0.228[b]	-0.230[b]	-0.226[b]	-0.232[b]	-0.172[c]	-0.062[b]
	(0.090)	(0.090)	(0.089)	(0.095)	(0.094)	
Ln Population density 500m					0.101[a]	
					(0.017)	
Characteristics of the establishment						
Ln Employment		-0.031[a]	-0.028[a]	-0.029[a]	-0.016	-0.031[a]
		(0.008)	(0.009)	(0.009)	(0.010)	
⊮ Headquarter			-0.051	-0.041	-0.055	-0.010
			(0.038)	(0.034)	(0.034)	
⊮ Exporter			-0.024	-0.019	-0.006	-0.007
			(0.020)	(0.020)	(0.020)	
# functions in the estab.			-0.035	-0.032	-0.040	-0.014
			(0.041)	(0.038)	(0.037)	
# 4-digit NAICS			0.004	0.001	0.000	0.001
			(0.016)	(0.015)	(0.016)	
# products			0.009[c]	0.009[b]	0.009[b]	0.019 [b]
			(0.004)	(0.004)	(0.004)	
Distance to transport infrastructure						
Ln Distance to major airport				0.211[a]	0.125	0.216[a]
				(0.074)	(0.079)	
Ln Distance to major seaport				-0.120[b]	-0.090[c]	-0.143[b]
				(0.054)	(0.053)	
Ln Distance to freight station				-0.077	-0.078[c]	-0.068
				(0.057)	(0.040)	
Ln Distance to junction				-0.030	-0.048	-0.027
				(0.051)	(0.049)	
Observations	8,513	8,513	8,513	8,513	8,513	8,514
R-squared	0.290	0.291	0.292	0.299	0.311	0.299
Industry (4-digit) fixed effects	yes	yes	yes	yes	yes	yes
Economic region fixed effects	yes	yes	yes	yes	yes	yes

Notes: All regressions include a polynomial function of degree 4 in the number of neighbors of the establishment on its parcel. Only observations with the highest reliable information on parcel size are included. ⊮ denotes $\{0,1\}$ dummy variables. Standard errors clustered at the CMA/CA level in parentheses. [a] p<0.01, [b] p<0.05, [c] p<0.1.

1.4.3 Robustness checks

Tables 1.3 provides several robustness checks on the determinants of both the parcel size per worker and the building-to-parcel ratio. We present the coefficients for three variables of interest only, namely city population, weighted distance to city centres and establishment size, but all the covariates of the benchmark specification are included in the regressions.

We propose eight different checks. We replicate in column (1) the benchmark results. In column (2), we expand the sample to all of the establishments for which we have information on the dependent variable irrespective of the quality of the geocoding and polygon assignment process. In column (3), we eliminate the 1% tails of the distribution in terms of the dependent variable. In column (4), the sample only contains observations for which the information on both the parcel size and the building footprint is of top quality. In column (5), we eliminate observations for which parcel size is smaller than building footprint (reflecting misidentified polygons or misassignment to parcel and/or building polygons). In column (6), we restrict the sample to manufacturing establishments with less than 50 employees to address the fact that large establishments may occupy several distinct adjacent parcels, in which case we under-estimate the amount of space they use (even though, as mentioned in section 1.2.1, this case is certainly rare). In column (7), we restrict the sample to those establishments that have no identified neighbors on the same parcel or in the same building. Indeed, despite the fact that we control for the number of neighbors in the benchmark regressions, it is still possible that we mismeasure the actual amount of land occupied by establishments when several manufacturing firms occupy the same parcel. In the same vein, in column (8), we replace the number of neighbors by the mean of the number of neighbors on the parcel and in the

building when both values are available. Indeed, the spatial junction between establishments and parcels on the one hand, and establishments and buildings on the other hand, being done independently, these two figures may not be the same. Finally, we focus in column (9) on establishments that are located farther than 5 kilometers from a city centre, to ensure the patterns we uncover are not driven by what happens in very central locations.

Irrespective of whether we explain the parcel size per worker or the building-to-parcel ratio, the results for city population and weighted distance to city centres are remarkably stable, both qualitatively (for all of them) and quantitatively (most of the time). The same applies to the relationship between parcel size per worker and establishment size, while the sign and significance of the correlation between building-to-parcel ratio and establishment size is less stable, but always close to zero. We are thus confident in the lessons we learn from the benchmark econometric analysis: controlling for establishment size, manufacturing establishments occupy parcels in a denser way in big cities and in central locations within cities. Moreover, controlling for location, bigger establishments occupy much fewer land per worker as measured by parcel size per worker.

Table 1.3 – Robustness checks

	ln PB-measure per worker								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Ln Population CMA	-0.227^a	-0.191^a	-0.224^a	-0.221^a	-0.142^a	-0.213^a	-0.222^a	-0.245^a	-0.165^c
	(0.065)	(0.055)	(0.059)	(0.066)	(0.049)	(0.066)	(0.066)	(0.067)	(0.085)
Weighted Distance to city centers	0.033^a	0.027^a	0.030^a	0.032^a	0.021^a	0.034^a	0.033^a	0.036^a	0.028^a
	(0.006)	(0.004)	(0.005)	(0.005)	(0.003)	(0.007)	(0.005)	(0.006)	(0.007)
Ln Employment	-0.626^a	-0.655^a	-0.568^a	-0.625^a	-0.630^a	-0.709^a	-0.542^a	-0.633^a	-0.620^a
	(0.016)	(0.017)	(0.014)	(0.016)	(0.019)	(0.018)	(0.015)	(0.019)	(0.020)
Observations	8,707	12,138	8,531	8,048	7,254	6,947	5,333	8,707	6,793
R-squared	0.568	0.488	0.549	0.570	0.596	0.547	0.458	0.538	0.579
	ln Building-to-parcel ratio								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Ln Population CMA	0.391^a	0.358^a	0.285^a	0.368^a	0.252^a	0.386^a	0.362^a	0.379^a	0.289^a
	(0.086)	(0.059)	(0.072)	(0.082)	(0.049)	(0.095)	(0.096)	(0.080)	(0.097)
Weighted Distance to city centers	-0.040^a	-0.036^a	-0.032^a	-0.038^a	-0.022^a	-0.043^a	-0.041^a	-0.040^a	-0.031^a
	(0.007)	(0.005)	(0.005)	(0.006)	(0.004)	(0.008)	(0.006)	(0.006)	(0.007)
Ln Employment	-0.029^a	-0.012	-0.010	-0.034^a	0.018^c	0.018	-0.032^b	-0.016	-0.013
	(0.009)	(0.012)	(0.010)	(0.008)	(0.010)	(0.018)	(0.014)	(0.010)	(0.010)
Observations	8,513	11,793	8,341	8,048	7,254	6,830	5,206	8,513	6,629
R-squared	0.299	0.251	0.285	0.292	0.220	0.312	0.314	0.300	0.283
Industry (4-digit) fixed effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
Economic region fixed effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
Controls for neighbors	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: All regressions include a polynomial function of degree 4 in the number of neighbors of the establishment on its parcel. Only observations with the highest reliable information on parcel size are included. $\mathbb{1}$ denotes $\{0,1\}$ dummy variables. Standard errors clustered at the CMA/CA level in parentheses. ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1. See main text for a description of the sample used in each regression.

1.5 Conceptual framework and structural interpretation

We now propose a simple conceptual framework to interpret the empirical regularities described in the previous section and use it to give a structural interpretation for some of the elasticities we estimated.

1.5.1 Setup

We assume that input markets are competitive and that firms are price takers in factor markets. Let i denote firms, s sectors, and z zones. We index firms by $i(s,z)$, meaning that firm i belongs to sector s and is located in zone z. When there is no confusion, we use i for short. Firm i produces with the following modified CES production function: [19]

$$Y_{i(s,z)} = A_i \left\{ \alpha_{i(s,z)} \left[\kappa_{i(s,z)} (P_i - \overline{P}_s) \right]^{\frac{\sigma_s - 1}{\sigma_s}} + (1 - \alpha_{i(s,z)}) L_i^{\frac{\sigma_s - 1}{\sigma_s}} \right\}^{\frac{\sigma_s}{\sigma_s - 1}} \tag{1.2}$$

where Y_i, P_i, and L_i stand for firm-level output, parcel (land) inputs, and the number of workers, respectively; A_i is a Hicks-neutral productivity shifter; $\alpha_{i(s,z)}$ is a technological parameters (with possibly both a sectoral and a firm component) that influences the intensity of the production function in land and labor; and $\kappa_{i(s,z)}$, is a land-augmenting productivity parameter specific to firm i (and/or its industry and zone). Observe that there is a minimum land requirement $\overline{P}_s$ in sector s, which captures the presence of some fixed costs or indivisibilities in the consumption of land. Furthermore, σ_s is the elasticity of substitution between land and labor. We assume this technological parameter is industry specific and $\sigma_s > 0$, i.e., land and labor inputs are imperfect substitutes in production. The production function exhibits constant returns to scale at the firm level in $P_i - \overline{P}_s$ and L_i. [20]

---

19. We can easily add capital as a third production factor. If we consider that its price is constant across space this does not change the analysis. With different prices across locations, the analysis becomes more involved. Since we are mostly interested in parcel size per worker, we do not develop the case with capital in more detail in this paper.

20. We do not rule out the presence of increasing returns to scale external to the firm. This would, e.g., be the case when $A_i = A_i(L_{s,z})$ depends on aggregate employment $L_{s,z}$ in sector s and zone z.

Letting p_z and w_z denote the unit price for parcels and labor in zone z, respectively, standard unit cost minimization yields:

$$p_z = \frac{\alpha_{i(s,z)}\kappa_{i(s,z)}[\kappa_{i(s,z)}(P_i-\overline{P}_s)]^{-\frac{1}{\sigma}}Y_{i(s,z)}}{\alpha_{i(s,z)}[\kappa_{i(s,z)}(P_i-\overline{P}_s)]^{\frac{\sigma_s-1}{\sigma_s}}+(1-\alpha_{i(s,z)})L_i^{\frac{\sigma_s-1}{\sigma_s}}} \tag{1.3}$$

$$w_z = \frac{(1-\alpha_{i(s,z)})L_i^{-\frac{1}{\sigma_s}}Y_{i(s,z)}}{\alpha_{i(s,z)}[\kappa_{i(s,z)}(P_i-\overline{P}_s)]^{\frac{\sigma_s-1}{\sigma_s}}+(1-\alpha_{i(s,z)})L_i^{\frac{\sigma_s-1}{\sigma_s}}} \tag{1.4}$$

We focus on the ratio P_i/L_i as this is the theoretical equivalent of the parcel size per worker used in our empirical analysis. Since

$$\frac{p_z}{w_z}=\frac{\alpha_{i(s,z)}}{1-\alpha_{i(s,z)}}(\kappa_{i(s,z)})^{\frac{\sigma_s-1}{\sigma_s}}\left(\frac{L_i}{P_i-\overline{P}_s}\right)^{\frac{1}{\sigma_s}},$$

we can express parcel size per unit of labor as follows:

$$\frac{P_i}{L_i}=\left(\frac{\alpha_{i(s,z)}}{1-\alpha_{i(s,z)}}\right)^{\sigma_s}(\kappa_{i(s,z)})^{\sigma_s-1}\left(\frac{p_z}{w_z}\right)^{-\sigma_s}+\frac{\overline{P}_s}{L_i}. \tag{1.5}$$

Observe that the ratio P_i/L_i is independent of L_i if land has no fixed-cost component (i.e., if $\overline{P}_s=0$). We have seen in our empirical analysis that this is not the case. Since empirically P_i/L_i decreases with L_i, this suggests that $\overline{P}_s>0$.

1.5.2 Determinants of P_i/L_i

Our conceptual framework highlights three main types of determinants of firm-level parcel size per worker: (i) the relative price of land; (ii) technological/productivity parameters; and (iii) an additional term that depends on firm size and the importance of land as a fixed factor of production.

Relative price of land. Conditional on technological and productivity parameters and firm size, the firm-level parcel size per worker is a decreasing

function of its relative price, its sensitivity being determined by the elasticity of substitution σ_s between production factors.

Technological/productivity parameters. Firm-level parcel size per worker also depends on technological parameters and due to the spatial sorting of plants, these are probably correlated in the data with the local relative price of production factors. For example, the land-intensity of the firm-level production function, as determined by the technological parameter $\alpha_{i(s,z)}$, matters for the relative quantity of land used by firms. For a given relative price of land, firms with low $\alpha_{i(s,z)}$ use relatively less land. We should then observe that firms with low $\alpha_{i(s,z)}$ sort into places where land is relatively expensive. The discussion is more involved for the land-augmenting productivity parameter $\kappa_{i(s,z)}$. Whether high or low $\kappa_{i(s,z)}$ firms sort into zones where land is expensive depends on the value of σ_s. Indeed, equation (1.5) above requires us to distinguish three cases:

(i) When $\sigma_s = 1$, any variation in factor-augmenting productivity $\kappa_{i(s,z)}$ leaves the relative demand P_i/L_i unaffected. There is no obvious spatial sorting of firms based on land-augmenting productivity in this case.

(ii) When $\sigma_s < 1$, firms cannot easily substitute workers for land. For a given relative price of land, firms with a high $\kappa_{i(s,z)}$ will then use less land per worker. Put differently, it is optimal for firms with a high land-augmenting productivity to tilt their demand towards non-land inputs when production factors are not easily substitutable. In this case, we should observe that firms with a high land-augmenting productivity sort into places where the relative price of land is high.

(iii) When $\sigma_s > 1$, firms can easily substitute workers for land. For a given relative price of land, firms with a high $\kappa_{i(s,z)}$ will then use more land per worker.

Put differently, it is optimal for firms with a high land-augmenting productivity to tilt their demand towards the land input when production factors are easily substitutable. In this case, we should see firms with a relatively low land-augmenting productivity sort into places where land is relatively expensive.

Before proceeding, it is also worth noting that high-A_i establishments are more likely to locate in zones with high factor costs (parce prices p_z and wages w_z) since only they can afford the high production costs there. We formally show this in Appendix 1.8. However, as equation (1.5) reveals, the parcel size per worker P_i/L_i used by establishments does not depend on their total factor productivity A_i, since the latter is a Hicks-neutral productivity shifter. [21]

Fixed land requirements. Last, for a given relative price of land and technological parameters, firms in sectors with larger fixed requirements mechanically use more land per worker.

1.5.3 Implications for empirical estimation

Because of the presence of the $\overline{P}_s$, we cannot readily log-linearize equation (1.5). However, we can proceed as follows. We first rewrite equation (1.5):

$$\begin{aligned}\frac{P_i}{L_i} &= \underbrace{\left(\frac{\alpha_{i(s,z)}}{1-\alpha_{i(s,z)}}\right)^{\sigma_s} (\kappa_{i(s,z)})^{\sigma_s-1} \left(\frac{p_z}{w_z}\right)^{-\sigma_s}}_{\equiv \xi_{i(s,z)}} + \frac{\overline{P}_s}{L_i} \\ &= \left(\frac{\alpha_{i(s,z)}}{1-\alpha_{i(s,z)}}\right)^{\sigma_s} (\kappa_{i(s,z)})^{\sigma_s-1} \left(\frac{p_z}{w_z}\right)^{-\sigma_s} \left(1 + \frac{\overline{P}_s}{\xi_{i(s,z)} L_i}\right). \qquad (1.6)\end{aligned}$$

21. Hence, external returns to scale that may be subsumed in the Hicks-neutral productivity shifter also do not affect firms' parcel size per worker. This vindicates the fact that we do not attempt to control for agglomeration effects in our empirical analysis.

We then log-linearize equation (1.6), introducing a constant term β_0 and a reduced-form error term $\tilde{\varepsilon}_i$, and obtain

$$\ln\left(\frac{P_i}{L_i}\right) = \beta_0 + \beta_1 \ln\left(\frac{p_z}{w_z}\right) + \varepsilon_i \tag{1.7}$$

where

$$\varepsilon_i = \sigma_s \ln\left(\frac{\alpha_{i(s,z)}}{1-\alpha_{i(s,z)}}\right) + (\sigma_s - 1)\ln\kappa_{i(s,z)} + \ln\left(1 + \frac{\overline{P}_s}{\xi_{i(s,z)} L_i}\right) + \tilde{\varepsilon}_i \tag{1.8}$$

is a structural error term.

In the previous section, we have estimated the elasticity of parcel size per worker to city population and the semi-elasticity of parcel size per worker to the weighted distance to city centres. Through the lens of the model, $\beta_1 = -\sigma_s$. It then follows that:

$$\frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \ln \text{Pop}_z} = -\sigma_s \frac{\partial \ln\left(\frac{p_z}{w_z}\right)}{\partial \ln \text{Pop}_z} \quad \textit{and} \quad \frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \text{Dist}_i} = -\sigma_s \frac{\partial \ln\left(\frac{p_z}{w_z}\right)}{\partial \text{Dist}_i}.$$

From the literature, $\partial \ln(p_z/w_z)/\partial \ln \text{Pop}_z > 0$ and $\partial \ln(p_z/w_z)/\partial \text{Dist}_i < 0$, and our regressions show $\partial \ln(P_i/L_i)/\partial \ln \text{Pop}_z < 0$ and $\partial \ln(P_i/L_i)/\partial \text{Dist}_i > 0$. Hence, our estimates imply a positive value for σ_s, which is reassuring. Quantitatively, considering for now that the elasticities $\partial \ln(p_z/w_z)/\partial \ln \text{Pop}_z$ and $\partial \ln(p_z/w_z)/\partial \text{Dist}_i$ are given, we can infer σ_s as:

$$\sigma_s = -\frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \ln \text{Pop}_z} \Big/ \frac{\partial \ln\left(\frac{p_z}{w_z}\right)}{\partial \ln \text{Pop}_z} = -\frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \text{Dist}_i} \Big/ \frac{\partial \ln\left(\frac{p_z}{w_z}\right)}{\partial \text{Dist}_i} \tag{1.9}$$

However, the theoretical discussion in section 1.5.2 shows that before structurally interpreting the elasticities estimated in the previous section, we need to discuss the endogeneity issues arising from the presence of $\alpha_{i(s,z)}$, $\kappa_{i(s,z)}$, and $\ln\left(1 + \frac{\overline{P}_s}{\xi_{i(s,z)} L_i}\right)$ in the structural error term (1.8).

Spatial sorting of firms and endogeneity. The type of bias arising from the firm-specific requirements in terms of land and labor is straightforward: low $\alpha_{i(s,z)}$ firms sort into high p_z/w_z zones and—in the absence of controls or valid instruments—the OLS estimate of $\partial \ln(P_i/L_i)/\partial \ln \text{Pop}_z$ is likely to be biased downward and the one of $\partial \ln(P_i/L_i)/\partial \text{Dist}_i$ is likely to be biased upward. The spatial sorting of firms based on $\kappa_{i(s,z)}$ induces the same type of biases, but this is less straightforward to establish. Indeed, as discussed before, three cases need to be distinguished:

(i) When $\sigma_s = 1$, spatial sorting of firms based on land-augmenting productivity is not an issue and there is no endogeneity bias.

(ii) When $\sigma_s < 1$, firms with a high land-augmenting productivity sort into places where the relative price of land is high. In this case, the naive estimates of $\frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \ln \text{Pop}_z}$ and $\frac{\partial \ln\left(\frac{P_i}{L_i}\right)}{\partial \text{Dist}_i}$ suffer from a downward and an upward bias repsectively.

(iii) When $\sigma_s > 1$, firms with a low land-augmenting productivity sort into places where land is relatively expensive, meaning that again the naive estimates of the coefficients suffer from the same biases.

To summarize, the direction of the bias related to the spatial sorting of firms based on their technological and productivity parameters is always the same: in the absence of adequate controls or valid instruments, $\partial \ln(P_i/L_i)/\partial \ln \text{Pop}_z$ is likely to be underestimated and $\partial \ln(P_i/L_i)/\partial \text{Dist}_i$ over-estimated. In both cases, we obtain an upper bound for σ_s .

Fixed costs and firm size. The second type of bias arises if land has a fixed-cost component. More productive and thus larger firms are more likely to

be found in high p_z/w_z zones. Since they also have a smaller parcel size per worker—because the fixed requirements are distributed over a larger workforce—not controlling for this leads to a downward biased estimate of $\partial \ln (P_i/L_i)/\partial \ln \text{Pop}_z$ and an upward biased estimates of $\partial \ln (P_i/L_i)/\partial \text{Dist}_i$.

1.5.4 Inferring σ_s

Following the above discussion, the possible fixed-cost components of land as well as the technological and productivity parameters need to be controlled for to infer meaningful values of σ_s from the estimation of $\partial \ln (P_i/L_i)/\partial \ln \text{Pop}_z$ and $\partial \ln (P_i/L_i)/\partial \text{Dist}_i$. The former is relatively straightforward to do: this is the reason why we included firm size as an additional explanatory variable into our econometric analysis. As we have shown before, the elasticity of P_i/L_i with respect to L_i is negative, highly significant, and very stable across specifications. Thus, land has a fixed component and we did control for it.

Let us now discuss the technological and productivity parameters, which are less straightforward to control for. We do neither directly observe firms' relative land requirements $\alpha_{i(s,z)}$ nor their land-augmenting productivity parameters $\kappa_{i(s,z)}$. However, the various controls we use in our regressions in Section 1.4 are likely correlated with these parameters and thus should be proxies for them.

Firm-level relative land requirements are certainly partly determined by some sectoral parameters of the production function. These are accounted for by the NAICS 4-digit industry fixed effects in our regressions. Still, the part of $\alpha_{i(s,z)}$ that is specific to establishments and $\kappa_{i(s,z)}$ are not controlled for by such fixed effects. However, we believe that four sets of our controls deal with this. First, land requirements vary with the functions that the establishment carries out

and with its international exposure. We actually showed that headquarters occupy less land per worker while the opposite is true for exporters. Second, local governments have various zoning policies that affect the quantity of land available for production. In particular, zones that are specifically dedicated to an industrial or commercial use are probably more attractive to firms that occupy a lot of land per worker, while the opposite should be true for residential zones where land suitable for production is likely to be scarce. This is why we also controlled for land-use fixed effects. Third, the proximity of transport infrastructure could affect the quantity of land available for production or attract firms that are specific in terms of their needs for land (e.g., the exporters we mentioned above). This is why we control for the distance to the closest major airport, major seaport, train freight station, and highway junction, even though these variables do not appear to be major determinants of parcel size per worker in the end. Finally, establishments involved in high number of activities in terms of products or sectors may require different types of facilities (multi-floor buildings for example to separate the various product lines), which could imply specific technological and productivity parameters regarding the land input. This is why we added to our regressions controls for the number of products, 4-digit NAICS industries and broad sectors covered by its operations.

Using the relationships highlighted in equation (1.9) and the estimation results in column (4) of Table 1.1, we now gauge the value of σ_s. We start with $\partial \ln(P_i/L_i)/\partial \ln \text{Pop}_z$, which we estimate to be equal to -0.227. To back out σ_s, we then need a value for $\partial \ln(p_z/w_z)/\partial \ln \text{Pop}_z$. Data on (commerical) land prices are notoriously difficult to find, and we do not have them for Canada. In other words, we do not observe the relative price p_z/w_z. We thus rely on estimates available in the existing literature. Based on French data, Combes et al. (2019) find that the elasticity of the price of parcels (per square metre) to city popula-

tion 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. [22] In France, the elasticity of relative land prices to city size/city population density is thus equal to 0.57. Taking this as a reference value for Canada, equation (1.9) implies a value of $\sigma_s = 0.227/0.57 = 0.4$.

We repeat the same exercise for the estimate of $\partial \ln(P_i/L_i)/\partial \text{Dist}_i$, equal to 0.033. Again, we are not aware of clean estimates of land price gradients for 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). This corresponds to a semi-log gradient of 0.197. [23] Since urban areas are delineated following commuting patterns, the average wage should not vary too much within them. Then, taking -0.197 as the reference value for $\partial \ln(p_z/w_z)/\partial \text{Dist}_i$ for Canada and using equation (1.9), it follows that $\sigma_s = 0.033/0.197 = 0.17$.

The two values of σ_s implied by the quantification exercises we propose, 0.17 and 0.4, are quite far from the ubiquitous Cobb-Douglas specification that has been used in the existing literature. They that labor and land, as measured by parcel size, are complements rather than substitutes in the production function

---

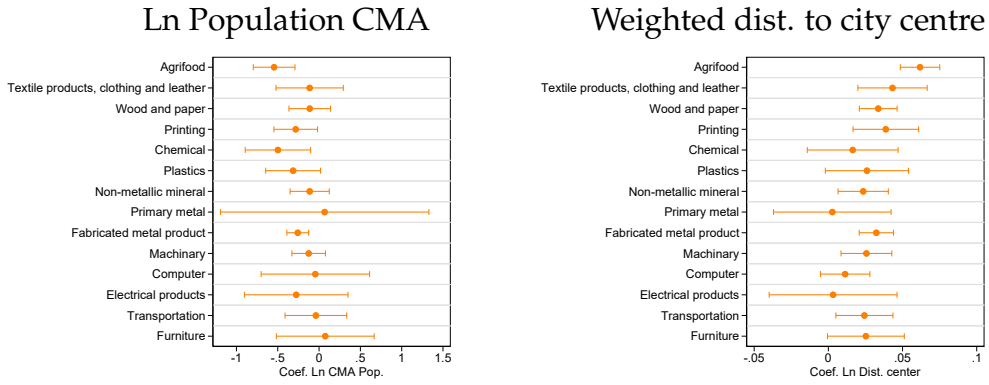
22. These two elasticities are not estimated over the same period of time and at the exact same spatial scale, but they are cleanly estimated with very detailed data. We are not aware of better estimates in the literature to obtain a measure of $\frac{\partial \ln\left(\frac{p_z}{w_z}\right)}{\partial \ln \text{Pop}_z}$. Moreover, the two regressions from which these estimates derive both contain the surface area of the unit over which population and population density are computed. In such an empirical framework, the elasticity to population and to population density are equivalent.

23. 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 $-\ln(6.5)/9.5 = -0.197$.

of manufacturing establishments.

In our 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 with our data. Figure 1.5 reports the sectoral estimates for the two covariates of interest, city population and weighted distance to city centres. In line with the pooled results, the coefficients we obtain are most of the time negative (but not always significant due to noisy estimates sometimes) for city population, and they are always positive (and most of the time significant) for weighted distance to city centres. However, once standard errors are accounted for, it is hard to see any significant heterogeneity across sectors. This suggests that σ_s is low in every sector, so that most of the heterogeneity across sectors highlighted in Section 1.3 comes from productivity/sectoral parameters.

Figure 1.5 – Heterogeneity of coefficients by sector - Parcel size per worker



Notes: The graphs shows the point estimate and the 10% confidence interval of the coefficient associated with the explanatory variable indicated in the heading. Regressions are run separately for the various sectors using the benchmark specification in column (4) of Table 1.1 .

1.6 Conclusion

To the best of our knowledge, we are the first to use data on the quantity of land used by manufacturing establishments to investigate the role of land as a production factor. We uncover several interesting stylized facts. First, there is substantial between- and within-sectors heterogeneity in establishments' land consumption per worker. Land per worker is strongly negatively related to establishment size, showing that land has a strong fixed cost component. Land consumption per worker is not significantly related to city size. This means that, even tough land prices are on average higher in big cities, there are still parts of big cities where land is cheap enough so that the consumption of land per worker does, on average, not depend on city size. Moreover, land consumption per worker is negatively related to local population density, but that correlation is small in absolute value. As the relative price of land increases with local population density, firms reduce their consumption of land per worker in denser places, but the small value of this correlation suggests a strong complementarity between land and other production factors. Finally, land as a production factor exhibits important adjustment costs: conditional on current size, firms that grew fast over last four year use fewer space per worker.

The facts we uncover may help explain why the rare attempts at quantifying the role of land for production of manufacturing firms have concluded to a limited role for the former. These studies usually rely on the estimation ofa Cobb-Douglas production function, which can hardly capture the fixed cost dimension of land. Moreover, these studies ignore adjustment costs for land inputs. Establishing and estimating a production function framework suited to the quantification of the contribution of land as a production factor should have a high priority but is beyond the scope of this paper.

1.7 Appendix to chapter 1

1.7.1 Geocoding

The geocoding process 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 provides in addition also 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 Scott's 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 Scott's 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. We merge the Scott's 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 Scott's 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 Scott's 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' (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.

1.7.2 Data sources and quality

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 Street Map (OSM): An open-source database built by people worldwide to create free editable geographic data. This data source has information on building footprints, yet the buildings polygons recorded in OSM on a voluntary basis are not comprehensive at the country level. As of March 02, 2019, the total number of Canadian buildings footprints recorder in OSM amounted to roughly 12.6 million out of a total number estimated to 15 million. The set of polygons that we manage to collect has a country-wide coverage and represents roughly 82 % of the sets of addresses in Canada.

The table below provides the complete list of polygon datasets that we collected along with the links where they can be accessed.

Table 1.4 – Overview of datasources

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: This table reports list of the open source data that we use to construct our land measure. Most of these sources are open to the public. In addition, we also proprietary data on parcels form the province of Quebec with the permission of our University for this research purpose.

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:

— 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.

— 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.

1.7.3 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 1.6 shows an example of how the geocoded Scott's plants are overlaid on the building polygon layer for the spatial join process.

Figure 1.6 – Example of the polygon layer, overlaid with geocoded establishments

As is well known, geocoding is 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 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". Note that this variable is built for each of our three land-

consumption measures as the results generated by the assignation process is not systematically the same for each plant across the three measures.

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 1.2.2); 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 "Acceptable" 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 size, the process leads to grade as "Excellent" the plants whose assignment category is "Within-Border-Center" or "Within". These plants with an "Excellent" parcel size 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, "Acceptable" 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.

1.7.4 Quality assessment

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.

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.[24] 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.

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".[25] We can, therefore, randomly draw a set

---

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

25. 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/`.

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%.

Table 1.5 – Quality of the assignment of establishments to polygons.

Assignment quality	Parcel (PB)		Building (BB)	
	N	%	*N*	%
Excellent	8,782	78.83	22,978	96.43
Good	720	6.46	487	2.04
Acceptable	1,639	14.71	363	1.52
Total	11,141	100.0	23,828	100.0

Notes: Distributions of geocoded establishments in 2017 across quality categories. This classification includes the quality of the geocoding and the quality of the assignment process. Regarding the geocoding quality, all observations with a less than excellent geocoding are removed, and the remaining are used to construct the three groups: Excellent, good, and acceptable. The final sample that we use is that of excellent quality where missing values of covariates used in the regression analysis are removed, that is a sample of 8,708 parel size. See Appendix 1.7.3 for further details.

As explained in Appendix 1.7.3, 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 acceptable (see Appendix 1.7.3 for additional details).

1.7.5 Step-by-step explanation of the dataset construction

Figure 1.7 summarizes the steps undertaken for the construction of the dataset. The color *orange* refers to steps, the color *blue* refers to inputs, and the color green refers to outputs (which in some cases are also inputs for other steps).

Step *1. Appending scotts 2001-2019*. The data from the Scott's for the odd years form 2001 to 2019 are processed one by one. The variables names are harmonized. Missing primary NAICS codes are replaced by secondary NAICS codes. Then plants located outside of the Canada are removed. The final dataset contains "*scotts_global*", the establishments operating in Canada for each of the recorded years, with almost 95% of the universe of the manufacturing, but less for others sectors.

Step *2. Creating a unique addresses*. From the appended dataset, unique addresses are identified since many plants can share the same location. We create an 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 (*scotts_address*) with variables, the detail address and the address identifier.

Step *3. Geocoding unique addresses*. We use the dataset of unique addresses as input for the geocoding process describe in appendix . The output file contains *address_geocoded*, in addition to the inputs variables, the geographic cordinates of each addresses, 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 cordinates.

Step *4. Extracting polygons' 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 a building areas to addresses. Three different spatial join approaches are used to associate polygon areas to addresses. The output (*address_land*), contains for each address, the associated polygon area from each of the three spatial join approach, as well as the distance between the each associated polygon and the geographic coordinates of the address.

Step *5. Extracting location characteristics*. Using a Geographic Information System, the geocoded addresses are overlaid on shapefiles of dissemination area, Census Metropolitan Areas(CMA), zoning restriction, highways, seaport and airport to compute variation location variables : populations and surfaces of dissemination areas, and CMA, distances to sea, airport, highways, zoning categories, distance to the nearest city, number of plants in a given radius (5km, 10km) and their corresponding employment.

Step *6. 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 and quality variables are constructed.

Step *7. Extracting location characteristics*. Using a Geographic Information System, the geocoded addresses are overlaid on shapefiles of zoning restriction retrieve zoning.

Step *8. Creating the final dataset*. The appended Scott's dataset is merged with location characteristics, zoning restriction, and land measures to obtain the final dataset of land.

Figure 1.7 – Step-by-step explanation of the dataset construction

1.7.6 Representativeness of the final dataset

Table 1.6 – Distribution of plants across industries in the final dataset

	Parcel (PB)		Building (BB)		Scott's data	
	N	%	N	%	N	%
311 Food mfg	745	8.6	1,550	7.6	2,875	8.9
312 Beverage and tobacco product mfg	77	0.9	183	0.9	340	1.0
313 Textile mills	32	0.4	58	0.3	96	0.3
314 Textile product mills	227	2.6	496	2.4	743	2.3
315 Clothing mfg	307	3.5	476	2.3	712	2.2
316 Leather, allied product mfg	40	0.5	87	0.4	127	0.4
321 Wood product mfg	353	4.1	858	4.2	1,884	5.8
322 Paper mfg	149	1.7	319	1.6	501	1.5
323 Printing, support activities	726	8.3	1,633	8.0	2,270	7.0
324 Petrol, coal product mfg	20	0.2	48	0.2	134	0.4
325 Chemical mfg	433	5.0	953	4.7	1,539	4.7
326 Plastics, rubber products mfg	545	6.3	1,254	6.1	1,907	5.9
327 Non-metallic mineral product mfg	387	4.4	968	4.7	1,951	6.0
331 Primary metal mfg	125	1.4	344	1.7	537	1.7
332 Fabricated metal product mfg	1,301	14.9	3,521	17.2	5,226	16.1
333 Machinery mfg	1,061	12.2	2,963	14.5	4,542	14.0
334 Computer, electronic product mfg	300	3.4	718	3.5	1,032	3.2
335 Electrical, appliance mfg	256	2.9	543	2.7	784	2.4
336 Transportation equipment mfg	297	3.4	666	3.3	1,099	3.4
337 Furniture, related product mfg	421	4.8	903	4.4	1,392	4.3
339 Miscellaneous manufacturing	906	10.4	1,902	9.3	2,726	8.4
Total	8,708	100.0	20,443	100.0	32,417	100.0

Notes: This table reports the distributions of the Scott's database along with our final sample for the three measures in 2017 across the different industries at the NAICS 3-digit level.

Table 1.7 – Distribution of plants across provinces in the final dataset

	Parcel (PB)		Building (BB)		Scott's data	
	N	%	N	%	N	%
AB	0	0.0	1,696	8.3	2,735	8.4
BC	2,211	25.4	2,513	12.3	3,812	11.8
MB	419	4.8	505	2.5	983	3.0
NB	214	2.5	252	1.2	708	2.2
NL	0	0.0	99	0.5	272	0.8
NS	0	0.0	313	1.5	765	2.4
NT	0	0.0	1	0.0	6	0.0
NU	0	0.0	0	0.0	6	0.0
ON	1,881	21.6	9,549	46.7	13,735	42.4
PE	0	0.0	52	0.3	144	0.4
QC	3,709	42.6	5,076	24.8	8,430	26.0
SK	274	3.1	371	1.8	799	2.5
YT	0	0.0	16	0.1	22	0.1
Total	8,708	100.0	20,443	100.0	32,417	100.0

Notes: This table reports the distributions of the Scotts dataset along with our final samples for the three measures in 2017 across the Canadian provinces. The three Territories $NorthWestTerritories, Yukon, and Nunavut$ have been removed because they contain few observations.

Table 1.8 – Plant-level parcel size by industry

	Parcel (PB)			
	N	Mean	Median	CV
311 Food mfg	745	14,906.3	5,711.1	2.6
312 Beverage and tobacco product mfg	77	29,819.0	8,312.4	2.9
313 Textile mills	32	11,587.3	3,524.2	1.9
314 Textile product mills	227	7,509.2	4,045.0	1.4
315 Clothing mfg	307	6,649.3	3,853.6	1.3
316 Leather, allied product mfg	40	6,261.0	2,615.9	1.4
321 Wood product mfg	353	20,172.9	6,918.5	5.0
322 Paper mfg	149	23,797.7	13,359.4	1.9
323 Printing, support activities	726	7,996.7	3,486.8	2.1
324 Petrol, coal product mfg	20	22,928.1	9,315.0	1.5
325 Chemical mfg	433	18,558.6	8,213.5	2.7
326 Plastics, rubber products mfg	545	13,352.8	7,914.5	1.6
327 Non-metallic mineral product mfg	387	16,901.6	8,058.6	2.5
331 Primary metal mfg	125	39,747.3	6,253.9	5.3
332 Fabricated metal product mfg	1,301	11,494.9	5,565.2	3.1
333 Machinery mfg	1,061	12,131.6	6,599.6	2.1
334 Computer, electronic product mfg	300	10,580.8	6,467.8	1.5
335 Electrical, appliance mfg	256	13,318.3	8,416.8	1.4
336 Transportation equipment mfg	297	28,172.2	6,806.6	4.1
337 Furniture, related product mfg	421	9,238.9	4,546.8	2.9
339 Miscellaneous manufacturing	906	8,939.5	2,710.3	3.3
Total	8,708	13,354.2	5,757.6	2.7

Notes: This table reports descriptive statistics for land intensity four our two land measures across 3-digit industry as well as population density categories, the sample is our final dataset

Table 1.9 – Plant-level parcel size per worker by industry

	Parcel (PB)			
	N	Mean	Median	CV
311 Food mfg	745	1,120.9	172.0	7.2
312 Beverage and tobacco product mfg	77	3,598.2	208.4	3.2
313 Textile mills	32	2,602.3	313.3	4.0
314 Textile product mills	227	1,503.2	351.6	2.2
315 Clothing mfg	307	661.5	218.8	1.8
316 Leather, allied product mfg	40	1,109.2	232.7	2.1
321 Wood product mfg	353	1,297.0	387.9	2.4
322 Paper mfg	149	967.9	389.4	2.0
323 Printing, support activities	726	1,512.4	333.2	3.6
324 Petrol, coal product mfg	20	1,249.1	693.3	1.3
325 Chemical mfg	433	1,204.2	368.4	3.3
326 Plastics, rubber products mfg	545	990.3	297.2	4.6
327 Non-metallic mineral product mfg	387	1,951.0	390.4	8.4
331 Primary metal mfg	125	787.7	295.3	2.0
332 Fabricated metal product mfg	1,301	1,134.9	316.7	3.9
333 Machinery mfg	1,061	1,063.1	327.1	2.9
334 Computer, electronic product mfg	300	1,060.5	302.9	3.3
335 Electrical, appliance mfg	256	955.8	297.7	2.2
336 Transportation equipment mfg	297	2,145.5	300.1	8.8
337 Furniture, related product mfg	421	1,564.4	316.4	7.8
339 Miscellaneous manufacturing	906	1,399.3	352.1	3.4
Total	8,708	1,281.0	310.3	4.3

Notes: This table reports descriptive statistics for land intensity four our two land measures across 3-digit industry as well as population density categories, the sample is our final dataset

Table 1.10 – Building to parcel ratio by industry

	Building to parcel-ratio			
	N	Mean	Median	CV
311 Food mfg	572	0.37	0.36	14.29
312 Beverage and tobacco product mfg	50	0.29	0.27	7.93
313 Textile mills	20	0.47	0.47	2.20
314 Textile product mills	168	0.35	0.34	3.38
315 Clothing mfg	205	0.42	0.42	3.08
316 Leather, allied product mfg	33	0.37	0.37	2.45
321 Wood product mfg	302	0.26	0.22	10.55
322 Paper mfg	121	0.37	0.38	12.44
323 Printing, support activities	531	0.39	0.38	7.62
324 Petrol, coal product mfg	14	0.09	0.07	2.40
325 Chemical mfg	370	0.32	0.31	9.52
326 Plastics, rubber products mfg	455	0.35	0.36	10.79
327 Non-metallic mineral product mfg	326	0.27	0.25	10.95
331 Primary metal mfg	97	0.30	0.31	2.39
332 Fabricated metal product mfg	1,093	0.34	0.33	17.63
333 Machinery mfg	921	0.31	0.28	14.72
334 Computer, electronic product mfg	255	0.34	0.31	13.07
335 Electrical, appliance mfg	216	0.33	0.34	2.81
336 Transportation equipment mfg	241	0.31	0.29	5.39
337 Furniture, related product mfg	338	0.37	0.37	3.77
339 Miscellaneous manufacturing	653	0.37	0.34	5.92
Total	6,981	0.34	0.33	10.61

Notes: This table reports descriptive statistics for Floor-to-parcel-ratio and the Building-to-parcel-ratio across 3-digit industry. For each variable, the sample corresponds to the subset for which both the numerator and the denominator are non-missing at the same time. The sample includes not only plants in CMA/CA but also those outside the CMA/CA. In addition we've constrained the sample such that the parcel size > building footprint size

Table 1.11 – Testing for selection on observable plant characteristics

Dep. var.:	⊮ in sample			
	(1)	(2)	(3)	(4)
Ln Employment			-0.009	0.005
			(0.014)	(0.018)
Headquarter			0.046	-0.014
			(0.031)	(0.033)
Exporter			0.003	-0.011
			(0.053)	(0.041)
⊮ Residential zoning		-0.152^{a}	-0.156^{a}	-0.114^{a}
		(0.057)	(0.058)	(0.042
⊮ Recreational zoning		-0.605^{a}	-0.605^{a}	-0.329^{a}
		(0.091)	(0.092)	(0.079)
Ln City population				0.253^{a}
				(0.058)
Ln Population density 500m				0.085
				(0.053)
Ln Distance to closest major airport				0.025
				(0.076)
Ln Distance to closest major seaport				-0.043
				(0.093)
Ln Distance to closest freight station				0.071
				(0.058)
Ln Distance to closest highway junction				-0.035^{a}
				(0.021)
Fixed effects:				
4-digit industry	Yes	Yes	Yes	Yes
Province	No	Yes	Yes	Yes
Observations	24,457	24,457	24,457	24,457
Pseudo R^2	0.016	0.272	0.272	0.326

Notes: This table reports the estimates of a probit model where the dependent variable equals 1 if the establishment is in the estimation sample. ⊮ denotes $\{0,1\}$ dummy variables. Standard errors clustered at the city-level in parentheses ^{c} $p<0.10$, ^{b} $p<0.05$, ^{a} $p<0.01$

1.7.7 Identification of city centers

To locate the centers of cities in Canada, we use a two-step procedure. First, we use dissemination areas (DA)—i.e., 'census blocks' with geographic coordinates and population density—to identify densely populated areas. We fol-

low the procedure of Behrens et al. (2020), who suggest an algorithm to construct clusters of manufacturing plants based on their spatial concentration. We apply this procedure to the DAs in the Census Metropolitan Areas (CMA) or Census Agglomerations (CA) (henceforth we use CMAs to mean either Census Metropolitan Areas or Census Agglomerations). For each CMA, we separately construct its centers (there may be several of them) from clusters of dense DAs. More precisely, we define town centers as the geographic centers of the identified population density clusters. Formally, we identify the population clusters as follows:

— we flag all DAs with population density greater than the third quartile of the population density distribution of the CMA;

— we draw a circle with 500 meters radius around each flagged DA and compute the hypergeometric probability of having the number of flagged DAs in that circle, given the overall number of flagged DAs in the CMA. We also compare the total population of the flagged DAs within the circle to the total population of the flagged DAs in the CMA;

— A DA is considered a focal point of population concentration if the hypergeometric probability we computed is below 1% and if the ratio of the total population of the flagged DAs in the circle compared to the total population of the flagged DAs in the CMA is greater than the median observed in the CMA;

— we finally construct population clusters by drawing a buffer of 1 kilometer around each DA identified as a focal point and merging together all the overlapping buffers.

We pinpoint the centers of each disjoint population cluster and consider them as town centers. There are six centers in the Toronto CMA, three centers in the

Montreal CMA, and one in the Vancouver CMA.

Using these city centers, we compute two different distances between each plant and the city centers. The first distance is the distance to the nearest city center and the second distance (the weighted distance) is the average of the distance between the plant and the centers of its CMA weighted by the population in a 500 meters radius around the center.

1.7.8 Additional empirical results

Table 1.12 – Determinants of building footprint

	ln Building footprint					
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of the local environment						
Ln Population CMA	0.050^{b}	0.104^{a}	0.106^{a}	0.116^{a}	0.138^{a}	0.131^{a}
	(0.021)	(0.025)	(0.025)	(0.030)	(0.034)	
Ln CMA surface area	0.016	-0.037	-0.036	-0.053^{c}	-0.063^{c}	-0.036^{c}
	(0.027)	(0.029)	(0.029)	(0.031)	(0.032)	
Weighted Distance to city centers	-0.001	-0.004^{c}	-0.004^{b}	-0.003	-0.005^{b}	-0.023
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
$\mathbb{1}$ Residential	-0.227^{a}	-0.585^{a}	-0.578^{a}	-0.557^{a}	-0.500^{a}	-0.172^{a}
	(0.030)	(0.036)	(0.036)	(0.030)	(0.028)	
$\mathbb{1}$ Recreational	-0.210^{a}	-0.226^{a}	-0.224^{a}	-0.216^{a}	-0.234^{a}	-0.060^{a}
	(0.031)	(0.042)	(0.042)	(0.039)	(0.037)	
Ln Population density 500m					-0.041^{a}	
					(0.011)	
Characteristics of the establishment						
Ln Employment		-0.620^{a}	-0.624^{a}	-0.630^{a}	-0.634^{a}	-0.584^{a}
		(0.015)	(0.014)	(0.013)	(0.014)	
$\mathbb{1}$ Headquarter			-0.116^{a}	-0.117^{a}	-0.111^{a}	-0.024^{a}
			(0.019)	(0.020)	(0.020)	
$\mathbb{1}$ Exporter			0.090^{a}	0.086^{a}	0.082^{b}	0.029^{a}
			(0.034)	(0.032)	(0.032)	
# functions in the estab.			-0.020	-0.017	-0.016	-0.007
			(0.013)	(0.013)	(0.013)	
# 4-digit NAICS in the estab.			-0.004	-0.004	-0.004	-0.004
			(0.007)	(0.008)	(0.008)	
# products produced in the estab.			-0.001	-0.001	-0.002	-0.003
			(0.003)	(0.003)	(0.003)	
Distance to transport infrastructure						
Ln Distance to major airport				-0.021	0.018	-0.019
				(0.024)	(0.032)	
Ln Distance to major seaport				0.046^{b}	0.033^{c}	0.049^{b}
				(0.022)	(0.020)	
Ln Distance to freight station				-0.047^{a}	-0.045^{a}	-0.039^{a}
				(0.018)	(0.015)	
Ln Distance to junction				-0.078^{a}	-0.075^{a}	-0.065^{a}
				(0.013)	(0.013)	
Observations	20,375	20,375	20,375	20,375	20,375	20,377
R-squared	0.258	0.533	0.534	0.539	0.540	0.539
Industry (4-digit) fixed effects	yes	yes	yes	yes	yes	yes
Economic region fixed effects	yes	yes	yes	yes	yes	yes
Controls for neighbors	yes	yes	yes	yes	yes	yes

Notes: All regressions include a polynomial function of degree 4 in the number of neighbors of the establishment on its parcel. Only observations with the highest reliable information on parcel size are included. $\mathbb{1}$ denotes $\{0,1\}$ dummy variables. Standard errors clustered at the CMA/CA level in parentheses. ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1.

1.8 Additional theoretical results

Conditions (1.3)–(1.5) in the conceptual framework can be reorganized as follows:

$$\begin{aligned}
\left(\frac{p_z}{A_i \kappa_{i(s,z)}}\right)^{1-\sigma_s} &= p_i^{1-\sigma_s} A_i^{\frac{\sigma_s-1}{\sigma_s}} Y_i^{\frac{-(\sigma_s-1)}{\sigma_s}} \alpha_{i(s,z)}^{T} \left[\kappa_{i(s,z)} P_i\right]^{\frac{\sigma_s-1}{\sigma_s}} \\
\left(\frac{w_z}{A_i}\right)^{1-\sigma_s} &= p_i^{1-\sigma_s} A_i^{\frac{\sigma_s-1}{\sigma_s}} Y_i^{\frac{-(\sigma_s-1)}{\sigma_s}} \alpha_{i(s,z)}^{L} L_i^{\frac{\sigma_s-1}{\sigma_s}} \\
\left(\frac{r}{A_i}\right)^{1-\sigma_s} &= p_i^{1-\sigma_s} A_i^{\frac{\sigma_s-1}{\sigma_s}} Y_i^{\frac{-(\sigma_s-1)}{\sigma_s}} \left(1-\alpha_{i(s,z)}^{T}-\alpha_{i(s,z)}^{L}\right) K_i^{\frac{\sigma_s-1}{\sigma_s}}
\end{aligned}$$

If we assume that the output market is competitive (i.e., $p_i = p^s$ which is taken as given by firms), by summing these conditions and using the definition of the production function, we get

$$p^s = \frac{1}{A_i}\left[\left(\frac{p_z}{\kappa_{i(s,z)}}\right)^{1-\sigma_s} + w_z^{1-\sigma_s} + r^{1-\sigma_s}\right]^{\frac{1}{1-\sigma_s}}. \tag{1.10}$$

This equation simply states that unit costs equal the output price which ensures zero profits at equilibrium.

In the case of imperfect competition on the output market, firm i charges a price-cost margin μ_i. Hence, equation (1.10) becomes

$$p_i = \frac{\mu_i}{A_i}\left[\left(\frac{p_z}{\kappa_{i(s,z)}}\right)^{1-\sigma_s} + w_z^{1-\sigma_s} + r^{1-\sigma_s}\right]^{\frac{1}{1-\sigma_s}}. \tag{1.11}$$

Most models consider that $\mu_i = \mu(A_i)$ is an increasing function of A_i, i.e., more productive firms (in terms of TFP) charge higher markups. As long as firm-level prices remain a decreasing function of productivity, i.e. the elasticity of markups to productivity is smaller than 1 so that there are pro-competitive effects, equations (1.10) and (1.11) allow us to make predictions that are qualitatively similar on how firms select into zones based on TFP.

When $p_i = p^s$, i.e., there is an industry-specific price common to all firms, equation (1.10) shows that a large A_i allows a firm to absorb higher production costs w_z and p_z (we assume that r does not vary between firms). Put differently, high-productivity firms are likely to sort into places where wages and/or land prices are high since only they can bear these higher production costs.

1.8.1 Fixed costs and adjustment costs

So far, we have assumed that land is, as labor and capital, a variable cost that firms can freely adjust to changes in the relative price of production factors (depending, of course, on their elasticity of substitution). This assumption might be violated in practice which has implications for the interpretation of our results.

First, if land is a variable cost only, with the CES production function we have assumed, the quantity of land per worker is independent of the size of the firm in terms of employees. However, part of the land used by firms has the nature of a fixed cost: corridors, bathrooms, office spaces, or production spaces have a size that is partly independent of the number of workers using them (there needs to be a bathroom whether there is one or five employees in the establishment for example). This implies that the amount of land per worker should be a decreasing function of the firm-level number of employees. Since it is well known that high-productivity/high-employment firms sort into bigger cities where the relative price of land is higher, not accounting for the fixed cost dimension of land will bias the coefficient on city size downward. On the other hand, if parcel size and building footprints are smaller in denser parts of urban areas, only smaller establishments in terms of employees will be able to locate there so that the coefficient on the population density in a 500m radius around the establishment will be upward biased. This is why we also introduce among the regressors the size of the firm in terms of employees in our benchmark regressions.

Moreover, if a firm can (quite) freely adjust its workforce when the relative cost of labor varies, the same does not apply to land: moving, opening, or closing a facility is costly. Hence, the location of a firm and the amount of space it uses is

hard to adjust, at least in the short- or medium-run. Only when shocks are large and permanent enough do firms adjust their land consumption, most likely by moving or by opening and closing some establishments. This means that when firms grow or shrink, they first do so by adjusting their number of employees only, especially if they face transitory shocks. This in turn affects the amount of land per worker they use. One way to detect the presence of these adjustment costs is to control for past employment growth of the firm. We conjecture that firms that grew more in the past will have a lower land-to-labor ratio, consistent with the existence of fixed costs and frictions that prohibit a quick adjustment of that ratio. We will do that in a separate section together with an analysis of the frequency and determinants of establishments' relocation decisions.

1.8.2 Evidence of adjustment costs

So far, we have considered that establishments can freely adjust the amount of land they consume. Obviously, this is unlikely to be the case in reality. To adjust the amount of land they use, plants have two options. First, on the 'intensive margin', they can stay on their current parcel and build additional buildings or expand to adjacent buildings or parcels. However, due to constraints on the amount of available space on the parcel they occupy and on the availability of adjacent parcels, this option seems often very limited. Second, on the 'extensive margin', firms can adjust the amount of space they consume by moving. However, moving is costly so that firms will do so only when there are important and permanent changes in their level of activity.

In our dataset, we consider the establishments that are present in both 2013 and 2017, and we define as 'movers' those for which: (i) we have information on their building footprint in both years; and (ii) whose building footprint has

changed over the period. Between 2013 and 2017, only 6.9% of the establishments have changed location (1,718 establishments out of 24,861), and the very vast majority of these relocations occur within the same urban area (1,650 out of 1,718). Relocations of establishments are thus rare and mostly local, which is in line with what Bergeaud and Ray (2021) find using French data. Among these relocations, 54% correspond to establishments moving to larger buildings (934 out of 1,718). Probit regressions reveal that initially larger establishments are less likely to move, which means that moving costs are more important when establishments employ many workers.[26] Several explanations can rationalize this finding: more people need to be convinced to change their workplace; the monetary costs related to the moving of the equipment are higher for larger establishments; there is more flexibility to expand space on site for bigger establishments that occupy larger parcels than for smaller ones; or there might be fewer production sites that are suitable for hosting large establishments. Also, quite intuitively, faster employment growth between 2013 and 2017 is positively correlated with the probability to move to a larger building, and negatively correlated with the probability to move to a smaller building, conditional on initial size.

As explained above, establishment relocations are rare and do mainly occur in the wake of large permanent shocks. Yet, firms often face smaller transitory shocks. In that case, firms will adjust (up to a certain point) their workforce without adjusting the amount of land they consume. These adjustments should translate into a negative correlation between land consumption per worker and the establishment's past employment growth, conditional on current employment. This is exactly what we find, as Table 1.13 shows. However, as Figure 1.8 shows, there is substantial heterogeneity in the elasticity of establishments'

26. These results are available upon request.

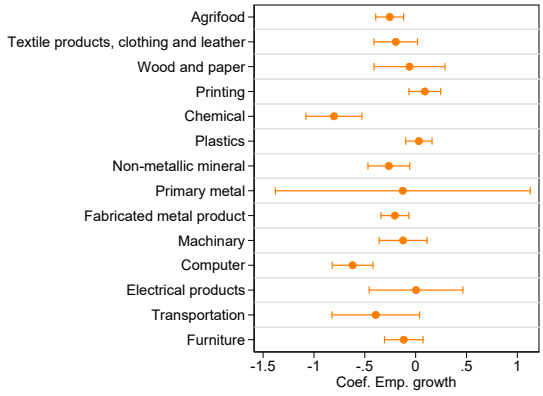
land consumption per worker to past employment growth. The coefficient is equal to -0.21 on average, with standard errors of 0.24. It is significant for half of the sectors only. While this might be due to a lack of precision of the estimates for some sectors, it also likely reflects heterogeneity in adjustment costs across sectors.

Table 1.13 – Past employment growth and adjustment costs

VARIABLES	Parcel size
Ln City population density	-0.008
	(0.066)
Ln City surface area	0.020
	(0.037)
Ln Population density 500m	-0.165[a]
	(0.021)
Ln Employment	-0.627[a]
	(0.014)
Δ Ln Employment 2013–2017	-0.153[a]
	(0.035)
Observations	7,314
R^2	0.578
Industry (4-digit) fixed effects	yes
Economic region fixed effects	yes

Notes: Δ Ln Employment is the establishment's employment growth. All other variables are measured in 2017. All regressions include the same controls as in our baseline regressions in Tables 9 and 10. Only observations with the highest reliable information on parcel size are included. Standard errors clustered at the CMA/CA level in parentheses. [a] p<0.01, [b] p<0.05, [c] p<0.1s

Figure 1.8 – Adjustment costs by sector



Note: The graphs display the point estimate and the 10% confidence interval of the coefficient associated to establishment past employment growth. Regressions are run separately for the various sectors using the benchmark specification in column (5) of Tables 1.1 augmented with past employment growth.

CHAPTER II

THE CAUSES OF THE AGGLOMERATION OF INNOVATION: EVIDENCE FROM COAGGLOMERATION PATTERNS

Abstract

We use the continuous measure of Duranton and Overman (2005) to describe the coagglomeration of innovation in Canada. The observed patterns reveal that Canadian innovation is concentrated and even more than production. Then, we analyze the effects of labor pooling, input sharing, and knowledge spillover on the coagglomeration of innovation. The analysis shows that on top of the coagglomeration of the production, only the knowledge spillover unambiguously causes the coagglomeration of innovation.

Keywords: (Co)agglomeration; innovation; input sharing; knowledge spillover; labor pooling; manufacturing; patents.
JEL Classification: R23; R32; O33; L14; L60.

2.1 Introduction

Innovation has a high tendency to be concentrated. Its central role in fostering long term growth, along with the advantages stemming from agglomeration economies make it important to understand how and why innovation agglomerates. [1] Answering this question is critical for both policy-makers and researchers. Indeed, the need of promoting innovation has retained the attention of public authorities worldwide these last years. For example, since 2010, the Commission of the European Union has put in place the Innovation Union to boost innovation in Europe. [2] In Canada in particular, the Government has elaborated an *"Innovation and Skill plan"* to accelerate innovation through super-clusters. [3] From an academic perspective, there is a mature literature on the microfoundations of agglomeration economies (Duranton and Puga, 2004). This literature well establishes that agglomeration externalities are the results of some mechanisms, such as the three Marshallian forces of input sharing, labor pooling, and knowledge spillover (Marshall, 1890). Empirically, some evidence supports the presence of these Marshallian forces in generating

1. On the role of innovation for long term growth see Krugman 1991; Segerstrom 1991; Aghion and Howitt 2005; Hall 2011; Aghion et al. 2014. In facts more than 50% of the USA growth since the WWII is attributable to innovation; Between 2000 to 2007 two-thirds of UK private-sector productivity resulted from innovation. Moreover, the private rate of return from innovation is estimated at 25 to 30 percent, and the social returns from innovation are typically 2 to 3 times larger than the private returns. See "National Innovation policies: What countries do best and how they can improve" from the Global Trade and Innovation Policy Alliance

2. The objective of this Innovation Union is to: (i) improve conditions and access to finance for research and innovation in the EU, (ii) create a genuine single European market for innovation, (iii) stimulate private sector investment and processes, to increase European venture capital investments

3. For more details see https://www.ic.gc.ca/eic/site/093.nsf/eng/home

the agglomeration of production (Rosenthal and Strange, 2001, 2004; Ellison et al., 2010; Faggio et al., 2017, 2020). However, there is no guarantee that this evidence of the role of the Marshallian forces on the agglomeration of production, translate into the agglomeration of innovation in the same way, for at least two reasons. First, innovation happens to be more concentrated than production (Feldman and Kogler, 2010), and second, firms do not locate production and Research and Development (R&D) at the same place (Kelly and Hageman, 1999; Duranton and Puga, 2001).

The few studies that exist on the determinants of the agglomeration of innovation focus on the effects of industrial characteristics, but hardly any study has looked at the effects of the micro-founded mechanisms that may explain the agglomeration of innovation. Some exception include Helsley and Strange (2002) who discuss the link between innovation and input sharing, Audretsch and Feldman (1996, 2004) who summarize the role of knowledge spillover for the agglomeration of innovation; and Gerlach et al. (2009); Simonen and McCann (2008) for the role of labor pooling on the agglomeration of innovation. Yet, there is still room for discussion on how the Marshallian mechanisms empirically operate on the agglomeration of innovation, and this work intends to contribute to this debate with a unified framework for all the three Marshallian forces similar to what has been done on the production side by Ellison et al. (2010).

This paper uses the coagglomeration of industry pairs to document new facts on the geographic concentration of innovation and tests the effects of the three Marshallian forces on this phenomenon. The motivation of the use of the colocation of industry pairs is twofold. First, diversity matters for innovation (Carlino and Kerr, 2015), and the colocation of industry pairs of innovation may be well appropriated to analyze the geographic concentration of innovation. Sec-

ond, previous works suggest that the variation in characteristics of industries that colocate better helps in understanding the microfoundations of agglomeration economies (Ellison et al., 2010; Faggio et al., 2017).

We use the Canadian patents applications for this empirical analysis and address two major challenges during the process. First, locating patents across space is not a straightforward task. Second, patents are categorized in the dataset at hand with the technological classification (IPC) whereas we need the North American Industrial Classification System (NAICS) for our analysis. We deal with the first challenge by assigning to each patent the geographic locations of its related innovators, and we address the second challenge by using a mapping matrix that relates the IPC to the NAICS. We consider many alternative ways of locating patents across space and in the NAICS classification for robustness checks. Then, we use the methodology of Duranton and Overman (2005) to provide a novel picture of the geographic concentration of the innovation in Canada and compare it to that of production. This primary exercise reveals that at the NAICS 3-digit, 65% of industry pairs in terms of innovation are colocated against 58% for production. Moreover, this colocation of industry pairs of innovation happens essentially at shorter distances.

Next, we estimate the causal effect of the three Marshallian externalities on the coagglomeration of innovation. More precisely, we regress the measure of coagglomeration of innovation at the NAICS 3-digit industry level on measures of input sharing, labor pooling, and knowledge spillover controlling for the coagglomeration of production. Our empirical strategy relies, for the three Marshallian forces, on measures constructed with USA data as proxies for the Canadian ones to deal with potential endogeneity, as well as the use of an instrumental variable for the input sharing measure. The results show that the Marshallian forces do not act on the formation of the coagglomeration of innovation in the

same way they do for that of production. Indeed, while previous studies have found that the three Marshallian forces positively cause the coagglomeration of the production (Ellison et al., 2010; Faggio et al., 2017), our results first suggest that a non-negligible part of the coagglomeration of innovation is the mere result of the coagglomeration of production. Second, once the coagglomeration of production is controlled for, only the knowledge spillover unambiguously determines the coagglomeration of innovation, and the effects of the two other forces are not significant. In terms of magnitudes, a one standard deviation of the coagglomeration of production is translated into a 0.36 standard deviation of the coagglomeration of innovation, and a one standard deviation of knowledge spillover results in a 0.26 standard deviation of the coagglomeration of the innovation. These results qualitatively survive a large variety of robustness checks.

This work is related to the large strand of literature on agglomeration economies and their determinants. First, the paper adds to previous studies which have provided valuable and useful insights into the agglomeration economies by describing the location and the colocation patterns of the production. Some examples include the description of the spatial patterns of the manufacturing industries in the USA (Ellison and Glaeser, 1997), the exploration of the location patterns of manufacturing industries in the UK (Duranton and Overman, 2005), the detailed analysis of the location patterns of the manufacturing in Canada (Behrens and Bougna, 2015), the agglomeration and coagglomeration of manufacturing in Russia (Aleksandrova et al., 2020), and the coagglomeration patterns of the manufacturing in Vietnam (Howard et al., 2016). On the innovation side, some studies have portraited the location pattern of knowledge spillover in the US (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014; Ganguli et al., 2020). This paper contributes to this literature by providing new

patterns of the coagglomeration of innovation in Canada.

Second, and more importantly, the paper builds on the theoretical and empirical literature of the sources of agglomeration. Theory on the microfoundation of agglomeration economies highlights the role of the three Marshallian forces in generating agglomeration. The possibility of sharing inputs and production facilities, the efficiency of the firm-worker matching, and the ease to create, accumulate and diffuse knowledge in a cluster are incentives for the agglomeration of economic activities (Duranton and Puga, 2004). Moreover, and directly related to the agglomeration of innovation, Helsley and Strange (2002) use a model to suggest that innovation may have an incentive to cluster since a dense network of input suppliers may reduce the cost of bringing new products to the market. Kelly and Hageman (1999) also use a model to show that knowledge spillover plays a key role in the agglomeration of innovation. On the empirical side, many studies have tested these agglomeration forces along with natural advantages and provided convincing evidence on their role in generating the agglomeration of the production (Ellison and Glaeser, 1999; Audretsch and Feldman, 1996; Rosenthal and Strange, 2001; Ellison et al., 2010). This paper contributes to this literature by looking at the effects of these forces on the coagglomeration of innovation at the industry level, where literature has focused essentially on the coagglomeration of the production. We use continuous measures that allow for more flexibility in detecting coagglomeration at various scales and provide new evidence on the role of Marshallian forces on the coagglomeration of innovation.

The rest of the document is organized as follows. Section 2.2 focuses on the pattern of the coagglomeration of innovation in Canada. It starts by presenting the innovation data, the measure used, and some key features of the coagglomeration of innovation in Canada. Section 2.3 discusses the rationale of the link

between the three Marshallian forces and the coagglomeration of innovation, along with the data and measures used to capture these determinants. Finally, it presents the empirical strategy and the results of the analysis. The document ends with some concluding remarks in section 2.5.

2.2 The coagglomeration of innovation in Canada

In this section, we present the dataset, and the methodology used to construct the coagglomeration measures. Then, we provide the patterns of the coagglomeration of innovation in Canada.

2.2.1 Innovation data

The ideal setting for measuring innovation for the purpose of our analysis should be an accurate location of the position where innovation happens. For example the actual location of the plant where the innovation is produced. Unfortunately, we do not have information on the place of work of innovators related to each patent. However, to analyze the geographic concentration of the innovation in Canada, we use the patent database digitized by the CD Howe institute.[4] This fine-grained level database, records the patent applications of innovation in Canada, with information on the postal code of the residences of all the inventors related to each patent (up to 10 inventors per patent), the International Patent Classification (IPC) code of the patent, the year of applica-

4. The C.D. Howe Institute is a registered charity and an independent not-for-profit research institute whose mission is to raise living standards by fostering economically sound public policies. see https://www.cdhowe.org

tion as well as the latest administrative status of the application. [5] The database also contains for each patent, a probability that links the IPC code to the North American Industry Classification System (NAICS) [6]. The concordance is based on the first 3 digits of the IPC codes and the first 3 digits of the NAICS codes for the manufacturing industries in particular. [7] More specifically, each patent has a probability to be related to each of the 21 manufacturing 3-digit-NAICS codes.

From that database, patent applications from Canadian resident inventors are extracted for the period spanning from 2001 to 2015 (odd years). This represents a total of 36,137 patents in the IPC classification. For the analysis, the patents need to be located across the space and related to the NAICS instead of the IPC. To have such a dataset, a first transformation is realized by assigning to the address of each inventor, the patent to which he or she is related. The average number of innovators related to each patent is 2.04. [8] This transformation allows to geographically locate each patent across the space. The resulting dataset contains 65,383 patents across the space. Then in a second transfor-

---

5. All the patent applications are considered irrespective of their descriptive status code (approved, granted, etc.) given that the mere fact of applying for a patent is enough to reflect an innovative activity

6. This probability link captures the industries that use or implement technologies rather than industries that perform research and development activities

7. While applying to protect an idea through a patent, some applications record the NAICS of the industry that manufactured the innovation to be protected and the NAICS of the industry of use. Using these cases where a clear correspondence is available, a probabilistic matrix is then built to match the IPC digit codes to the NAICS code which is the result of a combination of industry of use and the industry that manufacturer the patent. The NAICS code is therefore referred to as "NAICS of link". For more details on the methodology used to construct such concordance, see Evenson et al. (1991); Kortum and Putnam (1997)

8. see Figure 2.6 for the distribution of patents across the number of co-inventors.

mation, the probability that links IPC to NAICS is used to move from the IPC classification dataset of patents to the NAICS classification of patents. More precisely, each line of the dataset of the located patents is duplicated as many times as there are non-zero probabilities that link a patent of a specific IPC to the NAICS-3digits. Then, to each duplicated line, a single NAICS code is assigned, as well as the value of the corresponding probability that links the IPC code to the corresponding NAICS. The average number of NAICS with non-zero probability per patent is 2.96.[9] The final dataset after this second transformation contains 252,573 observations which are patents extended to NAICS and the addresses of all their related inventors.[10] Once these transformations are made, two concerns have to be dealt with in the resulting dataset. First, by locating each patent at all the residences of its related inventors, the number of patents is artificially inflated. Second, by assigning each patent to all the NAICS to which it is related, the colocation of industry pairs is also artificially augmented. To address these issues, alternative samples of patents, as well as weights are used in the analysis to alleviate the effects of the abovementioned concerns. In particular, for the first concern, the inverse of the number of inventors related to each patent is used to correct for duplicating each patent across the addresses of its related inventors. That is to consider only an equal fraction of the patent at each inventor address. For the second concern, the probability that links IPC to NAICS is used to correct for duplicating the patents across its

9. see Figure 2.7 for the distribution of patents across the number NAICS with non zero probability.

10. When all the probability values of a given patent are zeros, this indicates that the patent is not related to manufacturing at all, and the corresponding patent will be removed from the dataset. Such cases represent less than 1% of all the NAICS 3-digit innovators recorded. This value is not huge so that the restricted dataset still reflects the overall patenting activity in Canada

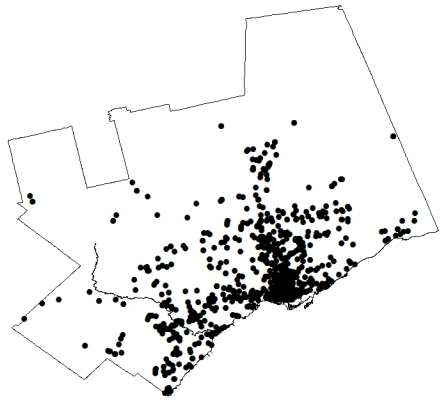
related NAICS, while moving from the IPC to the NAICS classification. That is to assign only a fraction of the patent to each of its related NAICS. In terms of alternative samples, the first one is created as follows: to move from the IPC to NAICS classification, instead of duplicating the patents as many times as there are non-zero probabilities that relate an IPC to a NAICS, the patent is assigned only to the NAICS with the maximum probability. This first, alternative dataset contains 71,760 patents. For the second alternative dataset, to move from the IPC to NAICS classification, instead of duplicating the patents as many times as there are non-zero probabilities that relate an IPC to a NAICS, the patent is assigned only to the NAICS with the maximum probability provided that the probability is greater than 0.5. This second, alternative dataset contains 54,000. Table 2.1 gives for each manufacturing 3-digit NAICS code, the mean annual flow of patents applications over the period 2001-2015 for the main, and the two alternative samples, and more details on these samples are presented in section 2.6.1 in the appendix.

A step further, the Postal Code Conversion File (PCCF) from Statistics Canada is used to geocode the resulting dataset.[11] The process consists in assigning geographic coordinates (longitudes and latitudes) to each located patent by merging the patents dataset with the PCCF. These geographic coordinates are key variables to the computation of the coagglomeration measures to be used in the analysis.

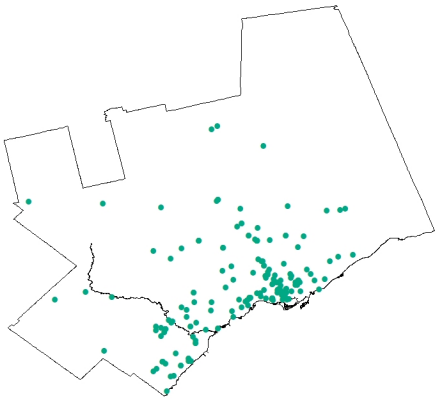
The final dataset is a table with all the patent applications in Canada from 2001 to 2015 (8 odd years), with 252,573 observations. The file is suitable for capturing the innovative activity for which explicit protection has been asked. As

11. The Postal Code Conversion File (PCCF) is a concordance dataset built and maintained by Statistics Canada. This dataset allows relating each postal code in Canada to its geographic cordinates

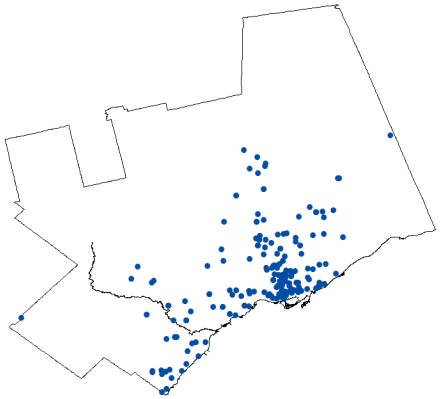
such, all the innovation for which there is not an effort of protection through a patent is not captured in this analysis. However, the dataset remains a good sample to describe the patterns of coagglomeration and discuss its causes.



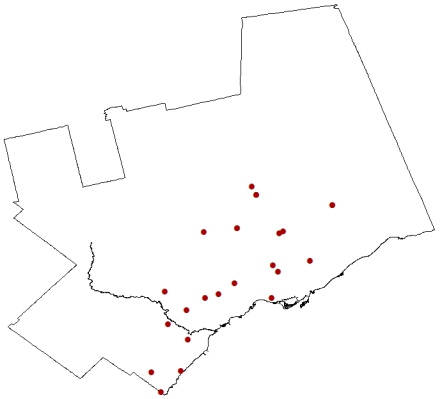
(a) 334 computer and electronic



(b) 312 Beverage and tobacco



(c) 315 Clothing



(d) 316 Leather

Figure 2.1 – Maps of four illustrative industries

Notes: The median bilateral distance between inventors of the same patent is 14km and the 3rd quartile is 35km indicating at a first glance a somehow spatial concentration of co-inventors.

Table 2.1 – Mean annual flow of patents applications

NAICS - 3 digits	Patent extended to innovators' addresses and all NAICS with prob.link $\neq 0$ (Sample 1 : Benchmark)		Patent extended to innovators' addresses and only the NAICS with max prob.link (Sample 2)		Patent extended to innovators' addresses and only one NAICS with prob.link>0.5 (Sample 3)	
	counts	%	counts	%	counts	%
311 Food manufacturing	1,971	6,2%	629	7.0%	66	1.0%
312 Beverage and tobacco product manuf	1,194	3,8%	27	0.3%	27	0.4%
313 Textile mills	56	0,2%	22	0.2%	22	0.3%
314 Textile product mills	2	0,0%	-	0.0%	-	0.0%
315 Clothing manufacturing	811	2,6%	47	0.5%	47	0.7%
316 Leather, allied product manuf	276	0,9%	80	0.9%	54	0.8%
321 Wood product manufacturing	93	0,3%	77	0.9%	77	1.1%
322 Paper manufacturing	776	2,5%	309	3.4%	61	0,9%
323 Printing, support activities	16	0.1%	-	0.0%	-	0.0%
324 Petrol, coal product manuf	816	2.6%	88	1.0%	88	1.3%
325 Chemical manufacturing	2,860	9.1%	664	7.4%	242	3,6%
326 Plastics, rubber products manuf	2,685	8.5%	224	2.5%	84	1.2%
327 Non-metallic mineral product manuf	1,352	4.3%	336	3.7%	321	4.8%
331 Primary metal manufacturing	518	1.6%	99	1.1%	99	1.5%
332 Fabricated metal product manuf	2,721	8.6%	245	2.7%	141	2.1%
333 Machinery manufacturing	4,466	14.1%	1,466	16.3%	899	13.3%
334 Computer, electronic product manuf	5,587	17.7%	2,893	32.3%	2 883	42.7%
335 Electrical, appliance manuf	1,501	4.8%	174	1.9%	131	1,9%
336 Transportation equipment manuf	1,122	3.6%	400	4.5%	400	5.9%
337 Furniture, related product manuf	524	1.7%	186	2.1%	186	2.8%
339 Miscellaneous manufacturing	2,226	7.1%	1,004	11.2%	924	13.7%
Total	**31,572**	**100%**	**8,970**	**100%**	**6,750**	**100%**

Notes: The original dataset in the International Patent Classification (IPC) contains 36,137 (Mean annual flow : 4,517) patents applications over the period 2001-2015 (odd years). When extended to NAICS and all the inventors' addresses, the number of patents amounts to 252,576 (Sample 1: Benchmark with a mean annual flow of 31,572). When extended to all the inventors' addresses and the NAICS with maximum probability, the number of patents amounts to 71,760 patents (Sample 2 with a mean annual flow of 8,970). When extended to all the inventors' addresses and the NAICS with maximum probability provided that the probability >0.5, the number of patents amounts to 54,000 patents (Sample 3 with a mean annual flow of 6,750).

Table 2.1 shows that less than 10% of the NAICS 3-digit manufacturing industries represent more than 33.3% of the number of patents. Indeed, the 2 industries "334 computer, electronic product" and "333 Machinery manufacturing" dominate in terms of patenting activity and represent together 32% of the total number of patents in the benchmark sample, 48% of the total number of

patents in sample 2, and 56% of the patents in sample 3. The least innovative sectors are "313 Textile mills", "314 Textile product mills", "315 Clothing", "316 Leather, allied product", "321 Wood product manufacturing", "323 Printing, support activities" and "331 Primary metal manufacturing". These seven industries altogether represent 30% of the number of 3-digits manufacturing industries but account for less than 6% of the total number of patents. For the rest of the analysis, the industries "314 Textile product mills" and "323 Printing, support activities" are removed given that they can induce noisy estimates particularly for the estimation of the coagglomeration of innovation. Thus, out of the 21 industries registered in the Canadian classification at 3-digit, 19 are left and represent 171 symmetric industry pairs and 1368 observations across the 8 odd years from 2001 to 2015.

2.2.2 Measuring the coagglomeration of innovation

The literature dedicated to assessing the extent of agglomeration usually uses two different and complementary approaches. The first approach relies on metrics that capture the degree of localization of entities from the same group, and the second approach uses metrics that measure the degree of co-location of entities from two different groups. While each of these approaches has its specificities, the advantage of the colocation approach over the location approach is the possibility to better understand the effects of the interrelatedness of industry pairs. In this analysis, the second approach is used to assess the extent of the agglomeration of innovation by the colocation patterns of patents from two different industries. More precisely, the K-density of the bilateral distances of patents from two different industries is used to measure the coagglomeration of industry pairs. This metric suggested by Duranton and Overman (2005) also has the advantage of dealing with the main issues related to studies on spatial

concentration. [12]

To put things in perspective, let consider two industries I and J, with N_I and N_J patents respectively. The observed coagglomeration of the industry pair I and J at a given distance d is estimated by :

$$\hat{K}_{IJ}(d) = \frac{1}{h \sum_{i=1}^{N_I} \sum_{j=1}^{N_J} (w(I) + w(J))} \sum_{i=1}^{N_I} \sum_{j=1}^{N_J} (w(I) + w(J)) f(\frac{d - d_{ij}}{h}) \tag{2.1}$$

Where d_{ij} is the euclidean distance between the location of the patent i and j from the industries I and J respectively, f is the kernel function with bandwith h, $w(I)$ and $w(J)$ are the weights for industry I and J respectively, used to account for the transformations made to move from the IPC to the NAICS classification. These weights are the probabilities that link IPC to NAICS of the industry I and J respectively.

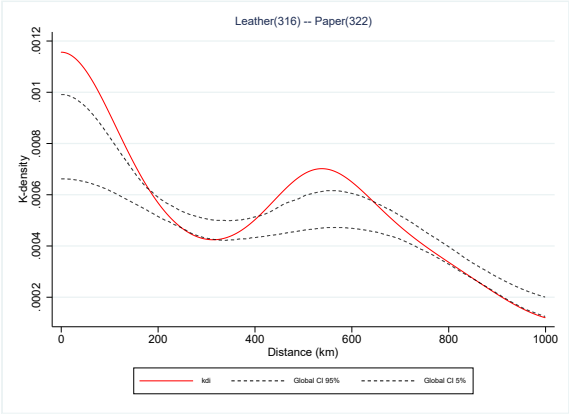
With this metric, the excess of coagglomeration for a given industry pair (I,J) is detected by constructing a counterfactual against which the observed coagglomeration is compared. More in detail, for the industry pair (I,J), a counterfactual coagglomeration is constructed using for each industry of the pair, a hypothetical distribution of locations for its patents. This hypothetical distribution of locations is obtained through a random draw of locations across the union of possible locations of the two industries. More precisely, imagine that $(i^1, i^2, i^3, ..., i^{N_I})$ are the N_I locations of patents in industry I and $(j^1, j^2, j^3, ..., j^{N_J})$ the N_J locations of patents in industry J. Then $(i^1, i^2, i^3, ..., i^{N_I}, j^1, j^2, j^3, ..., j^{N_J})$ will be the universe of the $N_I + N_J$ locations from which the random draws of

12. Duranton and Overman (2005) observed that their metric, over the previous ones, has the advantage to be : (i) comparable across industries, (ii) to control for the overall agglomeration of manufacturing, (iii) to control for industrial concentration, (iv) to be unbiased with respect to scales and borders, and (v) to be statistically testable.

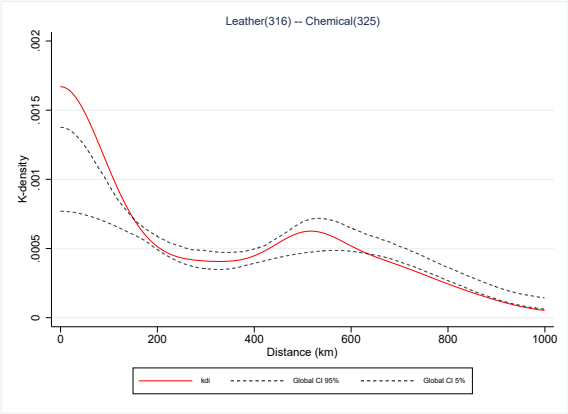
N_I hypothetical locations for the patents of the industry I and N_J hypothetical locations for patents of the industry J will be chosen. With these hypothetical distributions of locations, the counterfactual coagglomeration measure of the industry pair (I,J) is then constructed. The exercise consists in repeating 1000 times the draws of hypothetical distributions for I and J respectively and to estimate for each pair of these distributions, the K-density using the formula 2.1. Then global and local bandwidths are constructed along with the index of colocalization and that of co-dispersion. The details for the construction of these bandwidths and the computation of these indices are provided in appendix 2.6.2.

2.2.3 Key features of the coagglomeration of innovation in Canada

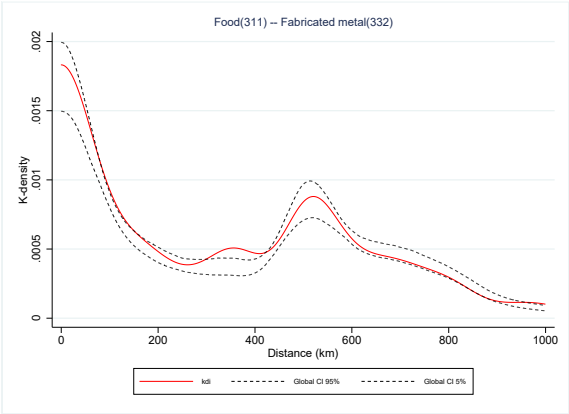
The above-mentioned measure is used, to estimate the coagglomeration for the 171 symmetrical industry pairs obtained by the combination of the 19 industries of Canadian classification at 3-digit NAICS code. In line with Duranton and Overman (2005), globally colocalized industry pairs can be graphically identified as those for which the observed K-density (solid red line) lies above the upper bound of the global confidence interval (upper dashed line) for at least one distance. Equivalently, globally codispersed industry pairs are those for which the observed K-density lies below the lower bound of the global confidence interval (lower dashed line) for at least one distance and is not colocalized at any distance. The graphs on figure 2.2 sum up key specificities observed in the coagglomeration patterns of innovation in Canada.



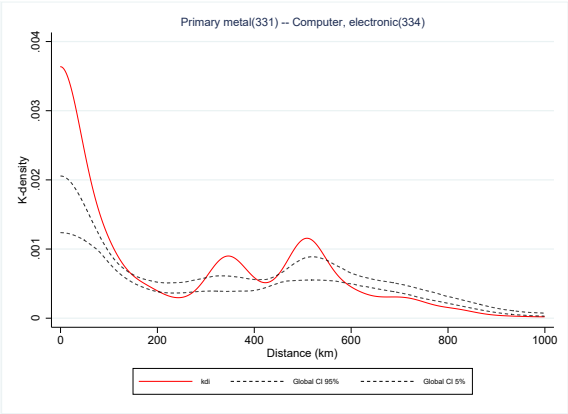
(a) Localized at short and long distances



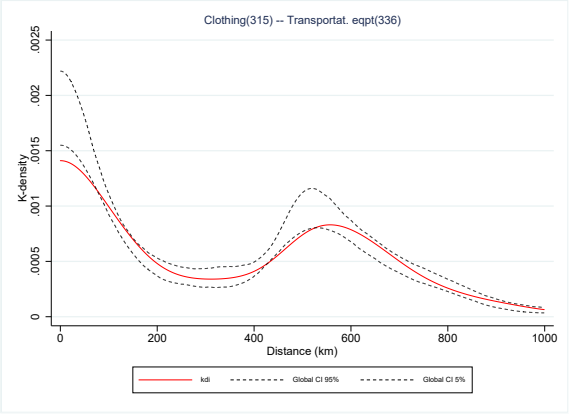
(b) localized at short distances



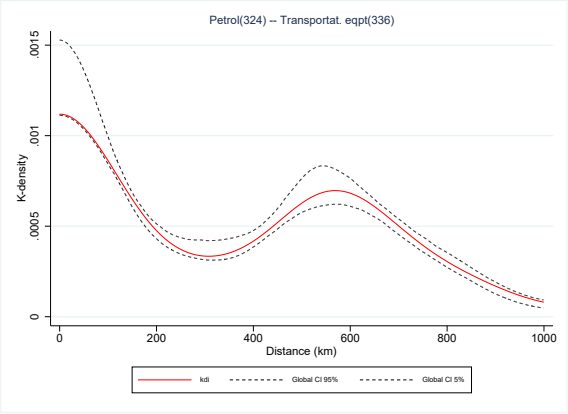
(c) localized at long distances



(d) localized at almost all distances



(e) Dispersed



(f) Random

Figure 2.2 – K-density of six illustrative coagglomeration patterns in 2015

The first four panels are cases of globally colocalized industry pairs. Panel (a) depicts industry pairs for which there is an overrepresentation of shorter bilateral distances as well as longer distances equivalent to the distances between major cities in Canada. This reflects the cases where a given industry pair is colocated in many cities to generate both within-cities coagglomeration and between-cities coagglomeration. The localization of innovative industries generating such patterns includes industry pairs that are simultaneously located in major cities. Some examples include the industry-pairs *"316 Leather – 322 Paper"*; *"335 Electrical appliances – 322 Paper"*, *"316 Leather – 335 Electric appliance"*. Panel (b) is representative of features for which the industry pair is present only in one major city. For example, industry pairs containing the *"325 Chemical"* or *"321 Wood"* industry that happen to be localized in very few locations across the country. Panel (c) depicts cases where the industries of the pair never co-locate in the same city so that co-localization only appears across cities. For example the innovation in the *"311 food"* industry is rarely localized in the same place with *"332 Fabricated metal"*. Finally, panel (d) represents special cases of patents in "cross-sectors" industries such as *"334 computer and electronic"*. Since this industry is likely to be connected to many other industries, co-location appears here almost at each distance with some peaks at the scales of the distances between major cities. This feature is also the case for patents in the industry pair of *"331 Primary metal – 335 Electrical appliances"* as well as the industry-pair *"316 leather – 331 Primary metal"*.

Panel (e) is a case of globally dispersed industry pairs. These are industry pairs that are scattered compared to a random distribution. For example *"311 Food – 333 Machinery"*, *"312 Beverage – 324 petrol"* as well as *"312 Beverage – 337 Furniture"*.

Finally, panel (f) depicts industry pairs for which innovation is neither coag-

glomerated nor codispersed, the co-location pattern, in this case, is not significantly different from one that would be obtained by a random co-location of innovation. Some examples are *"311 Food – 313 Textile"*, and *"311 Food – 322 Paper"*.

The scale of the coagglomeration patterns of the Canadian innovation

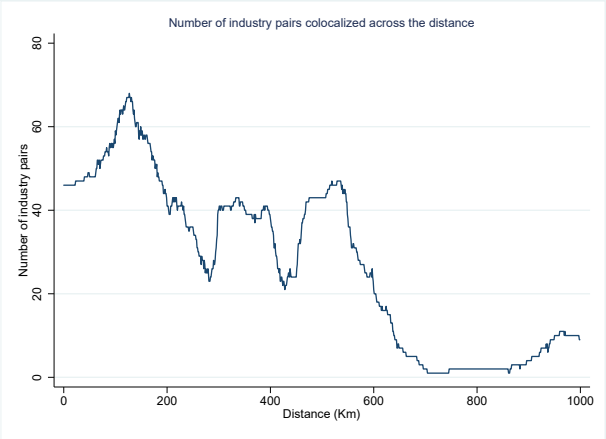


Figure 2.3 – Number of colocalized industry pairs at each distance in 2015

On the scale of coagglomeration and co-dispersion, the graphs in Figure 2.3 show the number of industry pairs colocated at each distance. It is immediately apparent on these graphs that colocalization happens essentially at almost all distances but much more at short distances compared to large distances. On the contrary, the extent of co-dispersion is much greater at long distances compared to short distances.

Coagglomeration of innovation and production compared

Table 2.2 reports on average over the period 2001 - 2015 (odd years), the numbers, and percentage of colocalized, codispersed, or randomly distributed industry pairs for both innovation and production. The table also provides in-

dexes that measure the extent of the strength of colocalization/co-dispersion. Details on the computation of these figures are presented in appendix 2.6.2. [13]

Table 2.2 – Localization-Dispersion-Randomness

	Innovation	*Production*
Global colocalization		
Number of industry pairs	112	100
% of industries pairs	65	58
Index of global colocalization (x 10-4)	253	122
Global codispersion		
Number of industry pairs	7	8
% of industries pairs	4	4
Index of global codispersion (x 10-4)	50	51
Random distribution		
Number of industry pairs	52	63
% of industries pairs	30	37
Total number of industry pairs	**171**	**171**

Notes: The formulas used to compute the Index of global colocalization and the Index of global co-dispersion are presented in Appendix 2.6.2. The values presented in this table are averages across the 8 odd years from 2001 to 2015. Table 2.13 in the appendix gives details of these numbers for each year.

Table 2.2 shows that global colocalization is more common than global co-dispersion for both innovation and production. On the innovation side, up to 65% of industry pairs on average are colocalized, compared to 4% of codispersed industry pairs, and 30% that are randomly distributed. On the produc-

13. The formula presented in equation 2.1 is used to estimate the coagglomeration of the production, with data of the manufacturing plants in Canada as recorded in the Scott's National Directories database. The weights are the number of workers of each plant so that the coagglomeration of production considered in the analysis is the coagglomeration of the employment.

tion side, the pattern of coagglomeration is quite similar with 58%, 4%, and 37% respectively for colocalized, codispersed, or randomly distributed industry pairs. However, it appears that the share of colocalized industry pairs, as well as the strength of colocalization, are larger for innovation. Put differently, on average colocalization is more frequent and more pronounced for the industry pairs in terms of innovation rather than in terms of production. This is in line with the stylized fact stating that innovation is more concentrated than production (Feldman and Kogler, 2010).

2.3 The causes of the coagglomeration of innovation

It is well established that the formation of agglomeration is attributable to some locational fundamentals, and some other micro-founded determinants. Marshall (1890) suggests that input-output, labor pooling, and knowledge spillover may well explain agglomeration. In addition, there is also evidence that natural advantage, home market effects, consumption opportunities, and rent-seeking all contribute to agglomeration (Rosenthal and Strange, 2004). In what follows, the intuition behind these Marshallian forces, as well as some other non-Marshallian determinants and how they connect to the agglomeration of innovation is briefly reviewed. Then, the measures used to account for each of these forces in the empirical analysis are also presented.

2.3.1 Marshallian externalities

In a discussion on innovation and agglomeration, Carlino and Kerr (2015) suggest that the traditional Marshallian externalities, of input sharing, labor pooling, and knowledge spillover are especially important for the spatial concentration of innovation activities. Although these marshallian externalities are

relevant to act on the coagglomeration of innovation, we discuss in the next sub-sections how their measures should differ from their equivalent on the production side. Moreover, in this analysis, the agglomeration of innovation is measured in terms of the geographic colocation of patents from different industry pairs. Hence, this measure of the coagglomeration of innovation will guide the following discussion on how the Marshallian mechanisms may be at play to induce the coagglomeration of the industry pairs.

Input sharing

At first glance, it could be tricky to figure out how the input sharing may act in generating the agglomeration of innovation, given that input for innovation is more likely to be other ideas than material. It may be useful to keep in mind that in many cases, even conceptual innovations need to be tested through experiences or be transformed into a new product. As such, innovation may gain from the sharing of common inputs, given that input-output linkages reduce the costs of outsourcing the inputs needed to implement new ideas (Helsley and Strange, 2002). Another way through which input sharing could potentially foster the formation of the agglomeration of innovation is through the buyer-supplier linkages induced by the exchange of inputs. For example, when a firm buys a piece of equipment for its production process, all the packages in terms of training, maintenance and customer service that accompany the equipment are many channels that can induce exchange of knowledge and ultimately innovation.

To measure input sharing between a given industry pair, three variables are constructed. Let define: $Input_{i \leftarrow j}$ the share of industry i's inputs that come from industry j. Similarly, let $Output_{i \rightarrow j}$ be the share of industry i's outputs that are sold to industry j. These shares are computed relative to all industries including non-manufacturing. The symmetric measures of input sharing are

then : $Input_{ij} = Max(Input_{i\to j}, Input_{i\leftarrow j})$; $Output_{ij} = Max(Output_{i\to j}, Output_{i\leftarrow j})$, and $InputOutput_{ij} = Max(Input_{ij}; Output_{ij})$. [14]

The data used to compute these metrics are the symmetric input-output matrices of the USA for the years 1997 and 2002, from the Bureau of Economic Analysis of the US department of commerce, and the symmetric input-output matrices of Canada for the years 2001 to 2009 from Statistics Canada. For both Canada and the USA, the measures are computed as described above at 3-digits and averaged across the years to have a unique cross-sectional variable. It is worthwhile noting that even if the material is not an input for the production of a patent, it remains a good proxy to capture the knowledge generated through the buyer-supplier connection as well as the role of material in the process of innovating before the application for the patent happens. We use measures constructed with the USA data first as a proxy of the measure for Canada. Then, we also use the direct measure of input-output constructed with the Canadian data, using the USA measure as an instrument to deal with potential endogeneity. We provide more details in the identification strategy in section 2.4

Labor pooling

The second Marshallian force is labor pooling. What motivates its role in generating agglomeration is essentially the fact that workers can move across industries provided that these industries use the same type of labor (Helsley and Strange, 1990; Combes and Duranton, 2006). The rationale specific to the agglomeration of innovation may be that by moving from one industry to another, workers bring with them knowledge and experience useful for the creation of innovation. In addition, the complementarity of knowledge across industries is also an advantage that fosters innovation. A clear example is "Biotechnology"

14. Alternative measures can be computed using minimums, averages, instead of maximum, but the underlying intuition remains the same.

which is an extensive innovative field that combines knowledge from both biology and technology.

What is common in this field to measure labor pooling is constructing some measures of labor similarities based on the correlation of the distribution of the employment of industries across different occupations. In this paper, the movement of workers across industries is rather used. This choice is motivated by the fact that the movement of laborers seems a more convenient way to connect labor to innovation creation as argued earlier. Thus, to construct the labor pooling measure, the labor movement pattern from the *Current Population Survey (CPS)* of the US Bureau of Labor Statistics is used. The available data from this source are used to construct the maximum of shares of labor movement across all the pairs, and we only consider the movement of the skilled workers. The same method used for Input-Output is replicated to construct symmetric measures of labor pooling as follow: $Share\_goers - comers_{ij} = Max(share\_comers_{ij}, share\_goers_{ij})$. Since the coagglomeration of innovation of industry pairs is observed at the 3-NAICS digit, these labor movement shares are also computed at the NAICS 3-digit level.

Knowledge spillover

On the knowledge spillover, the intuition on how it affects the agglomeration of innovation seems more direct given that clusters facilitate the creation, accumulation, and diffusion of knowledge among workers and firms (Duranton and Puga, 2004). Moreover, given that new knowledge is valuable only for a short period (Feldman, 1994), combined with the localized nature of the knowledge spillover (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014), innovative agents will have the incentive to cluster to take advantage of

the knowledge flow originating from one another. [15]

The knowledge spillover is measured using the United States Patent and Trademark Office citing-cited patent pairs. The available dataset records the average number of citing/cited patents over the period 1976-2006. These numbers are constructed using a probabilistic mapping of patents between the industrial sector with two variants. The first one is based on the Industry Of Manufacture (IOM) and the second is the Sector Of Use (SOU). The same method used for Input-Output is replicated to construct symmetric measures of knowledge spillover as following $Share\_Citing-Cited_{ij} = Max(share\_Cited_{ij}, share\_Citing_{ij})$ for IOM and SOU. Finally, it should be clear that the patents citation captures more the formal exchange of technology but not all forms of intellectual spillovers as well observed by Carlino and Kerr (2015). However, this may not be a concern given that the coagglomeration of innovation is also measured using patents applications.

2.3.2 Non-Marshallian determinants

In conjunction with the Marshallian forces, some other factors may also be at play in shaping the location of innovation. On the production side, natural advantages such as harbors, coal mines, natural resources are those considered as non-Marshallian determinants (Ellison et al., 2010). For innovation, one may think of local endowments of places such as universities, local policies, and the culture that may favor innovation. Indeed, authors have argued for instance that Silicon Valley and Boston became famous innovative places because

15. Kerr and Kominers (2015) argue that interactions occur at shorter distances and form clusters but eventually overlapping clusters form the agglomeration that is observed at higher spatial scale.

of their proximity to Standford University and MIT Saxenian (1994). Studies have also consider the role of the local culture on innovation (Chinitz, 1961; Saxenian, 1994; Landier, 2005). Finally, there are evidence on the role of local policies on innovation (Fallick et al., 2006). Accounting for these factors in this setting is quite challenging given that the measures used to assess the degree of agglomeration are based on pairs of industries. As such, it is not obvious to imagine a source of variation in some factors that may differently induce co-location of patents from two different industries. For example, one may think of some residential amenities that could attract inventors in a particular place and induce artificial co-location. But it is hard to explain how such residential amenities may vary across inventors from different industry pairs.

To control for possible non-Marshallian determinants of the coagglomeration of innovation, a measure of the coagglomeration of production will be used. Controlling for the coagglomeration of production also helps to capture only the externalities that are specific to innovation. The measure is built from the proprietary Scott's National All Business Directories database, which draws from the Business Register records and telephone surveys. This dataset provides good coverage of the plants operating in Manufacturing industries in Canada. The variables of the database used for the analysis are the precise postal address information of each plant, its industrial classifications (North American Industry Classification System, NAICS 3-digit level), the number of employees in the plant, the year of establishment of the plant. The period of observation spans from 2001 to 2017.[16] With this dataset, the coagglomeration of the production is constructed with the same metric used for the coagglomeration of innovation. Pairs of plants are formed with one plant from each industry of the pair. Next, bilateral distances are computed and then used to estimate the

16. Scott's data are not available for the year 2015. Instead, data from 2017 are used.

Kernel density exactly as presented in the subsection 2.2.2. The employment of the plant is used as the weights in the formula.

2.4 Empirical strategy and results

In this section, we discuss our empirical strategy, and then we present our main results followed by tests of robustness.

2.4.1 Specification and identification

An ideal setup for the estimation of the effects of the three Marshallian forces on the coagglomeration of industry pairs should be a panel specification that includes time-varying industry pairs specific controls, industry-pairs fixed effects as well as year fixed effects. In this paper, due to some limitations in the available data, the main empirical specification builds on Faggio et al. (2017). It is a pooled cross-section for eight years for each of the coagglomeration measures, while the Marshallian forces are averaged values across the available years, as follows:

$$\text{Coagg.Inv}_{ijt}(d) = \alpha \text{Know.Spil}_{ij} + \beta \text{InputOutput}_{ij} + \gamma \text{Labor}_{ij} + \lambda \text{Coagg.Prod}_{ijt}(d) + \eta_t + \varepsilon_{ijt}$$

Where *Coagg.Inv(d)* and *Coagg.Prod(d)* are the cumulative density function of the bilateral distances between the patents and plants respectively at a given distance d; *Know.Spil* is the knowledge spillover measure, *Labor* is the labor pooling measure, *InputOutput* is the input-output measure, η_t is the time fixed effect, ε_{ijt} is the error term.

The coagglomeration of industry pairs in terms of production is used as a control in the model to deal with potential spurious correlations which may originate from natural advantages, unobserved local policies, and other factors that

could jointly target innovative industries but not related to the coagglomeration of innovation.

One may wonder that in our setting, patens are used to measure co-agglomeration of innovation and at the same time patents are also used to compute the knowledge spillover measure of industry pairs. This may not be a major issue because one of the measures is based on the geographical proximity of the industry pairs, while the other is based on industrial interconnection. And our goal assessing the extent to which geographical proximity would have its origins in an interconnection in terms of knowledge exchange. Moreover, it is common in the field of research connected to our analysis to use this kind of measure. For example, Ellison et al. (2010) do a similar exercise from the point of view of production. On the one hand their dependent variable is based on plants and on the other hand the determinant of input-Output is based on the exchange of goods between industry pairs. Nevertheless, to deal with possible reverse causality, the Marshallian measures used in this specification are all constructed using information from the USA. There are good reasons to assume that the patterns of these measures across the manufacturing industries are likely to be the same in the USA and Canada since these two North American Countries share common industry standards such as the NAICS classification, the IPC classification. Moreover, the geographic proximity of the two countries favors many similarities in their manufacturing structures. As such, using the data from the United States as proxies for Canadian measures prevents to some extent the Marshallian measures to be entailed with endogeneity with the coagglomeration of the Canadian innovation. Thus, the constructed Marshallian measures can be assumed to be exogenous and well appropriated to capture the causal effects of these measures on the coagglomeration of innovation in Canada.

Another potential concern may be, the so-called *third industry effect*.[17] Some robustness checks will be carried out with subsamples free from industries that are more likely to generate the *third industry effect*.

Finally, the error terms are clustered at industry pair since only the coagglomeration measures have the time dimension. In addition, these error terms are bootstrapped to account for the fact that the Coagglomeration of the production on the right-hand side is also estimated.

In what follows the model specified as in equation 2.2 is first estimated and then some alternative specifications along with several checks for robustness are also considered. All the variables are normalized to have a zero mean and a unit standard deviation. This transformation allows for the comparison between the relative importance of the effects of the determinants. The model is estimated at a distance of 20 Km which is close to two times the median commuting distance in major cities in Canada.[18]

2.4.2 Main results

This section presents the estimates of the effects of the Marshallian forces on the coagglomeration of the production in Canada, followed by the results for the innovation.

---

17. To exemplify the third industry effect, let consider three industries A, B, and C are such that Marshallian mechanisms cause A to colocate with B and B to colocate with C, then A will be mechanically located near C even if they are no Marshallian mechanisms at play in that third relation.

18. The median distance to work among Canadian workers who had a usual workplace was 8.7 kilometers in 2016. See https://www150.statcan.gc.ca/n1/daily-quotidien/190225/dq190225a-eng.htm .

Table 2.3 – Regression of the coagglomeration of the production on Marshallian forces

	(1)	(2)	(3)	(4)
	Co-Agg.Prod	Co-Agg.Prod	Co-Agg.Prod	Co-Agg.Prod
Max input-Output - US	0.219***			0.180***
	(0.064)			(0.068)
Max labor movement of highly educated		0.160**		-0.005
		(0.066)		(0.072)
Max knowledge spillover - IOM			0.239***	0.212***
			(0.062)	(0.066)
Time FE	Yes	Yes	Yes	No
Nb of industry pairs x year	1368	1368	1368	1368
Adj R^2	0.04	0.02	0.05	0.08

Notes: The multivariate model is estimated as specified in equation 2.2, where innovation is replaced by production. It is a pooled cross-section from 2001 to 2015 for the coagglomeration of the production. The Marshallian forces are averaged across the years for which data are available. The years 2001, 2005, and 2009 for the Labor pooling; The odd years from 2001 to 2013 for the input-output; and the years 1976 to 2006 for the knowledge spillover. For all the specifications, variables are standardized to have zero mean and unit standard deviation. Standard errors are clustered at the industry pairs level and bootstrapped. Significant at ***1% **5% *10%

The first three columns of Table 2.3 give the estimates for the univariate specifications for each of the three forces. A one standard deviation increase in input sharing is associated with a 0.22 standard deviation increase in the actual coagglomeration of the production. The association is 0.16 for the labor pooling and 0.24 for the knowledge spillover. The coefficients for the input sharing and that of knowledge spillover are significant and stronger than the coefficient for the labor pooling. The specification in column (4) where all three forces are included yields positive and significant effects of input sharing and knowledge spillover, but the magnitude of these effects are smaller than those of the univariate specifications. In addition, the labor movement no longer has a significant effect.

The message that emerges from these results, is that only the input sharing and

the knowledge spillover positively and significantly affect the coagglomeration of the production in Canada. Let's keep in mind that the Marshallian forces are measured here in the perspective of the analysis of the coagglomeration of innovation. [19]

Table 2.4 gives the estimates of the effects of the input-output, the labor movement, the knowledge spillover, and the coagglomeration of the production on the coagglomeration of innovation.

---

19. When the measures used to account for the Marshallian forces are the same as in previous studies on the coagglomeration of the production e.g Ellison et al. 2010; Faggio et al. 2017, that is using the labor correlation instead of the labor movement, all the three Marshallian forces have positive and significant effects on the coagglomeration of the production for the Canadian case. See Table 2.17 in the appendix

Table 2.4 – Regression of the coagglomeration of innovation on the Marshallian forces

	(1)	(2)	(3)	(4)	(5)
	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv
Max input-Output - US	0.019				-0.049
	(0.051)				(0.051)
Max labor movement of highly educated		0.156**			-0.011
		(0.064)			(0.073)
Max knowledge spillover - IOM			0.358***		0.373***
			(0.076)		(0.098)
co-agglomeration of plants - cdf				0.424***	
				(0.065)	
Time FE	Yes	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1368	1368	1368	1368	1368
Adj R^2	-0.01	0.02	0.12	0.17	0.12

Notes: The multivariate model is estimated as specified in equation 2.2. It is a pooled cross-section from 2001 to 2015 for the coagglomeration measures of innovation and production. The Marshallian forces are averaged across the years for which data are available. The years 2001, 2005, and 2009 for the Labor pooling; The odd years from 2001 to 2013 for the input-output; and the years 1976 to 2006 for the knowledge spillover. For all the specifications, variables are standardized to have zero mean and unit standard deviation. Standard errors are bootstrapped to account for the estimated nature of the coagglomeration of the production on the right-hand side. Significant at ***1% **5% *10%

As can be seen, in table 2.4 the input output variable is not correlated to the coagglomeration of innovation indicating that input-out linkages has no effect on the spatial concentration of industry pairs as innovation is concerned. The labor movement has a positive association with the coagglomeration of innovation. As discussed earlier, the movement of workers from an industry to another may well stimulate the creation of innovation and thus its spatial concentration. The knowledge spillover is also positively associated with the concetration of innovation suggesting that patents citation is a channel of the coagglomeration of innovation. Finally, the coagglomeration of the production is positively correlated to the coagglomeration of innovation, this is in line with evidence of the coagglomeration of innovation and production provided in Lan

(2019).

Table 2.5 gives the estimates of the effects of the input-output, the labor movement, and the knowledge spillover on the coagglomeration of innovation, controlling for the coagglomeration of the production.

Table 2.5 – Regression of the coagglomeration of innovation on the Marshallian forces

	(1)	(2)	(3)	(4)
	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv
Max input-Output - US	-0.065			-0.102*
	(0.062)			(0.058)
Max labor movement of highly educated		0.087		0.002
		(0.058)		(0.071)
Max knowledge spillover - IOM			0.248***	0.260***
			(0.071)	(0.092)
co-agglomeration of plants - cdf	0.435***	0.407***	0.344***	0.357***
	(0.068)	(0.065)	(0.057)	(0.061)
Time FE	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1368	1368	1368	1368
Adj R^2	0.18	0.18	0.23	0.24

Notes: The multivariate model is estimated as specified in equation 2.2. It is a pooled cross-section from 2001 to 2015 for the coagglomeration measures of innovation and production. The Marshallian forces are averaged across the years for which data are available. The years 2001, 2005, and 2009 for the Labor pooling; The odd years from 2001 to 2013 for the input-output; and the years 1976 to 2006 for the knowledge spillover. For all the specifications, variables are standardized to have zero mean and unit standard deviation. Standard errors are bootstrapped to account for the estimated nature of the coagglomeration of the production on the right-hand side. Significant at ***1% **5% *10%

Columns (1) to (3) present univariate estimates for each of the Marshallian forces. Not surprisingly, the co-agglomeration of production is positively related to the co-agglomeration of innovation in line with Lan (2019) who provided evidence of the colocation of production and innovation. In addition, the estimates are insignificant and negative for the input-output, positive and

insignificant effects for the labor movement, but positive and significant effects for the knowledge spillover. More precisely, a one standard deviation of the input-output is translated in a -0.065 standard deviation of the coagglomeration of innovation. The association is 0.087 for the labor movement and 0.248 for the knowledge spillover.

Column (4) presents the estimates for the multivariate specification which includes the three Marshallian forces altogether with the time fixed effect, but no control. This specification shows no effect for both the input sharing and the labor pooling measures, but a positive and significant effect of the knowledge spillover. Column (5) adds the coagglomeration of the production as a control in the multivariate specification. With this specification, a negative but hardly significant effect of the input sharing pops up, there is still no effect for the labor pooling and only the knowledge spillover has a positive and significant effect, along with the coagglomeration of the production.

These results show that all things else equal, part of the coagglomeration of the innovation is driven by the coagglomeration of the production. Moreover, input sharing reduces the coagglomeration of innovation, but this negative effect is hardly significant. In addition, all things else equal, the labor movement of highly educated workers does not drive the coagglomeration of innovation in Canada. Finally, all things else equal, knowledge spillover has a positive and significant effect on the coagglomeration of innovation.

2.4.3 Robustness checks

In this section, checks for robustness are carried out in five directions. First, the influence of the distance at which the coagglomeration is captured. Second, different other specifications are discussed. Third, alternative measures of Mar-

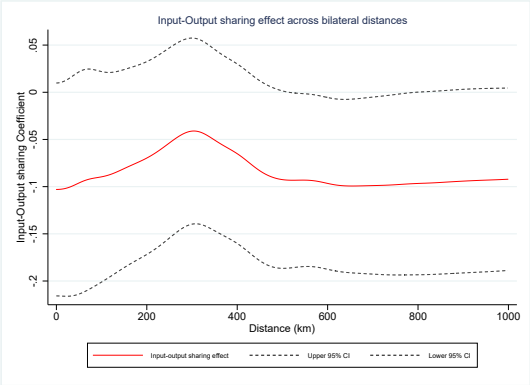
shallian forces are also considered. Fourth, some specific subsets of the dataset are considered, and fifth, we also rerun the model with the alternative samples presented in section 2.2.1.

The effect of the distance

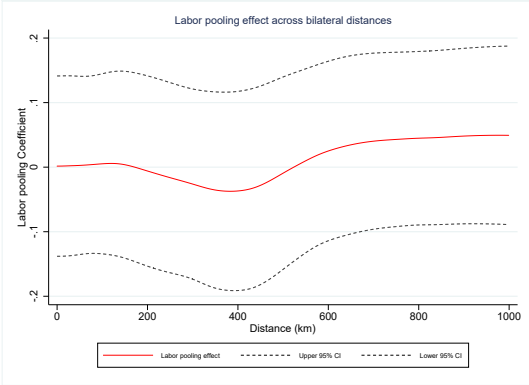
For the main specification, we estimate the effects of the Marshallian forces on the coagglomeration of innovation measured with the cumulative distribution up to a distance d=20 km. This choice has been guided by the mere fact that the median commuting distance estimated in 2016 in Canada is 8.7 km, and we use 20 km which is close to twice the value of that commuting. However, previous studies have shown that the scale at which agglomeration is affected is not the same for all the Marshallian forces (Rosenthal and Strange, 2001). We now check if our results change when we vary the value of this distance ranging from 0 to 999 km.

Figure 2.4 presents for each determinant, the variation of the estimated coefficients with the distance at which the coagglomeration is considered, along with their 5% confidence intervals at each distance.

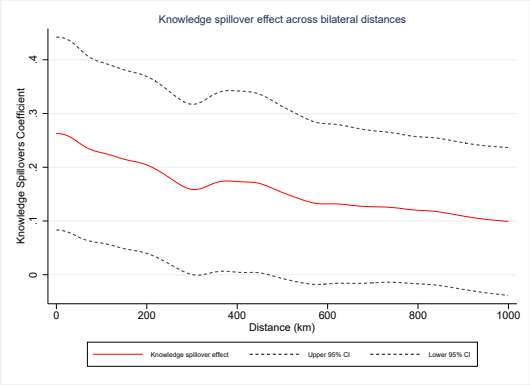
Figure 2.4 – Marshallian effects at different coagglomeration distances



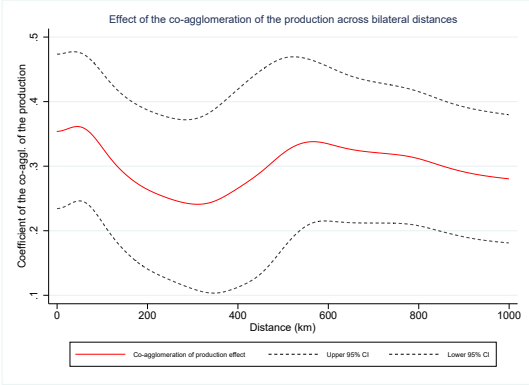
(a) input sharing



(b) Labor pooling



(c) Knowledge spillover



(d) Coagglomeration of production

Panel (a) shows a negative effect of the input-output measure, but this effect is hardly detectable. On Panel (b) the effect for the labor movement is also insignificant at each distance from 0 to 999 Km. Panel (c) indicates a positive and significant effect of the knowledge spillover which increases from 0 to about 300 km. Above that distance, the effect of the knowledge spillover is no longer detectable. Finally, the effect of the coagglomeration of the production –panel (d)– is also positive and significant at each distance. This check establishes that the main results found on the three Marshallian forces are stable for any dis-

tance lower than 300 Km, which is a reasonable upper bound to consider the effects of the Marshallian forces on the agglomeration of the innovation.

Using differents specifications

Another check is conducted using different specifications of equation 2.2.

First, in the original specification, the available data used are pooled cross-sections of Marshallian forces when both the coagglomeration of the production and the coagglomeration of innovation are panel data from 2001 to 20015 (odd years). This situation may potentially cause erroneous estimated standard errors and thus spurious significance for some coefficients. We cluster our errors at the industry pairs level to alleviate the effects of this data limitation. Now, at the expense of the number of observations, we check the extent of this situation by using a specification where all the measures including the coagglomeration ones are averaged across the years and the model is estimated as a single cross-section. The results are roughly the same as those of the benchmark, the negative but hardly significant effect of the input sharing disappears, and both the effects of the knowledge spillover and coagglomeration of the production increase (see column 2 in table 2.6).

Second, we add industry fixed effects to further try to get rid of all possible confounding effects specific to innovation that are not captured by the coagglomeration of the production.[20] The results in column (3) remain qualitatively the same. However, the effect of the input sharing becomes stronger and its magnitude drops. On the contrary, the effect of the knowledge spillover becomes weaker and its magnitude drops too.

Third, to measure input sharing, we use the variable constructed with data

20. To construct these fixed effects, for each industry a dummy variable equals 1 is generated when the industry belongs to the industry pair.

from the USA as a proxy of the Canadian measures. We check what the results are with the actual variable for Canada. In column 3 -Table 2.6- instead of the input sharing of the US as a proxy measure of that of Canada, we use the measure of Canada, that we instrument by the measure of the US to deal with potential endogeneity. The results are roughly the same as those from the benchmark both qualitatively and quantitatively, except the input-output effect which is no longer significant.

Table 2.6 – Estimates of the effects of the Marshallian forces for different specifications

	Benchmark	Alternative specifications		
	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	IV
Max input-Output - CA				-0.123
				(0.090)
Max input-Output - US	-0.102*	-0.111	-0.043**	
	(0.058)	(0.077)	(0.021)	
Max labor movement of highly educated	0.002	-0.012	0.016	0.017
	(0.071)	(0.077)	(0.029)	(0.071)
Max knowledge spillover - IOM	0.260***	0.291***	0.062**	0.252***
	(0.092)	(0.103)	(0.027)	(0.088)
co-agglomeration of plants - cdf	0.357***	0.383***	0.058	0.361***
	(0.061)	(0.063)	(0.043)	(0.063)
Industry FE	No	No	Yes	No
Time FE	Yes	No	Yes	Yes
Nb of industry pairs x year	1368	171	1368	1368
Adj R^2	0.24	0.27	0.79	0.24
Kleinbergen-Paap statistic				66.07

Notes: The estimates in column (1) are those from the multivariate model exactly as specified in equation 2.2. In column (2) all the variables are averaged to have a single cross-section. In column (3) we add industry fixed effects. Column (4) is an IV regression where the input sharing measure for Canada is instrumented by that of the US. Significant at ***1% **5% *10%

This check shows that, once the coagglomeration of the production is controlled

for, only the knowledge spillover affects significantly the coagglomeration of innovation. The labor movement of skilled workers consistently has no effect, and the negative effect of the input sharing which is hardly significant is not robust across the different specifications.

Using different measures

We now argue that our results do not merely reflect the very nature of the variables used. To that end, we consider different alternative measures of the Marshallian forces. In particular, the input-output is measured using the average of the inputs-outputs instead of the maximum of the inputs-outputs, and the results indicate that variable does not affect the coagglomeration of innovation (see column 1 in Table 2.7).

Next, labor pooling is measured using the labor movement of all the workers and not only the highly educated workers, and the effect is still insignificant for the *labor pooling* (see column 2 in Table 2.7).

Finally, the - Sector Of Use - version of the Knowledge Spillover is considered instead of the - Industry Of Manufacturing -(see column 3 in table 2.7). The estimates show a consistently positive effect of knowledge spillover, the labor pooling remains insignificant, and the negative but hardly significant effect of the input sharing becomes a bit stronger.

All in all, with these alternative measures, the knowledge spillover and the coagglomeration of the production consistently have a significant and positive effect on the coagglomeration of innovation, and the two other Marshallian forces have no robust effects.

Table 2.7 – Estimates of the effects of the Marshallian forces with alternative measures

	(1) Co-Agg.Inv	(2) Co-Agg.Inv	(3) Co-Agg.Inv
Max input-Output - US		-0.110*	-0.130**
		(0.058)	(0.053)
Avg input-Output - US	-0.083		
	(0.066)		
Max labor movement of highly educated	-0.000		0.031
	(0.074)		(0.061)
Max labor movement		0.024	
		(0.074)	
Max knowledge spillover - IOM	0.263***	0.250***	
	(0.092)	(0.091)	
Max knowledge spillover - SOU			0.221***
			(0.077)
co-agglomeration of plants - cdf	0.353***	0.357***	0.371***
	(0.060)	(0.061)	(0.063)
Time FE	Yes	Yes	Yes
Nb of industry pairs x year	1368	1368	1368
Adj R^2	0.23	0.24	0.23

Notes: In column (1) the Max of input-output is replaced by the Average of input-output. In column (2) The labor movement of the highly educated is replaced by the labor movement of all the workers. In column (3) the Sector Of Use -SOU option of the knowledge spillover is considered instead of the Industry Of Manufacturing-IOM. Significant at ***1% **5% *10%

Using different subsets of data

We discuss now issues related to the third industry effects and some other potential artificial coagglomeration of the innovation, not related to the Marshallian mechanism. The rationale behind the structure of the NAICS industrial classification may hide some form of connection between industries not related to the Marshallian mechanisms. To check this, we consider different subsets that exclude all industry pairs for which the two three-digits industries belong to the same two-digits groups. This is meant to make sure that the results are

not driven by some factors related to the classification of the industries. In addition, this may help to check for the third industry effect by removing in the estimation these industries which may potentially be geographically close. Column(1) excludes the pairs for which both industries are from the 2-digit "31". Column(2) excludes the pairs for which both industries are from the 2-digit "32", and Column(3) excludes the pairs for which both industries are from the 2-digit "33". Finally, column(4) excludes pairs for which both of the industries are from the same 2-digit industry. The results are presented in - Table 2.8-.

Table 2.8 – Estimates of the effects of the Marshallian forces for subsets of industry pairs

	Benchmark	Alternative subsets of industry pairs			
	(1)	(2)	(3)	(4)	(5)
	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv
Max input-Output - US	-0.102*	-0.036	-0.143*	-0.107	-0.078
	(0.058)	(0.061)	(0.082)	(0.090)	(0.087)
Max labor movement of highly educated	0.002	-0.017	0.023	0.005	0.019
	(0.071)	(0.053)	(0.064)	(0.076)	(0.097)
Max knowledge spillover - IOM	0.260***	0.229***	0.276***	0.213**	0.185*
	(0.092)	(0.087)	(0.105)	(0.103)	(0.112)
co-agglomeration of plants - cdf	0.357***	0.401***	0.369***	0.345***	0.406***
	(0.061)	(0.062)	(0.082)	(0.067)	(0.075)
Time FE	Yes	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1368	1288	1248	1144	944
Adj R^2	0.24	0.25	0.24	0.20	0.22

Notes: The Benchmark specification is considered. The estimates in column (2) exclude the pairs for which both of the industries are from the 2-digit "31". The estimates in column (3) exclude the pairs for which both of the industries are from the 2-digit "32". The estimates in column (4) exclude the pairs for which both of the industries are from the 2-digit "33". The estimates in column (5) exclude all the 497 pairs such that $31a-31b; 32c-32d; 33e-33f$ where $a,b\varepsilon\{1,2,3,4,5,6\}$ and $a \neq b$; $c,d\varepsilon\{1,2,3,4,5,6,7\}$ and $c \neq d$; $e,f\varepsilon\{1,2,3,4,5,6,7,9\}$ and $e \neq f$. Significant at ***1% **5% *10%

With these different subsets, the results for the knowledge spillover remain qualitatively the same, with some changes in its magnitudes and its significance. The effect of the labor movement is still insignificant and the negative

and hardly significant effect of the input sharing is not robust across all the subsets. We are thus, confident that our results are free from artificial effects which would have originated from the third industry effect and/or the NAICS structure. [21]

Using alternative samples of the patents

As discussed in section 2.2.1, the steps used to shape the initial dataset to the convenience of our analysis, have induced some concerns. In particular, the necessity of locating patents across space, and transforming them into NAICS classification has induced an artificial increase in the number of patents. We dealt with these issues in the main results by using weights while estimating the coagglomeration of the innovation. We now use alternative means of addressing the same issue to discuss the robustness of our main results. First, we re-estimate the coagglomeration of innovation using the alternative samples (2 and 3) presented in Table 2.1. Second, we consider a different way of computing the cumulative distribution of the Coagglomeration of the innovation based on the benchmark sample. Third, we consider different weights for the estimation of the coagglomeration of the innovation.

For *Sample 1* (benchmark), each patent is duplicated across the location of all its related inventors and duplicated across all the manufacturing NAICS with a non-zero probability used to move from the IPC to the NAICS classification. The weight used to estimate the coagglomeration of the innovation is the probability link.

For *Sample 2* each patent is duplicated across the location of all its related inventors, and only the NAICS with the maximum probability used to move from

21. A comprehensive check of the incidence of individual industry in the pairs is presented in the appendix 2.15 and 2.16. The results suggest that the effect of the knowledge spillover on the industry-pairs including at "324" is important for the coagglomeration of innovation.

the IPC to the NAICS classification is considered. The weight used to estimate the K-density is the probability link.

For *Sample 3*, each patent is duplicated across the location of all its related inventors, and then only the NAICS with the maximum probability used to move from the IPC to the NAICS classification is considered, provided that the value of the probability is greater or equal to 0.5. The weight used to estimate the K-density is the probability link.

Sample 4 is the same as Sample 1, but instead of the cumulative K-density of the innovation up to the distance d (d=20), the sum is computed without the value at the distance equal to zero. This is meant to correct for the artificial increase in zero bilateral distances induced by the transformation of the dataset.

Sample 5 is identical to the benchmark, but the weight used to estimate the K-density is the probability link times the inverse of the number of inventors related to each patent.

Sample 6 is Identical to sample 2, but the weight used to estimate the K-density is the probability link times the inverse of the number of inventors related to each patent.

Sample 7 is identical to sample 3, but the weight used to estimate the K-density is the probability link times the inverse of the number of inventors related to each patent. The results are presented in Table 2.9.

Table 2.9 – Estimates of the effects of the Marshallian forces for samples of industry pairs

	(1) Sample 1	(2) Sample 2	(3) Sample 3	(4) Sample 4	(5) Sample 5	(6) Sample 6	(7) Sample 7
Max input-Output - US	-0.102*	-0.033	0.006	-0.102*	-0.104	-0.054	-0.032
	(0.058)	(0.049)	(0.060)	(0.058)	(0.065)	(0.052)	(0.058)
Max labor movement of highly educated	0.002	-0.058	-0.113*	0.002	-0.013	-0.081	-0.133**
	(0.071)	(0.067)	(0.061)	(0.071)	(0.074)	(0.072)	(0.066)
Max knowledge spillover - IOM	0.260***	0.237***	0.146**	0.260***	0.221**	0.218**	0.144**
	(0.092)	(0.087)	(0.073)	(0.092)	(0.090)	(0.086)	(0.073)
co-agglomeration of plants - cdf	0.357***	0.350***	0.333***	0.357***	0.383***	0.339***	0.320***
	(0.061)	(0.063)	(0.064)	(0.061)	(0.066)	(0.065)	(0.062)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1368	1368	1368	1368	1368	1368	1368
Adj R^2	0.24	0.20	0.14	0.24	0.23	0.18	0.13

Notes: The multivariate model is estimated as specified in equation 2.2. Significant at ***1% **5% *10%.

Some differences pop up in the results across these samples, but the main conclusion remains the same. First, the input sharing has a negative but hardly significant effect for samples 1 and 4, but no effect with the other samples. There is a negative and weak effect of the labor movement for the specifications sample 3 and sample 7. The knowledge spillover has a positive and persistent effect for all the specifications and so is the coagglomeration of the production. The conclusion remains the same, the coagglomeration of the production and the knowledge spillover affect positively the coagglomeration of the innovation and the two other forces have no significant effects.

2.5 Conluding remarks

The paper uses the continuous metric of Duranton and Overman (2005) to describe the coagglomeration of innovation in Canada. The observed patterns reveal that Canadian innovation is concentrated and even more than production,

in line with what has been found in other countries. Then, building on Ellison et al. (2010); Faggio et al. (2017), the paper analyzes the effects of labor pooling, input sharing, and knowledge spillover on the coagglomeration of innovation. The analysis shows that only the knowledge spillover unambiguously causes the coagglomeration of innovation.

Two main cautions are to be acknowledged in this study. First, the measure of innovation only captures formal innovation for which protection is explicitly applied. As such, the actual pattern of the coagglomeration of innovation in Canada may be different to some extent, but more than likely stronger than what is uncovered in this paper, given that the actual innovation may be denser than what we capture here. Second, the Marshallian forces are expected to influence the coagglomeration of innovation through some form of knowledge diffusion, be it through the input sharing, the labor movement, or the knowledge spillover so that the results of this paper only inform on the fact that knowledge that spills formally through patents citations is the main driver of the coagglomeration of the formal innovation in terms of patents applications. Yet, the paper remains silent on the part of the effects of the Marshallian forces channeled through the flow of knowledge that *leaves no paper trails*.

However, the results remain informative for possible policy issues. The estimates of the difference in the effects of the Marshallian forces on the co-agglomeration of innovation compared to those on the co-agglomeration of the production are very useful on cost–benefit and cost-effectiveness analyses of urban and industrial policies. Through its spatial heterogeneity results, the paper also offers avenue to foster and promote regional innovation capacity in Canada.

2.6 Appendix to chapter 2

2.6.1 Additional tables on innovation data

Table 2.10 – Counts of patents extended to NAICS and addresses across the years

	2001	2003	2005	2007	2009	2011	2013	2015	Average
311 Food manufacturing	1936	2108	2208	2185	1978	1747	1789	1817	1971
312 Beverage and tobacco product manuf.	1215	1294	1308	1417	1251	1031	1063	972	1194
313 Textile mills	50	52	94	55	57	47	39	51	56
314 Textile product mills	0	1	4	1	0	0	1	2	2
315 Clothing manufacturing	773	821	921	878	734	791	726	846	811
316 Leather, allied product manuf.	308	324	297	265	213	268	276	256	276
321 Wood product manufacturing	95	143	106	120	51	73	91	61	93
322 Paper manufacturing	746	831	896	718	819	695	709	796	776
323 Printing, support activities	9	24	33	7	18	15	6	17	16
324 Petrol, coal product manuf.	629	826	953	1058	800	784	857	620	816
325 Chemical manufacturing	2661	2909	3215	3054	2688	2767	2857	2730	2860
326 Plastics, rubber products manuf.	2634	2750	2898	2825	2366	2708	2581	2719	2685
327 Non-metallic mineral product manuf.	1150	1321	1575	1547	1227	1324	1392	1283	1352
331 Primary metal manufacturing	601	559	617	535	392	485	509	448	518
332 Fabricated metal product manuf.	2785	2982	2985	2890	2436	2634	2454	2602	2721
333 Machinery manufacturing	4365	4637	4729	4523	3912	4428	4421	4712	4466
334 Computer, electronic product manuf.	5652	5527	6086	5841	5055	6146	5287	5099	5587
335 Electrical, appliance manuf.	1488	1666	1678	1514	1267	1557	1436	1403	1501
336 Transportation equipment manuf.	1115	1259	1309	1144	965	953	1061	1166	1122
337 Furniture, related product manuf.	519	588	638	681	517	420	409	421	524
339 Miscellaneous manufacturing	2148	2273	2483	2269	2178	2124	2131	2202	2226
Total	30879	32895	35033	33527	28924	30997	30095	30223	31572

Notes: This table reports patents and innovators counts across the years with duplicates innovators to convert the IPC database into a NAICS database without accounting for the probability link

Table 2.11 – Counts of patents extended to addresses and one unique NAICS 3-digit across the year

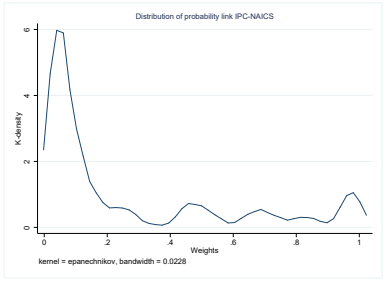
	2001	2003	2005	2007	2009	2011	2013	2015	Average
311 Food manufacturing	752	717	613	618	656	545	563	566	629
312 Beverage and tobacco product manuf.	36	40	38	29	15	19	13	25	27
313 Textile mills	25	17	24	25	29	21	16	19	22
315 Clothing manufacturing	39	53	57	41	45	51	50	40	47
316 Leather, allied product manuf.	76	72	80	84	88	58	99	79	80
321 Wood product manufacturing	80	124	88	89	46	53	82	53	77
322 Paper manufacturing	325	297	310	221	339	329	331	323	309
324 Petrol, coal product manuf.	35	60	48	123	73	169	111	81	88
325 Chemical manufacturing	559	663	828	755	671	601	757	480	664
326 Plastics, rubber products manuf.	195	253	238	265	144	266	157	272	224
327 Non-metallic mineral product manuf.	315	339	399	404	314	286	324	309	336
331 Primary metal manufacturing	219	101	96	82	46	73	102	74	99
332 Fabricated metal product manuf.	272	332	298	247	153	210	207	243	245
333 Machinery manufacturing	1245	1318	1402	1517	1324	1508	1678	1736	1466
334 Computer, electronic product manuf.	3085	2850	3315	3043	2781	3470	2490	2113	2893
335 Electrical, appliance manuf.	129	167	142	156	150	206	219	224	174
336 Transportation equipment manuf.	467	495	437	356	316	343	358	426	400
337 Furniture, related product manuf.	199	203	211	216	200	150	160	152	186
339 Miscellaneous manufacturing	991	1015	1141	1046	906	959	937	1039	1004
Total	9044	9116	9765	9317	8296	9317	8654	8254	8970

Notes: This table reports patents and innovators counts across the years with duplicates innovators to convert the IPC database into a NAICS database without accounting for the probability link

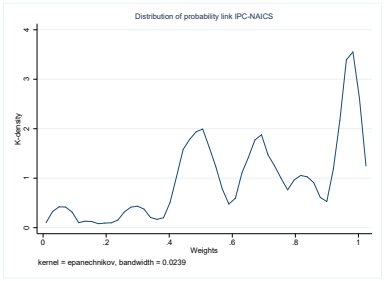
Table 2.12 – Counts of patent extended to addresses and one unique NAICS 3-digit with prob. > 0.5 across the years

	2001	2003	2005	2007	2009	2011	2013	2015	Average
311 Food manufacturing	103	69	59	48	86	68	55	40	66
312 Beverage and tobacco product manuf.	36	40	38	29	15	19	13	25	27
313 Textile mills	25	17	24	25	29	21	16	19	22
315 Clothing manufacturing	39	53	57	41	45	51	50	40	47
316 Leather, allied product manuf.	48	39	62	48	72	31	74	60	54
321 Wood product manufacturing	80	124	88	89	46	53	82	53	77
322 Paper manufacturing	70	79	87	39	65	53	38	53	61
324 Petrol, coal product manuf.	35	60	48	123	73	169	111	81	88
325 Chemical manufacturing	206	200	231	218	220	259	334	267	242
326 Plastics, rubber products manuf.	74	139	105	117	52	71	43	68	84
327 Non-metallic mineral product manuf.	308	323	381	362	310	281	311	293	321
331 Primary metal manufacturing	219	101	96	82	46	73	102	74	99
332 Fabricated metal product manuf.	123	186	172	116	99	135	137	162	141
333 Machinery manufacturing	767	745	835	958	805	891	1081	1106	899
334 Computer, electronic product manuf.	3075	2843	3294	3035	2761	3468	2481	2108	2883
335 Electrical, appliance manuf.	104	120	95	98	95	172	177	184	131
336 Transportation equipment manuf.	467	495	437	356	316	343	358	426	400
337 Furniture, related product manuf.	199	203	211	216	200	150	160	152	186
339 Miscellaneous manufacturing	918	940	1034	984	837	904	851	923	924
Total	6896	6776	7354	6984	6172	7212	6474	6134	6750

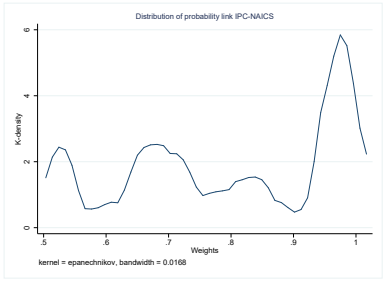
Notes: This table reports patents and innovators counts across the years with duplicates innovators to convert the IPC database into a NAICS database without accounting for the probability link



(a) All naics3



(b) Only naics3 with max prob. IPC-NAICS



(c) Only naics3 with max prob. > 0.5 IPC-NAICS

Figure 2.5 – Distribution of the probabilities linking IPC to NAICS

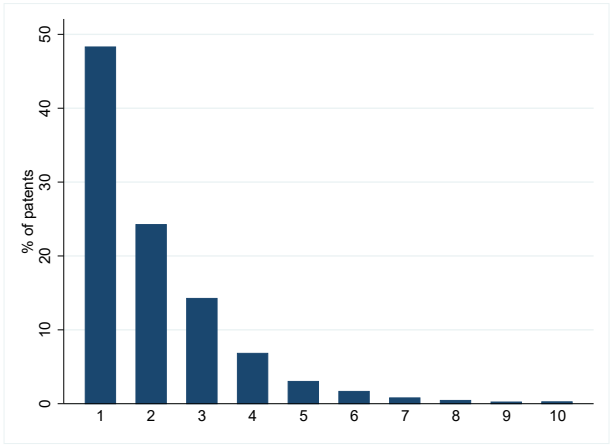


Figure 2.6 – % of patents across the number of co-inventors

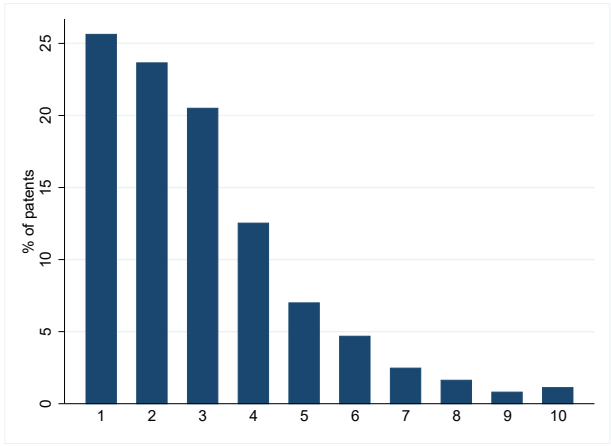


Figure 2.7 – % of patents across the of number naics with non zero probability

2.6.2 The Duranton and Overman methodology of coagglomeration

Global confidence bandwidths

Let $\hat{K}(d)$ be the estimated K-density of the bilateral distances between innovators of a given industry pairs, and $\bar{K}(d)$ its upper global confidence band. Following Duranton and Overman (2008), let consider that the upper band is hit by 5% of the 1000 simulations between 0 and 999 km. When $\hat{K}(d) > \bar{K}(d)$

for at least one d ε [0, 999], the innovation industry pair of are said to be colocalized (at 5% confidence level). For dispersion, let observe that if patents in an industry pair are very colocalized at short distances, they will likely show dispersion at larger distances. Hence, an industry pair will exhibit dispersion if the lower confidence band of the considered industry pairs $\underline{K}(d)$, is such that it is hit by 5% of the randomly generated K-densities that are not colocalized. Patents in that industry pairs are then said to exhibit dispersion (at 5% confidence level) when $\hat{K}(d) < \underline{K}(d)$ for at least one d ε [0, 999] and they do not exhibit localization. Dispersion is thus observed when there are fewer patents at short distances than what would be expected in a case of randomness.

Indices of localization and dispersion

The index of localization is therefore defined as :

$\Gamma(d)$ = max($\hat{K}(d)$ - $\bar{K}(d), 0$) (2.2)

The index of dispersion is defined as :

$$\Psi(d) = \begin{cases} max(\underline{K}(d) - \hat{K}(d), 0) & \text{if } \sum_{d=0}^{999} \Gamma(d) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

Graphically, the localization of patents in an industry pair is detected when the K-density lies above its upper global confidence bandwidth, and dispersion is detected when the K-density lies below the lower global confidence band and never lies above the upper global confidence band.

Table 2.13 – Coagglomeration patterns : Innovation Versus production across the years

	2001	2003	2005	2007	2009	2011	2013	2015	Average
Global colocalization - Innovation									
Number of industries pairs	111	93	98	118	112	130	120	114	112
% of industries pairs	65	54	57	69	65	76	70	67	65
$\Gamma\_i(x10^{-4})$	202	232	232	194	393	197	277	298	253
Global codispersion - Innovation									
Number of industries pairs	4	13	8	7	10	5	1	7	7
% of industries pairs	2	8	5	4	6	3	1	4	4
$\Psi\_i(x10^{-4})$	67	73	49	24	37	51	66	34	50
Random distribution - Innovation									
Number of industries pairs	56	65	65	46	49	36	50	50	52
% of industries pairs	33	38	38	27	29	21	29	29	30
Global colocalization - Production									
Number of industries pairs	124	89	101	135	90	81	70	110	100
% of industries pairs	73	52	59	79	53	47	41	64	58
$\Gamma\_p(x10^{-4})$	167	143	107	174	89	67	88	143	122
Global codispersion - Production									
Number of industries pairs	4	11	8	2	10	3	16	7	8
% of industries pairs	2	6	5	1	6	2	9	4	4
$\Psi\_p(x10^{-4})$	25	53	75	12	43	115	45	44	51
Random distribution - Production									
Number of industries pairs	43	71	62	34	71	87	85	54	63
% of industries pairs	25	42	36	20	42	51	50	32	37

Notes: This table reports statistics related to global localization and dispersion for both innovation and production. The scotts National Directories database is used to compute the figures for the production. The formula used to compute the indices are presented in appendix. The figures for the last column are averages across the years. The number of industry pairs for the production is 210 at 3-digit NAICS. For the innovation the number of industry pairs is 171 given that the industries 314-Textile mills and 323 Printing, support activities are removed

Table 2.14 – Classification of the NAICS 3-digits industry pairs in terms of innovation in 2015

NAICSCODE	NAICS311	NAICS312	NAICS313	NAICS315	NAICS316	NAICS321	NAICS322	NAICS324	NAICS325	NAICS326	NAICS327	NAICS331	NAICS332	NAICS333	NAICS334	NAICS335	NAICS336	NAICS337
NAICS312	Localized																	
NAICS313	Random	Random																
NAICS315	Localized	Localized	Random															
NAICS316	Random	Localized	Localized	Dispersed														
NAICS321	Random	Random	Random	Random	Localized													
NAICS322	Localized	Localized	Localized	Random	Localized	Random												
NAICS324	Dispersed	Localized	Localized	Localized	Localized	Localized	Dispersed											
NAICS325	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized										
NAICS326	Localized	Localized	Dispersed	Localized	Localized	Localized	Localized	Localized	Localized									
NAICS327	Localized	Localized	Random	Localized	Random	Localized	Localized	Localized	Localized	Localized								
NAICS331	Localized	Random	Random	Localized	Localized	Random	Localized	Localized	Localized	Localized	Random							
NAICS332	Localized	Localized	Random	Localized	Localized	Random	Localized	Random	Localized	Localized	Localized	Random						
NAICS333	Localized	Localized	Random	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized					
NAICS334	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized				
NAICS335	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized			
NAICS336	Localized	Localized	Random	Dispersed	Dispersed	Random	Localized	Random	Localized	Localized	Dispersed	Random	Localized	Localized	Localized	Localized		
NAICS337	Localized	Localized	Random	Dispersed	Random	Random	Localized	Localized	Localized	Localized	Dispersed	Localized	Dispersed	Localized	Localized	Localized	Dispersed	
NAICS339	Localized	Localized	Dispersed	Localized	Localized	Random	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized	Localized

Notes: This matrix reports the pattern of coagglomeration of industry pairs consisting of the industry in the row and in the column. The global indices of localization and dispersion are used to classify industry pairs as colocalized, Dispersed, or Randon. The pairs 323-314 are note classified since they only have a few observations to compute the comparison statistics.

Table 2.15 – Estimation excluding one NAIS-3 at the time, without industry fixed effects

	311	312	313	315	316	321	322	324	325	326	327	331	332	333	334	335	336	337	339
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv
Max input-Output - US	-0.097	-0.109	-0.005	-0.058	-0.098	-0.116	-0.106	-0.113	-0.152	-0.133	-0.103	-0.109	-0.129*	-0.093	-0.013	-0.109	-0.120	-0.127*	-0.118
	(0.076)	(0.067)	(0.070)	(0.053)	(0.076)	(0.073)	(0.069)	(0.069)	(0.097)	(0.083)	(0.070)	(0.087)	(0.076)	(0.082)	(0.059)	(0.076)	(0.074)	(0.077)	(0.078)
Max labor movement of highly educated	-0.001	0.006	-0.045	-0.027	-0.013	0.019	0.009	0.022	0.025	0.019	0.004	0.028	0.168	-0.084	0.046	-0.016	0.025	-0.006	-0.070
	(0.091)	(0.089)	(0.080)	(0.076)	(0.086)	(0.078)	(0.077)	(0.080)	(0.083)	(0.082)	(0.079)	(0.084)	(0.127)	(0.067)	(0.031)	(0.060)	(0.069)	(0.069)	(0.076)
Max knowledge spillover - IOM	0.271**	0.257**	0.268***	0.241**	0.263**	0.256**	0.233**	0.272**	0.283**	0.270**	0.255**	0.220**	0.208**	0.388***	0.028	0.252***	0.278***	0.256***	0.392***
	(0.112)	(0.111)	(0.102)	(0.106)	(0.114)	(0.114)	(0.113)	(0.119)	(0.122)	(0.112)	(0.113)	(0.112)	(0.104)	(0.111)	(0.033)	(0.090)	(0.093)	(0.096)	(0.111)
co-agglomeration of plants - cdf	0.356***	0.349***	0.361***	0.430***	0.363***	0.415***	0.360***	0.345***	0.332***	0.338***	0.366***	0.371***	0.351***	0.317***	0.249***	0.375***	0.351***	0.386***	0.381***
	(0.066)	(0.067)	(0.067)	(0.069)	(0.064)	(0.070)	(0.065)	(0.068)	(0.069)	(0.066)	(0.064)	(0.063)	(0.058)	(0.060)	(0.031)	(0.073)	(0.067)	(0.073)	(0.068)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224
Adj R^2	0.24	0.23	0.24	0.26	0.23	0.24	0.22	0.23	0.24	0.24	0.23	0.23	0.26	0.28	0.19	0.23	0.26	0.25	0.27

Notes: This table reproduces the main estimation. The industry specified in the column is excluded from the estimation and there is no industry fixed effects

Table 2.16 – Estimation excluding one NAIS-3 at the time, with industry fixed effects

	311	312	313	315	316	321	322	324	325	326	327	331	332	333	334	335	336	337	339
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv	Co-Agg.Inv
Max input-Output - US	-0.040	-0.044*	-0.022	-0.029	-0.037	-0.049**	-0.045**	-0.048**	-0.035	-0.033	-0.045**	-0.056**	-0.045*	-0.040	-0.048**	-0.047*	-0.044*	-0.051***	-0.044
	(0.028)	(0.023)	(0.034)	(0.025)	(0.026)	(0.019)	(0.018)	(0.019)	(0.032)	(0.028)	(0.021)	(0.022)	(0.025)	(0.026)	(0.021)	(0.025)	(0.024)	(0.018)	(0.029)
Max labor movement of highly educated	0.008	0.013	0.015	0.010	0.006	0.002	0.020	0.026	0.013	0.006	0.018	0.024	0.025	0.023	0.031	0.026	0.020	0.001	0.014
	(0.028)	(0.028)	(0.026)	(0.026)	(0.028)	(0.025)	(0.030)	(0.031)	(0.029)	(0.032)	(0.030)	(0.029)	(0.048)	(0.035)	(0.025)	(0.031)	(0.031)	(0.025)	(0.028)
Max knowledge spillover - IOM	0.058**	0.064**	0.060**	0.066**	0.067**	0.062**	0.053*	0.066**	0.044	0.064**	0.061*	0.055*	0.057*	0.046	0.074***	0.069**	0.064**	0.067**	0.078**
	(0.027)	(0.027)	(0.025)	(0.029)	(0.030)	(0.028)	(0.030)	(0.032)	(0.033)	(0.029)	(0.031)	(0.029)	(0.034)	(0.029)	(0.022)	(0.027)	(0.032)	(0.028)	(0.032)
co-agglomeration of plants - cdf	0.075*	0.033	0.060	0.089*	0.071*	0.099***	0.057	0.035	0.065	0.058	0.054	0.068	0.050	0.027	0.038	0.053	0.049	0.069*	0.057
	(0.043)	(0.044)	(0.052)	(0.050)	(0.042)	(0.037)	(0.039)	(0.042)	(0.043)	(0.042)	(0.042)	(0.042)	(0.043)	(0.038)	(0.035)	(0.043)	(0.043)	(0.041)	(0.038)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224	1224
Adj R^2	0.78	0.78	0.81	0.80	0.79	0.80	0.78	0.78	0.81	0.80	0.78	0.80	0.79	0.79	0.46	0.79	0.78	0.80	0.79

Notes: This table reproduces the main estimation. The industry specified in the column is excluded from the estimation and we add industry fixed effects

Table 2.17 – Regression of the coagglomeration of the production on the Marshallian forces

	(1) Co-Agg.Prod	(2) Co-Agg.Prod	(3) Co-Agg.Prod	(4) Co-Agg.Prod
Max input-Output - US	0.219***			0.145**
	(0.064)			(0.068)
Labor correlation - aggregated occupations		0.173**		0.120*
		(0.073)		(0.072)
Max knowledge spillover - IOM			0.239***	0.202***
			(0.062)	(0.062)
Time FE	Yes	Yes	Yes	Yes
Nb of industry pairs x year	1368	1368	1368	1368
Adj R^2	0.04	0.03	0.05	0.09

Notes: The multivariate model is estimated as specified in equation 2.2, where innovation is replaced by production. It is a pooled cross-section from 2001 to 2015 for the coagglomeration of the production. The Marshallian forces are averaged across the years for which data are available. The years 2001, 2005, and 2009 for the Labor pooling; The odd years from 2001 to 2013 for the input-output; and the years 1976 to 2006 for the knowledge spillover. For all the specifications, variables are standardized to have zero mean and unit standard deviation. Standard errors are clustered at the industry pairs level.

CHAPTER III

A COMPLEMENT TO THE TEST OF LOCALIZATION OF DURANTON AND OVERMAN (2005)

Abstract

We apply the methodology of Duranton and Overman (2005), using a new counterfactual to study the agglomeration patterns of Canadian manufacturing establishments. The new counterfactual which accounts more precisely for firms' possible location choices, better detects the departure from randomness, and generates substantial differences between the patterns that we uncovered compared to those obtained with the classical counterfactual. Our results suggest that localization and dispersion may be stronger than what is usually thought.

Keywords: Agglomeration, location, kernel density, counterfactual.

JEL Classification: R14, R15, L60

3.1 Introduction

A common feature of economic landscapes worldwide is geographic concentration, and understanding this tendency is essential in economics. Crucial to this process is the ability of economists to measure the extent of this concentration. In that wake, various endeavors of researchers to put numbers behind the degree of concentration have yielded interesting measures, starting from the simple Gini indices to more elaborated ones, such as the Ripley function (Marcon and Puech, 2003). One particular characteristic of these measures is that they do not have a meaning per se. They rather need to be compared to a benchmark to assess the extent of concentration of the reality that they intend to quantify. For example, the Gini index once computed is compared to the 45-degree line equivalent to a uniform distribution. The Herfindahl index is compared in many cases to questionable arbitrary thresholds. By doing so, these indices assume that an ideal world is one of uniform or arbitrary repartition. As such, many of these measures only capture unevenness. However, as well observed by Duranton and Overman (2005) unevenness does not necessarily mean concentration. For this reason, the recently proposed measures of concentration have suggested for comparison, a benchmark that accounts for what would have been in a total absence of tendency to agglomerate. For example, Ellison and Glaeser (1997), suggest a benchmark that controls for industrial concentration.

This paper moves in the same direction and suggests detecting localization on the basis of a better-informed benchmark. The paper builds on Duranton and Overman (2005, 2008) who propose an index, to assess the degree of localization of manufacturing in the United Kingdom. Their index is based on the estimation of the kernel density of the bilateral distances of pairs of industries.

In particular, to control for the overall agglomeration of manufacturing and the industrial concentration, they propose for each industry under investigation, a random draw among the possible locations as the hypothetical distribution of locations that they used to construct a counterfactual measure against which the observed measure is compared. However, their benchmark which is simple and straightforward opens new questions since it implies that a plant like "Bombardier Transport Canada" which enjoys a parcel of roughly "35,000 meters square", and employs more than 700 workers can switch locations with "El Mio Aliments Inc." a commercial bakery manufacturing of 6 employees located on a parcel of 250 meters square, in a residential neighborhood. In the real world, the former company will face space constraints at the location of the latter.[1] In addition, some zoning restrictions may exclude "Bombardier Transport Canada" from locating in a residential neighborhood whereas "El Mio Aliments Inc." may not face such constraint. Duranton and Overman (2008, p.236), themselves raised the issue since they acknowledge that this way of constructing the counterfactual assumes "*that the establishments, regardless of their size, face no restrictions on their location choice. The fact that large establishments require large sites to host them is, in practice, a binding constraint that prevents large manufacturing establishments from locating in many areas such as the central part of most cities, etc. These constraints may arise as a result of the workings of land markets (i.e., through prices) or as a result of government policy (e.g., zoning). These constraints could affect the results above and limit the opportunities for large establishments to cluster. More generally, the overall location patterns of industries could be affected by the availability of sites for their larger establishments*". To deal with the problem that many industries face constraints in their geographic location choice due to sites characteristics and zoning regulations, the authors refined their analysis by allowing

1. In the document, we use interchangeably the terms plant and establishment

establishments to be reallocated to a site occupied by an establishment in the same employment size class. They conclude that site size constraints do not appear to affect the tendency of U.K. manufacturing industries to localize or disperse.

While their conclusion is a good test of robustness for their results, it does not eliminate the necessity of accounting for location constraints in defining any relevant counterfactual, for at least two reasons. First, a plant's employment size may not necessarily be a good proxy for all the constraints that an establishment faces for its location choice. Second, there is substantial heterogeneity across industries in terms of plants location determinants. For example, on the land size, a plant with 250 employees in motor vehicle manufacturing will require a different square-footage than a plant with 250 employees in cut-and-sew clothing manufacturing. Finally, since the assessment of the departure from randomness entirely depends on the counterfactual, its definition is critical for any exercise of assessing agglomeration and thus deserves more attention.

We use the same metric than Duranton and Overman (2005) and propose a new way of constructing the counterfactual agglomeration against which the observed agglomeration should be compared for the detection of the departure from randomness. First, we start by selecting a set of variables suggested by the literature on firms' locations decisions, and we test their relevance in determining the locations of plants in Canada. Second, we use these variables as input in a clustering procedure to construct classes of locations that feature intra-class homogeneity and inter-class heterogeneity. Third, we use the output classes to construct the counterfactual agglomeration against which the observed agglomeration is compared. More precisely, while drawing a hypothetical location for a plant of a given industry, the process is constrained by the mere fact

that the new location should be of the same *Class* as the actual location of the plant.

Using this approach, we test for the departure from randomness of the geographic distribution of the manufacturing industries in Canada as recorded in the Scotts National Directories for the year 2017. We observe that accounting for firms' locations choice generates classes of locations that are quite polarized i.e, each *Class* is either dispersed or localized compared to the distribution of the universe of locations.[2] One important feature of such polarized classes is that they yield counterfactuals that are more precise in detecting departure from randomness. Indeed, on the agglomeration of employment for example, the approach with the new counterfactual identifies 81% of the industries as departing from randomness (52.4% of localization and 28.6% of dispersion), whereas the number is 66.7% for the unconstraint counterfactual (42.9% of localization and 23.8% of dispersion). Moreover, the new counterfactual suggests substantial changes in the classification for some industries, as nearly 20% of industries move from localized (respectively dispersed) with the classical counterfactual to dispersed (respectively localized) with our new counterfactual. Finally, localization happens with our new counterfactual at shorter distances compared to the classical counterfactual, while dispersion is more important at longer distances with the new counterfactual compared to the classical one.

The rest of the document is organized as follows: section 3.2 presents the manufacturing industries in Canada, section 3.3 presents the methodology of detecting agglomeration. Section 3.4 gives the results and section 3.5 concludes the document.

---

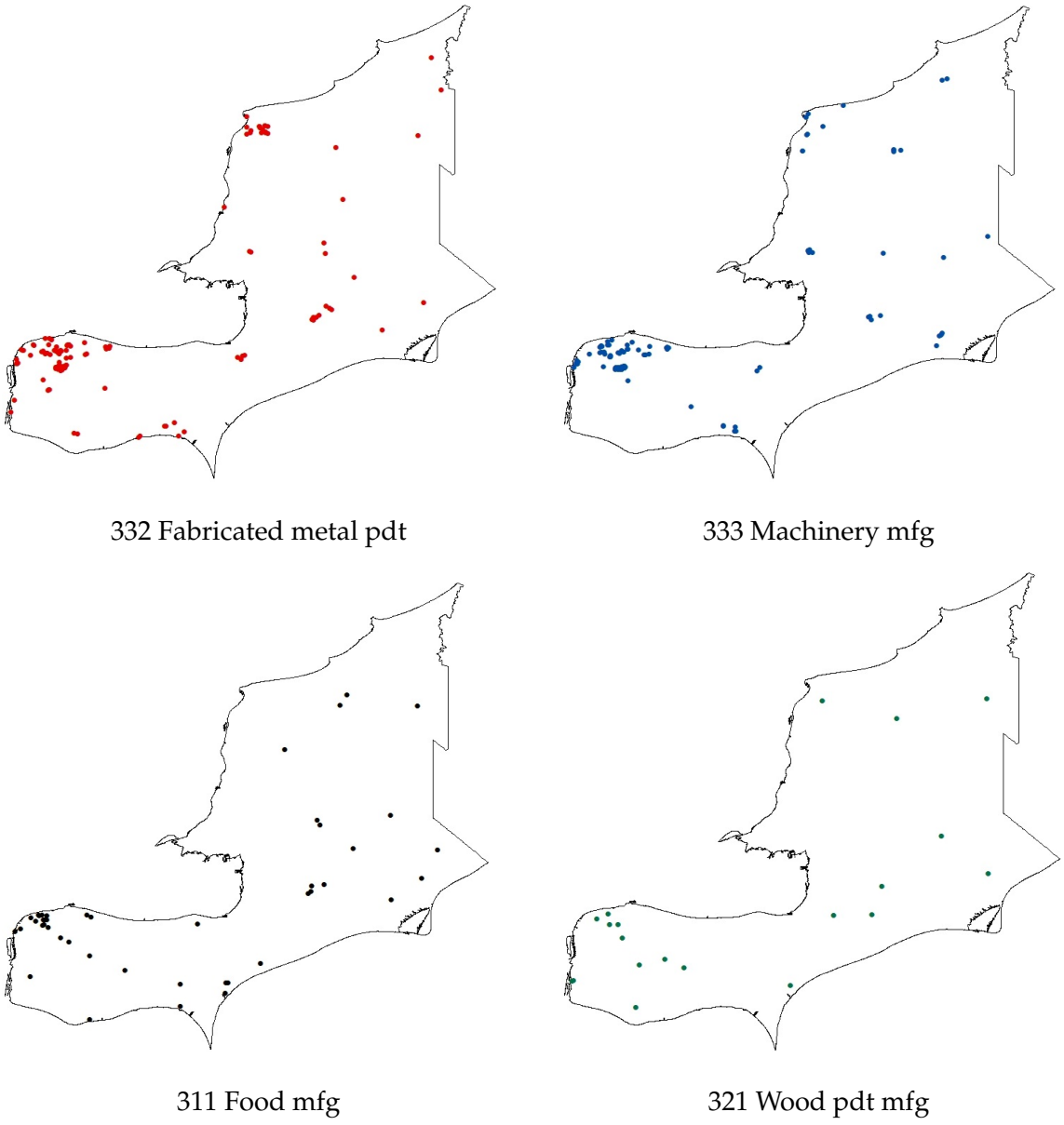
2. The universe of locations is what is usually used to construct previous counterfactuals.

3.2 Data

The main data source used in this paper is the Scott's National All Business Directories, a proprietary database that draws information on plants operating in Canada from Business Register records and telephone surveys. This dataset contains the address of each establishment including the postal code, the province, and the street number, as well as the year of the survey. It also contains the industrial classification (North American Industry Classification System, NAICS 6-digit level), the number of workers at the site, the business type of the activity of the plant, the headquarters, and the exports' status. From that database, we extract observations for manufacturing industries for the year 2017. Then, we use the addresses of establishments to retrieve the geographic coordinates of the location of each establishment, using a geocoding procedure (see appendix 1.7.5 for the geocoding). This first process is critical since our analysis relies on continuous metric constructed with these geographic coordinates. Thus, at the end of the geocoding process, we check for quality and remove observations with inaccurate coordinates. Next, we collect from statistics Canada, open-source data on polygons featuring the footprints of the buildings where plants produce and we use spatial join techniques to associate each plant to its building footprint (see appendix 1.7.3 for details on spatial join). These building footprints are used to proxy the size of the site where the plant is located. Finally, we also use other geographic datasets from Statistic Canada and DMTI that we merge with the Scott's dataset using Geographic Information System (GIS) tools to collect and construct for each establishment, the distance to its nearest junction, the zoning category of the location of the plant, the number of plants from the same industry in a radius of 10 km around the establishment, the total employment of the same industry in a radius of 10Km around the establishment, the population density in a radius of 1.5 Km, a dummy in-

dicating if the plant is located in a CMA or not.
We proceed with additional trimmings by removing observations with missing information and we also remove the 1% tails of the distribution of employment for each NAICS 3-digit industry. The reason for this trimming is that the variable is used as weight for the estimation of the agglomeration of employment, and inaccurate information can be critical in the process. At the end of this processing and cleaning stage, we end up with a dataset of 23,388 out of the 32,871 manufacturing plants registered in the Scott's National All Business Directories in 2017. More details on the construction of variables and the processing steps are presented in appendix 3.6.1.

Figure 3.1 (a)-(d) maps four selected NAICS 3-digit industries in the locality of Windsor. On the map, the industries "332 Fabricated metal product", and "333 Machinery manufacturing" look very localized and this is not surprising since Windsor is known for being the location of motor vehicle industries. On the contrary, the industries "311 Food manufacturing" and "321 Wood product manufacturing" are typical cases of dispersed industries.



332 Fabricated metal pdt

333 Machinery mfg

311 Food mfg

321 Wood pdt mfg

Figure 3.1 – Maps of four illustrative NAICS 3-digit industries in Windsor

3.3 Methodology

In this section, we present the methodology used, which builds on the approach of Duranton and Overman (2005, 2008). In this approach, the bilateral distances

observed for each pair of establishments in an industry are computed, and the "concentration" designates the density distribution of the observed bilateral distances. The exercise then consists in comparing the observed concentration of an industry to a counterfactual representing what we should observe following a random allocation of the plants across all the possible sites. To construct the counterfactual for a given industry, a random draw of sites is made and assigned to each establishment of the industry. Then, the process is repeated numerous times and confidence intervals are built. These confidence bands which feature the patterns of random allocations among the locations in the universe of possible sites are then used to test whether the observed distribution is significantly more or less concentrated.

In the original work by Duranton and Overman (2005) and subsequent works (e.g., Behrens and Bougna 2015), the set of locations that constitute the universe of possible sites for each plant is usually assumed to be the set of all locations with manufacturing plants, or a subset of those (e.g., all locations with plants in some industry; all locations with exporting plants, etc). Here, we include constraints on the universe of possible locations based on the determinants of plants' locations. For example, firms specialized in chemicals will not locate their productive plants in downtown Toronto because of some environmental and/or safety regulations. However, such firms may be allowed to locate their headquarters in a dense area without any restriction. Many other factors may constraint plants to locate or not in some places. As such, the construction of the counterfactual must account for these possible constraints.

In what follows, we present the approach that we suggest to detect localization. We start with the metric used to compute a measure of agglomeration, then we elaborate on the novel approach suggested for the construction of the counterfactuals.

3.3.1 Constructing the observed agglomeration

To measure the level of agglomeration of a given industry $\mathscr{I}$, the bilateral distances between all the pairs of plants of the industry are first computed. Then, the kernel density of these bilateral distances is estimated using the following formula :

$$\hat{K}(d) = \frac{1}{h \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e(i)e(j)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e(i)e(j) f(\frac{d - d_{ij}}{h}) \tag{3.1}$$

Where d_{ij} is the euclidean distance between the plants i and j, f is the kernel function with bandwith h, $e(i)$ and $e(j)$ are the weights used for the estimation. When the weights are equal to the number of workers of the plant i and j respectively, the estimated K-density is a measure of the concentration of the employment. When the weights are equal to one, the estimated K-density is a measure of the concentration of the establishments. In the rest of the document, this K-density of observed bilateral distance will be referred to as the "observed agglomeration".

3.3.2 Constructing counterfactuals

As mentioned earlier, while measuring concentration, the definition of the counterfactual is critical. Indeed, whether an industry is significantly concentrated or not is likely to hinge on the universe of possible sites we choose. Assume, for example, that we wish to measure the geographic concentration of industry $\mathscr{I}$. Denote by S the set of locations for which we observe plants in industry $\mathscr{I}$. The observed concentration is $\hat{K}_S(d)$ which is a function of the distance d. The theoretical reference distribution against which we want to judge that concentration

is the one observed if plants locate randomly across the set $\mathscr{S}$ of all sites that *could be chosen*. Consider a random draw of $S_j \subset \mathscr{S}$ sites, i.e., everything works as if all plants in the industry $\mathscr{I}$ make 'random' location choices within $\mathscr{S}$. We obtain a measured counterfactual concentration $\hat{K}_{S_j}(d)$. Repeating that process numerous times, e.g., $j = 1, 2, \ldots, 500$ draws, we can then judge how strongly the observed concentration $\hat{K}_S(d)$ deviates from that predicted by a random location process. More precisely, if we call $\overline{K}_S(d)$ the upper global confidence band and $\underline{K}_S(d)$ the lower confidence band. Following Duranton and Overman (2008), the upper band is such that it is hit by 5% of the 500 simulations between 0 and 999km, and the lower band is such that it is hit by 5% of the randomly generated K-densities that are not localized. Thus, when $K_S(d) > \overline{K}_S(d)$ for at least one d ε [0, 999], the industry under investigation is said to localized (at 5% confidence level). For dispersion, let observe that if an industry is localized at short distances, we should expect dispersion at larger distances. Hence, an industry will exhibit dispersion if when $K_S(d) < \underline{K}_S(d)$ for at least one d ε [0, 999] and they do not exhibit localization. Dispersion is thus observed when there are fewer establishments at short distances than what would be expected in a case of randomness.

The key problem with this approach is that we only observe the set $S \subset \mathscr{S}$ of actual choices but do not observe the set $\overline{S} = \mathscr{S} \setminus S$ of other possible choices. Hence, we need to determine the unobserved alternatives $\overline{S}$.[3] The results of the analysis will only tell us something interesting about the extent of geographic concentration of industry $\mathscr{I}$ if the set of unobserved potential sites, $\overline{S}$, is a plausible one. The initial idea in Duranton and Overman (2005) is to consider that $\mathscr{S}$ is the distribution of all sites with manufacturing plants. In subsequent work, they construct bins of employment and allow establishments to be re-

3. The set S always belongs to $\mathscr{S}$ by construction.

allocated only to sites occupied by an establishment in the same employment bins. By doing so, they intend to constraint each plant to switch only in the set of *possible sites*, assuming that the size of the site helps to identify these *possible sites*.

We extend and improve the work of Duranton and Overman (2008) in two ways. First, by accounting not only for the number of workers but also for many other factors that determine the location of plants. Second, by using a better-informed way of constructing clusters that allows building optimal bins based on one or many variables. The idea is to construct, bins that are heterogeneous across one another, but contain each, homogenous sites that are similar along the chosen variables. The procedure is realized in two major steps : the construction of the classes of locations, then the estimation of the counterfactual and the detection of localization. In what follows, unless explicitly stated, we use indifferently the words bins, categories, groups, or clusters to refer to the classes of locations, and the process of constructing the classes will be referred to as classification or clustering.

Constructing the classes of locations

To construct the classes of locations, the first step is to choose the variables that are relevant for plants' location. With these variables at hand, a procedure of classification is used to form the classes of locations.

i. Choosing the discriminant variables for the clustering

The relevant variables to use for the classification of locations are those that determine the choice of locations by plants. While choosing their locations, firms are more responsive to costs that they will be facing, as well as advantages that some places may offer. The theory has suggested that the location decision of individual firms is determined by factors such as market access and production costs (Fujita et al., 1999; Neary, 2001; Fujita and Thisse, 2002) Closely related to

these factors are the locations' characteristics, as well as plants' characteristics. On location characteristics, places, where agglomeration economies are at play, will attract more plants than others areas. In fact, agglomerated locations are places with higher production costs – land, labor, etc.-, but the gains stemming from the increasing returns outweigh these costs to generate more profitability for firms that locate in such places. Moreover, the type of agglomeration that is locally active will differently attract plants from different industries. In particular, localization economies will attract more plants from mature industries (Carlton, 1983; Autant-Bernard, 2006; Mota and Brandão, 2013) whereas modern and high tech firms will tend to locate where urbanization economies dominate (Henderson et al., 1995; Figueiredo et al., 2002; Head et al., 1995; Hansen, 1987). Besides these agglomerative forces which remain the main drivers of firm location decision, local policies may also influence the location decision of firms. The literature has shown that a high level of taxes is not attractive to firms (Coughlin and Segev, 2000; Basile et al., 2008; Head et al., 1999; McConnell and Schwab, 1990). However when local authorities spend high amounts on public goods and services, firms are willing to come even when these public expenditures are financed by an increase in local taxes (Gabe and Bell, 2004). Moreover, the quality of local institutions (Disdier and Mayer, 2004; Barrios et al., 2006), some promotional subsidies (Friedman et al., 1992), and the right to work (Schmenner et al., 1987; Klier and McMillen, 2008), also have an influence on the location decision of firms. In addition, firm characteristics may also interact with the characteristics of the location to increase or reduce the expected profitability of the firms (Schmenner et al., 1987; Blackley, 1985). These firms' characteristics include the establishment size (Arauzo Carod et al., 2010), the type of product, and the vocation of the establishment (Schmenner et al., 1987). Finally, local endowment such as the presence of highways is also important determinants of location choice (Klier and McMillen, 2008).

For the empirical part of this analysis, we use the population density around 1.5km of the location, as well as the size of the site measured by the footprint of the building where the plant is located as a proxy for land cost, the number of workers of the same industry around a 10 km radius of the location of the site as a measure of the local workforce. [4] The count of plants of the same industry is used as a measure of the localization economies. A dummy indicating whether or not a plant is located in a CMA will account for urbanization economies. The province and the zoning class of the site will capture the local policies, and the distance to the junction captures the locational fundamentals of the site. Finally, the exports status, the head-office status, the business type of the plant, as well as the NAICS industrial classification of the plant located on the site, will be the plants' characteristics.

To test whether or not these variables are relevant to construct the locations classes, their ability to predict the observed locations of the plants of a given industry is assessed in the spirit of Klier and McMillen (2008). But, instead of a conditional logit model as they did, we use a probit model. More precisely, let consider an industry $\mathscr{I}$, whose plants occupy $N_{\mathscr{I}}$ locations amount the N overall locations. A location dummy for the industry $\mathscr{I}$ is generated, with value 1 if the location hosts a plant belonging to the industry $\mathscr{I}$, and 0 if not. Then a probit model is estimated to explain the probability for a given location to host a plant from industry $\mathscr{I}$ by the location determinants suggested in the literature. Next, for the industry under scrutiny, the predicted probabilities across

4. Commuting distances can be quite dispersed across the country. According to statcan , "In 2016, Toronto had the greatest median distance at 10.5 km, followed by Ottawa?Gatineau at 9.2 km. The CMAs with the smallest median distance were Winnipeg (6.6 km) followed by Quc (7.5 km). For all eight CMAs, the highest proportions of commuters were travelling between 5 km and 14.9 km to get to work." see https://www150.statcan.gc.ca/n1/pub/75-006-x/2019001/article/00008-eng.htm

all the locations are estimated, and the locations with the $N_{\mathscr{I}}$ highest values are considered as the predicted locations for the industry $\mathscr{I}$ among the overall N locations.

The specification of the probit model is the following:

$$\begin{aligned}\text{Prob(Site k hosts a plant i } \varepsilon \text{ I)} \quad = \quad & \alpha_1 \text{EmploySameNaics10km}_k + \alpha_2 \text{NbPlantsSameNaics10km}_k \\ & + \alpha_3 \text{Density}_k + \alpha_4 \text{LandSize}_k + \alpha_5 \text{DistJunction}_k + \text{CMA}_k \\ & + \text{Province}_k + \text{Zoning}_k + \text{BusinessType}_{ki} + \text{Employees}_{ki} \\ & + \text{Exports}_{ki} + \text{Headquarter}_{ki} + \varepsilon_{ki} \end{aligned} \tag{3.3}$$

Where $EmploySameNaics10km_k$ is the total employment of the same industry 10km around site k, $NbPlantsSameNaics10km_k$ is the number of plants of the same industry 10km around the site k, $Density_k$ is the population density 1.5km around site k, $LandSize_k$ is the footprint size of the building grounded on site k, $DistJunction_k$ is the distance from the site k to the nearest junction, CMA_k is a dummy indicating whether or not site k is located in a Census Metropolitan area, $Province_k$ is the province where site k is located, $Zoning_k$ is the zoning type of the location of site k, $BusinessType_{ki}$ is the type of business conducted by the plant i which is actually on the site k, $Employees_{ki}$ is the number of workers of the plants i located on site k , $Exports_{ki}$, and $Headquarter_{ki}$ are respectively the headquarter and the export status of plant i actually located on site k. The model is estimated for each industry.

Table 3.1 presents the estimates of the marginal effects on the probability for a site to host a plant from a given industry. We illustrate the results with four industries and relegate the estimates for the other industries in appendix 3.7 and 3.8. As can be seen, variables do not affect the probabilities in the same way. For example, dense places have a higher probability to host plants from the "311 food manufacturing" whereas, they have a lower probability to host

plants from the "321 Wood manufacturing", "332 Fabricated metal manufacturing" or "333 Machinery manufacturing". On the footprint measure, locations with large footprints are more likely to host plants from the "311 Food" which usually uses indoor space, and having a large footprint affects negatively the probability of a location to host a plant from the "333 Machinery manufacturing". For the number of workers, the results indicate that a site where we observe a plant with a large number of workers has a higher probability to host a plant from the "311 Food manufacturing" or "321 Wood manufacturing", but a lower probability to host a plant from "333 Machinery" but these effects are hardly significant.

Table 3.1 – Marginal effects on the probability of location choice : four selected industries

	311 Food	321 Wood	332 Fabr metal	333 Machinery
main				
Density 1.5 km around	0.0000861***	-0.0000425***	-0.0000604***	-0.0000917***
	(0.00000746)	(0.0000117)	(0.00000712)	(0.00000828)
Building footprint size	0.0000243***	-0.0000187***	-0.00000742*	-0.0000327***
	(0.00000396)	(0.00000639)	(0.00000382)	(0.00000462)
Nb of workers of the plant	0.000168*	0.000110*	-0.000205	-0.000191*
	(0.0000864)	(0.0000657)	(0.000165)	(0.000108)
Employment, 10km around same naics	0.000961***	-0.0000245	-0.0000919***	0.0000116
	(0.0000362)	(0.0000279)	(0.0000148)	(0.0000154)
Nb plants, 10km around same naics	-0.0637***	-0.00338***	-0.0000705	0.00699***
	(0.00238)	(0.000968)	(0.000551)	(0.000582)
Distance to nearest junction	0.00499***	0.00135	-0.00394***	0.000906
	(0.00113)	(0.00128)	(0.00141)	(0.00122)
Business type, Distributor	0.184**	-0.250***	-0.302***	0.367***
	(0.0747)	(0.0823)	(0.0506)	(0.0503)
Business type, Manufacturer	0.296***	0.0391	-0.129***	-0.143***
	(0.0601)	(0.0590)	(0.0382)	(0.0424)
Industrial, commercial, institutional	0.00524	-0.179***	0.0711***	-0.174***
	(0.0345)	(0.0372)	(0.0266)	(0.0289)
Residential	-0.0493	-0.204***	-0.0755**	-0.0764**
	(0.0359)	(0.0405)	(0.0302)	(0.0333)
Exporter	-0.338***	-0.0753**	-0.0785***	0.527***
	(0.0282)	(0.0310)	(0.0210)	(0.0231)
Headquarter	0.0213	-0.000764	-0.134***	0.0316
	(0.0446)	(0.0503)	(0.0362)	(0.0359)
Prince Edward Islands	0.580***	-0.345*	-0.565***	-0.141
	(0.156)	(0.200)	(0.208)	(0.199)
New Scotia	0.303***	-0.262***	-0.178**	-0.296***
	(0.0785)	(0.0900)	(0.0766)	(0.0906)
New Brunswick	0.0229	-0.106	-0.228**	-0.0663
	(0.0972)	(0.0976)	(0.0894)	(0.0935)
Quebec	-0.0478	-0.288***	0.0595*	0.0101
	(0.0453)	(0.0451)	(0.0355)	(0.0392)
Ontario	-0.00633	-0.372***	0.0969***	-0.0774**
	(0.0435)	(0.0430)	(0.0333)	(0.0369)
Manitoba	0.0745	-0.402***	-0.0278	0.138*
	(0.0843)	(0.106)	(0.0722)	(0.0730)
Saskatchewan	0.0326	-0.546***	-0.0340	0.136*
	(0.0912)	(0.120)	(0.0798)	(0.0818)
Alberta	-0.0512	-0.342***	0.0800*	0.209***
	(0.0601)	(0.0626)	(0.0446)	(0.0468)
Census Metropolitan Area	-0.179***	-0.276***	0.162***	0.0551
	(0.0412)	(0.0458)	(0.0384)	(0.0432)
Observations	23388	23388	23388	23388
Pseudo R^2	0.139	0.057	0.022	0.114

Notes: This table reports the estimates of the marginal effects at the means on the probability for a site to host a plant from each of these industries from a probit regression. The three territories and the Newfoundland are removed due to very few observations and poor data quality. The reference categories are the following : Zoning(Industrial/commercial/institutional; Residential; ref=others), Business type=others, province(ref=British columbia). The Significance levels 0.10 * 0.05 ** 0.01 ***.

We now assess the goodness of the fit of our model to predict the locations of each industry. To that end, we first use the fitted model to predict the probabilities that each location hosts a plant of a given industry. Then, the highest predicted probabilities values, up to the total number of plants of the industry under investigation, are kept as the predicted locations for that industry. The observed locations are then compared to the predicted locations and the share of locations predicted as hosts of the plants from the industry, and which are actually hosts of plants from that industry is computed (henceforth Well predicted locations). Table 3.2 summarises these shares for each NAICS 3-digit industry.

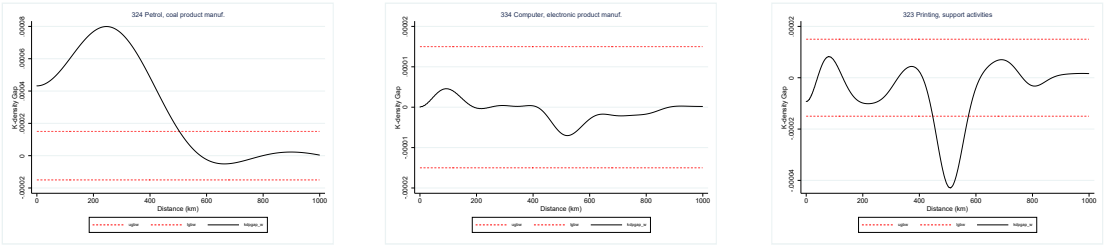
Table 3.2 – Shares of well predicted location across the NAICS 3-digit industries

	Well predicted locations
311 Food manufacturing	25.7
312 Beverage and tobacco product manuf.	8.3
313 Textile mills	5.6
314 Textile product mills	5.7
315 Clothing manufacturing	11.2
316 Leather, allied product manuf.	10.5
321 Wood product manufacturing	14.5
322 Paper manufacturing	12.1
323 Printing, support activities	46.3
324 Petrol, coal product manuf.	3.6
325 Chemical manufacturing	15.7
326 Plastics, rubber products manuf.	32.5
327 Non-metallic mineral product manuf.	13.0
331 Primary metal manufacturing	6.2
332 Fabricated metal product manuf.	22.7
333 Machinery manufacturing	37.2
334 Computer, electronic product manuf.	12.6
335 Electrical, appliance manuf.	7.0
336 Transportation equipment manuf.	17.1
337 Furniture, related product manuf.	9.6
339 Miscellaneous manufacturing	29.7

Notes: This table reports for each industry, the shares of sites hosting plants from that industry, and for which the model specified by equation 3.2 has predicted that there is a plant of that industry.

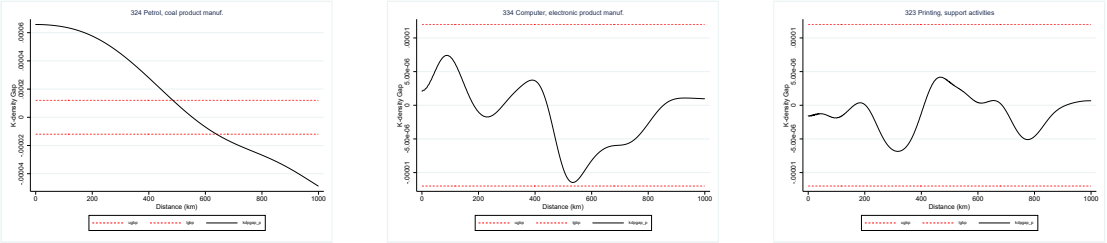
The results indicate that on the one hand, some industries have particularly good shares of *well-predicted location*. These industries are "323 Printing, support activities" (46.3%), "333 Machinery manufacturing" (37.2%), "332 Fabricated metal manufacturing" (22.7%), "311 Food manufacturing" (25.7%), and "326 Plastics, rubber product manufacturing" (32.5%). On the other hand, other industries exhibit low values of the shares of *well-predicted location*. These industries are "313 Textile mills" (5.6%), "314 Textile product mills" (5.7%), "312 Beverage and tobacco product manufacturing" (8.3%), "324 petrol, coal product manufacturing" (3.6%), "335 Electrical appliance manufacturing" (7.0%), and "331 Primary metal manufacturing" (6.2%).

A step further we compare the K-density estimated from the observed locations to that of the predicted location. We compute and plot the differences between the K-density of the observed distribution and the predicted one. We exemplify the results with three industries : "324 Petrol, coal product manufacturing", "334 Computer, electronic product manufacturing", and "323 Printing, support activities manufacturing" which are the three industries at the bottom, the median and the top of the shares of *well-predicted locations* in Table 3.2.



(a) 324 Petrol(3%) (b) 334 Computer(12.2%) (c) 323 Printing(45.2%)

Figure 3.2 – K-Densities of employment, predicted versus observed distributions



(a) 324 Petrol(3%) (b) 334 Computer(12.2%) (c) 323 Printing(45.2%)

Figure 3.3 – K-Densities of establishment, predicted versus observed distributions

The graphs on Figure 3.2 are for the employment, and those on Figure 3.3 are for the plants. The solid black line is the difference between the K-density of the observed distribution and the predicted one. The red dashed lines are bandwidths representing $+/-0.5\%$ of the maximum values of K-density of the observed distribution. Numbers in parenthesis with the labels are the shares of *Well predicted locations* for the considered industry.

Wrapping up, the above results validate the choice of some location characteristics (localization economies, local workforce, the density, the size of the site, the minimal distance to a junction, the CMA status of the location, the province, and the zoning categories), some plants characteristics (NAICS, the number of workers, the business Type of the plant, the export and the headquarter status) as relevant variables to use for the classification of the locations. These variables yield acceptable predictions of the locations of industries and do not affect the probability of location in the same way across the industries.

ii. Constructing the classes of locations

With these determinants of locations choice, we now move to the construction of the classes of locations using the Hierarchical Ascendant Classification (HAC) procedure. The exercise consists in organizing a set of N points into

groups such that, each group contains homogenous elements, but the groups are heterogeneous from one to another. We present here the procedure in simple terms and we relegate the details to appendix 3.6.6. The general scheme of this clustering procedure is as follows:

— Step 1: We start with N points representing all the locations of the plants. These initial N points are considered as original nodes.

— Step 2: We compute a pairwise dissimilarity measure between all the nodes. To compute the dissimilarity measure, all the variables should be numeric. So, categorical variables are transformed into dummies. Then, euclidean distance is used to compute the pairwise dissimilarity [5]

— Step 3: We identify the pair of nodes with minimal distances among all pairwise distances.

— Step 4: We join the two nodes with minimal distance into a new unique node and remove the two old nodes. The new nodes are labeled consecutively N, N + 1, . . . and constitute the first level of aggregation

— Step 5: We then, repeat N - 1 times the process from step 2, until there is one big node, which contains all the N original input points.

The output of this process can be represented as a tree that gives at each step the loss of information that occurs during the aggregation process. Figure 3.4 gives an illustrative classification tree generated by the process presented above with only 20 randomly selected plants to exemplify the results of the procedure. The numbers at the end of the branches are the plants' identifiers as recorded in the dataset. The actual trees for the construction of classes are presented in the appendix 3.6.3.

---

5. Various other metrics can be used to compute the dissimilarity measure between variables

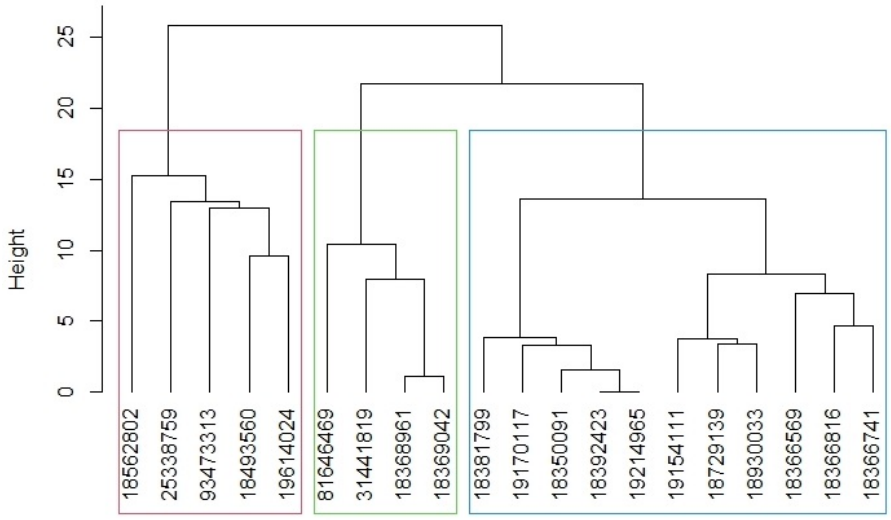


Figure 3.4 – Illustrative tree from a Hierarchical Ascendant Classification (HAC)

iii. Chosing the optimal number of classes

With this procedure, it is theoretically possible to classify the locations into 1, 2, 3,..., N classes. The choice of an optimal number of classes is based on a tradeoff between a parsimonious number of classes and a loss of information. At each stage of the clustering process, while aggregating the points, there is a loss of information that differentiates the locations. In simple words, the bottom of the tree corresponds to 20 individual classes with 0% of the discriminant power lost, and the top of the tree corresponds to one unique class with 100% of the discriminant power lost. An intermediate case is for example a horizontal line at a height of about 18 which indicates a choice of 3 classes as represented in Figure 3.4. The rectangles on the figure show the locations contained in each class. The optimal choice is therefore few classes with the maximum discriminant power.

In connection to our analysis, the extreme case with one class is the case of maximum within-class heterogeneity where all the locations are considered alike.

The opposite extreme case is the case with N classes where all the discriminant information is accounted for and no plant is allowed to switch from its original location because it is different from the other locations. In practice, to choose an optimal number of classes to consider, one way is to observe the classification tree from the top to the bottom. While moving from the top (One unique class) to the bottom (N distinct classes), the discriminant information increases (Or the loss of discriminant information decreases) and so does the number of classes. An optimal cutoff is a place where the increase in the number of classes is not worth the increase in the information. Some statistics also help to obtain the optimal number of classes to consider. These statistics are based on the within-class sum of squares. The optimal number of classes is obtained at a cutoff such that, increasing the number of classes does not decrease the within-class sum of squares.

We conduct the classification procedure for the 23,388 plants recorded in our dataset, using as discriminant variables the same as in model 3.2 and we also add the NAICS 3-digit code of the plant. Figure 3.5 represents the graph of the value of the gap of the within-class sum of squares, which guides the choice of an optimal number of classes. Only the variation of the gap statistic up to the first 25 classes is represented for visibility. Details on the computation of the within-class sum of squares are presented in Appendix 3.6.6. The curve suggests considering 6 classes. Hence, we chose 6 classes for the main results and we also provide results for four alternative options (4, 5, 7 and 8 classes) for robustness checks. The trees generated by the process for the entire dataset of 23,388 plants are presented in appendix 3.6.3.

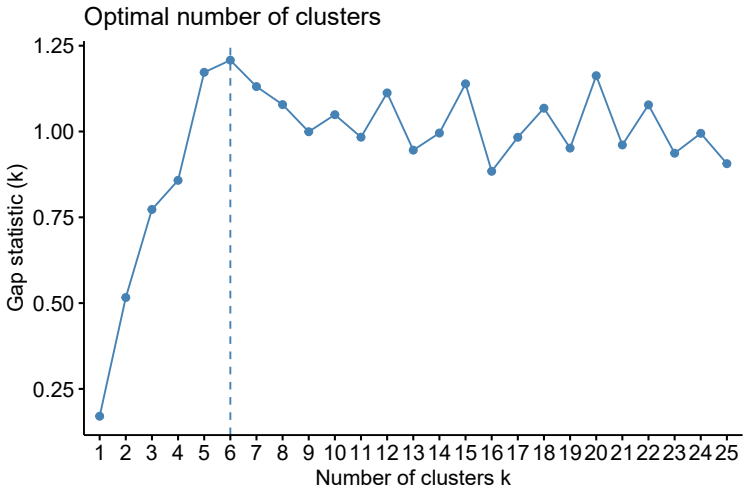


Figure 3.5 – Gap statistics curve

Table 3.3 gives descriptive statistics for some variables used in the clustering process, and the distributions of the numbers of plants, and the number of workers across the 6 Classes are provided in Table 3.9 in appendix. We use these statistics to characterize the six Classes generated by the clustering procedure.

iv. Characterizing the classes of locations- 3 digit

Table 3.3 – Characteristics of the classes of locations

	Class 1		Class 2		Class 3		Class 4		Class 5		Class 6	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Nb employees	37.2	86.1	67.1	323.0	41.6	79.1	24.3	39.7	16.7	34.0	17.6	42.3
Pop. density, 1.5km	1,271.1	1,897.3	1,135.2	1,571.2	1,403.8	1,821.0	1,273.5	1,608.4	2,873.5	3,640.4	2,074.7	2,366.8
Site's size(m2)	1,739.4	2,770.1	2,588.9	3,877.1	2,011.3	2,826.6	1,807.4	2,151.1	1,324.1	1,793.3	1,365.3	1,777.6
Nb workers, same naics, 10km	334.5	631.1	2,003.9	2,519.7	453.4	706.2	750.9	945.4	1,145.1	1,307.6	1,153.9	1,138.6
Nb plants, same naics, 10km	8.5	16.2	49.5	63.8	10.0	14.7	22.5	26.8	44.6	43.4	52.4	47.3
Dist.to junction	7.1	16.6	2.6	4.0	3.6	6.0	2.6	3.7	2.8	3.6	3.0	4.3
Share commercial sites	0.4	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5
Share residential sites	0.3	0.5	0.2	0.4	0.3	0.4	0.2	0.4	0.5	0.5	0.4	0.5
Share in CMA	0.7	0.4	1.0	0.2	0.9	0.3	1.0	0.1	1.0	0.2	0.9	0.2

Notes: The number of locations are as follows, Class 1: 5,322 ; Class 2: 6,293 ; Class 3: 7,052 locations, Class 4: 3,139 ; Class 5: 1,595

It appears that the locations from *Class 1*, that we label *"Sites in less agglomerated and moderately dense locations"* host plants that are surrounded by a few other plants from the same industry and a few workers from the same industry. These locations enjoy very few localization economies.

The locations from *Class 2*, that we label *"Sites in highly agglomerated and moderately dense locations"* are moderately dense places, which benefit from a high level of localization, and are essentially hosts of three industries: "326 Plastics, rubber products manufacturing", "332 Fabricated metal product manufacturing", and "333 Machinery manufacturing".

The locations from *Class 3*, and *Class 4* share similar features. They can be labeled *"Sites in less agglomerated and moderately dense locations, near main junctions"*. They differ in the nature of the industries that they host. *Class 3* hosts six industries "311 food manufacturing", "321 Wood product manufacturing", "325 Chemical manufacturing", "327 Non-metallic mineral product manufacturing", "334 Computer and electronic" and "337 Furniture, related product manufacturing", whereas *Class 4* hosts the majority of plants from "332 Fabricated metal product manufacturing".

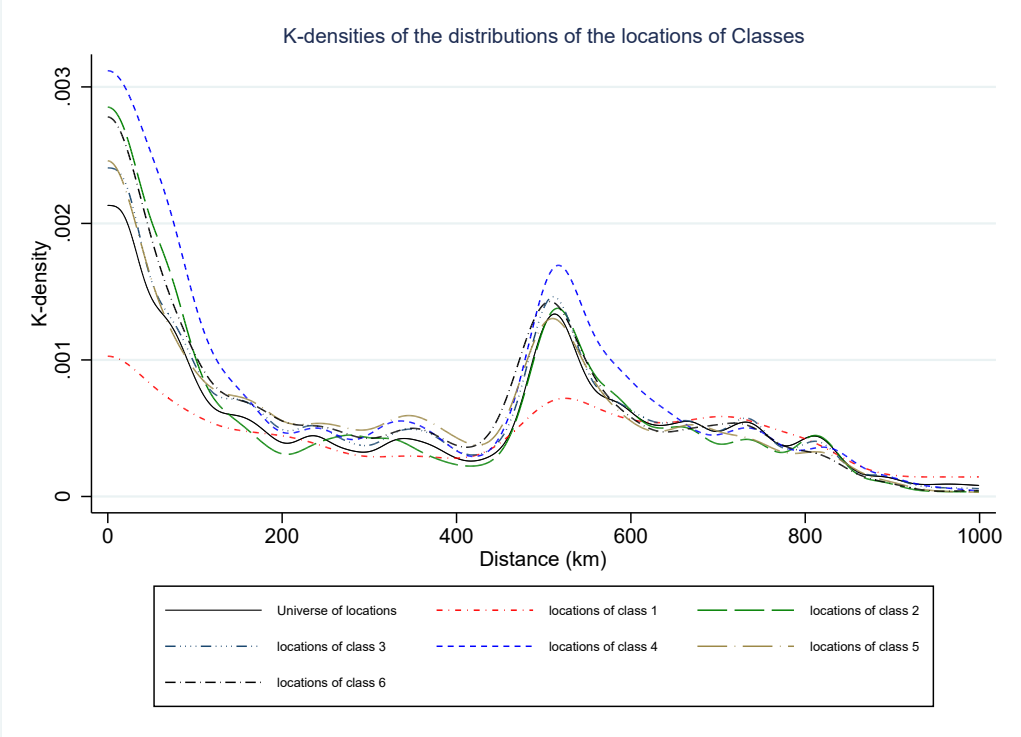
The locations from *Class 5*, that we label *"Sites of Miscellaneous manufacturing, in moderately agglomerated, and highly dense locations"* are places of high density in their neighborhood, they benefit from moderate localization economies. These locations are typically hosts for "339 "Miscellaneous manufacturing".

Finally, *Class 6*, that we label *"Sites of printing industries, in moderately agglomerated and highly dense locations"* hosts almost all the printing industries. They are dense locations with moderate agglomeration.

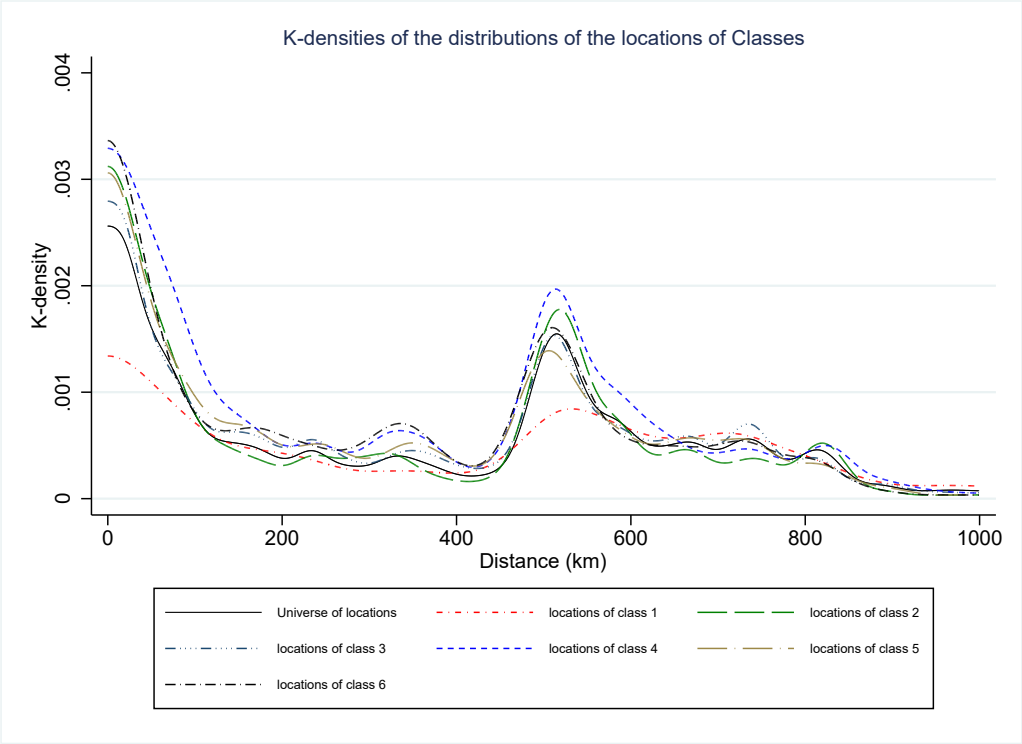
The characteristics of the classes in terms of the number of workers are a bit different for the employment. Classes 3, 4, 5, and 6 look specific to some in-

dustries whereas Classes 1 and 2 encompass a large variety of industries. In addition, *Class 2* appears to have more employees per plant compared to *Class 1*. All these differences will show up in the counterfactuals of the employment, compared to those of the establishments.

We compare the distributions of the locations of each of these classes to the distribution of the universe of locations. For each of these six classes, we estimate the K-densities of their distributions for both employment and establishment. We also estimate the distribution of the locations universe for both the distributions of establishments and employment. The graphs are presented in Figure 3.6.



(a) Establishments



(b) Employment

Figure 3.6 – K-densities of the locations of the 6 Classes and the universe

As can be observed, the K-densities of the locations of the six classes generated by the clustering procedure are quite "polarized", (ie. above the K-density of the universe of locations at each distance or below it at each distance). Indeed, the K-density of the distribution of the locations of *Class 1* is below the K-density of the distribution of the location universe at almost all the distances. Conversely, the K-density of the distribution of the locations of the other five *Classes 2, 3, 4, 5, and 6* are above the K-density of the distribution of the location universe at almost all the distances. This polarized feature holds for the distributions of both employment and establishments.

Furthermore, we proceed with a test of localization for each of the 6 classes

to assess whether their distributions of the locations depart from randomness compared to the distribution of the universe of the locations. The graphs of these tests provided in the appendix (see Figures 3.16 and 3.17) indicate that, compared to the distribution of the universe of the locations, the distribution of the locations in *Class 1* is dispersed and those in the others classes are localized.

Intuitively, a counterfactual based on a more dispersed distribution of locations should generate a more agglomerated pattern and the other way round. As such, an industry with a large share of plants/employees belonging to classes of dispersed (respectively localized) locations, is likely to be more agglomerated (respectively dispersed) with a test of localization based on a multivariate counterfactual, compared to a test based on an unconstrained counterfactual. For example, we should expect the industry "323 Printing, support activities" whose plants/employees are almost entirely in *Class 5* to be less localized with the multivariate counterfactual compared to the unconstraint counterfactual. Conversely, we should expect an industry such as "335 Electric, appliance manufacturing" which is essentially in *Class 1* - in terms of the number of plants and number of workers-, to be more localized compared to its test with an unconstraint counterfactual.

Globally, more than 50% of the industries have plants belonging almost entirely to *Class 1*, and the remaining industries have different distributions of plants/workers across the 5 other classes. Thus, the global picture with the new counterfactual is likely to be more localized and more dispersed than the picture with an unconstraint counterfactual. From an employment perspective, the feature is the same and we should also expect more localization and more dispersion for employment.

Estimating counterfactuals and detecting localization

With these classes of locations, we now construct for each industry a counterfactual agglomeration as follows: The plants of the industry are randomly reallocated across all the universe of locations and each plant is constrained to be relocated only on a site of the same *Class*.[6] This means that plants are to be assigned only on sites similar to their original location. Then, the bilateral distances of all the pairs of these hypothetical locations of the industry under scrutiny are computed anew, and the k-density of these distances are estimated with the metric specified by the formula 3.1. By repeating the process 500 times we can build local and global confidence bands used to detect localization, dispersion, or randomness. An industry is said to be globally localized if its "observed agglomeration" is above the upper bound of the global bandwidth for at least one distance. On the contrary, an industry is said to be globally dispersed if its "observed agglomeration" is below the lower bound of the global bandwidth for at least one distance and never above the upper bound of the global bandwidth for any distance. Finally, an industry is considered as not departing from randomness if it lays between the upper bound and the lower bound of the global bandwidth for all the distances.

3.4 The patterns of the agglomeration of industries

This section presents and compares the results for different approaches of constructing the counterfactuals used to detect the departure from randomness. The first approach uses as counterfactual the case with no constraint while

6. We do no constraint the locations where plants are randomly reallocated to be equal to the number of plants of the industry under scrutiny as we did for the assessment exercice with the probit model

reshuffling the plants across the universe of all the locations (henceforth "unconstrained counterfactual" or "benchmark" for short). The second approach uses for comparison, the counterfactual built by constraining the reshuffling of plants in five employment size classes as in Duranton and Overman (2008). These classes are explicitly 1-4, 5-19, 20-49, 50-250, 250+ (henceforth "Job counterfactual DO"). The third approach uses a counterfactual built by constraining the reshuffling of plants in classes obtained by the clustering procedure presented above, with employment as the only discriminant variable (henceforth "Job counterfactual HAC"). This third approach will help inform the extent to which the choice of bins for employment may change or not the final classification. Finally, the fourth approach uses classes constructed using all the determinants of plants location presented and discussed in the section 3.3.2 (henceforth "Multivariate counterfactual HAC").

3.4.1 Departure from randomness

For the NAICS 3-digit manufacturing industries, the clustering procedure has suggested 6 classes as observed above. For the "Job counterfactual HAC" where only the number of workers is the unique variable used for the classification, the optimal number of classes suggested by the procedure is 5. We first estimate the observed K-density, then we estimate the counterfactual represented by the confidence intervals for each of the four approaches: benchmark, "Job counterfactual DO", "Job counterfactual HAC", and "Multivariate counterfactual". The results are presented for both the concentration of establishments and that of employment. In what follows, we use the term "identification" to refer to the process of categorizing an industry as "localized", "dispersed" or "random".

Table 3.4 – Localization, dispersion and randomness, NAICS 3-digit industries

			Localized			**Dispersed**		**random**
Panel a : establishments	nb	%	Γ	nb	%	Ψ	nb	%
Benchmark, unconstrained	16	76.2	0.0253	2	9.5	0.0107	3	14.3
Job counterfactual, DO	16	76.2	0.0236	2	9.5	0.0083	3	14.3
Job counterfactual, HAC	16	76.2	0.0232	3	14.3	0.0044	2	9.5
Multivariate counterfactual	19	90.5	0.0594	1	4.8	0.0010	1	4.8
Panel b : employment	nb	%	Γ	nb	%	Ψ	nb	%
Benchmark, unconstrained	10	47.6	0.0210	4	19.0	0.0089	7	33.3
Job counterfactual, DO	9	42.9	0.0187	4	19.0	0.0127	8	38.1
Job counterfactual, HAC	9	42.9	0.0191	5	23.8	0.0106	7	33.3
Multivariate counterfactual	12	57.1	0.0702	6	28.6	0.0127	3	14.3

Notes: This table presents summary statistics on the test of localization resulting from different approaches of constructing counterfactual. Γ represents the average across the industries, of the Global index of localization, and Ψ is the average across the industries, of the Global index of dispersion. See appendix 2.6.2 for their computation. The total number of industries at 3 digit is 21.

Table 3.4-Panel (a)- gives for the establishments, the shares of localized, dispersed, and random industries for each of the four approaches used to construct the counterfactual. We also provide, two measures of the magnitudes of the degree of localization and dispersion : the global index of localization Γ and the global index of dispersion Ψ. The formulas for the computations of these indexes are provided in appendix 2.6.2. The figures in the table indicate that localization is the most important pattern for all four approaches.

We first discuss whether the clustering procedure adds some value to the testing process. To that end, we only compare the results of the "Job counterfactual DO" to those of the "Job counterfactual HAC". As can be observed, when the classes are generated from the Clustering procedure ("Job counterfactual HAC"), there is less randomness for both the agglomeration of the establishments (9.5% against 14.3% for the "Job counterfactual DO"), than for the agglomeration of employment (33.3% against 38.1% for the "Job counterfactual

DO"). This suggests that the way of constructing the classes of locations may have an incidence on the test of localization.

We turn now to the discussion on the differences between the benchmark and the multivariate counterfactual. Compare to the benchmark, it appears that the multivariate counterfactual generates more departure from randomness than all the three other approaches. Only 4.8% of industries on the establishment viewpoint exhibits randomness, and the number is 14.3% on the employment perspective. This may indicate that using the determinants of firms' locations choice to construct the counterfactual may be more efficient in detecting departure from randomness when testing for localization.

On the magnitudes, the constraint imposed by the classes while accounting for location choice yields narrower confidence bands that help to detect departure from randomness (see graphs in appendix 2.1). As a result, the excess of localization or dispersion will tend to be more important in magnitude compared to the benchmark. Indeed, for the establishments (see Table 3.4-Panel (a)) the average index of global localization Γ is much higher for the multivariate counterfactual (0.0594) than for the benchmark (0.0253). For the average index of global dispersion Ψ, the value is 0.001 for the multivariate counterfactual against 0.0107 for the benchmark.[7] For the employment, (see Table 3.4-Panel (b)) the two indices are higher for the multivariate counterfactual. The average index of global localization is 0.0702 against 0.0210 for the benchmark, and the average index of global dispersion Ψ is 0.0127 against 0.0089 for the benchmark.

7. This value is an average across all the dispersed industries. The average here for the multivariate may be lower than the average for the benchmark due to the number of observations involved in the computation of these averages for the dispersed industries

Table 3.5 – Switches from the benchmark to other approaches, NAICS 3-digit industries

	Switches from the Benchmark to other counterfactuals				
Panel b: Agglomeration of establishments					
	One step up	One step down	Two steps up	Two steps down	Unchanged
Benchmark to Job DO	0.0%	0.0%	0.0%	0.0%	100.0%
Benchmark to Job HAC	0.0%	4.8%	0.0%	0.0%	95.2%
Benchmark to Multivariate	9.5%	0.0%	9.5%	4.8%	76.2%
Panel a : Agglomeration of employment					
	One step up	One step down	Two steps up	Two steps down	Unchanged
Benchmark to Job DO	4.8%	9.5%	0.0%	0.0%	85.7%
Benchmark to Job HAC	0.0%	9.5%	0.0%	0.0%	90.5%
Benchmark to Multivariate	14.3%	14.3%	4.8%	4.8%	61.9%

Notes: This table presents the shares of changes of categorization from the benchmark to another approach.

Table 3.5- Panel a- gives statistics on the switches that occur from the benchmark counterfactual to other approaches of constructing the counterfactual for the agglomeration of the employment. In the table, "One step up" indicates that an industry identified as *Dispersed* with benchmark will turn to *Random* with another counterfactual, or an industry identified as *Random* will become *Localized* with another counterfactual. "One step down" indicates that an industry identified as *Localized* with benchmark will turn to *random* with another counterfactual, or an industry identified as *Random* will become *Dispersed* with another counterfactual. "Two steps up" indicates that an industry identified as *Dispersed* with benchmark will turn to *Localized* with another counterfactual. Finally, "Two steps down" means that an industry identified as *Localized* with benchmark will turn to *Dispersed* with another counterfactual.

For the agglomeration of establishments, the benchmark and the "Job counterfactuals, DO" are identical. From the benchmark to the "Job counterfactual, HAC", there are 4.8% of changes all in terms of "One step down". For the ag-

glomeration of the employment, there are 9.5% changes in terms of "One step down" for both "Job counterfactual, DO" and "Job counterfactual, HAC" and 4.8% additional "One step up" for the "Job counterfactual, DO".

The switches of identification from the benchmark to the multivariate counterfactual are more abundant and more pronounced. For the agglomeration on establishments (Table 3.5-Panel-a-), there are 23.8% of overall changes. These changes encompass 9.5% "One step up " and as many "Two steps up". These "up switch" originate from industries such as "312 Beverage and Tobacco product manufacturing" and "331 Primary metal manufacturing" which have all their plants in locations of *Class 1*, that exhibits dispersion compared to the universe of locations. The rest of the changes are entirely in terms of "Two steps down" and this is due to the industry "311 Food manufacturing" which has 73% of its plants on locations from *Class 3*, that exhibits "localization compared to the universe of locations.

For the agglomeration of employment (Table 3.5-Panel-b-), switches are much more important with 38.1% of changes in identifications from the benchmark to the multivariate counterfactual. These changes are split out in 19% of "up switches" originating from industries such as "335 Electrical, appliance manufacturing" and "322 paper manufacturing" which have respectively 90% and 70% of their employment on locations of *Class 1* that exhibits dispersion compared to the universe of locations. The remaining changes in terms of "down switches" are the result of industries such as "323 Printing manufacturing" with 91% of employment in Class 6, and "333 Machinery manufacturing" with 90% of employment in Class 2. These two classes have locations that are localized compared to the universe of locations.

The takeaway from these descriptives statistics is that, accounting for the determinants of locations choice yields a higher level of departure from randomness

compared to the case with counterfactuals without constraints. We check for robustness with different values for the optimal number of classes (4, 5, 7 and 8) and we find that the results are qualitatively the same, with more departure from randomness when accounting for the determinants of firms locations choice compared to an unconstrained counterfactual.
Additional results for the NAICS 4-digit industries provided in appendix 3.6.5 show that the global pattern of the differences between the benchmark and the multivariate counterfactuals are similar to that of the NAICS 3-digit industries. However, there are less drastic changes of the nature of "Two steps up" or "Two steps down" at four-digit than it appears at the NAICS 3-digit industries when moving from the benchmark to the multivariate counterfactuals.
In the rest of the analysis, we only consider the benchmark and the multivariate counterfactual.

3.4.2 Most localized and most dispersed industries

The previous description has already shown that the new counterfactual globally increases the mean value of the index of global localization as well as that of global dispersion. We now assess in more detail, the extent to which the counterfactual based on the location determinants changes the identification of industries in terms of the magnitude of their localization or dispersion. In other words, does the new counterfactual homogeneously increase the excess of agglomeration?

Table 3.6-panel a- presents, for the agglomeration of establishments, the most localized, and the most dispersed industries as generated by the multivariate counterfactual that we compare to the identification generated by the benchmark. For the 5 most localized industries 3 out of 5 appear in the order, in

the ranking of both approaches. These industries are "313 Textile mills", "322 Paper manufacturing", and "315 Clothing manufacturing". However, the two approaches differ for the identification of the two most dispersed industries.

On the agglomeration of employment, the 5 most localized industries show a similar picture. Three out of five industries appear in the ranking of both approaches. These industries are "313 Textile mills", "315 Clothing" and "316 Leather", but in a different order. For the most dispersed industries, 2 appear in the ranking of both approaches, these are "311 Food manufacturing", and "321 Wood manufacturing" but still not in the same order. Strikingly, "335 Electrical, appliance manufacturing" appears among the 4 most dispersed industries for the benchmark whereas it is among the 5 most localized industries for the multivariate counterfactual.

Table 3.6 – Most localized/dispersed, NAICS 3-digit industries

	Benchmark counterfactual			Multivariate counterfactual		
	NAICS	**Industry**	**Γ or Ψ**	**naics**	**Industry**	**Γ or Ψ**
Panel a : Establishments						
Most localized						
	313	Textile mills	0.0885	313	Textile mills	0.251
	322	Paper manufacturing	0.0804	322	Paper manufacturing	0.221
	315	Clothing manufacturing	0.0682	315	Clothing manufacturing	0.203
	326	Plastics, rubber products manuf.	0.0560	335	Electrical, appliance manuf.	0.133
	325	Chemical manufacturing	0.0349	331	Primary metal manufacturing	0.0800
Most dispersed						
	314	Textile product mills	0.0133	311	Food manufacturing	0.00104
	312	Beverage product manuf.	0.00823			
Panel b : Employment						
Most localized						
	326	Plastics, rubber products manuf.	0.0651	313	Textile mills	0.230
	325	Chemical manufacturing	0.0440	315	Clothing manufacturing	0.181
	315	Clothing manufacturing	0.0358	316	Leather, allied product manuf.	0.145
	313	Textile mills	0.0327	335	Electrical, appliance manuf.	0.0924
	316	Leather, allied product manuf.	0.0251	322	Paper manufacturing	0.0811
Most dispersed						
	321	Wood product manufacturing	0,0302	321	Wood product manufacturing	0.0461
	311	Food manufacturing	0,00333	333	Machinery manufacturing	0,0177
	334	Computer, electronic product manuf.	0,00133	311	Food manufacturing	0,00627
	335	Electrical, appliance manuf.	0,000681	332	Fabricated metal product manuf.	0.00405
				327	Non-metallic mineral product manuf.	0.00221

Notes: Only the five most localized or the five most dispersed industries are presented. For Panel a, the number of dispersed industries at the NAICS 3-digit is 2 for the benchmark and 1 for the multivariate counterfactual. For Panel b, only 4 industries are dispersed with the benchmark. procedure.

More generally, we perform a Spearman rank correlation test between the ranking generated by the two approaches. More precisely, industries are ranked based on their localization and dispersion index values. The ranking starts with the industry with the highest value of the index of localization Γ, then the industries which are identified as random are ranked in between with ex-aequo ranks. The ranking continues with the least dispersed industry and ends with the industry with the highest value of the index of dispersion. This ranking

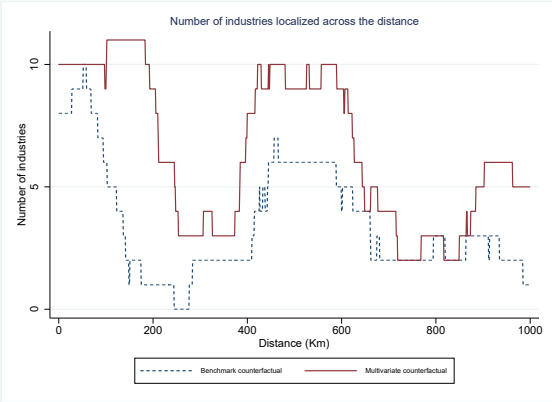
is performed for both the benchmark and the multivariate counterfactuals and a spearman rank test is run. The result rejects the independence of the two rankings. Put differently, the global picture from the two approaches does not significantly differ in terms of the identification of industries. However, a test of difference of the mean values of the indexes indicates a significant difference of the localization and dispersion indexes. [8]

3.4.3 Scale of localization and dispersion

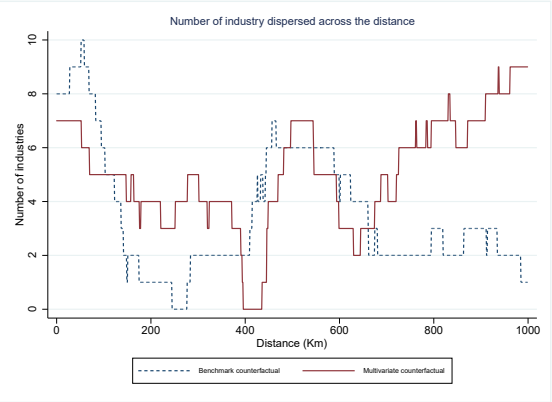
We now discuss the scale at which localization and dispersion occur. In Figure 3.7, we represent the number of the localized and dispersed at the NAICS 3-digit industries at each distance. The solid red line is for the multivariate counterfactual and the blue dashed line is for the benchmark.
For the localization of establishments -panel (a)-, the curve of each approach exhibits a multimodal shape. The number of localized industries is high at short distances and distances between big cities. As uncovered previously, the number of localized industries is higher with the multivariate counterfactual than with the benchmark at all distances. However, the multivariate counterfactual generates more concentration at shorter distances than the benchmark. For the dispersion of establishments-panel (b)-, the number of dispersed industries at short distance is slightly higher with the benchmark, and the majority of the dispersion in the industries happens at long distances with the multivariate counterfactual.

---

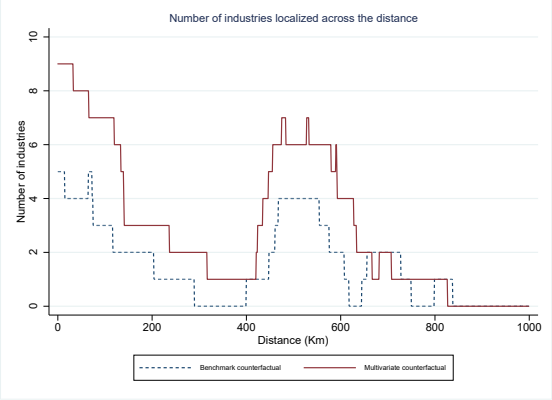
8. For the agglomeration of employment, the Rank correlation test of spearman rejects the independence of the two ranking with a p-value threshold of 0.0094. The t-test rejects the equality of the mean of the two distributions of indexes with a p-value threshold of 0.0297. These values are respectively, 0.0606 and 0.0068 for ranking from the agglomeration of establishments



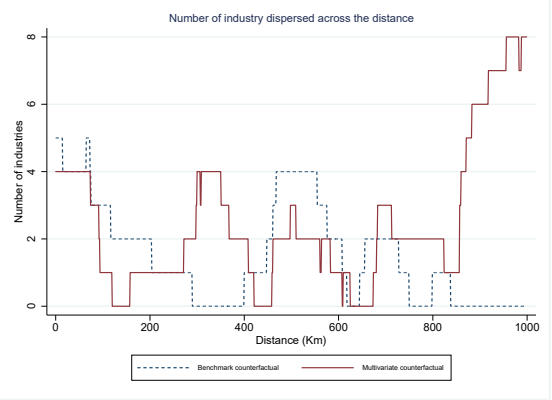
(a) Localized, establishments



(b) Dispersed, establishments



(d) Localized, employment



(d) Dispersed, employment

Figure 3.7 – Number of Localized / Dispersed 3 digit industries

For the scale of localization and dispersion of employment –panels c and d, the overall picture is similar to that of establishments and the distinction is even more clear. The multivariate counterfactual generates more localized industries at shorter distances compared to the benchmark. Concerning dispersion, the benchmark shows three peaks, with high dispersion at short, and medium distances whereas the multivariate approach yields a pattern with much more peaks, with a similar number of dispersed industries at short, and medium

distances and a higher number at long distances. Globally, the number of dispersion at shorter distances is higher for the benchmark compared to that of the multivariate counterfactual. Conversely, the highest peak of dispersion for the multivariate counterfactual is observed at long distances.

3.5 Concluding remarks

To study the location patterns of industries, we build on Duranton and Overman (2005, 2008) and suggest a new counterfactual to detect the departure from randomness. The construction of the counterfactual against which the observed concentration is to be compared is critical for the detection of localization and dispersion. Indeed, while testing for localization, any flaw entailed in the counterfactual will be reflected in the resulting classification of industries as localized, dispersed, or random.

We start by observing that the commonly used counterfactuals successfully capture what would be if the industries were randomly assigned across the actual locations. While these counterfactuals are relevant for controlling for the overall distribution of the industries, and possibly the distribution of some specific industries, they fail to account for the constraint that plants may face for reasons related to plants and locations characteristics. We propose a new counterfactual by constructing classes of locations, using a Hierarchical Ascendant Clustering, with inputs for the procedure, the key determinants of the choice of locations by firms suggested by the literature. Then, for each industry, we construct its counterfactual by reallocating each entity of that industry across the possible locations provided that the location belongs to the same *class* as that of the observed location of the entity. By doing so, we minimize the risk for a plant to switch on locations that are somehow impossible due to site size

constraints, potential regulatory constraints, or some others.

The approach with the new counterfactual is then used to assess the agglomeration of the manufacturing industries in Canada in 2017, and the results are compared to those obtained with the unconstraint counterfactual. We find for the concentration of the establishment that 90.5% (against 76.2% for the unconstraint counterfactual) of the industries are localized, 4.8% (against 9.5% for the unconstraint counterfactual) are dispersed, and 4.8% (against 14.3% for the unconstraint counterfactual) do not significantly depart from randomness. The concentration of employment exhibits similar patterns as 57.1% (against 47.6% for the unconstraint counterfactual) of industries are localized, 28.6% (against 19.0% for the unconstraint counterfactual) are dispersed and 14.3% (against 33.3% for the unconstraint counterfactual) do not depart significantly from randomness. Moreover, the degrees of concentration generated by the two approaches are quite different both qualitatively and quantitatively. Indeed, a large fraction of reclassification of industries as localized, dispersed, or random are observed across the two approaches (23.8% for the concentration of establishment, and 38.1% for the concentration of employment), and the new counterfactual yields, in general, higher magnitudes for localization and dispersions. Finally, the localization is more important at shorter distances with the new counterfactual compared to the classic counterfactual, while dispersion is more important at longer distances for the new counterfactual compared to the classic counterfactual. All these results also hold qualitatively at four-digit industries level for the concentration of both employment and establishments.

3.6 Appendix to chapter 3

3.6.1 Data

The main dataset use for the empirical part of this analysis is the Scott's National All Business Directories, a proprietary dataset that draws information on plants operating in Canada from Business Register records and telephone surveys. We start by geocoding the dataset to have the geographic coordinates of each plant. After this first step, we remove inaccurate geographic coordinates. Then we assign building footprint to establishments using the Geographic Information Systems tools. We clean the dataset again to remove observations for which there were no building polygons and those with inaccurate associations. To account for potential misreport of the number of workers of firms to their recorded plant and not only the part of the workers of each recorded establishment, we trim employment by removing the 1% upper and lower tails of employment across each industry. This last process also generates a loss of 4,438 observations. This cleaning process is important since the estimation of the agglomeration measures uses the geographic coordinates to compute bilateral distances and the number of workers of the establishment as weights to estimate the agglomeration of the employment. Finally, we remove the three territories of Nunavut, Yukon, and North Territories which only have very few observations, and we end up with a final dataset of 23,388 observations.

The second dataset is from the DMTI which records the zoning restrictions in Canada, i.e., the main type of activity allowed on each portion of space, namely, commercial, Government and Institutional, Open Area, Parks, and Recreational, Residential, Resource, and Industrial; and Waterboy, Not assigned. We aggregate these categories into three: The first group consists of commercial, Government, and Institutional, Resource, and Industrial; the second group

consists of: Residential, and the third group contains the remaining categories.

The third dataset is the Dissemination Area Boundary from Statistics Canada in 2016. This dataset contains information on the population at the level at the finest level at which Census data are disseminated. According to Statistic Canada, a dissemination area is a small area composed of one or more neighboring dissemination blocks and is the smallest standard geographic area for which all census data are disseminated. The files contain the boundaries of all dissemination areas which combine to cover all of Canada. We merge this file with the geocoded Scotts Dataset, then we use GIS tools to construct the population density at 1.5 km around each establishment. More precisely, we draw a circular buffer around each plant and aggregate all the dissemination areas whose centroids fall into the circle of 1.5km. Then we construct the population density of each of these aggregated entities. Next, we construct another buffer of 10km around each plant and count the number of plants from the same industry, as well as the total employment from the same industry.

The fourth dataset is the 2016 Census Road Network File from Statistics Canada that we use to compute for each plant the distance to its nearest junction.

3.6.2 Choice of discriminant variables

Probit estimation of location choice, NAICS 3-digit industries

Table 3.7 – Marginal effects on the probability of location choice, NAICS 3-digit industries (1/2)

	311	312	313	314	315	316	321	322	323	324	325
main											
Desnity 1.5 km around	0.0000861***	0.0000796***	0.0000999***	0.0000342***	0.000111***	0.000115***	-0.0000425***	0.0000244	-0.0000768***	-0.0000595	0.0000368***
	(0.00000746)	(0.0000131)	(0.0000153)	(0.00000848)	(0.00000661)	(0.0000144)	(0.0000117)	(0.0000168)	(0.00000861)	(0.0000452)	(0.00000934)
Building footprint size	0.0000243***	0.0000329***	0.0000416***	0.00000199	-0.00000828	0.0000160	-0.0000187***	0.0000496***	-0.0000188**	-0.0000280	-0.00000286
	(0.00000396)	(0.00000661)	(0.00000991)	(0.00000998)	(0.00000692)	(0.0000117)	(0.00000639)	(0.00000490)	(0.00000809)	(0.0000255)	(0.00000498)
Nb of workers of the plant	0.000168*	-0.000238	-0.00191*	-0.00412***	-0.000195	-0.00170	0.000110*	-0.00000531	-0.00114**	0.000198**	-0.000283
	(0.0000864)	(0.000191)	(0.000980)	(0.00103)	(0.000186)	(0.00109)	(0.0000657)	(0.000140)	(0.000483)	(0.0000892)	(0.000185)
Employment, 10km around same naics	0.000961***	-0.0000390	-0.000543***	-0.000715***	-0.0000600**	-0.000593**	-0.0000245	0.000454***	-0.00153***	0.00107***	0.000165***
	(0.0000362)	(0.000118)	(0.000144)	(0.0000622)	(0.0000238)	(0.000235)	(0.0000279)	(0.0000338)	(0.0000442)	(0.000194)	(0.0000461)
Nb plants, 10km around same naics	-0.0637***	-0.0309***	-0.0436***	0.0110***	-0.00102	-0.0805***	-0.00338***	-0.0278***	0.0517***	-0.198***	-0.0337***
	(0.00238)	(0.00594)	(0.00962)	(0.00119)	(0.000829)	(0.0142)	(0.000968)	(0.00216)	(0.00134)	(0.0279)	(0.00246)
Distance to nearest junction	0.00499***	-0.00163	-0.00911	-0.00141	-0.00177	-0.00123	0.00135	-0.00807	-0.00227	-0.0235	-0.00250
	(0.00113)	(0.00231)	(0.00780)	(0.00256)	(0.00227)	(0.00302)	(0.00128)	(0.00639)	(0.00169)	(0.0160)	(0.00296)
Business type, Distributor	0.184**		0.102	0.244**	0.100	-0.276*	-0.250***	0.0614	-0.645***	-0.0201	0.336***
	(0.0747)		(0.291)	(0.0993)	(0.112)	(0.156)	(0.0823)	(0.104)	(0.103)	(0.222)	(0.0765)
Business type, Manufacturer	0.296***	-0.0276	0.273	0.103	0.198**	-0.417***	0.0391	-0.140	0.109*	-0.157	0.00275
	(0.0601)	(0.0974)	(0.237)	(0.0830)	(0.0911)	(0.111)	(0.0590)	(0.0869)	(0.0589)	(0.173)	(0.0663)
Industrial, commercial, institutional	0.00524	-0.308***	0.247**	0.185***	0.235***	0.116	-0.179***	0.290***	0.211***	0.167	0.228***
	(0.0345)	(0.0697)	(0.116)	(0.0541)	(0.0579)	(0.108)	(0.0372)	(0.0645)	(0.0447)	(0.120)	(0.0418)
Residential	-0.0493	-0.273***	0.177	0.147***	0.147**	0.147	-0.204***	-0.00522	0.453***	-0.187	-0.0634
	(0.0359)	(0.0741)	(0.121)	(0.0554)	(0.0619)	(0.104)	(0.0405)	(0.0751)	(0.0447)	(0.147)	(0.0473)
Exporter	-0.338***	-0.303***	0.301***	-0.0648	0.119***	-0.00293	-0.0753**	0.252***	-0.471***	-0.271**	0.278***
	(0.0282)	(0.0616)	(0.0942)	(0.0409)	(0.0406)	(0.0824)	(0.0310)	(0.0493)	(0.0331)	(0.107)	(0.0331)
Headquarter	0.0213	0.109	-0.141	0.0414	-0.0245	-0.0712	-0.000764	-0.0237	-0.122*	0.113	0.312***
	(0.0446)	(0.0864)	(0.170)	(0.0719)	(0.0674)	(0.147)	(0.0503)	(0.0699)	(0.0632)	(0.157)	(0.0454)
Prince Edward Islands	0.580***	-0.433		-0.0782		0.440	-0.345*	0.320	0.220		0.453**
	(0.156)	(0.391)		(0.300)		(0.409)	(0.200)	(0.308)	(0.202)		(0.206)
New Scotia	0.303***	-0.0336		0.121	0.116	0.316	-0.262***	0.174	0.102	-0.0504	-0.00668
	(0.0785)	(0.141)		(0.119)	(0.157)	(0.250)	(0.0900)	(0.158)	(0.0952)	(0.286)	(0.123)
New Brunswick	0.0229	-0.552**	0.265	0.135	0.268	-0.0772	-0.106	0.216	0.0696	-0.0329	-0.187
	(0.0972)	(0.229)	(0.380)	(0.145)	(0.169)	(0.343)	(0.0976)	(0.170)	(0.113)	(0.281)	(0.159)
Quebec	-0.0478	-0.387***	0.579***	0.0364	0.365***	0.477***	-0.288***	0.134*	-0.118**	-0.277	0.269***
	(0.0453)	(0.0864)	(0.184)	(0.0625)	(0.0691)	(0.149)	(0.0451)	(0.0807)	(0.0498)	(0.174)	(0.0560)
Ontario	-0.00633	-0.214***	0.356*	0.0244	0.127*	0.297*	-0.372***	0.117	-0.0140	0.0510	0.217***
	(0.0435)	(0.0755)	(0.190)	(0.0611)	(0.0722)	(0.154)	(0.0430)	(0.0764)	(0.0473)	(0.156)	(0.0537)
Manitoba	0.0745	-0.499**		0.0154	0.0320	0.167	-0.402***	-0.125	0.369***	0.122	0.271***
	(0.0843)	(0.220)		(0.130)	(0.148)	(0.290)	(0.106)	(0.178)	(0.0814)	(0.300)	(0.101)
Saskatchewan	0.0326	-0.355*	0.204	-0.118	0.350**		-0.546***	-0.207	0.0720	0.0505	0.151
	(0.0912)	(0.206)	(0.374)	(0.145)	(0.142)		(0.120)	(0.244)	(0.100)	(0.297)	(0.118)
Alberta	-0.0512	-0.273**	0.0898	0.150*	0.00916	0.389**	-0.342***	-0.303**	0.153**	-0.0408	0.192***
	(0.0601)	(0.119)	(0.287)	(0.0795)	(0.104)	(0.185)	(0.0626)	(0.129)	(0.0599)	(0.222)	(0.0719)
Census Metropolitan Area	-0.179***	0.00486	0.0389	0.0829	-0.0999	0.216**	-0.276***	-0.0339	-0.0280	0.282**	0.331***
	(0.0412)	(0.0841)	(0.117)	(0.0701)	(0.0713)	(0.106)	(0.0458)	(0.0892)	(0.0570)	(0.139)	(0.0634)
Observations	23388	21200	22194	23388	23298	22946	23388	23388	23388	23298	23388
Pseudo R^2	0.139	0.114	0.168	0.068	0.074	0.170	0.057	0.123	0.275	0.186	0.109

Table 3.8 – Marginal effects on the probability of location choice, NAICS 3-digit industries (2/2)

	326	327	331	332	333	334	335	336	337	339
main										
Desnity 1.5 km around	-0.00000989	-0.00000822	0.0000174	-0.0000604***	-0.0000917***	0.0000298***	-0.0000179	-0.00000356	-0.0000188**	0.0000176***
	(0.00000858)	(0.0000101)	(0.0000159)	(0.00000712)	(0.00000828)	(0.00000928)	(0.0000124)	(0.0000111)	(0.00000820)	(0.00000555)
Building footprint size	0.0000308***	-0.0000145**	0.0000181***	-0.00000742*	-0.0000327***	-0.0000576***	-0.00000719	0.0000155***	0.0000217***	-0.0000510***
	(0.00000476)	(0.00000622)	(0.00000587)	(0.00000382)	(0.00000462)	(0.00000823)	(0.00000597)	(0.00000475)	(0.00000545)	(0.00000786)
Nb of workers of the plant	-0.000723***	0.0000168	0.000137	-0.000205	-0.000191*	0.000122	0.00000373	0.000344***	-0.000687	-0.000335
	(0.000243)	(0.0000713)	(0.0000847)	(0.000165)	(0.000108)	(0.0000836)	(0.0000852)	(0.0000953)	(0.000418)	(0.000261)
Employment, 10km around same naics	0.000852***	-0.0000642*	0.000216***	-0.0000919***	0.0000116	0.000336***	0.000217***	0.000837***	-0.0000384**	-0.000526***
	(0.0000229)	(0.0000385)	(0.0000775)	(0.0000148)	(0.0000154)	(0.0000463)	(0.0000347)	(0.0000546)	(0.0000190)	(0.0000289)
Nb plants, 10km around same naics	-0.0288***	-0.0191***	-0.0858***	-0.0000705	0.00699***	-0.0461***	-0.0235***	-0.0647***	0.000315	0.0225***
	(0.000837)	(0.00186)	(0.00614)	(0.000551)	(0.000582)	(0.00261)	(0.00198)	(0.00442)	(0.000716)	(0.000910)
Distance to nearest junction	0.0000137	0.000177	-0.00968**	-0.00394***	0.000906	-0.00385	-0.0159***	-0.00101	-0.00456***	-0.000408
	(0.00167)	(0.00156)	(0.00381)	(0.00141)	(0.00122)	(0.00322)	(0.00426)	(0.00184)	(0.00163)	(0.00170)
Business type, Distributor	-0.172**	0.0359	0.207	-0.302***	0.367***	-0.443***	0.120	0.196*	0.243**	-0.255***
	(0.0725)	(0.0699)	(0.127)	(0.0506)	(0.0503)	(0.0909)	(0.0878)	(0.105)	(0.104)	(0.0621)
Business type, Manufacturer	-0.112**	-0.149***	0.144	-0.129***	-0.143***	-0.198***	-0.106	0.267***	0.549***	-0.165***
	(0.0559)	(0.0560)	(0.107)	(0.0382)	(0.0424)	(0.0637)	(0.0737)	(0.0881)	(0.0857)	(0.0463)
Industrial, commercial, institutional	-0.0578	0.0337	0.141**	0.0711***	-0.174***	0.311***	0.156***	0.0209	0.0709*	0.0139
	(0.0393)	(0.0383)	(0.0582)	(0.0266)	(0.0289)	(0.0488)	(0.0501)	(0.0449)	(0.0393)	(0.0365)
Residential	-0.0537	-0.125***	-0.211***	-0.0755**	-0.0764**	0.200***	-0.0223	-0.0288	-0.0207	0.347***
	(0.0453)	(0.0411)	(0.0655)	(0.0302)	(0.0333)	(0.0532)	(0.0582)	(0.0485)	(0.0437)	(0.0371)
Exporter	0.250***	-0.275***	0.0366	-0.0785***	0.527***	0.508***	0.387***	0.0202	-0.298***	-0.152***
	(0.0308)	(0.0317)	(0.0461)	(0.0210)	(0.0231)	(0.0387)	(0.0400)	(0.0359)	(0.0319)	(0.0274)
Headquarter	0.00794	0.114**	-0.0255	-0.134***	0.0316	0.0351	0.148***	-0.0669	-0.0626	0.0178
	(0.0488)	(0.0483)	(0.0732)	(0.0362)	(0.0359)	(0.0571)	(0.0542)	(0.0595)	(0.0558)	(0.0462)
Prince Edward Islands	-0.272	-0.291		-0.565***	-0.141		0.348	-0.261	0.146	-0.0628
	(0.301)	(0.218)		(0.208)	(0.199)		(0.251)	(0.263)	(0.212)	(0.225)
New Scotia	-0.00374	-0.0995	-0.376*	-0.178**	-0.296***	0.0874	-0.454**	-0.108	-0.156	0.264***
	(0.108)	(0.0927)	(0.196)	(0.0766)	(0.0906)	(0.119)	(0.195)	(0.114)	(0.116)	(0.0838)
New Brunswick	-0.185	0.00955	-0.0594	-0.228**	-0.0663	-0.284*	-0.0609	-0.188	-0.0656	0.298***
	(0.138)	(0.0998)	(0.170)	(0.0894)	(0.0935)	(0.164)	(0.171)	(0.129)	(0.126)	(0.0959)
Quebec	0.0984**	-0.134***	0.0687	0.0595*	0.0101	-0.0922	0.0835	-0.259***	0.165***	-0.0816*
	(0.0500)	(0.0497)	(0.0783)	(0.0355)	(0.0392)	(0.0618)	(0.0634)	(0.0581)	(0.0499)	(0.0436)
Ontario	-0.113**	-0.0526	0.132*	0.0969***	-0.0774**	0.137**	0.0491	-0.101*	0.000858	0.0287
	(0.0499)	(0.0466)	(0.0743)	(0.0333)	(0.0369)	(0.0557)	(0.0595)	(0.0531)	(0.0495)	(0.0418)
Manitoba	0.114	-0.211*	-0.0290	-0.0278	0.138*	-0.0873	-0.0134	-0.0573	-0.264*	0.121
	(0.0960)	(0.116)	(0.164)	(0.0722)	(0.0730)	(0.131)	(0.130)	(0.106)	(0.137)	(0.0915)
Saskatchewan	-0.0379	0.0274	-0.187	-0.0340	0.136*	-0.287*	-0.0875	-0.300**	-0.0415	0.192**
	(0.126)	(0.101)	(0.189)	(0.0798)	(0.0818)	(0.158)	(0.152)	(0.135)	(0.116)	(0.0871)
Alberta	-0.161**	0.0297	0.204**	0.0800*	0.209***	0.182**	-0.0800	-0.303***	-0.239***	0.0924*
	(0.0692)	(0.0624)	(0.0992)	(0.0446)	(0.0468)	(0.0747)	(0.0871)	(0.0814)	(0.0734)	(0.0553)
Census Metropolitan Area	-0.0321	-0.00861	0.357***	0.162***	0.0551	0.553***	0.464***	0.161***	-0.104**	0.0920*
	(0.0568)	(0.0474)	(0.0766)	(0.0384)	(0.0432)	(0.0744)	(0.0900)	(0.0582)	(0.0523)	(0.0526)
Observations	23388	23388	23298	23388	23388	23298	23388	23388	23388	23388
Pseudo R^2	0.183	0.079	0.145	0.022	0.114	0.138	0.087	0.115	0.035	0.122

3.6.3 Classification trees, Gap statistics curves, distribution of classes

Figure 3.8 – Classification tree, worker HAC at 3 digit

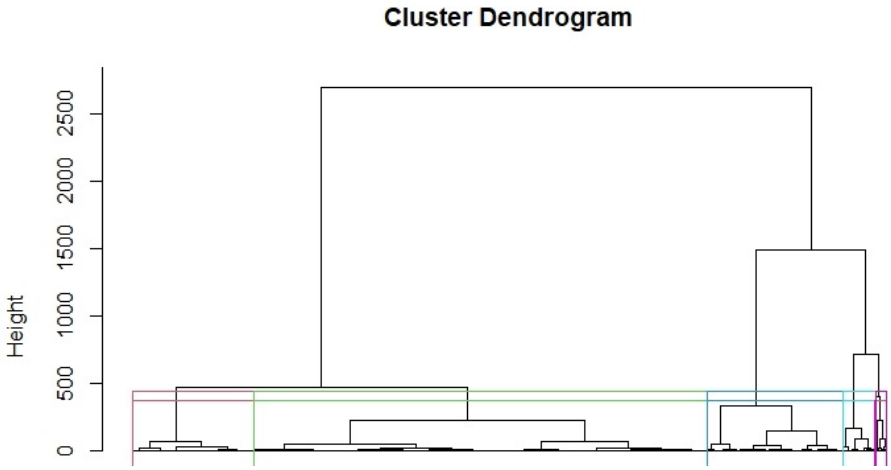


Figure 3.9 – Gap statistics curve, worker HAC at 3 digit

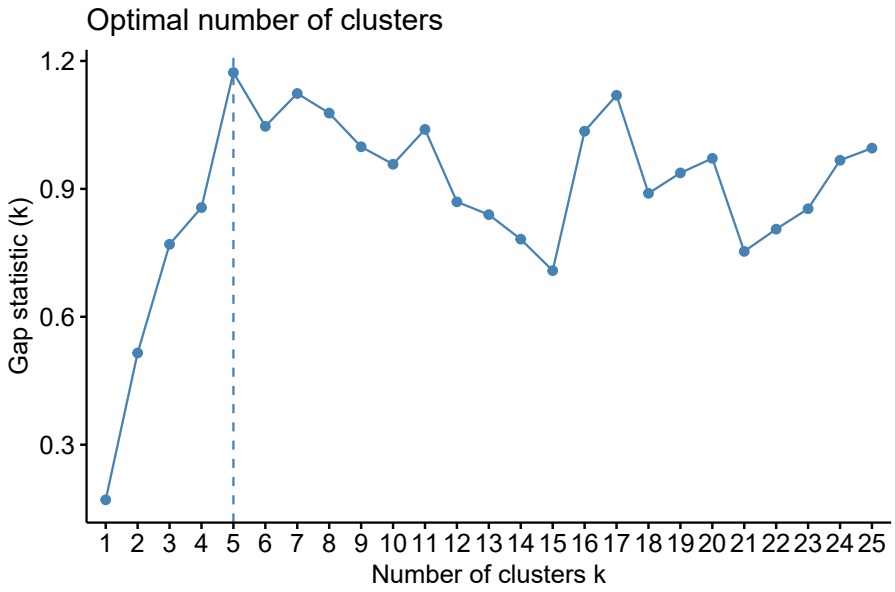


Figure 3.10 – Classification tree, multivariate HAC at 3 digit

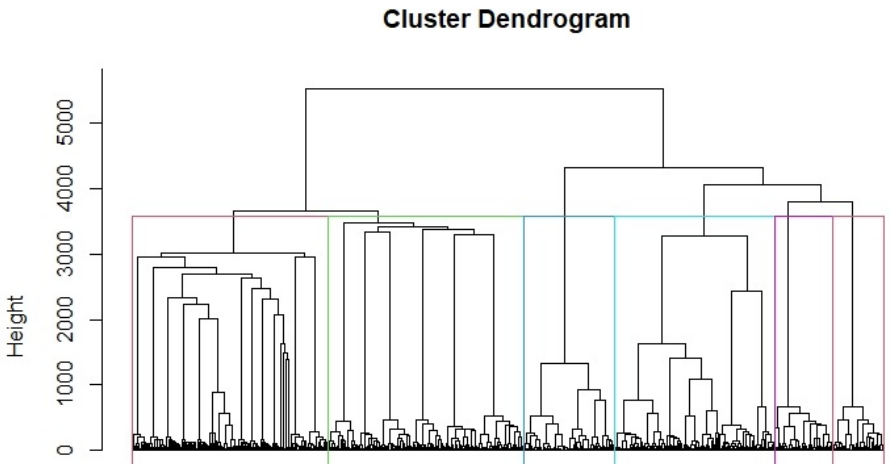


Figure 3.11 – Gap statistics curve, multivariate HAC at 3 digit

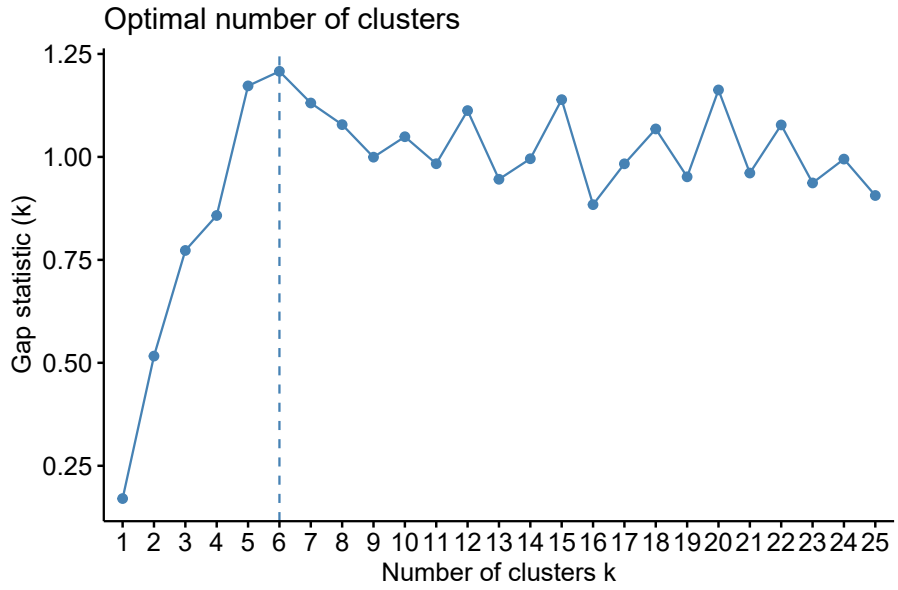


Figure 3.12 – Classification tree, worker HAC at 4 digit

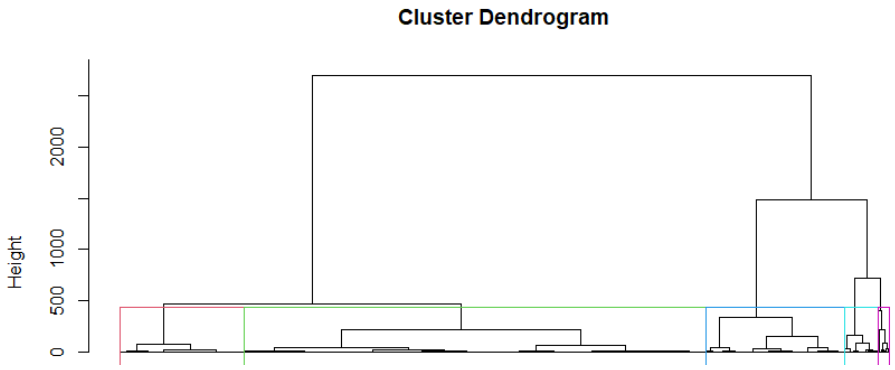
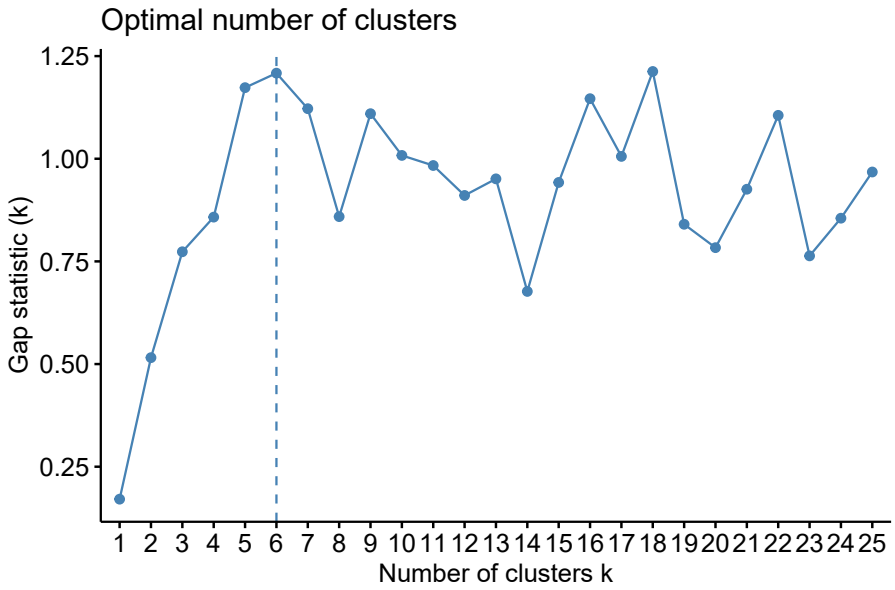


Figure 3.13 – Gap statistics curve, worker HAC at 4 digit



Notes: The gap statistics for the HAC procedure performed with workers suggests 6 bins. But we use 5 bins for the main results at the NAICS 4-digit for comparison to the 5 bins of Duranton and Overman (2008). This may not really alter results since the gap statistics between 5 and 6 classes is not very huge.

Figure 3.14 – Classification tree, multivariate HAC at 4 digit

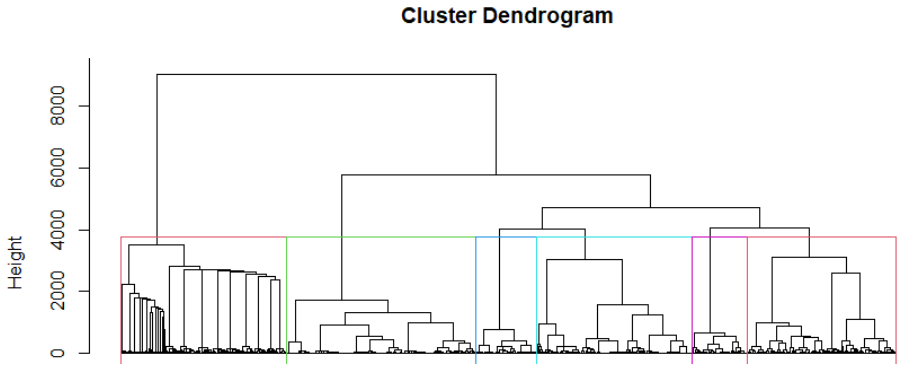


Figure 3.15 – Gap statistics curve, multivariate HAC at 4 digit

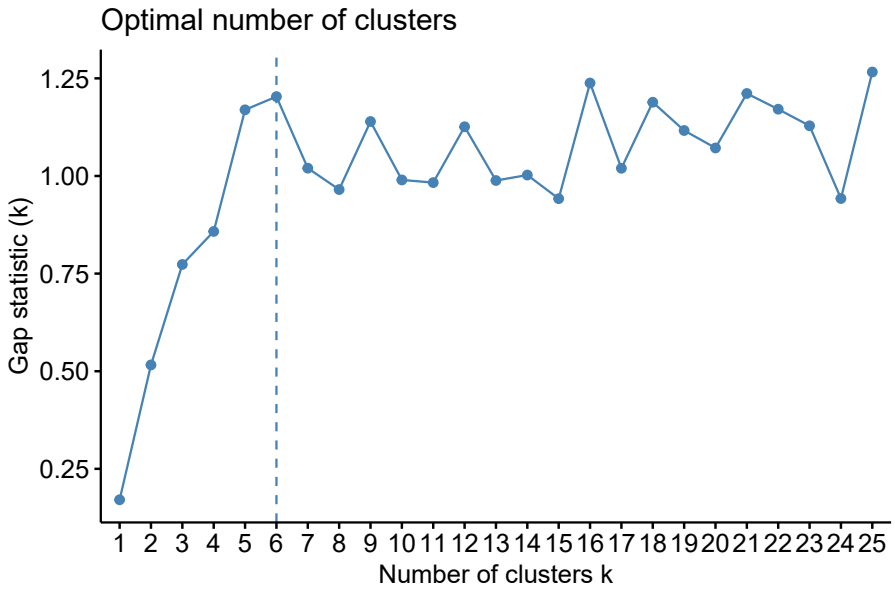
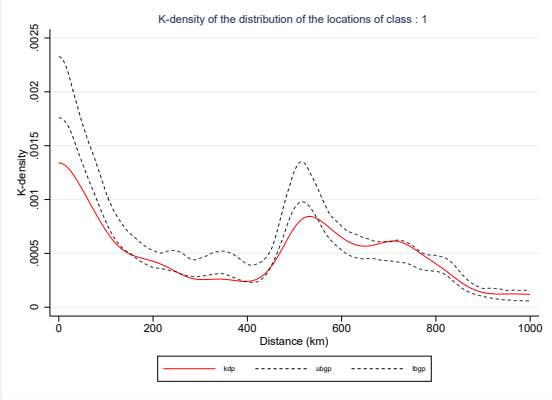
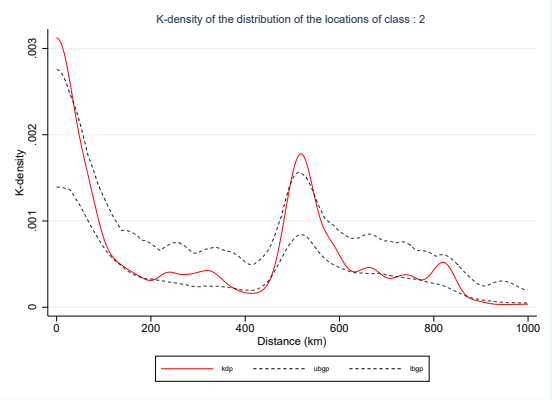


Table 3.9 – Distribution of the industries across the locations classes

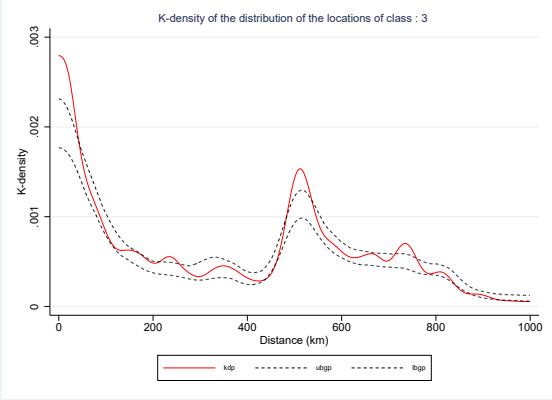
	Classes of locations						
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**	**Total**
Panel a : Establishment							
311 Food manufacturing	517	8	1,436	0	12	0	1,973
312 Beverage product manuf.	230	0	0	0	0	0	230
313 Textile mills	72	0	0	0	0	0	72
314 Textile product mills	558	0	0	0	0	0	558
315 Clothing manufacturing	535	2	0	0	0	0	537
316 Leather, allied product manuf.	105	0	0	0	0	0	105
321 Wood product manufacturing	136	4	1,028	0	3	0	1,171
322 Paper manufacturing	338	14	0	0	4	0	356
323 Printing, support activities	179	0	0	0	11	1,597	1,787
324 Petrol, coal product manuf.	55	0	0	0	0	0	55
325 Chemical manufacturing	89	5	946	0	4	0	1,044
326 Plastics, rubber products manuf.	91	1,332	0	0	4	0	1,427
327 Non-metallic mineral product manuf.	172	1	1,021	0	0	0	1,194
331 Primary metal manufacturing	386	2	0	0	0	0	388
332 Fabricated metal product manuf.	539	592	0	2,802	10	0	3,943
333 Machinery manufacturing	447	2,883	0	0	8	0	3,338
334 Computer, electronic product manuf.	49	3	704	0	0	0	756
335 Electrical, appliance manuf.	585	1	0	0	0	0	586
336 Transportation equipment manuf.	714	56	0	0	0	0	770
337 Furniture, related product manuf.	70	36	970	0	0	0	1,076
339 Miscellaneous manufacturing	211	66	0	0	1,745	0	2,022
Total	6,078	5,005	6,105	2,802	1,801	1,597	23,388
Panel b : Employment							
311 Food manufacturing	23,205	9,900	80,960	0	429	0	114,494
312 Beverage product manuf.	9,437	0	0	0	0	0	9,437
313 Textile mills	2,680	0	0	0	0	0	2,680
314 Textile product mills	8,919	0	0	0	0	0	8,919
315 Clothing manufacturing	17,242	2,900	0	0	0	0	20,142
316 Leather, allied product manuf.	2,408	0	0	0	0	0	2,408
321 Wood product manufacturing	4,905	6,425	39,901	0	82	0	51,313
322 Paper manufacturing	20,655	8,454	0	0	63	0	29,172
323 Printing, support activities	2,529	0	0	0	98	28,173	30,800
324 Petrol, coal product manuf.	7,130	0	0	0	0	0	7,130
325 Chemical manufacturing	3,199	7,200	38,374	0	53	0	48,826
326 Plastics, rubber products manuf.	3,020	64,282	0	0	20	0	67,322
327 Non-metallic mineral product manuf.	4,671	3,000	35,717	0	0	0	43,388
331 Primary metal manufacturing	22,735	5,500	0	0	0	0	28,235
332 Fabricated metal product manuf.	15,334	46,951	0	68,051	711	0	131,047
333 Machinery manufacturing	13,214	114,813	0	0	157	0	128,184
334 Computer, electronic product manuf.	1,978	8,800	32,746	0	0	0	43,524
335 Electrical, appliance manuf.	28,647	3,500	0	0	0	0	32,147
336 Transportation equipment manuf.	29,417	40,740	0	0	0	0	70,157
337 Furniture, related product manuf.	1,218	3,043	26,543	0	0	0	30,804
339 Miscellaneous manufacturing	3,635	10,347	0	0	28,439	0	42,421
Total	226,178	335,855	254,241	68,051	30,052	28,173	942,550



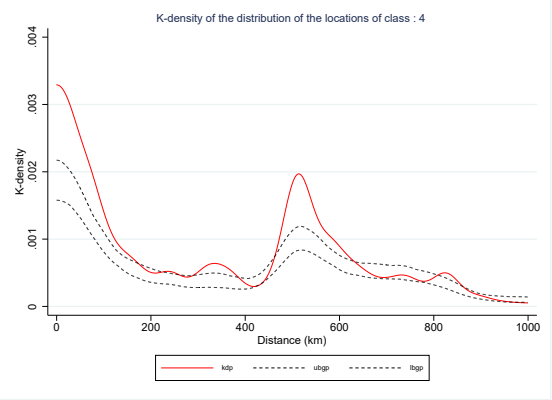
(a) Class 1, employment



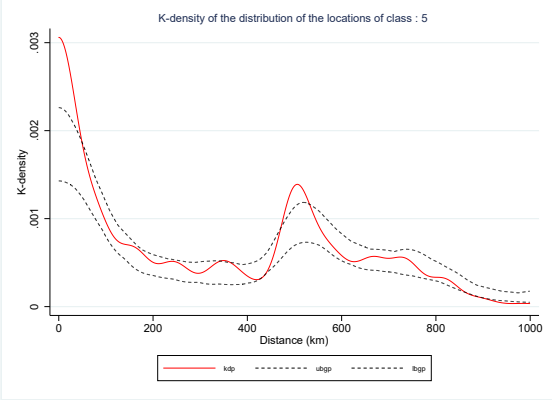
(b) Class 2, employment



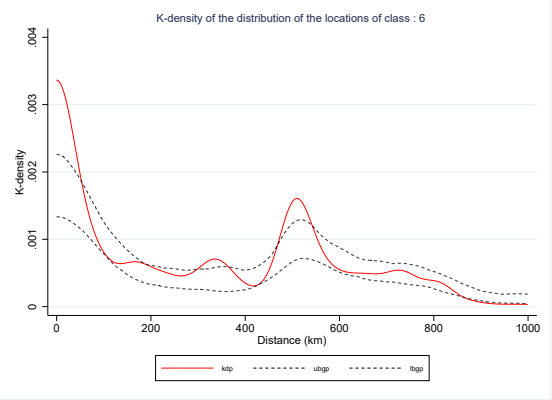
(c) Class 3, employment



(d) Class 4, employment

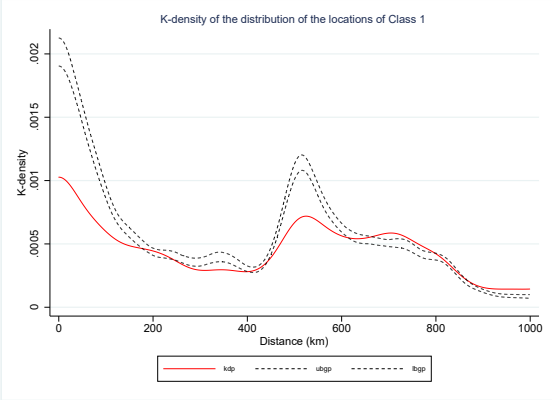


(e) Class 5, employment

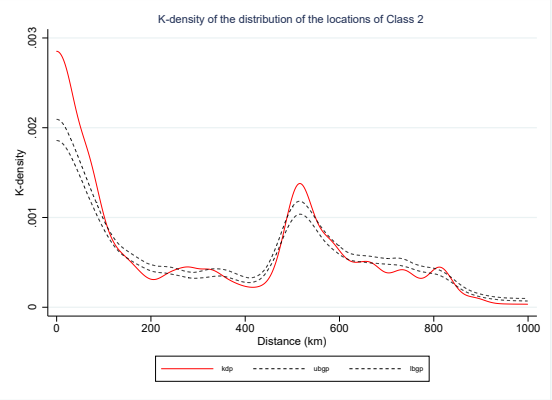


(f) Class 6, employment

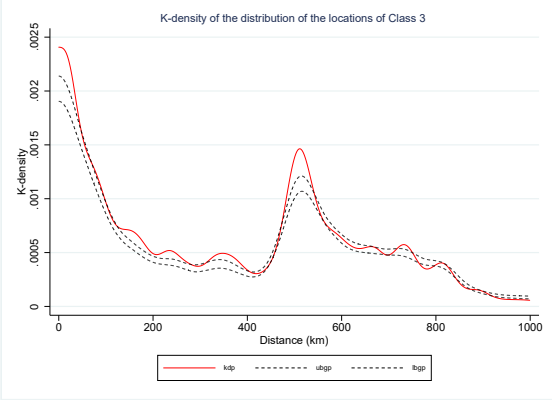
Figure 3.16 – K-densities of the distribution of locations of the classes - 3 digit, employment



(a) Class 1, establishments

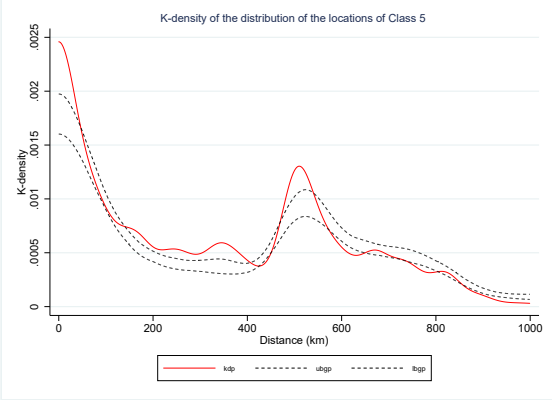


(b) Class 2, establishments



(c) Class 3, establishments

(d) Class 4, establishments



(e) Class 5, establishments

(f) Class 6, establishments

Figure 3.17 – K-densities of the distribution of locations of the classes - 3 digit, establishments

3.6.4 Detailed results for the NAICS 3-digit industries

Table 3.10 – Detailed classfication of the NAICS 3-digit industries, Establishments

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
311	Food manufacturing	agglo	agglo	agglo	agglo	agglo	disp	agglo	disp
312	Beverage product manuf.	disp	disp	disp	disp	disp	agglo	agglo	agglo
313	Textile mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
314	Textile product mills	disp	disp	disp	random	random	agglo	agglo	agglo
315	Clothing manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
316	Leather, allied product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
321	Wood product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
322	Paper manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
323	Printing, support activities	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
324	Petrol, coal product manuf.	random	random	random	random	random	random	random	random
325	Chemical manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
326	Plastics, rubber products manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
327	Non-metallic mineral product manuf.	agglo	agglo	agglo	disp	disp	agglo	agglo	agglo
331	Primary metal manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
332	Fabricated metal product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
333	Machinery manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
334	Computer, electronic product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
335	Electrical, appliance manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
336	Transportation equipment manuf.	random	random	disp	agglo	agglo	agglo	agglo	agglo
337	Furniture, related product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
339	Miscellaneous manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
	Share agglomerated	76,2%	76,2%	76,2%	81,0%	81,0%	90,5%	95,2%	90,5%
	Share random	14,3%	14,3%	9,5%	9,5%	9,5%	4,8%	4,8%	4,8%
	Share dispersed	9,5%	9,5%	14,3%	9,5%	9,5%	4,8%	0,0%	4,8%
	Percent switches between counterfactual								

Notes: This table reports the test for localization with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.11 – Changes of classification from the benchmark to other counterfatcuals, NAICS 3-digit: Establishments

NAICS3	NAMES	Switching						
		B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
311	Food manufacturing					2 steps down		2 steps down
312	Beverage product manuf.					2 steps up	2 steps up	2 steps up
313	Textile mills							
314	Textile product mills			1 step up	1 step up	2 steps up	2 steps up	2 steps up
315	Clothing manufacturing							
316	Leather, allied product manuf.							
321	Wood product manufacturing							
322	Paper manufacturing							
323	Printing, support activities							
324	Petrol, coal product manuf.							
325	Chemical manufacturing							
326	Plastics, rubber products manuf.							
327	Non-metallic mineral product manuf.			2 steps down	2 steps down			
331	Primary metal manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
332	Fabricated metal product manuf.							
333	Machinery manufacturing							
334	Computer, electronic product manuf.							
335	Electrical, appliance manuf.							
336	Transportation equipment manuf.		1 step down	1 step up	1 step up	1 step up	1 step up	1 step up
337	Furniture, related product manuf.							
339	Miscellaneous manufacturing							
	Percentage	**0,0%**	**4,8%**	**19,0%**	**19,0%**	**23,8%**	**19,0%**	**23,8%**

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.12 – Detailed classfication of the NAICS 3-digit industries, Employment

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
311	Food manufacturing	disp	disp	disp	agglo	agglo	disp	agglo	agglo
312	Beverage product manuf.	random	random	random	random	random	agglo	agglo	agglo
313	Textile mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
314	Textile product mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
315	Clothing manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
316	Leather, allied product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
321	Wood product manufacturing	disp	disp	disp	disp	disp	disp	disp	disp
322	Paper manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
323	Printing, support activities	agglo	random	random	random	random	random	random	random
324	Petrol, coal product manuf.	random	random	random	random	random	random	random	random
325	Chemical manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	disp
326	Plastics, rubber products manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
327	Non-metallic mineral product manuf.	random	disp	disp	random	random	disp	random	random
331	Primary metal manufacturing	random	random	random	random	random	agglo	agglo	agglo
332	Fabricated metal product manuf.	random	random	random	disp	disp	disp	disp	disp
333	Machinery manufacturing	agglo	agglo	agglo	disp	disp	disp	agglo	agglo
334	Computer, electronic product manuf.	disp	disp	disp	disp	agglo	disp	disp	disp
335	Electrical, appliance manuf.	disp	random	disp	agglo	agglo	agglo	agglo	agglo
336	Transportation equipment manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
337	Furniture, related product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	disp	disp
339	Miscellaneous manufacturing	random	random	random	random	random	random	random	random
	Share agglomerated	47,6%	42,9%	42,9%	52,4%	57,1%	57,1%	61,9%	57,1%
	Share random	33,3%	38,1%	33,3%	28,6%	28,6%	14,3%	19,0%	19,0%
	Share dispersed	19,0%	19,0%	23,8%	19,0%	14,3%	28,6%	19,0%	23,8%
	Percent switches between counterfactual								

Notes: This table reports the test for localization for with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.13 – Changes of classification from the benchmark to other counterfatcuals, NAICS 3-digit: Employment

NAICS3	NAMES	Switching						
		B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
311	Food manufacturing			2 steps up	2 steps up		2 steps up	2 steps up
312	Beverage product manuf.					1 step up	1 step up	1 step up
313	Textile mills							
314	Textile product mills							
315	Clothing manufacturing							
316	Leather, allied product manuf.							
321	Wood product manufacturing							
322	Paper manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
323	Printing, support activities	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down
324	Petrol, coal product manuf.							
325	Chemical manufacturing							2 steps down
326	Plastics, rubber products manuf.							
327	Non-metallic mineral product manuf.	1 step down	1 step down			1 step down		
331	Primary metal manufacturing					1 step up	1 step up	1 step up
332	Fabricated metal product manuf.			1 step down	1 step down	1 step down	1 step down	1 step down
333	Machinery manufacturing			2 steps down	2 steps down	2 steps down		
334	Computer, electronic product manuf.				2 steps up			
335	Electrical, appliance manuf.	1 step up		2 steps up	2 steps up	2 steps up	2 steps up	2 steps up
336	Transportation equipment manuf.							
337	Furniture, related product manuf.						2 steps down	2 steps down
339	Miscellaneous manufacturing							
	Percentage	**14,3%**	**9,5%**	**28,6%**	**33,3%**	**38,1%**	**38,1%**	**42,9%**

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

3.6.5 Key results for the NAICS 4-digit industries

Table 3.14 – Summary statistics : Localization, dispersion and randomness, NAICS 4-digit industries

			Localized			**Dispersed**		**random**
Panel a : establishments	nb	%	Γ	nb	%	Ψ	nb	%
Benchmark, unconstrained	41	47.7	0.0242	12	14.0	0.0130	32	38.4
Job counterfactual, DO 2008	38	44.2	0.0265	17	19.8	0.0122	30	36.0
Job counterfactual, HAC	40	46.5	0.0249	1	17.4	0.0110	30	36.0
Multivariate counterfactual	41	47.7	0.0550	19	22.1	0.0049	25	30.2
Panel b : employment	nb	%	Γ	nb	%	Ψ	nb	%
Benchmark, unconstrained	26	30.2	0.0355	13	15.1	0.0069	46	54.7
Job counterfactual, DO 2008	27	31.4	0.0375	15	17.4	0.0065	43	51.2
Job counterfactual, HAC	24	27.9	0.0381	16	18.6	0.0074	45	53.5
Multivariate counterfactual	21	24.4	0.0975	15	17.4	0.0036	49	58.1

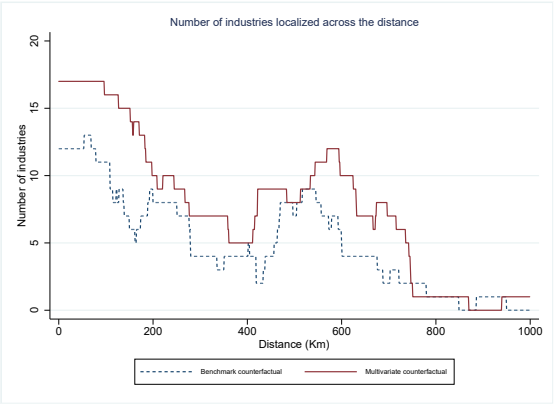
Notes:

Table 3.15 – Summary statistics : Changes from the benchmark to other counterfactuals, NAICS 4-digit industries

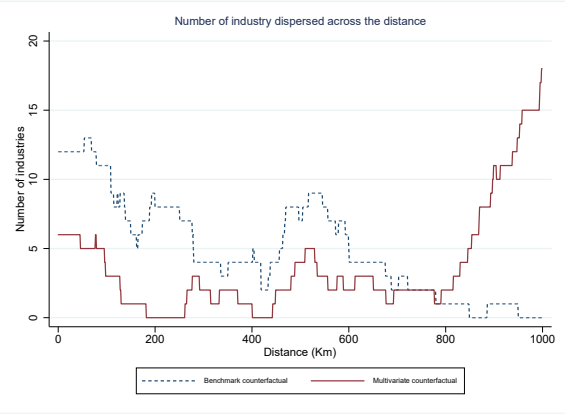
	Switches from the Benchmark to other counterfactuals				
Panel a : establishments					
	One step up	One step down	Two steps up	Two steps down	Unchanged
Benchmark to Job DO	1.2%	3.5%	1.2%	4.7%	89.5%
Benchmark to Job HAC	4.8%	3.5%	2.3%	3.5%	85.9%
Benchmark to Multivariate	11.6%	17.4%	5.8%	7.0%	58.1%
Panel b : employment					
	One step up	One step down	Two steps up	Two steps down	Unchanged
Benchmark to Job DO	3.5%	4.7%	0.0%	0.0%	91.9%
Benchmark to Job HAC	4.8%	7.0%	0.0%	0.0%	88.3%
Benchmark to Multivariate	14.0%	19.8%	3.5%	4.7%	58.1%

Table 3.16 – Most localized and most dispersed, NAICS 4-digit industries

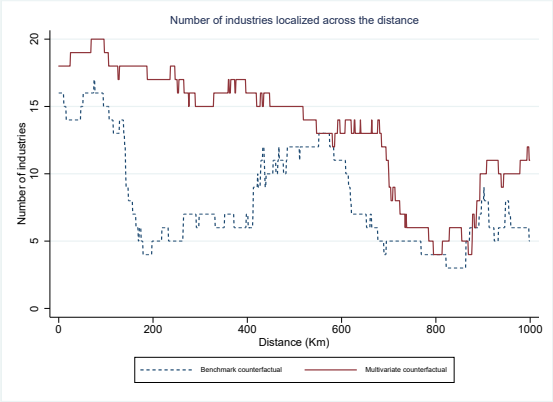
	Benchmark counterfactual			Multivariate counterfactual		
	naics-4	**Industry**	**Index**	**naics-3**	**Industry**	**Index**
Panel a : Establishments						
Most localized						
	3335	Metalworking machinery manufacturing	0.191	3133	Textile and fabric finishing and fabric coating	0.272
	3222	Converted paper product manufacturing	0.112	3132	Fabric mills	0.233
	3255	Paint coating and adhesive manufacturing	0.111	3152	Cut and sew clothing manufacturing	0.213
	3117	Seafood product preparation and packaging	0.0592	3113	Sugar & confectionery product manuf.	0.186
	3261	Plastic product manufacturing	0.0557	3151	Clothing knitting mills	0.166
	3152	Cut and sew clothing manufacturing	0.0557	3117	Seafood product preparation and packaging	0.154
	3372	Office furniture (including fixtures) manuf.	0.0504	3119	Other food manufacturing	0.135
	3312	Steel product manuf. from purchased steel	0.0478	3159	Clothing accessories and other clothing manufacturing	0.135
	3364	Aerospace product & parts manufacturing	0.0470	3162	Footwear manufacturing	0.128
	3344	Semiconductor & other electronic component manuf.	0.0373	3159	Clothing accessories and other clothing manufacturing	0.119
Most dispersed						
	3212	Veneer plywood and engineered wood product manuf.	0.0377	3329	Other fabricated metal product manufacturing	0.0152
	3366	Ship and boat building	0.0305	3161	Leather and hide tanning and finishing	0.0117
	3342	Communications equipment manufacturing	0.0235	3391	Medical equipment & supplies manuf.	0.0105
	3279	Other non metallic mineral product manuf.	0.0152	3342	Communications equipment manufacturing	0.00756
	3272	Glass and glass product manufacturing	0.0149	3311	Iron, steel mills & ferro alloy manuf.	0.00746
	3121	Beverage manufacturing	0.0133	3272	Glass and glass product manufacturing	0.00698
	3345	Navigational measuring	0.0117	3323	Architectural & structural metals manuf.	0.00680
	3211	Sawmills and wood preservation	0.00635	3255	Paint coating and adhesive manufacturing	0.00560
	3351	Electric lighting equipment manufacturing	0.00175	3361	Motor vehicle manufacturing	0.00463
	3391	Medical equipment & supplies manuf.	0.000969	3322	Cutlery & hand tool manufacturing	0.00381
Panel b : Employment						
Most localized						
	3335	Metalworking machinery manufacturing	0.202	3151	Clothing knitting mills	0.265
	3363	Motor vehicle parts manufacturing	0.121	3133	Textile and fabric finishing and fabric coating	0.228
	3255	Paint coating and adhesive manufacturing	0.115	3132	Fabric mills	0.227
	3344	Semiconductor & other electronic component manuf.	0.0711	3162	Footwear manufacturing	0.195
	3261	Plastic product manufacturing	0.0574	3152	Cut and sew clothing manufacturing	0.181
	3325	Hardware manufacturing	0.0423	3113	Sugar & confectionery product manuf.	0.167
	3252	Resin synthetic rubber	0.0392	3117	Seafood product preparation and packaging	0.160
	3372	Office furniture (including fixtures) manuf.	0.0368	3115	Dairy product manufacturing	0.134
	3117	Seafood product preparation and packaging	0.0329	3119	Other food manufacturing	0.115
	3113	Sugar & confectionery product manuf.	0.0289	3118	Bakeries and tortilla manufacturing	0.106
Most dispersed						
	3253	Pesticide fertilizer & other agricultural chemical manuf.	0.0325	3219	Other wood product manufacturing	0.0163
	3331	Agricultural construction & mining machinery manuf.	0.0225	3211	Sawmills and wood preservation	0.00940
	3219	Other wood product manufacturing	0.0145	3339	Other general purpose machinery manufacturing	0.00486
	3211	Sawmills and wood preservation	0.00677	3273	Cement and concrete product manufacturing	0.00454
	3324	Boiler tank & shipping container manuf.	0.00483	3121	Beverage manufacturing	0.00420
	3251	Basic chemical manufacturing	0.00402	3253	Pesticide fertilizer & other agricultural chemical manuf.	0.00354
	3341	Computer and peripheral equipment manufacturing	0.00163	3231	Printing and related support activities	0.00289
	3161	Leather and hide tanning and finishing	0.000802	3372	Office furniture (including fixtures) manuf.	0.00286
	3314	Non ferrous metal production & processing	0.000709	3255	Paint coating and adhesive manufacturing	0.00232
	3391	Medical equipment & supplies manuf.	0.000528	3327		0.00174



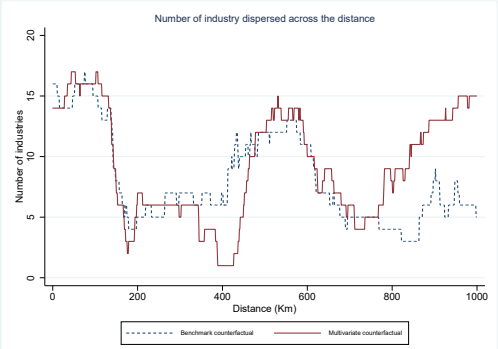
(a) Localized, employment



(b) Dispersed, employment



(c) Localized, establishments



(d) Dispersed, establishments

Figure 3.18 – Number of Localized / Dispersed, 4 digit industries

Table 3.17 – Detailed classfication of industries, NAICS 4-digit: Establishments (1/2)

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
3111	Animal food manufacturing	agglo	disp	disp	agglo	agglo	agglo	random	random
3112	Grain & oilseed milling	random	random	random	disp	disp	disp	disp	random
3113	Sugar & confectionery product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3114	Fruit, vegetable preserving & specialty food manuf.	random	random	random	agglo	agglo	agglo	random	agglo
3115	Dairy product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3116	Meat product manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
3117	Seafood product preparation and packaging	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3118	Bakeries and tortilla manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
3119	Other food manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3121	Beverage manufacturing	disp	disp	disp	agglo	agglo	agglo	disp	disp
3122	Tobacco manufacturing	random	random	random	agglo	agglo	agglo	random	agglo
3131	Fibre yarn and thread mills	random	random	random	random	random	random	random	random
3132	Fabric mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3133	Textile and fabric finishing and fabric coating	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3141	Textile furnishings mills	random	random	random	agglo	agglo	agglo	disp	disp
3149	Other textile product mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3151	Clothing knitting mills	random	agglo	random	agglo	agglo	agglo	agglo	agglo
3152	Cut and sew clothing manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3159	Clothing accessories and other clothing manuf.	random	random	random	agglo	agglo	agglo	agglo	agglo
3161	Leather and hide tanning and finishing	random	random	random	disp	random	disp	random	random
3162	Footwear manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3169	Other leather and allied product manufacturing	random	random	random	random	random	random	random	random
3211	Sawmills and wood preservation	disp	disp	agglo	agglo	agglo	agglo	agglo	agglo
3212	Veneer plywood and engineered wood product manuf.	disp	disp	disp	agglo	agglo	agglo	agglo	agglo
3219	Other wood product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3221	Pulp paper and paperboard mills	random	random	disp	disp	random	random	random	disp
3222	Converted paper product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	disp	agglo
3231	Printing and related support activities	agglo	agglo	agglo	disp	agglo	agglo	disp	disp
3241	Petroleum & coal product manufacturing	random	random	random	random	disp	disp	disp	random
3251	Basic chemical manufacturing	disp	agglo	agglo	agglo	random	random	random	random
3252	Resin synthetic rubber	random	random	agglo	random	random	random	random	random
3253	Pesticide fertilizer & other agricultural chemical manuf.	agglo	disp	agglo	agglo	agglo	agglo	disp	disp
3254	Pharmaceutical and medicine manufacturing	random	disp	random	disp	disp	disp	disp	disp
3255	Paint coating and adhesive manufacturing	agglo	agglo	agglo	agglo	disp	disp	agglo	disp
3256	Soap cleaning compound and toilet preparation manuf.	agglo	disp	disp	disp	disp	disp	disp	disp
3259	Other chemical product manufacturing	random	random	random	random	random	random	random	random
3261	Plastic product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3262	Rubber product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3271	Clay product and refractory manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3272	Glass and glass product manufacturing	disp	disp	disp	disp	disp	disp	disp	disp
3273	Cement and concrete product manufacturing	agglo	disp	agglo	agglo	agglo	agglo	agglo	agglo
3274	Lime and gypsum product manufacturing	random	random	random	random	random	random	random	random

Notes: This table reports the test for localization for with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.18 – Detailed classfication of industries, NAICS 4-digit: Establishments (2/2)

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
3279	Other non metallic mineral product manuf.	disp	disp	disp	agglo	agglo	agglo	agglo	agglo
3311	Iron, steel mills & ferro alloy manuf.	disp	disp	disp	disp	disp	disp	disp	agglo
3312	Steel product manuf. from purchased steel	agglo	agglo	agglo	random	agglo	random	disp	random
3313	Alumina, aluminum production & processing	agglo	random	random	random	random	random	random	random
3314	Non ferrous metal production & processing	random	random	random	random	random	random	random	random
3315	Foundries	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3321	Forging and stamping	random	random	random	random	random	random	random	random
3322	Cutlery & hand tool manufacturing	random	random	random	random	disp	disp	random	random
3323	Architectural & structural metals manuf.	random	disp	disp	disp	disp	disp	disp	disp
3324	Boiler tank & shipping container manuf.	random	random	random	random	random	random	random	random
3325	Hardware manufacturing	agglo	agglo	agglo	agglo	random	random	disp	random
3326	Spring & wire product manufacturing	random	random	random	random	random	random	random	random
3327	Machine shops turned product	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3328	Coating engraving	agglo	agglo	agglo	random	random	random	random	random
3329	Other fabricated metal product manufacturing	agglo	agglo	agglo	disp	disp	disp	disp	disp
3331	Agricultural construction & mining machinery manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3332	Industrial machinery manufacturing	agglo	agglo	agglo	agglo	random	random	random	random
3333	Commercial & service industry machinery manuf.	random	random	random	disp	random	random	random	random
3334	Ventilation heating	random	random	random	random	random	random	random	random
3335	Metalworking machinery manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3336	Engine turbine & power transmission equipment manuf.	random	random	random	disp	random	random	random	random
3339	Other general purpose machinery manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3341	Computer and peripheral equipment manufacturing	agglo	agglo	agglo	disp	disp	agglo	agglo	agglo
3342	Communications equipment manufacturing	disp	disp	disp	disp	disp	disp	disp	disp
3343	Audio and video equipment manufacturing	agglo	agglo	agglo	random	random	random	random	random
3344	Semiconductor & other electronic component manuf.	agglo	agglo	agglo	random	disp	disp	disp	random
3345	Navigational measuring	disp	disp	disp	disp	disp	disp	disp	disp
3346	Manufacturing, reproducing magnetic & optical media	random	random	random	random	random	agglo	random	random
3351	Electric lighting equipment manufacturing	disp	disp	disp	random	random	random	random	random
3352	Household appliance manufacturing	random	random	random	random	random	random	random	random
3353	Electrical equipment manufacturing	random	random	random	random	random	random	random	random
3359	Other electrical equipment & component manuf.	agglo	agglo	disp	random	random	random	random	random
3361	Motor vehicle manufacturing	agglo	agglo	agglo	agglo	random	disp	disp	disp
3362	Motor vehicle body & trailer manuf.	random	random	random	random	random	disp	random	disp
3363	Motor vehicle parts manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3364	Aerospace product & parts manufacturing	agglo	agglo	agglo	disp	disp	disp	disp	disp
3365	Railroad rolling stock manufacturing	random	random	random	random	random	random	random	random
3366	Ship and boat building	disp	disp	disp	disp	disp	agglo	agglo	agglo
3369	Other transportation equipment manufacturing	random	random	random	random	random	random	random	random
3371	Household, institutional furniture & kitchen cabinet manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3372	Office furniture (including fixtures) manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	disp
3379	Other furniture related product manuf.	random	random	random	random	disp	disp	disp	disp
3391	Medical equipment & supplies manuf.	disp	disp	disp	disp	disp	disp	disp	disp
3399	Other miscellaneous manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
	Share agglomerated	47,7%	44,2%	46,5%	48,8%	45,3%	47,7%	38,4%	40,7%
	Share random	38,4%	36,0%	36,0%	30,2%	33,7%	30,2%	36,0%	37,2%
	Share dispersed	14,0%	19,8%	17,4%	20,9%	20,9%	22,1%	25,6%	22,1%
	Percent switches between counterfactual								

Notes: This table reports the test for localization for with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.19 – Changes of classification from the benchmark to other counterfatcuals, NAICS 4-digit: Establishments (1/2)

		Switching						
NAICS3	NAMES	B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
3111	Animal food manufacturing	2 steps down	2 steps down				1 step down	1 step down
3112	Grain & oilseed milling			1 step down	1 step down	1 step down	1 step down	
3113	Sugar & confectionery product manuf.							
3114	Fruit, vegetable preserving & specialty food manuf.			1 step up	1 step up	1 step up		1 step up
3115	Dairy product manufacturing							
3116	Meat product manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3117	Seafood product preparation and packaging							
3118	Bakeries and tortilla manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3119	Other food manufacturing							
3121	Beverage manufacturing			2 steps up	2 steps up	2 steps up		
3122	Tobacco manufacturing			1 step up	1 step up	1 step up		1 step up
3131	Fibre yarn and thread mills							
3132	Fabric mills							
3133	Textile and fabric finishing and fabric coating							
3141	Textile furnishings mills			1 step up	1 step up	1 step up	1 step down	1 step down
3149	Other textile product mills							
3151	Clothing knitting mills	1 step up		1 step up	1 step up	1 step up	1 step up	1 step up
3152	Cut and sew clothing manufacturing							
3159	Clothing accessories and other clothing manuf.			1 step up	1 step up	1 step up	1 step up	1 step up
3161	Leather and hide tanning and finishing			1 step down		1 step down		
3162	Footwear manufacturing							
3169	Other leather and allied product manufacturing							
3211	Sawmills and wood preservation		2 steps up	2 steps up	2 steps up	2 steps up	2 steps up	2 steps up
3212	Veneer plywood and engineered wood product manuf.			2 steps up	2 steps up	2 steps up	2 steps up	2 steps up
3219	Other wood product manufacturing							
3221	Pulp paper and paperboard mills		1 step down	1 step down				1 step down
3222	Converted paper product manufacturing						2 steps down	
3231	Printing and related support activities			2 steps down			2 steps down	2 steps down
3241	Petroleum & coal product manufacturing				1 step down	1 step down	1 step down	
3251	Basic chemical manufacturing	2 steps up	2 steps up	2 steps up	1 step up	1 step up	1 step up	1 step up
3252	Resin synthetic rubber		1 step up					
3253	Pesticide fertilizer & other agricultural chemical manuf.	2 steps down					2 steps down	2 steps down
3254	Pharmaceutical and medicine manufacturing	1 step down		1 step down	1 step down	1 step down	1 step down	1 step down
3255	Paint coating and adhesive manufacturing				2 steps down	2 steps down		2 steps down
3256	Soap cleaning compound and toilet preparation manuf.	2 steps down	2 steps down	2 steps down	2 steps down	2 steps down	2 steps down	2 steps down
3259	Other chemical product manufacturing							
3261	Plastic product manufacturing							
3262	Rubber product manufacturing							
3271	Clay product and refractory manufacturing							
3272	Glass and glass product manufacturing							
3273	Cement and concrete product manufacturing	2 steps down						
3274	Lime and gypsum product manufacturing							

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.20 – Changes of classification from the benchmark to other counterfatcuals, NAICS 4-digit: Establishments (2/2)

		Switching						
NAICS3	NAMES	B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
3279	Other non metallic mineral product manuf.			2 steps up	2 steps up	2 steps up	2 steps up	2 steps up
3311	Iron, steel mills & ferro alloy manuf.							2 steps up
3312	Steel product manuf. from purchased steel			1 step down		1 step down	2 steps down	1 step down
3313	Alumina, aluminum production & processing	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down
3314	Non ferrous metal production & processing							
3315	Foundries							
3321	Forging and stamping							
3322	Cutlery & hand tool manufacturing				1 step down	1 step down		
3323	Architectural & structural metals manuf.	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down
3324	Boiler tank & shipping container manuf.							
3325	Hardware manufacturing				1 step down	1 step down	2 steps down	1 step down
3326	Spring & wire product manufacturing							
3327	Machine shops turned product							
3328	Coating engraving			1 step down	1 step down	1 step down	1 step down	1 step down
3329	Other fabricated metal product manufacturing			2 steps down	2 steps down	2 steps down	2 steps down	2 steps down
3331	Agricultural construction & mining machinery manuf.							
3332	Industrial machinery manufacturing				1 step down	1 step down	1 step down	1 step down
3333	Commercial & service industry machinery manuf.			1 step down				
3334	Ventilation heating							
3335	Metalworking machinery manufacturing							
3336	Engine turbine & power transmission equipment manuf.			1 step down				
3339	Other general purpose machinery manufacturing							
3341	Computer and peripheral equipment manufacturing			2 steps down	2 steps down			
3342	Communications equipment manufacturing							
3343	Audio and video equipment manufacturing			1 step down	1 step down	1 step down	1 step down	1 step down
3344	Semiconductor & other electronic component manuf.			1 step down	2 steps down	2 steps down	2 steps down	1 step down
3345	Navigational measuring							
3346	Manufacturing, reproducing magnetic & optical media					1 step up		
3351	Electric lighting equipment manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3352	Household appliance manufacturing							
3353	Electrical equipment manufacturing							
3359	Other electrical equipment & component manuf.		2 steps down	1 step down	1 step down	1 step down	1 step down	1 step down
3361	Motor vehicle manufacturing				1 step down	2 steps down	2 steps down	2 steps down
3362	Motor vehicle body & trailer manuf.					1 step down		1 step down
3363	Motor vehicle parts manufacturing							
3364	Aerospace product & parts manufacturing			2 steps down	2 steps down	2 steps down	2 steps down	2 steps down
3365	Railroad rolling stock manufacturing							
3366	Ship and boat building					2 steps up	2 steps up	2 steps up
3369	Other transportation equipment manufacturing							
3371	Household, institutional furniture & kitchen cabinet manuf.							
3372	Office furniture (including fixtures) manuf.							2 steps down
3379	Other furniture related product manuf.				1 step down	1 step down	1 step down	1 step down
3391	Medical equipment & supplies manuf.							
3399	Other miscellaneous manuf.							
	Percentage	10,5%	10,5%	36,0%	37,2%	41,9%	37,2%	41,9%

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.21 – Detailed classfication of industries, NAICS 4-digit: Employment (1/2)

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
3111	Animal food manufacturing	random	random	random	agglo	agglo	agglo	random	random
3112	Grain & oilseed milling	random	random	random	disp	disp	disp	random	random
3113	Sugar & confectionery product manuf.	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3114	Fruit, vegetable preserving & specialty food manuf.	random	random	random	agglo	agglo	agglo	disp	disp
3115	Dairy product manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
3116	Meat product manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
3117	Seafood product preparation and packaging	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3118	Bakeries and tortilla manufacturing	random	random	random	agglo	agglo	agglo	agglo	agglo
3119	Other food manufacturing	disp	disp	disp	agglo	agglo	agglo	agglo	agglo
3121	Beverage manufacturing	random	disp	disp	agglo	agglo	disp	disp	disp
3122	Tobacco manufacturing	random	random	random	random	random	random	random	random
3131	Fibre yarn and thread mills	random	random	random	random	random	random	random	random
3132	Fabric mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3133	Textile and fabric finishing and fabric coating	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3141	Textile furnishings mills	random	random	random	agglo	agglo	agglo	random	random
3149	Other textile product mills	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3151	Clothing knitting mills	agglo	agglo	random	agglo	agglo	agglo	agglo	agglo
3152	Cut and sew clothing manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3159	Clothing accessories and other clothing manuf.	random	random	random	agglo	agglo	agglo	random	random
3161	Leather and hide tanning and finishing	disp	disp	disp	random	random	random	disp	disp
3162	Footwear manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3169	Other leather and allied product manufacturing	random	random	random	random	random	random	random	random
3211	Sawmills and wood preservation	disp	disp	disp	agglo	disp	disp	agglo	disp
3212	Veneer plywood and engineered wood product manuf.	random	random	random	random	random	random	random	random
3219	Other wood product manufacturing	disp	disp	disp	disp	disp	disp	disp	disp
3221	Pulp paper and paperboard mills	random	random	random	random	random	random	random	random
3222	Converted paper product manufacturing	agglo	agglo	agglo	agglo	random	random	random	random
3231	Printing and related support activities	disp	random	random	disp	disp	disp	disp	disp
3241	Petroleum & coal product manufacturing	random	random	random	random	random	random	random	random
3251	Basic chemical manufacturing	disp	disp	disp	disp	disp	disp	disp	disp
3252	Resin synthetic rubber	agglo	agglo	agglo	random	random	random	random	random
3253	Pesticide fertilizer & other agricultural chemical manuf.	disp	disp	disp	disp	disp	disp	disp	disp
3254	Pharmaceutical and medicine manufacturing	agglo	agglo	random	agglo	random	random	random	random
3255	Paint coating and adhesive manufacturing	agglo	agglo	agglo	agglo	disp	disp	disp	disp
3256	Soap cleaning compound and toilet preparation manuf.	agglo	agglo	agglo	disp	random	random	random	random
3259	Other chemical product manufacturing	random	random	random	random	random	random	random	random
3261	Plastic product manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3262	Rubber product manufacturing	agglo	agglo	agglo	agglo	agglo	random	random	disp
3271	Clay product and refractory manufacturing	random	random	random	random	random	random	random	random
3272	Glass and glass product manufacturing	random	random	random	random	random	random	random	random
3273	Cement and concrete product manufacturing	random	disp	disp	disp	disp	disp	disp	disp
3274	Lime and gypsum product manufacturing	random	random	random	random	random	random	random	random

Notes: This table reports the test for localization for with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.22 – Detailed classfication of industries, NAICS 4-digit: Employment (2/2)

		Counterfactuals							
NAICS3	NAMES	Benchmark(B)	Job(DO5)	Job(J5)	Deter(D4)	Deter(D5)	Deter(D6)	Deter(D7)	Deter(D8)
3279	Other non metallic mineral product manuf.	random	random	random	random	random	disp	random	random
3311	Iron, steel mills & ferro alloy manuf.	random	random	random	random	random	random	random	random
3312	Steel product manuf. from purchased steel	agglo	agglo	agglo	agglo	random	random	random	agglo
3313	Alumina, aluminum production & processing	agglo	agglo	agglo	random	random	random	random	random
3314	Non ferrous metal production & processing	disp	disp	disp	random	disp	random	disp	disp
3315	Foundries	random	random	random	random	random	random	random	random
3321	Forging and stamping	random	random	random	random	random	random	random	random
3322	Cutlery & hand tool manufacturing	agglo	agglo	agglo	agglo	random	random	random	random
3323	Architectural & structural metals manuf.	random	random	random	random	random	random	random	random
3324	Boiler tank & shipping container manuf.	disp	disp	disp	disp	agglo	agglo	random	agglo
3325	Hardware manufacturing	agglo	agglo	agglo	agglo	random	random	random	random
3326	Spring & wire product manufacturing	random	random	random	random	random	random	random	random
3327	Machine shops turned product	random	random	random	random	random	disp	disp	random
3328	Coating engraving	random	random	random	random	random	random	random	random
3329	Other fabricated metal product manufacturing	random	random	random	random	random	random	random	random
3331	Agricultural construction & mining machinery manuf.	disp	disp	disp	disp	agglo	agglo	disp	agglo
3332	Industrial machinery manufacturing	random	random	random	random	random	random	random	random
3333	Commercial & service industry machinery manuf.	random	agglo	random	agglo	random	random	random	random
3334	Ventilation heating	disp	disp	disp	random	random	random	random	random
3335	Metalworking machinery manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3336	Engine turbine & power transmission equipment manuf.	random	random	random	random	random	random	random	random
3339	Other general purpose machinery manufacturing	agglo	agglo	agglo	agglo	disp	disp	agglo	agglo
3341	Computer and peripheral equipment manufacturing	disp	disp	disp	random	random	random	random	random
3342	Communications equipment manufacturing	random	random	random	random	random	random	random	random
3343	Audio and video equipment manufacturing	agglo	agglo	agglo	agglo	random	random	random	random
3344	Semiconductor & other electronic component manuf.	agglo	agglo	agglo	random	random	disp	random	random
3345	Navigational measuring	random	disp	disp	disp	disp	disp	random	random
3346	Manufacturing, reproducing magnetic & optical media	random	random	random	random	random	random	random	random
3351	Electric lighting equipment manufacturing	random	random	random	random	random	random	random	random
3352	Household appliance manufacturing	random	random	random	random	random	random	random	random
3353	Electrical equipment manufacturing	random	random	random	random	random	random	random	random
3359	Other electrical equipment & component manuf.	random	random	random	random	random	random	random	random
3361	Motor vehicle manufacturing	agglo	agglo	agglo	random	random	random	random	random
3362	Motor vehicle body & trailer manuf.	random	random	random	random	random	random	random	random
3363	Motor vehicle parts manufacturing	agglo	agglo	agglo	agglo	agglo	agglo	agglo	agglo
3364	Aerospace product & parts manufacturing	random	disp	random	random	random	random	random	random
3365	Railroad rolling stock manufacturing	random	random	random	random	random	random	random	random
3366	Ship and boat building	random	random	disp	random	random	random	random	random
3369	Other transportation equipment manufacturing	random	random	random	random	random	random	random	random
3371	Household, institutional furniture & kitchen cabinet manuf.	random	random	random	random	random	random	random	random
3372	Office furniture (including fixtures) manuf.	agglo	agglo	agglo	agglo	random	disp	disp	disp
3379	Other furniture related product manuf.	random	random	random	random	random	random	random	random
3391	Medical equipment & supplies manuf.	disp	random	disp	random	random	random	random	random
3399	Other miscellaneous manuf.	random	random	random	random	random	random	random	random
	Share agglomerated	30,2%	31,4%	27,9%	37,2%	26,7%	24,4%	19,8%	22,1%
	Share random	54,7%	51,2%	53,5%	51,2%	60,5%	58,1%	65,1%	62,8%
	Share dispersed	15,1%	17,4%	18,6%	11,6%	12,8%	17,4%	15,1%	15,1%
	Percent switches between counterfactual								

Notes: This table reports the test for localization for with different counterfactuals. Benchmark(B) uses the unconstrained counterfactual. Job(DO5) uses a counterfactual based on bins of number of workers as defined in Duranton and Overman (2008), Job(J5) uses a counterfactual based on 5 bins of workers as suggested by the HAC procedure. Deter(Dx) uses a counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.23 – Changes of classification from the benchmark to other counterfatcuals, NAICS 3-digit: Employment (1/2)

		Switching						
NAICS3	NAMES	B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
3111	Animal food manufacturing			1 step up	1 step up	1 step up		
3112	Grain & oilseed milling			1 step down	1 step down	1 step down		
3113	Sugar & confectionery product manuf.							
3114	Fruit, vegetable preserving & specialty food manuf.			1 step up	1 step up	1 step up	1 step down	1 step down
3115	Dairy product manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3116	Meat product manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3117	Seafood product preparation and packaging							
3118	Bakeries and tortilla manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3119	Other food manufacturing			2 steps up	2 steps up	2 steps up	2 steps up	2 steps up
3121	Beverage manufacturing	1 step down	1 step down	1 step up	1 step up	1 step down	1 step down	1 step down
3122	Tobacco manufacturing							
3131	Fibre yarn and thread mills							
3132	Fabric mills							
3133	Textile and fabric finishing and fabric coating							
3141	Textile furnishings mills			1 step up	1 step up	1 step up		
3149	Other textile product mills							
3151	Clothing knitting mills		1 step down					
3152	Cut and sew clothing manufacturing							
3159	Clothing accessories and other clothing manuf.			1 step up	1 step up	1 step up		
3161	Leather and hide tanning and finishing			1 step up	1 step up	1 step up		
3162	Footwear manufacturing							
3169	Other leather and allied product manufacturing							
3211	Sawmills and wood preservation			2 steps up			2 steps up	
3212	Veneer plywood and engineered wood product manuf.							
3219	Other wood product manufacturing							
3221	Pulp paper and paperboard mills							
3222	Converted paper product manufacturing				1 step down	1 step down	1 step down	1 step down
3231	Printing and related support activities	1 step up	1 step up					
3241	Petroleum & coal product manufacturing							
3251	Basic chemical manufacturing							
3252	Resin synthetic rubber			1 step down	1 step down	1 step down	1 step down	1 step down
3253	Pesticide fertilizer & other agricultural chemical manuf.							
3254	Pharmaceutical and medicine manufacturing		1 step down		1 step down	1 step down	1 step down	1 step down
3255	Paint coating and adhesive manufacturing				2 steps down	2 steps down	2 steps down	2 steps down
3256	Soap cleaning compound and toilet preparation manuf.			2 steps down	1 step down	1 step down	1 step down	1 step down
3259	Other chemical product manufacturing							
3261	Plastic product manufacturing							
3262	Rubber product manufacturing					1 step down	1 step down	2 steps down
3271	Clay product and refractory manufacturing							
3272	Glass and glass product manufacturing							
3273	Cement and concrete product manufacturing	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down	1 step down
3274	Lime and gypsum product manufacturing							

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

Table 3.24 – Changes of classification from the benchmark to other counterfatcuals, NAICS 3-digit: Employment (2/2)

		Switching						
NAICS3	NAMES	B - DO5	B - J5	B - D4	B - D5	B - D6	B - D7	B - D8
3279	Other non metallic mineral product manuf.					1 step down		
3311	Iron, steel mills & ferro alloy manuf.							
3312	Steel product manuf. from purchased steel				1 step down	1 step down	1 step down	
3313	Alumina, aluminum production & processing			1 step down	1 step down	1 step down	1 step down	1 step down
3314	Non ferrous metal production & processing			1 step up		1 step up		
3315	Foundries							
3321	Forging and stamping							
3322	Cutlery & hand tool manufacturing				1 step down	1 step down	1 step down	1 step down
3323	Architectural & structural metals manuf.							
3324	Boiler tank & shipping container manuf.				2 steps up	2 steps up	1 step up	2 steps up
3325	Hardware manufacturing				1 step down	1 step down	1 step down	1 step down
3326	Spring & wire product manufacturing							
3327	Machine shops turned product					1 step down	1 step down	
3328	Coating engraving							
3329	Other fabricated metal product manufacturing							
3331	Agricultural construction & mining machinery manuf.				2 steps up	2 steps up		2 steps up
3332	Industrial machinery manufacturing							
3333	Commercial & service industry machinery manuf.	1 step up		1 step up				
3334	Ventilation heating			1 step up	1 step up	1 step up	1 step up	1 step up
3335	Metalworking machinery manufacturing							
3336	Engine turbine & power transmission equipment manuf.							
3339	Other general purpose machinery manufacturing				2 steps down	2 steps down		
3341	Computer and peripheral equipment manufacturing			1 step up	1 step up	1 step up	1 step up	1 step up
3342	Communications equipment manufacturing							
3343	Audio and video equipment manufacturing				1 step down	1 step down	1 step down	1 step down
3344	Semiconductor & other electronic component manuf.			1 step down	1 step down	2 steps down	1 step down	1 step down
3345	Navigational measuring	1 step down	1 step down	1 step down	1 step down	1 step down		
3346	Manufacturing, reproducing magnetic & optical media							
3351	Electric lighting equipment manufacturing							
3352	Household appliance manufacturing							
3353	Electrical equipment manufacturing							
3359	Other electrical equipment & component manuf.							
3361	Motor vehicle manufacturing			1 step down	1 step down	1 step down	1 step down	1 step down
3362	Motor vehicle body & trailer manuf.							
3363	Motor vehicle parts manufacturing							
3364	Aerospace product & parts manufacturing	1 step down						
3365	Railroad rolling stock manufacturing							
3366	Ship and boat building		1 step down					
3369	Other transportation equipment manufacturing							
3371	Household, institutional furniture & kitchen cabinet manuf.							
3372	Office furniture (including fixtures) manuf.				1 step down	2 steps down	2 steps down	2 steps down
3379	Other furniture related product manuf.							
3391	Medical equipment & supplies manuf.	1 step up		1 step up	1 step up	1 step up	1 step up	1 step up
3399	Other miscellaneous manuf.							
	Percentage	8,1%	8,1%	27,9%	37,2%	41,9%	31,4%	29,1%

Notes: This table reports the changes in classification when comparing the test for localization from the benchmark counterfactuals to alternative counterfactual. *B-A* refers to the comaprison between Benchmark(B) counterfactual and the alternative *A*. *A* being one of the following : (i) DO5 : The counterfactual based on 5 bins of number of workers as defined in Duranton and Overman (2008), (ii) J5 : The counterfactual based on 5 bins of workers as suggested by the HAC procedure, (iii) Dx : The counterfactual based on x bins as suggested by the HAC procedure performed with firms' location determinants. x=4, 5, 6, 7, and 8.

3.6.6 Hierarchical Ascendant Classification

Procedure

We present here step by step the procedure of Hierarchical Ascendant Classification based on the technical package of the "Fastcluster" that we use on the statistical open-source software r (see Mullner 2021).

The starting point is the dataset of 23,388 Canadian establishments with information on the number of workers of the establishment, the employment from the same industry in 10km around the plant's location, the number of plants from the same industry in 10km around the plant's location, the population density in 1.5km around the plant, the building footprint of the site location, the distance to the nearest junction, a dummy indicating if the plant is located in a Census Metropolitan Area, the province of the site location, the zoning type of the location, the Business type of the plant, the export and the headquarter status of the plant.

We look now at each of the 23,388 observations as locations and the variables as their characteristics. Then the exercise is to construct classes of locations based on their characteristics. We use the Hierarchical Ascendant Classification which consists in aggregating progressively, the observations that are similar with respect to their characteristics, starting from individual observations up to a unique group. The process is the following: we represent our dataset of locations as a data frame with locations in rows and characteristics $X_i \, \varepsilon$ {Number of workers of the plant, Employment from the same industry in 10km around the plant's location, Number of plants from the same industry in 10km around the plant's location, Population density in 1.5km around the plant, Building footprint of the site location, Distance to the nearest junction, Dummy indicating if the plant is located in a Census Metropolitan Area, Province of the site location, Zoning type of the location, Business type of the plant, Dummy indi-

cating if the plant exports, Dummy indicating if the plant is the headquarter} in columns.

— We transform all non-numeric variables into dummies. For any given non-numeric variable, each of its categories becomes a dummy. The non-numeric variables concerned by this transformation are: Province; Zoning categories(industrial/commercial, residential, others); Business type of the plant(distribution, manufacturer, others). The other non-numeric variables are already dummies: Exporter, Headquarter, CMA. This gives a total of 45 variables

— We scale all the variables by standardizing each variables, obtain the following standardized dataset where $\tilde{x_i} = \frac{X_i - \bar{X_i}}{\sigma_X}$. So that the table becomes

id	$\tilde{x_1}$	$\tilde{x_2}$	...	$\tilde{x_i}$	...	$\tilde{x_{44}}$	$\tilde{x_{45}}$
1	$\tilde{x_{11}}$	$\tilde{x_{12}}$	...		.	...	.
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
k	$\tilde{x_{k1}}$	.	...	$\tilde{x_{ki}}$	...	$\tilde{x_{k44}}$	$\tilde{x_{k45}}$
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
23388	$\tilde{x_{23388.1}}$	.	...	$\tilde{x_{23388.i}}$	...	$\tilde{x_{23388.44}}$	$\tilde{x_{23388.45}}$

— We then compute the dissimilarity between sites as the euclidean distances between sites. For any given pair of sites (j,k). The dissimilarity distance between site j and site k is :

$$d_{jk} = \sqrt{\sum_{j=1}^{45} (\tilde{x_{ji}} - \tilde{x_{ki}})^2}$$

This gives us a symmetric matrix of 273,487,580 pairwise distances between all the pairs of site (j, k) ε 1, 2, .., 23388.

— The pair of sites with the minimal similarity distance are aggregated as one unique point in the dataset. The new dataset now has 23405 points and the two points that were aggregated constitute now a single point with characteristics.

— To move to the next level of aggregation, we need to define how to merge two nodes since from the 2 levels of aggregation and onward, a node can well be a cluster of many single sites. We chose the Ward method defined as follows: Let d be the dissimilarity distance, I and J the nodes (which may contain many sites) to be joined, and K the resulting node of the aggregation of I and J and L any other node. Let call $|I|$ the size of node I for example. With the Ward method, the distance between K and L is given by :

$$d(K,L) = \frac{(|I|+|L|).d(I,L)+(|J|+|L|).d(J,L)-|L|.d(I,J)}{|I|+|J|+|L|}$$

and the global cluster dissimilarity can be express as :

$$d(A,B) = \frac{2|A||B|}{|A|+|B|}.||\overrightarrow{C_A} + \overrightarrow{C_B}||^2$$

where $\overrightarrow{C_A}$ denotes the centroid of the points in cluster A.

— We continu the process of aggregating nodes up to the final aggregation into one unique group

The Gap statistics method for choosing the optimal number of clusters

The presentation of this section is inspired from Tibshirani et al. (2001) Suppose that we have p classes C1, C2, ..Cr, ...Cp, and let $nr = |Cr|$ be the number of observations in cluster r. We define :

$$D_r = \sum_{j,j' \varepsilon C_r} d_{jj'},$$

the sum of the pairwise distances for all points in cluster r, and

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

Since $d_{jj'}$ is the Euclidean distance between sites j and k, W_k is the pooled within sum of squares around the clusters means.

Tibshirani et al. (2001) suggest, standardizing the graph of $log(W_k)$by comparing it with its expectation under an appropriate null reference distribution of the data. Then, the optimal number of clusters will be the value of k for which $log(W_k)$ farthest away from the curve derived from the null reference distribution.

In other words, if

$$Gap_n(k) = E_n^* log(W_k) - log(W_k)$$

is the value of the difference between the $log(W_k)$ and its expected value under the null null reference distribution of the data, then k is the value that maximizes the value of $Gap_n(k)$. The reference distribution of the data used to solve this maximization problem is the uniform distribution.

3.6.7 Adjusted procedures of R packages to estimate the constrained counterfactuals

We use the package *dbmss* from the open-source software r to estimate the K-density (see Marcon et al. 2015). For the estimation of the new counterfactual, we modify some sections of the original code obtained online from the R package documentation (see https://rdrr.io/cran/dbmss/src/R/KdEnvelope.R). We present hereafter the adjustment that we made original and the adjusted codes.
We start by modifying the original code for the creation of objects of Class "Dtable" which are input for the K-density code in R. This code combines altogether three elements : (i) the symmetric bilateral *distances Matrix*, (ii) the *type* of the observations (here the NAICS of industries) and (iii) the *weight* to be used (here the employment). Then it converts all these elements into an object of Class "Dtable" usable in the procedure "KdEnvelop" that we present here below. We modified the code to obtain another one which includes the *Categories* of the observations (here the classes of locations obtained from the HAC procedure) in addition to the three previous elements.
We also modify the original code for the random selection of locations across the locations universe that generates the unconstraint counterfactual (Our benchmark). The original code yields a random draw of points across a given set of points. We modify it and obtain a new version of code which is such that, for each point of a given *Class*, the code realizes a random draw of points in the same *Class*.
Finally, we use the code that estimates the K-densities of bilateral distances, along with the confidence intervals. We turn it into a new code by adjusting the parameters and the procedures so as to include, the point categories and the modified version of the random draws of locations for counterfactual.
The original and the modified codes are available upon request.

CONCLUSION

Ce travail de recherche avait pour objectif, de contribuer à la compréhension des mécanismes sous-jacents à la concentration spatiale de l'activité économique. Nous nous sommes alors intéressés à trois aspects de ce phénomène : la mesure de la concentration, ses causes et ses conséquences.

Au sujet de la mesure de la concentration, des contributions récentes, notamment celles de Ellison and Glaeser (1997); Duranton and Overman (2005, 2008) ont ouvert la voie à une nouvelle génération de mesures de la concentration spatiale de l'activité industrielle. La pertinence de ces mesures tient essentiellement du fait qu'elles sont comparables à travers les industries, permettent de détecter la concentration au-delà de la simple inégalité spatiale, ne se sont pas sensibles au découpage spatial et sont statistiquement testables. Dans le cadre de cette thèse, nous avons raffiné le test statistique proposé par Duranton and Overman (2008) en proposant une méthode plus éclairée de séparer la concentration réelle de la simple inégalité spatiale.

Au sujet des causes de la concentration industrielle, nous nous sommes inspirés du travail de Ellison et al. (2010) sur les déterminants de la colocation des pairs d'industries du point de vue de la production, pour discuter de l'effet de l'échange des intrants, du partage d'un bassin commun de travailleurs et de la diffusion des connaissances sur la co-localisation de l'innovation. Ceci nous a permis de mettre en évidence des différences dans les déterminants de la concentration de l'innovation et celle de la production : une part non négligeable de la concentration de l'innovation est le fruit de la concentration de la production. Une fois la concentration de la production prise en compte, seule

la diffusion des connaissances a un effet positif et significatif sur la concentration de l'innovation. S'agissant des causes de la concentration de l'innovation, une avenue pour des travaux futurs serait de discuter l'hétérogénéité de l'effet des déterminants de l'innovation mis en évidence dans ce travail. Un tel travail pourrait s'inspirer d'un cadre conceptuel similaire à celui de Faggio et al. (2017, 2020) qui a discuté des effets hétérogènes des forces Marshalliennes sur le production.

Enfin, nous avons exploré les effets de la concentration industrielle sur l'activité productive des entreprises manufacturières au Canada. Nous avons proposé une méthode pour construire des données inexistantes sur le foncier utilisé par les entreprises manufacturières canadiennes pour leur besoin de production. Ensuite nous avons utilisé la mesure construite pour documenter des faits stylisés nouveaux sur l'utilisation du foncier par les firmes manufacturières. Nos résultats indiquent qu'il existe un coût fixe et un coût variable dans l'utilisation du foncier par les entreprises, et ces dernières n'ajustent pas facilement la quantité d'espace qu'elles utilisent pour produire. Une étape future serait d'utiliser la mesure de foncier utilisées dans ce travail pour estimer une fonction de production où le foncier est un facteur de production distinct du capital. Mais ceci nécessiterait de relever plusieurs défis économétriques et de collecte de données, notamment celles sur le capital des entreprises.

BIBLIOGRAPHY

Aghion, P., Akcigit, U. and Howitt, P. (2014). What do we learn from Schumpeterian growth theory? In *Handbook of Economic Growth*, volume 2 pp. 515–563. Elsevier.

Aghion, P. and Howitt, P. (2005). Growth with quality-improving innovations: An integrated framework. *Handbook of economic growth*, *1*, 67–110.

Ahlfeldt, G. M. and McMillen, D. P. (2015). The vertical city: the price of land and the height of buildings in Chicago 1870-2010.

Albouy, D., Ehrlich, G. and Shin, M. (2018). Metropolitan Land Values. *The Review of Economics and Statistics*, *100*(3), 454–466.

Aleksandrova, E., Behrens, K. and Kuznetsova, M. (2020). Manufacturing (co) agglomeration in a transition country: Evidence from Russia. *Journal of Regional Science*, *60*(1), 88–128.

Anas, A., Arnott, R. and Small, K. A. (1998). Urban spatial structure. *Journal of Economic literature*, *36*(3), 1426–1464.

Arauzo Carod, J. M., Liviano Solis, D. and Manjon-Antolin, M. (2010). Empirical Studies in Industrial Location: An Assessment of Their Methods and Results*. *Journal of Regional Science*, *50*(3), 685–711.

Audretsch, D. B. and Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *The American economic review*, *86*(3), 630–640.

Audretsch, D. B. and Feldman, M. P. (2004). Knowledge spillovers and the geography of innovation. In *Handbook of Regional and Urban Economics*, volume 4 pp. 2713–2739. Elsevier.

Autant-Bernard, C. (2006). Where do firms choose to locate their R&D? A spatial conditional logit analysis on French data. *European planning studies*, *14*(9), 1187–1208.

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

Barrios, S., Görg, H. and Strobl, E. (2006). Multinationals' location choice, agglomeration economies, and public incentives. *International Regional Science Review*, *29*(1), 81–107.

Basile, R., Castellani, D. and Zanfei, A. (2008). Location choices of multinational firms in Europe: The role of EU cohesion policy. *Journal of International Economics*, *74*(2), 328–340.

Behrens, K., Boualam, B. and Martin, J. (2020). Are clusters resilient? Evidence from Canadian textile industries. *Journal of Economic Geography*, *20*(1), 1–36.

Behrens, K. and Bougna, T. (2015). An anatomy of the geographical concentration of Canadian manufacturing industries. *Regional Science and Urban Economics*, *51*, 47–69.

Behrens, K., Duranton, G. and Robert-Nicoud, F. (2014). Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy*, *122*(3), 507–553.

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

Blackley, P. R. (1985). The demand for industrial sites in a metropolitan area: Theory, empirical evidence, and policy implications. *Journal of Urban Economics*, *17*(2), 247–261.

Brinkman, J. C. (2016). Congestion, agglomeration, and the structure of cities. *Journal of Urban Economics*, *94*, 13–31.

Brooks, L. and Lutz, B. (2016). From today's city to tomorrow's city: An empirical investigation of urban land assembly. *American Economic Journal: Economic Policy*, *8*(3), 69–105.

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

Carlino, G. and Kerr, W. R. (2015). Agglomeration and innovation. *Handbook of regional and urban economics*, *5*, 349–404.

Carlino, G. A., Chatterjee, S. and Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics*, *61*(3), 389–419.

Carlton, D. W. (1983). The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables. *The Review of Economics and Statistics*, pp. 440–449.

Carozzi, F. (2020). The role of demand in land re-development. *Journal of Urban Economics*, *117*(103244), 1–14.

Chaney, T., Sraer, D. and Thesmar, D. (2012). The collateral channel: How real estate shocks affect corporate investment. *The American Economic Review*, *102*(6), 2381–2409.

Cheshire, P. and Sheppard, S. (2004). Land markets and land market regulation: progress towards understanding. *Regional Science and Urban Economics*, *34*(6), 619–637.

Cheshire, P. C., Hilber, C. A. and Kaplanis, I. (2014). Land use regulation and productivity—land matters: evidence from a UK supermarket chain. *Journal of Economic Geography*, *15*(1), 43–73.

Chinitz, B. (1961). Contrasts in agglomeration: New york and pittsburgh. *The American Economic Review*, *51*(2), 279–289.

Combes, P.-P. and Duranton, G. (2006). Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics*, *36*(1), 1–28.

Combes, P.-P., Duranton, G. and Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of urban economics*, *63*(2), 723–742. Publisher: Elsevier.

Combes, P.-P., Duranton, G. and Gobillon, L. (2019). The Costs of Agglomeration: House and Land Prices in French Cities. *The Review of Economic Studies*, *86*(4), 1556–1589. http://dx.doi.org/10.1093/restud/rdy063

Coughlin, C. C. and Segev, E. (2000). Location Determinants of New Foreign-Owned Manufacturing Plants. *Journal of Regional Science*, *40*(2), 323–351.

Disdier, A.-C. and Mayer, T. (2004). How different is Eastern Europe? Structure and determinants of location choices by French firms in Eastern and Western Europe. *Journal of Comparative Economics*, *32*(2), 280–296.

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

Duranton, G., Ghani, S. E., Grover, A. G. and Kerr, W. R. (2015). *The misallocation of land and other factors of production in India*. Technical report.

Duranton, G. and Overman, H. G. (2005). Testing for localization using micro-geographic data. *The Review of Economic Studies*, *72*(4), 1077–1106.

Duranton, G. and Overman, H. G. (2008). Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *Journal of Regional Science*, *48*(1), 213–243.

Duranton, G. and Puga, D. (2001). Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*, *91*(5), 1454–1477.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4 pp. 2063–2117. Elsevier.

Duranton, G. and Puga, D. (2015). Urban land use. In *Handbook of regional and urban economics*, volume 5 pp. 467–560. Elsevier.

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of political economy*, *105*(5), 889–927.

Ellison, G. and Glaeser, E. L. (1999). The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review*, *89*(2), 311–316.

Ellison, G., Glaeser, E. L. and Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, *100*(3), 1195–1213.

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

Evenson, R. E., Putnam, J. and Kortum, S. (1991). Estimating patent counts by industry using the Yale-Canada concordance. *final report to the National Science Foundation*.

Faggio, G., Silva, O. and Strange, W. C. (2017). Heterogeneous agglomeration. *Review of Economics and Statistics*, *99*(1), 80–94.

Faggio, G., Silva, O. and Strange, W. C. (2020). Tales of the city: What do agglomeration cases tell us about agglomeration in general? *Journal of Economic Geography*, *20*(5), 1117–1143.

Fallick, B., Fleischman, C. A. and Rebitzer, J. B. (2006). Job-hopping in Silicon Valley: Some evidence concerning the microfoundations of a high-technology cluster. *The review of economics and statistics*, *88*(3), 472–481.

Feldman, M. P. (1994). *The Geography of Innovation*, volume 2. Springer Science & Business Media.

Feldman, M. P. and Kogler, D. F. (2010). Stylized facts in the geography of innovation. *Handbook of the Economics of Innovation*, *1*, 381–410.

Figueiredo, O., Guimaraes, P. and Woodward, D. (2002). Home-field advantage: Location decisions of Portuguese entrepreneurs. *Journal of Urban Economics*, *52*(2), 341–361.

Friedman, J., Gerlowski, D. A. and Silberman, J. (1992). What attracts foreign multinational corporations? Evidence from branch plant location in the United States. *Journal of Regional science*, *32*(4), 403–418.

Fujita, M. (1989). *Urban economic theory*. Cambridge University Press.

Fujita, M., Krugman, P. R. and Venables, A. (1999). *The Spatial Economy: Cities, Regions, and International Trade*. MIT press.

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

Fujita, M. and Thisse, J.-F. (2002). Agglomeration and market interaction. *Available at SSRN 315966*.

Fujita, M. and Thisse, J.-F. c. (2013). *Economics of Agglomeration: Cities, Industrial Location, and Globalization*. Cambridge University Press, 2nd edition.

Gabe, T. M. and Bell, K. P. (2004). Tradeoffs between local taxes and government spending as determinants of business location. *Journal of Regional Science*, *44*(1), 21–41.

Gan, J. (2007). Collateral, debt capacity, and corporate investment: Evidence from a natural experiment. *Journal of Financial Economics*, *85*(3), 709–734.

Ganguli, I., Lin, J. and Reynolds, N. (2020). The paper trail of knowledge spillovers: Evidence from patent interferences. *American Economic Journal: Applied Economics*, *12*(2), 278–302.

Gerlach, H., Rønde, T. and Stahl, K. (2009). Labor pooling in R&D intensive industries. *Journal of Urban Economics*, *65*(1), 99–111.

Glaeser, E. L., Gyourko, J. and Saks, R. (2005). Why is manhattan so expensive? regulation and the rise in housing prices. *Journal of Law and Economics*, *48*(2), 331–369.

Graham, D. J. (2007). Agglomeration, productivity and transport investment. *Journal of transport economics and policy (JTEP)*, *41*(3), 317–343.

Hall, B. H. (2011). *Innovation and Productivity*. Technical report, National bureau of economic research.

Hall, R. E. (2004). Measuring factor adjustment costs. *Quarterly Journal of Economics*, *119*(3), 899–927.

Hamermesh, D. S. and Pfann, G. A. (1996). Adjustment costs in factor demand. *Journal of Economic Literature*, *34*(3), 1264–1292.

Hansen, E. R. (1987). Industrial location choice in Sao Paulo, Brazil: A nested logit model. *Regional science and Urban economics*, *17*(1), 89–108.

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

Head, C. K., Ries, J. C. and Swenson, D. L. (1999). Attracting foreign manufacturing: Investment promotion and agglomeration. *Regional Science and Urban Economics*, *29*(2), 197–218.

Head, K., Ries, J. and Swenson, D. (1995). Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States. *Journal of international economics*, *38*(3-4), 223–247.

Helsley, R. W. and Strange, W. C. (1990). Matching and agglomeration economies in a system of cities. *Regional Science and urban economics*, *20*(2), 189–212.

Helsley, R. W. and Strange, W. C. (2002). Innovation and input sharing. *Journal of Urban Economics*, *51*(1), 25–45.

Henderson, V., Kuncoro, A. and Turner, M. (1995). Industrial development in cities. *Journal of political economy*, *103*(5), 1067–1090.

Hilber, C. A. and Robert-Nicoud, F. (2013). On the origins of land use regulations: theory and evidence from US metro areas. *Journal of Urban Economics*, *75*, 29–43.

Howard, E., Newman, C. and Tarp, F. (2016). Measuring industry coagglomeration and identifying the driving forces. *Journal of Economic Geography*, *16*(5), 1055–1078.

Jaffe, A. B., Trajtenberg, M. and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, *108*(3), 577–598.

Kelly, M. and Hageman, A. (1999). Marshallian externalities in innovation. *Journal of economic growth*, *4*(1), 39–54.

Kerr, W. R. and Kominers, S. D. (2015). Agglomerative forces and cluster shapes. *Review of Economics and Statistics*, *97*(4), 877–899.

Klier, T. and McMillen, D. P. (2008). Evolving Agglomeration in the U.s. Auto Supplier Industry*. *Journal of Regional Science*, *48*(1), 245–267.

Kortum, S. and Putnam, J. (1997). Assigning Patents to Industries: Tests of the Yale Technology Concordance. *Economic Systems Research*, *9*(2), 161–176.

Krugman, P. R. (1991). *Geography and Trade*. MIT press.

Lan, T. (2019). *The Coagglomeration of Innovation and Production*. Technical report, mimeo, University of Michigan.

Landier, A. (2005). Entrepreneurship and the Stigma of Failure. *Available at SSRN 850446*.

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

Marcon, E. and Puech, F. (2003). Evaluating the geographic concentration of industries using distance-based methods. *Journal of economic geography*, *3*(4), 409–428.

Marcon, E., Traissac, S., Puech, F. and Lang, G. (2015). Tools to characterize point patterns: Dbmss for R. *Journal of Statistical Software*, *67*(3), 1–15.

Marshall, A. (1890). *Principles of Economics, by Alfred Marshall*. Macmillan and Company.

Mayer, T., Mayneris, F. and Py, L. (2015). The impact of Urban Enterprise Zones on establishment location decisions and labor market outcomes: Evidence from France. *Journal of Economic Geography*, pp. lbv035.

McConnell, V. D. and Schwab, R. M. (1990). The impact of environmental regulation on industry location decisions: The motor vehicle industry. *Land Economics*, *66*(1), 67–81.

Mota, I. and Brandão, A. (2013). The determinants of location choice: Single plants versus multi-plants. *Papers in Regional Science*, *92*(1), 31–49.

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

Mullner, D. (2021). The fastcluster package: User's manual.

Murata, Y., Nakajima, R., Okamoto, R. and Tamura, R. (2014). Localized knowledge spillovers and patent citations: A distance-based approach. *Review of Economics and Statistics*, *96*(5), 967–985.

Neary, J. P. (2001). Of hype and hyperbolas: Introducing the new economic geography. *Journal of economic Literature*, *39*(2), 536–561.

Rosenthal, S. S. and Strange, W. C. (2001). The Determinants of Agglomeration. *Journal of Urban Economics*, *50*(2), 191–229.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. *Handbook of regional and urban economics*, *4*, 2119–2171.

Saiz, A. (2010). The geographic determinants of housing supply. *Quarterly Journal of Economics*, *125*(3), 1253–1296.

Saxenian, A. (1994). Regional networks: Industrial adaptation in Silicon Valley and route 128.

Schmenner, R. W., Huber, J. C. and Cook, R. L. (1987). Geographic differences and the location of new manufacturing facilities. *Journal of Urban Economics*, *21*(1), 83–104.

Segerstrom, P. S. (1991). Innovation, imitation, and economic growth. *Journal of political economy*, *99*(4), 807–827.

Simonen, J. and McCann, P. (2008). Firm innovation: The influence of R&D cooperation and the geography of human capital inputs. *Journal of Urban Economics*, *64*(1), 146–154.

Sood, A. (2020). *Land Market Frictions in Developing Countries: Evidence from Manufacturing Firms in India*. Technical report.

Tabuchi, T. and Yoshida, A. (2000). Separating Urban Agglomeration Economies in Consumption and Production. *Journal of Urban Economics*, *48*(1), 70–84.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, *95*(1), 450–460.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.