UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESSAIS EN ÉCONOMIE SPATIALE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN ÉCONOMIQUE

PAR

THÉOPHILE BOUGNA LONLA

SEPTEMBRE 2016

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESSAYS IN SPATIAL ECONOMICS AND CLUSTERING

THESIS

PRESENTED

AS A PARTIAL REQUIREMENT

OF DOCTORAL OF PHILOSOPHY IN ECONOMICS

BY

THÉOPHILE BOUGNA LONLA

SEPTEMBER 2016

REMERCIEMENTS

L'élaboration et l'aboutissement d'une thèse est un travail de longue haleine. Au terme de cette belle et enrichissante expérience, j'aimerais remercier les personnes qui, de près ou de loin, directement ou indirectement, y ont contribué. En premier et tout particulièrement, je voudrais sincèrement remercier mon directeur de thèse Kristian Behrens pour la rigueur, la discipline, la disponibilité, l'appui financier et le soutien sans faille qui m'ont permis de surmonter les obstacles rencontrés au cours de cette expérience. J'en profite également pour lui exprimer toute ma gratitude pour l'appui qu'il m'a apporté lorsque je me suis présenté sur le marché du travail. Deux chapitres de cette thèse sont co-écrits avec Kristian. Ces deux travaux ont été pour moi l'occasion d'apprendre et de comprendre les exigences, les efforts, la discipline et la rigueur que nécessitent le choix d'une approche méthodologique et l'écriture d'un article de recherche de bonne qualité. Au-delà de la qualité d'encadrement, travailler sous sa supervision a été un réel plaisir, une formidable expérience d'apprentissage et de collaboration tant sur le plan humain que scientifique. Merci Kristian !

Je remercie Gilles Duranton qui a accepté d'évaluer cette thèse et pour tout l'appui qu'il m'a apporté lorsque je me suis présenté sur le marché du travail. La plupart des travaux de cette thèse sont inspirés et basés sur les mesures de concentration spatiale développées par Gilles Duranton et Henry Overman en 2005. Mes remerciements vont également à l'endroit de Wilfried Koch et Martin Julien Frédéric, professeurs au département des sciences économiques de l'UQAM qui ont accepté d'évaluer cette thèse et qui m'ont soutenu lorsque je me suis présenté sur le marchée du travail. Je tiens également à remercier Mark Brown, Chef - spécialiste, à la Direction des études analytique de Statistique Canada. En plus de faciliter mon intégration à Statistique Canada, il a toujours été disponible pour échanger. Le deuxième article de cette thèse a été co-écrit avec Mark Brown et Kristian Behrens.

Mes remerciements vont également à l'endroit des professeurs du département des sciences économiques de l'UQAM. En particulier, Pavel Sevcik, Pierre-Yves Yanni, Louis Phaneuf, Charles Séguin, Dalibor Stevanovic, Jean-Denis Garon, Duppe Till et Raquel Fonseca qui m'ont soutenu et préparé avant l'entrée sur le marché du travail. Je dis merci à mes enseignants : Max Blouin, Alain Delacroix, Claude Fluet, Arianna Degan, Alain Guay, Pierre-Carl Michaud, Stephane Pallage, Alessandro Barattieri, Alain Paquet, Douglas-James Hodgson et Victoria Miller. Merci à tout le personnel administratif du Département pour leur disponibilité et leur soutien : Martine Boisselle, Lorraine Brisson, Hélène Diatta, Francine Germain, Julie Hudon, Jacinthe Lalonde et Josée Parenteau. J'ai été chanceux de faire partie d'une belle cohorte à l'UQAM. Je remercie mes collègues et amis étudiants : Thomas, Yorou, Hervé, Charles-Olivier, Boubacar, Sam Aguey, Juan, Afef, Oualid, Koffi, William, Marie-Mélanie, Mouna, Tran, Aurelie, Jean-Blaise, Edem, Jonathan, Constant, Jean-Paul, Adil, Adamou, Hermann, Franckel, Hamidou, Cathérine, François et Katherine.

J'aimerais également souligner le soutien financier de l'École Supérieure de Gestion et du département de sciences économiques de l'UQAM, du Centre inter-universitaire sur le risque, les politiques économiques et l'emploi (CIRPEE), de la chaire de recherche sur les répercussions régionales de la mondialisation et de la fondation de l'UQAM.

Ma famille et mes amis ont joué un rôle important dans la réalisation de ce projet. Je ne saurais jamais assez remercier mes parents Augustin et Odette Bougna, pour l'éducation et le soutien sans cesse. J'ai toujours pu compter sur les conseils, les prières et le soutien de mes frères et soeurs et de leurs familles respectives : Duplex, Virgile, Sylvanie, Virginie, Emmanuel, Ruth-Aurelie, Florette, Kepler, Gaüs et Franck. Un grand merci à ma belle famille pour les encouragements et le soutien affectif : Jean-Claude et Thérèse Taningpa, les familles Kwadjambou, Douo, Nouchi et Makem. Aux familles Mbiakoup, Keutiben, Chupenga, Agbanou, Fouopi, Tatsa, Kamdem et Peumi je dis merci pour toutes les prières et l'affection. Je remercie les familles et amis suivants : Tatoutchoup, Takam, Wandji, Dkeudo, Nguimkeu, Nzokem, Chekouo, Lonkeng, Bahel, Tchouaket, Fongang, Djithe, Daga, Njike, Djoumbissi, Tsimpo, Tiomo, Bobda, Nkabkob, Samen, Mbialeu, Mokoué, Fankou, Feunou, Nganou, Zeufack, Blimpo, Yutaka, Ouambea, Enock, Biahe, Évrard, Ornella, Some, Fotzeu, Djiotsa, Tonlio, Nepetsoun, Dongmo, Momeya, Sango, Wekape, Konhawa, Belinga, Tchuente, Kaffo, Ndzouetchoum. Je remercie toute ma famille chrétienne de Siloé.

Finalement, mes derniers remerciements vont à l'endroit de mon épouse Dany Joëlle et de mes enfants : Kim, Seraya et Jayden. Dany a été d'une patience et d'un soutien indéfectibles durant ces années de thèse. Elle a su partager mes moments de doute, supporter et combler mes nombreuses absences à la maison afin que je puisse mener à bien mes travaux de recherche. À mes enfants, vous êtes mes rayons de soleil, merci pour l'amour, la fierté et la joie que vous m'apportez au quotidien. Je vous aime !

DEDICACE

"To God be the glory forever and ever ! Amen." Galatians 1 : 5

TABLE DES MATIÈRES

LISTE DES FIGURES

Figure Page

LISTE DES TABLEAUX

Tableau Page

RÉSUMÉ

Le trait le plus frappant de la géographie de l'activité économique est le fort degré de concentration spatiale, et ce dans la majorité des pays et à divers échelles géographiques. Michael Porter (1998, p.197) a introduit la notion de cluster qu'il définit comme étant « une concentration géographique d'entreprises liées entre elles, de fournisseurs spécialisés, de prestataires de services, de firmes d'industries connexes et d'institutions associées (universités, agences de normalisation ou organisations professionnelles, par exemple) dans un domaine particulier, qui s'affrontent et coopèrent ». Au cours des vingt dernières années, cette notion de cluster a connu un regain d'intérêt auprès des décideurs politiques, des agences de développement et des universitaires. Plusieurs pays et agences de développement ont construit leur stratégie de développement industriel sur les modèles de pôles de compétitivité. Malgré quelques succès de leur implantation au Brésil, aux États-Unis, au Japon, en France, en Finlande et en Italie, plusieurs études s'interrogent sur l'efficacité coût-bénéfices de telles politiques. En effet, bien que contribuant à l'augmentation de la productivité, des salaires et de l'emploi, la concentration spatiale de l'activité économique entraine des coûts qui sont très souvent ignorés ex ante : la congestion, la rareté de l'espace, la criminalité, la pollution, etc. Il y'a donc très peu d'évidences empiriques sur l'impact des politiques de promotion des clusters au niveau macroéconomique (voir Duranton, 2011 ; Duranton, Martin, Mayer, and Mayneris, 2012).

Afin de mieux comprendre les causes et les implications des politiques de promotion des clusters, il est important de bien mesurer l'ampleur et le degré de concentration observé. Bien que cette question ait fait l'objet de nombreuses recherches, à ce jour, il n'existe pas d'indicateur statistique idéal de mesure de la concentration spatiale de l'activité économique. Cette thèse propose trois chapitres qui utilisent des données micro-géographiques et les méthodes d'estimation paramétrique et non-paramétrique afin (i) de fournir un portrait complet de l'état, l'ampleur et la dynamique de la concentration spatiale de l'activité manufacturière au Canada ; (ii) de fournir des évidences empiriques qui permettent de mieux comprendre les déterminants de la concentration spatiale des industries et comment ces déterminants ont influencé les changements observés dans la concentration spatiale ; et (iii) de proposer une nouvelle approche non-paramétrique de mesure de la concentration spatiale de plusieurs industries technologiquement liées.

Le premier chapitre intitulé "*An anatomy of the geographical concentration of Canadian manufacturing industries*" est un travail empirique dans lequel nous analysons à l'aide de mesures de concentration spatiales récentes, les tendances de la concentration spatiale de l'activité économique et les changements observés au cours de la première décennie de 2000. Nos résultats montrent qu'en fonction des années et du niveau d'agrégation des secteurs, 40 à 60% des industries sont concentrées dans l'espace géographique. Au cours de cette même période, on a observé une tendance à la dé-concentration des activités manufacturières au Canada. La contribution majeure du chapitre est qu'il permet de suivre l'évolution dans le temps de la concentration spatiale et fait ressortir les schémas de localisation des exportateurs, des petits et des jeunes établissements qui sont considérés comme vitaux pour la création d'emploi, et le développement local et régional. Ce chapitre permet également de faire une comparaison directe entre les mesures discrètes et les mesures continues de concentration spatiale. Il contribue ainsi à l'évaluation de l'ampleur du Problème des Unités Spatiales Modifiables (MAUP). De plus, à notre connaissance, c'est la première fois que les mesures continues de concentration spatiale sont appliquées aux données canadiennes en particulier et Nord-américaines en général (voir Holmes et Stevens, 2004). Cependant, il reste silencieux sur les causes et les déterminants de cette tendance à la dé-concentration des industries manufacturières au Canada.

Dans le second chapitre intitulé *"The world is not yet flat : Transport costs matter !*" nous nous intéressons aux déterminants de la concentration spatiale. En utilisant un long panel (1992-2008), nous régressons la mesure de concentration spatiale de Duranton et Overman (2005), sur des mesures micro-géographiques et spatiales des coûts de transport, de l'exposition au commerce international, et des liens en amont et en aval. Nos résultats montrent que l'augmentation des coûts de transport, la concurrence accrue du fait des importations en provenance des pays à faibles coûts et l'accroissement de la distance vers les clients et les fournisseurs sont tous fortement associés à une baisse de la concentration spatiale des industries manufacturières au Canada. Ces effets sont importants. En effet, sur la période 1992 – 2008, les changements observés dans les coûts de transport, les importations en provenance des pays à faibles coûts, et l'accès aux intrants intermédiaires expliquent entre 20 et 60% de la baisse observée dans la concentration spatiale des industries manufacturières au Canada. La principale contribution du chapitre est qu'il propose des évidences empiriques sur les déterminants de la concentration spatiale de l'activité économique en utilisant des mesures micro-géographiques et spatiales construites à des échelles industrielle et spatiale très fines. Ce chapitre révèle également un fait important. En effet, malgré la baisse historique observée dans les coûts de transport, ils continuent d'être un facteur déterminant de la structure industrielle et de la répartition spatiale des industries.

Le dernier chapitre, intitulé "*The determinants of localization : a conditional distance-based approach*" s'appuie sur les deux premiers chapitres. Nous apportons un raffinement à l'approche de Duranton et Overman (2005). L'idée étant de proposer une approche qui permet d'atténuer le problème du découpage sectoriel et de se rapprocher ainsi de l'indice de concentration spatial idéal. Nous combinons d'une part l'approche de mesure de la concentration spatiale (à la Duranton et Overman, 2005) et l'approche de co-localisation (à la Ellison, Glaeser et Kerr, 2010), et d'autre part, nous associons ces mesures au degré avec lequel les industries échangent les biens, les travailleurs et les idées. L'objectif est de combiner des mesures de distances technologiques (non-géographiques) à des mesures de distances géographiques entre secteurs. Plus précisément, nous proposons une nouvelle approche non-paramétrique de mesure de la localisation de plusieurs industries similaires. Conditionnellement à la similarité des établissements dans un espace non-géographique (liens en amont et en aval, type de travailleurs ou technologie utilisée), notre approche permet de vérifier si ces établissements sont concentrés ou non dans l'espace géographique. Puisque l'espace non-géographique est construit à base des mécanismes de la 'Trinité Marshallienne', le test permet également de jauger leur importance.

Nos résultats permettent de relever l'importance des liens en amont et en aval, et de l'accès à un bassin d'employés spécialisés dans les décisions de localisation des industries manufacturières. Nos résultats ne soutiennent pas l'importance de la technologie dans les décisions de co-localisation des industries. La contribution majeure de ce chapitre à la littérature sur la mesure de la concentration spatiale est qu'il propose un cadre unique qui permet de mesurer la co-agglomération des industries et de jauger de manière non-paramétrique l'importance des facteurs Marshalliens. Cette approche permet également d'atténuer le problème de la sensibilité des mesures existantes à un changement de nomenclature industrielle. Cependant, elle demeure sensible au découpage sectoriel en ce sens que la similarité des industries est mesurée à partir des données sectorielles agrégées. Un moyen de s'affranchir complètement du découpage sectoriel serait de mesurer la similarité à partir des données établissements et se rapprocher ainsi de l'indice idéal de concentration spatiale.

Mots-clés : Concentration géographique ; Micro-données géographiques ; Canada ; Industries manufacturières ; Estimation paramétrique ; Estimation non-paramétrique ; Densité de Kernel Conditionnelle ; Agglomération ; Coûts de transport ; Exposition au commerce international ; Liens en amont et en aval ; Trinité Marshallienne.

ABSTRACT

The most striking feature of industrial location patterns is geographical concentration. This has been of interest to economists since Marshall (1890). Clusters can be defined as a group of firms, related economics actors, and institutions that are located near each other and have reached a sufficient scale to develop specialized expertise, services, resources, suppliers, and skills. Over the last two decades, clusters have attracted interest from policy makers, academics, economic development practitioners, and development agencies. Many countries and economic development initiatives have built their industrial development strategies on cluster-based models. Despite successful implementation in the US, Brazil, Japan, France, Italy, and Finland, recent economic studies increasingly question the use of cluster policies: there is indeed little evidence that more clustering will have significant effects on average productivity or wages in manufacturing industries (e.g., Duranton, 2011; Duranton, Martin, Mayer, and Mayneris, 2012).

The starting point to better understand the drivers and implications of cluster-based development is to measure correctly the observed degree of clustering. Many studies have empirically defined and measured industrial localization, however, the ideal index of spatial concentration still seems out of reach. My thesis addresses these challenges through the use of micro-geographic data, parametric and non-parametric techniques to measure and to explain changes in the spatial distribution of economic activity.

Specifically, this thesis proposes three chapters that use micro-geographic data, parametric and non-parametric estimation methods in order to (i) provide a comprehensive anatomy of the geographical concentration of manufacturing in Canada and its dynamics, (ii) provide empirical evidence that allows to better understand the determinants of the geographical concentration of industries and how these determinants have influenced changes in that concentration, and (iii) propose a new non-parametric approach to measuring the localization of 'closely related' multiple industries – i.e., a multidimensional way to assess coagglomeration – in continuous space.

The research proceeds along three chapters. In the first chapter entitled "*An anatomy of the geographical concentration of Canadian manufacturing industries*" we use detailed micro-geographic data to dissect the location patterns of Canadian manufacturing industries and changes in those patterns during the first decade of 2000. Our results show that, depending on industry classifications and years, 40 to 60 percent of industries are geographically localized i.e., are spatially clustered relative to overall manufacturing. This chapter's main contribution is that it allows to follow the time evolution of the pattern of industries localization in Canada and provides a detail location trend of exporters, small and young plants. These plants are perceived as being vital for employment growth and local regional development, thus making them prime targets for cluster policy. I also allows for direct comparison between the results based on discrete versus continuous measures of localization. Finally, to the best of our knowledge, continuous localization measures have until now neither been applied to Canadian data in particular, nor to North American data in general (see Holmes and

Stevens, 2004).

In chapter two, entitled *"The world is not yet flat: Transport costs matter!*" , we provide evidence for the effects of changes in transport costs, international trade exposure, and input-output linkages on the geographical concentration of Canadian manufacturing industries. We document that increasing transport costs, stronger import competition, and the spreading out of upstream suppliers and downstream customers are all strongly associated with declining geographical concentration of industries. The effects are large: changes in trucking rates, in import exposure, and in access to intermediate inputs explain between 20% and 60% of the observed decline in spatial concentration over the 1992 – 2008 period. This chapter makes two contributions. First, we construct new and finer measures of the costs of trading goods across space than in the previous literature. Second, we are – to the best of our knowledge – among the first to exploit the time-series variation in the data to shed light on what drives *changes* in the spatial concentration of industries

The last chapter, entitled *"The determinants of localization: a conditional distance-based approach*" draws upon the first two chapters. The key idea is to first combine the measurement approach of localization in continuous space with a coagglomeration approach, and then relate them to the degree to which industries share goods, people, and ideas. More precisely, I propose a new non-parametric approach to measuring the localization of 'closely related' multiple industries – i.e., a multidimensional way to assess coagglomeration – in continuous space. Conditional on belonging to industries with similar characteristics (in terms of input-output linkages, types of workers employed, or technology), I check whether plants are lo-

cated near one another in space. Since the non-geographic space is built upon Marshallian proxies, my test allows me to gauge non-parametrically their importance. It allows to answer the following questions: Do pairs of plants with 'close or similar' input-output linkages, types of workers employed, and that use or exchange similar technology locate near one another in space? My results show that plants which belong to industries with similar input-output linkages and which employ similar types of workers tend to co-locate near one another. I find little evidence that plants that share similar technologies, as measured using patent citations, cluster geographically. This chapter makes three contributions into the literature on measuring localization. First, we propose an approach that allows to assess coagglomeration of industries and the importance of Marshallian forces non-parametrically and in a unified framework. Second, we propose an approach that accounts for both spatial and technological distances between industries. Third, our approach allows to alleviate the problem of change in industrial classification. In order to get truly away from industrial classifications, we need to use detailed plant-level data to build finer non-geographic distance measures, and therefore move towards an ideal index of localization.

Keywords: Geographical concentration; Micro-geographic data; Canada; Manufacturing industries; Parametric Estimation; Non-parametric Estimation; Conditional Kernel density; Agglomeration; Transport costs; International trade exposure; Input-output linkages; Marshall Trinity.

INTRODUCTION

La description des phénomènes d'agglomération d'entreprises a été popularisée dans les années 1990 par Michael Porter, professeur à la Harvard Business School. Porter (1998, p.197) a introduit la notion de cluster qu'il définit comme étant *"une concentration géographique d'entreprises liées entre elles, de fournisseurs spécialisés, de prestataires de services, de firmes d'industries connexes et d'institutions associées (universités, agences de normalisation ou organisations professionnelles, par exemple) dans un domaine particulier, qui s'affrontent et coopèrent*". Dans son livre "Geography and Trade" (1991, p.5), Krugman relève qu'avec un peu de recul la concentration spatiale est probablement le trait le plus frappant de la géographie de l'activité économique. Fujita et Thisse (2002) soulignent également que l'un des faits marquants du paysage économique est la concentration des activités humaines sur une faible portion du territoire, et en particulier les villes.

Au cours des dernières années, la concentration spatiale de l'activité économique a eu un certain regain d'intérêt auprès des chercheurs en économie et en géographie. C'est un phénomène observé à plusieurs échelles (mondiale, régionale et locale) et dans la grande majorité des pays. Au Canada par exemple, les données de Statistique Canada révèlent que les provinces de l'Ontario et du Québec qui représentent moins de 25% du territoire concentrent environ 62% de la population, 69% des sites d'installations et 72% de la main d'oeuvre du secteur manufacturier. Pour ce qui est de la concentration à l'intérieur des provinces, les données de l'Institut de la

Statistique du Québec révèlent que la région de Montréal concentre près de 35% du PIB du Québec avec une superficie de moins de 0.04% de la superficie totale du Québec. De plus, la région de Toronto (0.06% du territoire de l'Ontario) concentre environ 45% du PIB de l'Ontario. Il apparaît ainsi que la majorité des activités économiques (analysées sous l'angle de l'emploi ou du PIB) est polarisée au niveau provincial et à l'intérieur des provinces canadiennes. En France, l'Île-de-France qui représente seulement 2,2% de la superficie du territoire, concentre sur 12,2% de sa surface 18,9% de la population française et contribue à hauteur de 30% du PIB de la nation (Fujita et Thisse, 2002). Même à des échelles spatiales plus fines, on observe également une forte concentration de certaines industries. Les exemples les plus frappants concernent: (i) les activités de haute technologie dans la Silicon Valley en Californie, la Massachusetts Route 128 près de Boston, et dans le North Carolina Research Triangle; (ii) les activités du secteur automobile dans le corridor Détroit-Windsor.

Pourquoi est-il important de mesurer la concentration spatiale des activités économique?

La question de la mesure des inégalités – de revenu ou le degré de concentration spatiale – intéresse les économistes et les géographes depuis très longtemps. Évaluer ou mesurer la concentration industrielle est important car ceci permet d'une part d'appréhender le degré d'inégalité régionale ou sectorielle et d'autre part, d'analyser les différences entre unités spatiales, secteurs, ou à travers le temps. En effet, autant il est important pour les économistes de mesurer la croissance économique (par exemple à travers la mesure de la variation du PIB), autant il est important qu'ils soient capables de mesurer le niveau et l'ampleur de la concentration spatiale des activités économiques afin d'informer l'opinion et les décideurs publics notamment dans un contexte où le débat sur la subvention des clusters comme outil de développement régional est d'actualité. Cette question est également importante dans la mesure où il est nécessaire d'apporter des évidences sur les mécanismes et les déterminants de cette concentration spatiale.

Comment mesurer la concentration des activités économiques?

Au cours des vingt dernières années, la notion de cluster a connu un regain d'intérêt auprès des décideurs politiques, des agences de développement et des universitaires. Plusieurs pays et agences de développement ont construit leur stratégie de développement industriel sur les pôles de compétitivité. Malgré quelques succès au Brésil, aux États-unis, au Japon, en France, en Finlande et en Italie, plusieurs études s'interrogent sur l'efficacité coûts-bénéfices de telles politiques. En effet, bien que contribuant à l'augmentation de la productivité, des salaires et de l'emploi, la concentration spatiale de l'activité économique entraine des coûts qui sont très souvent ignorés ex-ante: la congestion, la rareté de l'espace, la criminalité, la pollution, etc. Il y a donc très peu d'évidences empiriques sur l'impact des politiques de promotion des clusters au niveau macroéconomique (voir Duranton, 2011; Duranton, Martin, Mayer, and Mayneris, 2012). Ainsi, afin de mieux comprendre les causes et les implications des politiques de promotion des clusters, il est important de bien mesurer l'ampleur et le degré de concentration observé.

Bien que cette question ait fait l'objet de nombreuses recherches, à ce jour, il n'existe pas d'indicateur statistique idéal de mesure de la concentration spatiale. Mesurer la concentration spatiale nécessite l'élaboration de mesures qui permettent de comprendre quels sont les secteurs les plus concentrés, ceux qui ne le sont pas et d'utiliser l'inférence statistique afin de tester les évidences empiriques trouvées. Idéalement, la construction d'un indice de concentration spatiale requiert des informations précises sur la localisation exacte de chaque firme. Cependant, ces informations sont très coûteuses et ne sont pas facilement accessibles. À la suite des travaux de Duranton et Overman (2005), Combes et al. (2008) ont définit six propriétés que devrait avoir un indicateur idéal de concentration géographique: (i) *la mesure de concentration spatiale doit être comparable entre secteurs*; (ii) *la mesure de concentration spatiale doit être comparable entre zones géographiques*; (iii) *la mesure de concentration spatiale doit être insensible à un changement de définition des unités spatiales*; (iv) *la mesure de concentration spatiale doit être insensible à un changement de définition des secteurs*; (v) *la mesure de concentration spatiale doit être effectuée par rapport à une référence clairement établie*; (vi) *la mesure doit permettre de déterminer si des différences significatives par rapport à la référence ou entre deux situations (zones, périodes ou secteurs) existent*.

Une manière de mesurer la concentration spatiale d'un secteur consiste à comparer la distribution spatiale de son emploi à la distribution spatiale de l'emploi total (indice de Gini par exemple). Cependant, l'emploi ou l'activité économique est réparti entre un nombre limité d'établissements, il convient alors de corriger pour la concentration industrielle. Ellison et Glaeser (1997), Maurel et Sédillot (1999) proposent des indices de concentration qui tiennent compte du nombre d'établissements et de la distribution de l'emploi entre établissements. Ces mesures satisfont seulement trois des six critères d'un indice idéal et reposent sur un découpage géographique prédéfinit du territoire (provinces, régions économiques, divisions de recensement, etc.). Ce découpage ne dépend pas des caractéristiques économiques, ce qui rend ces mesures sensibles à un changement dans le découpage géographique. L'indice devient alors sensible à la position relative des unités spatiales de ce découpage et incapable de capter les niveaux de concentration qui s'étendent aux unités adjacentes. Ce problème est plus connu sous le nom de "Problème des Unités Spatiales Modifiables" (MAUP). Duranton et Overman (2005) et Marcon et Puech (2003) ont résolu ce problème en proposant de tester la concentration spatiale de l'activité économique en utilisant une approche continue. Cette approche est basée sur la distribution des distances bilatérales entre paires d'établissements au sein d'un secteur d'activité. Ils testent si la densité observée d'un secteur est proche ou non d'une densité provenant d'une hypothèse où les établissements du secteur seraient distribués aléatoirement. Cette mesure a le mérite de satisfaire cinq des six propriétés d'un indice idéal de concentration spatiale énumérées par Combes et al. (2008). La seule exception étant la sensibilité de la mesure à un changement de nomenclature industrielle.

Cette thèse comprend trois chapitres qui auront pour objectif: (i) de construire les mesures de concentration spatiale afin de fournir un portrait complet de l'état, l'ampleur et la dynamique de la concentration spatiale de l'activité économique au Canada; (ii) de s'interroger sur les facteurs explicatifs ou les déterminants (question fondamentale pour la mise en oeuvre des politiques publiques) et sur comment ces déterminants ont influencé les changements observés dans la concentration spatiale; et (iii) de construire un test qui permet de se rapprocher de la mesure idéale de concentration spatiale en mesurant la co-agglomération de plusieurs industries.

Le chapitre 1 est un travail empirique dans lequel nous analysons à l'aide de mesures de concentration spatiales récentes (discrète et continue), les tendances de la concentration spatiale des activités manufacturières entre 2001 et 2009 au Canada. De manière spécifique, nous utilisons d'une part l'indice discret d'Ellison et Glaeser (1997) et sa version pondérée d'une correction spatiale – qui permet de tenir compte de la position des régions dans l'espace – et d'autre part, la mesure continue de Duranton et Overman (2005) qui permet de s'affranchir du "MAUP". Nos résultats montrent qu'en fonction des années et du niveau d'agrégation des secteurs, 40 à 60% des industries manufacturières sont concentrées dans l'espace géographique. De plus, la plupart de ces industries sont concentrées soit à de faibles distances (moins de 150km) soit à des distances intermédiaires (400 – 600km). Au cours de cette même période, on a observé une tendance à la dé-concentration des activités manufacturières au Canada. Ce chapitre a le mérite de faire ressortir le portrait de la géographie des activités manufacturières au Canada. La contribution majeure du chapitre est qu'il permet de suivre l'évolution dans le temps de la concentration spatiale en plus de faire ressortir les schémas de localisation des exportateurs, des petits et des jeunes établissements qui sont considérés comme vitaux pour la création d'emploi, et le développement local et régional. Ce chapitre permet également de faire une comparaison directe entre les mesures discrètes et les mesures continues de concentration spatiale. Il contribue ainsi à une évaluation indirecte de l'ampleur du MAUP. Finalement, à notre connaissance, c'est la première fois que les mesures continues de concentration spatiale sont appliquées aux données canadiennes en particulier et Nord-américaine en général (voir Holmes et Stevens, 2004). Cependant, il reste silencieux sur les causes et les déterminants de cette tendance à la dé-concentration des industries manufacturières au Canada. De plus, le degré de concentration spatiale mesuré demeure sensible à un changement de la nomenclature industrielle utilisée. Ces deux problèmes seront abordés dans les chapitres 2 et 3.

Dans le chapitre 2, nous nous intéressons aux déterminants de la concentration spatiale des activités manufacturières, avec une emphase sur le rôle des coûts de transport et du commerce international dans les changements observés sur la concentration des activités manufacturières au Canada. En utilisant un long panel (1992-2008), nous régressons la mesure de concentration spatiale de Duranton et Overman (2005), sur des mesures micro-géographiques et spatiales des coûts de transport, de l'exposition au commerce international, et des liens en amont et en aval. Nos résultats montrent que l'augmentation des coûts de transport, la concurrence accrue du fait des importations en provenance des pays à faibles coûts et l'accroissement de la distance vers les clients et les fournisseurs sont tous fortement associés à une baisse de la concentration spatiale des industries manufacturières au Canada. Ces effets sont importants. En effet, sur la période 1992 – 2008, les changements observés dans les coûts de transport, les importations en provenance des pays à faibles coûts, et l'accès aux intrants intermédiaires expliquent entre 20 et 60% de la baisse observée dans la concentration spatiale des industries manufacturières au Canada. La principale contribution de chapitre est qu'il propose des évidences empiriques sur les déterminants de la concentration spatiale de l'activité économique en utilisant des mesures micro-géographiques et spatiales construites à des échelles industrielle et spatiale très fines. Ce chapitre révèle également un fait très important. En effet, malgré la baisse historique observée dans les coûts de transport, ils continuent d'être un facteur déterminant de la structure industrielle et de la répartition spatiale des industries.

Le dernier chapitre de cette thèse apporte un raffinement à l'approche continue de mesure de la concentration spatiale. L'idée étant de proposer une approche qui permet d'atténuer le problème du découpage sectoriel et de se rapprocher ainsi de l'indice de concentration spatial idéal. De manière spécifique, nous combinons d'une part, l'approche de mesure de la concentration spatiale (à la Duranton et Overman, 2005) et l'approche de co-localisation (à la Ellison, Glaeser et Kerr, 2010), et d'autre part, nous associons ces mesures au degré avec lequel les industries échangent les biens, les travailleurs et les idées. L'objectif étant de combiner des mesures de distance technologiques (non-géographiques) à des mesures de distances géographiques entre secteurs. Plus précisément, nous proposons une nouvelle approche non-paramétrique de mesure de la localisation de plusieurs industries similaires. Conditionnellement à la similarité des établissements dans un espace non-géographique (liens en amont et en aval, type de travailleurs ou technologie utilisée), notre approche permet de vérifier si ces établissements sont concentrés ou non dans l'espace géographique. Puisque l'espace non-géographique est construit à base des mécanismes de la 'Trinité Marshallienne', notre test permet également de jauger leur importance. Une application de ce test au secteur manufacturier canadien, permet de relever l'importance des liens en amont et en aval, et de l'accès à un bassin d'employés spécialisés dans les décisions de localisation des industries manufacturières. Nos résultats ne soutiennent pas l'importance de la technologie dans les décisions de co-localisation des industries. La contribution majeure de ce chapitre à la littérature sur la mesure de la concentration spatiale est qu'il propose un cadre unique qui permet de mesurer la co-agglomération des industries et de jauger de manière non-paramétrique l'importance des facteurs Marshalliens. Cette approche permet également d'atténuer le problème de la sensibilité des mesures existantes à un changement de nomenclature industrielle. Cependant, elle demeure sensible au découpage sectoriel en ce sens que la similarité des industries est mesurée à partir des données sectorielles agrégées. Un moyen de s'affranchir complètement du découpage sectoriel et se rapprocher ainsi de l'indice idéal de concentration spatial serait d'utiliser des données au niveau établissements.

CHAPITRE I

AN ANATOMY OF THE GEOGRAPHICAL CONCENTRATION OF CANADIAN MANUFACTURING INDUSTRIES

Abstract

We use detailed micro-geographic data to document the location patterns of Canadian manufacturing industries and changes in those patterns during the first decade of 2000. Depending on industry classifications and years, 40 to 60 percent of industries are geographically localized, i.e., are spatially clustered relative to overall manufacturing. Although some industries are increasingly clustered, localization has generally decreased in Canada according to our measures. We further document the locational trends of small plants, young plants, and exporters. Their location patterns do not differ significantly from that of the other plants in their industries.

Keywords : Location patterns ; manufacturing industries ; micro-geographic data ; Canada.

JEL classification : R12 ; L60.

1.1 Introduction

One of the most salient features of the economic landscape is the strong geographical concentration of economic activity. That concentration is observed in most countries and at various spatial scales. Famous examples of 'clusters' include the high-technology concentrations of Silicon Valley, Boston's Route 128, the North Carolina research triangle, as well as concentrations of more mature industries like the automotive cluster in the Detroit-Winsor corridor or the Italian manufacturing 'districts'. In Canada, economic activity – measured by either GDP or employment – is strongly concentrated *across and within* provinces. Ontario and Quebec, for example, host about 60 percent of Canadian GDP and 75 percent of manufacturing employment. Within those two provinces, the Toronto metropolitan area, about 0.06 percent of Ontario's surface, generates 45 percent of Ontario's GDP ; whereas the Montreal metropolitan area generates almost 35 percent of Quebec's GDP on about 0.04 percent of Quebec's surface. [1]

The resurgence of spatial analysis in economics has led to a renewed interest in empirically analyzing and theoretically explaining the strong geographical concentration of industries. Clusters and regional development have also often been – and are becoming increasingly more – a matter of concern for policy makers around the world. Quebec's government, for example, has recently launched the 'Plan Nord', with the aim to invest around $80 billion over the next 25 years to create 20,000 jobs, generate $14 billion in government revenue, and $162 billion for Quebec's GDP.

[1] These figures for 2013 are from Statistics Canada and the Institut de Statistiques du Québec.

Such huge investment plans – which have a clear regional development component – are unlikely to leave the geography of economic activity unchanged. It is, therefore, important to understand which industries tend to cluster, what location patterns we observe for specific types of plants that are important targets for economic development (e.g., young plants, small plants, and exporters), and what the broad trends of geographical concentration have been over the last decade. This is the focus of the present paper.

There is a substantial literature dealing with the measurement of *industrial localization*, i.e., the geographical concentration of industries in excess of the concentration of economic activity in general. Ellison and Glaeser (1997; henceforth EG) have developed an index that has been widely applied to that issue. Despite its numerous advantages and appealing theoretical properties, that index has no strong spatial flavor as it does not take into account the relative positions of the geographical units. We address that issue using two alternative strategies. First, we exploit the micro-geographic nature of our data to compute point pattern based continuous measures following Ripley (1976, 1977), Duranton and Overman (2005, 2008; henceforth DO), and Marcon and Puech (2003, 2010). Using continuous measures allows us to sidestep the need for pre-defined administrative units, which give rise to the well-known *modifiable areal unit problem* (henceforth MAUP; Openshaw and Taylor, 1979; Openshaw, 1983). Second, we analyze the geographical concentration in Canada by explicitly integrating 'neighborhood effects' into the EG index, following recent work by Guimarães, Figueiredo, and Woodward (2011).

To the best of our knowledge, continuous localization measures have until now neither been applied to Canadian data in particular, nor to North

American data in general (see Holmes and Stevens, 2004).[2] The empirical literature on localization using micro-geographic data, though growing, is still relatively limited. Using the EG and DO indices, we identify the most and the least localized manufacturing industries in Canada. Consistent with previous findings for the UK, France, and Japan, industries related to textiles and to the extraction of natural resources rank among the most localized industries. We also provide a broad picture of the main trends for the first decade of 2000. Our key findings can be summarized as follows. First, depending on industry definitions and years, 40 to 60 percent of manufacturing industries are clustered, mainly at distances of less than 150 kilometers, and at distances of about 500 kilometers. These figures suggest that there is less industrial localization in Canada as compared to other developed countries like France or the UK. Second, since, our dataset spans a ten year period, we can look at the 'dynamics' of localization. We are not aware of any other study looking at the changes in localization over time using large micro-geographic plant-level datasets. We find that localization is decreasing, i.e., manufacturing industries have become less geographically concentrated in Canada. Yet, there is a lot of heterogeneity across industries, and some of the most strongly localized industries are becoming even more localized. The changes in spatial concentration through time are negatively correlated with changes in industrial concentration.

Two advantages of our dataset is that it contains a large number of

[2] Ellison, Glaeser, and Kerr (2010) use a 'lumpy approximation' of the DO index for the US. Riedel and Hyun-Ju (2014) do the same for Germany. It is unclear whether using a discrete approximation of a continuous measure helps in solving the fundamental spatial aggregation problems.

small and young plants, and that it reports plant-level information on export status. This allows us to document in detail the location trends for those subgroups, and in particular to look at trends specific to exporter plants involved in international business. Understanding those trends is relevant from a policy perspective, since these groups of plants are perceived as being vital for employment growth and local regional development, thus making them prime targets for cluster policy. Our findings suggest that they are, in general, not more strongly concentrated than all plants in their respective industries. The only exception is for exporters, but their 'excess concentration' tends to significantly decrease over the first decade of 2000.

The remainder of the paper is organized as follows. Section 1.2 provides a snapshot of manufacturing in Canada. Section 1.3 presents our empirical results using continuous measures of localization. Section 1.4 summarizes our empirical results using discrete measures as a robustness check, controlling for the relative position of the spatial units. Finally, Section 1.5 concludes and places our results into the policy debate about industry clusters and regional development. We relegate all technicalities, the description of our datasets, and additional results to an extensive set of appendices.

1.2 A snapshot of Canadian manufacturing, 2001–2009

To set the stage, we first provide a quick overview of the sectoral and geographical structure of manufacturing in Canada from 2001 to 2009. Total salaried employment in Canada in 2001 was 12,978,258 jobs, of which 1,974,636 – or 15.21 percent – were in manufacturing. In 2005, the corresponding numbers were 13,931,343 and 1,837,828 jobs – or 13.19 percent – respectively ; whereas they were 14,570,025 and 1,473,472 jobs – or 10.11 percent – in 2009.[3] The downwards trend in manufacturing can also be seen from Table 1.1, which shows that the number of plants in our data has fallen from 54,379 in 2001 to 46,391 in 2009. This 'de-industrialization' is not specific to Canada and affects most developed countries in a similar way (see, e.g., Duranton, Martin, Mayer, Mayneris, 2012, for the French case). As can be seen from Table 1.1, the decrease in the number of plants went hand-in-hand with an increase in average plant size – as measured by employment – except for the Atlantic provinces (see Appendix A for details on the data).

Table 1.1 Descriptive statistics by province.

	2001		2005		2009	
Province	# of plants	Avg. empl.	# of plants	Avg. empl.	# of plants	Avg. empl.
Alberta	3,933	36.100	3,455	44.430	3,581	52.780
British Columbia	6,219	31.930	5,371	33.730	4,991	34.370
Manitoba	1,654	43.330	1,481	55.230	1,263	57.790
New Brunswick	1,395	35.660	1,258	40.080	1,175	36.940
Newfoundland and Labrador	576	43.830	540	44.830	472	42.500
Nova Scotia	1,676	29.930	1,495	37.140	1,296	35.020
Ontario	21,306	45.010	20,966	46.080	19,637	46.760
Prince Edward Island	328	25.350	327	24.410	280	25.430
Quebec	15,939	41.640	14,166	45.690	12,560	49.550
Saskatchewan	1,353	27.360	1,305	32.520	1,091	36.230
Territories	–	–	40	5.940	45	10.140
Total	**54,379**	**36.01**	**50,404**	**37.28**	**46,391**	**38.86**

Source : Authors' computations using Scott's National All Business Directories.

Table 1.2 summarizes industry-level details of our data, including the

[3] Source : Statistics Canada, CANSIM.

average plant size by industry and the number of exporting plants. There is clearly substantial cross-industry variation, as extensively documented by previous studies (e.g., Bernard and Jensen, 1995). Observe that, although the number of plants has decreased substantially, the share of exporting plants has increased from 42.3% in 2001 to 45.1% in 2009 in the wake of increasing globalization.

Turning to the spatial dimension, population is strongly concentrated geographically in Canada. Indeed, because of historical settlement patterns, the climatic conditions in the north, and access to the large US market to the south, about 90 percent of the Canadian population lives less than 100 miles from the US border. Quite naturally, the overall distribution of manufacturing is thus also strongly concentrated geographically in Canada – namely in Ontario and Quebec and, more generally, along the Canada-US border – as can be seen from Figure 1.7 in Appendix E. We show in Appendix D that the overall 'shape' of the distribution of bilateral distances between manufacturing plants in Canada has remained – in the aggregate – fairly stable between 2001 and 2009. This suggests that the localization measures we compute in what follows for individual industries are comparable between the years of our analysis.

Since manufacturing is strongly concentrated geographically in Canada, we will use its overall distribution as the benchmark against which we assess localization in a given sector. This avoids picking up localization patterns that are solely driven by the overall concentration of industries in large metropolitan areas (Combes, Mayer, and Thisse, 2008) or, in the case of Canada, in the traditional manufacturing corridor running from Quebec City to Windsor via Montreal and Toronto (see Figure 1.7 in Appendix

Table 1.2 Breakdown of plants by NAICS 3-digit industries.

NAICS3	Industry name	# NAICS6	# of plants			Avg. plant size (empl.)			# of exporters		
			2001	2005	2009	2001	2005	2009	2001	2005	2009
311	Food Manufacturing	33	4,807	4,327	3,929	50.114	56.711	62.158	1,667	1,591	1,404
312	Beverage & Tobacco Product Mfg	6	477	426	462	64.522	77.345	64.036	129	134	126
313	Textile Mills	7	539	356	277	51.986	53.858	53.359	246	198	162
314	Textile Product Mills	4	1,413	1,307	1,146	18.340	17.568	17.147	422	488	430
315	Apparel Manufacturing	17	2,364	1,905	1,354	40.631	38.855	36.349	932	819	642
316	Leather & Allied Product Mfg	3	382	308	238	36.728	28.454	29.091	203	163	131
321	Wood Product Manufacturing	14	3,919	3,546	3,127	42.826	48.239	48.557	1,733	1,690	1,436
322	Paper Manufacturing	12	911	854	775	119.594	114.557	115.001	582	588	546
323	Printing & Related Support activ.	6	5,091	4,577	4,089	18.600	22.935	23.964	1,063	1,174	1,041
324	Petroleum & Coal Products Mfg	4	347	318	301	100.009	135.365	130.882	123	115	106
325	Chemical Manufacturing	20	2,183	2,034	1,982	47.907	56.685	63.959	1,231	1,205	1,146
326	Plastics & Rubber Products	14	2,206	2,227	2,084	48.950	57.802	54.252	1,375	1,423	1,334
327	Nonmetallic Mineral Products	12	2,608	2,618	2,473	27.539	27.651	42.394	778	808	766
331	Primary Metal Manufacturing	13	927	820	805	113.145	115.373	106.953	587	534	484
332	Fabricated Metal product Mfg	21	8,018	7,521	7,255	26.504	30.020	31.093	3,014	3,085	2,975
333	Machinery Manufacturing	17	5,237	4,758	4,583	34.210	37.538	41.780	3,160	3,147	2,994
334	Computer & Electronic Products	9	2,130	1,654	1,643	61.658	59.794	63.845	1,433	1,205	1,201
335	Electrical Equip. & Appliances	12	1,193	1,047	1,007	43.489	50.602	47.018	777	749	707
336	Transportation Equipment Mfg	18	2,008	1,907	1,839	116.297	129.609	125.060	990	1,010	918
337	Furniture & Related Product Mfg	10	3,526	3,351	2,869	25.192	29.308	32.065	1,126	1,198	1,001
339	Miscellaneous Manufacturing	7	4,093	4,543	4,153	17.337	16.022	15.934	1,434	1,467	1,353
		259	**54,379**	**50,404**	**46,391**	**52.647**	**57.347**	**57.347**	**23,005**	**22,791**	**20,903**
									42.3%	**45.2%**	**45.1%**

Source : Authors' computations using Scott's National All Business Directories.

E). We will compute both discrete and continuous measures of localization – for industries in general, but also for certain types of plants like small plants, young plants, and exporters – and analyze their trends over time. When looking at specific types of plants, we will use an even more restrictive benchmark, namely the spatial distribution of all plants in *their own industry*. In other words, we will look at the 'excess concentration' of small plants, young plants, and exporters as compared to the concentration of plants in their industry in general. Doing so will provide a very fine picture of the 'state of geography' of manufacturing in Canada, both in terms of industries and in terms of specific plant types.

1.3 Continuous measures : Methodology and results

While discrete measures of localization, such as the EG index, are very popular and have been widely used, they are known to be sensitive to the choice of geographical units. They are also independent of the relative position of those units. To deal with those two problems, we exploit the micro-geographic nature of our data and compute continuous measures of localization, namely the Duranton-Overman index (Duranton and Overman, 2005, 2008). This index is based on the kernel density of the distribution of bilateral distances across all plants in an industry – or, in its weighted version, of all employees in an industry – and compares that distribution to a counterfactual one that is obtained under the assumption of 'spatial randomness'. Concerning the weighted version of the DO index that we use, we need to point out that, contrary to Duranton and Overman (2005) who use a multiplicative weighting scheme, we use an additive one. Methodological details and a discussion of the implications of the weighting scheme are provided in Appendix B.

The key advantage of the DO index is that it retains the desirable properties of the EG index – namely to control for the size distribution of plants in an industry – while getting rid of the need to choose specific spatial units for the analysis.[4] Another important advantage of the DO index is that its statistical significance can be tested. Two-sided confidence intervals that contain 90 percent of the estimated K-densities can be constructed by using bootstrap. The upper bound of this interval is given by the 95th percentile of the generated values, and the lower bound by the 5th percentile of these values. If we observe a higher K-density than that of randomly drawn distributions, we consider the industry as localized. Similarly, if we observe a lower K-density than that of randomly drawn distributions, we consider the industry as dispersed. We can also measure the strength of localization and dispersion by the 'area' between the observed distribution and the upper- and lower-bounds of the confidence bands. We denote these measures by Γ_i and Ψ_i for each industry i. They can intuitively be interpreted as the 'excess probability' to find another firm in the same industry closer than some distance d when controlling for the reference distribution and accepting a 5% risk level.[5]

[4] That methodology has been recently extended and can be applied to many economic problems where space matters and where micro-geographic data is available (see, e.g., Murata, Nakajima, Okamoto, and Tamura, 2014, for an application to the localization of patents).

[5] Consider a sector that is localized (see, e.g., the upper-left panel of Figure 1.1). The area below the actual curve up to a distance d, the CDF at d, is the probability for a firm that a randomly drawn 'neighbor' in the *same industry* is less than d apart. The same area under the upper bound of the envelope is the maximum probability for a firm

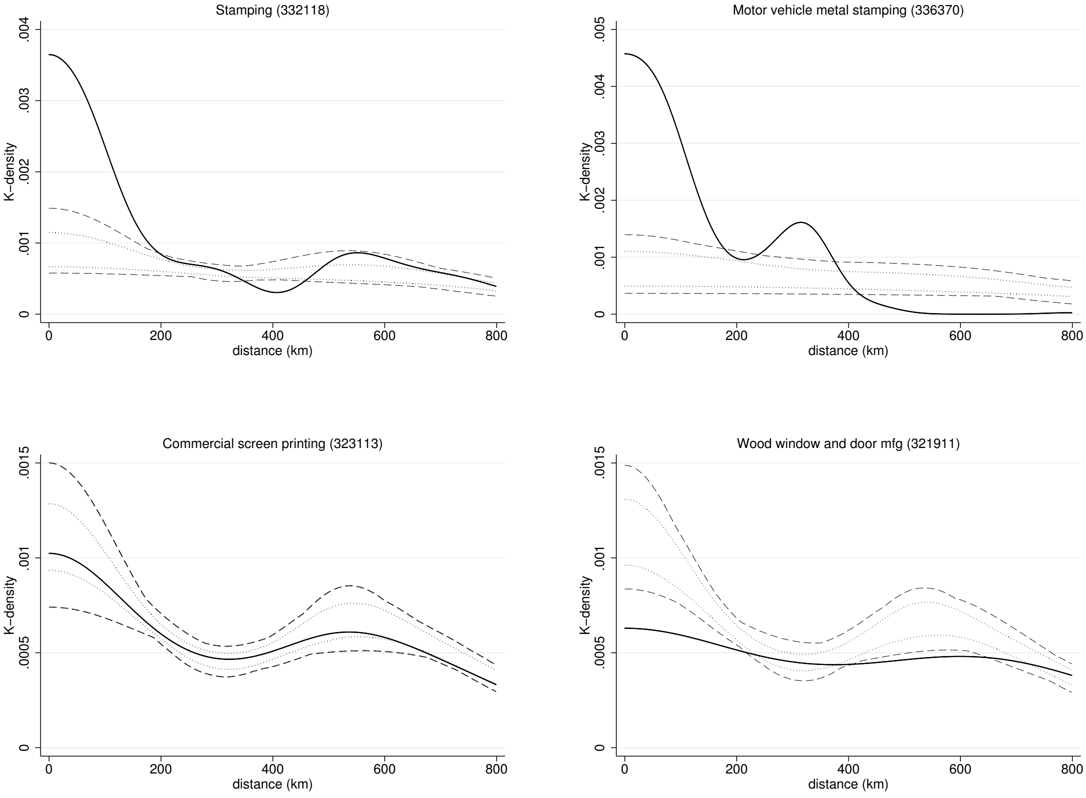


Figure 1.1 Selected location patterns of industries in 2009 (unweighted K-densities).

To understand the logic underlying the DO index, we illustrate the possible patterns with the help of Figure 1.1. The observed distribution of distances in the industry is depicted by the solid line, which we refer to as the K-density. The figure also depicts the 'local' (dashed) and the 'global' (dotted) confidence bands (see Appendix B for details). These bands contain 90 percent of the counterfactual distributions, so that when the solid line lies within them we cannot reject – at the 5 percent level – the null hypothesis that the observed location pattern of the industry is one of 'spatial randomness'. If the solid line lies above the upper bound of the confidence band, distances between plants are over-represented as compared to spatial randomness, which is interpreted as *localization*; whereas when the solid line lies below the lower bound of the confidence band, distances between plants are under-represented as compared to spatial randomness, which is interpreted as *dispersion*.

The four industries depicted in Figure 1.1 display four different geographical patterns. The top-left panel depicts an industry that is localized at a regional scale (up to 200 kilometers), however dispersed at longer distances (around 400 kilometers). This corresponds to the 'classical' location pattern where plants are disproportionately located at short distances, i.e., the industry is localized. The top-right panel depicts an industry that is both significantly concentrated at short distances, and also significantly agglomerated in between major urban areas – 400–500 kilometers corresponds approximately to the distance between the peripheries of the greater me-

that a randomly drawn 'neighbor' in *any industry* is less than d apart, accepting a 5% risk level. The difference between the two, which we call Γ, is therefore the 'excess probability' to find a neighbor in the same industry less than d apart, controlling for the reference distribution. We thank a referee for suggesting this interpretation.

tropolitan regions of Toronto and Montreal. The bottom-left panel depicts an industry that is neither significantly localized nor significantly dispersed. The location pattern of that industry is not significantly different from one that would be obtained by a purely random location process of the plants. Last, the bottom-right panel depicts an industry that is significantly dispersed, both at short and at long distances.

1.3.1 Baseline results

We first examine the number of industries that are localized or dispersed according to the DO index. As can be seen from Table 1.3, using a strict definition of manufacturing plants (see Appendix A), we find that roughly 31 percent and 55 percent of industries were significantly localized in 2001 at the 6-digit and the 4-digit levels, respectively. These numbers were quite stable between 2001 and 2005, but they fall below 25 percent at the 6-digit level and below 49 percent at the 4-digit level in 2009. On average, the share of localized manufacturing industries in Canada is smaller than the ones reported for the UK (52 percent), France (63 percent), Germany (71 percent), and Japan (50 percent) in earlier studies by Duranton and Overman (2005), Barlet, Briant, and Crusson (2013), Riedel and Hyun-Ju (2014), and Nakajima, Saito, and Uesugi (2012), respectively.

There is a clear tendency towards less localization between 2001 and 2009 : the number of localized industries decreases, as well as the strength of localization (as measured by the average $\overline{\Gamma}$ across all localized sectors ; see Appendix B for details). This trend affects both the 4- and the 6-digit industries, with and without employment weights. Although industries tend to display less localization when using the employment-weighted K-

Table 1.3 Summary statistics for K-density estimates.

4-digit industries

	2001, unweighted				2001, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	47	54.651	52	60.465	47	54.651	49	56.977
random	26	30.233	20	23.256	29	33.721	29	33.721
dispersed	13	15.116	14	16.279	10	11.623	8	9.302
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.051		0.050		0.057		0.059	
$\overline{\Psi}\|_{\Psi_i>0}$	0.024		0.026		0.018		0.027	

	2005, unweighted				2005, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	48	55.814	50	58.140	42	48.837	46	53.488
random	24	27.907	17	19.767	33	38.372	30	34.884
dispersed	14	16.279	19	22.093	11	12.791	10	11.628
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.043		0.038		0.050		0.045	
$\overline{\Psi}\|_{\Psi_i>0}$	0.027		0.027		0.020		0.023	

	2009, unweighted				2009, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	42	48.837	47	54.651	34	39.535	36	41.860
random	29	33.721	23	26.744	39	45.349	40	46.512
dispersed	15	17.442	16	18.605	13	15.116	10	11.628
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.039		0.035		0.044		0.036	
$\overline{\Psi}\|_{\Psi_i>0}$	0.029		0.028		0.017		0.030	

6-digit industries

	2001, unweighted				2001, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	79	30.620	100	38.610	88	34.109	105	40.541
random	153	59.302	120	46.332	157	60.853	132	50.965
dispersed	26	10.078	39	15.058	13	5.039	22	8.494
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.082		0.062		0.072		0.059	
$\overline{\Psi}\|_{\Psi_i>0}$	0.018		0.018		0.008		0.016	

	2005, unweighted				2005, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	78	30.116	105	40.541	69	26.641	96	37.066
random	150	57.915	108	41.699	170	65.637	139	53.668
dispersed	31	11.969	46	17.761	20	7.722	24	9.266
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.069		0.044		0.085		0.047	
$\overline{\Psi}\|_{\Psi_i>0}$	0.016		0.019		0.012		0.014	

	2009, unweighted				2009, weighted			
	Strict		Extended		Strict		Extended	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
localized	64	24.710	94	36.293	62	23.938	80	30.888
random	163	62.934	120	46.332	180	69.498	148	57.143
dispersed	32	12.355	45	17.375	17	6.564	31	11.969
$\overline{\Gamma}\|_{\Gamma_i>0}$	0.071		0.044		0.077		0.047	
$\overline{\Psi}\|_{\Psi_i>0}$	0.016		0.018		0.012		0.012	

Notes : See the Appendix for details on how to compute Γ_i and Ψ_i. We denote their arithmetic average by $\overline{\Gamma}$ and $\overline{\Psi}$.

densities than in the unweighted case, the key results remain very similar. Note, however, that employment-weighted K-densities tend to decrease less through time. This may either be due to the geographical dispersion of small firms, or to changes in the industrial concentration of industries, or a mix of both. If, for example, geographically close firms merge, the clusters loose points (in terms of plant counts), which decreases the unweighted localization measures. The employment weighted measures would, instead, not be strongly affected by these mergers since the clusters do not loose employment. Table 1.4 below summarizes changes in the plant-level Herfindahl indices of industries over time. As can be seen from that table, the Herfindahl indices increase, on average, over our study period – the joint result of fewer plants and larger average plant sizes. As can further be seen from Table 1.4, there is a systematic pattern in the data : industries that experienced more dispersion (measured here by a switch from being either significantly localized to being random, or from being random to being significantly dispersed) saw their Herfindahl indices increase, whereas industries that experienced more localization (measured here by a switch from being either random to being significantly localized, or from being significantly dispersed to being random) saw their Herfindahl indices decrease. This provides suggestive evidence that changes in industrial concentration – through, e.g., mergers and acquisitions of spatially proximate firms – correlate with changes in industrial localization. Hence, the tendency towards more dispersion may not be solely driven by the dispersion of small firms as compared to large firms.

It is worth noting that the number of industries that do not significantly depart from randomness is quite large in our samples – around 59

Table 1.4 Changes in the plant-level Herfindahl indices (HI) over time.

	Change in HI		
	Mean	Std. dev.	Obs.
2001-2005, all industries	0.014	0.097	259
2005-2009, all industries	0.007	0.059	259
2001-2005, increasing localization	-0.010	0.050	19
2005-2009, increasing localization	-0.007	0.128	14
2001-2005, decreasing localization	0.010	0.038	21
2005-2009, decreasing localization	0.040	0.087	23

Notes : Changes in the plant-level Herfindahl indices over time.

percent in 2009 – which may be due to either the fine level of sectoral disaggregation, or to the presence of a large number of small plants in our samples, or to the specific structure of the Canadian economy.[6] Table 1.3 summarizes our results for the different sample definitions (strict vs extended), different weighting schemes (unweighted vs weighted), and different industrial aggregation levels (6-digit vs 4-digit).

Since the raw value of the DO index is hard to interpret, we report results using the cumulative distribution function (CDF) associated with the K-density, evaluated at a distance of 50 kilometers. These results are summarize in Tables 1.5 and 1.6 below. Consider, e.g., 'Knit Fabric Mills' (NAICS 313240) in 2001. As can be seen from Table 1.5, the CDF at a distance of 50 kilometers is 0.417. In words, 41.7 percent of plant pairs are located less than 50 kilometers apart in that sector. Alternatively, we can view this as the probability that two randomly drawn plants from that industry are less than 50 kilometers away from each other. Clearly, more than two chances

[6] Previous studies for the UK, France, Germany, or Japan, focus on 'compact countries', whereas Canada is geographically all but 'compact'.

in five is a large value given the geographical extent of Canada. As can be seen from Tables 1.5 and 1.6, various textile and metal-related sectors rank among the most strongly localized industries in the different years.

One advantage of the continuous measures is that they allow us to finely assess at *what distances* localization or dispersion actually occur. The top panel of Figure 1.2 depicts the number of 6-digit industries that are globally localized at each distance between 1 and 800 kilometers, both in the unweighted (left panel) and the weighted (right panel) case in 2001. As one can see, most industries are localized at relatively short distances (up to 150–180 kilometers) or at intermediate distances (about 500 kilometers). The reason is that some industries cluster predominantly in an urban environment – short distances, or distances of about 500 kilometers between major urban centers – whereas other industries cluster in more rural and semi-rural areas between major urban centers (about 200–400 kilometers). These industries are then naturally underrepresented at short distances, because dispersion at some distances is the flip-side of agglomeration at other distances. Observe also that : (i) less industries are localized in 2009 than in 2001, especially at short distances and at intermediate inter-city distances ; and (ii) this trend is stronger in the unweighted case, thereby suggesting that the change in the pattern is driven by smaller plants that either disappear (exit or M&As) or change location.

Last, Figure 1.3 plots the rank-ordered distribution of the Γ_i (solid line) and the Ψ_i (dashed line) measures of the strength of localization and dispersion. As one can see, there are only a small number of highly localized or dispersed industries. Furthermore, most of the industries do not have extreme spatial patterns, which is similar to results for the UK and

Table 1.5 Ten most localized industries according to the DO CDF (unweighted).

NAICS6	Industry name	CDF
	2001	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.471
313240	Knit Fabric Mills	0.417
315210	Cut and Sew Clothing Contracting	0.258
315292	Fur and Leather Clothing Manufacturing	0.234
333220	Rubber and Plastics Industry Machinery Manufacturing	0.206
333519	Other Metalworking Machinery Manufacturing	0.204
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.178
325991	Custom Compounding of Purchased Resins	0.175
332118	Stamping	0.170
336370	Motor Vehicle Metal Stamping	0.159
NAICS6	Industry name	CDF
	2005	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.536
333220	Rubber and Plastics Industry Machinery Manufacturing	0.369
332118	Stamping	0.237
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.230
312210	Tobacco Stemming and Redrying	0.200
315292	Fur and Leather Clothing Manufacturing	0.188
333519	Other Metalworking Machinery Manufacturing	0.188
336370	Motor Vehicle Metal Stamping	0.168
325991	Custom Compounding of Purchased Resins	0.166
315110	Hosiery and Sock Mills	0.158
NAICS6	Industry name	CDF
	2009	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.513
312210	Tobacco Stemming and Redrying	0.282
333220	Rubber and Plastics Industry Machinery Manufacturing	0.256
332991	Ball and Roller Bearing Manufacturing	0.252
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.241
336370	Motor Vehicle Metal Stamping	0.228
315292	Fur and Leather Clothing Manufacturing	0.186
333519	Other Metalworking Machinery Manufacturing	0.180
332118	Stamping	0.180
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.151

Notes : The CDF at distance d is the cumulative sum of the K-densities up to distance d. Results in this table are reported for a distance $d = 50$ kilometers.

Table 1.6 Ten most localized industries according to the DO CDF (employment weighted).

NAICS 6	Industry name	CDF
	2001	
325110	Petrochemical Manufacturing	0.344
313240	Knit Fabric Mills	0.309
333220	Rubber and Plastics Industry Machinery Manufacturing	0.254
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.247
336370	Motor Vehicle Metal Stamping	0.216
315110	Hosiery and Sock Mills	0.207
332118	Stamping	0.199
333519	Other Metalworking Machinery Manufacturing	0.169
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.166
315233	Women's and Girls' Cut and Sew Dress Manufacturing	0.166

NAICS 6	Industry name	CDF
	2005	
333220	Rubber and Plastics Industry Machinery Manufacturing	0.277
312210	Tobacco Stemming and Redrying	0.241
336370	Motor Vehicle Metal Stamping	0.192
313240	Knit Fabric Mills	0.179
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.169
332118	Stamping	0.162
315210	Cut and Sew Clothing Contracting	0.157
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.157
333519	Other Metalworking Machinery Manufacturing	0.156
333511	Industrial Mould Manufacturing	0.155

NAICS 6	Industry name	CDF
	2009	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.459
312210	Tobacco Stemming and Redrying	0.249
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	0.209
336370	Motor Vehicle Metal Stamping	0.207
333220	Rubber and Plastics Industry Machinery Manufacturing	0.188
332118	Stamping	0.158
333519	Other Metalworking Machinery Manufacturing	0.156
333511	Industrial Mould Manufacturing	0.142
332991	Ball and Roller Bearing Manufacturing	0.135
325520	Adhesive Manufacturing	0.132

Notes : The CDF at distance d is the cumulative sum of the K-densities up to distance d. Results in this table are reported for a distance $d = 50$ kilometers.

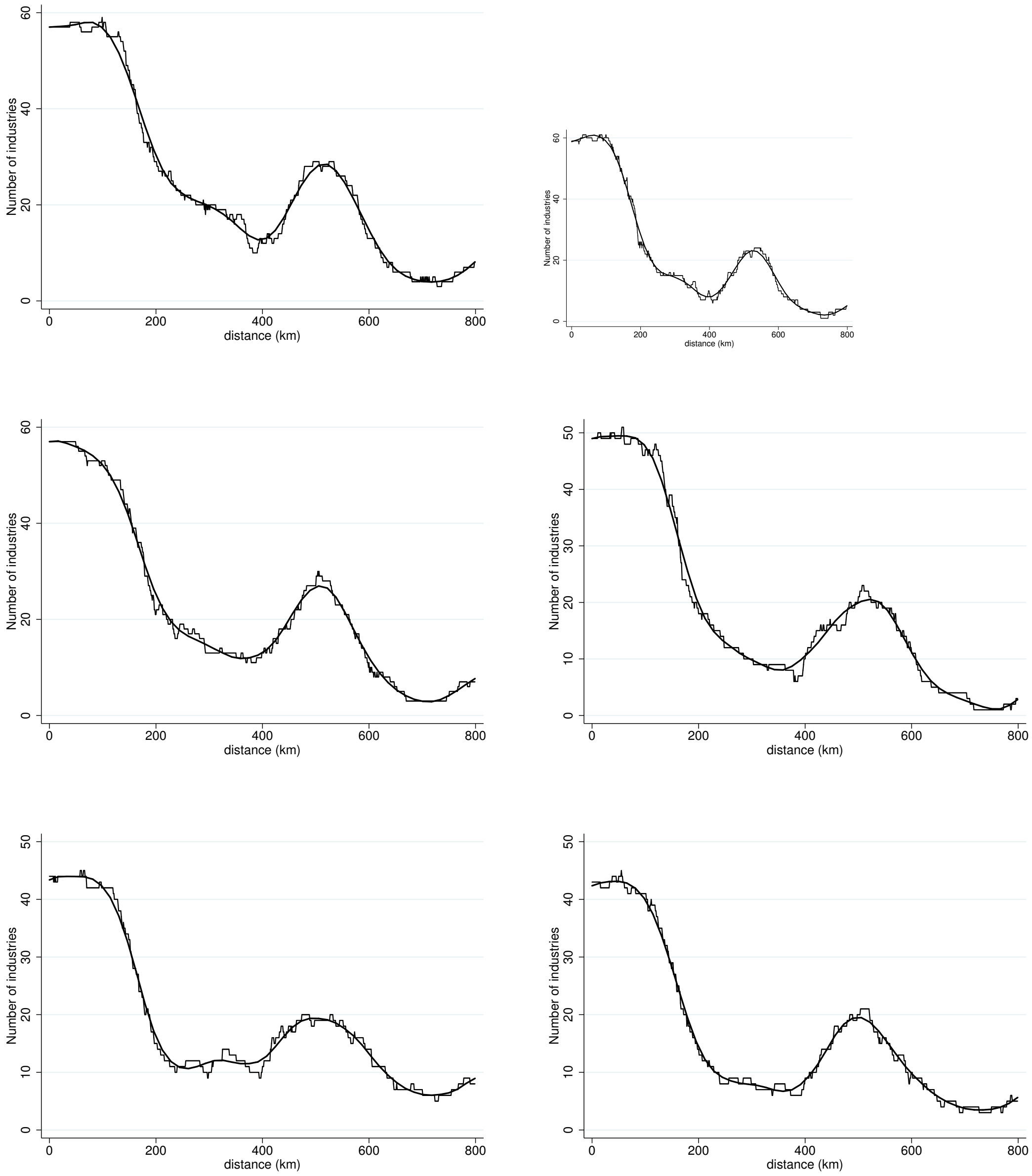


Figure 1.2 Global localization, unweighted (left panel) and weighted (right panel) K-densities for 2001 (top), 2005 (middle), and 2009 (bottom).

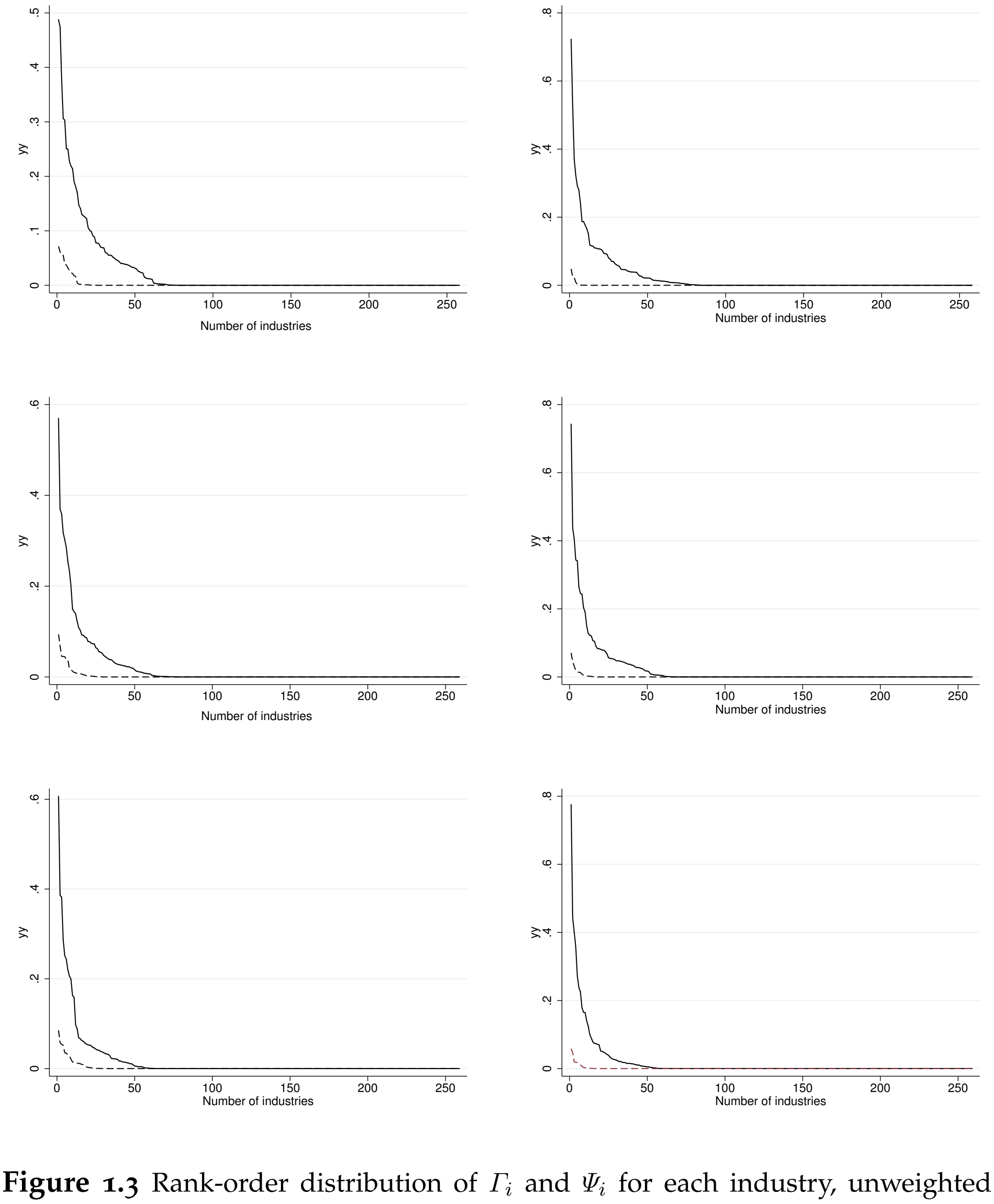


Figure 1.3 Rank-order distribution of Γ_i and Ψ_i for each industry, unweighted (left panel) and weighted (right panel) K-densities for 2001 (top), 2005 (middle), and 2009 (bottom).

Japan. One can also see that the number of localized industries decreases over time, both in the unweighted and in the weighted case, while there is not much change in the degree and strength of dispersion, as well as in the number of dispersed industries. Last, it is worth noting that some of the most strongly localized industries tend to get even more strongly localized. These findings suggest an interesting insight : over the 2001–2009 period, manufacturing industries got generally less localized in Canada, but localization increased at the very top of the distribution. The general trend of spatial deconcentration thus does not affect all industries in the same way and there is substantial cross-industry heterogeneity in locational dynamics

1.3.2 Sectoral scope of localization

Does the level of sectoral aggregation matter for our results ? Do NAICS 4-digit industries exhibit comparable location patterns than NAICS 6-digit industries ? The short answers to those two questions are 'yes' and 'no'. As can be seen from Table 1.3, as we move to a more aggregate definition of industries, the degree of concentration changes. There are two reasons for this. The first is that, as explained in detail in Section 1.4.2 later on, aggregation tends to mix sub-industries that exhibit different location patterns. This is problematic, especially since location patterns are often strong for those sub-industries (see Figure 1.3 ; and Duranton and Overman, 2005). The second reason is that, when breaking down industries into sub-industries, the number of plants gets smaller. This makes the test weaker against the reference distribution, i.e., the K-density confidence bands become wider and localization is more difficult to detect.

In Table 1.7, we compute the ratio of localized 6-digit industries in the total number of 6-digit industries that make up a particular 3-digit industry. The results of this exercise are summarized in Table 1.7. As can be seen from that table, 6-digit industries belonging to the 3-digit industries 313 ('Textile Mills'), 315 ('Clothing Manufacturing'), 323 ('Printing and Related Support Activities'), 333 ('Machinery Manufacturing'), and 334 ('Computer and Electronic Product Manufacturing') are made up of subindustries that display strong localization patterns. On the contrary, 6-digit sectors belonging to industries 324 ('Petroleum and Coal Products'), 312 ('Beverage and Tobacco'), and 321 ('Wood products') display only very weak patterns of localization. These findings are similar to those for the UK, where textile (SIC 17-19) and publishing (SIC 22) industries are among the most localized industries, while food and drink (SIC 15), wood (SIC 20), and petroleum (SIC 23) industries are among the least localized ones (see Duranton and Overman, 2005). The pattern is also similar to that observed in Japan by Nakajima, Saito, and Uesugi (2012), where the most localized industries are related to 'Textile Mill Products' (JSIC 11), 'Electrical Machinery' (JSIC 27), whereas the least localized are related to 'Petroleum and Coal Products' (JSIC 18), and 'Lumber and Wood Products' (JSIC 13).

These results are useful for two reasons. First, as already mentioned, they show that industrial aggregation often mixes sub-industries that display quite different – and fairly strong – location patterns. As we argue in Section 1.4.2, this can significantly affect the outcome, similar to the MAUP in the case of spatial aggregation. Second, it shows that some industries are characterized by either production processes or outputs that display a general tendency to localization. It seems, e.g., that 'textile related' industries generally rely on similar inputs, techniques, and labor that is conducive to the spatial concentration of plants operating in those industries. However, in most 3-digit industries, localized, random, and dispersed sub-industries coexist. Hence, the analysis should be carried out at a detailed industrial level (or even the product level) in order to pick up the fine sectoral location patterns.

In a nutshell, Table 1.7 suggests that the finest 6-digit classification is probably the most appropriate for looking at location patterns. Moving to a more detailed industry classification allows us to pick up more detailed location patterns. The cost of this disaggregation is, however, less precision of the tests against the reference distribution as reflected by the width of the confidence bands.

1.3.3 Location patterns of small plants, young plants, and exporters

We now look at the location patterns of specific subsets of plants : small plants, young plants, and exporting plants. There are good theoretical reasons to look at those plants in particular. Rosenthal and Strange (2003, 2010) document, e.g., that the marginal effect on the entry of new plants in an industry generated by an employee at a small establishment is greater than that generated by an employee at a large establishment. The intuition is that small firms rely more on their external environment, whereas larger firms 'do their own business' (see also Alcácer and Chung, 2013, who link the clustering of small plants to the industrial structure of incumbents in a cluster). Rosenthal and Strange (2010) also provide an extensive review of

Table 1.7 Localization patterns by broad industry groups.

NAICS3	Industry name	#subsectors	#localized	#random	#dispersed	% localized
	Unweighted K-density estimates					
311	Food Manufacturing	32	3	26	3	9.375
312	Beverage and Tobacco Product Manufacturing	6	0	5	1	0.000
313	Textile Mills	7	3	4	0	42.857
314	Textile Product Mills	4	0	3	1	0.000
315	Clothing Manufacturing	17	13	4	0	76.471
316	Leather and Allied Product Manufacturing	3	1	2	0	33.333
321	Wood Product Manufacturing	14	4	6	4	28.571
322	Paper Manufacturing	12	3	8	1	25.000
323	Printing and Related Support Activities	6	3	3	0	50.000
324	Petroleum and Coal Products Manufacturing	4	0	4	0	0.000
325	Chemical Manufacturing	20	7	12	1	35.000
326	Plastics and Rubber Products Manufacturing	14	4	10	0	28.571
327	Non-Metallic Mineral Product Manufacturing	12	1	9	2	8.333
331	Primary Metal Manufacturing	13	3	10	0	23.077
332	Fabricated Metal Product Manufacturing	21	8	12	1	38.095
333	Machinery Manufacturing	17	9	5	3	52.941
334	Computer and Electronic Product Manufacturing	9	5	2	2	55.556
335	Electrical Equipment, Appliance and Component Manufacturing	12	2	10	0	16.667
336	Transportation Equipment Manufacturing	18	3	9	6	16.667
337	Furniture and Related Product Manufacturing	10	4	6	0	40.000
339	Miscellaneous Manufacturing	7	3	3	1	42.857
	Weighted K-density estimates					
311	Food Manufacturing	32	3	28	1	9.375
312	Beverage and Tobacco Product Manufacturing	6	0	6	0	0.000
313	Textile Mills	7	6	1	0	85.714
314	Textile Product Mills	4	1	3	0	75.000
315	Clothing Manufacturing	17	16	0	1	94.118
316	Leather and Allied Product Manufacturing	3	1	2	0	33.333
321	Wood Product Manufacturing	14	2	5	7	14.286
322	Paper Manufacturing	12	7	4	1	58.333
323	Printing and Related Support Activities	6	0	3	1	0.000
324	Petroleum and Coal Products Manufacturing	4	0	4	0	0.000
325	Chemical Manufacturing	20	10	10	0	50.000
326	Plastics and Rubber Products Manufacturing	14	7	7	0	50.000
327	Non-Metallic Mineral Product Manufacturing	12	1	9	2	8.333
331	Primary Metal Manufacturing	13	3	10	0	23.077
332	Fabricated Metal Product Manufacturing	21	9	12	9	42.857
333	Machinery Manufacturing	17	9	5	3	52.941
334	Computer and Electronic Product Manufacturing	9	8	1	0	88.889
335	Electrical Equipment, Appliance and Component Manufacturing	12	6	6	0	50.000
336	Transportation Equipment Manufacturing	18	8	9	1	44.444
337	Furniture and Related Product Manufacturing	10	2	7	1	20.000
339	Miscellaneous Manufacturing	7	3	2	2	42.857

Notes : Results are reported for the year 2001. The measures are computed using the unweighted K-densities (top panel) and the employment-weighted K-densities (bottom panel). Subsectors are identified at the 6-digit level.

the theoretical mechanisms that explain the importance of small and young establishments for clustering and industry dynamics, especially the entry and clustering of new firms through, e.g., 'spin offs' or their greater reliance on locally sourced external services. Turning to the importance of young firms for growth, there is abundant evidence that clusters and cities with younger firms and more entrepreneurship have higher growth rates (see, e.g., Faberman, 2011 ; Glaeser, Kerr, and Kerr, 2014), thus suggesting that small and young firms are important for economic development. Given the widely documented and large effects of small plants and of young plants on industry dynamics and growth, it seems worthwhile to investigate in more detail their geographical location patterns. Their spatial concentration may indicate that clustering is conducive to the creation of new plants and jobs.

Looking at location patterns, industrial concentration, and the propensity of small US firms to export, Mittelstaedt, Ward, and Nowlin (2006) find that the greater the geographic concentration of an industry, the higher the likelihood that firms will export. Greenaway and Kneller (2008) reach a similar conclusion in a study covering fifteen years of firm-level data in the UK. Export spillovers – and thus the tendency for exporters to concentrate geographically – are also documented at length in Koenig (2009) and Koening, Mayneris, and Poncet (2010).[7] Note that the clustering of exporters (if

[7] Not all studies find evidence for the existence of export spillovers. By using relatively aggregated measures of agglomeration (regions are approximated by US states and industries at the 2-digit level), Bernard and Jensen (2004) find no role for either geographic spillovers or for export activity of other firms in the same industry for a panel of large US plants. Another example is the paper of Barrios, Goerg, and Strobel (2003), who use a panel of Spanish firms to document that there is no evidence for spillover effects through the presence of other exporters or multinationals.

there are export spillovers) can imply that attracting exporters may have a beneficial effect on other plants which may subsequently also engage in export activity. Given the widely-documented fact that exporters pay higher wages (e.g., Bernard and Jensen, 1995), these plants seem like prime targets for economic policy.

We define young plants and small plants as those plants that are below the median employment size or year of establishment in their industry.[8] Instead of using the overall distribution of manufacturing as the benchmark, we now consider the distribution of all plants in their particular industry as the benchmark. The question is hence : Do small plants, young plants, or exporters locate closer to each other than plants in the industry in general ?

Table 1.8 summarizes our results. Across years, we find that only 7 to 11 industries (3 to 4 percent) exhibit localization of small plants, whereas 13 to 19 industries (5 to 7 percent) exhibit dispersion of small plants. This leaves more than 90 percent of industries with location patterns of small plants that do not differ significantly from randomness. These findings sug-

[8] Note that using the median may seem a priori arbitrary. In unreported results, we have also split the sample at the first quartile. The results (available upon request), are very similar. Note, however, that stricter definitions yield smaller sample sizes, so that the estimates are less precise. Note also that we cannot use an absolute criterion to split the samples. The reason is that the efficient size of plants vastly differs across industries. A ’small plant’ in a chemical industry may correspond to a ’huge plant’ in a textile sector. For example, in the "Fur and Leather Clothing Manufacturing" industry, around 90 percent of plants have less than 20 employees while in the "Alkali and Chlorine Manufacturing" industry, the average establishment size is 200 employees. Clearly, using an absolute threshold to classify plants is not meaningful.

gest that small plants in an industry do not locate differently than its plants in general. This weak tendency for clustering of small plants is consistent with Duranton and Overman's (2008) findings for the UK. We obtain very similar results for young plants, as can be seen from Table 1.8. Turning to exporters, these plants exhibit somewhat more localization. There are indeed 36 to 41 industries (14 to 16 percent) that exhibit localization of exporters, whereas only 11 to 28 (4 to 11 percent) exhibit dispersion of exporters. Even though these figures are larger than for small plants and young plants, three-quarter of industries display no clear pattern with respect to the geographical distribution of their exporters. Hence, there is little evidence that small plants, young plants, or exporters are more localized than their industries (see also Table 1.18 in the appendix, which reports the conditional probabilities of young plants, small plants, or exporters to be localized/dispersed/random conditional on whether the industry they belong to is localized/dispersed/random).

One may worry that our finding that many industries display random patterns is driven by small sample sizes. To check the robustness of our results, we thus restrict our industries conservatively to subsamples with at least 25 plants and run our estimations again. Doing so leaves us with 170 to 190 industries – depending on the year and the subsample. As one can see from the right part of Table 1.8, the results are similar, thus suggesting that they are not biased because of sectors with small sample sizes.

When looking at the specific industries that underlie the foregoing figures, we find again a very heterogeneous group of industries. The three industries with the most localized subgroups of plants in 2009, for example, are : (i) 'All Other Plastic Product Manufacturing' (NAICS 326198), 'Other

Motor Vehicle Parts Manufacturing' (NAICS 336390), and 'Coating, Engraving, Heat Treating and Allied Activities' (NAICS 332810) for small plants ; (ii) 'Pottery, Ceramics and Plumbing Fixture Manufacturing' (NAICS 327110), 'All Other Industrial Machinery Manufacturing' (NAICS 333299), and 'All Other Plastic Product Manufacturing' (NAICS 326198) for young plants ; and (iii) 'Sawmills (except Shingle and shake Mills)' (NAICS 321111), 'Prefabricated Wood Building Manufacturing' (NAICS 321992), and 'Other Animal Food Manufacturing' (NAICS 311119) for exporters.

Table 1.8 Summary statistics for small, young, and exporter subsamples.

	Small plants		Young plants		Exporters		Small plants		Young plants		Exporters	
	2001, all 6-digit industries						2001, restricted 6-digit industries					
Status	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
localized	10	3.891	16	6.226	41	15.953	9	4.945	11	6.077	37	19.271
random	228	88.716	239	92.996	205	79.767	153	84.066	168	92.818	146	76.042
dispersed	19	7.393	2	0.778	11	4.280	20	10.989	2	1.105	9	4.688
$\overline{\Gamma}\vert_{\Gamma_i>0}$	0.021		0.003		0.023		0.021		0.004		0.025	
$\overline{\Psi}\vert_{\Psi_i>0}$	0.008		0.006		0.006		0.007		0.006		0.007	
	2005, all 6-digit industries						2005, restricted 6-digit industries					
Status	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
localized	11	4.264	8	3.113	36	13.900	10	5.464	4	2.210	30	15.306
random	232	89.922	242	94.163	195	75.290	158	86.339	171	94.475	141	71.939
dispersed	15	5.814	7	2.724	28	10.811	15	8.197	6	3.315	25	12.755
$\overline{\Gamma}\vert_{\Gamma_i>0}$	0.006		0.062		0.013		0.007		0.123		0.015	
$\overline{\Psi}\vert_{\Psi_i>0}$	0.006		0.003		0.002		0.006		0.003		0.002	
	2009, all 6-digit industries						2009, restricted 6-digit industries					
Status	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
localized	7	2.713	11	4.280	37	14.341	6	3.550	3	1.775	29	15.847
random	238	92.248	238	92.607	198	76.744	152	89.941	159	94.083	133	72.678
dispersed	13	5.039	8	3.113	23	8.915	11	6.509	7	4.142	21	11.475
$\overline{\Gamma}\vert_{\Gamma_i>0}$	0.002		0.008		0.020		0.002		0.011		0.026	
$\overline{\Psi}\vert_{\Psi_i>0}$	0.005		0.002		0.002		0.006		0.002		0.003	

Notes : See the Appendix for details on how to compute Γ_i and Ψ_i. We denote their arithmetic average by $\overline{\Gamma}$ and $\overline{\Psi}$. The restricted industries case includes only industries with samples of more than 25 plants of that specific type.

Figure 1.4 depicts the global localisation (left panel) and the rank-order distribution of localized and dispersed industries (right panel) for small plants, young plants, and exporters in 2009. Despite some differences – especially for small and young plants, where there is only very little localization – the general shape of these graphs is similar to the baseline case : most industries are localized at relatively short or at intermediate distances. The number of industries in these ranges is far smaller than the number in the baseline case. The number of dispersed industries (not shown here) is increasing over the entire range of distances between 0 and 800 kilometers. It is also increasing across years. This mirrors our general finding that industries – and specific subgroups of plants – have had a tendency to geographically disperse in Canada over the first decade of 2000. As one can also see from Figure 1.4, the rank-order distributions of localized and dispersed industries are quite similar to those in the baseline case. It is worth noting that exporters are both more strongly localized in terms of the number of industries that display a significant localization (bottom left panel of Figure 1.4), and also substantially more in the strength of localization of the industries with the most clustered exporting plants (bottom right panel of Figure 1.4). Thus, there is some evidence that exporters 'locate differently'. [9]

1.4 Robustness analysis : Results with discrete measures

To check the robustness of our key findings, we now provide results on the geographical concentration of industries using discrete measures of localization. More precisely, we start by computing the ubiquitious Ellison-Glaeser index (Ellison and Glaeser, 1997). This measure, though somewhat

[9] Our analysis does not allow us to assess whether exporting plants have a tendency to cluster, or whether clustering makes plants export. See, e.g., Koenig, Mayneris, and Poncet (2010) for evidence on the latter.

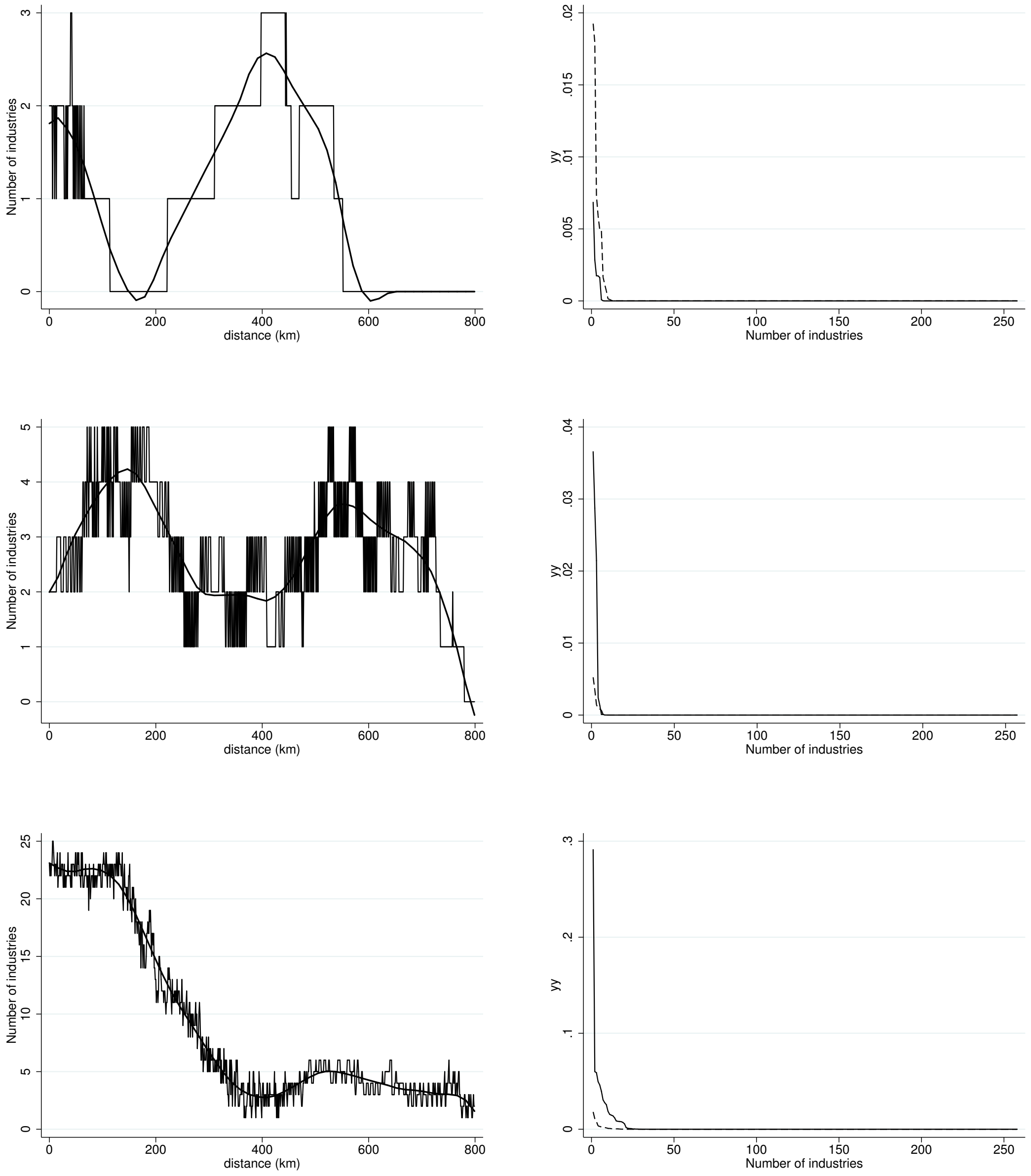


Figure 1.4 Global localization, unweighted K-densities (left panel) and rank-order distribution of Γ_i and Ψ_i (right panel) for small plants (top), young plants (middle), and exporters (bottom) in 2009.

sensitive to the way space is subdivided into administrative units, has been widely used in the literature and will allow us to compare our results to existing ones. We also compute a 'spatially weighted' version of the EG index to take into account 'neighborhood effects', i.e., the fact that industry concentrations may stretch across several adjacent administrative units. This spatially weighted version of the EG index, due to Guimarães, Figueiredo, and Woodward (2011) and henceforth denoted by EGspat, has not been much used in the literature until now (see Appendix C for methodological details).

1.4.1 Baseline results

We compute the EG index – and its spatially weighted version – for 2001, 2005, and 2009 at the NAICS 6-digit level using three different spatial scales : provinces (PROV), economic regions (ER), and census divisions (CD). We implement two spatial weighting schemes. The first is based on the geographical distance between the centroids of the spatial units. The second one – which more accurately captures the fact that agglomerations may extend across *borders* – is based on the common length of the border between two adjacent units computed from GIS data. Our key findings, shown in Table 1.9, can be summarized as follows.

First, about 70 to 75 percent of manufacturing industries are localized in Canada according to the EG index. This fraction is lower than the one reported for the US (97 percent), France (95 percent), and the UK (94 percent) in earlier studies by Ellison and Glaeser (1997), Maurel and Sédillot (1999), and Duranton and Overman (2005). [10]

Second, the number of localized manufacturing industries in Canada has decreased between 2001 and 2009. This can be seen in terms of numbers, but also from the decrease in the mean value of the EG index at all spatial scales, safe for the smallest one (CD). We also find that there is a sizeable share of sectors for which the EG index is negative, thus suggesting that dispersion prevails – and increases over time – in some industries. When taken together, these results show that manufacturing industries have become less geographically concentrated over the first decade of 2000, thus corroborating our findings using continuous measures.

Third, despite some changes across industries, the EG index is, on average, smaller than its spatially weighted counterpart (see the two bottom panels of Table 1.9). Put differently, spatial concentration extends over multiple adjacent spatial units, and this fact has to be taken into account when computing the EG index. Note that all our results are fairly robust across years, spatial scales, and to the use of the chosen weighting scheme

[10] Duranton and Overman (2005) note that the definition of 'weak localization' by Ellison and Glaeser (1997) picks up manufacturing industries in the UK which have a pattern that is not significantly different from that of spatial randomness. Our mean value for the unweighted index at the ER level is very close to the one of 0.034 reported by Duranton and Overman (2005), whereas our median is somewhat higher. We performed a one-sided statistical test following Ellison and Glaeser (1997) by assuming that $\widehat{\gamma}$ in the EG index and $\widehat{\gamma}_S$ in the EGspat index are asymptotically normally distributed (see Appendix C for further details). At a 5% significance level we find that, on average, 40–60% of the γ and γ_S parameters of industries are significant. Hence, location choices of plants are not independent in 40–60% of the industries. Note that these figures are lower than the shares reported in Table 1.9 which are based on Ellison and Glaeser's (1997) 'rule of thumb'.

for computing the EGspat index (see Table 1.17 in Appendix E).

Figure 1.5 summarizes the distributions of the EG and EGspat indices for the 259 6-digit manufacturing industries in 2001, 2005, and 2009. Observe that these distributions are quite skewed towards zero, i.e., only few industries are highly localized, whereas a majority of them are weakly localized – the EG index is positive but less than 0.05. These results are similar to the ones reported by Maurel and Sédillot (1999) for French industries, and by Ellison and Glaeser (1997) for US industries. We can also see that, despite the general trend towards a decrease in localization between 2001 and 2009, the overall distributions of the EG and EGspat indices have remained fairly stable over time. [11]

Table 1.10 lists the ten most and the ten least localized industries at the NAICS 6-digit level for the year 2009. As can be seen from that table, and in accordance with the results we established using the DO index of localization in the previous section, various industries related to either textiles or to the extraction and processing of natural resources dominate the group of the most localized industries. This result is robust across localization measures, which suggests that those measures identify the same 'most concentrated' industries. Note that the hierarchy of individual industries is unchanged when using the EGspat index. Indeed, the Spearman-rank correlation between the EG and the EGspat indices is 0.96. This suggests that,

[11] The correlation of the EG indices across industries in 2001 and 2009 varies from about 0.83 at the province level to 0.73 at the census division level. One reason for the differences across geographical scales is that the processes generating province-level agglomeration are likely to be different from the ones generating agglomeration at the economic region and census division levels (see, e.g., Rosenthal and Strange, 2001, 2003).

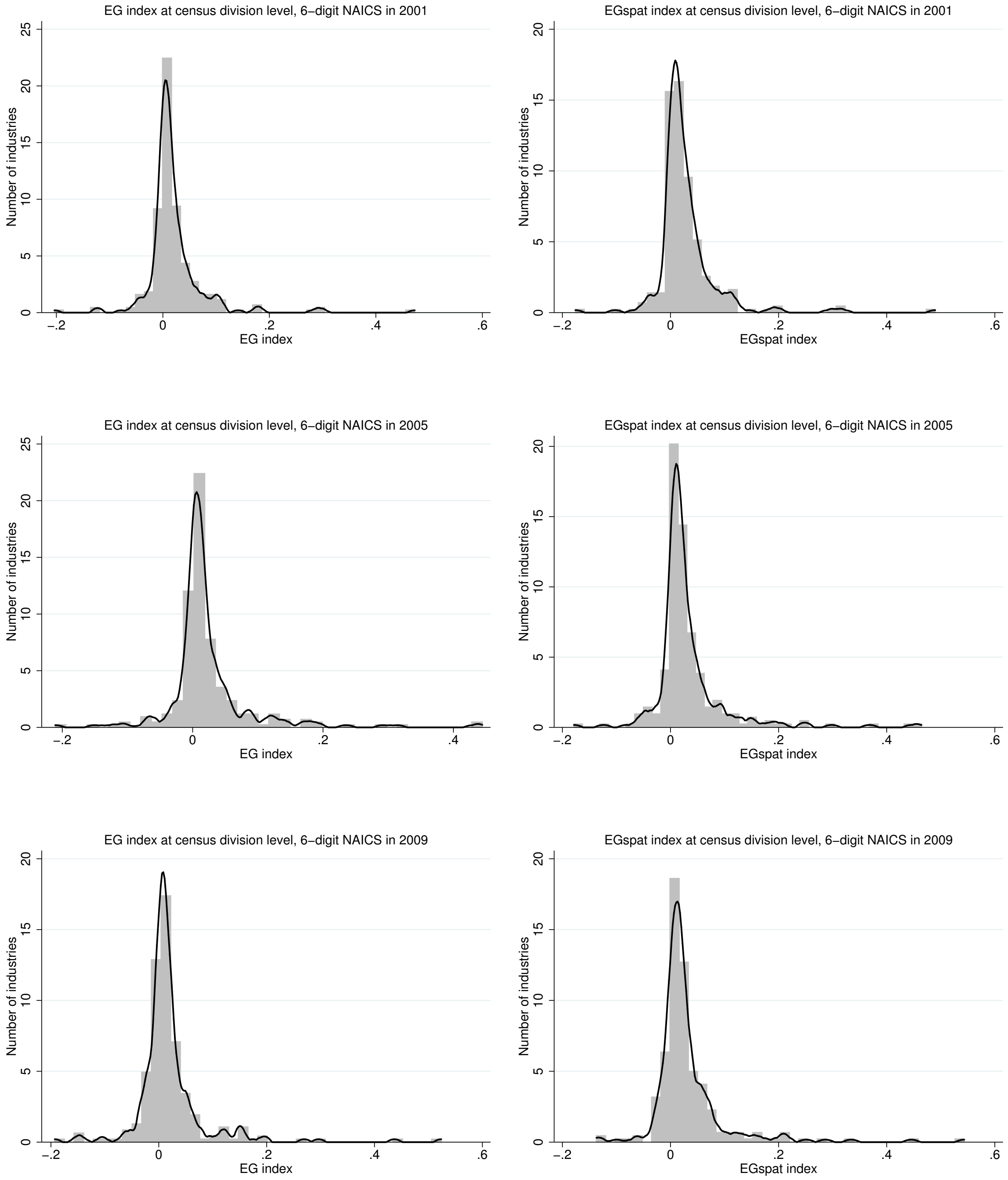


Figure 1.5 Distribution of the EG index at the CD level (NAICS 6-digit), unweighted EG (left panel) and spatially-weighted EGspat (right panel).

Table 1.9 Mean and median EG and EGspat indices at different spatial scales (NAICS 6-digit).

	2001			2005			2009		
Geography	PROV	ER	CD	PROV	ER	CD	PROV	ER	CD
	Unweighted EG								
Mean	0.074	0.036	0.021	0.073	0.035	0.023	0.060	0.032	0.020
Median	0.023	0.021	0.010	0.023	0.018	0.010	0.019	0.015	0.010
Share < 0	31.660	23.552	26.255	35.521	25.483	25.483	36.154	29.615	29.231
Share $\in (0, 0.05]$	26.255	47.876	58.301	23.552	47.876	59.459	27.692	44.231	56.538
Share > 0.05	42.085	28.571	15.444	40.927	26.641	15.058	36.154	26.154	14.231
	EGspat, weighted by the inverse distance matrix								
Mean	0.080	0.047	0.029	0.086	0.049	0.032	0.077	0.048	0.031
Median	0.025	0.026	0.017	0.028	0.024	0.0157	0.024	0.024	0.016
Share < 0	31.660	17.375	16.602	34.363	18.533	16.602	33.846	20.769	20.000
Share $\in (0, 0.05]$	25.869	47.876	65.251	23.552	47.104	64.479	26.538	45.769	60.000
Share > 0.05	42.471	34.749	18.147	42.085	34.363	18.919	39.615	33.462	20.000
	EGspat, weighted by the common border length								
Mean	0.077	0.051	–	0.093	0.054	–	0.085	0.052	–
Median	0.027	0.030	–	0.026	0.027	–	0.021	0.024	–
Share < 0	32.432	17.761	–	31.274	19.691	–	33.462	23.846	–
Share $\in (0, 0.05]$	26.641	45.946	–	27.027	44.402	–	28.077	41.538	–
Share > 0.05	40.927	36.293	–	41.699	35.907	–	38.462	34.615	–

Notes : Mean and median values for 259 (resp., 260 in 2009) NAICS 6-digit industries. Share < 0 means 'not clustered'. Share $\in (0, 0.05]$ means 'weakly clustered'. Share > 0.05 means 'strongly clustered'. See Ellison and Glaeser (1997) for details.

in the case of Canada, industrial concentrations do not extend 'too much' across geographical units.

Table 1.10 Ten most and least localized 6-digit industries in 2009, EG and EGspat indices.

NAICS 6	Most localized industries in 2009	EG	EGspat
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.524	0.543
315233	Women's and Girls' Cut and Sew Dress Manufacturing	0.437	0.446
315221	Men's and Boys' Cut and Sew Underwear and Nightwear Manufacturing	0.296	0.334
333130	Mining and Oil and Gas Field Machinery Manufacturing	0.269	0.286
313240	Knit Fabric Mills	0.196	0.216
315232	Women's and Girls' Cut and Sew Blouse and Shirt Manufacturing	0.195	0.204
315190	Other Clothing Knitting Mills	0.174	0.202
325181	Alkali and Chlorine Manufacturing	0.155	0.251
321112	Shingle and Shake Mills	0.154	0.214
311111	Dog and Cat Food Manufacturing	0.151	0.180

NAICS 6	Least localized industries in 2009	EG	EGspat
315227	Men's and Boys' Cut and Sew Trouser, Slack and Jean Manufacturing	-0.056	-0.034
339930	Doll, Toy and Game Manufacturing	-0.059	-0.056
336330	Motor Vehicle Steering and Suspension Components (except Spring) Manufacturing	-0.063	-0.033
335110	Electric Lamp Bulb and Parts Manufacturing	-0.072	-0.056
311830	Tortilla Manufacturing	-0.100	0.021
333611	Turbine and Turbine Generator Set Unit Manufacturing	-0.109	-0.099
327990	All Other Non-Metallic Mineral Product Manufacturing	-0.139	-0.137
312210	Tobacco Stemming and Redrying	-0.148	-0.072
325110	Petrochemical Manufacturing	-0.155	-0.012
321217	Waferboard Mills	-0.193	-0.129

Notes : EG and EGspat indices computed at the 6-digit NAICS level. The spatial scale used is census divisions (CD), and the weighting is inverse distance between CD centroids.

1.4.2 Sectoral scope of localization

We next look again at the sectoral scope of localization.[12] They are similar to the results using the continuous measures. We find less concentration across both years and geographical scales. At the census division level, less than 13 percent – around 34 industries – are found to be localized in terms of small plants, young plants, and exporters. Most of these industries are, however, strongly localized. Table 1.11 summarizes our results for the 86 NAICS 4-digit industries. As can be seen from the table, there are fewer dispersed industries (share < 0) at the 4-digit level as compared to the 6-digit level (11 percent on average in Table 1.11, compared to 17 percent on average in Table 1.9). There are also fewer strongly localized sectors, but to a smaller extent than for dispersed sectors. As pointed out by Haedo and Mouchart (2012), when sectors are aggregated, some dispersed ones are mixed up with concentrated ones to give a 'medium' distribution (Table 1.7 shows that the variation of 'localization types' within 3-digit industries is generally fairly strong ; the same holds true for 4-digit industries). The share of concentrated sectors decreases less because localization is easier to detect and has higher values (see Figure 1.3), while the share of dispersed sectors decreases a lot more during aggregation. More generally, at higher levels of industrial aggregation, it is more difficult to find departures from the reference distribution.

This result is contrary to findings by Rosenthal and Strange (2001),

[12] We also computed the results for young firms, small firms, and exporters, taking the distribution of industry employment as the benchmark. To save space, these results are available upon request.

who find that the average level of agglomeration increases as one moves from 4- to 6-digit industries when computing the EG index. Concerning the geographical scale, we find that agglomeration increases as we go from economic regions to provinces, and from census divisions to economic regions. This is a manifestation of the MAUP that we have mentioned earlier. This finding is in accord with what is know from other studies and countries (Rosenthal and Strange, 2001), and they are aligned with our findings using continuous measures of localization.

Finally, Table 1.12 reveals that there are systematic localization patterns by broad industry groups, as in the case of continuous measures. Some 3-digit industries are made up of many concentrated 6-digit subindustries (e.g., 'Apparel manufacturing' or 'Chemical manufacturing'), whereas others are mostly dispersed (e.g., 'Beverage and Tobacco Product Manufacturing'). This shows again that localization extends across different 3-digit groupings.

1.5 Discussion and concluding remarks

We have used extensive micro-geographic data to provide what we believe is to date the most comprehensive anatomy of the geographical concentration of manufacturing industries in Canada. Looking at the changes between 2001 and 2009 allowed us also to examine the 'dynamics' of localization in a detailed way. The following key results stand out.

First, depending on industry definitions and years, 40 to 60 percent of manufacturing industries are clustered, mainly at short distances and at distances of about 400-500 kilometers. This finding suggests that there

Table 1.11 Mean and median EG and EGspat indices at different spatial scales, NAICS 4-digit industries.

	2001			2005			2009		
Geography	PROV	ER	CD	PROV	ER	CD	PROV	ER	CD
	Unweighted EG								
Mean	0.064	0.033	0.019	0.065	0.031	0.020	0.056	0.027	0.015
Median	0.018	0.023	0.012	0.017	0.016	0.011	0.018	0.010	0.007
Share < 0	16.279	6.977	6.977	22.093	11.628	11.628	23.256	16.279	15.116
Share $\in (0, 0.05]$	45.349	68.605	82.558	43.023	69.767	79.070	46.512	62.791	76.744
Share > 0.05	38.372	24.419	10.465	34.884	18.605	9.302	30.233	20.930	8.140
	EG weighted by the inverse distance matrix								
Mean	0.066	0.036	0.022	0.068	0.035	0.022	0.060	0.031	0.018
Median	0.019	0.027	0.014	0.017	0.021	0.013	0.020	0.013	0.011
Share < 0	18.605	5.814	6.977	23.256	8.140	5.814	22.093	12.791	11.628
Share $\in (0, 0.05]$	43.023	67.442	82.558	41.860	72.093	82.558	48.837	65.116	77.907
Share > 0.05	38.372	26.744	10.465	34.884	19.767	11.628	29.070	22.093	10.465
	EG weighted by the common border length								
Mean	0.064	0.040	–	0.071	0.039	–	0.064	0.033	–
Median	0.022	0.029	–	0.022	0.023	–	0.021	0.015	–
Share < 0	18.605	5.814	–	22.093	8.140	–	23.256	15.116	–
Share $\in (0, 0.05]$	44.186	63.953	–	43.023	65.116	–	38.372	61.628	–
Share > 0.05	37.209	30.233	–	34.884	26.744	–	38.372	23.256	–

Notes : Mean and median values for 86 NAICS 4-digit industries. Share < 0 means 'not clustered'. Share $\in (0, 0.05]$ means 'weakly clustered'. Share > 0.05 means 'strongly clustered'. See Ellison and Glaeser (1997) for details.

Table 1.12 Localization patterns by broad industry groups.

NAICS3	Industry name	Subsectors	# of localized subsectors			# of dispersed subsectors		
			2001	2005	2009	2001	2005	2009
	EG index, unweighted							
311	Food Manufacturing	33	17	19	18	16	14	15
312	Beverage and Tobacco Product Manufacturing	6	3	4	4	3	2	2
313	Textile Mills	7	5	6	5	2	1	2
314	Textile Product Mills	4	2	3	3	2	1	1
315	Apparel Manufacturing	17	15	15	14	2	2	3
316	Leather and Allied Product Manufacturing	3	3	2	3		1	
321	Wood Product Manufacturing	14	12	13	11	2	1	3
322	Paper Manufacturing	12	9	11	5	3	1	7
323	Printing and Related Support Activities	6	6	6	3			3
324	Petroleum and Coal Products Manufacturing	4	4	3	2		1	2
325	Chemical Manufacturing	20	15	16	16	5	4	4
326	Plastics and Rubber Products Manufacturing	14	8	11	12	6	3	3
327	Nonmetallic Mineral Product Manufacturing	12	7	9	7	5	3	5
331	Primary Metal Manufacturing	13	9	9	9	4	4	4
332	Fabricated Metal Product Manufacturing	21	15	13	16	6	8	5
333	Machinery Manufacturing	17	15	14	15	2	3	2
334	Computer and Electronic Product Manufacturing	9	8	7	6	1	2	3
335	Electrical Equipment, Appliances and Components	12	9	6	7	3	6	5
336	Transportation Equipment Manufacturing	18	14	12	14	4	6	4
337	Furniture and Related Product Manufacturing	10	9	8	9	1	2	1
339	Miscellaneous Manufacturing	7	6	6	5	1	1	2
	Total	259	191	193	184	68	66	76
	% of localized or dispersed		73.745	74.517	70.769	26.255	25.483	29.231
	EGspat index, weighted by inverse distance							
311	Food Manufacturing	33	23	22	26	10	11	7
312	Beverage and Tobacco Product Manufacturing	6	4	5	4	2	1	2
313	Textile Mills	7	5	7	6	2		1
314	Textile Product Mills	4	3	3	3	1	1	1
315	Apparel Manufacturing	17	15	16	14	2	1	3
316	Leather and Allied Product Manufacturing	3	3	3	3			
321	Wood Product Manufacturing	14	14	14	13			1
322	Paper Manufacturing	12	10	12	8	2		4
323	Printing and Related Support Activities	6	6	6	4			2
324	Petroleum and Coal Products Manufacturing	4	4	3	3		1	1
325	Chemical Manufacturing	20	19	19	17	1	1	3
326	Plastics and Rubber Products Manufacturing	14	10	13	12	4	1	2
327	Nonmetallic Mineral Product Manufacturing	12	8	9	8	4	3	4
331	Primary Metal Manufacturing	13	11	11	10	2	2	3
332	Fabricated Metal Product Manufacturing	21	17	15	17	4	6	4
333	Machinery Manufacturing	17	16	15	15	1	2	2
334	Computer and Electronic Product Manufacturing	9	8	8	6	1	1	3
335	Electrical Equipment, Appliances and Components	12	9	7	8	3	5	4
336	Transportation Equipment Manufacturing	18	15	14	15	3	4	3
337	Furniture and Related Product Manufacturing	10	10	8	9		2	1
339	Miscellaneous Manufacturing	7	6	6	7	1	1	1
	Total	259	216	216	208	43	43	52
	% of localized or dispersed		83.398	83.398	80.000	16.602	16.602	20.000

Notes : The measures are computed using the EG index at the CD level (NAICS 6-digit) unweighted (top panel) and weighted by inverse distance EGspat (bottom panel). Subsectors are identified at the 6-digit level. Blank cells indicate that there are no subsectors in the respective category (localized or dispersed or random).

is less industrial localization in Canada than in other developed countries. Second, according to all measures we computed – continuous, discrete, and spatially weighted discrete – localization has been decreasing from 2001 to 2009. Third, industries related to textiles and to the extraction and processing of natural resources dominate the group of the most localized industries. This finding is in accord with previous results for other countries. Fourth, while there has been a general trend towards less geographical concentration, some of the most strongly localized industries tend to become even more localized. Last, small plants and young plants are, in general, not more strongly concentrated than all plants in their respective industries – there is little evidence that these plants obey a location logic that is different from that of their industry in general. There is some evidence for 'excess concentration' of exporters, but that effect tends to get weaker during the first decade of 2000.

Our analysis leaves three issues unresolved. First, our paper remains silent on the causes for localization and the changes therein. Yet, we need to better understand what agglomeration forces contribute to the clustering of Canadian manufacturing industries. Previous studies – such as Rosenthal and Strange (2001, 2003) or Ellison, Glaeser, and Kerr (2010) – have addressed that question for the US. Disentangling the relative contribution of the different sources of agglomeration in Canada – labor market pooling, input-output linkages, transportation costs, and knowledge spillovers – is the next item on our research agenda but clearly beyond the scope of this paper. We tentatively correlated selected industry characteristics with 'localization status' (localized vs dispersed), but did not pick up significant differences. [13] Localized industries seem marginally more skilled-labor intensive, but do not differ notably from dispersed industries in terms of either intermediate input intensity or capital intensity. As explained below, dispersed industries do, however, seem to be more intensive in Information and Communication Technologies (ICT) capital, thus suggesting that lower communication costs may partly be associated with industrial dispersion.

Second, although continuous measures of localization obviate the need for using rather arbitrary spatial subdivisions, they still do rely on equally arbitrary subdivisions of industries. As shown by our analysis, the results do somewhat depend on industrial classifications. Hence, extending our measures to analyze location patterns in terms of 'plant similarity', like similarity in terms of labor requirements or in terms of input-output structures, seems a necessary step for deriving more robust results on agglomeration patterns and may provide valuable insights into what is driving agglomeration more generally. We leave this very important question again open for future work.

Third, our analysis remains silent on the driving forces for the observed downward trend in the geographical concentration of industries. This issue is partially addressed in Behrens, Bougna, and Brown (2015) and in Behrens (2013). In these studies, we show that declining localization in Canadian manufacturing industries is strongly associated with import

[13] We do not report the results, but they are available upon request. For transportation costs in Canada, see Behrens, Bougna, and Brown (2015). Using a different dataset and methodology, we show in that paper that changes in trucking rates, in input-output linkages, and in international trade exposure drive substantial changes in industrial location patters in Canada between 1992 and 2008.

competition (especially from low-wage Asian countries). After the end of the 'Multi-Fibre Arrangement' in 2005, a surge in textile imports led to a significant decrease in the number of textile-related firms and a strong fall in the degree of localization (as can be seen from the results in Table 4, textile-related industries are among the most geographically concentrated industries ; see also Ellison, Glaeser, and Kerr, 2010). Holmes and Stevens (2014) document similar findings for the case of the furniture industry in the US. Overall, decreasing localization is largely driven by exit of firms, especially in clusters of industrial activity. Relocations and increased geographical mobility of workers can likely be ruled out as explanations for the decrease in localization.[14] Turning to communication costs, we are not aware of any study that convincingly establishes the impact of ICT on the geographical concentration of industries (though this is a channel that is often used in theoretical models). We have tentatively looked at how industry-wide quantity indices of ICT capital services correlate with industrial location patterns.[15] We pick up an effect of this variable : dispersed industries have a significantly larger average value for that variable for the pooled sample of years than localized industries.[16] While we obviously cannot read any

[14] Under the caveat that relocations are very difficult to measure in the data, there is only little relocation of manufacturing plants between provinces in Canada. Yet, the general westward shift of population and manufacturing activity following the development of the oil industry in the Canadian west may help explain a part of the increasing de-concentration.

[15] To this end, we have used the variable ifqk2 ('Quantity Index of ICT Capital Services') from Statistics Canada's KLEMS database to proxy for 'communication costs'.

[16] The value for localized industries is 140.96, while that for dispersed industries is 165.63. The T-statistic of a two-sided equality-of-means test is 2.7107, thus showing that

causal statement out of this simple correlation, it suggests that industries that operate with a more dispersed structure invest more in ICT capital.

We also cannot rule out the potential impact of cluster policies on our results. In Canada, like in many other countries, the federal government has put in place a nation-wide cluster policy program through the National Research Council (NRC). The main objective of these policies is to stimulate lagging regions, to bolster highly performing ones, and to diversify older industrial areas into higher technology ones. The NRC has initiated the 'Technology Cluster Initiatives' to foster the development and growth of technological clusters across Canada. These initiatives may partly affected the observed trends. Note, however, that cluster policies do in general favor the *concentration* of industries, not their dispersion. Yet, we observe a tendency to dispersion over our study period. Thus, de-concentration may have been even stronger in the absence of these cluster policies. Unfortunately, we cannot test these propositions directly, and doing so is beyond the scope of this paper.

One may finally wonder whether the increasing trend towards dispersion between 2005 and 2009 is linked to the financial crisis. We do not think that this is the case. The Lehman Brothers collapse occurred in September 2008, so that our concentration measures in 2009 will hardly be affected by the financial crisis that really hit off in late 2008 (e.g., the 'Great Trade Collapse' of 2008-2009). Firm exit was gradual over a 2-3 year period after the collapse, and this should not affect significantly our 2009 results.

industries that disperse have a significantly larger 'Quantity Index of ICT Capital Services' than industries that localize.

To conclude, our findings have a number of implications for 'cluster policy' and 'regional development'. As countries and regions strive to remain competitive in the face of globalization, governments – both local and national – seek increasingly to support competitive regional clusters – see, e.g., Canada's NRC 'Technology Cluster Initiatives', the French 'Pôles de compétitivité' Program, and the German 'BioRegio' Program. The 2007 OECD report on 'National Policy Approaches to Cluster Strategies' highlights the increasing focus on building strategic research capacity in selected regions as the basis for promoting clusters. Recent economic studies, however, increasingly question the use of cluster policies. There is indeed little evidence that more clustering will have significant effects on average productivity or wages in manufacturing industries (e.g., Duranton, 2011; Duranton, Martin, Mayer, and Mayneris, 2012; Behrens, 2013).[17] Our findings show that the general trend in Canada is towards less industrial localization during the last decade. Although this does not provide per se evidence that localization economies have become less valuable to firms, it suggests at least that implementing clusters against this tendency towards more dispersion might be an uphill battle.

Acknowledgements. We thank Mark Brown and two anonymous referees for very helpful comments and suggestions. We also thank participants at the 2013 NARSC meetings in Atlanta and the 13th annual meetings of CIRPÉE in Lac Beauport for helpful comments. Katherine Pineault provided

[17] Similarly, the general public support for 'entrepreneurship' is getting increasingly criticized. See Shane (2009) for a detailed overview of why the 'average small and young' firm should not be the target of public policy.

excellent research assistance during the early stages of the project. Financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization* is gratefully acknowledged. We also gratefully acknowledge financial support from FQRSC Québec (Grant NP-127178), and from CIRPÉE. All remaining errors are ours.

1.6 Appendix to Chapter 1

This appendix is structured as follows. Appendix A describes in detail our datasets and sources. Appendix B provides details on the Duranton-Overman K-density approach. Appendix C briefly presents the Ellison-Glaeser and the spatially weighted Ellison-Glaeser indices. Appendix D provides information on the comparability of the K-densities through time. Last, Appendix E contains additional tables and results.

A. Data and data sources

This appendix provides details on the data used in this paper and the sources.

Plant-level data and industries. Our analysis is based on the *Scott's National All Business Directories Database*. This establishment-level database contains information on plants operating in Canada, with an extensive coverage of the manufacturing sector. It comprises 54,379 manufacturing plants in 2001, 50,404 in 2005, and 46,391 in 2009 (see Table 3.7 below for a breakdown by province). Our data cover the years 2001, 2005, and 2009. For every etablishment, we have information on its primary 6-digit NAICS code and up to four secondary 6-digit NAICS codes ; the opening year of the establishment ; its employment ; whether or not it is an exporter ; and its 6-digit postal code. The latter allows us to effectively geo-locate the plants.

The Scott's database constitutes probably the best alternative to Statistics Canada's proprietary *Annual Survey of Manufacturers Longitudinal Mi-*

crodata File or the micro-level *Canadian Business Patterns*. As can be seen from Tables 1.13 and 3.7, which provide a comparison of the Scott's National All 2001, 2005, and 2009 databases with Statistics Canada's province-level data from the 2003 and 2005 Annual Survey of Manufacturers (ASM; CANSIM Tables 301–0003 and 301–0006) and from the 2001, 2005, and 2009 Canadian Business Patterns, it has a wide and similar coverage. Those tables also show that, despite the good coverage of manufacturing plants, plants in the economic core provinces (Ontario, Quebec, British Columbia, and Alberta) seem slightly under-represented (at 83% when weighted by employment). The bottom panel of Table 1.13 also provides summary statistics across industries for the two datasets. The cross-industry correlations of the Scott's Data and the CBP data are very high (about 0.93), thus showing that the industrial composition of our large samples is very representative. To summarize, our data are very similar to those of the ASM and the CBP in terms of coverage and both province- and industry-level breakdown of plants and, therefore, provide a fairly accurate picture of the overall manufacturing structure in Canada.[18] are of course free to not do so. Also, small/new establishments may appear in the base with a lag only (and establishments may exit with a lag only), but this is not a big issue for our purpose since we do not exploit the time-series variation of the database.

We consider that a plant is a manufacturer in the strict sense if it reports a manufacturing sector (NAICS 31–33) as its primary sector of activity.

[18] There is no 'sampling frame' strictly speaking (though Scott's uses the Canadian Business Register – which contains the universe of entities – to contact the different establishments in a systematic way to include them into their database). There may be some selection and updating biases, since firms are contacted to sign up but

Since plants in our dataset also report up to four secondary NAICS codes, we can construct two different industry-level samples for the analysis : (i) a *strict sample*, restricted to plants that report a manufacturing sector as their primary sector of activity ; and (ii) an *extended sample* that includes all plants that report a manufacturing sector as one of their sectors of activity, either primary or se condary. We thus can associate plants with industries at different levels of detail.

Table 1.13 Comparing the Scott's National All databases to the Canadian Business Patterns (CBP).

Province	CBP 2001	Scott's 2001	%	CBP 2005	Scott's 2005	%	CBP 2009	Scott's 2009	%
Alberta	5,843	3,933	67.311	5,416	3,455	63.792	5,351	3,581	66.922
British Columbia	8,797	6,219	70.695	8,261	5,371	65.016	7,697	4,991	64.843
Manitoba	1,883	1,654	87.839	1,741	1,481	85.066	1,605	1,263	78.692
New Brunswick	1,446	1,395	96.473	1,195	1,258	105.272	1,018	1,175	115.422
Newfoundland	757	576	76.090	629	540	85.851	508	472	92.913
Nova Scotia	1,832	1,676	91.485	1,483	1,495	100.809	1,225	1,296	105.796
Ontario	25,006	21,306	85.204	23,220	20,966	90.293	21,673	19,637	90.606
Prince Edward Island	354	328	92.655	292	327	111.986	256	280	109.375
Quebec	18,349	15,939	86.866	17,026	14,166	83.202	15,238	12,560	82.426
Saskatchewan	1,378	1,353	98.186	1,259	1,305	103.654	1,151	1,091	94.787
Territories	68	–	–	63	40	63.492	57	45	78.947
Canada	**65,713**	**54,379**	**82.752**	**60,585**	**50,404**	**83.196**	**55,779**	**46,391**	**83.169**
Cross-industry corr CBP/Scott's	0.908			0.939			0.937		
Cross-industry average	253.718	209.958		233.919	194.610		214.535	178.431	
Cross-industry min	4	1		2	2		3	2	
Cross-industry max	3316	3604		3047	2738		2695	2378	
Cross-industry std dev.	380.011	346.940		359.400	310.664		339.320	282.503	

Notes : Province-level breakdown of manufacturing plants (NAICS 31–33) in the 2001, 2005, and 2009 Scott's National All databases versus Statistics Canada's 2001, 2005, and 2009 Canadian Business Patterns (CBP). The descriptive statistics reported as 'cross-industry' in the bottom panel of the table are computed across all industries.

Geographical data. To geolocate plants, we used latitude and longitude data of postal code centroids obtained from Statistics Canada's Postal Code Conversion Files (PCCF). These files associate each postal code with different Standard Geographical Classifications (SGC) that are used for reporting census data. We match plant-level postal code information with geographical

Table 1.14 Comparing Scott's National All to the Annual Survey of Manufacturers.

Province	Statcan ASM 2003	Statcan ASM 2005	Scott's 2001	Scott's 2005	Scott's 2009
Alberta	4,882	7,750	3,933	3,455	3,581
British Columbia	6,933	11,942	6,219	5,371	4,991
Manitoba	1,481	2,307	1,654	1,481	1,263
New Brunswick	963	1,533	1,395	1,258	1,175
Newfoundland and Labrador	522	765	576	540	472
Nova Scotia	1,106	1,944	1,676	1,495	1,296
Ontario	21,470	34,184	21,306	20,966	19,637
Prince Edward Island	211	351	328	327	280
Quebec	15,251	23,042	15,939	14,166	12,560
Saskatchewan	1,008	1,804	1,353	1,305	1,091
Territories	–	–	–	40	45
Total	53,827	85,622	54,379	50,404	46,391

Notes : Province-level breakdown of manufacturing plants (NAICS 31–33) in the 2001, 2005, and 2009 Scott's National All databases versus Statistics Canada's 2003 Annual Survey of Manufacturers (ASM; CANSIM Table 301–0003) and 2005 ASM (CANSIM Table 301–0006). The 2003 ASM reports only employer plants with sales exceeding C$30,000 whereas the 2005 ASM reports information for manufacturing plants (including forestry, which is absent in the 2003 ASM) without any sales threshold (thus including small establishments that would qualify as 'self-employed'). The Canadian Business Patterns 2009 of Industry Canada report 55,779 employer plants in manufacturing (see Table 1.13).

coordinates from the PCCF, using the postal code data for the next year in order to consider the fact that there is a six months delay in the updating of postal codes. The census geography of 1996 and the postal codes as of May 2002 (818,907 unique postal codes) were associated with our 2001 sample. We also matched our 2005 sample with the 2001 Census geography and the postal codes as of January 2007 (861,765 unique postal codes). Finally, our 2009 sample was matched with the census geography of 2006, and the postal codes as of October 2010 (890,317 unique postal codes). Table 3.8 summarizes the geographical structure for the three years and provides details on postal codes and census geographies.

Table 1.15 Geographical structure of the Census and PCCF data.

	Census 1996 in the PCCF	Census 2001 in the PCCF	Census 2006 in the PCCF
Provinces and territories	13	13	13
Economic regions	74	76	76
Census divisions	285	288	288
Census subdivisions	4,410	4,088	3,692
Dissemination areas	34,940	42,297	45,904
Geographical concordance :			
Scott's All year	2001	2005	2009
PCCF version	May 2002	Jan 2007	Oct 2010
Census geography	1996	2001	2006
#unique postal codes	818,907	861,765	890,317

Notes : Geography of the 1996, 2001, and 2006 Censuses and concordances between *Scott's National All* databases and Statistic Canada's PCCFs.

The highest level of geographical aggregation is that of the 10 provinces and 3 territories (PR); the second-highest level is that of the 76 economic regions (ER); the third-highest level is the 293 census divisions (CD); the fourth-highest level is the 5253 census subdivisions (CS); and finally, the finest level is dissemination areas (DA). Census subdivisions, census divisions, and economic regions are useful spatial scales for computing discrete measures of localization like the Ellison and Glaeser (1997) index. Provinces are too coarse a spatial scale, whereas dissemination areas are too fine – most of the time, they contain no plants for any 4- or 6-digit NAICS industries. Note also that each postal code can be associated with multiple DAS. In that case, only one DA figures in the PCCF, so that the total number of DAS in the PCCF is smaller than that in the Census. This problem does not arise for larger geographical scales (provinces, regions, census divisions, and census subdivisions).

Subsamples. We construct three industry subsamples. The first relates to small-scale plants. Instead of using Statistics Canada's definition of small-scale business – a plant with less than 50 full-time equivalent employees or having annual sales of less than $2 million – we consider a plant as being small if its size – as measured by the number of employees – is less than the industry median. Using a fixed employment threshold makes little sense, as the minimum operational scale varies widely across different industries. Based on this criterion, and depending on the year, about 52% of plants in our database are small. We repeat the same exercise to construct our young plants subsample. We consider a plant as being young if its age – measured since the year of its establishment – is less than the industry median. Our last subsample is for exporting plants. Here, we simply select all plants that report some exporting activity.

B. The distance-based Duranton-Overman approach

In this appendix, we briefly recall the logic underlying our continuous measure of localization. Duranton and Overman (2005) propose a methodology that uses bilateral distances across pairs of plants to identify localized industries. The idea is to apply sampling and bootstrapping techniques to determine the distribution of bilateral distances between the plants in an industry, and to compare it to a set of bilateral distances obtained from samples of randomly drawn plants. There are four steps. First, we compute the pairwise distances between all plants in an industry and estimate a kernel density function of the distance distribution. Second, we construct a distribution of counterfactuals to assess whether the location pattern of a given industry departs statistically significantly from randomness. The counterfactuals are constructed on the basis that the plants in a given industry are located randomly among all possible locations where we do observe manufacturing activity. Third, we construct confidence intervals using our counterfactual random location distributions. Last, we test whether an industry is localized or dispersed, by comparing the actual distribution of bilateral distances with the confidence bands derived from the sampling. We provide more information on the four steps in what follows.

First step (kernel densities). Consider industry A with n plants. We compute the great circle distance, using postal code centroids, between each pair of plants in that industry. This yields $n(n-1)/2$ bilateral distances for industry A. Let us denote the distance between plants i and j by d_{ij}. Given n etablishments, the kernel-smoothed estimator of the density of these pairwise distances, which we henceforth call K-density as in Duranton and Overman (2005), at any distance d is :

$$\widehat{K}(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d-d_{ij}}{h}\right), \tag{B.1}$$

where h is the optimal bandwidth, and f a Gaussian kernel function. The distance d_{ij} (in kilometers) between plants i and j is computed as :

$$d_{ij} = 6378.39 \cdot \text{acos}\left[\cos(|\text{lon}_i - \text{lon}_j|)\cos(\text{lat}_i)\cos(\text{lat}_j) + \sin(\text{lat}_i)\sin(\text{lat}_j)\right].$$

We also compute the employment-weighted version of the K-density, which is given by

$$\widehat{K}_W(d) = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(e_i+e_j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (e_i+e_j) f\left(\frac{d-d_{ij}}{h}\right), \tag{B.2}$$

where e_i and e_j are the employment levels of plant i and j, respectively. As can be seen from (B.3), contrary to Duranton and Overman (2005) who use a multiplicative weighting scheme, we use an additive one. The additive scheme gives less weight to pairs of large plants and more weight to pairs of smaller plants than the multiplicative scheme does. Since our sample features many small plants and some very large plants – a well-known structural characteristic of the Canadian economy – this seems preferable to us. A multiplicative weighting scheme in equation (B.2) gives more weight to large establishments close to one another. Also, it assumes the equivalence between industrial and geographical concentration : n firms of 1 employee at the same place yield the same K_d value as 1 firm with n employees. [19] This could imply that our results may be too strongly driven by a few very large plants in a given industry. The downside of the additive weighting scheme is that its interpretation (in terms of distance between employees of an industry) is no longer strictly speaking correct.

The weighted K-density thus describes the distribution of bilateral distances between employees in a given industry, whereas the unweighted K-density describes the distribution of bilateral distances between plants in that industry. Since the K-density is a distribution function, we can also compute its cumulative (CDF) up to some distance d. The CDF at distance d thus tells us what share of plant pairs is located less than distance d from each other. Alternatively, we can view this as the probability that two randomly drawn plants in an industry will be at most d kilometers away.

Second step (counterfactual samples). Using the overall sample of manufacturing plants located in Canada, we randomly draw as many locations

[19] We thank a referee for bringing this point to our attention.

as there are plants in industry A. To each of these locations, we assign randomly a plant from industry A, using its observed employment. This procedure ensures that we control for the overall pattern of concentration in the manufacturing sector as a whole, as well as for the within-industry concentration. We then compute the bilateral distances of this hypothetical industry and estimate the K-density of the bilateral distances. Finally, for each industry A, we repeat this procedure 1,000 times. This yields a set of 1,000 estimated values of the K-density at each distance d.

Third step (confidence bands). To assess whether an industry is significantly localized or dispersed, we compare the actual K-density with that of the counterfactual distribution. We consider a range of distances between zero and 800 kilometers.[20] We then use our bootstrap distribution of K-densities, generated by the counterfactuals, to construct a two-sided confidence interval that contains 90 percent of these estimated values. The upper bound, $\overline{K}(d)$, of this interval is given by the 95th percentile of the generated values, and the lower bounds, $\underline{K}(d)$, by the 5th percentile of these values. Distributions of observed distances that fall into this confidence

[20] The interactions across 'neighboring cities' mostly fall into that range in Canada. In particular, a cutoff distance of 800 kilometers includes interactions within the 'western cluster' (Calgary, AB; Edmonton, AB; Saskatoon, SK; and Regina, SK); the 'plains cluster' (Winnipeg, MN; Regina, SK; Thunder Bay, ON); the 'central cluster' (Toronto, ON; Montreal, QC; Ottawa, ON; and Quebec, QC); and the 'Atlantic cluster' (Halifax, NS; Fredericton, NB; and Charlottetown, PE). Setting the cutoff distance to 800 kilometers allows us to account for industrial localization at both very small spatial scales, but also at larger interregional scales for which market-mediated input-output and demand linkages, as well as market size, might matter much more.

band could be 'as good as random' and are, therefore, not considered to be either localized or dispersed.

Fourth step (identification of location patterns). The bootstrap procedure generates a confidence band, and any deviation from that band indicates localization or dispersion of the industry. If $\widehat{K}(d) > \overline{K}(d)$ for at least one $d \in [0, 800]$, whereas it never lies below $\underline{K}(d)$ for all $d \in [0, 800]$, industry A is defined as globally localized at the 5 percent confidence level. On the other hand, if $\widehat{K}(d) < \underline{K}(d)$ for at least one $d \in [0, 800]$, industry A is defined as globally dispersed. We can also define an index of global localization, $\gamma_i(d) \equiv \max\{\widehat{K}(d) - \overline{K}(d), 0\}$, as well as an index of global dispersion

$$\psi_i(d) \equiv \begin{cases} \max\{\underline{K}(d) - \widehat{K}(d)\} & \text{if} \quad \sum_{d=0}^{800} \gamma_i(d) = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

Intuitively, if we observe a higher K-density than that of randomly drawn distributions, we consider the industry as localized. Similarly, if we observe a lower K-density than that of randomly drawn distributions, we consider the industry as dispersed. Last, the strength of localization and dispersion can be measured by $\mathit{\Gamma}_i \equiv \sum_d \gamma_i(d)$ and $\mathit{\Psi}_i \equiv \sum_d \psi_i(d)$, which corresponds roughly to a measure of the 'area' between the observed distribution and the upper- and lower-bounds of the confidence band.

C. The Ellison-Glaeser and spatially weighted Ellison-Glaeser indices

In this appendix, we briefly recall the logic underlying our discrete measures of localization. The Ellison-Glaeser index (Ellison and Glaeser, 1997), computed using employment data, is given by the following formula :[21]

$$\widehat{\gamma} = \frac{G - H_i(1 - X^{'}X)}{(1 - H_i)(1 - X'X)}, \tag{C.1}$$

where :

— H_i is a Herfindahl index measuring the industry concentration in terms of plant-level employment ;

— G = $(S - X)^{'}(S - X)$ is the raw concentration index ;

— S is a vector containing the regional shares of our measure of interest (employment) ;

— $X^{'}$ = [x_1 x_2 $\ldots$ x_J] is a vector containing the elements of the reference distribution (employment).

Given one well-known limit of the EG index – namely that it ignores the geographical positions of regions in space, the so-called 'checkerboard problem' – Guimaraes, Figueiredo, and Woodward (2011) derived from a probabilistic plant location decision model a 'spatially weighted' version of the EG index. To this end, they introduce 'neighborhood effects' in the EG index, which we henceforth refer to as EGspat when it is weighted. The matrix notation of the *spatially weighted* version of the EG is given by :

$$\widehat{\gamma}_S \equiv \frac{G_S - H_i(1 - X^{'}\Psi X)}{(1 - H_i)(1 - X'\Psi X)}, \tag{C.2}$$

[21] See Maurel and Sédillot (1999) for the definition of a very similar measure.

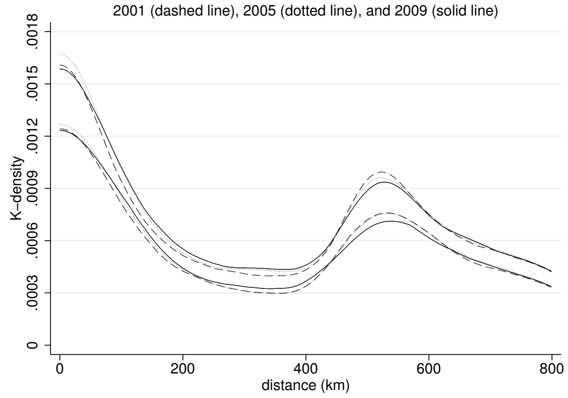
where :

- — H_i and $X^{'}$ are defined as previous.
- — $G_S = (S - X)^{'}\Psi(S - X)$ is the spatially weighted version of the raw concentration index ;
- — Ψ is a spatial weight matrix with generic element Ψ_{ij} and non-zero elements on the main diagonal. It is designed to account for spillovers that extend outside of the areal boundaries for which the EG index is computed. In general, $\Psi = \mathbf{I} + \mathbf{W}$, where $\mathbf{I}$ is the the identity matrix, and where $\mathbf{W}$ is a weight matrix for adjacent units. Adjacent units – also called contiguous units – are usually considered neighbors. In this study, we use two different matrices for Ψ, where the coefficients are either the inverse distance or the length of the common border between adjacent areal units. The latter measure has been computed using Canadian GIS data. A larger coefficient means that two adjacent units share a larger common border, so that there is greater potential that economic activity in one sector straddles the border. The latter effect increases the EGspat coefficient, which takes into account the spatial concentration across geographical units.

We can also perform a one-sided statistical test by assuming that the parameters $\widehat{\gamma}$ in the EG index and $\widehat{\gamma}_S$ in the EGspat index are asymptotically normally distributed. Following Ellison and Glaeser (1997, footnote 13), it can be shown that under the assumption of asymptotic normality of the vector $S - X$, the variance of $\widehat{\gamma}_S$ under the null hypothesis that $\gamma_S = 0$ is given by :

$$V(\widehat{\gamma}_S) = \frac{2H_i^2 tr[\Psi[\text{diag}(X) - XX^{'}]]\Psi[\text{diag}(X) - XX^{'}]]}{[(1 - H_i)(1 - X^{'}\Psi X)]^2}.$$

Figure 1.6 Distribution of distances between plants, 95% confidence bands, 5 percent sample of plants.



We test whether the EG and the EGspat indices are larger than 0, whereas Ellison and Glaeser (1997) suggest the 'rule of thumb' to check whether the indices are larger than 0.05 to assess whether or not an industry is 'strongly localized'.

D. Comparability of K-densities across years

Despite their numerous advantages, it is unclear whether and how continuous localization measures are comparable across either time or countries. The reason is that the underlying benchmark against which we want to detect localization can be very different. How this impacts on the likelihood to detect agglomeration/dispersion is theoretically and empirically unclear. Can we compare the evolutions of K-densities across time ? We believe that in our case the answer is 'yes'. To see why this is so, Figure 1.6 plots the overall distribution of bilateral distances in Canada across all industries for the years 2001, 2005, and 2009. Figure 1.6 depicts the 95 percent confidence bands for the Duranton-Overman measures of localization applied to a 5 percent random sample of all manufacturing plants (it is, unfortunately, computationally infeasible to compute the measure for all 50,000+ plants). The confidence bands in the three years overlap substantially, i.e., the observed distributions are not 'substantially' different from one another. Thus, between 2001 and 2009 – and within Canada – the reference distribution has not changed much, thus suggesting that the results are comparable across time. As can be seen from the figure, the overall spatial distribution of manufacturing in Canada has remained fairly stable between 2001 and 2009. There is no clear trend towards increasing concentration or dispersion of manufacturing in general, despite a substantial decrease in the number of manufacturing plants between 2001 and 2009.

E. Additional figures, tables, and results

This appendix provides additional tables for the DO measures for the most and the least localized industries in 2009 (Table 1.16) ; for the EG index for young plants, small plants, and exporters (Table 1.17) ; and for the frequency of localization, dispersion, or randomness of young plants, small plants, and exporters conditional on the localization pattern of all plants in the industry (Table 1.18).

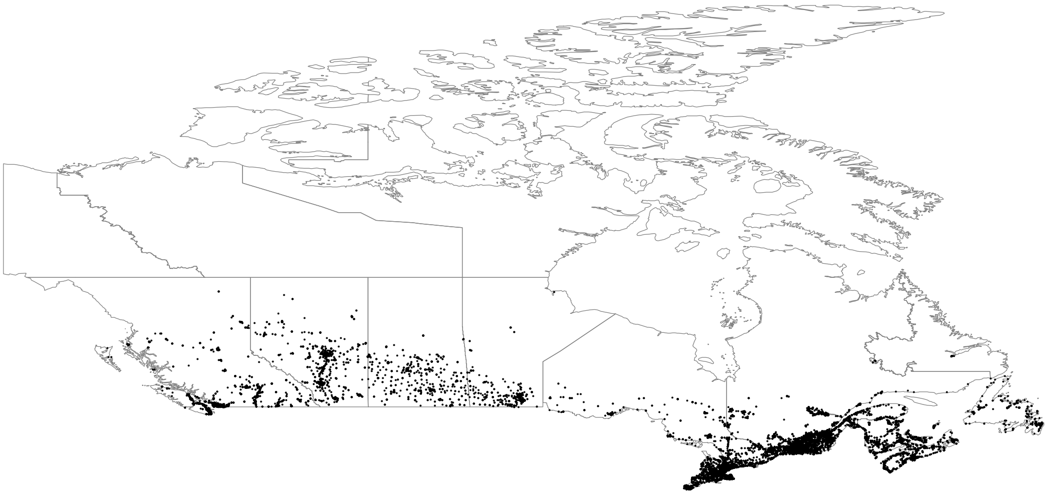


Figure 1.7 Spatial distribution of manufacturing plants in 2001 (Scott's All).

Table 1.16 Ten most and least localized industries according to the DO index in 2009.

NAICS4	Industry name		DO
	Most localized industries	# of plants	Γ_A
3335	Metalworking Machinery Manufacturing	829	0.268
3321	Forging and Stamping	161	0.171
3152	Cut and Sew Clothing Manufacturing	1096	0.153
3361	Motor Vehicle Manufacturing	169	0.145
3372	Office Furniture (including Fixtures) Manufacturing	588	0.092
3359	Other Electrical Equipment and Component Manufacturing	260	0.076
3222	Converted Paper Product Manufacturing	579	0.074
3341	Computer and Peripheral Equipment Manufacturing	154	0.059
3344	Semiconductor and Other Electronic Component Manufacturing	358	0.059
3328	Coating, Engraving, Heat Treating and Allied Activities	645	0.043
	Least localized industries	# of plants	Ψ_A
3273	Cement and Concrete Product Manufacturing	980	0.078
3331	Agricultural, Construction and Mining Machinery Manufacturing	661	0.074
3366	Ship and Boat Building	276	0.062
3219	Other Wood Product Manufacturing	1953	0.044
3121	Beverage Manufacturing	409	0.041
3212	Veneer, Plywood and Engineered Wood Product Manufacturing	284	0.037
3116	Meat Product Manufacturing	682	0.029
3362	Motor Vehicle Body and Trailer Manufacturing	324	0.022
3149	Other Textile Product Mills	764	0.020
3241	Petroleum and Coal Products Manufacturing	276	0.006
NAICS6	Industry name		DO
	Most localized industries	# of plants	Γ_A
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	37	0.607
336370	Motor Vehicle Metal Stamping	39	0.386
336110	Automobile and Light-Duty Motor Vehicle Manufacturing	115	0.382
333519	Other Metalworking Machinery Manufacturing	619	0.287
333220	Rubber and Plastics Industry Machinery Manufacturing	33	0.252
315292	Fur and Leather Clothing Manufacturing	147	0.244
332118	Stamping	140	0.220
315110	Hosiery and Sock Mills	21	0.206
335920	Communication and Energy Wire and Cable Manufacturing	39	0.199
333511	Industrial Mould Manufacturing	210	0.163
	Least localized industries	# of plants	Ψ_A
327320	Ready-Mix Concrete Manufacturing	559	0.085
321215	Structural Wood Product Manufacturing	183	0.058
336612	Boat Building	235	0.054
312110	Soft Drink and Ice Manufacturing	162	0.052
321911	Wood Window and Door Manufacturing	489	0.035
333110	Agricultural Implement Manufacturing	275	0.034
311614	Rendering and Meat Processing from Carcasses	326	0.031
314990	All Other Textile Product Mills	514	0.029
336310	Motor Vehicle Gasoline Engine and Engine Parts Manufacturing	170	0.021
312130	Wineries	140	0.015

Notes : The measures of localization and dispersion are defined as in Duranton and Overman (2005) : $\Gamma = \sum_d \Gamma(d)$, where $\Gamma(d)$ is the maximum between zero and the difference between the empirical K-density and the upper bound of the global confidence band at distance d. Analogously, $\Psi = \sum_d \Psi(d)$, where $\Psi(d)$ is the maximum between zero and the difference between the lower bound of the global confidence band and the empirical K-density at distance d, provided that the empirical K-density does not exceed the upper bound over the whole distance range. See Appendix C.

Table 1.17 Mean and median EG indices at different spatial scales.

	2001			2005			2009		
Geography	PROV	ER	CD	PROV	ER	CD	PROV	ER	CD
Unweighted EG : NAICS 6-digit industries									
Results for small plants									
Mean	0.301	-0.332	-17.652	-0.518	-5.526	-0.314	0.286	-3.101	-0.727
Median	0.295	-1.010	-1.041	0.281	-1.012	-1.051	0.281	-1.007	-1.045
Share < 0	16.602	61.776	89.922	22.780	65.251	91.120	21.154	59.615	87.692
Share $\in (0, 0.05]$	0.772	2.703	0.388	0.772	1.931	0.772	0.769	3.462	1.923
Share > 0.05	82.625	35.521	9.690	76.448	32.819	8.108	78.077	36.923	10.385
Results for young plants									
Mean	0.061	-0.833	-1.984	-3.623	-0.626	-2.449	0.254	-0.695	-0.802
Median	0.364	-1.072	-1.122	0.328	-1.052	-1.158	0.336	-1.049	-1.140
Share < 0	20.463	67.829	88.372	23.166	62.162	88.417	23.462	62.308	86.538
Share $\in (0, 0.05]$	1.158	1.938	1.550	0.772	2.317	0.386	0.000	1.923	0.385
Share > 0.05	78.378	30.233	10.078	76.062	35.521	11.197	76.538	35.769	13.077
Results for exporters									
Mean	1.335	-1.060	-0.654	0.629	-1.210	-1.319	-0.452	-0.864	-1.112
Median	0.366	-1.061	-1.117	0.358	-1.078	-1.138	0.329	-1.066	-1.127
Share < 0	16.602	65.251	88.372	18.533	66.023	91.506	21.923	63.462	87.692
Share $\in (0, 0.05]$	0.772	2.703	0.775	1.158	3.089	1.544	1.154	1.923	0.385
Share > 0.05	82.625	32.046	10.853	80.309	30.888	6.950	76.923	34.615	11.923

Notes : Mean and median values for 259 (resp., 260 in 2009) NAICS 6-digit industries. Share < 0 means 'not clustered'. Share $\in (0, 0.05]$ means 'weakly clustered'. Share > 0.05 means 'strongly clustered'. See Ellison and Glaeser (1997) for details.

Table 1.18 Location patterns of small plants, young plants, and exporters by sectoral location patterns.

			Small			Young			Exporters		
Year	Industries that are :	%	R	D	L	R	D	L	R	D	L
	Random	59.302	144	6	3	145	0	8	129	7	16
2001	Dispersed	10.078	24	2	0	23	0	3	21	0	5
	Localized	30.620	60	11	7	71	2	5	55	4	20
	Cond. prob. (R/R, D/D, L/L)	%	94.118	7.692	8.974	94.771	0.000	6.410	84.868	0.000	25.316
	Random	57.915	142	5	2	141	3	4	122	14	14
2005	Dispersed	11.969	28	3	0	31	0	0	25	2	4
	Localized	30.116	62	7	9	70	4	4	48	12	18
	Cond. prob. (R/R, D/D, L/L)	%	95.302	9.677	11.538	95.270	0.000	5.128	81.333	6.452	23.077
	Random	62.934	153	8	1	153	2	6	129	13	20
2009	Dispersed	12.355	31	0	1	30	2	0	22	3	7
	Localized	24.710	54	5	5	55	4	5	47	7	10
	Cond. prob. (R/R, D/D, L/L)	%	94.444	0.000	7.813	95.031	6.250	7.813	79.630	9.375	15.625

Notes : R, D and L stand for 'Random', 'Dispersed', and 'Localized', respectively. The line 'Cond. prob. (R/R, D/D, L/L)' denotes the conditional probability of a particular subset of firms (small, young, exporter) to be of type $y = \{D, L, R\}$ conditional of belonging to an industry that displays that same location pattern y. For example, the value 94.118% in the top panel of the table for small firms indicates that the probability of small firms in an industry to display a random pattern *conditional on being in an industry that has a random location pattern* is about 94%.

CHAPITRE II

THE WORLD IS NOT YET FLAT : TRANSPORT COSTS MATTER !

Abstract

We provide evidence for the effects of changes in transport costs, international trade exposure, and input-output linkages on the geographical concentration of Canadian manufacturing industries. Increasing transport costs, stronger import competition, and the spreading out of upstream suppliers and downstream customers are all strongly associated with declining geographical concentration of industries. The effects are large : changes in trucking rates, in import exposure, and in access to intermediate inputs explain between 20% and 60% of the observed decline in spatial concentration over the 1992–2008 period.

Keywords : transport costs ; international trade exposure ; input-output linkages ; trucking rates ; geographical concentration.

JEL classification : R12 ; C23 ; L60

2.1 Introduction

We provide evidence for the effects of changes in the costs of trading goods across space – as proxied by domestic trucking rates, international trade exposure, and customer-supplier linkages – on the geographical concentration of Canadian manufacturing industries. Using measures constructed from micro-geographic data, we find that increasing trucking rates, stronger import competition, and the spreading out of upstream suppliers and downstream customers are all strongly associated with declining geographical concentration of industries. The effects are large : holding all other variables fixed at their 1992 levels, changes in domestic trucking rates and in import exposure up to 2008 explain about 20% and 60% of the observed decline in spatial concentration, respectively. Hence, contrary to the widespread belief that the world has become 'flat' in the wake of the fall in transport, trade, and communication costs over the past two centuries, our key message is the opposite : even though the costs of trading goods across space may have hit their historical lows, changes in those costs still drive to a sizable extent changes in the economic geography of countries. [1]

[1] The fallacy of equating 'low' with 'unimportant' is reminiscent of the 'kaleidoscopic comparative advantage' debate in international trade : "[. . .] I was arguing that we now had "kaleidoscopic" comparative advantage – what we call in economic jargon, "knife-edge" specialization – so that specialization would shift among countries with small changes in cost conditions. The factors that had produced this situation were several, e.g. interest rates were less unequal across countries with integrated capital markets ; technology used by multinationals located in different countries became more available across nations ; the spread of technical education also meant that many in India and China read the same textbooks as Americans and Europeans ; and so on. So, with kaleidoscopic (or "thin" or "knife-edge") comparative advantage in many activities, we were now confronted with

The world is not yet flat : transport costs matter ! These results hold up to a variety of robustness checks and to instrumental variables estimations that deal with potential endogeneity concerns.

Assessing empirically the impact of transport costs on the spatial concentration of industries is important for several reasons. First, it is fair to say that, despite their fundamental theoretical role in spatial modeling, little is still known empirically on how transport costs drive the geographical concentration or dispersion of industries. Whereas many models tackle the questions of why and how spatial structure changes due to changes in the trading environment, much less is know empirically.[2] Second, assessing the direction of change in the geographical concentration of industries is important as there may be a tension between domestic policies that aim at growing clusters or at alleviating regional imbalances, and policies that aim at increasing international trade. Should trade be, for example, dispersive,

volatility in, not the end of, comparative advantage." (Jagdish Bhagwati, "Why the world is not flat", 2010; available at `http://www.worldaffairsjournal.org/blog/jagdish-bhagwati/why-world-not-flat`).

[2] Even theory reaches different conclusions on the effects of changes in trade costs on the spatial structure of an economy. Krugman and Livas Elizondo (1996), Helpman (1998), and Behrens, Mion, Murata, and Südekum (2013) all find that decreasing trade costs are dispersive. However, Krugman (1991), Krugman and Venables (1995), and Fujita, Krugman, and Venables (1999) reach the opposite conclusion. Using a richer spatial structure involving two countries and four regions, Behrens, Gaigné, Ottaviano, and Thisse (2007) find that increasing international trade exposure is dispersive within countries, whereas falling domestic transport costs are agglomerative. The reasons underlying these diverging results are differences in the agglomeration and dispersion forces in the models, as well as in the modeling frameworks and the spatial structure used.

pushing both domestic cluster policies and international trade agendas simultaneously may not deliver the expected results. Last, disentangling the effects of domestic shipping costs, international trade exposure, and access to both customers and suppliers on geographical concentration will also allow us to assess which components of transport costs are more likely to affect location patterns. Having an idea on this is important since all three components usually move simultaneously, thereby making assessments on the overall effects a rather complex endeavor.

Assessing empirically the impact of transport costs on the spatial concentration of industries is also a complicated task. First, we need fine measures of said spatial concentration across time to assess its changes. In this paper, we employ – for the first time to our knowledge – a long panel of continuous micro-geographic localization measures, computed from geo-coded plant-level data using the approach of Duranton and Overman (2005).[3] Using panel data allows us to go beyond existing studies that have mainly looked at the cross-sectional variation in the geographical concentration of industries. Instead, we look at the time-series variation over a nearly 20 year period to better understand what changes in covariates drive changes in the geographical concentration of industries. Dynamic analyses of agglomeration and changes therein are rare in the literature.[4] Yet, they

[3] See Holmes and Stevens (2004) for an exhaustive survey of location patterns in North America. They do, however, not report results using continuous measures. Ellison, Glaeser, and Kerr (2010) use a 'lumpy approximation' of the Duranton and Overman (2005) measure and apply it to US manufacturing data.

[4] Dumais, Ellison, and Glaeser (2002) is one exception. They analyze the impact of entry, exit, and firm growth on the geographic distribution of manufacturing employment

are required if we want to control for unobserved heterogeneity and omitted variable bias in the estimations.

Secondly, we devote substantial effort to the construction of more sophisticated measures of transport costs – proxied by domestic trucking rates, international trade exposure, and input-output linkages among firms. We build trucking rates time series from the micro-data files on truck shipments within Canada. These measures capture time-changes in domestic transport costs and are invariant to the spatial structure of industry, thereby side-stepping the often endogenous nature of standard transportation measures (e.g, transportation margins from input-output accounts). Turning to trade exposure, we investigate in detail the impacts of international trade – broken down by imports and exports and by trading partners – on industry location. Last, concerning input-output linkages, we propose a novel and much more detailed micro-geographic measure than what has been used before in the literature. Loosely speaking, we construct plant-level measures that reflect the 'minimum distance' of a plant from a dollar of inputs, or the minimum distance it has to ship a dollar of outputs. Our proxies will allow us to derive more detailed evidence on the impacts of transport costs, international trade, and input-output linkages on the spatial structure of the economy.

Finally, as the analysis is at the industry level, we also need to deal with the possible endogeneity of our main covariates. For example, it is well documented that productivity rises as an industry concentrates geographically (see, e.g., Rosenthal and Strange, 2004; Combes and Gobillon,

in the US between 1972 and 1992.

2014). If the productivity gains from agglomeration are passed on to consumers and affect also trucking rates, the causality may actually run from agglomeration to transport costs and not the other way round. Furthermore, agglomeration may lead to imbalances in shipping patterns, and the latter may increase the cost of transportation due to standard logistics problems like 'backhaul' of empty trucks (e.g., Jonkeren, Demirel, van Ommeren, and Rietveld, 2009; Behrens and Picard, 2011). Turning to trade exposure, the spatial concentration of an industry may drive export partipation (via productivity gains) or may reduce import penetration (via lower prices), thus potentially biasing the estimated coefficient. To deal with endogeneity, we require some form of instrumental variables. Since we have a large number of industries and a fairly large time dimension, our setting lends itself well to the construction of internal instruments. We implement the method suggested by Lewbel (2012), which exploits heteroscedasticity and variance-covariance restrictions to obtain identification with 2SLS when some variables are endogenous and when external instruments are either weak or not available. We also follow Ellison, Glaeser, and Kerr (2010) and use US industry price indices – for the transportation sector and for manufacturing industries – to construct external instruments for the trucking rate series.

Our paper contributes to the growing literature that investigates how the geographical structure of national economies changes as trading goods – both within and across borders – becomes cheaper. Trade influences the spatial structure of economic activity via changes in market access (e.g., Redding and Sturm, 2008; Brülhart, Carrère, and Trionfetti, 2012; Brülhart, Carrère, and Robert-Nicoud, 2014), firm entry and exit (e.g., Dumais, Ellison, and Glaser, 2002; Behrens, 2014), tougher competition in product markets (e.g., D'Costa, 2010; Holmes and Stevens, 2014), infrastructure investments (e.g., Duranton and Turner, 2012; Duranton, Morrow, and Turner, 2014), cheaper access to foreign-sourced intermediates, changes in local labor market (e.g., Autor, Dorn, and Hanson, 2013; Dauth, Findeisen, and Suedekum, 2014), or any combination of these. See Brülhart (2011) for a review of the ambiguous theoretical and empirical effects of increased trade openness on the internal geography of countries.

The remainder of the paper is structured as follows. Section 2.2 briefly documents the evolutions of the geographical concentration of Canadian manufacturing industries. Section 2.3 describes our empirical strategy, constructs our key variables, and discusses the various identification issues we face. Section 3.4 presents our key results on the impacts of trade costs and measures related to customer and supplier access on the geographical concentration of Canadian manufacturing industries. We provide a large number of robustness checks and instrumental variables estimates. Section 3.6 concludes. Technical details are relegated to the appendix.

2.2 Trends in industrial localization from 1990 to 2009

As a prelude to the econometric analysis to follow, we first briefly describe the data and the measures of geographical concentration we use in this paper. We then provide a quick overview of the broad trends in the localization of Canadian manufacturing industries from 1990 to 2009.

2.2.1 Measuring localization

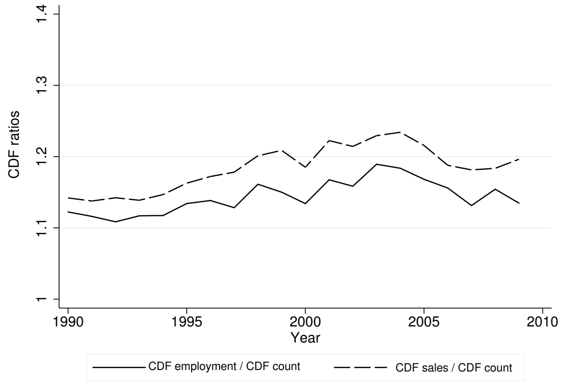
Our analysis is based on Statistics Canada's Annual Survey of Manufacturers (ASM) Longitudinal Microdata file from 1990 to 2009. This file contains between 32,000 and 53,000 plants per year, covering 257 NAICS 6-digit manufacturing industries. For every plant, we have information about : its primary NAICS industry ; its employment ; its sales ; and its 6-digit postal code. The latter allows us to effectively geo-locate the plants using latitude and longitude coordinates of postal code centroids. A detailed description of the data is relegated to Appendix A.

We exploit the micro-geographic nature of our data and measure the geographical concentration of industries using the Duranton and Overman (2005, 2008 ; henceforth, DO) K-densities (see Appendix B for technical details). The DO K-densities look at how close plants are relative to each other by considering the kernel-smoothed distribution of bilateral distances between them. We explain in Section 2.3.2 why we use a kernel-smoothed distribution of bilateral distances and not on the raw distribution. The DO K-densities provide a very detailed micro-geographic description of location patterns, and allow for statistical testing of whether those patterns may be due to chance or not. We estimate the K-densities year-by-year for all industries at the NAICS 6-digit level. For each pair of plants, we compute the bilateral great circle distance between them using their geographical coordinates. Since the K-density is a distribution function, we can also compute its cumulative (CDF) up to some distance d. The CDF of the K-density at distance d tells us what share of plant pairs in an industry is located less than distance d from each other. Since we are not interested in identifying at which specific distances localization of firms occurs, the CDF of the K-

density provides a better measure of the 'overall degree' of geographical concentration.

Table 2.1 summarizes the K-density CDF for the most localized industries in 1990, 1999, and 2009, respectively. To understand how to read that table, take 'Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing' (NAICS 315231) as an example. In 1990, 62 percent of the distances between plants in that industry are less than 50 kilometers. Put differently, if we draw two plants in that industry at random, the probability that these plants are less than 50 kilometers apart is 0.62. If we, however, draw two plants at random among *all* manufacturing plants, that same probability would only be about 0.08 (see Table 2.2 below). Clearly, this large difference suggests that the location patterns of plants in the 'Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing' industry are very different from those of manufacturing in general. Plants in that industry are much closer than they 'should be' if they were distributed like overall manufacturing.

Figure 2.1 Year-on-year changes in the CDF ratios at 50 kilometers.



Whereas the standard K-densities are computed based on plant counts,

Table 2.1 Ten most localized NAICS 6-digit industries (based on plant counts).

NAICS	Industry descripition	CDF
	1990	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.62
315233	Women's and Girls' Cut and Sew Dress Manufacturing	0.55
313240	Knit Fabric Mills	0.53
315292	Fur and Leather Clothing Manufacturing	0.42
315291	Infants' Cut and Sew Clothing Manufacturing	0.32
315210	Cut and Sew Clothing Contracting	0.30
337214	Office Furniture (except Wood) Manufacturing	0.21
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.21
313110	Fibre, Yarn and Thread Mills	0.19
333511	Industrial Mould Manufacturing	0.18
	1999	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.63
313240	Knit Fabric Mills	0.47
315210	Cut and Sew Clothing Contracting	0.22
333220	Rubber and Plastics Industry Machinery Manufacturing	0.20
336370	Motor Vehicle Metal Stamping	0.18
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.18
336330	Motor Vehicle Steering and Suspension Components (except Spring) Manufacturing	0.17
333519	Other Metalworking Machinery Manufacturing	0.16
337214	Office Furniture (except Wood) Manufacturing	0.15
315291	Infants' Cut and Sew Clothing Manufacturing	0.14
	2009	
315231	Women's and Girls' Cut and Sew Lingerie, Loungewear and Nightwear Manufacturing	0.61
322299	All Other Converted Paper Product Manufacturing	0.29
337214	Office Furniture (except Wood) Manufacturing	0.17
336370	Motor Vehicle Metal Stamping	0.17
332720	Turned Product and Screw, Nut and Bolt Manufacturing	0.16
337215	Showcase, Partition, Shelving and Locker Manufacturing	0.15
321112	Shingle and Shake Mills	0.14
331420	Copper Rolling, Drawing, Extruding and Alloying	0.13
336360	Motor Vehicle Seating and Interior Trim Manufacturing	0.13
315110	Hosiery and Sock Mills	0.13

Notes : The CDF at distance d is the cumulative sum of the K-densities up to distance d. Results in this table are reported for a distance $d = 50$ kilometers.

i.e., distances between pairs of plants without any weighting scheme, we can also compute weighted versions (see Duranton and Overman, 2005). In particular, we can weight pairs of plants by either plant-level employment or plant-level sales. For these weighted versions, the foregoing interpretations remain true, except that the unit of observation is now the employee or a dollar of sales. We generally report results for the weighted measures only as robustness checks, since the qualitative patterns are similar to the ones obtained from using the unweighted measures. However, comparing the unweighted to the employment- or sales-weighted K-densities reveals some interesting patterns. As can be seen from Figure 2.1, industries are on average always more concentrated in terms of employment than in terms of plant counts, and even more concentrated in terms of sales than in terms of employment. This is a manifestation of agglomeration economies, and it is consistent with the findings of Holmes and Stevens (2002, 2014) and others that more localized plants tend to be larger and more productive than less localized plants. Note that the ratios are increasing until about 2004, and slightly decreasing afterwards. In 2009, within 50 kilometer distance, the concentration of employment exceeds that of plant counts by about 13%, whereas the concentration of sales exceeds that of plant counts by about 20%.

2.2.2 Decreasing localization

There is evidence that the geographical concentration of manufacturing industries has decreased over the first decade of the years 2000 in Canada (see Behrens and Bougna, 2013 ; Behrens, 2014). This de-concentration trend can clearly be seen in our data from Table 2.2. There has been a nearly monotonic decline in the mean value of the CDF across industries between 1990 and 2009. For example, the average CDF at 50 kilometers distance was 0.076 in 1990, 0.062 in 1999, and 0.056 in 2009, a 27.1% decrease over a twenty year period. Whereas concentration has decreased at all distances, the greatest declines, however, were at shorter distances : plants are dispersing, but less so at longer distances. [5] This finding suggests that the incentives for plants to locate in very close proximity to each other are lessening over time. It also likely reflects the fact that manufacturing industries have been 'bid out' of cities because of higher land and labor costs there, and that they are moving to smaller nearby urban, sub-urban, or rural areas as a consequence (see, e.g., Henderson, 1997). Still, the fact that the CDF continues to fall at 500 km suggests a broader geographic dispersion of manufacturing activity, which is likely driven by the rising manufacturing output in western Canada and the associated fundamental shifts in manufacturing location away from the 'traditional corridor' that runs through Quebec and Ontario.

Observe that the de-concentration trend also affects the employment-weighted and the sales-weighted measures of localization (see Table 2.2).

[5] Whereas the CDF of the K-density is easily interpretable and provides a natural measure to track the changing concentration of industries, it cannot tell us anything about whether or not industries are statistically significantly concentrated or not. Table 2.9 in Appendix E summarizes location patterns by year, based on their statistical significance (see Duranton and Overman, 2005, and Appendix B for more information). As can be seen from Table 2.9, the share of statistically significantly localized industries has been decreasing over our study period, thus mimicking the downward trend in the K-density CDFs. In a nutshell, there is a clear trend towards less localization, and that trend is captured by both the CDF and the statistical tests for localization.

Yet, as can be seen from Figure 2.2, although industries have in general become more geographically dispersed according to all three measures, the size of plant pairs in close proximity has tended to increase in relative terms regardless of whether size is measured by employment or by sales. Put differently, the process of dispersion is less pronounced when measured by either employment or sales, thus suggesting that smaller plants drive a substantial part of the dispersion process, either through entry and exit or through relocation.

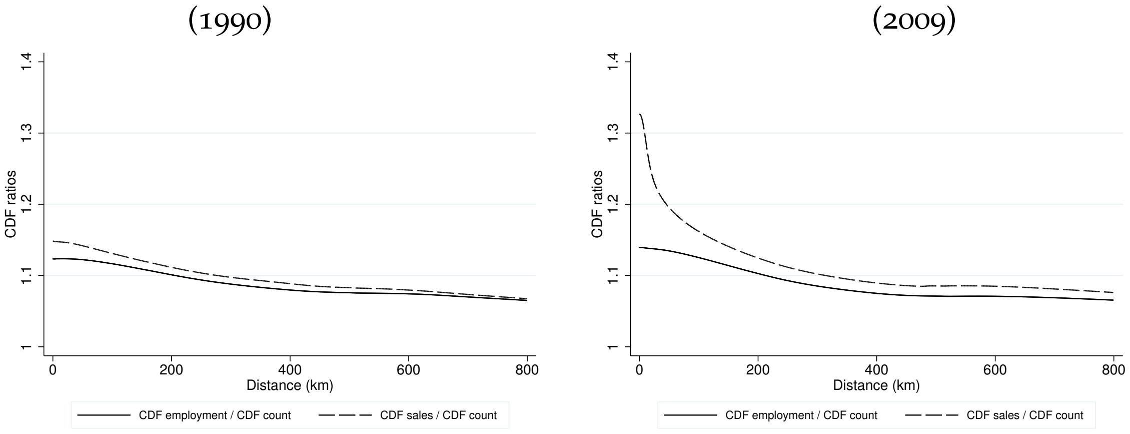
Table 2.2 Mean of the Duranton-Overman CDFs across industries, 1990 to 2009.

	Unweighted				Employment weighted				Sales weighted			
	CDF at a distance of											
Year	10 km	50 km	100 km	500 km	10 km	50 km	100 km	500 km	10 km	50 km	100 km	500 km
1990	0.020	0.076	0.139	0.420	0.021	0.083	0.151	0.449	0.022	0.086	0.156	0.453
1991	0.019	0.076	0.139	0.423	0.022	0.083	0.152	0.447	0.023	0.087	0.156	0.453
1992	0.020	0.074	0.135	0.418	0.020	0.079	0.147	0.442	0.022	0.084	0.151	0.448
1993	0.019	0.072	0.132	0.416	0.020	0.079	0.145	0.440	0.021	0.082	0.148	0.446
1994	0.017	0.071	0.131	0.413	0.020	0.077	0.143	0.438	0.021	0.081	0.147	0.443
1995	0.017	0.068	0.126	0.402	0.019	0.076	0.141	0.432	0.020	0.080	0.145	0.438
1996	0.016	0.065	0.122	0.402	0.019	0.073	0.136	0.428	0.020	0.076	0.140	0.435
1997	0.016	0.066	0.123	0.401	0.017	0.072	0.135	0.427	0.019	0.077	0.140	0.433
1998	0.016	0.064	0.120	0.396	0.019	0.074	0.135	0.425	0.019	0.078	0.141	0.433
1999	0.015	0.062	0.118	0.398	0.017	0.072	0.134	0.426	0.018	0.076	0.139	0.434
2000	0.014	0.063	0.120	0.383	0.016	0.073	0.135	0.411	0.016	0.075	0.140	0.421
2001	0.013	0.061	0.118	0.383	0.015	0.072	0.136	0.412	0.016	0.076	0.142	0.421
2002	0.013	0.062	0.119	0.383	0.016	0.073	0.137	0.413	0.017	0.078	0.143	0.422
2003	0.013	0.060	0.117	0.384	0.015	0.072	0.137	0.416	0.016	0.075	0.141	0.422
2004	0.013	0.060	0.115	0.379	0.015	0.070	0.132	0.412	0.017	0.074	0.137	0.418
2005	0.012	0.059	0.113	0.379	0.014	0.068	0.130	0.409	0.016	0.072	0.134	0.415
2006	0.013	0.061	0.116	0.378	0.015	0.069	0.131	0.406	0.015	0.072	0.135	0.412
2007	0.012	0.057	0.110	0.374	0.015	0.064	0.122	0.399	0.017	0.069	0.127	0.406
2008	0.012	0.057	0.110	0.376	0.017	0.067	0.125	0.400	0.017	0.069	0.128	0.405
2009	0.013	0.056	0.107	0.373	0.015	0.063	0.121	0.397	0.017	0.068	0.126	0.403
Mean	0.015	0.064	0.121	0.394	0.017	0.073	0.136	0.422	0.019	0.077	0.141	0.428
Change	-36.0%	-27.1%	-22.6%	-11.3%	-28.7%	-23.3%	-20.3%	-11.4%	-21.5%	-21.2%	-19.3%	-11.0%

Notes : Authors' computations based on the Annual Survey of Manufacturers Longitudinal Microdata file, 1990–2009. The means of the CDF are based on 257 industries and are not weighted (but the CDFs for each industry are weighted by either employment in the middle columns, or by sales in the right columns ; see Appendix B). 'Mean' refers to the mean of the K-densities over the 1990–2009 period. 'Change' is the percentage change between 1990 and 2009.

To conclude, the descriptive evidence points to a significant decrease in the geographical concentration of manufacturing industries in Canada over the last 20 years, no matter whether concentration is measured in terms of plant counts, employment, or sales. The pace of decline, however, differs across industries in systematic ways. Understanding which factors drive that decrease to what extent and for which industries, with a special focus on transportation costs, trade, and input-output linkages between plants, is the objective of the remainder of this paper.

Figure 2.2 Ratios of mean employment- and sales-based CDFs to count-based CDF by distance.



2.3 Empirical methodology

While the patterns highlighted in Section 2.2 show that there are clear trends in changes in the geographical concentration of industries, they do not allow us to isolate the factors that drive those changes. We therefore now turn to multivariate analysis to identify the sources of those changes and to measure their relative contribution. We first briefly spell out our empirical specification. We then explain the construction of our main variables and discuss the different identification problems.

2.3.1 Econometric specification

We work at the industry-year level and take advantage of the panel nature of our data. More precisely, we estimate the following baseline model :

$$\gamma_{m,t}(d) = \mathbf{T}_{m,t}\beta_T + \mathbf{C}_{m,t}\beta_C + \alpha_t + \mu_m + \varepsilon_{m,t} \tag{E.1}$$

where $\gamma_{m,t}(d)$ is the K-density CDF for industry m in year t at distance d; where $\mathbf{T}_{m,t}$ is a vector of 'trade cost' correlates that constitute our main variables of interest; where $\mathbf{C}_{m,t}$ is a vector of time-varying industry controls; where α_t and μ_m are time and industry fixed effects, respectively; and where $\varepsilon_{m,t}$ is the error term. The latter is assumed to be independently and identically distributed with the usual properties for consistency of OLS.

One may be worried by the fact that identification in (E.1) comes from the *within* variation in the data. The latter may be small given yearly data, especially for the spatial variables. This point has been raised in other studies (e.g., Ellison, Glaeser, and Kerr, 2010, p.1200), but those studies usually use more aggregated measures of agglomeration like the Ellison and Glaeser (1997) index or similar discrete indices. Those measures change much more slowly over time than the K-densities, especially at short distances. The reason is that the micro-geographic measures are constructed from geo-coded data, and that there is a lot of churning at short distances that is not picked up by spatially more aggregated measures. This churning creates a tension. One the one hand, there is *substantial year-on-year variation*, which allows for identification using this within variation. On the other hand, there is also *a lot of noise* at a small geographical scale, which makes the estimates imprecise. As we argue in Section 2.3.2 below, the K-density CDF measures provide the right tools to balance these two conflicting points.

Table 2.3 Key variables and summary statistics.

	Industry	Mean	Standard deviation		
Variable names and descriptions	detail		Overall	Between	Within
$\mathbf{T}_{m,t}$: *Trade, transportation, and input-output variables*					
Share of industry imports from Asian countries (excluding OECD members)	NAICS6	0.12	0.23	0.17	0.06
Share of import s from OECD member countries (excluding U.S. and Mexico)	NAICS6	0.16	0.18	0.13	0.05
Share of impors from NAFTA countries (U.S. and Mexico)	NAICS6	0.66	0.33	0.26	0.07
Share of industry exports from Asian countries (excluding OECD members)	NAICS6	0.03	0.08	0.05	0.03
Share of export from OECD member countries (excluding U.S. and Mexico)	NAICS6	0.09	0.13	0.08	0.05
Share of exports from NAFTA countries (U.S. and Mexico)	NAICS6	0.83	0.26	0.19	0.07
Ad valorem trucking costs for an avg. load shipped 500km as a share of goods shipped	L-level	0.034	0.035	0.030	0.005
Industry mean of the avg. distance to a dollar of inputs from the 5 nearest plants (km)	NAICS6	242.99	152.33	95.94	56.39
Industry mean of the avg. distance to ship a dollar of output to the 5 nearest plants (km)	NAICS6	244.86	171.87	104.36	67.51
Minimum average distance to 5 × 257 closest plants	NAICS6	64.54	56.63	42.44	14.19
$\mathbf{C}_{m,t}$: *Industry-year control variables*					
Share of input from natural resource-based industries	L-level	0.11	0.2	0.17	0.03
Sectoral energy inputs as a share of total sector output L-level	L-level	0.03	0.057	0.044	0.013
Total industry employment	NAICS6	6938	9749.88	7744.11	2005.76
Herfindahl index of enterprise-level employment concentration	NAICS6	0.1	0.126	0.092	0.034
Mean plant size	NAICS6	74	181	139	42
Share of plants controlled by multi-plant firms	NAICS6	0.21	0.248	0.183	0.065
Share of foreign controlled plants	NAICS6	0.15	0.2	0.14	0.06
Share of hours worked by all workers with post-secondary education	NAICS6	0.4	0.115	0.07	0.045
Intramural research and development expenditures as a share of industry sales	L-level	0.0111	0.039	0.027	0.012

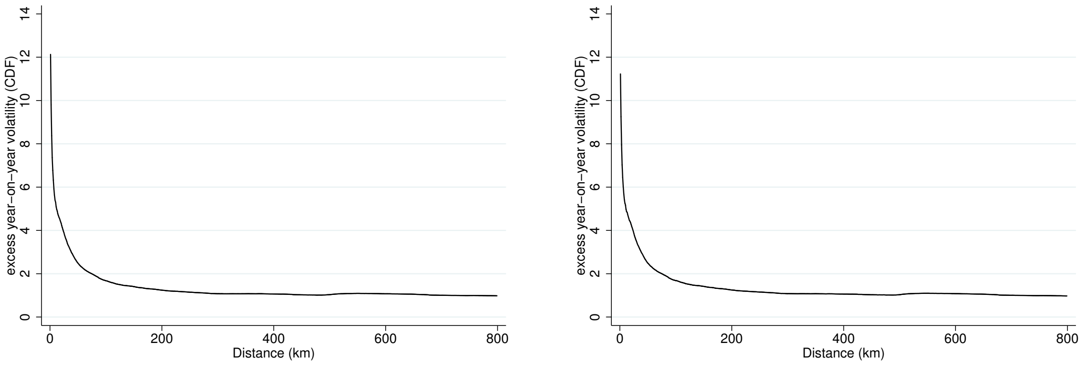
Notes : All descriptive statistics are based on the sample we use in the regression analysis, which includes 4,369 observations covering 257 industries and 17 years. The standard deviation is decomposed into between and within components, which measure the cross sectional and the time series variation, respectively. Some industry-level data are available at the L-level only, which is the finest level of data for public release in Canada (between the NAICS 3- and 4-digit levels of aggregation). Additional information regarding our data sources and the construction of our key variables is provided in Appendix A and in Section 2.3.2.

Table 2.3 summarizes our main variables, provides descriptive statistics, and reports the within and between components of the variance. As can be seen, there is substantial time variation in our data, although the bulk of the variation remains cross-sectional, as expected.

2.3.2 Construction of the key variables

We now describe in detail the construction of our key variables : (i) our K-density geographical concentration measures ; (ii) our industry measures of transportation costs ; (iii) our micro-geographic input-output linkages ; and (iv) our measures of industries' international trade exposure. We also discuss a number of methodological issues related to their construction.

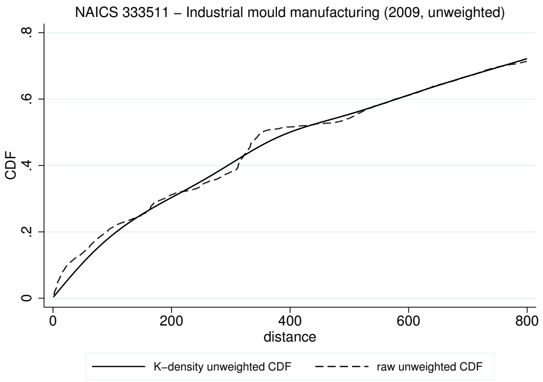
Figure 2.3 'Excess volatility' of the raw CDFs, linear trend (left) and autoregressive (right).



K-density CDFS

The technical details concerning the construction of the K-density CDFS are given in Section 2.2.1 and in Appendix B. Here, we discuss a number of issues linked to the time variability and the smoothing that we mentioned above. Starting with the former point, Figure 2.3 depicts the year-on-year 'excess volatility' at each distance d between 1 kilometer and 800 kilometers. The excess volatility is defined as the ratio of the year-on-year volatility of the raw distribution and that of the kernel-smoothed distribution.[6] As can be seen from Figure 2.3, the raw distribution is always more volatile than the smoothed distribution, and *especially so at short distances*. Whereas for distances greater than about 200 kilometers the volatility of the raw and the smoothed CDFS are roughly identical, the raw distribution is up to 11 or 12 times more volatile at short distances. In other words, due to substantial churning at the plant level, the micro-geographic measures contain a lot of noise in the time-series at short distances, though it is at those distances that the effects of transport costs and trade that we intend to identify are most likely to operate. Thus, smoothing is important to reduce the noise in the time series.[7]

Figure 2.4 Example of raw vs kernel-smoothed CDFS for plant counts.



Smoothing has, however, the drawback to alter the raw distribution. Figure 2.4 depicts the 'raw' (unsmoothed) CDF of the bilateral distances as

[6] See Appendix B for the formal definition of the 'raw' distribution. We use standard measures of volatility based on the year-on-year variance, the fitting of a linear trend, or an autoregressive AR(1) model.

[7] We ran our analysis using the raw CDFS as dependent variables, but the results for short distances become very imprecise. Most coefficients are not statistically significant due to their large standard errors.

a dashed line, and the K-density CDF (smoothed) as a solid line for a representative industry – 'Industrial mould manufacturing'. Two comments are in order. First, as can be seen, the smoothed CDFS are less volatile and more regular than the unsmoothed CDFS, though the two become very similar at longer distances starting at about 200 kilometers. As can also be seen from Figure 2.4, the smoothed CDFS tend to underestimate the degree of geographical concentration at short distances. This point has been recently made by Murata, Nakajima, and Tamura (2014), who show that there is a downward bias in the Duranton-Overman K-density estimates at short distances due to 'reflection' and the use of a differentiable kernel function.

To summarize, there are costs and benefits of using the smoothed CDFS compared to the unsmoothed CDFS. The benefit is that the smoothed densities exhibit substantially less year-on-year variability at short distances, thus reducing the noise due to plant-level churning that shows up in the data and that affects the micro-greographic concentration measures. The cost is that the smoothed densities underestimate the degree of geographical concentration at short distances, thus potentially biasing the estimated coefficients on the trade cost covariates towards zero. Since identification stems from the time-series variation in our approach, we believe that the benefits of using the smoothed CDFS outweight the costs. [8]

[8] In a cross-sectional analysis, we would rather use the raw CDFS since there is no need to smooth out any time-series volatility. However, Duranton and Overman (2005) argue that even in a cross section smoothing may be required to cope with unobserved variation in, e.g., the density of the road network.

Transportation costs

Transportation costs loom large in the theoretical literature on industry location and geographical concentration. Industries with high transportation costs – either for their inputs, for their outputs, or for both – should agglomerate production in locations close to their suppliers or customers to minimize those costs. Despite their dominant theoretical role, it is fair to say that limited work has gone thus far into the elaboration of good measures of transportation costs, and even less into their application to the analysis of changes in agglomeration. Rosenthal and Strange (2001), for example, use the ratio of inventories to sales at the end of the year as a proxy for 'perishability of output', itself a proxy for transportation costs. Lu and Tao (2009) use a similar proxy, namely the finished goods to output ratio, where finished goods are inventories not yet sold. Ellison, Glaeser, and Kerr (2010) do not even talk about the possible role of transportation costs in their analysis, the reason being that these costs are assumed to have become 'negligible'. While this may be the case in a cross-section of industries – with transport costs on average around 3–4% of the value of the shipment according to our estimates – our results show that their time-series variation is a major driver of the changes in the location patterns of industries. In other words, transport costs matter!

Our work aims to improve our understanding of how changes in transportation costs influence changes in the geographical concentration of industries. To this end, we use *direct measures* of transportation costs constructed from detailed micro-data files on shipments within Canada. To estimate ad valorem rates, we first use a pricing model to predicted trucking firm revenues for a 500 kilometers trip by commodity for the average tonnage using shipment (waybill) data from Statistics Canada's Trucking Commodity Origin-Destination Survey (see Brown and Anderson, 2015, for details). We estimate the 'prices' charged by trucking firms as a function of distance shipped, tonnage, and a set of commodity and firm fixed effects.[9] The prices are then converted into ad valorem trucking costs by estimating the value of each shipment. This value is derived by multiplying the tonnage of the average shipment on a commodity basis by their respective value per tonne derived from an 'experiment export trade file' produced only in 2008. The ad valorem estimates at the commodity level in 2008, in turn, are used to estimate ad valorem rates $\tau_{m,2008}$ for L-level industries in 2008 using a set of industry-commodity concordances. Yearly trucking industry price indices $p_{\text{trans},t}$ and manufacturing industry price indices $p_{m,t}$ from Statistics Canada's KLEMS database are then used to project the ad valorem rates backwards and forwards in time, thereby creating an industry-specific ad valorem transportation rate time series $\tau_{m,t}$:

$$\tau_{m,t} = \frac{p_{\text{trans},t}}{p_{m,t}} \tau_{m,2008}. \tag{E.2}$$

Although our measures of transport costs are much more direct and detailed than those used before in the agglomeration literature, they are by construction unlikely to be fully exogenous to industrial location patterns since they depend on price indices. We come back to this point in Section 2.3.3 below when we discuss the different identification issues. Note, however, that we estimate transportation costs for a 'representative shipment' by truck, holding distance fixed at 500 kilometers. Hence, variable

[9] While we do not directly control for the time costs of transportation they will be, at least partially, embedded in the transportation prices (which would capture quality of service for time-dependent trips).

shipping distances that result from optimal location choices of plants in an industry have a priori no direct influence on our measures.

Figure 2.5 Changes in average transportation costs, 1990–2009.

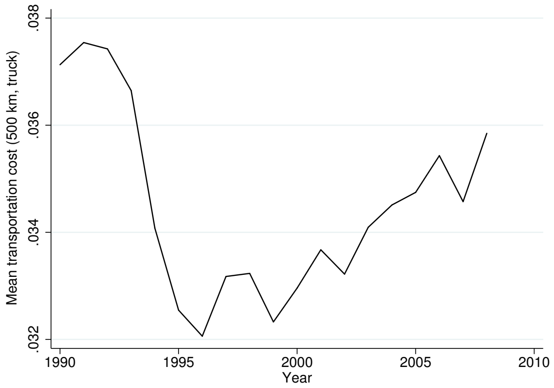


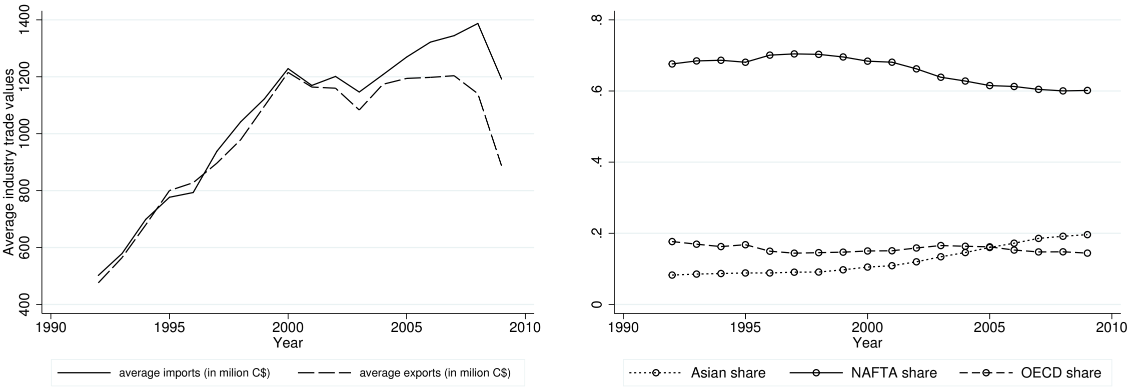
Figure 2.5 depicts the year-on-year changes in the (unweighted) cross-industry average transportation costs for a 500 kilometers shipment. As can be seen, transport costs are first decreasing – due, essentially, to reductions in labor costs at constant fuel prices – and then increasing – due, essentially, to increasing fuel prices at constant labor costs. They range from about 3.8% of the value of the shipment in the early nineties, to about 3.2% in the mid-nineties. Since industries tend to localize when their shipping costs are either high (market access) or low (to exploit other sources of agglomeration economies), we expect transportation costs to have a non-linear and negative effect on the degree of industrial agglomeration, especially for industries characterized by intermediate values of transport costs. Since there is significant time- and cross-industry variation in transportation costs in our data (see Table 2.3), we will be able to estimate precisely the effect of transportation costs on the geographical patterns of industries.

International trade exposure

While transportation costs capture the 'domestic' part of trade in our model, we also control finely for the role of international trade in the location of industries. It is indeed well known theoretically – though less so empirically – that trade influences the spatial structure of economic activity via firm entry and exit, tougher competition in product markets, cheaper access to foreign-sourced intermediates, and changes in local labor markets (e.g., D'Costa, 2010; Brülhart, Carrère, and Trionfetti, 2012; Autor, Dorn, and Hanson, 2013; Behrens, 2014; Brülhart, Carrère, and Robert-Nicoud, 2014; Holmes and Stevens, 2014). We use detailed yearly data on imports and exports by industry and country of origin and destination to control for industries' import and export exposure (the ratio of industry imports or exports to industry sales). To disentangle the different effects that depend on whether trade is in intermediates or final goods (on which we have unfortunately no information in our data), and on whether trade is 'North-North' or 'North-South', we break these measures down by countries of origin : low-cost Asian countries; OECD countries; and NAFTA countries.

The left panel of Figure 2.6 depicts the changes in the average import and export values by industry over our study period. The right panel provides a snapshot of how import and export shares change across broad groups of trading partners. As one can see, the importance of international trade has dramatically increased – at least up to the trade collapse starting 2008 – and there has been a progressively increasing re-orientation of trade towards Asian countries (especially for imports).

Figure 2.6 Changes in import- and export trade values (left), and import shares (right).



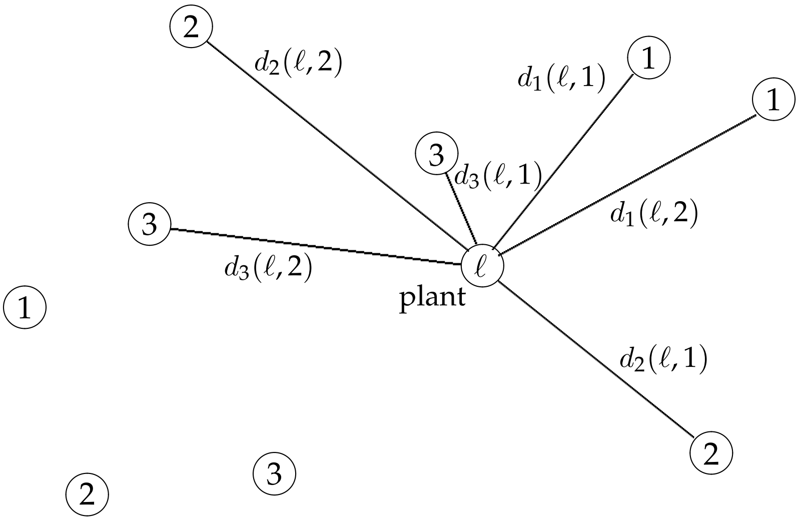
Input-output linkages

Another important trade-related source of agglomeration are input and output linkages. Many studies find that customer-supplier relationships is the most important mechanism to explain the co-location of industries, which is suggestive of their importance for geographical concentration.[10] Despite their importance, the empirical treatment of input-output linkages has been rather limited until now. Rosenthal and Strange (2001) use manufacturing and non-manufacturing inputs purchased by the industry per dollar of output. Lu and Tao (2009) use the export-intensity of a

[10] Holmes (1999) documents that plants in US manufacturing industries that are geographically more concentrated are more vertically disintegrated. Their purchased inputs as a percent of the value of outputs is higher in areas where the industry concentrates, thus suggesting that input-output linkages may drive industry localization. Note, however, that he cannot rule out reverse causality : plants in industries that concentrate geographically for some unobserved reason may vertically disintegrate more because of that concentration.

sector as a proxy for input sharing.[11] Another approach to modelling input sharing – the most widely adopted in the literature – is to use input-output accounts to measure the extent that industries buy and sell from one another (e.g., Duranton and Overman, 2005, 2008; Ellison, Glaeser, and Kerr, 2010). The drawbacks of all these approaches is that the input-output measure is potentially endogenous, and that it does not take into account any geographical information.

Figure 2.7 Constructing input-output distances and 'minimum distance' measures.



Our measures of input and output linkages are very different and make use of the micro-geographic nature of our data. Consider a plant ℓ active in sector $\Omega(\ell)$. Let Ω denote the set of sectors and Ω_s the set of plants

[11] The rationale for this proxy is that, when compared to other industries, export industries strongly rely on inputs and information sharing like the information on procedures and international markets where they sell their products. This measure thus cannot disentangle information externalities from input sharing.

in sector s. Let $k_s(i,\ell)$ denote the ith closest sector-s plant to plant ℓ. Our micro-geographic measures of input- and output linkages are constructed as weighted averages as follows :

$$\mathcal{I}\text{dist}(\ell) = \sum_{s \in \Omega \setminus \Omega(\ell)} \omega^{\text{in}}_{\Omega(\ell),s} \times \frac{1}{N} \sum_{i=1}^{N} d(\ell, k_s(i,\ell)), \tag{E.3}$$

for inputs, and

$$\mathcal{O}\text{dist}(\ell) = \sum_{s \in \Omega \setminus \Omega(\ell)} \omega^{\text{out}}_{\Omega(\ell),s} \times \frac{1}{N} \sum_{i=1}^{N} d(\ell, k_s(i,\ell)), \tag{E.4}$$

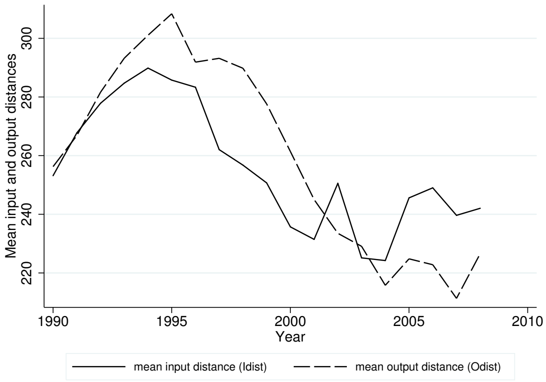
for outputs, where $d(\cdot,\cdot)$ is the great circle distance between the plants' postal code centroids, and where $\omega^{\text{in}}_{\Omega(\ell),s}$ and $\omega^{\text{out}}_{\Omega(\ell),s}$ are sectoral input- and output shares.[12] Figure 2.7 illustrates the construction for the case where $N = 2$ and with three industries.

Since by construction $\sum_s \omega^{\text{in}}_{\Omega(\ell),s} = \sum_s \omega^{\text{out}}_{\Omega(\ell),s} = 1$, we can interpret $\mathcal{I}\text{dist}(\ell)$ as the minimum average distance of plant ℓ to a dollar of inputs from its N closest suppliers. Analogously, $\mathcal{O}\text{dist}(\ell)$ is the minimum average distance plant ℓ has to ship a dollar of outputs to its N closest (industrial) customers.[13] The larger are $\mathcal{I}\text{dist}(\ell)$ or $\mathcal{O}\text{dist}(\ell)$, the worse are plant ℓ's input or output linkages – it is, on average, further away from a dollar of intermediate inputs or a dollar of demand emanating from the other industries.

[12] Appendix C provides additional details on the input and output shares.

[13] Unfortunately, we have no micro-geographic information on final demand and thus cannot include it in our output linkage measures. Using a population-weighted market potential measure as a proxy is infeasible because of the very strong persistence in time. However, our industry fixed effects are likely to control for slow-changing final demand due to changes in the population distribution.

Figure 2.8 Changes in average input-output distances, 1990–2009.



Note that our input and output linkages make use of plant-level location information, but only of *national* input and output shares. The latter is due to the fact that we do not directly observe input-output linkages at the plant level. Yet, given this, our procedure has the advantage to sidestep problems of endogeneity of those measures. Note also that our input-output measures are computed across all industries except the one the plant belongs to. Thus, our measures capture finely the whole cross-industry location patterns, but do not pick up industrial localization of the sector itself since it is excluded from the computation. This is important to not confound input-output linkages with other drivers of geographical concentration.

We compute the measures (E.3) and (E.4) for all years and for all plants, using the $N = 3, 5, 7, 10$ nearest plants in each industry. We then average them across plants in each industry and each year to get an industry-year specific measure of both input and output distances :

$$\mathcal{O}\text{dist}_s = \frac{1}{|\Omega_s|} \sum_{\ell=1}^{|\Omega_s|} \mathcal{O}\text{dist}(\ell) \quad \text{and} \quad \mathcal{I}\text{dist}_s = \frac{1}{|\Omega_s|} \sum_{\ell=1}^{|\Omega_s|} \mathcal{I}\text{dist}(\ell), \qquad \text{(E.5)}$$

where $|\Omega_s|$ denotes the number of plants in industry s. As expected, these measures are strongly correlated. Yet, despite that correlation we can include them simultaneously into our regressions and still identify their effect on industrial localization.

Figure 2.8 depicts the time-series changes in the (unweighted) average input and output measure across all industries. As one can see, in 2000 for example, plants were on average located about 235 kilometers from a dollar of inputs, and had to ship a dollar of their output on average over a distance of 260 kilometers. [14]

One potential problem with the measures (E.3) and (E.4) is that they tend to be mechanically smaller in denser areas. To control for this fact, we also compute a 'minimum distance measure', i.e., the distance of plant ℓ from the $M = N \times 257$ closest plants regardless of their industry. Including that measure into our regressions then controls for the overall plant density in a location, which implies that our input-output linkage measures pick up the effect of being closer to a dollar of inputs or outputs conditional on the overall density of the area the plant is located in. Formally, we compute for each plant ℓ the following measure :

$$\mathcal{M}\text{dist}(\ell) = \frac{1}{M} \sum_{i=1}^{M} d(\ell, k_{\setminus \Omega(\ell)}(i, \ell)), \tag{E.6}$$

where $d(\ell, k_{\setminus \Omega(\ell)}(i, \ell))$ denotes the distance to the ith closest plant in any

[14] Time-series changes in the input- and output-distance measures may reflect three things : (i) entry or exit of potential suppliers ; (ii) changes in the geographical location of input suppliers and/or clients ; and (iii) changes in the input-output coefficients, i.e., the technological relationships. We cannot dissociate the sources (i) and (ii) in our analysis, but entry and exit are vastly more important than relocation when looking at plant-level data.

industry but $\Omega(\ell)$. We then average this measure across all plants in the same industry as before.

Industry-level controls

The literature on industrial localization has identified many important sources of externalities that cause the spatial concentration of industries and changes therein (see Duranton and Puga, 2004, for a review). Knowledge spillovers and labor market pooling are among the most important 'Marshallian' factors, but various other structural characteristics like industry size, an industry's dependence on raw materials, the presence of multi-unit firms, or foreign ownership also affect their spatial structure.

In the subsequent analysis, we control for these confounding time-varying agglomeration factors as follows. First, we control for knowledge spillovers using as a proxy an industry's research and development (R&D) intensity, i.e., the ratio of R&D expenditure to total output of that industry. By their very nature, knowledge spillovers are very hard to measure directly. The literature has often proxied them using patent citation data, i.e., patents originating in industry i that are cited by patents of industry j. While useful in a cross-sectional context, our twenty year panel does not allow us to exploit patent citation data. Second, along with knowledge spillovers, labor market pooling is another important source of agglomeration. To construct good proxies for labor market pooling, it is important to identify industry characteristics that are related to the specialization of the industry's labor force (see Rosenthal and Strange, 2001 ; and Lu and Tao, 2009). The literature suggests that agglomeration occurs because workers are able to move across firms and industries, thus improving the average quality of firm-worker matches. Furthermore, idiosyncratic productivity shocks at the firm level can be better hedged in locations where firms using similar workers concentrate. Firms also agglomerate to take advantage of scale economies associated with a large labor pool that allows industries to use the same type of workers. Since it is difficult to identify these characteristics, we employ a proxy related to workers' occupations. More specifically, we use the share of hours worked by all workers with post-secondary education in the total number of hours worked. [15]

We finally construct numerous time-varying controls that proxy for the remaining agglomeration factors in our econometric analysis. We firstly control for the importance of natural advantage in the agglomeration process. The importance of doing so has been pointed out, among others, by Kim (1995) and by Ellison and Glaeser (1999). We use the share of inputs from natural resource-based industries, and the sectoral energy inputs as a share of total sector output, as proxies for natural advantage. We secondly control for basic industry structure and scale effects by including the following controls : total industry employment ; mean plant size ; the Herfindahl index of firm-level concentration (employment based) ; [16] the share of plants controlled by multi-unit firms ; and the share of plants controlled by

[15] We also tried to construct proxies for labor market conditions using the non-production to production worker ratio and others educational characteristics of the work-force. The latter are available at a more aggregated industry level (L-level) from Statistics Canada's KLEMS database (e.g., the share of hours worked by all workers with a university degree, and the labor productivity index). These measures, however, proved to not give significant results in the time series because they change quite slowly over time.

[16] Estimates using a Herfindahl index of plant-level concentration are qualitatively similar.

foreign firms (see Table 2.3). These controls proxy for sectoral differences in the size distribution of firms and plants, for potential differences in the location patterns of multinationals and multi-unit firms, as well as for differences in 'business culture' (Rosenthal and Strange, 2003).

Note that all these controls are time varying and industry specific. When combined with both time and industry fixed effects, they will control for a wide range of factors that may drive changes in the degree of geographical concentration of industries that are unrelated to changes in transportation costs, trade, or input and output linkages. This will provide better identification. We now discuss remaining identification issues.

2.3.3 Identification issues

The three main problems that plague the identification of agglomeration effects are unobserved heterogeneity, omitted variable bias, and simultaneity bias. All studies based on cross-sectional data at the industry level (e.g., Rosenthal and Strange, 2001 ; Ellison, Glaeser, and Kerr, 2010) are potentially prone to these identification problems and use different strategies to overcome them. The panel nature of our data allows us to control for industry-specific time-invariant factors and general macroeconomic trends. Furthermore, the inclusion of a large set of time-varying industry controls for natural advantage, industry structure, ownership structure, and proxies for labor demand conditions and knowledge spillovers (see Section 2.3.2) substantially reduces the risk of omitted variable bias when estimating our key coefficients β_T for the trade cost correlates. However, neither the panel structure nor the controls will help with potential problems of reverse causality. These may affect our three variables of interest, namely transportation costs, trade exposure, and input-output linkages.

Transportation costs. It is well documented that productivity rises as an industry concentrates geographically (see, e.g., Rosenthal and Strange, 2004; Combes and Gobillon, 2014). Because our measure of transportation costs is computed on an ad valorem basis and includes the industry price index, the causality may run from agglomeration to lower prices and, therefore, lower ad valorem transportation costs. At the same time, agglomeration may lead to imbalances in shipping patterns, and the latter may increase the cost of transportation due to standard logistics problems like 'backhaul' of empty trucks (e.g., Jonkeren, Demirel, van Ommeren, and Rietveld, 2009; Behrens and Picard, 2011). Agglomeration would thus increase the transportation price index and affect our estimates. In a nutshell, $p_{\text{trans},t}/p_{m,t}$ in expression (E.2) is likely to be endogenous to the degree of geographical concentration of an industry, with stronger concentration increasing that ratio due to a combination of rising freight prices and lower output prices. Thus, the estimated OLS coefficient for transportation costs is likely to be upward biased in our model.[17]

To deal with that problem, we adopt three different strategies. First, we clear out the effect of productivity growth on prices (the presumed source of endogeneity) by regressing our transportation cost series on industry multi-factor productivity indices (from the KLEMS database), as well

[17] Industries that agglomerate are also likely to ship their output over different distances than industries that are less concentrated because of their location choices. This problem does not affect our estimates since our measure of transportation costs is constructed for a representative shipment over a fixed distance of 500 kilometers.

as industry and year fixed effects. We then use the residual from that regression as a proxy for the transportation cost series. By definition, that residual is orthogonal to any productivity driven price changes that could stem from the changing geographic concentration of industries. This strategy does, however, not deal directly with the transportation price index.

Second, as we have a large number of industries and a fairly large time dimension, our setting lends itself well to the construction of internal instruments. We implement the method suggested by Lewbel (2012), which exploits heteroscedasticity and variance-covariance restrictions to obtain identification with 2SLS when some variables are endogenous and when external instruments are either weak or not available.

Third, we use US manufacturing industry price indices as external instruments for the transportation cost series. The instrumentation strategy is similar to that of Ellison, Glaeser, and Kerr (2010), who instrument the US input-output matrix and the US industry labor requirements with those of the UK. The underlying idea is the following. Assume that the geographical concentration of an industry increases over time because of unobserved factors that we cannot control for in our analysis. The increasing geographical concentration then raises ad valorem transportation costs via price decreases of the industry's output. Provided that the changes for the US are not driven by the same unobserved factors that affect the spatial concentration of the industry in Canada, but that the US price series are correlated with the changes in $p_{\text{trans},t}/p_{m,t}$, they will provide valid instruments for the Canadian transportation cost series. Two potential limitations of these instruments are the following : (i) there may be common underlying unobserved factors that drive changes in the concentration of the same industries in Canada and the US; or (ii) the geographical concentration of an industry in Canada affects directly the productivity – and, therefore, the price indices – in the US. While we cannot completely rule out those possibilities, neither strikes us as extremely plausible. First, the panel nature and the extensive set of time-varying controls should pick up most of the unobserved factors that may drive the increasing concentration of the industry; and second, the Canadian economy is small compared to the US economy, so that changes in the degree of concentration in Canada are very unlikely to have substantial productivity impacts in the US. [18]

Trade exposure. As argued above, the geographical concentration of plants increases productivity and, therefore, may increase the propensity of an industry to export and to import. For example, the agglomeration of an industry may reduce prices, which makes import penetration harder. In that case, the dispersion of an industry may be associated with increasing imports since productivity falls. Also, the agglomeration of an industry may be associated with rising exports due to productivity gains – although the productivity gains reduce unit export values, the total value of exports may increase. We deal with the potential endogeneity of trade flows using the Lewbel (2012) estimator with internal instruments.

[18] The empirical elasticity of productivity to the density or size of economic activity is usually in the 3–8 percent range, and thus huge changes in the geographical structure would be required to obtain large productivity changes. Furthermore, empirical work has documented that the effects of shocks to Canadian productivity have very limited effects on the US, safe for a couple of states relatively close to the border or a couple of border-spanning industry networks (like the automotive industry).

Input-output linkages. Our measures of input-output linkages are, by construction, reasonably exogenous to the spatial stucture of the economy. First, observe that we compute those measures using national input-output shares instead of plant-level input-output shares. Hence, we do not pick up spuriously large values for inputs or outputs – due to substitution effects – when plants are located in close proximity to plants in related industries. Second, we exclude the own industry from the computation, so that the measure only picks up cross-industry links and not the geographical concentration of the industry itself (which is on the left-hand side of our regressions). Last, for each plant, the input and output distance is computed using *all other 256 industries in Canadian manufacturing*. For the geographical concentration of one industry to drive the input-output linkage measure, that industry would need to substantially affect the whole location patterns of most other related industries, which strikes us as fairly unlikely (though we cannot completely rule out this possibility). Although the input- and output-measures should be reasonably exogenous, we will also instrument them following Lewbel (2012) in the subsequent regressions. As we will see, our results are very stable across specifications.

As should be clear from the foregoing discussion, it is virtually impossible to fully solve all endogeneity issues given the level of aggregation at which we carry out our analysis. Yet, the panel nature of our data, our extensive set of time-varying controls, as well as the construction and instrumentation strategies for our main variables of interest – transportation costs, trade exposure, and input-output linkages – all help us to be reasonably confident that we identify causal effects of changes in those covariates $\mathbf{T}_{m,t}$ on our measure $\gamma_{m,t}(d)$ of geographical concentration.

2.4 Empirical results

We estimate four specifications based on equation (E.1), which differ by the set of industry characteristics and controls that they include.[19] **Model 1** includes a measure of industry size, proxies for industry structure (the Herfindahl index of the firm-size distribution, mean plant size, the share of plants controlled by multiplant firms, and the share of plants controlled by foreign-owned firms), and proxies for natural advantages (the share of inputs from natural resource-based industries, and the share of energy inputs in total output). It also includes the 'Marshallian covariates', namely the proxies for the skill composition of the workforce and for knowledge spillovers. **Model 2** adds our trade variables (import and export shares by broad trading partner groups) to the baseline case. **Model 3** includes transportation costs and our input-output distances – the industry mean of the average minimum distance to a dollar of inputs or outputs computed using the five nearest plants in each industry – as well as our minimum distance (density) control.[20] Finally, **Model 4** – our preferred specification – includes all the variables and uses the residual transport cost obtained from a first-stage regression of that cost on industry multi-factor productivities and a

[19] We performed the Hausman test for (E.1) to confirm that the appropriate estimator is a fixed-effects estimator and not a random-effects estimator. The result of the test strongly confirms (at the 1% level) that the fixed-effects estimator is the preferred specification. Note also that we work with the universe of manufacturing industries, so that there is no sampling variability with respect to industries.

[20] Using $N = 3, 5, 10$ yields qualitatively very similar results.

set of industry and year fixed effects (see Section 2.3.3 for details).[21]

2.4.1 Baseline results

Our baseline results are presented in Table 2.4, which uses the unweighted (plant count) CDF at 50 kilometers distance as the dependent variable. Robustness checks with respect to that distance are provided in the next section, whereas robustness checks using the employment- and sales-weighted CDFs are relegated to Appendix E (see Table 2.10). All variables except the trade shares and the shares of plants controlled by multiplant and by foreign firms enter as natural logarithms into the regressions, so that their coefficients can be interpreted as elasticities.

As can be seen from Table 2.4, in Model 1, which includes only control variables, only total industry employment and the share of plants controlled by foreign firms are statistically significant. Put differently, growing industries and industries with an increasing share of foreign-controlled plants tend to become more localized. The first finding is at odds with results by Dumais, Ellison, and Glaeser (2002), who document that growing US manufacturing industries tend to disperse, whereas shrinking ones concentrate (see also Behrens, 2014, for the case of textiles in Canada). The second finding is in line with previous evidence which documents that foreign firms tend to locate within existing clusters (see, e.g., Head, Ries, and

[21] When using the 'ad valorem trucking cost residual' from the first-stage regression, we need to bootstrap the standard errors to control for the presence of an estimated regressor. We did this for the baseline specification (see Model 4 in Table 2.8), and it makes virtually no difference. We hence report non-bootstrapped standard errors in most specifications.

Table 2.4 Baseline estimation results for specification (E.1).

	Dependent variable is the CDF at 50 kilometers			
Variables	**(Model 1)**	**(Model 2)**	**(Model 3)**	**(Model 4)**
Total industry employment	0.179^{b}	0.150^{b}	0.288^{a}	0.289^{a}
	(0.070)	(0.067)	(0.039)	(0.039)
Firm Herfindahl index (employment based)	-0.028	-0.038	0.002	0.001
	(0.036)	(0.035)	(0.021)	(0.021)
Mean plant size	-0.026	-0.029	-0.280^{a}	-0.282^{a}
	(0.078)	(0.077)	(0.045)	(0.044)
Share of plants affiliated with multiplant firms	-0.301	-0.203	-0.006	-0.005
	(0.191)	(0.164)	(0.100)	(0.099)
Share of plants controlled by foreign firm	0.584^{a}	0.660^{a}	0.338^{a}	0.340^{a}
	(0.216)	(0.214)	(0.125)	(0.124)
Natural resource share of inputs	0.024	0.034^{c}	0.008	0.008
	(0.023)	(0.020)	(0.014)	(0.014)
Energy share of inputs	-0.037	-0.024	0.054	0.037
	(0.052)	(0.026)	(0.040)	(0.040)
Share of hours worked by all workers with post-secondary education	0.032	0.013	0.036	0.032
	(0.078)	(0.069)	(0.045)	(0.045)
In-house R&D share of sales	-0.031	0.006	0.011	0.014
	(0.020)	(0.022)	(0.015)	(0.015)
Asian share of imports		-1.570^{a}	-1.132^{a}	-1.119^{a}
		(0.456)	(0.380)	(0.383)
OECD share of imports		-1.032^{b}	-0.491	-0.476
		(0.412)	(0.344)	(0.345)
NAFTA share of imports		-1.114^{a}	-0.562^{c}	-0.549^{c}
		(0.382)	(0.327)	(0.327)
Asian share of exports		0.473	0.482	0.482
		(0.500)	(0.405)	(0.412)
OECD share of exports		0.412^{c}	0.440^{b}	0.443^{b}
		(0.237)	(0.189)	(0.193)
NAFTA share of exports		0.353	0.319	0.318
		(0.267)	(0.196)	(0.201)
Ad valorem trucking costs		-0.291^{b}	-0.208^{b}	
		(0.135)	(0.088)	
Ad valorem trucking costs (residual)				-0.260^{a}
				(0.079)
Input distance			-0.361^{a}	-0.358^{a}
			(0.055)	(0.055)
Output distance			-0.313^{a}	-0.318^{a}
			(0.042)	(0.043)
Average minimum distance			-0.296^{a}	-0.294^{a}
			(0.039)	(0.039)
Number of NAICS industries	257	257	257	257
Number of years	17	17	17	17
Year dummies	yes	yes	yes	yes
Industry dummies	yes	yes	yes	yes
Observations (NAICS× years)	4,369	4,369	4,369	4,369
R^2	0.089	0.137	0.516	0.518

Notes : The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. Standard errors are clustered at the industry level and given in parentheses. Our measures of input and output distances, as well as average minimum distance, are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported.

Swenson, 1995; and Guimaraes, Figueiredo, and Woodward, 2000). The natural resource share of inputs variables are basically never significant across all four models, i.e., changes in natural advantage is not strongly associated with changes in localization. One of the reasons for this is that their time variation is small. The same holds true for the 'Marshallian covariates', which are not significant either. Again, lack of time-series variation may explain that result.

Turning to Model 2, rising shares of imports are across the board associated with falling localization. The (non-OECD) Asian share of imports, which we use as a proxy for low-wage countries, has the largest estimated coefficient in absolute value and is the most statistically significant. One explanation for the dispersive effect of import competition is that firms become more footlose as they source a larger share of their intermediates from abroad and no longer rely on (localized) domestic suppliers. Another explanation, for which Holmes and Stevens (2014) provide empirical evidence, is that import competition from low-wage countries leads to significant exit of large plants that produce standardized 'main segment' goods.[22] If those plants are the ones that are predominantly clustered at short distances, their exit will significantly reduce the extent of measured localization.[23] As can be also seen from Model 2 in Table 2.4, rising export

[22] We cannot disentangle the impact of exit vs relocation on the spatial structure. However, we control for the size of the industry, which at least partly picks up entry and exit dynamics. Note that relocations are quite rare and should have little impact on our results. The bulk of the variation is driven by entry and exit.

[23] This is a somewhat surprising result, because we would expect the productivity enhancing effects of localization to shelter firms from low-wage competition. Yet, one

shares are across the board associated with increasing localization, though the effect is only significant for the share of exports to OECD countries. This pattern may be driven by the fact that more isolated non-exporting plants have a higher chance to exit the market, or that localization increases the export participation and performance of plants (e.g., Koenig, Mayneris, and Poncet, 2010).

Regarding transportation costs, we have no clear prior as to their impact, as stated before. In theory, the effects of changes in transportation costs on the geographical concentration of economic activity depend on the underlying dispersion forces in the economy. If, on the one hand, firms tend to serve a predominantly dispersed immobile demand, lower transportation costs would tend to be agglomerative, as in Krugman (1991). If, on the other hand, all demand is a priori mobile and dispersion stems from urban costs due to agglomeration, lower transportation costs would tend to be dispersive (Helpman, 1998; Behrens, Mion, Murata, and Suedekum, 2012). As can be seen from Model 2 in Table 2.4, lower transportation costs are associated with more geographical concentration in our estimations.[24]

should keep in mind that clustering provides firms with benefits as long as clusters grow (positive shocks), but that the unravelling of clusters (negative shocks) may lead to a domino effect as the agglomeration benefits dissipate with the exit of firms. Also, as shown by Holmes and Stevens (2014), plants in clusters operate on different market segments than non-clustered plants, and they are more vulnerable to import competition.

[24] We also experimented with different non-linear transportation cost specifications. More precisely, we estimated the effect of transportation costs with a spline, allowing the coefficients to vary between ad valorem rates of 0 to 0.05% (low), 0.05 to 15% (moderate), and 15% or greater (high). These are admittedly arbitrary categories, but ones that we believe make intuitive sense. The results are, by and large, consistent with the sim-

Model 3 adds our input- and output-linkage measures, whereas Model 4 uses the residual transport cost instead of the original variable. The input-output coefficients are highly significant and negative in all specifications, and they tend to be of similar magnitude (as in Ellison, Glaeser, and Kerr, 2010) : industries tend to follow their suppliers and customers. If supplier industries tend to become more dispersed (in the sense of being, on average, further away from plants in the downstream industry), the downstream industry becomes less concentrated too. This result suggests that the geographic concentration of upstream supply and downstream demand goes hand-in-hand with increasing localization of an industry. Note that this effect is not driven by changes in overall density, since we control for this (and the associated variable is highly significant). The coefficient for transportation costs remains fairly stable when introducing the input-output linkages, as can be seen from Model 3, albeit it slightly decreases in absolute value, as expected. Last, as can be seen from Model 4, the coefficient on transportation costs becomes larger in absolute value when using the productivity-purged residual. This is in line with our expectations discussed in Section 2.3.3, where we have argued that endogeneity concerns due to reverse causality are likely to bias the coefficient upwards (towards zero in this case). Observe that the endogeneity bias does not seem to be too severe, which is in line with findings related to the endogeneity of wages in standard 'wage-density' regressions (see, e.g., Combes, Duranton, and Gobillon, 2011, for a discussion). Last, as can be seen from our prefered specification (Model 4 in Table 2.4), about half of the time-series variation

pler specification that we use. Yet, we find that at low levels, the effect of transportation costs is positive or insignificant. At moderate levels, the coefficient is negative and always significant, and at high levels the coefficient is negative and insignificant.

in localization is explained by the model.

As shown in Section 2.2.2, the degree of localization of manufacturing industries has significantly fallen in Canada between 1990 and 2009. How much of that change is explained by changes in transportation or trade costs ? To see how much of the observed change can be attributed to changes in those variables, we compute the predicted change in the CDFS by holding, one-by-one, the : (i) ad valorem trucking costs ; (ii) different import shares ; and (iii) the input or output distances to their 1992 values, while still allowing the other variables to change through time. The results are summarized in Table 2.5.

Table 2.5 Predicted contributions to changes in geographical concentration.

Observed avg. CDF changes 1992–2008	Counterfactual avg. CDF changes 1992–2008 for changes in			
	Ad valorem trucking costs	Import shares	Input distances	Output distances
-23.37%	-28.36%	-14.63%	-30.32%	-31.86%

Notes : Observed and predicted changes in the unweighted cross-industry average CDFS at 50 kilometer distance.

As can be seen, the observed change in the cross-industry average CDF between 1992 and 2008 at a distance of 50 kilometers is -23.37%. Holding the ad valorem trucking rate fixed at its 1992 level, the change would have been -28.36%. Thus, had transportation costs not decreased, the geographical concentration would have fallen by about 5 percentage points more (about 20% of the overall change). Turning to imports, holding all import shares constant at their 1992 level, the change in the CDF would have been -14.63%. In words, had imports remained at their 1992 levels, the geographical concentration would have fallen by about 9 percentage points (i.e., 60%) less than what we observed. Clearly, these are large effects, thus showing that *transportation costs and trade exposure have sizable effects on the spatial structure of economic activity*. Last, turning to input and output distances, in the former case the change would have been -30.32% (about 7 percentage points more) and in the latter case the change would have been -31.86% (about 8.5 percentage points more). Had supplier and customer access not changed – these distances fell through time, as can be seen from Figure 2.8 – the dispersion of industries would have been even greater than the one we observed.

2.4.2 Robustness checks

We now provide evidence on the robustness of our key findings. To this end, we run five main types of robustness checks. First, we investigate the robustness of our results to the choice of the dependent variable. Table 2.10 in Appendix E shows that the effect of transportation costs on localization is weaker – and the explanatory power of the model lower – when the latter is measured using either employment- or sales-weighted CDFS. Although the key qualitative flavor of the results and the sign and significance of our key coefficients remain largely unchanged, the estimates using employment- or sales-weighted K-densities are less sharp. Furthermore, the effect of import competition tends to be more limited to imports from Asia, and the coefficient tends to be smaller. This suggests that much of the adaptation to import competition, particularly from low wage countries which are responsible for the bulk of exit in Canadian manufacturing (Behrens, 2014), occurs for smaller plants and firms. Turning to the residual transportation cost variable, it remains significantly negative in all specifications that we estimate, irrespective of how we construct the dependent variable. The same holds true for the input-output distances and the overall density control. In a nutshell, changes in transportation costs and in input-output linkages have a significant effect on the spatial concentration of economic activity, no matter whether we consider plants, employment, or sales to measure that concentration.

Table 2.6 Estimation results for specification (E.1) by distance and by incremental change in the CDF.

	Model (4), by distance			Model (4), by incremental CDF			
Variables	CDF 10km	CDF 100km	CDF 500km	$\Delta\gamma_m(10,25)$	$\Delta\gamma_m(25,50)$	$\Delta\gamma_m(50,100)$	$\Delta\gamma_m(100,500)$
Asian share of imports	-1.359^a	-0.923^a	-0.307^b	-1.029^b	-0.724^b	-0.352	0.583
	(0.467)	(0.299)	(0.139)	(0.433)	(0.337)	(0.235)	(0.429)
OECD share of imports	-0.666	-0.334	0.018	-0.451	-0.174	0.102	0.721
	(0.425)	(0.271)	(0.158)	(0.374)	(0.285)	(0.211)	(0.455)
NAFTA share of imports	-0.710^c	-0.411	-0.037	-0.527	-0.284	0.007	0.587
	(0.396)	(0.254)	(0.135)	(0.359)	(0.268)	(0.190)	(0.372)
Asian share of exports	0.399	0.415	0.096	0.630	0.658	0.421	-0.782
	(0.439)	(0.345)	(0.123)	(0.426)	(0.404)	(0.264)	(0.714)
OECD share of exports	0.366^c	0.419^b	0.265^a	0.545^a	0.662^a	0.470^a	-0.112
	(0.219)	(0.166)	(0.094)	(0.197)	(0.224)	(0.156)	(0.304)
NAFTA share of exports	0.217	0.314^c	0.139^c	0.440^b	0.541^b	0.431^a	-0.191
	(0.231)	(0.174)	(0.080)	(0.211)	(0.215)	(0.162)	(0.274)
Ad valorem trucking costs (residual)	-0.269^a	-0.250^a	-0.212^a	-0.253^a	-0.238^a	-0.229^a	-0.105
	(0.080)	(0.073)	(0.048)	(0.079)	(0.080)	(0.069)	(0.090)
Input distance	-0.382^a	-0.340^a	-0.242^a	-0.332^a	-0.322^a	-0.315^a	-0.193^a
	(0.063)	(0.049)	(0.033)	(0.061)	(0.055)	(0.054)	(0.041)
Output distance	-0.307^a	-0.307^a	-0.197^a	-0.341^a	-0.340^a	-0.302^a	-0.122^a
	(0.046)	(0.040)	(0.027)	(0.045)	(0.045)	(0.045)	(0.039)
Average minimum distance	-0.322^a	-0.268^a	-0.137^a	-0.298^a	-0.243^a	-0.204^a	-0.038
	(0.046)	(0.035)	(0.024)	(0.041)	(0.043)	(0.038)	(0.036)
R^2	0.473	0.540	0 .545	0.481	0 .417	0.436	0.168

Notes : All estimations for 257 industries and 17 years (4,369 observations). The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF at the reported distance. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. All specifications include industry and year fixed effects. Standard errors, given in parentheses, are clustered at the industry level. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported. All industry controls (Total industry employment; Firm Herfindahl index (employment based); Mean plant size; Share of plants affiliated with multiplant firms; Share of plants controlled by foreign firms; Natural resource share of inputs; Energy share of inputs; Share of hours worked by all workers with post-secondary education; In-house R&D share of sales) are included but not reported.

Second, we check the robustness of our results to the choice of the distance d at which the K-density CDF is evaluated. Doing so allows us to highlight how our key covariates influence the localization of industries at different geographical scales. Furthermore, we can provide plots of the marginal effects of our variables of interest over the whole distance range, thus allowing for a fine analysis of the spatial dimensions of the changes in agglomeration due to changes in the trading environment. The left half of Table 2.6 summarizes our results for different distances. To save space, we only report results for Model 4 at three selected distances : 10, 100, and 500 kilometers. As can be seen, the qualitative results do not depend on the distance threshold d. This holds true for all our key variables, thus showing that transportation costs, trade, and input-output linkages matter at most spatial scales. Furthermore, there is a general tendency for the values and significance of the covariates to attenuate as the CDF increases in distance. This can be seen from the right half of Table 2.6, where we define the incremental distance of the CDF between distance d_1 and distance $d_2 > d_1$ as follows : $\Delta\gamma_m(d_1, d_2) = \gamma_m(d_1) - \gamma_m(d_2)$. We estimate the marginal effects of our variables by 'distance bands'. As one can see, there is basically no more additional effect of our covariates on the degree of localization beyond about 100 kilometers, except for our input-output measures. Furthermore, the largest (and statistically most significant results) occur in the distance bands between either 10 and 25 kilometers, or between 25 and 50 kilometers. This result suggests that many of the agglomeration mechanisms linked to transportation, trade, and input-output linkages operate at the scale of metropolitan areas.[25] At longer distances – beyond about 200 kilometers – other factors that do not figure in our model drive the clustering of firms, or incremental clustering becomes weak and fairly unim-

[25] For example, the island of Montreal is about 50 kilometers long.

portant.[26] The decrease in the marginal effects can be clearly seen from Figure 2.9, which depicts the incremental change in coefficients of our key variables by 10 kilometers steps increases in distances (since all marginal coefficient changes are statistically zero after 200 kilometers, we limit the plots to that range).

Third, we first re-estimate the model by averaging all variables over five year periods. Doing so reduces the year-on-year volatility of some variables (e.g., the trade variables), and allows for slowly moving variables like R&D expenditures or localization patterns to be potentially better identified in the regressions. It also deals potentially with business cycle aspects that may drive the changes in the geographical concentration of industries. The last three columns of Table 2.10 in Appendix E show that our basic findings are unchanged when replacing year-on-year variations with five-year averages.

Fourth, our results may be partly driven by sectoral 'outliers'. For example, as documented by Behrens (2014), the textile industries in Canada experienced a remarkable downward trend in terms of number of plants and the geographical dispersion of activity in the wake of the end

[26] This result is not really surprising. There are two possible explanations. First, the determinants of localization may operate at 'small' spatial scales, whereas they are no longer very relevant at longer distances. Second, the CDFs across industries tend to display less variation the longer is the distance d. The reason is that they are bounded from above by unity, and we converge by construction to that value for all industries if we compute them over sufficiently large distances. This problem is similar to the spatial scale of aggregation issue when using different spatial scales to compute discrete measures like the Ellison and Glaeser (1997) index used by Rosenthal and Strange (2001).

Figure 2.9 Transportation, trade, and input-output coefficients (marginal effect by distance).

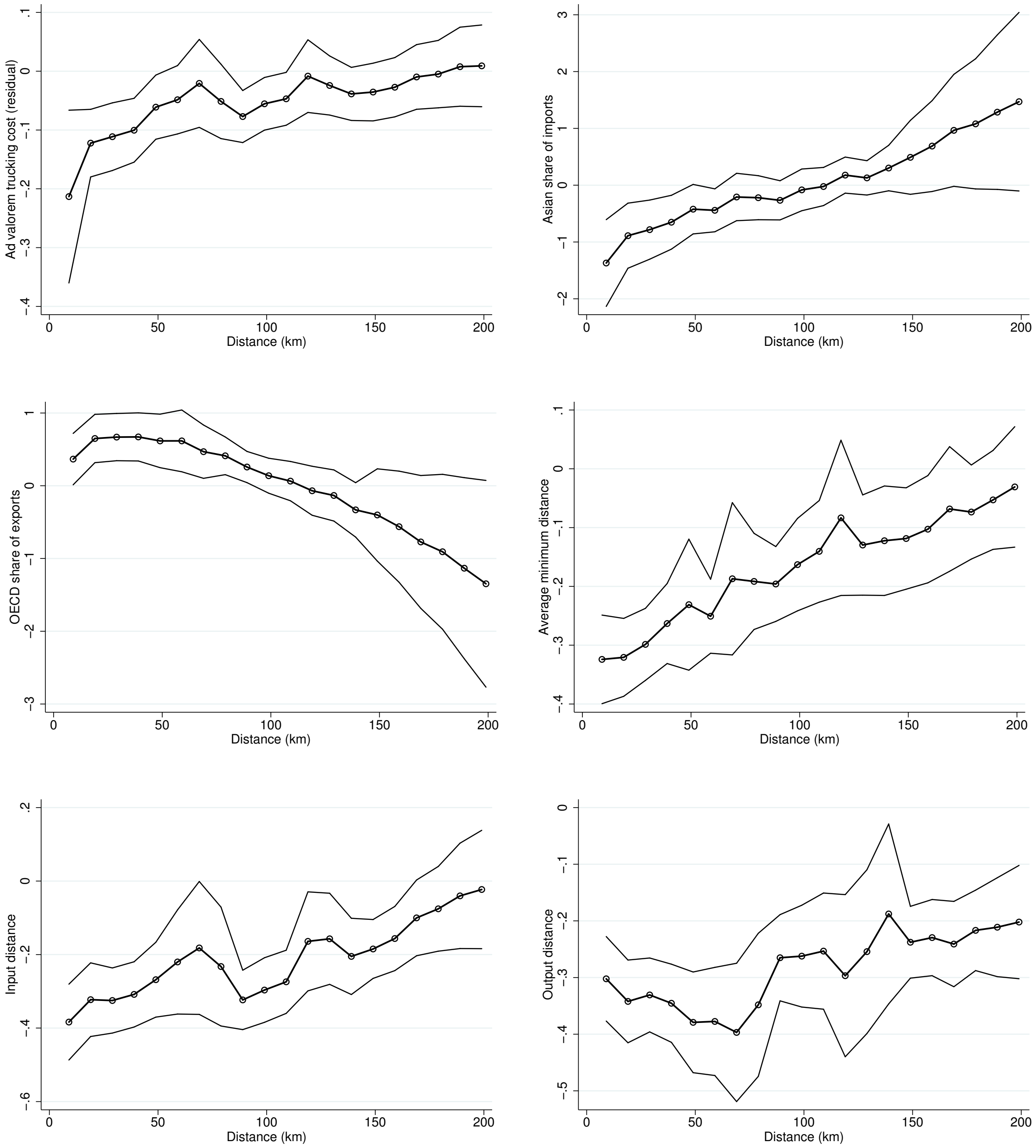


Table 2.7 Estimation of specification (E.1) excluding textile and high-tech industries.

	Excluding textiles industries			**Excluding high-tech industries**		
Variables	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km
Asian share of imports	-0.568^c	-0.508^c	-0.211	-1.517^a	-1.035^a	-0.380^b
	(0.322)	(0.282)	(0.174)	(0.554)	(0.350)	(0.155)
OECD share of imports	-0.035	0.007	0.137	-0.860	-0.474	-0.084
	(0.275)	(0.241)	(0.181)	(0.530)	(0.333)	(0.177)
NAFTA share of imports	-0.097	-0.062	0.076	-0.878^c	-0.531^c	-0.133
	(0.251)	(0.221)	(0.156)	(0.499)	(0.317)	(0.157)
Asian share of exports	0.627	0.505	0.096	0.468	0.469	0.111
	(0.440)	(0.358)	(0.130)	(0.490)	(0.378)	(0.121)
OECD share of exports	0.471^b	0.413^b	0.249^b	0.346	0.424^b	0.271^a
	(0.186)	(0.161)	(0.097)	(0.236)	(0.170)	(0.098)
NAFTA share of exports	0.400^b	0.348^b	0.128	0.149	0.275	0.124
	(0.196)	(0.170)	(0.080)	(0.246)	(0.179)	(0.085)
Ad valorem trucking costs (residual)	-0.213^a	-0.210^a	-0.193^a	-0.396^a	-0.324^b	-0.205^a
	(0.077)	(0.072)	(0.049)	(0.145)	(0.128)	(0.068)
Input distance	-0.458^a	-0.439^a	-0.315^a	-0.387^a	-0.346^a	-0.245^a
	(0.051)	(0.049)	(0.036)	(0.075)	(0.057)	(0.038)
Output distance	-0.265^a	-0.245^a	-0.155^a	-0.333^a	-0.336^a	-0.216^a
	(0.043)	(0.040)	(0.029)	(0.051)	(0.044)	(0.030)
Average minimum distance	-0.289^a	-0.265^a	-0.142^a	-0.321^a	-0.257^a	-0.128^a
	(0.041)	(0.038)	(0.026)	(0.053)	(0.038)	(0.026)
R^2	0.516	0.532	0.539	0.481	0.556	0.553

Notes : All estimations for 257 industries and 17 years (4,369 observations). ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. All specifications include industry and year fixed effects. Standard errors are clustered at the industry level and given in parentheses. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. A constant term is included but not reported. All industry controls (Total industry employment ; Firm Herfindahl index (employment based) ; Mean plant size ; Share of plants affiliated with multiplant firms ; Share of plants controlled by foreign firms ; Natural resource share of inputs ; Energy share of inputs ; Share of hours worked by all workers with post-secondary education ; In-house R&D share of sales) are included but not reported.

of the Multi-Fibre Arrangement in 2005. Given that these sectors were initially among the most strongly localized ones (see Table 2.1), and given that these industries have a tendency to display very strong co-agglomeration patters (see Ellison, Glaeser, and Kerr, 2010, p.1199), the large changes in these sectors may drive some of the results. That this is not the case, and that all of our main findings are robust to the exclusion of those sectors, is shown in Table 2.7. The left panel provides results when excluding the textile sectors, whereas the right panel provides results when excluding the high-tech sectors. [27] In both cases, our key coefficients are qualitatively unchanged. Note, however, two differences. First, the input-output linkages become more negative when excluding the textile industries. Second, the transport cost variable becomes more negative when excluding the high-tech industries. The former result suggests that textile industries are less dependent on input-output linkages than other industries (e.g., manufacturing durables). The latter result suggests that spatial patterns of high-tech industries are less impacted by changes in transportation costs, so that their inclusion tends to reduce the estimated coefficient on transport costs.

As a final series of robustness checks, we ran a number of experiments that we do not report in detail. We used, for example, the ICT investment variables from the KLEMS database, interacted with the other variables of the model, to check whether changes in communication costs have the

[27] Our definition of high-tech sectors is based on the US Bureau of Labor Statistics classification by Hecker (2005). This definition of high-tech industries is 'input based'. An industry is 'high-tech' if it employs a high proportion of scientists, engineers or technicians. As shown by Hecker (2005), these industries are also usually associated with a high R&D-to-sales ratio, and they also largely – but not always – produce goods that are classified as 'high-tech' by the Bureau of Economic Analysis.

same effect than changes in transportation costs. We did not get any significant coefficients – neither for the direct effects, nor for the interaction terms. We also estimated models with heterogeneous coefficients since transportation costs differ across industries. To this end, we split our sample into high-vs-low transport cost industries, using a 'below median'–'above median' criteria. The two coefficients were statistically identical. We also treated decreasing/increasing transportation costs in an asymmetric way as they may have asymmetric impacts. Again, the two coefficients were fairly close. We also replaced our measures of input and output linkages with the industry 'material share to sales' ratio, a proxy for reliance on intermediate inputs. That variable turns out to be insignificant in our regressions, whereas the other coefficients are largely unaffected. We also ran the model in a pooled cross-section and by year using a between estimator and found roughly the same signs and significant coefficients for transportation costs and the input and output distance measures. The cross-sectional results are summarized by Table 2.12 in Appendix E. It is worth noting that, although the levels of trade costs do seem to matter for the geographical concentration of industries, the time-series changes in those costs are much more strongly associated with changes in that concentration. Last, we also tried to control for the 'labor intensity' of an industry (not just highly skilled workers vs low-skilled workers). We constructed different measures using the quantity index of labor and the quantity index of capital from the KLEMS data, but these variables turned out again to be insignificant in our regressions.

To summarize, our key findings are fairly robust and continue to hold true in a variety of alternative specifications. Imports are mostly dispersive, whereas exports play in the opposite direction. Sectors that see their transportation costs increase tend to disperse more. [28] Last, our microgeographic measures of input and output linkages are across the board the most significant and stable variables. Since they are computed by taking into account the *relative positions of all industries with respect to each other*, our findings suggest that there are very strong regularities in how industries relate spatially to one another and on how changes in the spatial structure of some industries shape changes in the spatial structure of linked industries.

2.4.3 Controlling for endogeneity

We finally address the potential endogeneity concerns that we discussed at length in Section 2.3.3. The results of the different estimations are summarized in Table 2.8.

Model 4 replicates column 4 of Table 2.4. As explained previously, we use the residual of a regression of ad valorem trucking costs on sectoral multifactor productivity – including a set of industry and year fixed effects – in that specification. The residual from that regression is, by construction, orthogonal to multifactor productivity. Observe that Model 4 in Table 2.8 differs from Model 4 in Table 2.4 only by the standard errors, which are bootstrapped using 200 replications. Comparing the results in the two tables

[28] Holmes and Stevens (2014) document for the case of US manufacturing that import competition is dispersive for big firms that produce 'primary segment goods' in clusters, whereas small firms outside are less affected since they produce 'specialty segment goods' that are more costly to transport. Higher transport costs shield those small firms, whereas more trade exposes the larger firms. Our results concerning the impacts of changes in transportation costs and trade exposure are broadly in line with those findings.

shows that no coefficient changes its significance level. The coefficient on the residual ad valorem trucking rate is larger in absolute value than the coefficient that is not purged from productivity effects (-0.260 instead of -0.208). The direction of the bias is consistent with an industry price-decreasing effect of agglomeration ($p_{m,t}$ decreases in (E.2)) or a transportation sector price-increasing effect ($p_{\text{trans},t}$ increases in (E.2)). Both of these effects could underlie the upward bias in the coefficient on transportation costs that we estimate.

Model 5 is a standard 2SLS instrumental variable regression. We instrument the ad valorem trucking rate using formula (E.2), where we replace Canadian price indices with their US counterparts to construct our instrument. The rationale underlying this instrumentation strategy was explained before in Section 2.3.3 and is similar in spirit to that in Ellison, Glaeser, and Kerr (2010). The first-stage results are summarized in Table 2.11 in Appendix E. As can be seen from that table, the instrument is strong (with a first-stage F-test value of 19.07 and an R^2 of 0.62). Table 2.8 shows that the instrumented coefficient is substantially more negative than the coefficient for the residual ad valorem trucking rate, itself more negative than the coefficient using the unpurged trucking rate. The direction of the bias in the estimated coefficients is the same in Models 4 and 5, which suggests that OLS estimates significantly underestimate the impact of changes in transportation costs on the spatial concentration of industries.

Finally, models 6 and 7 in Table 2.8 use the Lewbel (2012) estimator with internal instruments for the input-output distances and a set of the trade shares (see Appendix D for more details on the implementation).[29] The excluded external instrument is the US price-based ad valorem trucking costs as before. As can be seen from the results, the instrumented coefficient on the Asian share of imports increases, as do most of the other trade share coefficients. At the same time, both the magnitude of transportation costs and of the input and output distances decreases slightly. However, these variables remain significant and their magnitude is in the same ballpark than in the case of OLS (-0.194 vs -0.208 from Model 3 in Table 2.4). Thus, our results appear to be robust. Changes in transportation costs, in international trade exposure, and in access to suppliers and clients all affect the geographical concentration of manufacturing industries even when potential endogeneity concerns are taken into account.

2.5 Concluding remarks

Using a long panel of micro-geographic concentration measures, we have substantiated evidence for the causal effects of changes in transport costs – broadly defined – on the geographical concentration of Canadian manufacturing industries. We find large effects. Holding all other variables fixed at their 1992 levels, changes in trucking rates explain about 20%, changes in input-output linkages about 30%, and changes in import exposure about 60% of the observed decline in spatial concentration over the 1992–2008 period. Our qualitative results are robust to endogeneity

[29] Since there is an insignificant correlation between the OECD export share and the squared residuals, we did not include it. We substituted instead the NAFTA import share because it is consistently significant in the baseline set of models and it meets the criteria for being internally instrumented.

Table 2.8 Controlling for potential endogeneity of $\mathbf{T}_{m,t}$ in specification (E.1).

	Dependent variable is the CDF at 50 kilometers			
	(Model 4)	**(Model 5)**	**(Model 6)**	**(Model 7)**
Variables	Base	IV-2SLS	Lewbel 1	Lewbel 2
Asian share of imports	-1.119^{a}	-1.110^{a}	-1.589^{a}	-1.621^{a}
	(0.420)	(0.377)	(0.533)	(0.495)
OECD share of imports	-0.476	-0.486		-0.673
	(0.393)	(0.341)		(0.416)
NAFTA share of imports	-0.549	-0.558^{c}	-0.756^{c}	-0.850^{b}
	(0.374)	(0.323)	(0.435)	(0.419)
Asian share of exports	0.482	0.452		0.641
	(0.409)	(0.398)		(0.580)
OECD share of exports	0.443^{b}	0.422^{b}		0.638^{c}
	(0.202)	(0.189)		(0.360)
NAFTA share of exports	0.318	0.297		0.532
	(0.206)	(0.194)		(0.365)
Ad valorem trucking costs		-0.346^{a}	-0.180^{b}	-0.194^{b}
		(0.095)	(0.091)	(0.089)
Ad valorem trucking costs (residual)	-0.260^{a}			
	(0.083)			
Input distance	-0.358^{a}	-0.359^{a}	-0.132^{c}	-0.223^{a}
	(0.053)	(0.054)	(0.077)	(0.076)
Output distance	-0.318^{a}	-0.314^{a}	-0.385^{a}	-0.349^{a}
	(0.040)	(0.042)	(0.086)	(0.086)
Average minimum distance	-0.294^{a}	-0.293^{a}		
	(0.041)	(0.039)		
R^2	0.518	0.514	0.316	0.328

Notes : The dependent variable is the unweighted (count based) Duranton-Overman K-density CDF. ^{a}, ^{b} and ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. Our measures of input and output distances are computed using $N = 5$. 'Ad valorem trucking costs (residual)' denotes the residual of the regression of 'Ad valorem trucking costs' on industry multi factor productivity. Model 4 replicates our preferred model but the standard errors are bootstrapped because of the generated regressor. Model 5 instruments the 'Ad valorem trucking costs' using costs constructed from US price indices. Models 6 and 7 use the Lewbel (2012) methodology to instrument input-output distances and trade shares. In model 6 only a subset of the import shares is instrumented, while all trade shares are instrumented in model 7. See Appendix D for details. A constant term is included but not reported. All industry controls (Total industry employment ; Firm Herfindahl index (employment based) ; Mean plant size ; Share of plants affiliated with multiplant firms ; Share of plants controlled by foreign firms ; Natural resource share of inputs ; Energy share of inputs ; Share of hours worked by all workers with post-secondary education ; In-house R&D share of sales) are included but not reported.

concerns and to the way we measure the spatial concentration of industries – in terms of plants, employment, or sales.

Our research makes three distinct contributions. First, we construct new and finer measures of the costs of trading goods across space than in the previous literature. We use detailed microdata on freight transportation to estimate industry-level time-varying measures of transport costs, and we propose a new way of constructing micro-geographic input-output linkages based on location patterns and national input-output tables. Second, we are – to the best of our knowledge – among the first to exploit the time-series variation in the data to shed light on what drives *changes* in the spatial concentration of industries. The panel nature of the data allows us to control for unobserved heterogeneity and a battery of other time-varying factors. We have highlighted a hitherto unnoticed tradeoff when using time-varying geographical concentration measures constructed from micro-geographic data : the need to smooth out the time-series volatility at short distances versus the potential underestimation bias of the concentration measures due to the smoothing. More work is called for here to propose better measures of concentration in the presence of substantial plant-level churning in the data. Last, by exploiting the spatially continuous nature of our data, we have also shed light on the spatial scale at which the aforementioned effects operate. In line with previous research that has looked at the geographical scale of knowledge spillovers, labor market pooling, and input-output linkages, we find that the costs of trading goods influence the spatial structure of industries at small geographical scales : whereas the effects are sizable at short distances up to 50 kilometers, they basically vanish beyond about 100–200 kilometers.

We believe that our results are important because they show that, although the costs of trading goods across space have hit historical lows, changes in those costs still do shape location patterns of industries. In a world where profit margins have become tiny, even small changes in trade costs can have large effects on firm location, specialization patterns, and trade. In a nutshell, the often heralded 'death of distance' is premature. The world is not yet flat : transport costs matter !

Acknowledgements. We thank Gabriel Ahlfeldt, John Baldwin, Antoine Bonnet, Pierre-Philippe Combes, Wulong Gu, Julien Martin, Se-il Mun, Yasusada Murata, Frédéric Robert-Nicoud, Mathieu Parenti, Eugenia Shevtsova, Peter Warda, and conference participants at the 2013 NARSC Meetings in Atlanta, the 2014 Meetings of the Société canadienne de science économique in Ottawa, the 2014 Western Regional Science Meetings in San Diego, the 2014 African Econometric Society Meetings, the Summer School on 'New Structural Economics and Development in Africa' at Peking University, the Center for Market Studies and Spatial Economics in Saint Petersburg, the University of Geneva, the University of Quebec at Montreal, and the 2014 ERSA–UEA Meetings in Saint Petersburg, for very helpful comments and suggestions. Financial support from Statistics Canada's 'Trade Cost Analytical Projects Initiative' (API) is gratefully acknowledged. Tim Pendergast provided excellent technical assistance with program codes. Behrens and Bougna gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*. Bougna further gratefully acknowledges financial support from CIRPÉE. This work was carried out while Behrens and Bougna were Deemed Employees of Statistics Canada. The views expressed in this paper and all remaining errors are ours. This paper has

been screened to ensure that no confidential data are revealed.

2.6 Appendix to Chapter 2

This set of appendices is structured as follows. Appendix A describes our datasets, data sources, and key variables. Appendix B provides details on the Duranton-Overman K-density computations. Appendix C describes the construction of the weights used in our input-output measures. Appendix D provides details for the implementation of the Lewbel (2012) estimates. Last, Appendix E contains supplemental tables and results.

A. Data and data sources

This appendix provides details on the data used and the data sources. A description of the key variables and the associated descriptive statistics are given in Table 2.3 in the main text.

Plant-level data and industries. Our analysis is based on the Annual Survey of Manufacturers (ASM) Longitudinal Microdata file. This data cover the years from 1990 to 2010. Our focus is on manufacturing plants only. For every plant we have information on : its primary 6-digit NAICS code (the codes are consistent over the 20 year period) ; its year of establishment ; its total employment ; whether or not it is an exporter in selected years ; its sales ; the number of non-production and production workers ; and its 6-digit postal code. The latter, in combination with the Postal Code Conversion files (PCCF), allows us to effectively geo-locate the plants by associating them with the geographical coordinate of their postal code centroids.

The survey frame of the ASM has evolved over time. Early in the period, it was relatively stable with, on average, about 32,000 plants per sample year. The sample of plants was restricted to those with total employment (production plus non-production workers) above zero, and plants must have sales in excess of $30,000. Also, aggregate records were excluded. These records represent multiple (typically small) plants without latitudes and longitudes. In 2000, however, the number of plants in the survey increased substantially as the ASM moved from its own frame to Statistics Canada's centralized Business Register, increasing the sample to an average of 53,000 plants. In 2004, however, the number of plants in the frame was once again restricted, with many of the small plants once again excluded, or included in aggregate records. With this in place, the sample returned to near previous levels, averaging about 33,000 plants between 2004 and 2009. The expanded survey scope in the early 2000s had little effect on trends in the CDFS, but there was an effect on the number of industries found to be localized or dispersed (see Table 2.9 in the Appendix). Our econometric analysis deals with the change in the sample frame through the inclusion of year fixed effects.

We also use the ASM to construct controls for the labor market variables, for some natural advantage proxies, and for industry ownership structure variables that we include in the regressions. All variables are constructed by aggregating plant-level data to the industry level.

L-level input-output tables. We use these tables to construct our plant-level proxies for the importance of input and output linkages (see Appendix C and Section 2.3.2 for more details). The *L*-level tables are at a more aggregate level than the 6-digit NAICS level. We break them down to the 6-digit level based on industries' weights in terms of sales.

KLEMS database. This database, which covers the period from 1961 to 2008, contains various industry-level informations useful for constructing proxies for natural advantage (e.g., energy intensity, water usage etc.).

Trucking micro-data. The trucking micro-data comes from Statistics Canada's Trucking Commodity Origin-Destination Survey and from the 'experiment export trade file' produced in 2008 (see Brown and Anderson, 2015, for details). Section 2.3.2 provides details on the methodology used to estimate ad valorem rates by industry and year.

Geographical data. To geolocate firms, we use latitude and longitude data of postal code centroids obtained from Statistics Canada's Postal Code Conversion files (PCCF). These files associate each postal code with different Standard Geographical Classifications (SGC) that are used for reporting census data in Canada. We match firm-level postal code information with geographical coordinates from the PCCF.

Trade data. The industry-level trade data come from Industry Canada and cover the years 1992 to 2009. The dataset reports imports and exports at the NAICS 6-digit level by province and by country of origin and destination. We aggregate the data across provinces and compute the shares of exports and imports that go to or originate from a set of country groups : Asian countries, OECD countries, and NAFTA countries. Since the trade data is available from 1992 on, whereas the KLEMS data is available until 2008, we restrict our sample to the 1992–2008 period in all estimations to maintain comparability of results.

US price indices. We use detailed year-by-year NAICS 6-digit price indices from the NBER-CES Manufacturing Productivity Database (`http://nber.org/data/nberces5809.html`) to construct instruments for Canadian industry-level transportation costs. Methodological details are provided in Sections 2.3.3 and 2.4.3.

B. The distance-based approach to measuring localization

Following Duranton and Overman (2005, 2008), hereafter DO, the estimator of the kernel density (probability density function or PDF) of bilateral distances between plants at a given distance d, is given by :

$$\widehat{K}(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d-d_{ij}}{h}\right), \tag{B.1}$$

where h is Silverman's optimal bandwidth and f is a Gaussian kernel function. The distance d_{ij} (in kilometers) between plants i and j is computed as :

$$d_{ij} = 6378.39 \cdot \text{acos}\left[\cos(|\text{lon}_i - \text{lon}_j|)\cos(\text{lat}_i)\cos(\text{lat}_j) + \sin(\text{lat}_i)\sin(\text{lat}_j)\right]. \tag{B.2}$$

Alternatively, rather than using plant counts as the unit of observation in (B.1), we can characterize the localization of employment or sales at the industry level. This can be accommodated by adding weights to (B.1) :

$$\widehat{K}_W(d) = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(e_i+e_j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (e_i+e_j) f\left(\frac{d-d_{ij}}{h}\right), \tag{B.3}$$

where e_i and e_j are the employment or sales levels of plants i and j, respectively.[30] The weighted K-density thus describes the distribution of bilateral distances between plants weighted by either employees or sales in a given industry, whereas the unweighted K-density describes the distribution of bilateral distances between plants in that industry. When required, as in Table 2.9, we follow Duranton and Overman (2005) and implement a Monte Carlo approach for measuring the statistical significance of localization of industries.

To construct the K-densities, we need to fix a cutoff distance. Following Behrens and Bougna (2014), we choose a cutoff distance of 800 kilometer for computing the K-densities. The interactions across 'neighboring cities' mostly fall into that range in Canada. In particular, a cutoff distance of 800 kilometer includes interactions within the 'western cluster' (Calgary, AB ; Edmonton, AB ; Saskatoon, SK ; and Regina, SK) ; the 'plains cluster' (Winnipeg, MB ; Regina, SK ; Thunder Bay, ON) ; the 'central cluster' (Toronto, ON ; Montréal, QC ; Ottawa, ON ; and Québec, QC) ; and the 'Atlantic cluster' (Halifax, NS ; Fredericton, NB ; and Charlottetown, PE). Setting the cutoff distance to 800 kilometer allows us to account for industrial localization at both very small spatial scales, but also at larger interregional scales for which market-mediated input-output and demand linkages, as well as market size, might matter much more.

[30] Contrary to Duranton and Overman (2005), who use a multiplicative weighting scheme, we use an additive one. The additive scheme gives less weight to pairs of large plants and more weight to pairs of smaller plants than the multiplicative scheme does. Using a multiplicative scheme would imply that our results may be too strongly driven by a few very large firms in a given industry.

While the K-density PDF provides a clear picture of localization at every distance d, and while it allows for statistical testing, it is not well suited in capturing globally the location patterns of industries up to some distance d. This can, however, be achieved by using the K-density cumulative distribution up to distance d. In all our econometric estimations, we use as dependent variable the CDF of the K-densities. Those are given by :

$$\text{CDF}(d) = \sum_{\delta=1}^{d} \widehat{K}(\delta) \quad \text{and} \quad \text{CDF}_W(d) = \sum_{\delta=1}^{d} \widehat{K}_W(\delta). \tag{B.4}$$

Finally, for the purpose of comparision of our results, we also compute the 'raw' unweighted CDFS of the distribution of bilateral distances, which are given by

$$\text{RAW}(d) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \chi(d_{ij} \leq d), \tag{B.5}$$

where n is the number of plants in the industry and where $\chi(\cdot)$ is an indicator function that takes value 1 if the bilateral distance d_{ij} is less than d and zero otherwise. While (B.4) provides a kernel-smoothed distribution, (B.5) provides a raw distribution.

Table 2.1 provides the (unweighted) K-density CDFS in 1990, 1999, and 2009 for the most strongly localized industries in Canada ; while Table 2.2 summarizes the industry-average K-densities across years and using different weighting schemes. Last, Table 2.9 summarizes the year-on-year location patterns of industries based on the formal significance test of Duranton and Overman (2005) that we have described in the foregoing.

C. Input-output shares

We use the L-level national input-output tables from Statistics Canada at buyers' prices. These tables – which constitute the finest sectoral public release – feature 42 sectors that are somewhere in between the NAICS 3- and NAICS 4-digit levels. For each industry, i, we allocate total inputs purchased or outputs sold in the L-level matrix to the corresponding NAICS 6-digit sectors. We allocate total sales to each subsector in proportion to that sector's sales in the total sales to obtain a 257×257 matrix of NAICS 6-digit inputs and outputs, which we use in constructing the linkages.[31] From that table, we compute the share α_{ij} that sector i sells to sector j. We also compute the share β_{ij} that sector i buys from sector j. We systematically exclude within-sector transactions where $i = j$, as those may be capturing all sorts of intra-sectoral agglomeration economies that are conducive to clustering but not correlated with input-output linkages. Thus, the weights we use in equations (E.3) and (E.4) are given by

$$\omega^{\text{in}}_{\Omega(\ell),s} \equiv \alpha_{\Omega(\ell),s} \quad \text{and} \quad \omega^{\text{out}}_{\Omega(\ell),s} \equiv \beta_{\Omega(\ell),s}. \tag{C.1}$$

Using the L-level matrix provides smoother series of input-output linkages than those obtained using the confidential W-level national input-output tables (which are directly in the 257×257 industries format).

[31] Because of confidentiality reasons, we do not use the finer W-level matrices since this would make disclosure of results more problematic. However, the tests we ran using those matrices yield very similar results to the ones we report in this paper.

D. Applying the Lewbel (2012) method

To apply the Lewbel (2012) procedure, we need to verify two conditions : heteroscedasticity and correlation. First, we regress the potentially endogeneous variables (input and output distances, trade shares, and trucking costs) on all other exogeneous variables of the model. We then predict the residuals of that regression and run a standard heteroscedasticity test. We need to reject the homoscedasticity assumption for the Lewbel method to be applicable. In our case, we strongly reject the null hypothesis of homoscedasticity for all series of residuals (the p-value is zero in all tests). Second, we take the square of the predict residuals from the foregoing regression, and check the correlation between the dependent variable of the regression (input distances, or output distances, or the different trade shares, or trucking costs) and those squared residuals. The correlation needs to be 'strong' and statistically strongly significant for the instruments to work properly. In our case, this condition holds true for transportation costs, the input and output distances, and for all import shares : the correlation of the squared residuals with the variable itself is significant at 1% in all cases. It is 0.067 for transportation costs, -0.081 for input distances, -0.089 for output distances, 0.130 for the Asian share of imports, and -0.079 for the NAFTA share of imports. We find no statistically significant correlation for the export shares.

Since the two conditions (heteroscedasticity of the residuals and correlation of the squared residuals with the variable) are met in our case, we can apply the Lewbel estimator. Since fixed effects cannot be included in the estimation (see `ivreg2h` in Stata), we de-mean all variables by industry first. The exogeneous variables are partialled-out for the Lewbel estimator and so their coefficients are not reported. Since we have an exogeneous instrument for transportation costs, we apply the Lewbel estimator only to deal with potential endogeneity concerns of trade shares and input-output distances.

E. Additional tables and results

Table 2.9 summarizes the location patterns by year and by statistical significance following the methodology developed by Duranton and Overman (2005). It contains information on the percentage of industries with random, localized, and dispersed point patterns for all years between 1990 and 2009. Table 2.10 contains robustness checks for the estimation of model (1) using the employment- and sales-weighted K-density CDFS, respectively. It also replicates our main results by averaging all variables over five-year intervals to reduce the volatility of some variables, and to allow slow-changing variables to be better identified. Table 2.11 contains the first-stage estimates for the IV regression, whereas Table 2.12 contains the cross-sectional estimates (both pooled and year-by-year) for transportation costs.

Table 2.9 Percentage of industries with random, localized, and dispersed point patterns, 1990 to 2009.

	Unweighted (plant counts)			Employment weighted			Sales weighted		
Year	Random	Localized	Dispersed	Random	Localized	Dispersed	Random	Localized	Dispersed
1990	52.53	34.63	12.84	52.53	36.96	10.51	54.86	37.35	7.78
1991	51.36	36.19	12.45	52.92	38.52	8.56	55.25	36.19	8.56
1992	53.70	36.19	10.12	56.42	35.02	8.56	58.37	33.46	8.17
1993	53.70	34.24	12.06	58.37	33.46	8.17	59.53	31.52	8.95
1994	49.81	36.96	13.23	57.20	33.07	9.73	60.70	30.74	8.56
1995	55.25	33.46	11.28	58.37	33.07	8.56	59.53	32.30	8.17
1996	54.09	35.41	10.51	56.03	35.41	8.56	59.53	33.46	7.00
1997	55.25	35.41	9.34	60.70	32.30	7.00	61.09	32.68	6.23
1998	55.64	34.24	10.12	58.37	35.02	6.61	61.87	32.68	5.45
1999	55.25	34.63	10.12	58.75	35.41	5.84	61.48	32.30	6.23
2000	47.86	37.74	14.40	51.75	40.47	7.78	53.31	40.47	6.23
2001	43.58	41.25	15.18	52.92	40.86	6.23	50.58	42.41	7.00
2002	45.91	39.69	14.40	50.97	41.63	7.39	54.86	37.35	7.78
2003	47.47	36.58	15.95	50.58	40.86	8.56	55.64	35.41	8.95
2004	60.31	30.35	9.34	60.31	33.07	6.61	60.70	32.30	7.00
2005	58.75	33.46	7.78	62.65	31.13	6.23	64.20	31.52	4.28
2006	60.31	30.35	9.34	60.31	33.46	6.23	62.26	33.85	3.89
2007	57.59	33.46	8.95	60.70	33.85	5.45	62.65	32.30	5.06
2008	56.03	34.24	9.73	61.48	31.91	6.61	64.59	29.96	5.45
2009	59.53	33.07	7.39	63.04	31.52	5.45	63.04	31.13	5.84

Source : Authors' computations using the *Annual Survey of Manufacturers* Longitudinal Microdata file. The statistical significance of the location patterns is computed using Monte Carlo simulations with 1,000 replications following the procedure developped by Duranton and Overman (2005).

Table 2.10 Estimation of specification (E.1) using employment-weighted CDFS, sales-weighted CDFS, and five year averages.

Dependent variable	Employment weighted CDF			Sales weighted CDF			Unweighted CDF, five year averages		
Variables	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km	CDF 10km	CDF 100km	CDF 500km
Total industry employment	0.289^a	0.235^a	0.074^b	0.309^a	0.257^a	0.091^a	0.313^a	0.242^a	0.077^b
	(0.049)	(0.041)	(0.029)	(0.050)	(0.043)	(0.029)	(0.052)	(0.040)	(0.033)
Firm Herfindahl index (employment based)	-0.001	-0.009	0.003	0.023	0.015	0.021	0.011	-0.004	0.003
	(0.028)	(0.025)	(0.020)	(0.029)	(0.025)	(0.020)	(0.037)	(0.027)	(0.022)
Mean plant size	-0.230^a	-0.177^a	-0.044	-0.258^a	-0.199^a	-0.062	-0.286^a	-0.233^a	-0.069
	(0.055)	(0.049)	(0.039)	(0.054)	(0.048)	(0.037)	(0.063)	(0.052)	(0.043)
Share of plants affiliated with multiplant firms	-0.010	-0.133	-0.204^b	0.024	-0.113	-0.199^b	-0.003	-0.115	-0.213^b
	(0.123)	(0.110)	(0.081)	(0.126)	(0.112)	(0.080)	(0.143)	(0.118)	(0.086)
Share of plants controlled by foreign firm	0.233	0.274^b	0.261^a	0.238	0.271^c	0.223^a	0.163	0.256^c	0.264^b
	(0.144)	(0.128)	(0.081)	(0.161)	(0.141)	(0.085)	(0.169)	(0.137)	(0.105)
Natural resource share of inputs	-0.005	0.007	0.003	-0.010	0.003	-0.001	0.027	0.035^b	0.010
	(0.017)	(0.011)	(0.008)	(0.017)	(0.012)	(0.008)	(0.025)	(0.016)	(0.011)
Energy share of inputs	0.058	0.033	0.020	0.048	0.024	0.013	0.045	0.032	0.037
	(0.051)	(0.048)	(0.035)	(0.053)	(0.049)	(0.035)	(0.058)	(0.047)	(0.033)
Share of hours worked by all workers with post-secondary education	0.028	0.041	0.035	0.014	0.023	0.021	-0.214	-0.126	-0.058
	(0.064)	(0.054)	(0.033)	(0.071)	(0.056)	(0.033)	(0.137)	(0.114)	(0.087)
In-house R&D share of sales	0.004	0.019	0.018^b	-0.005	0.011	0.016^c	0.019	0.041^b	0.032^a
	(0.017)	(0.013)	(0.009)	(0.018)	(0.014)	(0.009)	(0.023)	(0.018)	(0.012)
Asian share of imports	-0.684^b	-0.531^b	-0.241^c	-0.713^b	-0.604^b	-0.285^c	-1.463^b	-1.012^a	-0.383^c
	(0.312)	(0.252)	(0.145)	(0.349)	(0.276)	(0.162)	(0.579)	(0.357)	(0.202)
OECD share of imports	-0.377	-0.232	0.008	-0.305	-0.186	0.043	-0.770	-0.351	-0.006
	(0.264)	(0.217)	(0.164)	(0.286)	(0.236)	(0.176)	(0.566)	(0.336)	(0.236)
NAFTA share of imports	-0.312	-0.208	-0.018	-0.262	-0.195	0.003	-0.821	-0.477	-0.104
	(0.244)	(0.198)	(0.141)	(0.276)	(0.226)	(0.159)	(0.518)	(0.317)	(0.201)
Asian share of exports	0.264	0.368	0.065	0.217	0.299	0.082	0.322	0.366	0.051
	(0.483)	(0.389)	(0.130)	(0.507)	(0.398)	(0.106)	(0.539)	(0.439)	(0.211)
OECD share of exports	0.212	0.330	0.181^c	0.349	0.424^c	0.280^a	0.360	0.450	0.266
	(0.295)	(0.210)	(0.094)	(0.288)	(0.216)	(0.096)	(0.386)	(0.314)	(0.191)
NAFTA share of exports	0.111	0.276	0.098	0.190	0.318	0.169^b	0.265	0.442	0.180
	(0.310)	(0.206)	(0.075)	(0.303)	(0.213)	(0.076)	(0.383)	(0.296)	(0.149)
Ad valorem trucking costs (residual)	-0.158^b	-0.150^b	-0.148^a	-0.134^c	-0.127^c	-0.137^a	-0.377^a	-0.361^a	-0.315^a
	(0.077)	(0.072)	(0.053)	(0.076)	(0.070)	(0.045)	(0.085)	(0.076)	(0.060)
Input distance	-0.256^a	-0.238^a	-0.186^a	-0.256^a	-0.239^a	-0.180^a	-0.258^a	-0.246^a	-0.221^a
	(0.063)	(0.054)	(0.032)	(0.064)	(0.056)	(0.033)	(0.073)	(0.059)	(0.043)
Output distance	-0.234^a	-0.222^a	-0.127^a	-0.200^a	-0.193^a	-0.113^a	-0.374^a	-0.383^a	-0.239^a
	(0.053)	(0.048)	(0.030)	(0.056)	(0.048)	(0.029)	(0.069)	(0.062)	(0.044)
Minimum distance	-0.312^a	-0.246^a	-0.119^a	-0.327^a	-0.249^a	-0.131^a	-0.400^a	-0.297^a	-0.141^a
	(0.050)	(0.039)	(0.026)	(0.054)	(0.039)	(0.026)	(0.067)	(0.043)	(0.032)
Number of NAICS industries	257	257	257	257	257	257	257	257	257
Number of years	17	17	17	17	17	17	4	4	4
Industry dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations (NAICS × years)	4,369	4,369	4,369	4,369	4,369	4,369	1,028	1,028	1,028
R^2	0.318	0.371	0.381	0.294	0.359	0.376	0.517	0.599	0.598

Notes : ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. We use simple OLS. Standard errors, given in parentheses, are clustered at the industry level. Our measures of input and output distances are computed using $N = 5$. A constant term is included but not reported.

Table 2.11 First-stage results for the IV regression.

Dependent variable : Ad valorem trucking costs	
Variables	
Total industry employment	0.017
	(0.014)
Firm Herfindahl index (employment based)	0.002
	(0.010)
Mean plant size	0.006
	(0.014)
Share of plants affiliated with multiplant firms	0.026
	(0.039)
Share of plants controlled by foreign firm	0.055
	(0.044)
Natural resource share of inputs	-0.008
	(0.006)
Energy share of inputs	0.084^a
	(0.018)
Share of hours worked by all workers with post-secondary education	-0.057^a
	(0.014)
In-house R&D share of sales	0.024^a
	(0.009)
Asian share of imports	-0.056
	(0.107)
OECD share of imports	0.067
	(0.095)
NAFTA share of imports	0.021
	(0.109)
Asian share of exports	-0.156^c
	(0.089)
OECD share of exports	-0.104
	(0.072)
NAFTA share of exports	-0.065
	(0.069)
Ad valorem trucking costs US (instrument)	0.485^a
	(0.111)
Input distance	0.035^c
	(0.020)
Output distance	-0.011
	(0.015)
Average minimum distance	0.005
	(0.014)
First-stage R^2	0.628
First-stage F test of excluded instruments	19.07

Notes : ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. OLS regression of 'ad valorem trucking cost' on the ad valorem trucking cost US (our instrument) and all control variables. We report the first-stage R^2 and note from the first-stage test that the instrument is strong.

Table 2.12 Cross-sectional estimates, pooled and year-by-year.

Dependent variable : CDF at 50 kilometers			
Pooled cross section		Yearly cross sections (ad valorem trucking costs (residual))	
Asian share of imports	-0.044	1992	-0.128^a
	(0.272)		(0.045)
OECD share of imports	-0.094	1993	-0.116^b
	(0.268)		(0.046)
NAFTA share of imports	-0.062	1994	-0.097^b
	(0.207)		(0.041)
Asian share of exports	0.531	1995	-0.109^b
	(0.552)		(0.043)
OECD share of exports	0.288	1996	-0.090^b
	(0.336)		(0.041)
NAFTA share of exports	0.201	1997	-0.074^c
	(0.248)		(0.040)
Ad valorem trucking costs (residual)	-0.065^b	1998	-0.064
	(0.031)		(0.041)
Input distance	-0.306^a	1999	-0.060
	(0.098)		(0.046)
Output distance	-0.428^a	2000	0.008
	(0.099)		(0.042)
Average minimum distance	-0.380^a	2001	-0.039
	(0.062)		(0.040)
Observations	4,369	2002	-0.038
R^2	0.773		(0.041)
		2003	-0.041
			(0.039)
		2004	-0.043
			(0.047)
		2005	-0.028
			(0.045)
		2006	-0.044
			(0.044)
		2007	-0.062
			(0.040)
		2008	-0.068^c
			(0.036)

Notes : ^{a}, ^{b}, ^{c} denote coefficients significant at the 1%, 5% and 10% levels, respectively. OLS regressions, dependent variables is the CDF at 50 kilometers distance. All specifications include the same controls than in the main text. There are no time fixed effects in the pooled cross section. Huber-White robust standard errors in parentheses.

CHAPITRE III

THE DETERMINANTS OF LOCALIZATION : A CONDITIONAL DISTANCE-BASED APPROACH

Abstract

Do pairs of plants with 'close or similar' input-output linkages, types of workers, and that use or exchange similar technology locate near one another in space ? To answer these questions, I propose a new non-parametric approach to measuring the localization of 'closely related' multiple industries – i.e., a multidimensional way to assess coagglomeration – in continuous space. More precisely, I combine the measurement approach of localization in continuous space with a coagglomeration approach, and then relate them to the degree to which industries share goods, people, and ideas. My results show that plants which belong to manufacturing industries with similar input-output linkages or workforces tend to locate near one another. I find little evidence that plants that share similar technologies cluster geographically.

Keywords : Industrial localization ; Agglomeration ; Manufacturing industries ; Non-parametric statistics ; Conditional kernel density.

JEL classification : R12 ; L60 ; R30 ; R32 ; C140.

"Together these three [input-output links, labor similarity, and technological similarity] explain more of the variation in coagglomeration than does natural advantage, which supports the view that agglomeration economies is a more important determinant of geographic location." (Ellison, Glaeser, and Kerr, 2010, p.1205).

3.1 Introduction

The most striking feature of industrial location patterns is geographical concentration. This has been of interest to economists since 1890 when Alfred Marshall pointed out the stylized fact that some industries tend to cluster geographically whereas others do not. [1] Marshall identified three sources of agglomeration : firms want to be near their customers and suppliers in order to economize on transport costs of goods (goods), to reap the benefits of a thicker labor market (people), and to learn from others and speed their own innovations (ideas). The extent of this concentration of economic activities is surely the reason why interest in agglomeration has grown in recent years. Over the last two decades, clusters have attracted interest from policy makers, academics, economic development practitioners, and development agencies. Many countries and economic development initiatives have built their industrial development strategies on cluster-based models. Despite successful implementation in the US, Brazil, Japan, France, Italy, and Finland, recent economic studies increasingly question the use of cluster policies : there is indeed little evidence that more clustering will have significant effects

[1] Famous examples of industry clusters include information technology firms in Silicon Valley and Boston's Route 128 (Saxenian, 1996), the Manufacturing Belt in the U.S, the Blue Banana' in Europe, industrial districts in Italy (Pyke et al., 1990), Toronto's biopharmaceutical cluster (Martin et al., 2004), advertising firms in Manhattan (Arzaghi and Henderson, 2008), and furniture producers in western North Carolina (Acharya et al., 2009).

on average productivity or wages in manufacturing industries. The starting point to better understand the drivers and implications of cluster-based development is to measure correctly the observed degree of clustering. Consequently, many studies have empirically defined and measured industry localization using different spatial concentration indices. This is relevant from a policy perspective because there is an increasing need for cluster-based data to support research, facilitate comparisons of clusters across regions and support policymakers in defining regional strategies.

Rigorous empirical tests of industrial agglomeration in space depend on the availability of micro-geographic data at a fine spatial scale. In most cases, these data, which enable researchers to determine precise agglomeration patterns, are not widely available to the public and are fairly expensive. Combes et al. (2008) discussed six ideal properties for a spatial concentration index. According to these authors, any test of localization should rely on a measure of spatial concentration which : (i) is comparable across industries ; (ii) is comparable across spatial scales ; (iii) is unbiased with respect to arbitrary changes to a spatial classification ; (iv) is unbiased with respect to arbitrary changes to industrial classification ; (v) is carried out with respect to a well-established benchmark ; and (vi) allows one to determine whether significant differences exist between an observed distribution and this benchmark. In other words, the measure should provide an indication of the significance of the results through a variety of statistical tests. The ideal index of spatial concentration still seems out of reach.

Two main approaches have been followed in the literature. The first treats space as discrete and the second considers space as continuous. The first approach was developed by Ellison and Glaeser (1997), it was followed by Maurel and Sédillot (1999), Brülhart and Traeger (2005), and Mori, Nishikimi, and Smith (2005). Ellison and Glaeser built an index (the EG index) of industrial agglomeration that is comparable across industries with different industrial levels of concentration (i.e., different numbers of plants and different plant size distributions). Their index takes the value of zero when an industry is as concentrated as one would expect to result from a random location process; the index takes a positive value when an industry is more concentrated than what one would expect to occur randomly. Cassey and Smith (2014) recently improved the interpretation of the EG index by simulating confidence intervals that can be used for a statistical test. Another discrete measure is the D-index developed in Mori et al. (2005). This index is a statistical test based on the Kullback-Leibler divergence measure, derived from a discrete-space axiomatic model. Despite several advantages, the indices of Ellison and Glaeser (1997), Maurel and Sédillot (1999), and Mori et al. (2005), rely on a discrete space (i.e., arbitrary spatial units) and hence are vulnerable to the well-known *modifiable areal unit problem* (MAUP).

The MAUP has been first addressed by Openshaw and Taylor (1979), Arbia (1989), and recently by Duranton and Overman (2005, 2008, henceforth DO), and Marcon and Puech (2003, 2010, 2014). Based on the seminal work by Ripley (1976, 1977) who introduced the K function (see Diggle, 1983 and Cressie, 1993), a famous distance-based method widely used in ecology, Duranton and Overman (2005) construct a non-parametric test of localization that uses micro-geographic data and treats space as continuous, thereby effectively eliminating the MAUP. The main idea behind this approach is to determine the distribution of bilateral distances between all pairs of plants in each industry and to compare that distribution to a randomly drawn set of bilateral distances. An industry is localized or dispersed if its distribution of bilateral distances significantly deviates from a series of simulated random draws. This approach has gained increasing acceptance because it derives more reliable results and respects five of the six properties of an ideal measure of concentration (the only exception is the property related to the arbitrary changes to industrial classification).

Although the Duranton and Overman (2005) index respects most of the properties of an ideal measure of concentration, it is still sensitive to industry characteristics (i.e., the index is biased with respect to arbitrary changes in industrial classification). Just as the arbitrary carving up of spatial units leads to the MAUP, by defining a limited number of sectors, an industrial classification may also arbitrarily separate closely related economic activities or reunite activities despite their differences. Haedo and Mouchart (2012) pointed out that when sectors are aggregated, some dispersed sectors are mixed up with the concentrated sectors to provide a 'medium' distribution. Applying the Duranton and Overman (2005) methodology to Canadian data, Behrens and Bougna (2015) documented that 50-60% of the manufacturing industries are localized 'at the NAICS 4-digit level', whereas only 30-40% are localized at NAICS 6-digit. In the case of the United Kingdom (U.K.), Duranton and Overman (2005) document different proportions of localized industries 'at the 4-digit level' (i.e., 52 percent at a 5 percent confidence level) and 3-digit industries (i.e., 58 percent at a 5 percent confidence level). Riedel and Hyun-Ju (2014), and Nakajima et al. (2012) find similar results respectively for Germany and Japan. Concentration levels change because at the NAICS 6-digit level, industries are close in space. This suggest that they are subject to the same agglomerative forces, therefore, they should not be separated within industrial classifications. Clearly, the measured levels of industrial concentration are sensitive to changes in industrial aggregation (or alternatively, industrial classification). My paper partially corrects this by introducing a new conditional test of localization that accounts for both the spatial and technological distances between industries.

The literature on empirically defining and measuring industry localization is growing. However, there are only few rigorous papers about the microfoundations that go beyond assessing Marshall's three forces.[2] The main reason

[2] Exceptions include Duranton and Puga (2001), Strange et al. (2006), Ellison, Glaeser, and Kerr (2010), Rosenthal and Strange (2005, 2010), Strange, Faggio, and Silva (2014), and Behrens, Bougna and Brown (2015).

lies with the 'Marshallian equivalence' (Duranton and Puga, 2004), i.e., all agglomeration mechanisms predict that plants tend to locate near other plants that share similar characteristics. Plants do this for productivity gains, irrespective of the channels through which these gains materialize. Moreover, there are currently few studies on the coagglomeration of industries into business clusters. Ellison and Glaeser (1997) document U.S. coagglomeration patterns. Ellison et al. (2010) use coagglomeration measures to assess the relative importance across industries of potential sources of agglomeration economies. Klier and McMillen (2008) use the Duranton and Overman (2005) index to explain concentrations in the U.S. auto supplier industry. Billings and Johnson (2012) introduce a non-parametric test for industrial specialization. This specialization test refers to the concentration of an industry within a given urban area (Denver-Boulder-Greeley). There is also a literature on industrial complex analysis, that looks at the co-location of plants based on their input-output (and other) relationships. The basic idea is the identification of clusters and complexes, or of groups of industries linked by flows of goods and services, or showing significant mutual locational attraction (Czamanski and Ablas, 1979; Feser and Bergman, 2000; Feser, 2003; Delgado, Porter and Stern, 2015).

My paper adds to this growing literature on industrial agglomeration. I exploit information contained in coagglomeration patterns to construct a non-parametric statistical test of colocalization derived from micro-geographic data. Unlike other studies on coagglomeration – which only look at pair-wise coagglomeration – my 'conditional test' is a modified version of the Duranton and Overman (2005, 2008) test for localization and can be viewed as a non-parametric multidimensional approach to the measurement of coagglomeration. The key idea of my test is to first combine the measurement approach of localization in continuous space with a coagglomeration approach, and then relate them to the degree to which industries share goods, people, and ideas. More precisely, I propose a new non-parametric approach to measuring the localization of 'closely related' multiple industries – i.e., a multidimensional way to assess coagglomeration – in continuous space. My approach allows to measure concentration and explains the relationship between industrial concentration and its determinants in a single framework. Conditional on belonging to industries with similar characteristics (in terms of input-output linkages, types of workers employed, or technology used), l check whether plants are located near one another in space. To do so, I use Marshallian proxies to measure the proximity of plants in some non-geographic space (in order to select a subset of plants with similar characteristics), and I then use a non-parametric estimation method to see if these similar pairs of plants are located close to one another in geographical space. Similarity of industries are measured through Euclidian distances and Pearson correlation coefficients. Since the non-geographic space is built upon Marshallian proxies, my test allows me to gauge non-parametrically their importance. It allows to answer the following questions : Do pairs of plants with 'close or similar' input-output linkages, types of workers employed, and that use or exchange similar technology locate near one another in space ?

My results show that two out of three Marshallian forces find support in coagglomeration patterns. I find that plants which belong to industries with similar input-output linkages are localized at short distances and dispersed at long distances – similar to Rosenthal and Strange (2010). I further show that pairs of large plants are localized at short and intermediate distances, while pairs of small plants are localized at short distances and dispersed at long distances. My results also suggest that large plants co-locate with large plants (Holmes and Stevens, 2014 ; Behrens and Sharunova, 2015). Regarding the role of labor market pooling, I document that plants which employ similar types of workers (in terms of skills and expertise) tend to co-locate near one another in space. Similarly to Ellison et al. (2010), and Behrens and Guillain (2015), I find little evidence that plants that used or share similar technologies, as measured by patent citations, cluster geographically. I also find that input-output linkages play a more important role than labor market pooling in manufacturing location decisions. Last, my results also reveal that industries are, on average, always more colocalized in terms of employment than in terms of plant counts. These results are robust to the choice of the similarity metric, i.e., Euclidean distances or correlation coefficients.

The rest of the paper is organized as follows. Section 3.2 describes the data and variables used to generate my conditional kernel density measures of coagglomeration. Section 3.3 outlines the methodology. Section 3.4 contains my main empirical results and deals with potential heterogeneity in agglomeration benefits across plants and industries (co-location patterns of large and small plants). I provide several robustness checks in Section 3.5. Finally, Section 3.6 serves as the conclusion.

3.2 Data and measurement

3.2.1 Industries, plants, and geographical data

I briefly discuss my data and explain how I construct my Marshallian proxies. I relegate a more detailed description of the data to the appendix – in particular the comparison between the Scott's National All and the Canadian Business Patterns data of Statistics Canada and information on the geographical structure of the census and PCCF data.

Industries and plants : My empirical analysis is based on data from the Scott's National All Business Directories Database. The biannual data covers the period from 2001 to 2013. A plant is considered a manufacturer in the extended sense, if it reports a manufacturing sector (NAICS 31–33) as its primary or secondary sector of activity. The Scott's database contains between 41,000 and 54,000 plants per year, covering 242 concorded NAICS 6-digit manufacturing industries – see the appendix for more details on industry concordances. For every plant or establishment, I have information about : its primary NAICS 6-digit industry code ; up to four 6-digit secondary NAICS codes ; the year of establishment ; its employment ; whether or not the plant is an exporter ; and its 6-digit postal code. Table 3.1 presents the descriptive statistics of the data by province. In 2013, the average plant size by province and/or territory varied between 11.57 in the Canadian Territories (i.e., Northwest Territories, Yukon and Nunavut) and 57.35 in Manitoba. In 2005, it was between 6.24 in the Canadian Territories and 52.25 in Manitoba. Quebec and Ontario concentrated more than 70 of the total employment across the years. The size distribution of manufacturing plants was very skewed towards small establishments across the years. On average, only 15% of plants had more than 50 employees ; the majority of plants employed between 1 and 20 workers. This is consistent with what we know from other countries (Lafourcade and Mion, 2007 ; Holmes and Stevens, 2004).

Table 3.1 Descriptive Statistics of Canada Manufacturing Industries by province : 2001 – 2013.

	2001		**2003**		**2005**		**2007**		**2009**		**2011**		**2013**	
Provinces	Plants	Emp.	Plants	Emp.	Plants	Emp.	Plants	Emp.	Plants	Emp.	Plants	Emp.	Plants	Emp.
Alberta	3,924	36.09	3,674	39.92	3,557	44.46	3,779	48.90	3,722	52.53	3,481	58.29	3,312	50.25
British Columbia	6,165	32.09	5,957	32.19	5,464	33.61	5,328	34.40	5,123	35.01	4,941	34.84	4,440	36.37
Manitoba	1,665	45.95	1,569	46.54	1,515	52.25	1,429	54.98	1,301	57.62	1,265	53.62	1,175	57.35
New Brunswick	1,424	35.60	1,401	37.51	1,286	39.92	1,196	40.09	1,201	39.24	997	38.05	920	43.30
Newfoundland & Lab.	578	43.46	582	42.40	549	44.41	515	47.93	484	42.97	409	44.20	388	46.64
Nova Scotia	1,720	29.54	1,613	32.29	1,563	32.82	1,396	36.49	1,356	34.44	1,143	33.57	1,041	37.10
Ontario	20,518	45.35	22,225	46.61	21,488	45.59	20,704	47.82	20,318	46.63	18,958	45.84	17,189	47.63
Prince Edward Island	331	25.69	306	25.08	331	24.42	310	26.06	286	25.42	236	27.12	225	28.70
Quebec	15,822	44.29	14,930	47.27	14,348	45.44	13,175	46.53	12,914	48.59	11,943	48.01	11,118	50.59
Saskatchewan	1,393	27.70	1,309	27.86	1,343	32.29	1,231	34.30	1,144	36.53	1,139	38.45	1,021	44.59
Territories	–	–	–	–	41	6.24	49	8.29	46	8.50	40	12.63	35	11.57
Total and average	**53,540**	**41.48**	**53,566**	**43.43**	**51,485**	**43.55**	**49,112**	**45.28**	**47,895**	**45.70**	**44,552**	**45.58**	**40,864**	**47.14**

Notes : Emp. is the average plant size by province. Data on Territories are not available in the database for the years 2001 and 2003.

The Scott's database probably constitutes the best alternative to Statistics

Canada's micro-level *Canadian Business Patterns*. Table 3.7 in the appendix provides a comparison between the Scott's National 2001, 2003, 2005, 2007, 2009, 2011, and 2013 databases and Statistics Canada's province-level data from the Canadian Business Patterns (CBP). Considering that the CBP database is close to the universe of manufacturing plants, the coverage of manufacturing plants in the Scott's database is very good. On average, it covers about 83% of the plants across the years. Note also that by using cross-industry correlations, Behrens and Bougna (2015) illustrated that there is no strong industrial bias in the data. This implies that the Scott's database yields geographical results that are comparable to what can be obtained with other datasets like the Annual Survey of Manufacturing (ASM) longitudinal microdata file. [3]

Geographical data : The 6-digit postal codes are useful for geo-locating plants. To this end, I use the latitude and longitude coordinates of the postal code centroids obtained from Statistics Canada's Postal Code Conversion Files (PCCF). These PCCF files associate each postal code with different Standard Geographical Classifications (SGC) that are used by Statistics Canada. I match postal code information with geographical coordinates by using the postal code data for the following year in order to consider the fact that there is a six month delay in the updating of the postal code data. Table 3.8 in the appendix provides more information on the geographical structure of the census and PCCF data.

3.2.2 Data for the Marshallian agglomeration proxies

According to Marshall (1920), firms tend to locate near one another for three reasons : (i) to reduce the costs of obtaining intermediate inputs and shipping

[3] For example, Behrens, Bougna, and Brown (2015), find similar results than Behrens and Bougna (2015), using Statcan and Scott's data, respectively.

goods to downstream customers (*goods*); (ii) to take advantage of workers with similar skills (*people*); and (iii) to speed the flows of ideas or technology across industries (*ideas*). In order to assess the relative importance of these three Marshallian forces (the flows of goods, people, and ideas) across industry pairs, I use data from three different sources. I now describe in detail, the construction of my Marshallian agglomeration proxies.

Firstly, I use Statistics Canada's yearly L-level input-output tables (henceforth I-O), disaggregated to the W-level (NAICS 6-digit level) from 1998 to 2010.[4] In these tables, I am primarily interested in industry inputs (i.e., the value of intermediate goods, services and other factors of production that were used to produce the output). Since the national I-O tables are produced on an annual basis with a 30 months lag from the reference year, I apply a three year lag to this data when matching it to Scott's data. These tables help to build the Euclidian distances and the Pearson correlation metrics in order to capture the similarity of plants in terms of I-O linkages.

Secondly, I use the *Occupational Employment Statistics* (OES) from the U.S. *Bureau of Labor Statistics* (BLS). The BLS occupation tables provide industry level (NAICS 4-digit) employment data for 555 occupations in the manufacturing industries. Most of these data are obtained from employer or establishment surveys. Since OES data span two different Occupational Classifications, all occupational codes are adjusted to reflect changes between the 2000 and 2010 Standard Occupational Classification (SOC). These data are used to build the Euclidian distance

[4] The L-level of the national input-output tables from Statistics Canada is the most detailed sectoral public release level – featuring 42 sectors that are somewhere in between the NAICS 3- and NAICS 4-digit levels – that allows to construct consistent time series of annual data. The W-level is the most detailed level (not publicly released) which represents 300 industries and 727 commodities.

and the Pearson metrics in order to capture the similarity of plants in terms of skills and expertise of their workers. One may raise the problem of using U.S. data. However, occupation data by industries are not publicly available in Canada and I do not think this is a significant problem since the U.S. and Canadian NAICS 4-digit levels are the same. In addition, the U.S. and Canada are structurally and technologically similar, hence I expect no significant differences between industries at the NAICS 4-digit level. A few studies also employed OES data from other countries as instruments for domestic measures (see Ellison et al., 2010).

Thirdly, I use NBER U.S. Patent Citations Data. This database contains information on all patent applications between 1976 and 2006 (3,209,376 patents). I also use all of the citations made to these patents between 1976 and 2006 : 23,650,891 citations. I first use the concordance between the U.S. patent class and the U.S. SIC code provided by Kerr (2008). This link is built upon a mapping correspondence developed by Brian Silverman (2002) and researchers at Statistics Canada. That mapping helps to build the corresponding concordance between the SIC codes and the NAICS codes. I use this patent citations data to build the Euclidian distance and the Pearson correlation metrics in order to capture the similarity of plants in terms of technology.

Data limitations : First, there are two concerns with using patents as a measure of innovation : (i) patents reflect the first stage of innovation, that is, invention ; and (ii) the value of patents is highly skewed. However, patent citations are a good measure of innovation because they are the direct outcome of the invention process, and these data are released at the micro-level and are the most widely used data in empirical approaches. See Carlino and Kerr (2015) for a thorough discussion on the advantages and disadvantages of using patents citations data.

Second, I use industry data to construct my proxies. Specifically, I use NAICS

6-digit level data for input-output linkage proxies, NAICS 4-digit level data labor market pooling proxies, and NAICS 5-digit data level for knowledge spillover proxies. Since by doing so all plants in the same industry have similar I-O linkages, worker skills and expertise, and patent profiles, there is lumpiness in my data in the sense that observations are not similar on a plant-by-plant basis, but only on an industry-by-industry basis (as in Ellison et al., 2010). Hence, my procedure will select plants in terms of proximity of their industries, but not in terms of proximity of their plant-level characteristics (as would be desirable in an ideal world). The lumpiness is then that all plants in two industries will enter my measure at the same time, or none of them. One could get rid of that lumpiness by using plant-level data on detailed input-output links, the detailed composition of the workforce, or the patent output and citation patterns. As should be clear, these data – which are required to compute proximity between pairs of plants at the microlevel – are basically non-existent. Hence, the ideal test remains for now out of reach.

3.3 Estimation methodology

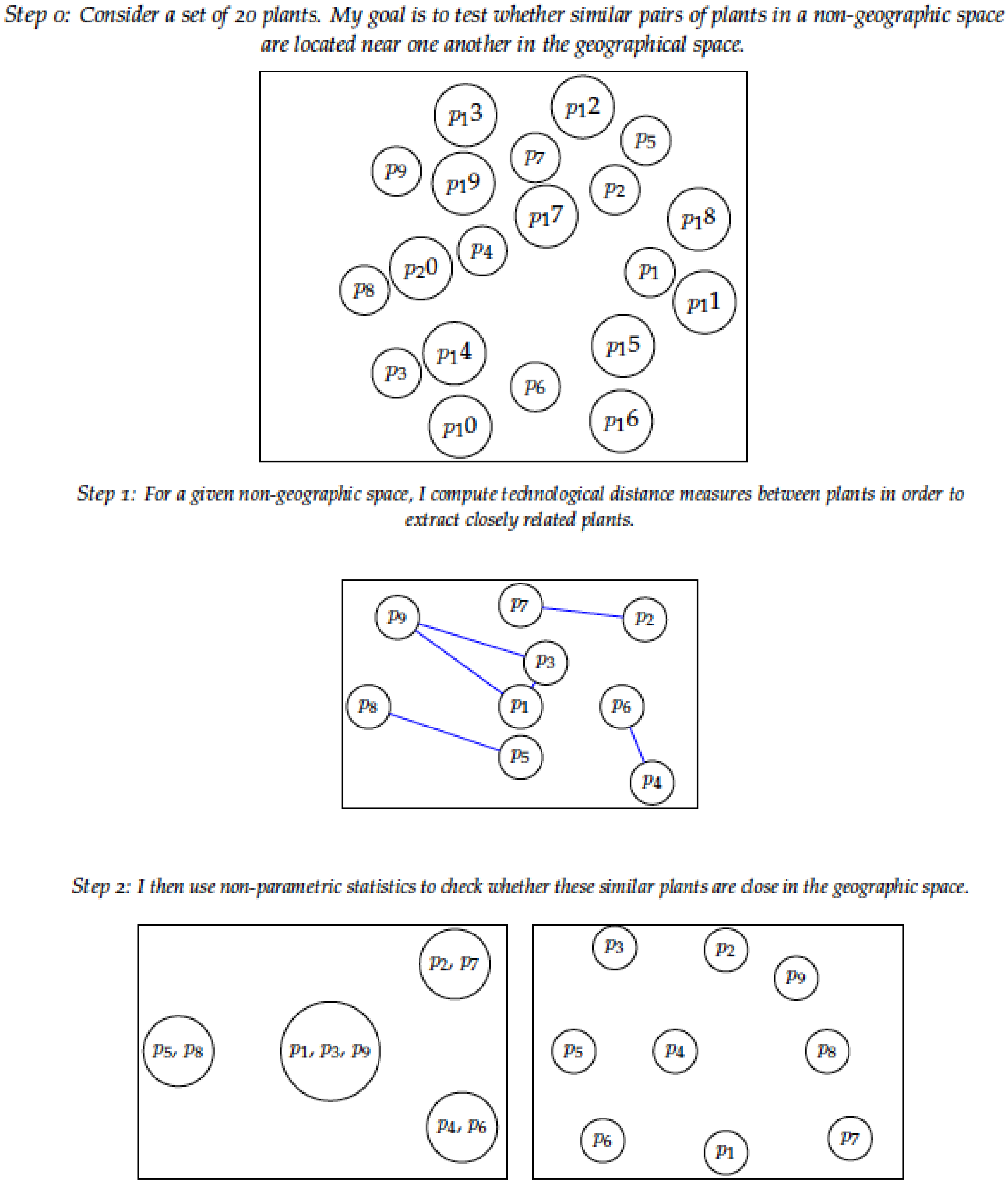
I propose a non-parametric test of localization in continuous space, based on the Duranton and Overman (2005, 2008) test for localization. My 'conditional test' can be viewed as a non-parametric multidimensional approach to the measurement of coagglomeration (Duranton and Overman, 2008, Ellison et al., 2010, and Strange et al., 2014). One drawback of Duranton and Overman (2005) is that their methodology is silent on the potential causes of localization. In my paper, I stratify the sample in such a way that I can better reveal the underlying causes. To accomplish this, I generate conditional kernel density (henceforth CK-density) measures of industry coagglomeration and apply them to Canadian manufacturing plants. The main idea behind my methodology is to *measure the similarity of plants in a non-geographic space* in order to select a subset of plants with similar characteristics in that space. Then, I use a non-parametric approach to test whether these similar pairs of plants are located close to one another in geographical space. Since the non-geographic space is built upon Marshallian proxies, this test allows me to assess the importance of Marshallian forces in shaping the localization of industries (something that the unconditional Duranton and Overman, 2005, test cannot do). My test is a 'conditional test', because I decide to focus only on pairs of plants that belong to industries with significantly similar I-O linkages, worker skills and expertise, and patent profiles. There are on average more than 50,000 plants per year and it would be cumbersome to estimate my measure for all possible pairs of industries (see Scholl and Brenner (2014, 2015) for a discussion of the limitations).

The intuition behind my methodology is best illustrated by means of an example. Let us consider a set of 20 plants denoted by p_1, p_2,..., p_20 (see the top box of Figure 3.3). For each of these plants I have informations on their relationships in a given non-geographic space. I use these informations to compute the bilateral distances between each pairs of plants. My first goal is to measure the similarity of these 20 plants in a given non-geographic space (input-output linkages, labor market pooling, or knowledge spillover) and select a subset of plants with relatively close or similar characteristics. To do so, I use the Euclidian distance (as similarity measure) to compute the 190 unique bilateral distances between my 20 plants in the non-geographic space. I denote these non-geographic distances by g_{i-j}. For a given threshold distance g, two plants i and j are relatively similar if their bilateral distances $g_{i-j} < g$. Let us assume that the following bilateral distances satisfies this condition : g_{1-3}, g_{1-9}, g_{3-9}, g_{5-8}, g_{4-6}, and g_{2-7}. Thus, out of my 20 plants, 9 plants denoted by p_1, p_2,..., and p_9 are related – either by input-output linkages, or by similar types of workers, or by similar technology – in a non-geographic space (see the middle box of Figure 3.3). Since my unit of observation is the bilateral distances, only 6 pairs of distances will enter my estimations. My second goal is to see if plants with relatively similar characteristics in the non-geographic space are locate near one another in the geographic space. Two possible location patterns are depicted in Figure 3.3. The bottom-left box of Figure 3.3 illustrates a pattern where my 9 plants with similar characteristics in the non-geographic space are localized in the geographic space, while the bottom-right box of Figure 3.3 illustrates a pattern where they are not localized.

My approach allows to alleviate the problem related to the changes in industrial classification since the non-geographic space is built using the Marshallian proxies. It is, therefore, possible to measure the technological distance between plants in that non-geographic space. However, its implementation require information on Marshallian characteristics at the plant level. Unfortunately, data for these information are just available at the industry level. Thus, I select plants in terms of proximity of their industry characteristics. By doing so, I assume that all plants in the same industry have similar Marshallian characteristics : I-O linkages, worker skills and expertise, and patent profiles. Hence, my test remains somewhat sensitive to changes in industrial classifications. However, this industrial lumpiness is attenuated by the within industry variation observed.

Conceptually, my test has five steps. In the first step, I design the similarity space. In the second step, I define a preselection procedure. In the third step, I compute the CK-densities of the bilateral distances between all pairs of establishments with similar characteristics. In the fourth step, I compute counterfactuals : the same number of establishments are randomly reallocated across existing manufacturing sites. In the last step, I construct local confidence bands and global confidence bands. These allow for the comparison between the actual distribution and the counterfactuals in order to assess the significance of departures from randomness. I now describe these five steps in greater details.

Figure 3.1 Similarity of plants (non-geographic space) and location patterns (geographic space).



Step 0: Consider a set of 20 plants. My goal is to test whether similar pairs of plants in a non-geographic space are located near one another in the geographical space.

Step 1: For a given non-geographic space, I compute technological distance measures between plants in order to extract closely related plants.

Step 2: I then use non-parametric statistics to check whether these similar plants are close in the geographic space.

Notes : Similarity of Industries in non-geographic space (middle-panel) and two different location patterns : Concentrated (bottom-left panel) and non-concentrated (bottom-right panel).

Step 1 : Design of the similarity space

The non-geographic similarity space is built upon proxies for 'Marshall's trinity' : input sharing, labor market pooling, and knowledge spillovers. I use the Canadian I-O tables to measure the extent to which industries buy and sell intermediate inputs from one another. I have 242 concorded NAICS 6-digit manufacturing industries.[5] I use the 242 × 242 I-O square matrix, to compute the shares IO_{ij} and IO_{ji} of inputs that each industry buys from others, as fractions of their total intermediate inputs. I will later use these shares to build my similarity measures. I consider only the input relationships i.e., I use the column standardization.[6] I apply a three-year lag for the I-O tables to match with Scott's data (e.g., the 2001 Scott's data is match with the 1998 input-output table and the 2013 Scott's data with the 2010 input-output table).

Finding a proxy for labor market pooling is one of the most difficult tasks.[7] In order to assess the importance of labor market pooling as a micro-foundation of agglomeration, I use the occupational tables from the BLS to measure the extent to which sectors that use the same types of workers are located near one another.

[5] Canadian manufacturing industries are classified into 259 to 260 NAICS industries, depending on the classification year. My data span four different industrial classifications : NAICS 1997, NAICS 2002, NAICS 2007, and NAICS 2012. I have concorded those classification to a stable set of 242 manufacturing industries.

[6] Another possibility is to use output relationships in order to look at the extent to which industries sell intermediate outputs. As a robustness check, I ran some estimations with output relations and my results remain fairly similar. This is consistent with Ellison et al. (2010) who show that the input and output coefficients remain similar in magnitude.

[7] The problem in proxying for the importance of pooling in an industry is that it is difficult to identify industry characteristics that are related to the specialization of the industry's labor force (Rosenthal and Strange, 2001, p. 204).

More precisely, for a given pair of industries i and j and a given occupation o, I compute the shares of employees of occupation o in total employment in industries i and j, respectively (LP_{oi} and LP_{oj}). I will use these shares to assess how similar industries are in terms of labor requirements.

Marshall (1920) argued that firms tend to locate where they are likely to learn from other firms. However, it is difficult to observe and to measure patterns of knowledge spillovers and to assess them empirically – see Dumais et al. 2002; Ellison and Glaeser, 1999; and Carlino and Kerr, 2015, for a recent survey. I use the NBER patent citations data to measure the extent to which industries use or exchange similar technologies i.e., patents from industry i cite patents from industry j, and vice versa. Using patent citations data, I build a square matrix that contains either the number or the shares of citations that a patent in sector i is receiving from patents in sector j (KS_{ij}) and the number or the shares of patents in sector j that a patent from sector i is citing (KS_{ji}). Following Ellison et al. (2010), my citation measure is a proxy for the importance of exchanging technology (ideas), rather than a proxy for all forms of intellectual spillovers which are hardly identifiable.[8] In addition, it is hard to dissociate labor mobility from knowledge spillovers. I use patent citation flows that cover the period 1976 through 2006. I match this data with my 2001, 2003, 2005, and 2007 samples. NBER data contain the flows of citations made and received between 1976 and 2006, so I cannot use that data for later years.

[8] Even if many authors employ patent citations to assess intellectual spillovers, it remains that they are an imperfect measure of intellectual spillovers – see Jaffe, Trajtenberg, and Henderson (1993), Jaffe, Trajtenberg, and Fogarty (2000), and Thompson and Fox-Kean (2005) for more details.

Step 2 : Preselection procedure

I use the Marshallian proxies built in step 1 to precompute the empirical distances between all industry-pairs in non-geographic space. It is important to note that my unit of observation is the pair of industries, more precisely, the bilateral distances between the pairs of industries. Ideally, I should use the universe of plants to preselect pairs of plants. However, I need to build a 'relevant subset of plants' (restricted sample) with relatively similar characteristics to run my test for the determinants of localization in continuous space. The importance of this relevant subset of plants is two-fold. First, it allows to select only plants with relatively similar industrial characteristics in my first step (and so avoids the inclusion of many dissimilar pairs of plants in my test, since my goal is to look at the location pattern of pairs of plants with relatively similar industrial characteristics). Second, it proved computationally infeasible to work with the distribution of bilateral distances between all 50,000 plants (see Scholl and Brenner, 2015 for a thorough discussion).

Let me define Ω_t as the universe of plants in year t, $\Omega_{d,t}$ the relevant subset of plants with relatively similar characteristics, and $g_{i,j}$ the Euclidian distance between industries i and j in non-geographic space (with $i \neq j$). For each year t, and for a given threshold distance in non-geographic space g, I impose the following two restrictions for plant selection :

— $\Omega_{g,t} = \{(i,j) \in \Omega \times \Omega$, such as $0 < g_t(i,j) < g\}$; and

— $\Omega_{g,t}$ >= 0.1×N_t , where $N_t = \mid \Omega_t \mid$.

Formally,

$$g_{i,j} = \begin{cases} \sqrt{\sum_{l=1}^{k}(IO_{il} - IO_{jl})^2} & \text{in the case of I-O linkages (NAICS 6-digit, k = 242)} \\ \sqrt{\sum_{l=1}^{k}(LP_{il} - LP_{jl})^2} & \text{case of labor market pooling (k = 555 occupations)} \\ \sqrt{\sum_{l=1}^{k}(KS_{il} - KS_{jl})^2} & \text{case of knowledge spillovers (NAICS 5-digit, k = 180)} \end{cases}$$

In the case of labor market pooling, I have employment data at the NAICS 6-digit level for 555 occupations, therefore, k = 555 occupations. As a robustness check, I compute the Pearson correlation coefficient (ρ) as measure of industries similarity following Glaeser and Kerr (2009), Ellison et al., (2010), and Strange et al. (2014). I exclude all pairs of plants within the same NAICS industry in my computation i.e., I systematically set to zero the own industry elements as those may capture all sorts of intra-sectoral agglomeration forces that push toward clustering but are not correlated with the input-output linkages, labor market pooling, or the knowledge spillovers (e.g., a cluster policy promotion). [9] I then use these precomputed distributions of distances to define a selection criterion for industries with similar characteristics. As stated above, I use two criteria to generate my relevant subsets of plants with relatively 'close or similar' industry characteristics :

— the threshold selection distance g between pairs of industries should allow the selection of plants that belong to industries that are relatively similar in non-geographic space ;

— the threshold selection distance g should also allow to select at least 10% of the universe of plants each year.

I apply these two criteria to the universe of plants to obtain a restricted

[9] Duranton and Overman (2008) pointed out that the colocalization test may fail despite strong forces pushing toward colocalization if own industry concentration forces dominate.

sample of plants with similar characteristics (see Table 3.3). If the non-geographic threshold distance $g = \infty$, then my restricted sample of plants equals the universe i.e., all plants will enter my test and therefore I will always find a random location pattern. If $g = 0$, the restricted sample will be empty. However, setting a small threshold distance will lead to a sample of plants with relatively similar characteristics. The trade off between sample size and similarity is not trivial.

My strategy for the choice of the threshold distance g is to loop over the non-geographical distance distributions and to choose the smallest threshold distance that allows to fulfill the second requirement. Table 3.3 summarizes my preselected subsets of plants and the average distance in non-geographic (similarity) space across years.

Table 3.2 Preselection sample and average distance in the similarity space across years.

Year	Universe	I-O links			Labor			Knowledge		
		Sample	Avg.	%	Sample	Avg.	%	Sample	Avg.	%
2001	53,540	6,101	0.260	11.4	7,097	0.674	13.3	8,513	0.617	15.9
2003	53,566	5,674	0.246	10.6	5,691	0.614	10.6	10,819	0,582	20,2
2005	51,485	5,661	0.236	11.0	5,222	0.569	10.2	10,496	0.585	20.4
2007	49,112	4,923	0.254	10.0	7,734	0.590	15.8	10,011	0.582	20.4
2009	47,896	5,393	0.261	11.3	5,087	0.610	10.62			
2011	44,552	4,547	0.271	10.2	5,149	0.618	11.6			
2013	40,864	5,371	0.246	13.1	5,830	0.657	14.3			

Notes : 0.01, 0.2, 0.2 correspond to the threshold distance between industries set in the input-output linkages, labor market pooling and knowledge spillovers spaces respectively. Patent citations data flows covers the period 1976-2006, this explain why there is no sample information in 2009, 2011, and 2013 for knowledge spillovers.

The main advantage of my conditional procedure is that it allows for the construction of CK-density measures of industry coagglomeration between industries with relatively 'close or similar' characteristics. This is the same idea than coagglomeration, but while coagglomeration is limited to two industries – Ellison and Glaeser (1997), Duranton and Overman (2005, 2008), Ellison et al. (2010), Strange et al. (2014) – my approach allows to compute these measures using more than two industries. In addition, my approach to measure the proximity of plants (Euclidian distance) in the non-geographic space is different. Ellison et al. (2010), and Strange et al. (2015) define the maximum between two industries input shares to assess the importance of I-O linkages ($\max IO_{ij}, IO_{ji}$), and the correlation between industries shares of employees of a given occupation in the total employment in two industries ($LaborCorrelation_{i,j}$) to measure the similarity of employments in industries i and j.

Step 3 : Estimating the conditional K-densities of industries

As explained previously, I follow the methodology proposed by Duranton and Overman (2005, 2008). The main idea is to determine the distribution of the bilateral distances between plants with relatively similar characteristics and to compare this distribution to a randomly drawn set of bilateral distances.

Let me denote the geographical distance between plants i and j by d_{ij}. The *unconditional* estimator of the density of the bilateral distances at any distance d is given by equation (E.1). The Duranton and Overman (2005) estimator is also conditional, since it is computed conditional on the plants being in one industry.

$$\widehat{K}(d|\Omega_N) = \frac{1}{N(N-1)h} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} f\left(\frac{d-d_{ij}}{h}\right) \tag{E.1}$$

This is the Duranton and Overman (2005) kernel density estimator, where :

— Ω_N is the sampling universe where firms selection occurs.

— N is the number of plants, N = | Ω_N | ;

— f is a (gaussian) kernel function ;

— h is Silverman's (1982) optimal bandwidth i.e., the smoothing parameter ;

— d_{ij} is the great circle distance (in kilometers) between plants i and j.[10]

For a sample of plants that fulfill my two selection criteria, *the conditional kernel density estimator* (CK-density) is defined by

$$\widehat{K}(d|\Omega_{d,t}) = \frac{1}{N_t(N_t-1)h} \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} f\left(\frac{d-d_{ij}}{h}\right) \tag{E.3}$$

All CK-densities are computed using a Gaussian kernel with a bandwidth set according to the recommendations in Silverman (1986).[11]

Step 4 : Constructing counterfactuals

The key question in this step is to assess, for every given distance, to what extent the location patterns of industries with similar industrial characteristics depart significantly from randomness. Following Duranton and Overman (2005, 2008), I compute counterfactuals of the conditional kernel density estimates. I then compare these counterfactuals with the actual conditional kernel density determined in step three. Since I sample from the overall population of existing manufacturing plants, by doing this, I am implicitly controlling for the overall tendency of economic activity to agglomerate.

Basically, I randomly draw as many plants as the relevant subset of plants with similar characteristics had and assign each of them to one of any possible

[10] The Great circle distance (in kilometers) between plants i and j is given by the formula :

$$d_{ij} = 6378.39 * \text{acos}\left[\cos(|\text{lon}_i - \text{lon}_j|)\cos(\text{lat}_i)\cos(\text{lat}_j) + \sin(\text{lat}_i)\sin(\text{lat}_j)\right] \tag{E.2}$$

[11] See Silverman (1986) for details concerning the choice of the kernel function.

locations where I observe manufacturing firms. As a robustness check, and following Duranton and Overman's (2008) coagglomeration measures, I also restrict my counterfactual universe by assuming that my hypothetical similar industries randomly choose their locations in the existing locations where I observe industries with similar characteristics. I then compute – conditional on a distance d – the distribution of the hypothetical sample of pairs of plants and estimate the conditional kernel density of the bilateral distances. Finally, I repeat the first and second steps a thousand times. This yields a set of 1,000 estimated values for each distance.

Step 5 : Constructing local and global confidence bands

For each relevant subset of plants that are similar in non-geographic space, I test the statistical significance of their departure from randomness. In order to make a statement about the statistical departure of the localization pattern from randomness, I compute local and global confidence bands, as in Duranton and Overman (2005). To do so, I use the simulated counterfactual distributions from the previous steps to construct two sided confidence intervals that contain 90% of these estimates. The upper bound of this interval is given by the 95 percentile of the generated values ; the lower bound is given by the 5 percentile of the generated values. This procedure generates two smooth curves. Hence, any deviation from randomness can be concluded as indicating localization or dispersion.

Local confidence bands : For each distance d between 0 and 800 kilometers,[12] and conditional on a predetermined cutoff distance (in the non-geographic

[12] Duranton and Overman (2005) consider a threshold distance of 180 kilometers for the United Kingdom, which refers to the median plant to plant distance in their sample. The median plant to plant distance is much larger for Canada. See Behrens and Bougna (2015) for details concerning the choice of the cutoff distance of 800 kilometers in Canada.

space), if the distribution of the distances between the pairs of plants observed after the smoothing procedure exceeds the upper bound of the confidence bands, the selected pairs of plants are said to be *locally concentrated* at distance *d* with a confidence level of 95%. In other words, the location patterns of plants that use the same types of workers, or share inputs and technology is significantly different from a purely random process in space (i.e., pairs of plants with similar types of workers, similar input-output linkages or that use or exchange technology tend to locate near one another). If the distribution of distances between plants is smaller than the lower limit, the selected pairs of plants are said to be *locally dispersed* at the distance under consideration.

Global confidence bands : For each sample of pairs of plants, the previous intervals only allow to make a local statement (i.e., at a given distance) about the departures from randomness. What about the global location patterns of the conditional distribution ? The key point here is to find out which local upper and lower bounds would include 90% of the estimated values across all distances. This requirement will allow all statements to be valid for the overall location pattern. Thus, *global localization* is detected when the CK-density of one particular conditional distribution lies above its upper confidence band and *global dispersion* occurs when the CK-density lies below the lower confidence band and never exceeds the upper confidence band. These bands contain 90 percent of the counterfactual distributions. When the observed distribution lies within them, we cannot reject, at the 5 percent level, the null hypothesis that the observed location pattern of pairs of plants with similar characteristics is one of spatial randomness. If the observed distribution lies above the upper bound of the confidence bands, the distances between plants are over-represented, as compared to spatial randomness, which

Ellison et al. (2010) choose a cutoff distance between 100 and 1,000 miles (around 161 and 1610 km) in the U.S. case.

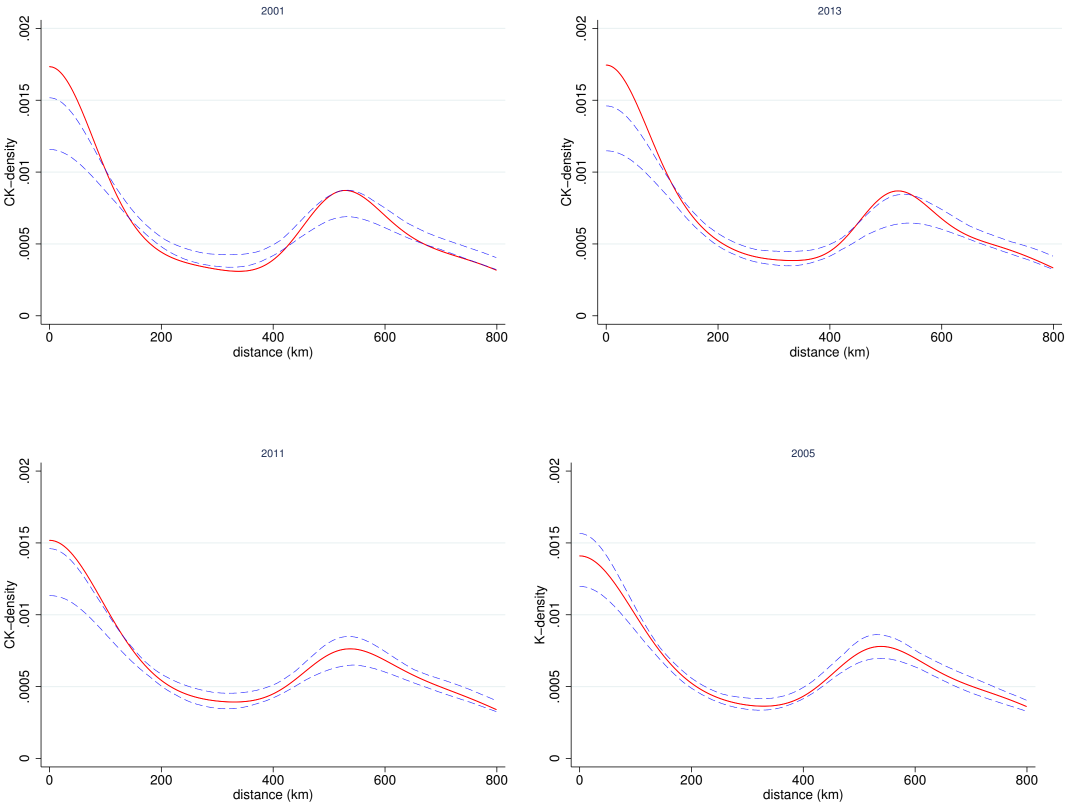
is interpreted as *localization*. Whereas when the observed distribution lies below the lower bound of the confidence band, the distances between plants are under-represented, as compared to spatial randomness, which is interpreted as *dispersion* (see Duranton and Overman, 2005, 2008 and Behrens and Bougna, 2015).

Interpretation and examples

I provide four examples of possible localization patterns in Figure 3.2 to explain what my conditional kernel density measures of coagglomeration capture. The observed distribution of distances in the sample of pairs of plants with similar characteristics is depicted by the solid line (CK-density). The dotted lines depict the global confidence bands.

Figure 3.2 illustrates four different geographical patterns. The top-left panel represents a location pattern where plants with similar I-O linkages are localized at short distances and dispersed at intermediate distances. The distribution of these pairs of plants illustrates a high density for distances between zero and approximately 150 km. In the top-right panel, we observe location pattern with two peaks in the distance density. These are pairs of plants with similar I-O linkages that are localized at short and intermediate distances (i.e., around 500 km), which corresponds to the distance between the two main urban centers in Canada : Montréal and Toronto. The bottom-right panel represents the location patterns of pairs of plants that use or exchange similar technology – as proxied by patent citations. This pattern is not different from one that would arise if location was random. The bottom-left panel shows the location patterns of pairs of plants with similar types of worker that are localized at short distances and random at long distances. All these location patterns illustrate the importance for plants to located near one another in order to reduce the costs of obtaining their intermediate inputs. This result on the role of I-O linkages is in line with Marshall (1920) : when inputs are far away from the market, firms will trade off the distance between customers and suppliers based on the costs of moving inputs and finished goods.

Figure 3.2 CK-density and global confidence bands of select industries with similar characteristics.



Notes : Estimations are based on : 11.3% of the universe of plants in 2001, 13.0% in 2013, 11.6% in 2011 and 13.0% in 2005.

3.4 Industrial colocalization patterns in Canadian manufacturing

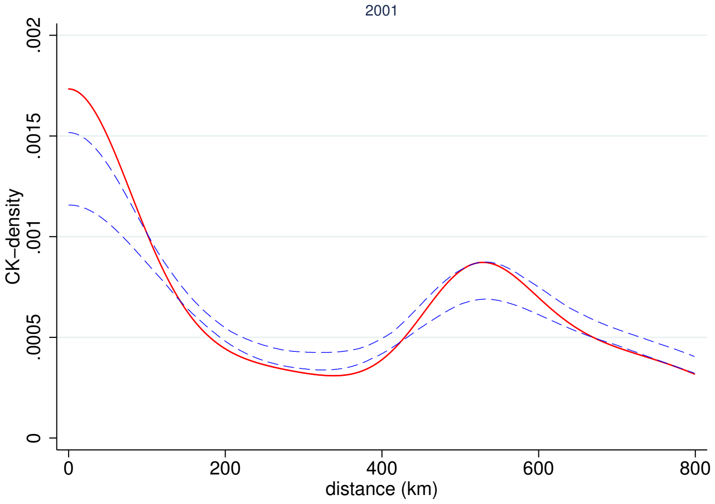
I now use equation (E.3) to look at the location patterns of plants with similar characteristics. I first consider input sharing as my non-geographic space and analyze whether pairs of plants with similar I-O linkages are located near one another in geographic space. I then look at the location patterns of plants with relatively similar types of workers and ask whether they are located near one another or not. I finally look at the location patterns of plants that use or exchange similar technology (similar patent profiles) and ask whether they are located near one another.

3.4.1 Location patterns of plants with similar input-output linkages

I first use the I-O tables to compute the bilateral distances between industries. To fulfill my two sampling requirements – choice of a threshold distance that allows for (i) the selection of plants that belong to industries that are relatively close in non-geographic space ; and (ii) to capture more than 10% of the universe of plants – I fix a threshold distance that allows me to select between 4,547 (10.2%) plants in 2011 and 6,101 (11.3%) plants in 2001 for my relevant subsets of plants. I then use these relevant subsets of plants with similar I-O characteristics to estimate my CK-density measures of localization. I later fix a more restrictive selection distance to see if my results continue to hold. As stated previously and for reasons of simplicity, only global confidence bands are reported.

My results illustrate that plants tend to reduce the costs of obtaining intermediate inputs and of shipping goods. Figure 3.3 shows the CK-density of pairs of plants with similar upstream-downstream linkages. As can be seen, they are located near one another at short distances in 2001 (less than 150 km). This result is in line with the findings in Ellison and Glaeser (1999), Ellison et al. (2010) and Strange et al. (2014), who document that the I-O factor is an important determinant of geographic location. My results also illustrate that plants with similar input-output characteristics are dispersed at intermediate (between 200 and 400 km) and at long distances (beyond 700 km). Figure 3.4 shows that the observed location patterns in 2001 also consistently hold for other years, except in 2013 where we observe a second peak. This second peak corresponds to pairs of plants with similar I-O linkages that are localized at intermediate distances (around 500 km), which roughly corresponds to the distance between Montréal and Toronto.

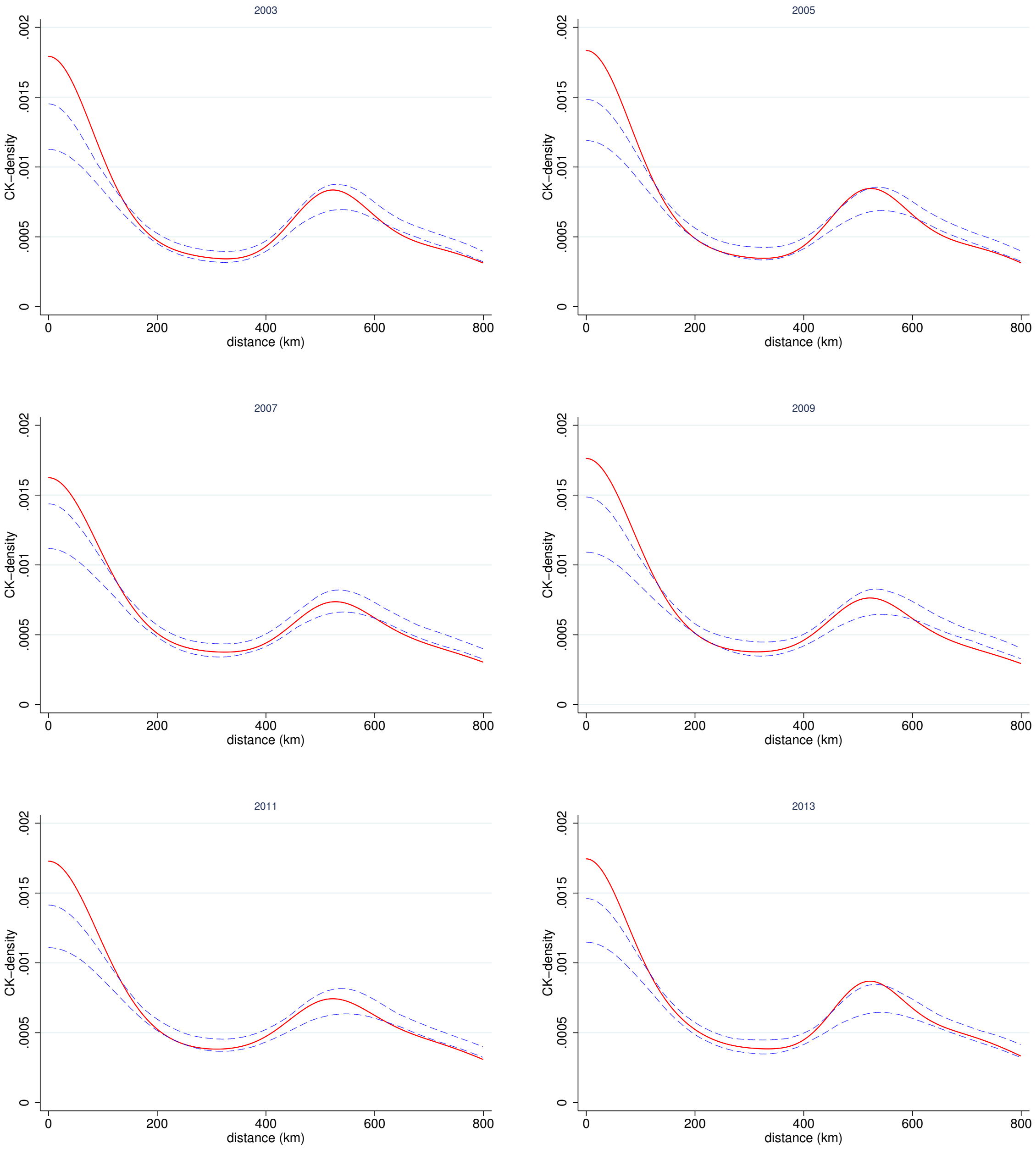
Figure 3.3 Location patterns of plants with similar I-O linkages in 2001.



Notes : Estimations are based on the relevant subset of plants that account for 11.3% of the universe of plants.

Despite the restrictions applied to the selection of industries, my relevant subsets of plants with similar input-output linkages have a good coverage of manufacturing sectors. As can be seen from Table 3.9 in the Appendix, my sample covers 28% of the 242 NAICS 6-digit industries in the strict definition (the plant reports the manufacturing sector as its primary sector of activity).

Figure 3.4 Location patterns of plants with similar I-O linkages.



Notes : Estimations are based on the relevant subsets of plants that represent : 10.3% of the universe of plants in 2003, 10.8% in 2005, 10.0% in 2007, 11.2% in 2009, 10.2% in 2011, and 13.0% in 2013.

To get an idea of the composition of the relevant subsets of plants, I now look at its industrial composition. My unit of observation is the bilateral distance between the plants. In 2001, my relevant subset of plants with similar I-O linkages contains 921,265 unique bilateral distances, where 'Cutlery and Hand Tools' (332210) and 'Metal Valve Manufacturing' (332910) are the most represented industries, with 22.14% of the bilateral distances. In 2007 and 2013, the most represented industries are related to the printing sector. The two most frequently coagglomerated industries are 'Quick Printing (323114) and 'Commercial Screen Printing' (323113), with 28.4% of bilateral distances in 2007 and 33.1% in 2013. Table 3.4.1 illustrates that 6-digit industries mostly source their intermediate inputs within their NAICS 3-digit sectors. This result is in line with Ellison and Glaeser (1999) and Ellison et al. (2010), where the two highest pairwise coagglomerated industries are within the same two digit SIC. More information on the most frequently co-localized industries in 2001, 2007, and 2013 is provided in Table 3.4.1 where I report the two most frequently co-localized industries (in columns) and industries with which they have similar input-output relationship (in lines).

3.4.2 Location patterns of plants using a similar workforce

Alfred Marshall's (1920) ideas about labor market pooling suggest that "employers locate around workers with the skills which they require" and workers seek out places "where there are many employers who need such skill as their" (Marshall, 1920 p. 225). In order to assess the importance of labor market pooling, I use the *Occupational Employment Statistics* (OES) of the *U.S. Bureau of Labor Statistics* (BLS). To fulfill my two sampling requirement – choice of a threshold distance that allows for (i) the selection of plants that belong to industries that are similar in non-geographic space; and (ii) to capture more than 10% of the universe of plants – I fix a threshold distance that allows me to select between 5,087 (10.6% of the universe) plants in 2009 and 7,097 (13.3%) plants in 2001 for my relevant

Table 3.3 Most frequently co-localized industries in 2001, 2007, and 2013 : Similarity in I-O space

Two Most Co-Localized Industries in 2001

	332210 : Cutlery and hand tool mfg	13.57%			332910 : Metal valve manufacturing	6.67%	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
332910	Metal valve manufacturing	34,235	27.39	332210	Cutlery and hand tool mfg	25,219	41.05
332720	Turned product and screw, nut mfg	19,951	15.96	332311	Prefabricated metal building and comp.	8,892	14.48
332311	Prefabricated metal building and comp.	17,363	13.89	332420	Metal tank (heavy gauge) mfg	7,777	12.66
332420	Metal tank (heavy gauge) mfg	15,112	12.09	332439	Other metal container mfg	4,988	8.12
332439	Other metal container mfg	9,700	7.76	332611	Spring (heavy gauge) mfg	4,388	7.14
	Others industries	28,616	22.90		Others industries	10,165	16.55
	Total	124,977	100.00		Total	61,429	100.00

Two Most Co-Localized Industries in 2007

	323114 : Quick printing	14.18%			323113 : Commercial screen printing	14.17%	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
323113	Commercial screen printing	49,806	55.47	323114	Quick printing	48,287	53.82
323115	Digital printing	35,854	39.93	323115	Digital printing	37,879	42.22
323116	Manifold business forms printing	4,137	4.61	323116	Manifold business forms printing	3,550	3.96
	Total	89,797	100.00		Total	89,716	100.00

Two Most Co-Localized Industries in 2013

	323114 : Quick printing	16.6%			323120 : Support activities for printing	16.6%	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
323113	Commercial screen printing	28,715	21.47	323115	Digital printing	66,986	50.09
323115	Digital printing	65,801	49.20	323114	Quick printing	33,029	24.70
323116	Manifold business forms printing	2,810	2.10	323113	Commercial screen printing	30,991	23.17
323120	Support activities for printing	36,407	27.22	323116	Manifold business forms printing	2,723	2.04
	Total	133,733	100.00		Total	133,729	100.00

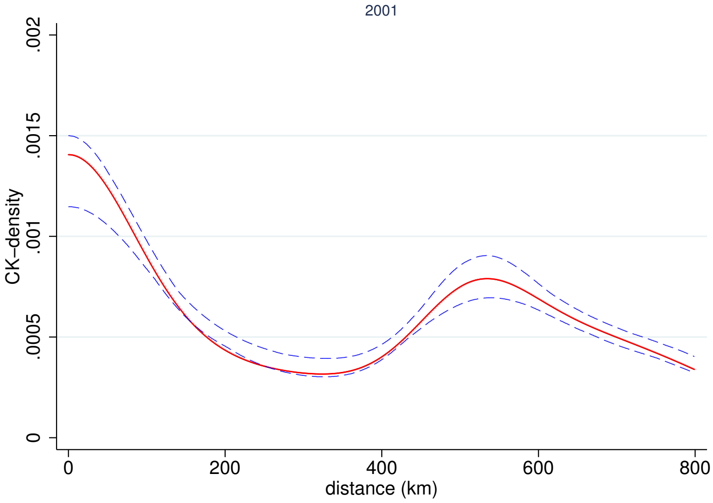
Notes : I present only the two most co-localized industries which represent more than 20% of the total bilateral distances across years : 20.2% in 2001 ; 28.4% in 2007 and 33.1% in 2013. These two industries are displayed in columns, while their related industries are displayed in lines.

subsets of plants with a similar workforce. I then use these subsets to estimate my CK-density measure of coagglomeration. Like in the previous case, I will also fix a more restrictive selection distance to see if my results continue to hold.

My results show that in 2001, at both short and long distance, the location pattern of pairs of plants using a similar workforce was not significantly different from one that would be obtained by a purely random location process (see Figure 3.5). This result also holds in 2003, 2007, and 2009. In 2001, these plants are significantly dispersed at intermediate distances (200 km). However, in 2005, 2011, and 2013, I find that manufacturing plants tend to take advantage of groups of workers with similar skills and expertise. As can be seen from the top right and the bottom left-panel of Figure 3.6, pairs of plants that use workers with similar skills and expertise are located near one another at short distances in 2005, 2011, and 2013. This result is reminiscent of the findings in Ellison and Glaeser (1999), Ellison et al. (2010), and Strange et al. (2014). However, at intermediate and long distances, the location patterns are not significantly different from those that would be obtained by a purely random location process. This last result is in line with the findings by Ellison et al. (2010), who document that labor market pooling is important at a small spatial scale, but has much less of an effect when we look at coagglomeration at a broader geographic scale. My result is also reminiscent of that by Kolko (2010), who shows that labor market effects are larger for either zip codes or counties, and that an industry benefits from labor market pooling as long as it is agglomerated within a state.

As can be seen from Table 3.9 in the Appendix, the coverage of industries within my relevant subsets of plants is also good in the labor case. Across years, I cover 29% of the 86 NAICS 4-digit industries using the strict definition. Like in the I-O case, I now focus my analysis on industries with the highest frequencies to investigate the most co-localized industries. In 2005, my sample contains 505,282 unique bilateral distances, where 'Power, distribution and specialty transformers'

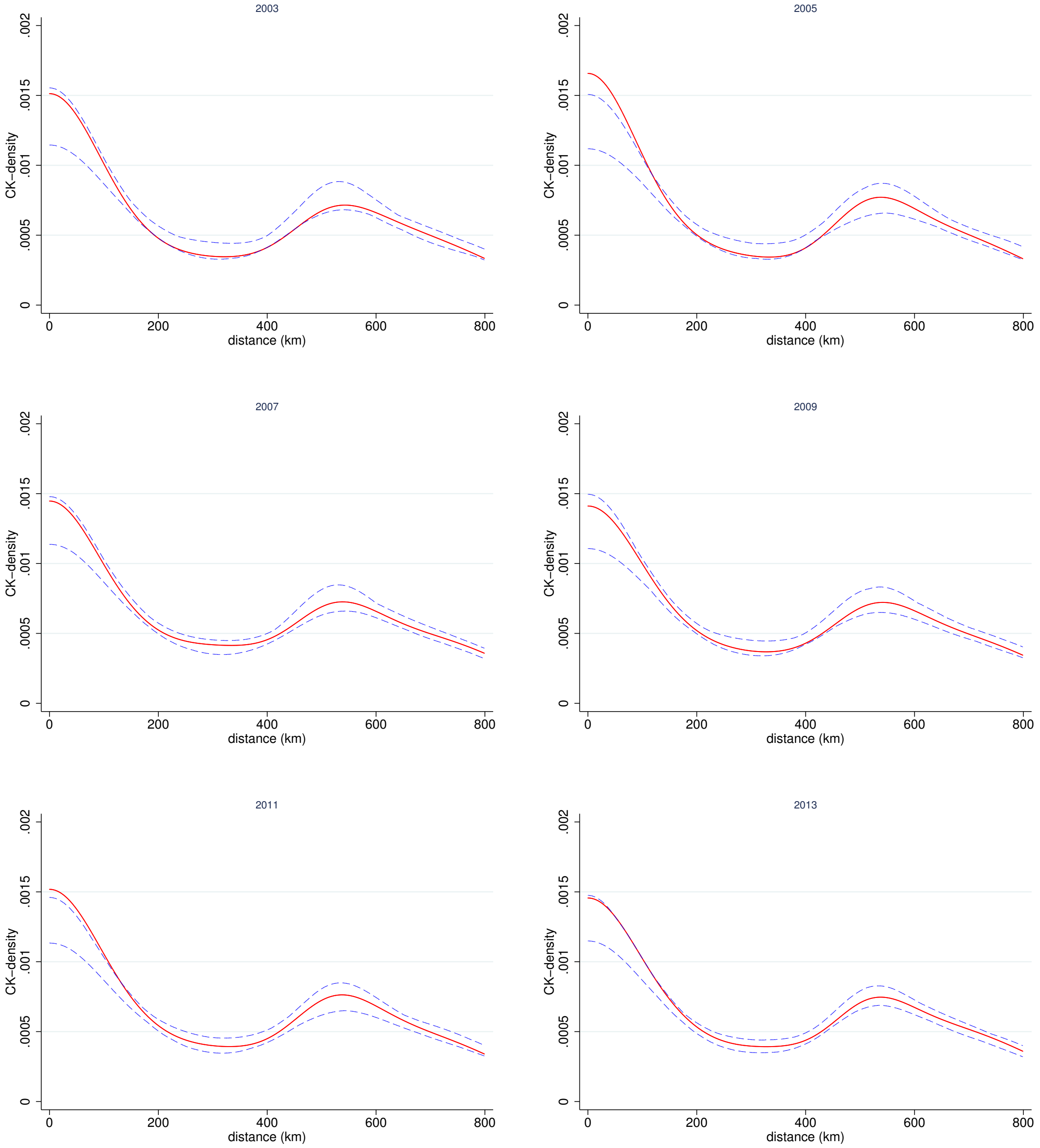
Figure 3.5 Location patterns of a plants using a similar workforce in 2001.



Notes : Estimations are based on the relevant subsets of plants that account for 13.3% of the universe of plants.

(3353) and 'Navigational and Guidance Instruments Manufacturing' (3345) were the most represented industries, with 49.9% of the bilateral distances. Conditional on their similarity in the types of skills and expertise they use, these two industries are mostly coagglomerated with 'Mattress manufacturing' (3379) and 'Coating, engraving, cold and heat treating and allied activities' (3328). In 2011 and 2013, these two industries accounted for 38.8% and 35.4% of the total bilateral distances respectively. As can be seen from Table 3.9, my CK-density measure of coagglomeration also allows to capture interactions between industries with similar types of workers that belong to *different* NAICS 3-digit sectors. More information on the most frequently co-localized industries in 2005, 2011, and 2013 is provided in Table 3.4.2.

Figure 3.6 Location patterns of plants with a similar workforce.



Notes : Estimations are based on the relevant subsets of plants that represent : 10.6% of the universe of plants in 2003, 10.2% in 2005, 15.8% 2007, 10.6% in 2009, 11.6% in 2011, and 14.3% in 2013.

Table 3.4 Most co-localized Industries in 2005, 2011, and 2013 : similarity in the workforce Space

Two Most Co-Localized Industries in 2005

	3353 : Power, distribution & specialty transformers	23.62%			3345 : Navigational & guidance instruments	21.31%	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
3345	Navigational & guidance instruments	85,462	71.61	3353	Power, distribution & specialty transformers	69,573	71.61
3379	Mattress manufacturing	33,599	28.15	3379	Mattress manufacturing	36,431	33.83
3328	Coating, engraving, cold and heat	291	0.24	3328	Coating, engraving, cold and heat	1,031	0.96
				3372	Wood office furniture mfg	664	0.62
Total		119,352	100	Total		107,699	100

Two Most Co-Localized Industries in 2011

NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
	3353 : Power, distribution & specialty transf.	21.04%			3345 : Navigational & guidance instruments	19.37%	
3345	Navigational & guidance instruments	89,79	80.51	3353	Power, distribution & specialty transformers	58,899	57.83
3379	Mattress manufacturing	21,562	19.49	3379	Mattress manufacturing	24,853	24.40
				3339	Pump and compressor mfg	16,736	16.43
				3341	Computer and peripheral equipment mfg	1,295	1.27
				3271	Pottery, ceramics & plumbing fixture	69	0.07
Total		110,641	100	Total		101,852	100

Two Most Co-Localized Industries in 2013

	3345 : Navigational & guidance instruments	19.25%			3353 : Power, distribution & specialty transf.	16.13%	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
3339	Pump and compressor mfg	66,987	37.35	3345	Navigational & guidance instruments	69,698	78.64
3353	Power, distribution & specialty transformers	48,005	26.76	3379	Mattress manufacturing	14,509	16.37
3341	Computer and peripheral equipment mfg	27,147	15.13	3339	Pump and compressor mfg	3,528	3.98
3379	Mattress manufacturing	19,205	10.71	3341	Computer and peripheral equipment mfg	893	1.01
3271	Pottery, ceramics & plumbing fixture	16,927	9.44				
3372	Wood office furniture mfg	1,101	0.61				
Total		110,641	100	Total		101,852	100

Notes : I present only the two most co-localized industries which represent more than 44.9% of the total bilateral distances across years : 40.4% in 2005 ; 38.8% in 2007 and 35.4% in 2013. These two industries are displayed in columns while their related industries are displayed in lines.

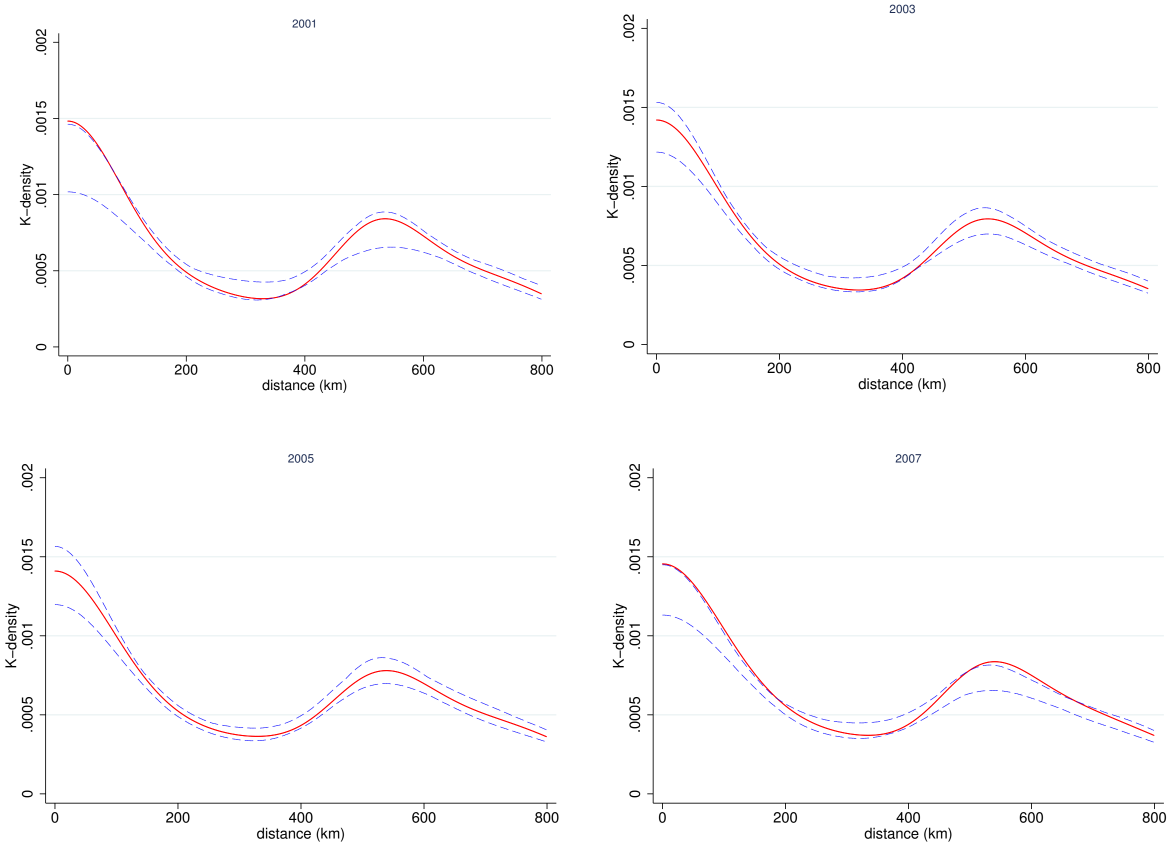
3.4.3 Location patterns of plants using or exchanging similar technologies

I use the NBER U.S. Patent Citations Data to compute the bilateral distances as proxied by patent citations between industries in technology space. I fix a threshold distance that allows me to fulfill my sampling requirements. This threshold distance allows to select between 8,513 (15.9%) plants in 2001 and 10,819 (20.2%) plants in 2003. I then use these relevant subsets of plants to estimate my CK-density measure of coagglomeration.

My results show that plants tend to take advantage of the speeding of the flows of ideas across sectors in 2001. As can be seen from the top left panel of Figure 3.7, the pairs of plants that use or exchange similar technology are located near one another at short distances in 2001 (less than 150 km). This result is reminiscent of the findings in Strange et al. (2014) who document that knowledge spillovers are positively associated with coagglomeration. Looking at intermediate and long distances, the location patterns of pairs of plants with similar technology flows are not significantly different from the ones that would be obtained by a purely random location process, except in 2007. This result is in line with Ellison et al. (2010) who find that patent citation measures were uncorrelated with coagglomeration at long distances. Figure 3.7 illustrates that the location patterns of pairs of plants that use or exchange similar technologies are not significantly different from ones that would be obtained by a purely random location process in 2003, 2005, and 2007. However, my results are always borderline significant as found in other studies (e.g., Behrens and Guillain, 2015). Results for others years are missing because of data issues. Patent citation flows cover the period 1976-2006, and I match these data with my Scott's 2001, 2003, 2005, and 2007 samples.

Table 3.9 in the Appendix summarizes the industry coverage within my

Figure 3.7 Location Patterns of Plants using or exchanging similar technologies.



Notes : Estimations are based on the relevant subsets of plants that represent : 15.9% of the universe of plants in 2001, 20.2% in 2003, 20.4% in 2005, and 20.4% in 2007.

relevant subsets of plants. Across years, my CK-density measures are based on plants that belong to 42% of the 185 NAICS 5-digit industries in the strict definition of manufacturing. Turning to the most co-localized industries, my relevant subsets of plants using similar technology contain 643,790 unique bilateral distances, where 'Soap and cleaning compound manufacturing' (32561) and 'Plastic plumbing fixture manufacturing' (32619) are the most represented industries, with 59.87% of the total distances. Conditional on their similarity in technology space, these two industries are mostly coagglomerated with 'Clay building material' and 'Refractory manufacturing' (32712) and 'Synthetic dye and pigment manufacturing' (32513). These two industries remain the most represented in 2003, 2005, and 2007, where they account for 41.2%, 40.6% and 40.2% of the total bilateral distances, respectively. More information on the most frequently co-localized industries in 2001 and 2007 is provided in Table 3.4.3 below.

Table 3.5 Most frequently co-Localized Industries in 2001, and 2007 : similarity in technology space

	Two Most Co-Localized Industries in 2001						
	32561 : Soap and cleaning compound mfg	**30.30%**			**32619 : Plastic plumbing fixture mfg**	**29.57%**	
NAICS	NAICS name	Freq.	%	NAICS	NAICS name	Freq.	%
32619	Plastic plumbing fixture mfg	195,007	99.97	32561	Soap and cleaning compound mfg	190,083	99.86
32712	Clay building material and refractory mfg	44	0.02	32712	Clay building material and refractory mfg	160	0.08
32513	Synthetic dye and pigment mfg	25	0.01	33122	Cold-rolled steel shape mfg	114	0.06
	Total	195,076	**100**		**Total**	**190,357**	**100**
	Two Most Co-Localized Industries in 2007						
	32561 : Soap and cleaning compound mfg	**21.57%**			**32619 : Plastic plumbing fixture mfg**	**18.65%**	
32619	Plastic plumbing fixture mfg	119,875	99.44	32561	Soap and cleaning compound mfg	103,321	99.13
32712	Clay building material and refractory mfg	571	0.47	33122	Cold-rolled steel shape mfg	838	0.80
32513	Synthetic dye and pigment mfg	99	0.08	32712	Clay building material and refractory mfg	70	0.07
	Total	**120,546**	**100**	**Total**		**104,229**	**100**

Notes : I represent only the two most frequently co-localized industries which represent more than 40% of the total bilateral distances across years : 59.9% in 2001and 40.22% in 2007. These two industries i are displayed in column while their related industries are displayed in lines.

3.4.4 Location patterns of small and large plants

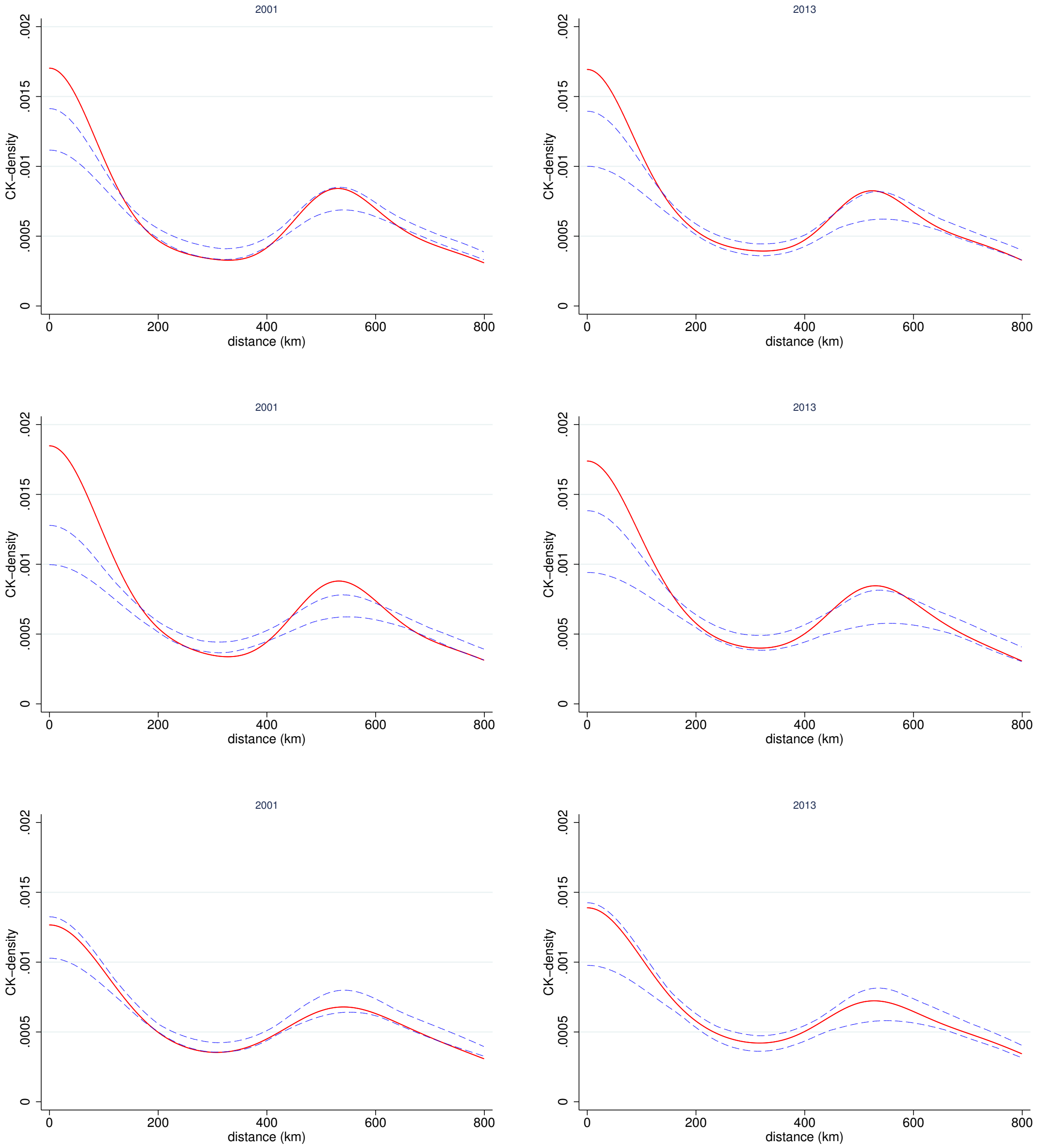
This section deals with potential heterogeneity in agglomeration benefits across plants and industries. My main focus is on plant size. I will thus look at the location patterns of establishments of different sizes. There are good theoretical and empirical reasons to look at small and large plants. Chinitz (1961) points out the differential importance of agglomeration effects for small and large firms. Behrens and Sharunova (2015) find that large plants tend to cluster with other large plants. Holmes and Stevens (2014) find that large plants tend to cluster, whereas small plants are more dispersed. Rosenthal and Strange (2005, 2010) and Rigby and Brown (2015) find that industries differ in how they benefit from clustering, and within industries, large and small plants display different location patterns. Holmes and Stevens (2002, 2014) suggest that clustering in the United States is driven mostly by large establishments. Barrios, Bertinelli, and Strobl (2006b) provide similar findings for Ireland. Rosenthal and Strange (2003, 2010) document that the marginal effect on the entry of new plants in an industry generated by an employee in a small establishment is greater than that generated by an employee in a large establishment. This should drive a stronger clustering of small plants. Alcácer and Chung (2014) find that small firms locate where supplier agglomeration economies exist, but large firms do not. When excluding the smaller establishments in the U.K. data, Duranton and Overman (2005) find that localization tends to become stronger in some industries, but weaker in others. Given the effects of small and large plants on industry dynamics and growth, it seems worthwhile to investigate in more detail their geographical colocation patterns.

I define small plants and large plants as those plants that are below or above the median employment size in their industry. I then split my relevant subsets of plants into small and large establishments and use equation (E.3) to perform a number of exercises in order to answer the following questions : Do pairs of small plants, pairs of large plants, and pairs of 'mixed' (small and large) plants with similar characteristics locate near one another in space? To do so, I look at the location patterns of small-small pairs of plants, large-large pairs of plants, and small-large or large-small pairs of plants.

Looking at the location patterns of pairs of 'mixed' plants conditional on their similarity in the I-O linkage space, the top panel of Figure 3.8 shows that pairs of big-small or small-big plants are strongly localized at short distances and dispersed at long distances (beyond 600 km) in 2001 and 2013. This result consistently holds across years (see Figure 3.17 in the Appendix). When looking at the location patterns of pairs of small versus pairs of large plants, my results illustrate that their location patterns are starkly different, with pairs of large plants exhibiting more localization than pairs of small plants.

As can be seen from the middle panel of Figure 3.8, pairs of large plants with similar I-O linkages are localized at short distances and at intermediate distances. The bottom panel of Figure 3.8 shows that in 2001 and 2013, the location patterns of pairs of small plants are not significantly different from the ones that would be obtained by a random location process. This result suggests that big plants drive the colocation patterns of industries with similar I-O linkages. This result is in line with that of Strange et al. (2014), who find that input-output linkages increase as we move from small to big sector pairs. In addition, pairs of plants that belong to industries with large entrants and incumbents are more geographically concentrated than 'mixed' and small pairs of plants. At long distances, both small and large pairs of plants are significantly dispersed. These results suggest that large plants exhibit more localization than small plants (Holmes and Stevens, 2002, 2014). However, my analysis shows a slightly different result for pairs of small and big plants in 2005, 2009, and 2011. As can be seen from Figure 3.18 in the appendix, pairs of small plants with similar I-O linkages exhibit localization at short distances and dispersion at long distances. Hence, I-O links are important for all types of

Figure 3.8 Location patterns of plants with similar I-O linkages in 2001 and 2013 : small-large (Top Panel), large-large (Middle Panel), and small-small (Bottom Panel).



Notes : Estimations are based on the relevant subsets of plants that represent : 11.0% of the universe of plants in the top left panel, 12.6% in the top right panel, 5.2% in the middle left panel, 5.9% in the middle right panel, 5.8% in the bottom left panel, and 6.7% in the bottom right panel.

plants, but especially for big ones and to a lesser extent for small ones.

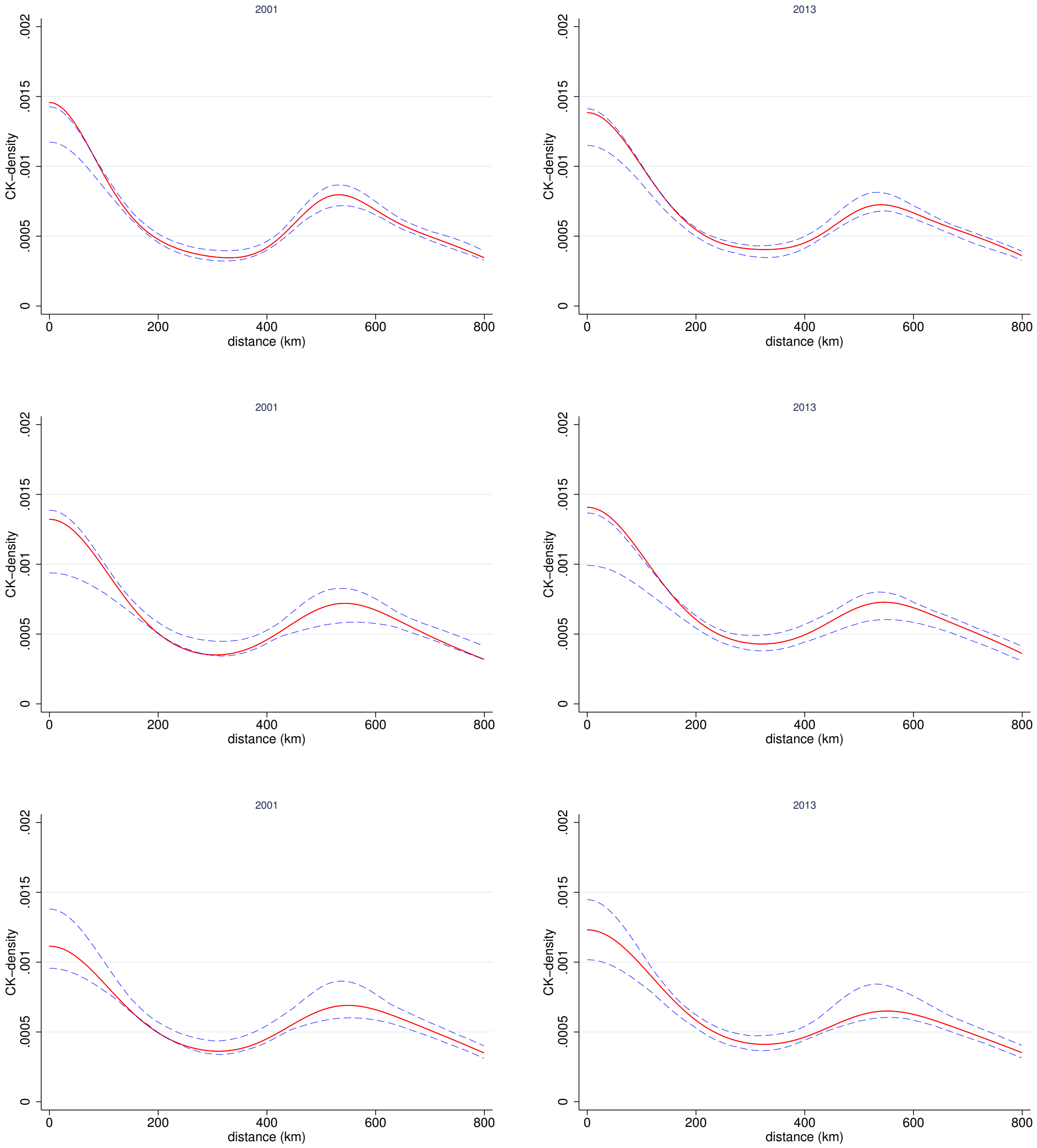
Regarding the localization patterns of pairs of small plants, pairs of large plants, and pairs of 'mixed' plants that belong to industries that use the same types of workers, the top panel of Figure 3.9, shows that pairs of big-small and small-big plants that use workers with similar skills and expertise are slightly localized at short distances in 2001. There is no strong evidence for this pattern in other years. At intermediate and long distances, their location patterns are not significantly different from what would be obtained by a random location process. This result is consistent across years (see the right panel of Figure 3.17 in the Appendix).

Looking at the location patterns of pairs of small-small and large-large plants, my results illustrate that pairs of small and pairs of large plants that use workers with similar skills and expertise are located differently at short distances and at long distances. As can be seen from Figure 3.9, pairs of big plants tend to exhibit localization at short distances while the location patterns of pairs of small plants are not significantly different from randomness. At intermediate and long distances, the location patterns of small and of large plants are not significantly different from randomness (all these patterns are consistent across years, see Figure 3.18 and Figure 3.19 in the Appendix). These results are also in line with previous findings by Holmes and Stevens (2002, 2014) who show that clustering in the U.S. is driven mostly by large establishments.

3.4.5 The strength of input-output linkages versus labor market pooling

I now look at the *strength* of localization, i.e., the area between the observed distribution (solid line) and the upper-bound of the confidence band (in dash) in Figure 3.3. It is computed by summing the difference between the upper bound

Figure 3.9 Location patterns of plants with a similar workforce in 2001 and 20013 : small-large (Top Panel), large-large (middle panel), and small-small (bottom panel).



Notes : Estimations are based on the relevant subsets of plants that represent : 12.4% of the universe of plants in the top left panel, 13.4% in the top right panel, 5.6% in the middle left panel, 6.6% in the middle right panel, 6.5% in the bottom left panel, and 7.2% in the bottom right panel.

and the localization measure across all distances (see Duranton and Overman, 2005 and Behrens and Bougna, 2015 for details on this measure). Intuitively, this measure can be interpreted as the excess probability of finding another plant in the same relevant subset of industries (with similar characteristics) at some distance d when controlling for the overall distribution of manufacturing and accepting a 5% risk level. As can be seen from Table 3.6 below, the CK-density measure of localization with respect to input-output linkages is greater than the CK-density measure with respect to labor market pooling. This result suggests that input-output linkages play a more important role than labor market pooling in manufacturing location decisions. It is in line with the findings in Ellison et al. (2010), who show that input-output linkages are particularly important.

Another interesting result that can be seen from Table 3.6 is that industries are always more concentrated in terms of employment that in terms of plant counts. This is consistent with the findings of Behrens et al. (2015), and of Holmes and Stevens (2012, 2014). When looking at small and big plants with similar input-output linkages or labor market pooling, Table 3.6 show that pairs of large-large plants are on average always more concentrated than pairs of 'mixed' plants, and even more concentrated than pairs of small-small plants.

Table 3.6 Strength of localization across years

	Input-output						Labor market pooling				
Year	All	Restricted	Weighted	Small/Big	Big	Small	All	Restricted	Small/Big	Big	Small
2001	0.013	0.018	0.057	0.022	0.062			0	0.001		
2003	0.028	0.018	0.023	0.024	0.048				0.006	0.001	
2005	0.025	0.026	0.020	0.023	0.048	0.011	0.010	0.016	0.003	0.005	
2007	0.014	0.019	0.021	0.012	0.045			0.007		0.004	
2009	0.022	0.034	0.034	0.022	0.039	0.013				0.005	
2011	0.024	0.017	0.036	0.025	0.033	0.011	0.005		0.001	0.001	
2013	0.020	0.016	0.018	0.023	0.033					0.004	

Notes : The strength of localization is the average localization index across all distances.

3.5 Robustness checks : restricted distance and correlation coefficients

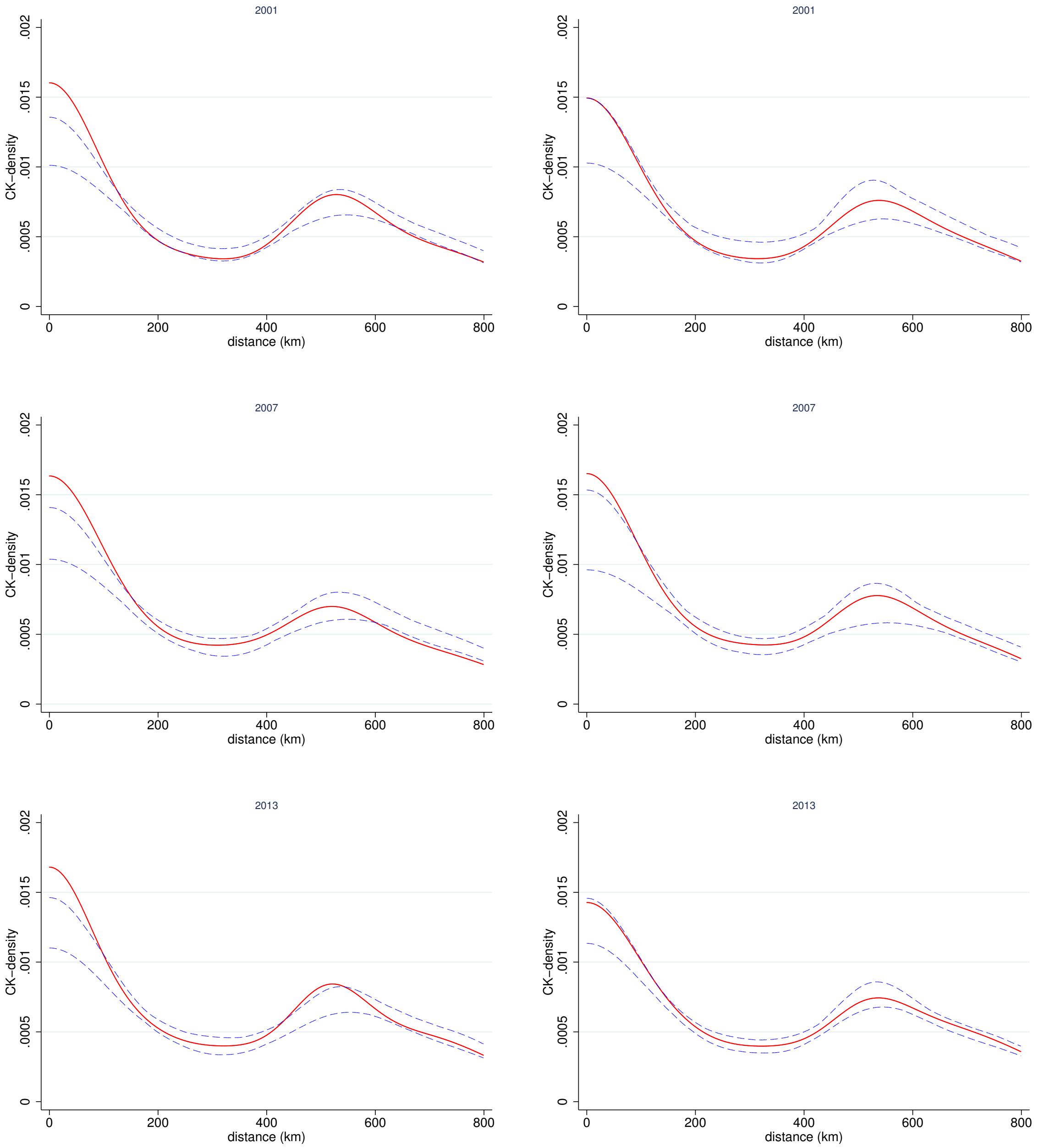
I perform four types of robustness checks. In the first robustness check, I use a restricted selection distance. In my second robustness check, I estimate my CK-density measures using Pearson correlation coefficients to measure industries similarity, as in Glaeser and Kerr (2009), Ellison et al. (2010), and Strange et al. (2014). In a third robustness check, I compute employment weighted CK-density measures of coagglomeration. Finally, as a last robustness check I test whether my results are robust to a restricted counterfactual location universe. This last restriction allows me to compare my measure to the coagglomeration measures used by Duranton and Overman (2008).

3.5.1 Localization patterns of plants using a restricted distance selection

The key idea behind the use of a restricted selection distance is to see if my main results are robust when I restrict my samples to pairs of plants that are 'very' close in the non-geographic space. Figure 3.16 illustrates my previous results on the importance of I-O linkages (left panel) and labor market pooling (right panel), using a more restrictive threshold g in non-geographic space and using the same relevant samples as before. The results remain robust to the use of a restricted selection distance. The pairs of manufacturing plants with similar I-O linkages are localized at short distances and dispersed at long distances. The importance of labor market pooling on coagglomeration remains significant at short distances and for the years 2005, and 2007.

Looking at the colocation patterns of pairs of small-large and large-small plants with similar I-O linkages and types of workers, my results remain robust

Figure 3.10 Location Patterns of Plants with Similar I-O linkages (left panel) and Workforce (right panel) : Restricted Selection Distance.



Notes : Estimations are based on the relevant subsets of plants that represent : 8.4% of the universe of plants in the top left panel, 5.3% in the top right panel, 7.8% in the middle left panel, 7.2% in the middle right panel, 11.1% in the bottom left panel, and 8.0% in the bottom right panel.

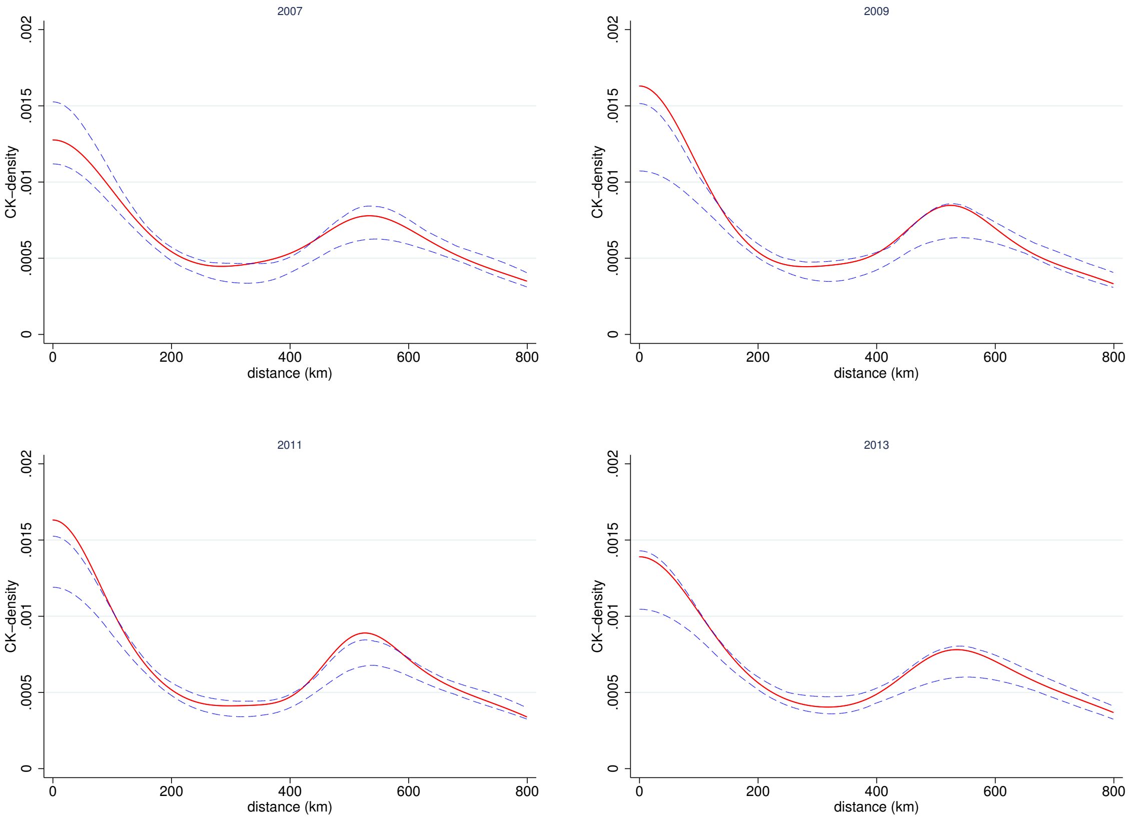
to the use of a restricted selection distance. As can be seen from Figure 3.16 in the Appendix, pairs of big-small or small-big plants with similar I-O linkages are strongly localized at short distances and dispersed at long distances (beyond 600 km). Finally, as can be seen from Table 3.6 the strength of input-output get stronger when I use a restricted selection distance (except in 2003 and 2013). As observed, the confidence bands when using restricted selection distances are wider compared to the baseline case, which is due to less observations in the restricted samples. This shows that localization get stronger (Table 3.6).

3.5.2 Localization of plants using correlation coefficients as similarity measures

Ellison et al. (2010) and Strange et al. (2014) use pairwise industry correlation coefficients to measure the similarity of employment in industries. I compute this correlation coefficient to see if my main results are robust to the choice of the similarity measure. To do so, I correlate vectors of occupational employment shares across industries and I re-estimate my similarity metric using the Pearson correlation coefficient. Contrary to the Euclidian distance, two industries are similar if the Pearson correlation coefficient of their vectors of occupation shares are high. I fix a threshold correlation coefficient of 0.9 – i.e., all pairs of plants with a correlation coefficient greater than or equal to 0.9 are included in my relevant subsets of plants. Across years, this threshold coefficient allows for the selection of 5,104 plants in 2009 (10.7%) and 8,322 (16.9%) plants in 2007. I then use these relevant subsets of plants to estimate my CK-density measure of coagglomeration.

My results confirm the previous findings that manufacturing plants tend to take advantage of groups of workers with similar skills and expertise. As can be seen from Figure 3.11, pairs of plants that belong to industries with a similar workforce, as measured by the Pearson correlation coefficient, are located near one another at short and intermediate distances in 2009 and 2011. At long distances, their location pattern is not significantly different from one that would be obtained by a purely random location process. In less evidence in other years.

Figure 3.11 Location patterns of similar pairs of plants in the workforce space : correlation coefficient.



Notes : Estimations are based on the relevant subsets of plants that represent : 16.9% in 2007, 10.7% of the universe of plants in 2009, 18.5% in 2011, and 14.9% in 2013.

3.5.3 Localization of plants using employment-weighted coagglomeration measures

I finally provide results for the employment-weighted version of the CK-density measure of coagglomeration. The weighted CK-density describes the distribution of bilateral distances between employees, whereas the unweighted CK-density describes the distribution of bilateral distances between plants. The main difference, therefore, lies in their interpretation : weighting plants by their employment gives a measure of colocalization of employment, which is different from a measure of plant colocalization. Contrary to Duranton and Overman (2005), who use a multiplicative weighting scheme, I use an additive scheme. Basically, the additive scheme gives less weight to pairs of large plants and more weight to pairs of smaller plants than the multiplicative one does. Methodological details and a discussion of the implications of the weighting scheme are provide in Behrens and Bougna (2015).

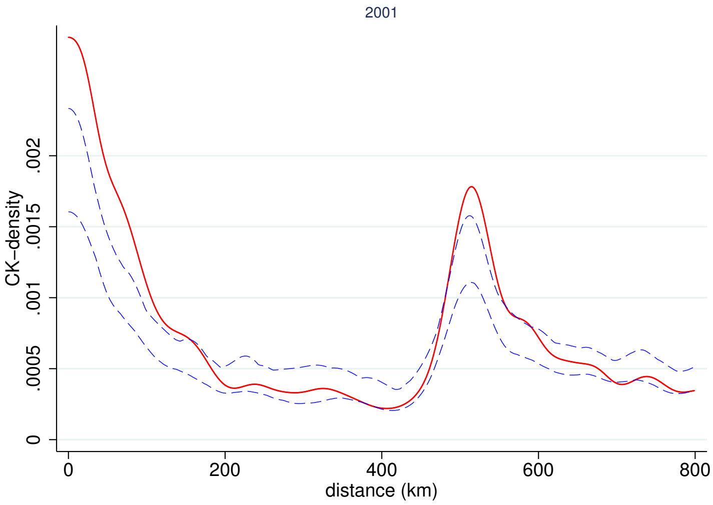
Let me denote the employment of plant i by e_i. The *the employment-weighted conditional kernel density estimator* becomes :

$$\widehat{K}_W(d|\Omega_{d,t}) = \frac{1}{h\sum_{i=1}^{N_t-1}\sum_{j=i+1}^{N_t}(e_i+e_j)} \sum_{i=1}^{N_t-1}\sum_{j=i+1}^{N_t}(e_i+e_j) f\left(\frac{d-d_{ij}}{h}\right) \tag{E.4}$$

I use equation (E.4) to look at the location pattern of employees that belong to plants of industries with similar I-O linkages. As can be seen from Figure 3.12, employees of pairs of plants with similar I-O linkages are localized at short and intermediate distances in 2001. At long distances, their location pattern is not significantly different from one that would be obtained by a purely random location process. As can be seen from Figure 3.13, these location patterns continue to hold in other years. My results also reveal that industries are, on average, always more colocalized in terms of employment than in terms of plant counts. This latter result is in line with findings by Behrens and Bougna (2015), Behrens et al. (2015), and

Holmes and Stevens (2002).

Figure 3.12 Location patterns of similar pairs of plants in the I-O linkages space : employment weighted.

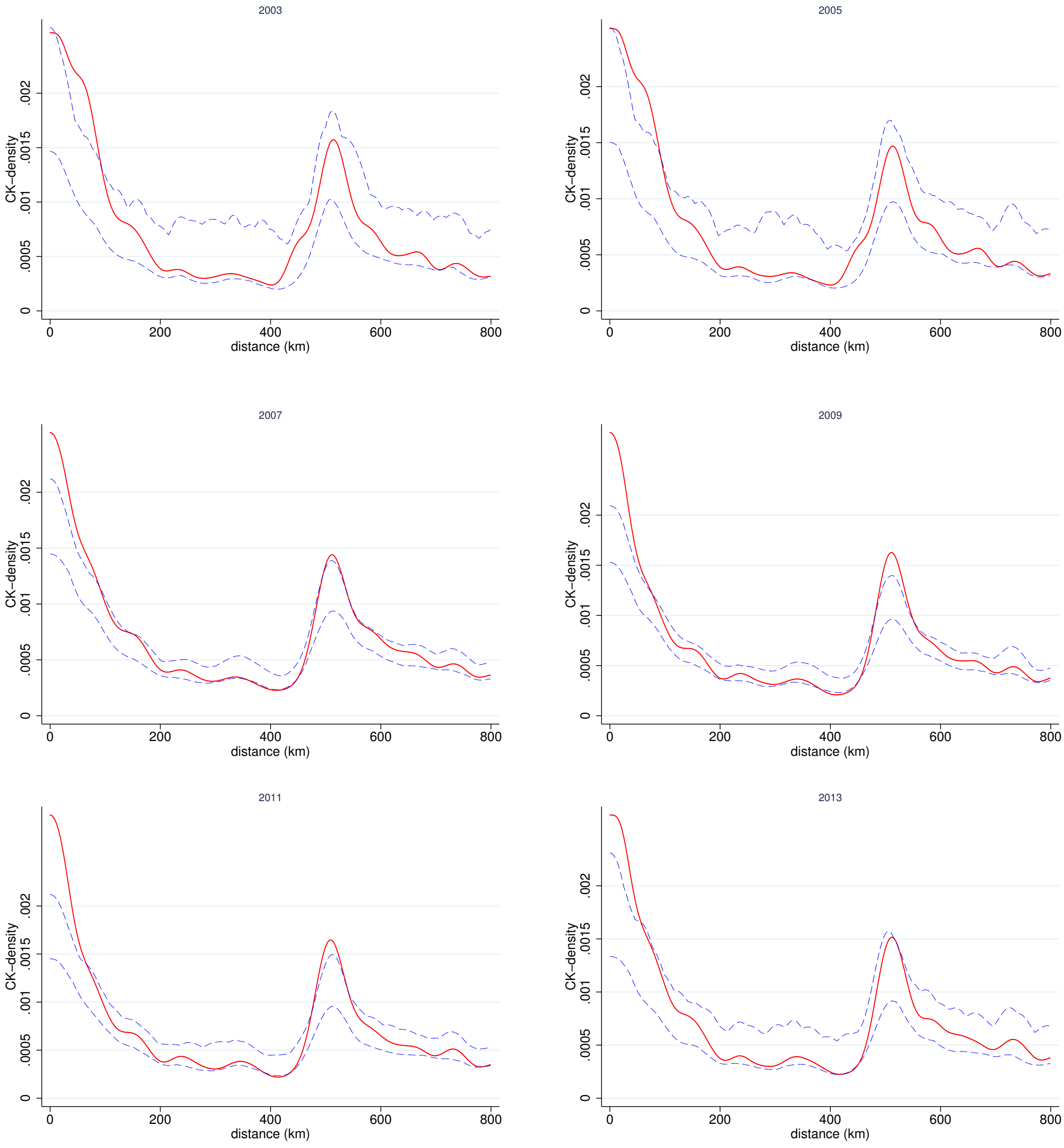


Notes : Estimations are based on the relevant subsets of plants that account for 11.3% of the universe of plants.

3.5.4 Localization of plants using a restricted location universe

In this last robustness check, I compare the location patterns of plants that belong to industries with similar characteristics to a hypothetical counterfactual distribution. Rather than using the whole manufacturing universe for the counterfactuals, I will restrict the counterfactual location universe to the population of sites occupied by plants that belong to industries with similar characteristics, i.e., the relevant subsets of plants. By doing so, I can compare my results to the coagglomeration measures in Duranton and Overman (2008). For example, in 2001, I have 53,540 plants in the universe with 6,101 that belong to industries with similar I-O linkages. In my previous analysis, my counterfactuals were generated by ran-

Figure 3.13 Location patterns of similar pairs of plants in the I-O linkages space : Employment weighted.



Notes : Estimations are based on the relevant subsets of plants that represent : 10.6% of the universe of plants in the top left panel, 10.8% in the top right panel, 10.0% in the middle left panel, 11.2% in the middle right panel, 10.2% in the bottom left panel, and 13.0% in the bottom right panel.

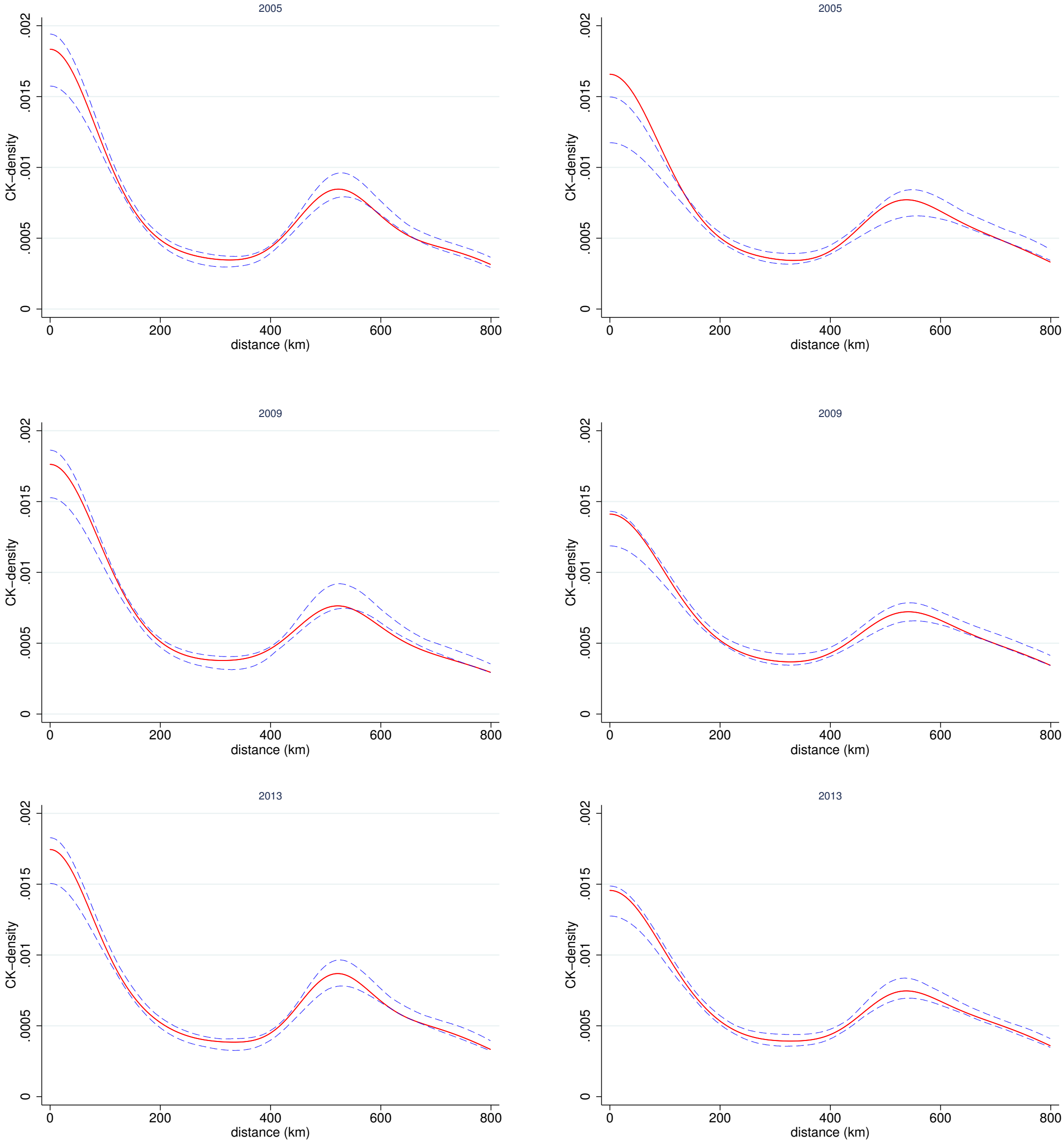
domly reallocating plants across the sites used by the universe of manufacturing plants (54,540 sites). In this robustness check, I randomly reallocate plants that belong to industries with similar characteristics across the 6,101 sites occupied by plants with similar I-O linkages.

As can be seen from Figure 3.14, the observed distribution of bilateral distances between plants that belong to vertically-linked industries and industries with a similar workforce is not different from a distribution derived from a random location process. The only exception is in 2005 and for plants with similar types of workers and expertise. However, pairs of big plants exhibit again different location patterns. Figure 3.15 confirms my previous findings where big plants tend to locate near one another at short distances to take advantage of pools of workers and to reduce the cost of obtaining their intermediate inputs. I find similar results with the employment-weighted version of the CK-density measure of coagglomeration.

3.6 Concluding remarks

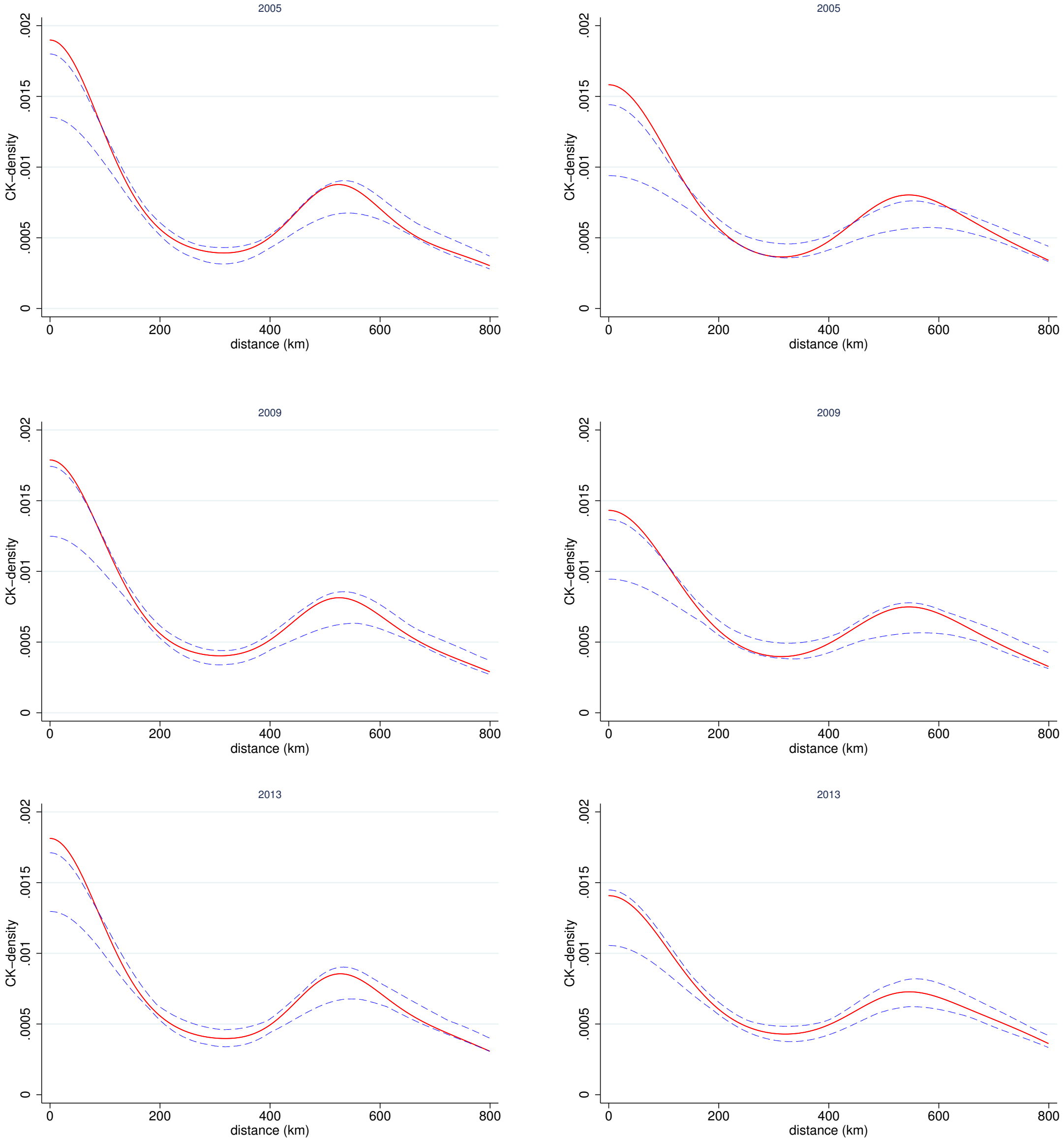
In this paper, I propose a new non-parametric approach to measuring the localization of 'closely related' multiple industries in continuous space. This leads to a new CK-density measure of coagglomeration. My 'conditional test' can be viewed as a non-parametric multidimensional way to assess coagglomeration. I apply these CK-density measures to Canadian manufacturing data in order to answer the following questions : Do pairs of plants with similar I-O linkages locate near one another in space ? Do pairs of plants with similar types of workers locate near one another in space ? Are pairs of plants that use or exchange similar technologies, as proxied by patent citations, locate near one another in space ? My results provide evidence for two of the three Marshallian mechanisms (i.e., input sharing and labor market pooling) in Canadian manufacturing industries. By examining the coagglomeration patterns, it can be seen that I-O flows and labor market pooling

Figure 3.14 Location patterns of similar pairs of plants with restricted counterfactuals : I-O linkages (left panel) and Workforce (right panel).



Notes : Estimations are based on the relevant subsets of plants that represent : 11.3% of the universe of plants in the top left panel, 13.3% in the top right panel, 10.0% in the middle left panel, 15.8% in the middle right panel, 13.0% in the bottom left panel, and 14.3% in the bottom right panel.

Figure 3.15 Location patterns of similar large pairs of plants with restricted counterfactuals : I-O linkages (left panel) and workforce (right panel).



Notes : Estimations are based on the relevant subsets of plants that represent : 3.3% of the universe of plants in the top left panel, 3.3% in the top right panel, 4.7% in the middle left panel, 4.7% in the middle right panel, 5.9% in the bottom left panel, and 6.6% in the bottom right panel.

are the most important Marshallian forces regarding industrial agglomeration. Similarly to Ellison et al. (2010), and Behrens and Guillain (2015), I find less evidence for knowledge spillovers. This result is probably due to the mis-measurement of knowledge and the fact that it is hard to dissociate labor mobility from knowledge spillovers (see Gabe and Jaison, 2013).

In summary, I find that plants tend to reduce the costs of obtaining intermediate inputs and shipping goods in their location decisions. Large plants exhibit more localization than small plants. My results further show that pairs of large plants are localized at short and intermediate distances, while pairs of small plants are localized at short distances and dispersed at long distances. I also document that plants tend to take advantage of groups of workers with similar skills and expertise, specifically at short distances. I find little evidence that plants that use or exchanged similar technologies, as measured using patent citations, cluster geographically.

My results also confirm previous findings for the role of large plants in clustering (Duranton and Overman, 2008) and the importance of Marshall's forces for industrial agglomeration (Ellison et al., 2010 and Strange et al., 2014). Big plants tend to co-locate with big plants. These insights are important, because plant colocation decisions and strategies are driven, in part, by firms' responses to agglomeration economies which, in turn, can be a source of competitive advantage (Alcácer and Chung, 2014). Regarding the more important of the three mechanisms, my results from the measure of the strength of localization indicate that input-output linkages are a significant factor in Canadian manufacturing. Labor market pooling is also important, specifically at smaller spatial scales, but it has less of an effect when we look at coagglomeration at a broader geographical scale. All of my main findings are robust to the use of a more restricted threshold distance and correlation coefficients as a similarity measure.

Further extensions of my empirical methodology are to develop an approach that allows one to test and to quantify the relative contribution of each of the Marshallian forces in a unified framework. Another possible extension of my CK-density approach is to use detailed plant-level data to build finer non-geographic distance measures in order to get truly away from industrial classifications and, therefore, move towards an ideal index of localization. Finally, my approach can also be used to investigate where precisely plants that use workers with a particular set of skills or specific technologies locate.

Acknowledgements. I am extremely grateful to my advisor, Kristian Behrens, for the guidance and support he provided to me. I also thank William Kerr for sharing data. Frédéric Robert-Nicoud, Yasusada Murata, Julien Martin, Florian Mayneris, Évarist Nguimkeu, Didier Tatoutchoup, Mark Brown, Eric Bahel, and Octave Keutiben provided discussions and comments. I finally thank seminar participants at the 2015 North American Regional Science Council (NARSC) Meetings in Portland, the 2015 Meetings of the Société Canadienne de Science Économique (SCSE) in Montréal, the January 2016 World Bank DEC Seminar, Department of Economics of Suffolk University, and the Department of Economics of the Université du Québec à Montréal (ESG-UQAM). Financial support from the *Canada Research Chair in Regional Impacts of Globalization* and CIRPÉE are gratefully acknowledged. The views expressed in this paper, as well as all errors, are my own.

3.7 Appendix to Chapter 3

Appendix A : Data quality, additional tables, and results

Table 3.7 provides a comparison of the Scott's National All 2001, 20133, 2005, 2007, 2009, 2011, and 2013 databases with Statistics Canada's province-level data from the Canadian Business Patterns. As can be seen, the database I use has a wide and similar coverage than the CBP.

Table 3.7 Comparing Scott's National All to the Canadian Business Patterns data of Statistics Canada.

	2001		2003		2005		2007		2009		2011		2013	
Provinces	CBP	**Scotts (%)**	CBP	**(%)**	CBP	**(%)**	CBP	**(%)**	**CBP**	**(%)**	CBP	**(%)**	CBP	**(%)**
Alberta	6170	63.60	6,001	61.22	5,416	65.68	5,435	69.53	5,351	69.56	4,880	71.33	5,361	61.78
British Columbia	9170	67.23	8,734	68.20	8,261	66.14	8,021	66.43	7,697	66.56	7,146	69.14	7,500	59.20
Manitoba	2007	82.96	1,942	80.79	1,741	87.02	1,643	86.98	1,605	81.06	1,462	86.53	1,502	78.23
New Brunswick	1488	95.70	1,342	104.40	1,195	107.62	1,092	109.52	1,018	117.98	932	106.97	939	97.98
Newfoundland and L.	799	72.34	734	79.29	629	87.28	536	96.08	508	95.28	455	89.89	448	86.61
Nova Scotia	1912	89.96	1,749	92.22	1,483	105.39	1,327	105.20	1,225	110.69	1,105	103.44	1,105	94.21
Ontario	25935	79.11	25,182	88.26	23,220	92.54	22,450	92.22	21,673	93.75	20,063	94.49	21,188	81.13
Prince Edward I.	361	91.69	318	96.23	292	113.36	262	118.32	256	111.72	221	106.79	234	96.15
Quebec	18902	83.71	18,341	81.40	17,026	84.27	15,904	82.84	15,238	84.75	14,390	83.00	14,570	76.31
Saskatchewan	1472	94.63	1,407	93.03	1,259	106.67	1,191	103.36	1,151	99.39	1,063	107.15	1,125	90.76
Territories	79	–	83	–	63	65.08	54	90.74	57	80.70	49	81.63	50	70.00
Total	**68,295**	**78.40**	**65,833**	**81.37**	**60,585**	**84.98**	**57,915**	**84.80**	**55,779**	**85.87**	**51,766**	**86.06**	**54,022**	**75.64**

Notes : Province-level breakdown of manufacturing firms (NAICS 31–33) in the 2001, 2003, 2005, 2007, 2009, 2011 and 2011 Scott's National All databases versus the Canadian Business Patterns data of Statistics Canada.

Geographical data. The census geography of 1996 and the postal codes as of May 2002 were associated with my 2001 sample. I also match my 2003 and 2005 samples with the 2001 Census geography and the postal codes as of December 2003 (for 2003) and January 2007 (for 2005). My 2007, 2009, and 2011 samples are matched with the census geography of 2006, and the postal codes as of March 2008, October 2010, and May 2011 respectively. Finally, my 2013 sample is matched with the census geography of 2011, and the postal codes as of June 2013. Table 3.8 below provides information on the geographical structure of the census (2001, 2006, and 2011) and the corresponding PCCF data.

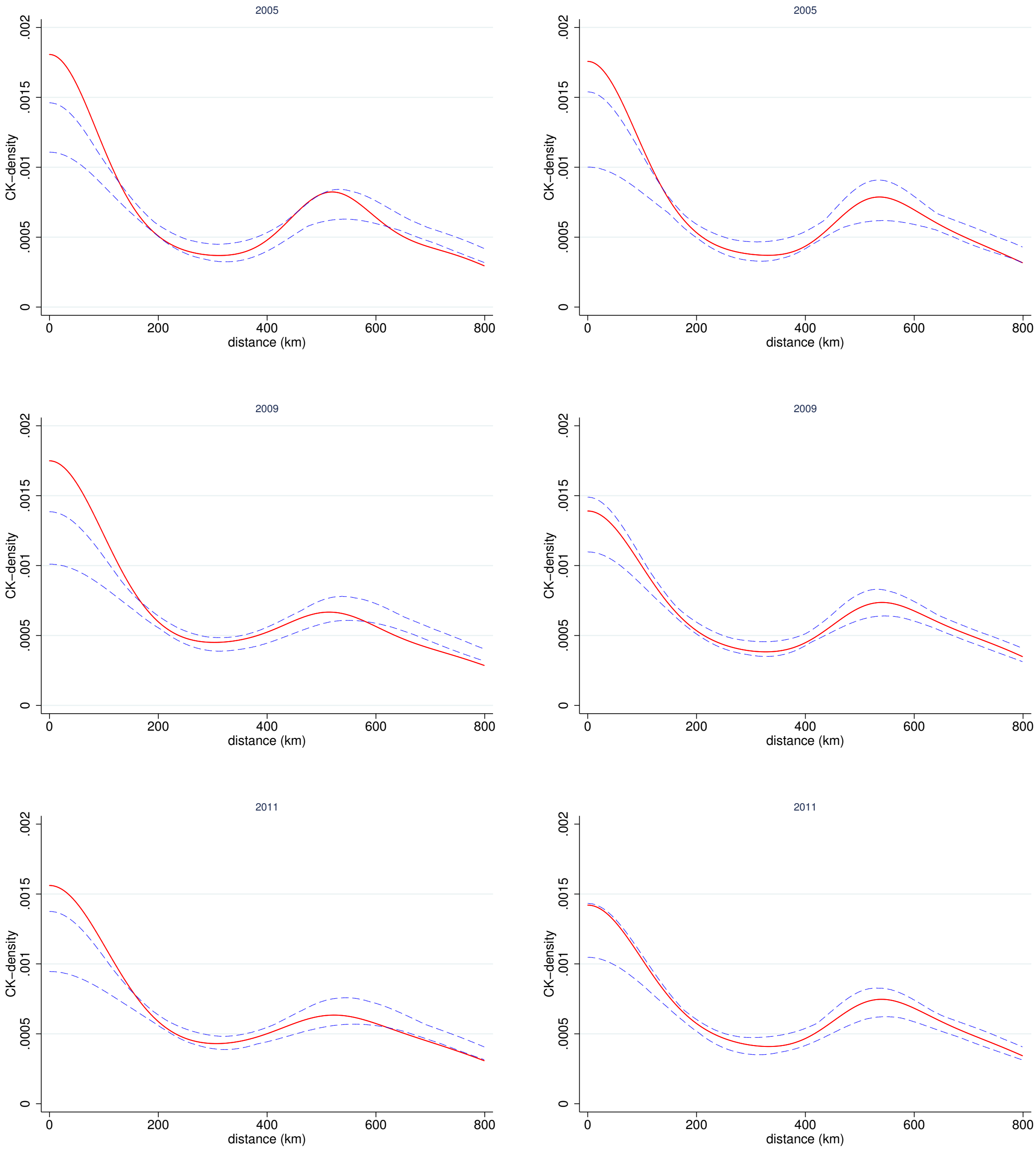
Table 3.8 Geographical Structure of the Census and PCCF Data.

	Census 1996 in the PCCF	Census 2001 in the PCCF	Census 2006 in the PCCF	Census 2011 in the PCCF
Provinces and territories	13	13	13	13
Economic regions	74	76	76	76
Census divisions	285	288	288	293
Census subdivisions	4,410	4,088	3,692	3,671
Dissemination areas	34,940	42,297	45,904	47,179
Geographical concordance :				
Scott's All year	2001	2005	2009	2011
PCCF version	May 2002	Jan 2007	Oct 2010	June 2013
Census geography	1996	2001	2006	2011
#unique postal codes	818,907	861,765	890,317	848,476

Notes : Geography of the 1996, 2001, 2006, and 2011 Censuses and concordances between *Scott's National All* databases and Statistic Canada's PCCFs.

Occupational Employment Statistics The OES survey samples approximately 200,000 establishments semi-annually in November and May of each year. This yields a combined sample size of 1.2 million establishments over six semiannual panels. Workers are categorized according to the U.S. Standard Occupational Classification (SOC) system, which contains over 800 occupations, of which 555 are related to manufacturing industries i.e., I have non-zero employment in at least one manufacturing sector.

Figure 3.16 Location patterns of plants with similar I-O linkages (left panel) and workforce (right panel) : restricted Selection Distance.



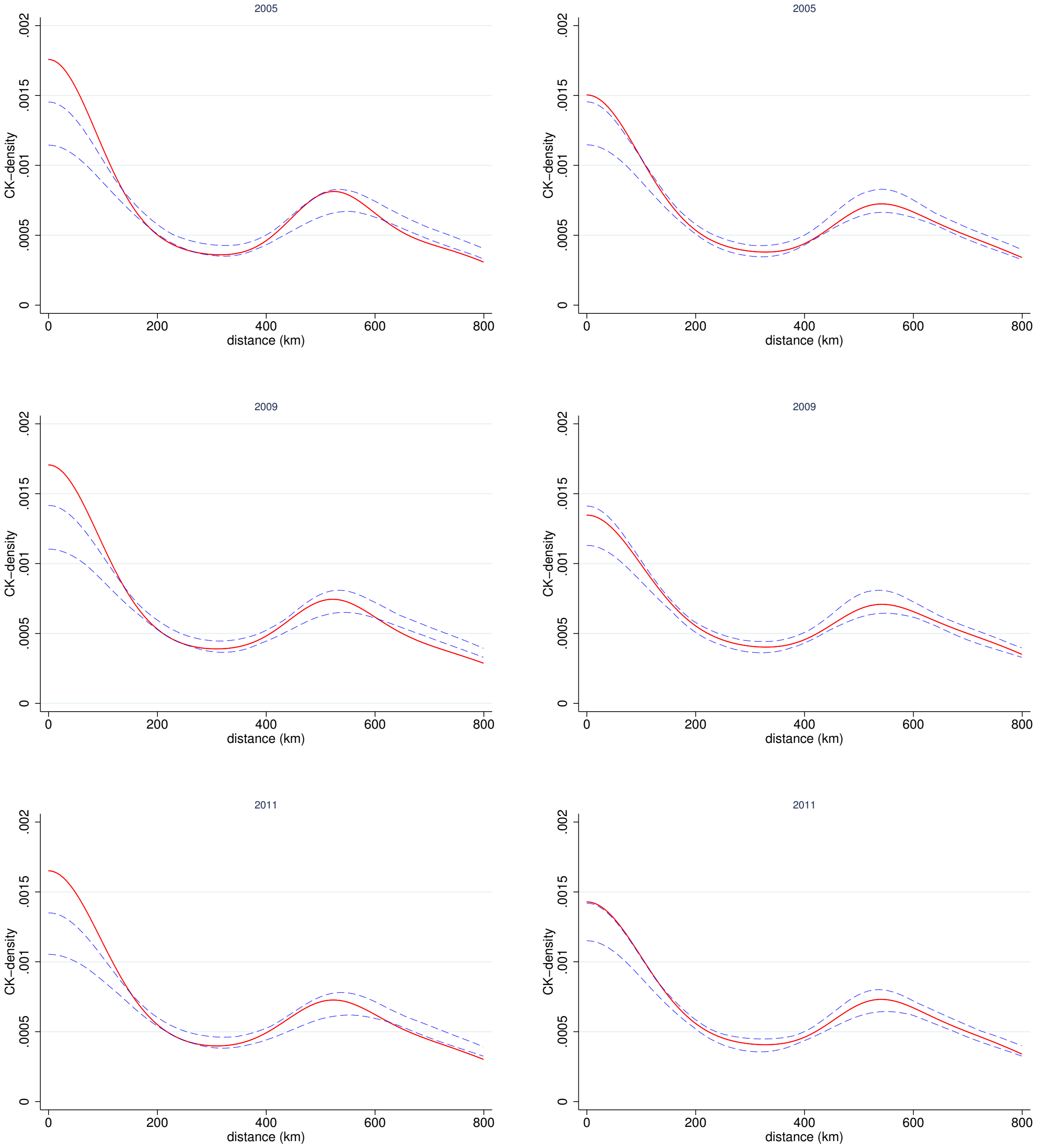
Notes : Estimations are based on the relevant subsets of plants that represent : 7.7% of the universe of plants in the top left panel, 6.1% in the top right panel, 6.6% in the middle left panel, 8.9% in the middle right panel, 7.0% in the bottom left panel, and 8.0% in the bottom right panel.

Table 3.9 Industries coverage in the relevant subsets of plants by years,

	Number of industries					
Year	Input-output		Labor market pooling		Knowledge spillover	
	strict	%	strict	%	strict	%
2001	74	30.6	25	29.4	73	40.6
2003	67	27.7	22	25.9	80	44.4
2005	69	28.5	24	28.2	77	42.8
2007	60	24.8	33	38.8	79	43.9
2009	63	26.0	23	27.1		
2011	65	26.7	24	28.2		
2013	78	32.2	21	24.7		

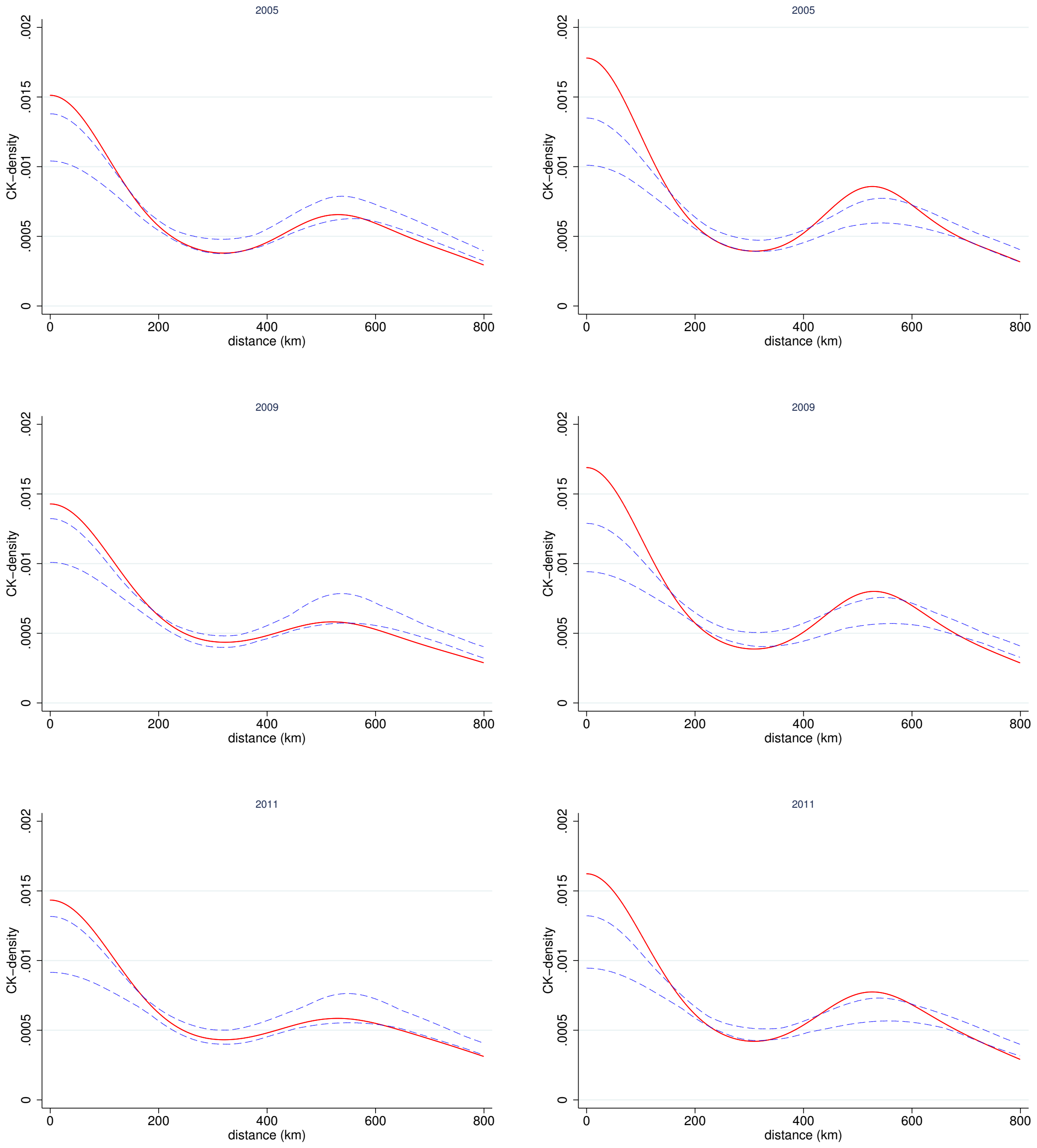
Notes : Strict refers to plants that report a manufacturing sector as their primary sector of activity and extended is for plant that report a manufacturing sector as one of their sector of activities (primary or secondary). There are 242 concorded NAICS industries at the 6-digit level (I-O linkages), 85 at the 4-digit level, and 180 at the 5-digit level.

Figure 3.17 Location patterns of small-large plants with similar I-O linkages (left panel) and workforce (right panel).



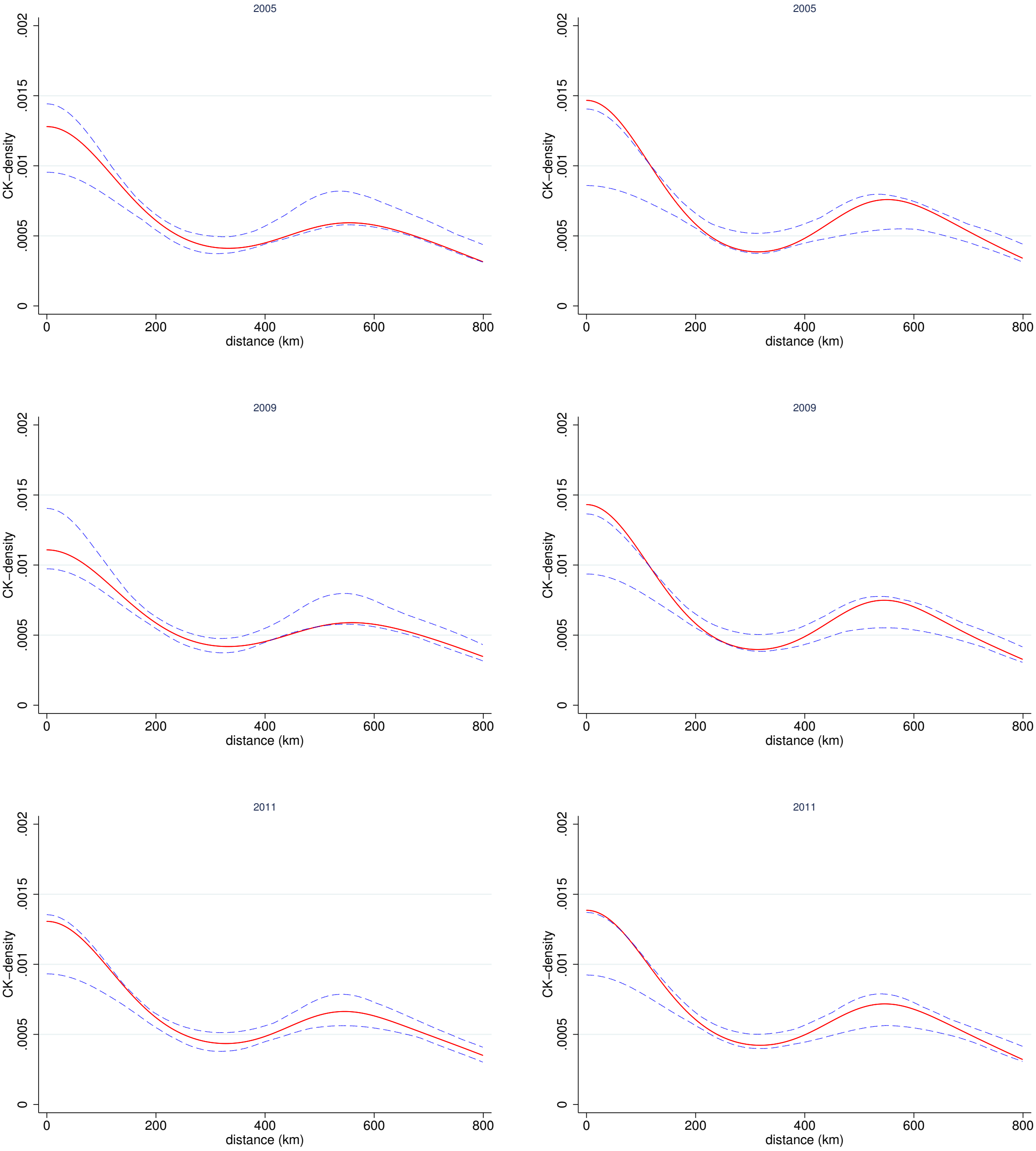
Notes : Estimations are based on the relevant subset of plants that represent : 11.0% of the universe of plants in the top left panel, 7.9% in the top right panel, 9.7% in the middle left panel, 12.9% in the middle right panel, 9.8% in the bottom left panel, and 10.8% in the bottom right panel.

Figure 3.18 Location patterns of pairs of small (left panel), and big plants (right panel) with similar I-O linkages.



Notes : Estimations are based on the relevant subsets of plants that represent : 5.5% of the universe of plants in the top left panel, 5.0% in the top right panel, 5.6% in the middle left panel, 5.1% in the middle right panel, 5.2% in the bottom left panel, and 4.7% in the bottom right panel.

Figure 3.19 Location patterns of similar pairs of small (left panel), and big plants (right panel) with a similar workforce.



Notes : Estimations are based on the relevant subsets of plants that represent : 6.5% of the universe of plants in the top left panel, 5.6% in the top right panel, 5.1% in the middle left panel, 4.7% in the middle right panel, 5.5% in the bottom left panel, and 5.0% in the bottom right panel.

CONCLUSION

Cette thèse propose trois essais sur la concentration spatiale de l'activité économique au Canada. Comment mesurer la concentration spatiale de l'activité économique ? Quels sont les déterminants de cette concentration spatiale, en particulier, quel est le rôle des coûts de transport et des trois facteurs Marshalliens dans les changements observés dans la concentration spatiale des industries ? L'objectif de la thèse est d'une part de bien mesurer la concentration spatiale de l'activité économique au Canada – afin de comprendre les mécanismes et les déterminants de cette concentration spatiale et d'informer l'opinion et les décideurs publics – et d'autre part, de proposer une mesure qui permet d'atténuer le problème du découpage sectoriel et de se rapprocher de l'indice idéal de concentration spatiale.

Dans le premier chapitre, nous construisons les mesures discrètes et continues de concentration spatiale afin de fournir un portrait complet de l'état, l'ampleur et la dynamique de la concentration spatiale de l'activité économique au Canada. Nos résultats montrent qu'au cours de la première décennie des années 2000, 40 à 60% des industries manufacturières sont concentrées de courtes et à des distances intermédiaires. Nos résultats font également ressortir une tendance à la dé-concentration des activités manufacturières au Canada. La contribution majeure du chapitre est qu'il permet de suivre l'évolution dans le temps de la concentration spatiale et fait ressortir les schémas de localisation des exportateurs, des petits et des jeunes établissements qui sont considérés comme vitaux pour la création d'emplois, et le développement local et régional.

Dans le deuxième chapitre, nous nous sommes intéressés aux déterminants de la concentration spatiale des industries manufacturières au Canada. En utilisant un long panel (1992-2008), nous régressons la mesure de concentration spatiale de Duranton et Overman (2005), sur des mesures micro-géographiques et spatiales des coûts de transport, de l'exposition au commerce international, et des liens en amont et en aval. Nos résultats montrent que l'augmentation des coûts de transport, la concurrence accrue du fait des importations en provenance des pays à faibles coûts et l'accroissement de la distance vers les clients et les fournisseurs expliquent entre 20 et 60% de la baisse observée dans la concentration spatiale des industries manufacturières au Canada. La principale contribution du chapitre est qu'il propose des évidences empiriques sur les déterminants de la concentration spatiale de l'activité économique en utilisant des mesures micro-géographiques et spatiales construites à des échelles industrielle et spatiale très fines. Malgré la baisse historique observée dans les coûts de transport, ce chapitre révèle également l'importance et le rôle qu'ils continuent de jouer dans la structure industrielle et la répartition spatiale des industries.

Dans le dernier chapitre, nous apportons un raffinement à l'approche de Duranton et Overman (2005). L'idée étant de proposer une approche qui permet d'atténuer ou de s'affranchir du problème du découpage sectoriel et de se rapprocher ainsi de l'indice de concentration spatiale idéal. De manière spécifique, nous combinons d'une part, l'approche de mesure de la concentration spatiale (à la Duranton et Overman, 2005) et l'approche de co-localisation (à la Ellison, Glaeser et Kerr, 2010), et d'autre part, nous associons ces mesures au degré avec lequel les industries échangent les biens, les travailleurs et les idées. L'objectif étant de combiner des mesures de distance technologiques (non-géographiques) à des mesures de distances géographiques entre secteurs. Conditionnellement à la similarité des établissements dans un espace non-géographique (liens en amont et en aval, type de travailleurs ou technologie utilisée), notre approche permet de vérifier si ces établissements sont concentrés ou non dans l'espace géographique. Nos résultats permettent de confirmer l'importance des liens en amont et en aval, et de l'accès à un bassin d'employés spécialisés dans les décisions de localisation des industries manufacturières. La contribution majeure de ce chapitre à la littérature sur la mesure de la concentration spatiale est qu'il propose un cadre unique qui permet de mesurer la co-agglomération des industries et de jauger de manière non-paramétrique l'importance des facteurs Marshalliens. Cette approche permet également d'atténuer le problème de la sensibilité des mesures existantes à un changement de nomenclature industrielle. Cependant, elle demeure sensible au découpage sectoriel en ce sens que la similarité des industries est mesurée à partir des données sectorielles agrégées. Un moyen de s'affranchir complètement du découpage sectoriel serait de mesurer la similarité à partir des données établissements et se rapprocher ainsi de l'indice idéal de concentration spatiale.

Au terme de cette expérience enrichissante, nous avons appris que toute recherche empirique doit être guidée par de solides fondements théoriques. Les divers travaux, nous ont permis de développer des outils et de comprendre la complexité et les difficultés liés à la construction d'un indice idéal de concentration spatiale. Dans un futur proche, nous allons poursuivre cet agenda qui contribue à mesurer la concentration spatiale et à expliquer les changements observés dans la structure spatiale des activités économiques. Dans un premier temps, nous allons construire des mesures de similarité à partir des données sur les établissements afin de s'affranchir définitivement du découpage sectoriel. Nous allons également enrichir l'approche développée afin d'être capable de tester et de quantifier la contribution des mécanismes Marshalliens. Finalement, la logique qui soutient les schémas de localisation dans les pays en développement est fort probablement différente de celle observée dans les pays développés. Il ne serait donc pas surprenant de constater que les clusters puissent jouer un rôle plus important dans le développement local et régional des pays en développement. Ainsi, dans un contexte où les données à des échelles spatiales très fines commencent à être disponibles dans ces pays, il devient primordial de construire des mesures et de rendre disponible les informations sur les clusters afin de faciliter la recherche et d'informer les décideurs politiques. Notre défi sera d'utiliser les outils développés dans le cadre de cette thèse afin de faire ressortir un portrait de la géographie des activités économiques et d'expliquer les changements observés dans la structure spatiale.

BIBLIOGRAPHIE

[1] Alcácer, Juan, and Wilbur Chung. (2014). "Location Strategies for Agglomeration Economies." *Strategic Management Journal*, **(35)** : 1749–1761.

[2] Acharya, S., Clayton, Z., Eriksson Giwa, S., Malinger, E. and Moura, A. (2009). "The North Carolina Furniture Cluster - the Microeconomics of Competitiveness." *Harvard University Press, Cambridge, Massachusetts*.

[3] Arbia, Giuspepe. (1989). " Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems." *Kluwer, Dordrecht*.

[4] Arzaghi, Mohammad, and J. Vernon Henderson. (2008). "Networking off Madison Avenue." *Review of Economic Studies*, 75**(4)** : 1011–1038.

[5] Autor, David H., David Dorn, and Gordon H. Hanson. (2013). "The China syndrome : Local labor market effects of import competition in the United States." *American Economic Review* 103**(6)** : 2121–2168.

[6] Barlet, Muriel, Anthony Briant and Laure Crusson. (2013). Location patterns of service industries in France : A distance-based approach, *Regional Science and Urban Economics* 43**(2)**, 338–351.

[7] Barrios, Salvador, Holger Goerg, and Eric Strobl. (2003). Explaining Firms' Export Behaviour : R&D, Spillovers and the Destination Market, *Oxford Bulletin of Economics and Statistics* 65**(4)**, 475–496.

[8] Barrios, Salvador, Luisito Bertinelli, and Eric Strobl. (2006b). "Geographic Concentration and Establishment Scale : An Extension Using Panel Data." *Journal of Regional Science*, 46 : 733–746.

[9] Behrens, Kristian, and Rachel Guillain. (2015). "Coagglomeration patterns and functional specialization in Canada." In progress.

[10] Behrens, Kristian, Vera Sharunova. (2015). “Inter- and intra-firm linkages : Evidence from microgeographic location patterns." *CEPR Discussion Papers*, DP 10921.

[11] Behrens, Kristian, Théophile Bougna, and W. Mark Brown. (2015). The World is not yet Flat : Transportation Costs matter ! CEPR Discussion Paper #10XXX, *Centre for Economic Policy Research*, London, UK.

[12] Behrens, Kristian, and Théophile Bougna. (2015). *“An anatomy of the geographical concentration of Canadian manufacturing industries*" (published in *Regional Science and Urban Economics*, 2015 **(51)** : 47–69.

[13] Behrens, Kristian. (2014). “Unweaving clusters : Manufacturing localization and international trade in Canada, 2001–2009.” *In progress*, Univ. of Quebec at Montreal, Canada.

[14] Behrens, Kristian. (2013). Strength in Numbers ? The Weak Effect of Manufacturing Clusters on Canadian Productivity. Commentary #377, CD Howe Institute, Toronto.

[15] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Südekum. 2012. “Spatial frictions.” CEPR Discussion Paper #8572, Center for Economic Policy Research, London, UK.

[16] Behrens, Kristian, and Pierre M. Picard. (2011). “Transportation, freight rates, and economic geography.” *Journal of International Economics* 85**(2)** : 280–291.

[17] Behrens, Kristian, Carl Gaigné, Gianmarco I.P. Ottaviano, and Jacques-François Thisse. (2007). “Countries, regions, and trade : On the welfare impacts of economic intergration.” *European Economic Review* 51**(5)**, 1277–1301.

[18] Bernard, Andrew B., and J. Bradford Jensen. (2004). Why Some Firms Export, *Review of Economics and Statistics* 86**(2)**, 561–569.

[19] Bernard, Andrew B., and J. Bradford Jensen. (1995). Exporters, Jobs, and Wages in U.S. Manufacturing : 1976-1987, *Brookings Papers on Economic Activity, Microeconomics* 1995, 67–119.

[20] Billings, Stephen B., and Johnson, Erik B. (2012). "A Nonparametric Test for Industrial Specialization." *Journal of Urban Economics*, 71**(3)** : 312–331.

[21] Brown, W. Mark, and William P. Anderson. (2015). "How thick is the border : the relative cost of Canadian domestic and cross-border truck-borne trade, 2004-2009." *Journal of Transportation Geography* 42 : 10–21.

[22] Brown, W. Mark, and David L. Rigby. (2015). "Who benefits from agglomeration ?" *Regional Studies* 49**(1)** : 28–43.

[23] Brülhart, Marius, Céline Carrère, and Frédéric Robert-Nicoud. (2014). "Trade and towns : On the uneven effects of trade liberalization." *In progress*, Université de Lausane *and* Université de Genève, Switzerland.

[24] Brülhart, Marius, Céline Carrère, and Federico Trionfetti. (2012). "How wages and employment adjust to trade liberalization : Quasi-experimental evidence from Austria." *Journal of International Economics* 86**(1)** : 68–81.

[25] Brülhart, Marius. (2011). "The spatial effects of trade openness : a survey." *Review of World Economics* 147**(1)** : 59–83.

[26] Brülhart, Marius, and Rolf Traeger. (2005). "An account of Geographic Concentration Patterns in Europe." *Regional Science and Urban Economics*, **(35)** : 597–624.

[27] Carlino, Gerald A. and Kerr, William R. (2015). "Agglomeration and Innovation." *Handbook of Regional and Urban Economics.*, 5A and 5B.

[28] Cassey, J. Andrew, Smith O. Ben. (2014). "Simulating confidence for the Ellison-Glaeser index". *Journal of Urban Economics*, **(81)** : 85–103.

[29] Chinitz, Benjamin. (1961). "Contrasts in agglomeration : New York and Pittsburgh". *American Economic Review Papers and Proceedings* 51**(2)** : : 279–289.

[30] Combes, Pierre-Philippe, and Laurent Gobillon. (2014). "The empirics of agglomeration economies." In : G. Duranton, J.V. Henderson, and W.C. Strange (eds.), *Handbook of Regional and Urban Economics, vol. 5*. North-Holland : Elsevier, *forthcoming*.

[31] Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. (2011). "The identification of agglomeration economies." *Journal of Economic Geography* 11**(2)** : 253–266.

[32] Combes, Pierre-Philippe, Thierry Mayer, and Jacques-François Thisse. (2008). "Economic Geography : The Integration of Regions and Nations". *Princeton, NJ : Princeton Univ. Press.*

[33] Cressie, Noel. (1993). "Statistics for spatial data". *John Wiley & Sons*, New York.

[34] Czamanski, Stan, and Luiz Augusto de Q. Ablas. (1979). "Identification of industrial clusters and complexes : a comparison of methods and findings." *Urban Studies* 16**(1)** : 61–80.

[35] D'Costa, Sabine. (2010). "Trade liberalization and the geographical concentration of industries : Evidence from Hungarian manufacturing industries." *PhD Dissertation*, London School of Economics, UK.

[36] Dauth, Wolfgang, Sebastian Findeisen, and Jens Suedekum. (2014). "The rise of the east and the far east : German labor markets and trade integration." *Journal of the European Economic Association* 12**(6)** : 1643–1675.

[37] Delgado, Mercedes, Porter, Michael E., Stern, Scotts. (2016). "Defining clusters of related industries", *Journal of Economic Geography* 16**(1)**, 1–38.

[38] Diggle, Peter J. (1983). "Statistical analysis of spatial point patterns." *Academic Press*, London.

[39] Dumais, Guy, Glenn D. Ellison, and Edward L. Glaeser. (2002). "Geographic concentration as a dynamic process." *Review of Economics and Statistics* 84**(2)**, 193–204.

[40] Duranton, Gilles, Peter M. Morrow, and Matthew Turner. (2014). "Roads and trade : Evidence from the US." *Review of Economic Studies* 81**(2)**, 681–724.

[41] Duranton, Gilles, and Matthew Turner. (2012). "Urban growth and transportation." *Review of Economic Studies* 79**(4)**, 1407–1440.

[42] Duranton, Gilles. (2011). California Dreamin' : The Feeble Case for Cluster Policies, *Review of Economic Analysis* 3, 3–45.

[43] Duranton, Gilles, Philippe Martin, Thierry Mayer and Florian Mayneris. (2011). *The Economics of Clusters : Lessons from The French Experience*. Oxford Univ. Press and CEPREMAP.

[44] Duranton, Gilles, and Henry G. Overman. (2008). Exploring the Detailed Location Patterns of UK Manufacturing Industries Using Microgeographic Data, *Journal of Regional Science* 48**(1)**, 213–243.

[45] Duranton, Gilles, and Henry G. Overman. (2005). Testing for Localisation Using Micro-Geographic Data, *Review of Economic Studies* 72**(4)**, 1077–1106.

[46] Duranton, Gilles, and Diego Puga. (2004). "Micro-foundations of urban agglomeration economies." In : J. Vernon Henderson, and Jacques-François Thisse (eds.), *Handbook of Regional and Urban Economics, vol. 4*. North-Holland : Elsevier B.V., pp. 2063–2117.

[47] Duranton, G. and D. Puga. (2001). "Nursery cities : Urban Diversity, Process Innovation, and the Life Cycle of Products ?, *American Economic Review* 91**(5)** : 1454 –1477.

[48] Ellison, Glenn D., Edward L. Glaeser, and William R. Kerr. (2010). "What Causes Industry Agglomeration ? Evidence from Coagglomeration Patterns", *American Economic Review* 100**(3)**, 1195–1213.

[49] Ellison, Glenn D., and Edward L. Glaeser. (1999). "The geographic concentration of industry : Does natural advantage explain agglomeration ?" *American Economic Review*, 89**(2)** : 311–316.

[50] Ellison, Glenn D., and Edward L. Glaeser. (1997). "Geographic concentration in U.S. manufacturing industries : A dartboard approach." *Journal of Political Economy* 105**(5)** : 889–927.

[51] Faberman, R. Jason. (2011). The Relationship Between The Establishment Age Distribution And Urban Growth, *Journal of Regional Science, Wiley Blackwell* 51**(3)**, 450–470.

[52] Faggio, Giulia, Silva, Olmo, and William C. Strange. (2014). "Heterogeneous agglomeration." SERC *DiscussionPapers,* SERCDP0152. *Spatial Economics Research Center (*SERC, London School of Economics and Political Science, London, U.K..

[53] Feser, Edward J. (2003). "What regions do rather than make : A proposed set of knowledge-based occupation clusters.' *Urban studies*, 40**(10)** : 1937–1958.

[54] Feser, Edward J., and Edward M. Bergman. (2000). "National industry cluster templates : a framework for applied regional cluster analysis.' *Regional studies*, 34**(1)** : 1–19.

[55] Gabe, Todd M., and Jaison, R. Abel. (2012). "Specialized Knowledge and the Geographic Concentration of Occupations.' *Journal of Economic Geography*, 12**(3)** : 435–453.

[56] Glaeser, Edward L., Sari Pekkala Kerr, and William R. Kerr. (2014). Entrepreneurship and Urban Growth : An Empirical Assessment with Historical Mines, *Review of Economics and Statistics*, forthcoming.

[57] Glaeser, Edward L., and William R. Kerr. (2009). "Local industrial conditions and entrepreneurship : how much of the spatial distribution can we explain ?" *Journal of Economics& Management Strategy*, 18**(3)** : 623–663.

[58] Fujita, Masahisa, Krugman, Paul R., and Anthony J. Venables. (1999). *The Spatial Economy : Cities, Regions and International Trade*. MIT Press, Cambridge, MA.

[59] Guimarães, Paulo, Octávio Figueiredo, and Douglas Woodward. (2011). Accounting for Neighboring Effects in Measures of Spatial Concentration, *Journal of Regional Science* 51**(4)**, 753–774.

[60] Guimaraes, Paulo, Octávio Figueiredo, and Douglas Woodward. (2000). “Agglomeration and the location of foreign direct investment in Portugal.” *Journal of Urban Economics* 47**(1)** : 115–135.

[61] Greenaway, David, and Richard Kneller. (2008). Exporting, Productivity and Agglomeration, *European Economic Review* 52**(5)**, 919–939.

[62] Haedo, C., Mouchart, M. (2012). “A stochastic independence approach for different measures of concentration and specialization.” *Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).*

[63] Jaffe, Adam B., Manuel Trajtenberg, and Michael S. Fogarty. (2000). “Knowdledge Spillovers and Patent Citations : Evidence from a Survey Inventors. ?" *American Economic Review* 90**(2)** : 215-218.

[64] Jaffe, A.B., M. Trajtenberg and R. Henderson. (1993). “Geographic localization of knowledge spillovers as evidenced by patent citations. ?" *Quarterly Journal of Economics* 108**(3)** : 577-598.

[65] Head, Keith, John Ries, and Debrorah Swenson. (1995). “Agglomeration benefits and location choice : Evidence from Japanese manufacturing investments in the United States.” *Journal of International Economics* 38**(3-4)** : 223–247.

[66] Hecker, Daniel E. (2005). “High-technology employment : a NAICS-based update.” *Monthly Labor Review* 128**(7)** : 57–72.

[67] Helpman, Elhanan. (1998). “The size of regions.” In : D. Pines, E. Sadka, I. Zilcha (eds.), *Topics in Public Economics. Theoretical and Empirical Analysis*, Cambridge University Press, pp. 33–54.

[68] Henderson, J. Vernon. (1997). “Medium-sized cities.” *Regional Science and Urban Economics* 27**(6)** : 583–612.

[69] Holmes, Thomas J., and John J. Stevens. (2014). “An alternative theory of the plant size distribution, with geography and intra- and international trade.” *Journal of Political Economy* 122**(2)** : 369–421.

[70] Holmes, Thomas J. and John J. Stevens (2004). "Spatial distribution of economic activities in North America." In : J.V. Henderson, J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, vol. 4*. North-Holland : Elsevier B.V., pp. 2797–2843.

[71] Holmes, Thomas J., and John J. Stevens. (2002). "Geographic concentration and plant scale." *Review of Economics and Statistics* 84**(4)** : 682–690.

[72] Holmes, Thomas J. (1999). "Localization of industry and vertical disintegration." *Review of Economics and Statistics* 81**(2)** : 314–325.

[73] Jonkeren, O., Erhan Demirel, Jos van Ommeren, and Piet Rietveld. (2009). "Endogenous transport prices and trade imbalances." *Journal of Economic Geography* 11**(3)** : 509–527.

[74] Kerr, William R. (2008). "Ethnic Scientific Communities and International Technology Diffusion." *Review of Economics and Statistics* 90 **(3)** : 518–537.

[75] Kim, Sukkoo. (1995). "Expansion of markets and the geographic distribution of economic activities : The trends in U.S. regional manufacturing structure, 1860-1987." *Quarterly Journal of Economics* 110**(4)** : 881–908.

[76] Klier, Thomas, and McMillen, Daniel P. (2008). "Evolving Agglomeration in the U.S. Auto Supplier Industry." *Journal of Regional Science* 48**(1)** : 245–267.

[77] Koenig, Pamina, Florian Mayneris, and Sandra Poncet. (2010). Local Export Spillovers in France, *European Economic Review* 54**(4)**, 622–641.

[78] Koenig, Pamina. (2009). Agglomeration and the Export Decisions of French Firms, *Journal of Urban Economics* 66**(3)**, 186–195.

[79] Kolko, Jed. (2010). "Urbanization, agglomeration, and the coagglomeration of Services industries." *Agglomeration Economics.* NBER BOOKS, University of chicago Press, pp. 151–180.

[80] Krugman, Paul R., and R. Livas Elizondo. (1996). "Trade policy and the third world metropolis." *Journal of Development Economics* 49**(1)** : 137–150.

[81] Krugman, Paul R., and Anthony J. Venables. (1995). "Globalization and the inequality of nations." *Quarterly Journal of Economics* 110**(4)** : 857–880.

[82] Krugman, Paul R. (1991). Increasing Returns and Economic Geography, *Journal of Political Economy* 99**(3)**, 483–499.

[83] Lewbel, Arthur. (2012). "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models." *Journal of Business and Economic Statistics* 30**(1)** : 67–80.

[84] Lu, Jiangyong, and Zhigang Tao. (2009). "Trends and determinants of China's industrial agglomeration." *Journal of Urban Economics* 65**(2)** : 167–180.

[85] Lafourcade, Miren, Giordano, Mion. (2007). "Concentration, agglomeration and the size of plants." *Regional Science and Urban Econmics*, 37**(1)** : 46–68.

[86] Marcon, Eric, and Florence Puech. (2015). "A typology of distance-based measures of spatial concentration." *<halshs-00679993v4>*.

[87] Marcon, Eric, and Florence Puech. (2010). "Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods." *Journal of Economic Geography* 10**(5)** :, 745–762.

[88] Marcon, Eric, and Florence Puech. (2003). "Evaluating the Geographic Concentration of Industries using Distance-based Methods." *Journal of Economic Geography* 3**(4)** : 678–693.

[89] Marshall, Alfred. (1890). "Principles of Economics." *8th edition (1920). London, uk : Macmillan and Co., Ltd.*

[90] Martin, Roger L., Dean, Joseph, Milway, James. (2004). "Assessing the Strength of the Toronto Biopharmaceutical Cluster." *The Institute for Competitiveness & Prosperity.*

[91] Maurel, Françoise, and Beatrice Sedillot. (1999). "A Measure of the Geographic Concentration in French Manufacturing Industries." *Regional Science and Urban Economics* 23**(5)** : 575–604.

[92] Marcon, Eric, and Florence Puech. (2010). "Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods, *Journal of Economic Geography* 10**(5)**, 745–762.

[93] Marcon, Eric, and Florence Puech. (2003). "Evaluating the Geographic Concentration of Industries using Distance-based Methods", *Journal of Economic Geography* 3**(4)**, 678–693.

[94] Maurel, Françoise, and Beatrice Sedillot. (1999). A Measure of the Geographic Concentration in French Manufacturing Industries", *Regional Science and Urban Economics* 23**(5)**, 575–604.

[95] Mittelstaedt, John D., William A. Ward, and Edward Nowlin. (2006). "Location, Industrial Concentration and the Propensity of Small US Firms to Export", *International Marketing Review* 26**(2)**, 486–503.

[96] Mori, Tomoya, Nishikimi, Koji, and Smith, Tony E. (2005). "A Divergence Statistic for Industrial Localization." *The Review of Economics and Statistics* 87**(4)** : 635–651.

[97] Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura. (2014). Localized Knowledge Spillovers and Patent Citations : A Distance-Based Approach, *Review of Economics and Statistics* 96**(5)** : 967–985.

[98] Murata, Yasusada, Ryo Nakajima, and Ryuichi Tamura. (2014). "Testing for localization using micro-geographic data : A new approach." *In progress*, Nihon University, Japan.

[99] Nakajima, Kentaro, Saito, U. Yukiko, and Uesugi, Lichiro. (2012). "Measuring economic localization : Evidence from Japanese firm-level data." *Journal of the Japanese and International Economies*, 26**(2)** : 201–22.

[100] Openshaw, Stan. (1983). *The Modifiable Areal Unit Problem*. Geo Books : Norwick, Norfolk, UK.

[101] Openshaw, Stan, and Peter Taylor. (1979). A Million or so Correlation Coefficients : Three Experiments on the Modifiable Areal Unit Problem. In : Neil

Wrigley (ed.), *Statistical Applications in the Spatial Sciences*. Pion, London : pp. 127–144.

[102] Porter E. Michael. (1998). “On competition.” *Boston, Harvard Business Review Books*.

[103] Pyke, F., Becattini, G., and Sengenberger, W. (eds.). 1(990). “Industrial Districts and Inter-Firm Co-operation in Italy.” *Geneva : International Institute of Labour Studies*.

[104] Riedel, Nadine, Hyun-Ju, Koh. (2014). “Assessing the Localization Pattern of German Manufacturing and Service Industries." *Regional Studies,* 48 **(5)** : 823–843.

[105] Ripley, Brian David. (1977). “Modelling Spatial Patterns.” *Journal of the Royal Statistical Society*, B 39**(2)** : 172–212.

[106] Ripley, Brian David. (1976). “The Second-Order Analysis of Stationary Point Processes.” *Journal of Applied Probability*, 13**(2)** : 255-266.

[107] Redding, Stephen J., and Daniel M. Sturm. (2008). “The Costs of Remoteness : Evidence from German Division and Reunification." *American Economic Review*, 98**(5)**, 1766–1797.

[108] Rosenthal, Stuart S., and William C. Strange. (2010). “Small Establishments/Big Effects : Agglomeration, Industrial Organization and Entrepreneurship.” In : Edward L. Glaeser (ed.), *Agglomeration Economics* (NBER Book) : University of Chicago Press, pp. 277–302.

[109] Rosenthal, Stuart S., and William C. Strange. (2005). “The Geography of Entrepreneurship in the New York Metropolitan Area.” Processed, Syracuse University and Rotman School of Management, University of Toronto.

[110] Rosenthal, Stuart S., and William C. Strange. (2004). “Evidence on the nature and sources of agglomeration economies.” In : J.V. Henderson, J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, vol. 4*. North-Holland : Elsevier B.V., pp. 2119–2171.

[111] Rosenthal, Stuart S., and William C. Strange. (2003). Geography, Industrial Organization, and Agglomeration, *Review of Economics and Statistics* 85**(2)**, 377–393.

[112] Rosenthal, Stuart S. and William C. Strange. (2001). The Determinants of Agglomeration, *Journal of Urban Economics* 50**(2)** 191–229.

[113] Scholl, Tobias, and Thomas, Brenner. (2015). "Optimizing distance-based methods for large data sets." *Journal of Geographical Systems* 17**(4)** : 333–351.

[114] Scholl, Tobias, and Thomas, Brenner. (2014). "Detecting spatial clustering using a firm-level cluster index." *Regional Studies* 0**(0)** : 1–15, doi = 10.1080/003343404.2014.958456.

[115] Shane, Scott. (2009). Why Encouraging More People to Become Entrepreneurs is Bad Public Policy ? *Small Business Economics* 33**(2)**, 141–149.

[116] Silverman, Brian, S. (2002). "Technological Resources and the Logic of Corporate Diversification." *London,* U.K. *: Routledge Press* .

[117] Silverman, Bernard, W. (1986). "Density Estimation for Statistics and Data Analysis." *New York : Chapman and Hall.*

[118] Strange, W.C., W. Hejazi and J. Tang. (2006). "The Uncertain City : Competitive Instability, Skills, Innovation, and the Strategy of Agglomeration ?", *Journal of Urban Economics,* **(3)** : 331-351.

[119] Thompson, Peter, and Melanie Fox-Kean. (2005). "Patent Citations and the Geography of Knowdledge Spillovers : A Reassessment. ?" *American Economic Review* 95**(1)** : 450-60.