Agglomeration within an Urban Area

Stephen B. Billings*
and Erik B. Johnson†

January 24, 2014

Abstract

This paper utilizes a newly created index for colocalization to estimate the determinants of industrial agglomeration within a single urban area. Our new index directly incorporates the location of individual establishments relative to the population of all establishments to create this measure of spatial similarity between two industries. We estimate that proximity to transportation infrastructure and consumers as well as knowledge spillovers largely explain patterns of agglomeration. We find a smaller role for input-output linkages and consumption externalities for retail and consumer service industries. Results highlight the importance of accurately measuring industrial agglomeration.

JEL Classification: L14, L60, L80, R32
Keywords: Agglomeration; Colocalization; Urban Area Industrial Concentration

Acknowledgements: We thank Andrew Cassey, John Clapp, Gilles Duranton, Edward Glaeser, William Kerr, Scott Duke Kominers, Tomoya Mori, Alvin Murphy and Henry Overman for helpful comments as well as seminar participants at the NBER SI 2013 Workshop for Urban Economics, American Real Estate and Urban Economics Association 2014 Annual Meeting, University of Massachusetts-Lowell, the Urban Economic Association sessions for the European Regional Science Association 2012 and the North American Regional Science Association 2012 annual meetings. This paper previously circulated under the title "Localization and Colocalization within an Urban Area".

*University of North Carolina Charlotte, Department of Economics. Email: stephen.billings@uncc.edu
†Quinnipiac University, Department of Economics. Email: Erik.Johnson2@quinnipiac.edu

1 Introduction

Urban scholars have long tried to explain such phenomena as the clustering of information technology firms in Silicon Valley or auto manufacturing in Michigan. While there is a large literature on the agglomeration of manufacturing industries across cities, states or nations, there is a paucity of studies that examine the distribution of industrial activity within a single urban area.[1] This paper attempts to fill this gap by using the city as our spatial scale and by broadening the industrial sectors we examine to include both manufacturing and service.

For an establishment that is mobile across cities, the choice of location within an urban area may be viewed as the second stage of a two-stage decision process.[2] In the first stage, an establishment selects a region or city and in the second stage determines its location within an urban area. In choosing a site, establishments are influenced by agglomerative forces that vary within the urban area, which leads to substantially different patterns of spatial concentration relative to national location decisions. Industrial agglomeration within an urban area limits such traditional reasons for spatial concentration as improved labor matching, but does support other agglomerative forces such as consumption based externalities due to consumer shopping behavior or input producers serving an export-oriented industry (e.g. lawyers clustered around banks downtown).

For the most part, the second stage of the location decision process is confined to the domain of theory and has not been subject to empirical tests.[3] In terms of agglomeration, scholars focus on theoretically modeling site selection as a combination of the agglomerative forces due to consumer uncertainty and shopping trip costs and dispersing forces due to spatial competition in consumption-based sectors such as retail trade and consumer services.[4] These models predict a range of possible spatial distributions of establishments that vary by the strength of agglomerative forces as well as highlight the importance of place amenities that often relate to commercial density and industry attributes.

[1] See Glaeser, Edward (Ed.) (2010), Rosenthal and Strange (2003), McCann and Folta (2009) for reviews of the agglomeration literature.

[2] For establishments that lack mobility across cities, location decisions simply involve only the second stage of site selection. This lack of mobility is likely common to a number of small businesses whose ownership is tied to specific cities.

[3] Exceptions such as the work by Arzaghi and Henderson (2008) and Fu (2007) exist, but they usually focus on one industry (e.g. advertising) or a specific determinant of agglomeration (e.g. knowledge spillovers) within an urban area.

[4] See Eaton and Lipsey (1979), Konishi (2005), Sheppard et al. (1992), Mulligan and Fik (1994) and Anderson and Engers (1994)

While the focus of this paper is on industrial agglomeration within an urban area, the empirical literature provides extensive support that agglomerative forces play a key role in the spatial concentration of industries across cities or regions. Ellison et al. (2010) and Rosenthal and Strange (2001) find that input-output linkages, knowledge spillovers and labor sharing all positively contribute to industrial concentration within US manufacturing industries and a number of recent papers[5] lend support to these traditional Marshallan determinants of manufacturing agglomeration in the US and Europe.

In addition to furthering our understanding of the role of agglomerative forces, the literature provides methodological contributions to the accurate measurement of industrial concentration. Most of the metrics in this literature focus on localization, which involves scaling industry concentration by the general concentration of all manufacturing. Measures of localization have evolved from simple indices of inequality such as Gini coefficients to theoretically grounded measures such as those constructed by Ellison and Glaeser (1997) to the nonparametric pairwise point based measure of Duranton and Overman (2005). As scholars incorporate these indices in applied research, the relative strengths and weakness of different methods has been highlighted in the literature. For example, aggregated measures such as Ellison and Glaeser (1997) may suffer from the Modifiable Areal Unit Problem (MAUP) and lack statistical tests of significance. The continuous metric of Duranton and Overman (2005) overcomes the MAUP and offers statistical tests of significance, but may conclude statistically significant localization in relatively dispersed industries for study areas with multiple commercial centers.[6] While these measures may have shortcomings, they are continuously refined and improved upon by researchers.[7]

Beyond own industry spatial concentration, Ellison et al. (2010) discuss the advantages of using across industry spatial relationships in order to better identify determinants of agglomeration. Examining across industry relationships can be quantified as colocalization or the degree of spatial similarity between two industries relative to that of general industry concentration. Duranton and Overman (2005) and Ellison and Glaeser (1997) both develop colocalization indices as an extension of their original localization indices. Conceptually,

[5] e.g. Jofre-Monseny et al. (2011), Martin et al. (2011), Greenstone et al. (2010) and Rosenthal and Strange (2008)

[6] Billings and Johnson (2013) show the potential for false positive localization with the Duranton & Overman index and that it is directly related to the distance between commercial centers and the choice of pairwise distance threshold.

[7] For example, Cassey and Smith (2012) outline a bootstrap based method to assign statistical significance to the Ellison and Glaeser (1997) index and Barlet et al. (2012) highlight and correct for potential bias due to industry sample size in the Duranton and Overman (2005) index.

these extensions contain similar strengths and shortcomings as localization indices but do contain some unique properties that have been underexplored in the literature. Controlling for the 'dartboard effect' is more complicated under colocalization and it is unclear if one should use all establishments, the joint distribution of the establishments within the industry pair or some combination thereof to create a relevant counterfactual. Furthermore, Ellison et al. (2010) show that estimates of the determinants of agglomeration are sensitive to the use of the Duranton and Overman (2005) or Ellison and Glaeser (1997) indices.

Our contribution to the literature is twofold. First, we introduce a new colocalization index with a number of desirable attributes. Our index represents an innovation over Ellison and Glaeser (1997) by incorporating establishment level data and advances Duranton and Overman (2005) by directly incorporating the places that contain spatially similar industries. We construct our index by treating the location of all establishments in an industry as a spatial density distribution and compare the bivariate distributions of two industries using an established statistical tool, the Wasserstein distance. We scale the Wasserstein distance between two industries by general industry concentration to construct an index based on p-values. One can naturally interpret our colocalization index as the probability of correctly rejecting the null hypothesis that a given industry colocates with a randomly located industry. From descriptive results, we show that statistically significant (at 5% level) colocalization is relatively unusual within an urban area and limited to 8.1% of all ordered pairs of industries. Four digit industries in the Transportation and Warehousing and Manufacturing sectors exhibit the greatest prevalence of colocalization while industries in the Retail Trade and Accommodation and Food Services sectors rarely colocate with other industries.

Our second contribution incorporates our new index of colocalization to formally estimate the role of natural advantage, traditional Marshallan determinants of agglomeration and consumption externalities in explaining urban area patterns of industry concentration. Results show a strong positive role for access to transportation infrastructure, access to consumers and knowledge spillovers in determining industrial agglomeration within an urban area. We find smaller positive effects from input-output linkages and consumption externalities. When we increase our measures of natural advantage, input-output relationships, knowledge spillovers, labor pooling/sharing and trip chaining by one standard deviation, our new colocalization index increases by 0.475 standard deviations.

How one measures industrial agglomeration between industries matters. The magnitude and significant of coefficients varies when using our new colocalization index, the Ellison and Glaeser (1997) index or the Duranton and Overman (2005) index. This variation is

attributable to data aggregation, the choice of geographical units used to quantify industrial concentration as well as how each index controls for the general concentration of industry. Results are robust to just focusing on service industries as well as models that weight by industry size or limit analysis to only within sector industry pairs. Conclusions highlight the larger role of natural advantage and knowledge spillovers and the smaller role of input-output linkages and labor sharing in explaining industrial agglomeration at the urban area scale relative to national studies.

Our paper continues by describing how we measure the spatial concentration of specific industries in Section 2. We then detail the construction and properties of our new index for colocalization in Section 3. Estimating various determinants of agglomeration from the literature as well as new measures of consumption externalities is the focus of Sections 4 and 5. We compare results to the commonly used Ellison and Glaeser (1997) (E-G) and Duranton and Overman (2005) (D-O) indices in Section 5 with conclusions drawn in Section 6.

2 Dataset of Establishments

The examination of industrial concentration within the small scale of a single urban area requires the use of spatially disaggregate data. We begin with a dataset of 79,038 individual establishments in the Denver-Boulder-Greeley CMSA.[8] Specifically, we take a fourth quarter extract from the 2006 Quarterly Census of Employment and Wages (QCEW) Program (formerly know as ES-202) dataset. The QCEW program produces a comprehensive tabulation of employment and establishments for workers covered by state unemployment insurance laws and includes the geographical coordinates of each establishment.[9] In Colorado, any business that paid wages of at least $1,500 in any quarter of the previous year, or employed at least one person for any part of a day for 20 weeks during the previous year must pay state unemployment insurance and thus is included in our dataset. This dataset provides the latitude and longitude of each establishment for all manufacturing and service industries. We focus our analysis on all industries with at least 10 establishments and classified in the range of NAICS 3111 through NAICS 8139.

We define this set of establishments as our population of businesses within a single urban area and measure spatial relationships for individual industries based on an establishment's

[8] This urban area contains 2.6 million people over 13,679 square kilometers.

[9] Excluded employees include members of the armed forces, the self-employed, proprietors, domestic workers, unpaid family workers, and railroad workers covered by the railroad unemployment insurance system.

point location. Our analysis only focuses on establishments rather than individual employees because employment location is dependent on establishment location. Therefore, we prefer not to treat each employee as an independent unit.[10] We can visualize the full population of all 79,038 establishments in Figure 4a. The main area of establishment concentration is centered on downtown Denver and extends to secondary commercial centers in the south, west and northwest. Figure 4 b provides shows NAICS 5411 - Legal Services. From this figure, one can see the large number of establishments concentrated in downtown Denver and also that this industry generates greater concentration downtown relative to the population of all establishments.

3 Colocalization Index

Our new measure of colocalization is based upon the premise that colocalization, or the tendency of establishments from two different industries to locate together, can be modeled as the similarity of spatial density distributions between two industries. Our measure of colocalization assumes that establishments locate in order to maximize profits based on both natural advantage and spillovers across industries. This idea is formalized in the theoretical model of Ellison et al. (2010) and we approach our index of colocalization as simply a tool to improve the measurement of spatial relationships between establishments.

Existing indices often characterize colocalization by reducing the dimensionality of establishment locations within a study area. For example, the Ellison and Glaeser (1997) (E-G) index assigns establishment concentration to areal units such as states or counties while Duranton and Overman (2005) (D-O) use pairwise distances between establishments. In either case, the (X,Y) coordinates of establishments are being reduced to individual geographic units or pairwise distances. This aggregation of information allows for desirable measures of spatial similarity between industries, but the E-G index does omit information about establishments in adjacent geographies. The D-O index provides distances between establishments and does not depend on the aggregation of establishments to geographies. Since the D-O index is built on the pairwise distance between establishments, it removes any information about specific locations and is unable to incorporate information on the degree to which establishments line up with employment centers. Apriori, it is not clear how

[10] Rather, heterogeneity in employment across establishments in the same industry may represent substitution between factors of production, productivity differences or heterogeneity in establishment output within the same industry category. Ideally, we would jointly model establishment and employment, but leave this issue for future research.

important this omitted information is regarding the measurement of colocalization, but the fact that results show significant variation when Ellison et al. (2010) incorporate the E-G versus D-O indices suggest that it may be a nontrivial issue.[11]

Therefore, we turn to an established metric for comparing the degree of similarity between two distributions that retains the spatial relationships between industries in two dimensions: the Wasserstein metric.[12] The Wasserstein metric is calculated by measuring the distance between two density functions in a given space. We define our space as the Denver-Boulder-Greely CMSA and our distributions as the unique spatial density for the set of coordinates $\mathbf{x}$ representing establishments in a given industry j or k.

Formally, let f_j and f_k be two functions that characterize the spatial distribution on $\mathbb{R}^2$ for industries j and k and let the area under each density function be normalized to one. A map $M : \mathbb{R}^2 \to \mathbb{R}^2$ realizes a transfer of f_j to f_k if, for all bounded subsets A of $\mathbb{R}^2$,

$$\int_{\mathbf{x}\in A} f_k(\mathbf{x})dx = \int_{M(\mathbf{x})\in A} f_j(\mathbf{x})dx \tag{1}$$

Using this, we can define the Wasserstein distance as

$$W(\hat{f}_j, \hat{f}_k) = \inf_M \int |M(\mathbf{x}) - \mathbf{x}| f_j(\mathbf{x})dx \tag{2}$$

for all maps M which transport f_j to f_k.

Conceptually, one can visualize the Wasserstein distance (W) as the minimum 'cost' of turning one distribution into the other, where the cost is density moved times distance.[13] In the case of a set of points assigned to two dimensional space, one can visual the Wasserstein distance as the number of points times distance traveled to place each point in one industry on top of the closest point in another industry. For two industries with the same number of establishments, one would calculate the shortest aggregate distance to move all points in industry j to coincide with the points in industry k. Since we want to compare industries which vary in the number of establishments, we need to normalize our spatial distribution of points into a density function such that each point in industry j is given a density equal to $\frac{1}{N_j}$ and each point in industry k is given a density equal to $\frac{1}{N_k}$ where N indicates the

[11] These issues are formally discussed and tested with respect to localization in Billings and Johnson (2013), where the D-O index may be problematic in a study area with multiple city centers and E-G is sensitive to the number of geographic units.

[12] Kantorovich (1940), Kantorovich and Rubinstein (1958) and Wasserstein (1969)

[13] This metric is often termed earth mover's distance in computer science because it can be visualized as moving dirt from one pile to another in order to create two identical piles of dirt.

number of establishments in a given industry j or k. These densities are represented by spatial density functions $f_j(\mathbf{x})$ and $f_k(\mathbf{x})$.

We let $W_{j,k}$ represent the Wasserstein distance between industry j and industry k. Where $W_{j,k}$ involves a computational algorithm that minimizes the total amount of density and distance in moving industry j's density distribution to become the density distribution of industry k. We compute $W_{j,k}$ for all pairwise permutations of industries for which $j \neq k$ and measure distance in kilometers based on the Euclidean distance between points representing the location of each establishment. Figure 1 provides a visualization of the Wasserstein distance algorithm with the dark area for industry j being moved to the right to become identical to industry k. $W_{j,k}$ is this context would be the area of the dark rectangle times the distance moved.

Converting this measure of spatial similarity into a colocalization index requires comparing $W_{j,k}$ to a counterfactual of industry j's spatial similarity to a randomly located (pseudo) industry k. In this case, industry j may have a number of agglomerative forces that influence its spatial density, but pseudo industry k is not subject to any agglomerative forces including benefits from locating proximate to industry j. Our null hypothesis is no spatial similarity between industry j and industry k conditional on the spatial density of industry j.[14]

Similar to Duranton and Overman (2005) and Billings and Johnson (2012), we construct our counterfactual of randomly located (pseudo) industries based on two specific criteria: 1) the sample should be drawn from the set of locations where a establishment could potentially locate, and 2) the sample size used in constructing the counterfactual must be equal to the number of establishments in the industry. This strategy helps control for undevelopable land as well as other unobservable constraints on industrial location. In selecting the set of locations for our counterfactual, we restrict pseudo industry construction to the sites of our full population of establishments. Therefore, we allow for mixed commercial land-use as well as the possibility of commercial land-use regulation to be influenced by a establishment's location decision.[15] For each ordered pair (j, k), we construct pseudo industry $\tilde{k}$ based on randomly selecting N_k locations from the set of all establishments.[16]

Therefore, our counterfactual measure of spatial similarity ($W_{j,\tilde{k}}$) incorporates actual

[14] One could also model colocalization conditional on the spatial density of k and we later include this model by comparing industry j to industry k as well as industry k to industry j in our measure of colocalization.

[15] There are a number of possible ways to select our counterfactual choice set based on different industrial classification, but we use the least restrictive counterfactual definition to allow us to compare spatial relationships across manufacturing and service industries.

[16] Restricting out counterfactual to the same number of establishments as our industry of interest accounts for any variation in the estimated density due to the sample size of the point process.

industry j compared to pseudo industry $\tilde{k}$ of size N_k. The computation of $W_{j,\tilde{k}}$ is repeated for a $1,000$ pseudo industries to generate an empirical null distribution for colocalization of industry j to pseudo industry $\tilde{k}$. We construct our colocalization index by determining the number of pseudo industries for which $W_{j,k}$ is less than $W_{j,\tilde{k}}$. If $W_{j,k}$ is less than $W_{j,\tilde{k}}$ 950 times then we assign our index as $Coloc_{j,k} = 0.95$. Therefore, the magnitude of our colocalization index represents the statistical significance of spatial similarity for industry j to industry k. Even though $W_{k,j} = W_{j,k}$, $Coloc_{j,k} \neq Coloc_{k,j}$ since the null distributions are direction specific and will vary due to differences in the degree to which industry j and industry k are localized as well as how an industry varies from the population of all establishments. This process is repeated for all $(201 * 200) = 40,200$ industry ordered pairs with corresponding industry specific null distributions. Results provide two p-values of colocalization for each industry pair, one for each direction of colocalization.[17]

3.1 Properties of the Index

To justify the development of a new colocalization index, we highlight the properties of $Coloc_{j,k}$. Our index shares the desirable properties of existing colocalization indices including enabling comparisons across industries, controlling for the overall concentration of industrial activity and grouping industries by production oriented classification. The dimensions for which indices tend to vary is in terms of spatial aggregation or the MAUP, measures of statistical significance and the ability to retain place specific information. We now focus on the these later dimensions to better understand the properties of $Coloc_{j,k}$.

A concern with existing indices is the appropriate scale upon which to consider spatial concentration. The Ellison and Glaeser (1997) (E-G) index uses geographical units such as states or counties as the distance criteria or scale upon which to measure industrial agglomeration. However, the ability to avoid any aggregation to geographical units requires not only disaggregate data, but the construction of an index that directly incorporates point based data in its construction. By construction, the E-G index requires the aggregation of points into geographical units and thus is unable to avoid this type of aggregation. The D-O index avoids issues related to MAUP by incorporating point data, but requires an explicitly

[17] We considered generating a more traditional index which would allow us to rank industries based on the relative values of $(W_{j,\tilde{k}})$ and $(W_{j,k})$. For example, one could have created a measure of colocalization equal to $\frac{W_{j,\tilde{k}}}{W_{j,k}}$ or $W_{j,\tilde{k}} - W_{j,k}$. Our results are sensitive to the choice between these two scaled Wasserstein measures of colocalization. Furthermore, ranking industries by either of these scaled Wassersteins are dominated by industries with large values of $W_{j,\tilde{k}}$. We even tested these scaled Wasserstein indices in later regression analysis and consistently found weaker coefficients in terms of magnitude as well as significance.

stated pairwise distance criteria in order to conclude localization or colocalization. These assumptions regarding the scale of agglomeration have nontrivial consequences. Duranton and Overman (2005) (D-O) finds as little as 39% of UK manufacturing industries localized at a distance criteria of 5km and as much as 52% using a median pairwise distance criteria of 180km. Using the D-O colocalization index, Ellison et al. (2010) find between 87% and 99% of industry pairs to be colocalized for 250 mile and 1,000 mile criteria. Additionally, one may be concerned that using larger distance criteria may capture the distance between employment centers or industry clusters.[18]

Since our index directly incorporates point data, we avoid any concern regarding the MAUP. We also avoid the use of adhoc distance criteria by measuring colocalization as the degree to which the spatial similarity of an industry pair deviates from an industry colocating with a randomly located (pseudo) industry. Our measure of colocalization removes the choice of distance criteria and frames the index in terms of statistical significance. Therefore, our index indicates the probability of correctly rejecting the null hypothesis of an industry being spatial similar to a randomly located industry. An index based on p-values allows statistical tests of significance without imposing a significance level to generate non-zero index values. The cost of basing our index in terms of spatial similarity is that we have no measure of the distance between establishments, and thus we cannot conclude anything about the distance up to which industries cluster together.[19]

Beyond the properties of scaling our measure of colocalization in terms of statistical significance and no apriori distance criteria, our measure of colocalization offers some additional attributes. First, our index is computationally easier than the Duranton & Overman (D-O) colocalization index because it does not require the creation of a large number of pseudo industries to identify global confidence bands in order to control for multiple hypothesis testing across pairwise distances. The computational burden of computing the D-O colocalization index is nontrivial and for implementation requires a subset of all pairwise industries, a random sampling of large industries and other simplifying assumptions.[20]

Second, existing indices for colocalization are symmetric in construction with the colocalization of industry j to industry k generating identical measures of colocalization as industry

[18]For example, the distance from New York City to Washington, DC, Philadelphia or Boston are all less than the 250 mile distance criteria incorporated by Ellison et al. (2010).

[19]Existing work by Duranton and Overman (2005) and Duranton and Overman (2008) and recent work by Kerr and Kominers (2010) provide some insight into the scale upon which industries cluster.

[20]Both Duranton and Overman (2005) and Ellison et al. (2010) compute colocalization for pairwise 3 and 4 digit manufacturing industries using a number of simplifying assumptions and sampling of establishments within an industry/subset of industries and more than a month of computing time.

k to industry j. Recent work by Leslie and Kronenfeld (2011) shows that the direction of comparisons between sets of points generates different expected values as well as critical values for a class of spatial association metrics based on nearest neighbors. These directional differences are directly related to the number of points in each set. Given that any colocalization index involves comparing establishments between two industries of different size, this issue may be problematic in existing indices. We avoid this concern because our new colocalization index is non-symmetric and incorporates counterfactuals that are consistent with the direction of the comparison. This property also allows us to incorporate the direction of input-output flows in discussing the determinants of colocalization.

Overall, we find that 8.1% of all industry ordered pairs are colocalized at the 5% significance level ($Coloc_{j,k} \geq 0.95$) and 44.5% of colocalized ordered pairs are colocalized in both directions. This symmetry in colocalization among industry pairs is not suprising given that as industry j and industry k converge in space, $Coloc_{j,k}$ converges to $Coloc_{k,j}$. Table 1 provides a basic summary of colocalization based on significant ordered pairs of colocalization for three major industry sectors.[21] One can see that same sector colocalization is more prevalent than cross industry sector colocalization and that results are not symmetric between industry sectors. Non-symmetric results are due to the degree to which industry j or k's spatial density follows the density of all establishments. Figure 2 highlights this point by showing two cases for two industries. Both cases hold constant the spatial similarity between industry j and industry k, but vary in their relationship to the population of all establishments. In the top case, both industries are spatially similar to the population of all establishments, which gives the counterfactual of general industry concentration a high degree of spatial similarity. Therefore, two industries have a higher expected degree of spatial similarity due to chance and thus require greater spatial similarity for us to conclude colocalization. In the bottom case, industry j is spatially dissimilar from the population leading one to conclude colocalization even when industry j and k do not exhibit a high degree of spatial similarity. Thus, our index treats the case of establishments clustering in suburban or exurban portions of our city as more colocalized than the same amount of clustering within commercial centers.

This directional component of our colocalization index is observed by the fact that only 2.2% of ordered pairs in the non-business services sector colocalize with manufacturing industries, while 11.4% of ordered pairs in manufacturing colocate with non-business services

[21] We classify NAICS 3111 through NAICS 3399 as Manufacturing; NAICS 4231 through NAICS 4251 as well as NAICS 4811 through NAICS 6244 as Business Services; and NAICS 4411 through NAICS 4543 as well as NAICS 7111 through NAICS 8139 as Non-Business Services.

industries. Thus, our measure of colocalization takes into account the possibility that manufacturing may colocate with non-business services while non-business services tend to almost never co-locate with manufacturing. These different results are consistent with variation in the spatial distribution of non-business services which follows the overall population distribution while manufacturing locates away from major commercial centers. For example, colocalization is not symmetric when comparing 5411 Legal Services with 7224 Drinking Places. When we assign 5411 Legal Services to be industry j and 7224 Drinking Places to be industry k, we find significant colocalization. This relationship indicates that lawyers are located significantly closer to bars than the overall population of all industries. Interestingly, the reverse case of Drinking Places locating proximate to bars is not significantly colocalized. In this case, Drinking Places locate towards a wide range of potential customers, not just lawyers. The fact that Drinking Places more closely follows the spatial distribution of all industries weakens the colocalization of Drinking Places to Legal Services.

3.2 Colocalization within an Urban Area

Across all industries, we find a mean colocalization index of 0.43 with a standard deviation of 0.33. Of the 8.1% of colocalized industry ordered pairs, 407 (1%) are assigned a colocalization value of one and $1,630$ (4%) have a colocalization value of zero. The large number of industry ordered pairs assigned values of zero is not surprising given that we include industries paired across the manufacturing and service sectors.[22] Figure 3 provides the distribution of $Coloc_{j,k}$ and highlights the higher densities at the tales of the distribution indicating the role of both dispersion and agglomeration in this dataset. The former case of dispersion is larger, but is primarily driven by comparison across industry sectors and the dispersion of land-use across the city.

Table 2 provides the percentage of industry pairs with statistically significant colocalization ($Coloc_{j.k} \geq 0.95$) by the two digit industry sectors of industry j. The Transportation and Warehousing industries generate the largest portion of industry pairs with significant colocalization which is consistent with trucking and warehousing facilities clustering around Interstates and rail infrastructure. Since other industries locate for transportation access, colocalization may simply be due to shared natural advantage. Other sectors with higher amounts of colocalization include Finance and Insurance as well as Information. These indus-

[22]The colocalization of industry pairs across sectors may be limited since manufacturing establishments may never locate at the same sites as service establishments due to land use restrictions. We may still find these comparisons meaningful due to consumption externalities or input-output relationships. For example, restaurants may cluster next to manufacturing plants to serve workers lunch.

tries are typically discussed as higher skilled business service industries that may benefit from a number of traditional determinants of agglomeration including knowledge spillovers, input-output linkages or shared labor. For example, NAICS 5182 Data Processing, Hosting and Related Services may benefit from being proximate to their clients, NAICS 5232 Securities and Commodities Exchanges. The small portion of colocalized industries in Accommodation and Food Services and Retail Trade is consistent with the presence of a number of industries that spread through the urban area to serve a range of businesses and consumers and thus do not generate statistically significant differences from the population of all establishments.

In order to explore the properties of our colocalization index, we compare our results to the Duranton-Overman (ψ) and Ellison-Glaeser (λ) indices. We provide details of the construction of the D-O colocalization index and the E-G coagglomeration index in Appendices A and B. We provide two values for each index to indicate two geographical units for E-G and two distance criteria upon which to conclude colocalization for D-O. We correlate $Coloc_{j,k}$ with these four indices in Table 3. Our index has the largest correlation with the E-G index for zip codes at 0.31 and the lowest correlation with the D-O index using a 10 km distance criteria at 0.09. This range of correlations is similar when comparing the E-G and D-O indices with the smallest correlation of 0.03 between the D-O index using 30km (median pairwise distance) and the E-G index for zip codes. The largest correlation is 0.24 for the D-O index using a 10 km distance criteria and the E-G index for CBGs. The fact that correlations across indices are not that strong is indicative of the nature of measuring spatial relationships, where a number of index attributes influence results.

To further compare these indices, we report the portion of four digit industries that are considered colocalized. $Coloc_{j,k}$ finds 8.1% of industry pairs to be colocalized, while using the D-O index, we find a range of results that depend on the distance criteria. For example, we find as little as 8.7% using a distance criteria of 1km and as much as 57.1% using the median pairwise distance of 30 km.[23] The stronger correlation and similarity in number of significant ordered pairs using the smallest D-O distance criteria indicates that smaller pairwise distances approach the results of $Coloc_{j,k}$. The use of smaller distance criteria is similar to the nature of $Coloc_{j,k}$ because any differences in the dimensionality of information about the location of an establishment is minimized at small pairwise distances. In the case of the E-G index, we find between 4.8% and 15.4% of industry ordered pairs to be colocalized for zip codes and census block groups respectively based on a cutoff of greater than 0.01.[24]

[23]This is considerably less colocalization than Ellison et al. (2010) because we include service industries as well as adopt a completely different scale upon which to measure colocalization.

[24]See Table E.1 for summary statistics of the E-G index.

4 Agglomerative Forces

Since industry classification is production oriented, it does not highlight specific agglomerative forces. Therefore, we estimate the impact of hypothesized determinants of industrial agglomeration to our urban area measure of colocalization. We include traditional Marshallan determinants of agglomeration such as input-output linkages, labor pooling/matching and knowledge spillovers as well as variables to quantify consumption externalities such as comparison shopping and multi-purpose shopping behavior. We also provide three different natural advantage factors that are unique to our scale of analysis: access to transportation infrastructure, access to consumers and industrial mix.

4.1 Input-Output Relationships

One of the most concrete benefits from establishments locating near one another is to reduce transport costs for commodities used as inputs in the production of other goods or for consumers traveling to a business to obtain a final product. Initially discussed by Marshall (1920) and expanded as part of "new economic geography" models (Krugman (1999)), the reduction in transport costs of raw inputs and finished products is a large driver of industrial agglomeration. The presence of input and output relationships at the urban area scale may still drive industrial agglomeration as establishments colocate to reduce distances for intraurban transport or access. Beyond standard manufacturing relationships, service industries may require personal interaction to facilitate input-output exchanges and thus benefit from spatial distances as small as neighboring businesses. One can imagine the provision of tax consulting services to banks likely requires a substantial personal presence and numerous face-to-face transactions.

To quantify input-output relationships between industries, we use the 2002 Benchmark Input-Output (IO) tables from the Bureau of Economic Analysis (BEA). The input-output tables provide a correspondence to determine commodity flows between most pairs of four digit NAICS industries while the remaining industries incorporate two or three digit NAICS input-output relationships.[25] We focus our analysis on direct input-output flows between industries and incorporate the fact that we assign directionality to our colocalization index.

In terms of colocalization, we develop a directional measure of input-output linkages

[25]The BEA provides a correspondence between commodities and NAICS industry classifications. In most cases, we use specific four digit industries or aggregate five and six digit classifications to four digit levels. In the remaining cases, we assign four digit industries their associated two or three digit NAICS IO relationship. See Appendix C for more details.

($ColocIO_{j\to k}$) by making the assumption that input suppliers benefit from locating proximate to their purchasers.[26] In this context, ($ColocIO_{j,k}$) provides the share of industry k's commodity input value that is attributed to the output of industry j. ($ColocIO_{j,k}$) averages 0.009 with a maximum of 0.84 for NAICS 5231 - Securities & Commodity Contracts Intermediation & Brokerage to NAICS 5251 - Insurance & Employee Benefit Funds and 27.4% of industry ordered pairs contain a value of 0.

4.2 Labor Markets

The second highly cited determinant of agglomeration is a reduction in the cost of labor through a large labor pool or better matches between a job's needed skill set and a worker's abilities.[27] Reduced labor costs due to industrial clustering occur because large labor pools minimize the risk to employers and workers from productivity changes in individual establishments that may lead to changes in employment levels. A large pool of establishments in the same industry minimize the risk of longer term job vacancies or loss due to individual establishments. In the end, these theories predict benefits from colocalization in industries that employ similar types of workers. Even though our analysis focuses on a single urban area, residential sorting to jobs and the existence of neighborhood referrals for job opportunities (Bayer et al. (2008)) may impart labor market benefits from agglomeration.

In order to operationalize similar worker types between pairs of industries, we incorporate data from the 2002 National Industrial-Occupation Employment Matrix (NIOEM) published from the Bureau of Labor Statistics (BLS). NIOEM provides industry level employment for approximately 395 NAICS industries and 980 occupations. We aggregate to 121 four digit occupation groups given the similarity in skill sets between some of the highly disaggregate occupation codes. Similar to Kolko (2010), our measure of occupational similarity between industry j and k is given by

$$LaborSim_{j,k} = \frac{2 - \sum_l |occ_{jl} - occ_{kl}|}{2} \tag{3}$$

with occ_{jl} indicates the share of industry j's workforce in occupation l. Therefore, $LaborSim_{j,k}$ equals one if industries j and k contain identically distributed occupations and zero for non-

[26] Of course, one could assume that purchasers may locate proximate to their input suppliers. Given a competitive selling environment, the assumption of sellers moving proximate to buyers appears more realistic, especially in the case of consumer based industries where the development of suburban retail malls closer to residential development has become commonplace. We also highlight the reverse case in our analysis since we include each industry pair twice, once as (j, k) and again as (k, j),

[27] See Helsley and Strange (1990) and Ellison et al. (2010)

overlapping distributions of occupations between industries j and k. $LaborSim_{j,k}$ averages 0.35 with a maximum of 0.66 for NAICS 6114 - Business & Computer & Management Training with NAICS 6117 - Educational Support Services.

4.3 Knowledge Spillovers

Beyond labor matching and pooling, knowledge may be shared between workers and establishments across industries. The spatial concentration of industries may facilitate the learning of new skills as well as lower costs related to the transfer of new ideas. Scholars often cite the important of knowledge spillovers and information exchange in explaining industrial concentration especially in higher technology industries.[28] There are two main approaches to measure knowledge spillovers: 1) linking detailed data on patents from the United States Patent and Trademark Office (USPTO) to industry categories or 2) using occupational categories and worker education levels to proxy for knowledge sharing. The former is preferable given that patents and their corresponding citations represent tangible evidence of knowledge and information sharing, but patent technology classifications are difficult to match to industry classifications and previously have been limited to probabilistic matching and three-digit SIC manufacturing industries. In order to include all of our four digit NAICS industries, we adopt the second approach and follow Kolko (2010) by using a simple measure of potential for knowledge spillovers based on the share of an industry's workers that have graduate college degrees ($Grad(\%)$).[29] Highly educated workforces are more likely to be involved in technology or knowledge activities and thus more likely to share ideas or specialized training with other workers.

In order to construct our variable for knowledge spillovers, we simply average $Grad(\%)$ for industry j and industry k to create $\overline{Grad}(\%)_{j,k}$. We then multiply $\overline{Grad}(\%)_{j,k}$ with $LaborSim_{j,k}$ to generate our measure of knowledge spillovers ($KnowSpill_{j,k}$). In our dataset, $KnowSpill_{j,k}$ has a mean of 0.17 with a standard deviation of 0.07 and a maximum of 0.60 shared with a number of industry pairs in the Information and Education Services sectors.

[28] See Jaffe et al. (1993), Audretsch and Feldman (1996), Saxenian (1996)

[29] We used a tabulation of educational attainment by occupation from the 2002 National Industrial-Occupation Employment Matrix (NIOEM) to determine the percent of workers with graduate degrees by industry.

4.4 Natural Advantage

As discussed by Ellison and Glaeser (1997), place specific factors (natural advantage) may influence industrial concentration. For example, establishments in NAICS 4841 General Freight Trucking only locate proximate to interstate highways and thus spatial concentration in this industry or with another industry may simply be due to the location of transportation infrastructure. The corresponding benefits of reduced transportation costs is likely industry specific and thus a major agglomerative force for only some industries. At the urban area scale, proximity to consumers may be an important attribute of location choices for establishments in retail and consumer service industries. Locating close to consumers will lower consumer travel costs and allow stores near consumers to provide lower cost items at the same retail price thus increasing profitability.

Beyond transportation costs and access to consumers, establishments may benefit from the mix of industries within a location. This story of economic diversity was made popular by Jacobs (1969) and is often touted as an important component of urban growth and development. Beyond just commercial density, the presence of multiple types of industries may highlight additional place specific benefits on an establishment's productivity. Since we cannot isolate what attributes of industrial diversity may influence specific industry spatial concentration, we use this variable as a residual for other elements of natural advantage beyond transportation infrastructure and access to consumers. Given the presence of Marshallan externalities and a number of place specific factors that influence locations decisions, we want to disentangle the contribution of multiple sources of natural advantage from the Marshallan factors discussed above.

In order to measure natural advantage along these three dimensions, we construct a measure of colocalization based on the location of establishments if industry location was based only on natural advantage. Therefore, we conduct a series of regressions where we allow location variables to predict the number of establishments for a given industry in a given Census Block Group (CBG). This first stage model estimates the pattern of establishments for a given industry based solely on natural advantage. In our case, we construct our measure of natural advantage by estimating separate binomial regressions for each industry for a cross section of CBGs that contain at least one establishment (in any industry). Our dependent variable for these first stage regressions is establishment counts for each CBG in each industry. Each regression provides a predicted number of establishments for a given industry based only on location variables. We do this series of 201 industry regressions uniquely for access to transportation variables (distance to Interstate, distances to railroads and their quadratics);

access to consumers variables (population density, aggregate income); and industrial diversity (Herfindahl index based on the share of establishments in each industry within a CBG). These predicted points are then used to construct our colocalization index and thus represent the share of colocalization determined by natural advantage.[30] Since we estimate our first stage predicted distribution of establishments uniquely for each industry, we allow each set of location variables to vary by industry. Therefore, access to transportation infrastructure may be a strong predictor of establishments in NAICS 4841 General Freight Trucking, but may only weakly predict establishments in NAICS 5411 Legal Services.

Table 4 provide correlations between the $Coloc_{j,k}$ based on actual establishments and three measures of colocalization that indicate the degree of colocalization if only location variables determined the spatial distribution of establishments within an industry. According to simple correlation coefficients, access to transportation infrastructure ($NATrans_{j,k}$) explains up to 13.4% , access to consumers ($NAPop_{j,k}$) explains up to 12.4% and industrial mix ($NAHerf_{j,k}$) explains up to 8.8% of the variation in colocalization across industry ordered pairs. These three natural advantage variables do vary from one another with correlations of between 0.24 and 0.26.

4.5 Consumption Externalities

Since we hypothesize that urban area agglomeration is influenced by the relationship between consumers and businesses within an urban area, we want a way to measure the importance of consumers shopping behavior on the location of businesses. At the urban scale, consumers frequently visit establishments in the retail trade (NAICS 44-45) and consumer services industries (NAICS 71-81) as part of shopping trips. The spatial proximity of stores across these industries may impart externalities on consumers by minimizing the cost of shopping trips as consumers visit multiple stores in a single trip. Gould et al. (2002) highlight these consumption externalities in the form of capitalized rents for tenants of retail shopping malls. These same spatial relationships are not confined to shopping malls and likely generate a number of benefits to establishments that locate near complementary store types. The literature formalizes consumption benefits across stores in different industries as trip–chaining, which is characterized as consumers lowering travel costs by visiting multiple stores selling a variety of goods in the same shopping trip.[31] This shopping behavior has been modeled

[30] We provide some additional details on the construction of our natural advantage variables in the appendix.

[31] See Thill and Thomas (1987) for a more detailed discussion of this literature.

by Ingene and Ghosh (1990) and Anas (2007) and highlights the potential for travel cost reduction from multi-purpose shopping.

In order to operationalize trip–chaining, we construct a measure of purchase frequency for a range of consumer goods. The premise being that goods consumed in similar frequencies are more apt to being the subject of multistop shopping as consumers need to purchase goods at about the same intervals of time. We define shopping frequency for a good,$Freq_j$, to equal the portion of respondents that purchased a good produced by industry j over a three month period. To generate $Freq_j$, we take a detailed extraction from the Census 2002 National Survey of Consumer Expenditures. This dataset provides the frequency of purchase for over 200 consumption good categories and ask respondents to include the amount of each good that was purchased in the last three months. We match each consumption good category to its corresponding industrial classification based on the description of consumption goods and industry classifications.[32]

Our measure of trip–chaining indicates the similarity in frequency of purchase between a pair of industries. For example, we would expect greater trip–chaining benefits between the purchase of groceries and gasoline than between groceries and a new car. The former pair of industries, Grocery Stores (NAICS 4451) and Gasoline Stations (NAICS 4471), involve frequent consumer visits and thus customers are more likely to need gasoline as well as groceries at the same time and thus could benefit from a shopping trip that purchases both goods. In the case of groceries and a new car (NAICS 5511), consumers are unlikely to decide to buy a new car while out shopping for groceries. Consumer goods such as new cars and electronics are destination goods and typically involve a dedicated shopping trip, while gasoline and groceries are more a convenience good and often part of other shopping trips or in conjunction with commuting.

We define the benefits of trip-chaining as a consumption externality and define it as $ConsumptionExt_{j,k} = \overline{Freq}_{j,k} * (1 - |Freq_j - Freq_k|)$ where $\overline{Freq}_{j,k}$ equals the average of $Freq_j$ and $Freq_k$. For industry pairs with high frequency shoppers and similar frequencies of purchase this metric approaches one and is bounded on the lower end by zero. This variable is set equal to zero for industry pairs where are least one industry does not sell directly to consumers or has a frequency of purchase equal to zero since they are not captured in the National Survey of Consumer Expenditures. The industry pair with the highest value of $ConsumptionExt_{j,k}$ is 4471 Grocery Stores with 4451 Gas Stations and 24% of our industries sell directly to consumers and represent expenditures in the National Survey of Consumer

[32]See the Appendix for details on this matching procedure.

Expenditures.

5 Empirical Results

We formally examine the role of each determinant of industrial agglomeration using our new measure of colocalization as a dependent variable. Specifically, we estimate Equation 4 based on the variables described above.[33]

$$\begin{aligned} CoLoc_{j,k} = \beta_0 &+ \beta_1 NATrans_{j,k} + \beta_2 NACons_{j,k} + \\ &+ \beta_3 NAHerf_{j,k} + \beta_4 CoLocIO_{j,k} + \beta_5 ConsumptionExt_{j,k} + \\ &+ \beta_6 KnowSpill_{j,k} + \beta_7 LaborSim_{j,k} + \epsilon_{j,k} \end{aligned} \tag{4}$$

A concern in estimating the relationship between Marshallan and consumption externalities and our index for colocalization is that clustering may change industrial relationships between industries. For example, the use of insurance as a input in the NAICS 6221 - General Medical & Surgical Hospitals industry may reflect the presence of NAICS 5242 - Agencies, Brokerages, & Other Insurance Related Activities in close proximity. Hospitals may substitute to more legal representation and consume lower levels of insurance if lawyers happen to locate proximate to hospitals due to other factors. The simultaneous relationship between colocalization and their determinants is discussed in Ellison et al. (2010)'s analysis of U.S. manufacturing industries.

The scale of analysis adopted in our research limits concerns regarding simultaneity between industry characteristics and colocalization for two main reasons. First, the Denver-Boulder-Greeley CMSA contains only 1% of all establishments in the U.S. and thus our pattern of colocalization has minimal contribution to national industrial characteristics. In fact, our use of national industry characteristics to proxy for local industry characteristics is similar in spirit to Ellison et al. (2010), who adopt United Kingdom industry attributes as instruments for US industry attributes in constructing variables to capture Marshallan externalities. Second, our measure of colocalization incorporates different counterfactuals than national studies, which limits any relationship to national patterns of industrial concentration and thus any simultaneity with industry attributes. Additionally, the similarity between OLS and IV estimation results for the determinants of agglomeration in Ellison

[33] Appendix Table E.1 summarizes regression variables as well as provides descriptive statistics.

et al. (2010) further alleviates concern regarding simultaneity.

Since our measure of natural advantage represents place specific factors for the Denver urban area, one may be concerned that transportation infrastructure is built to serve existing industry or that consumers locate proximate to retail or consumer service industry clusters. To address these concerns, we construct our measure of access to transportation infrastructure using only Interstate highways as well as rail lines. The location of this infrastructure is plausibly exogenous since highways were established as part of the Interstate highway system and both highways and rail were built prior to the development of Denver as a major urban area. Additionally, large scale residential subdivisions are limited to non-commercial and large tracts of land that often begin development prior to rather than after commercial development. Since we cannot eliminate all elements of simultaneity due to natural advantage, we estimate a number of models with and without the inclusion of different measures of natural advantage and show almost identical results. This result highlights that natural advantages variables are largely orthogonal to other determinants of agglomeration in an urban area.

Since different assumptions regarding the use of industry versus sector or even the overall population as our control for the general concentration of industry may influence estimates of our determinants of agglomeration, we incorporate a series of fixed effects models. In all regressions, we include major sector j by major sector k fixed effects where we define major sector as manufacturing, business services and non-business services. These sector fixed effects allow identification to be obtained from within sector j by sector k variation and thus control for issues related to comparing industries across institutional constraints such as land-use regulation. We further control for elements of place that may be idiosyncratic to individual industries by implementing a series of NAICS 4-digit fixed effects for industry j as well as industry k in a number of regression models. These fixed effects control for the fact that certain industries may exhibit more or less colocalization with any industry due to their spatial distribution within our study area.

6 Results

We present our initial results for estimating the determinants of agglomeration in Table 5. All models incorporate $Coloc_{j,k}$ as the dependent variable and the first two models exclude industry fixed effects while remaining models include fixed effects uniquely for industry j and k. All variables are transformed so that coefficients indicate marginal effects in terms of

standard deviations of independent and dependent variables. The first two columns provide some modest impacts with approximately a 0.09 standard deviation increase in colocalization for each standard deviation increase in access to transportation infrastructure or consumers. Most of the traditional Marshallan determinants of agglomeration have a weak relationship with colocalization. Input-Output linkages have a small and weakly significant positive impact on colocalization, but labor similarity generates a positive impact of 0.15 standard deviations.

We implement a series of four digit NAICS fixed effects for industry j and industry k in columns 3 through 5 to control for the fact that establishments may choose locations based on idiosyncratic factors unique to a given industry. Across these models, we find our measures of natural advantage to be largely orthogonal to our other determinants of agglomeration. The exclusion of our three natural advantage variables as well as the exclusion of a subset of our natural advantage variables provide negligible changes in the remaining coefficients. The main results we find from these industry fixed effects models is that transportation and consumers generate coefficient estimates of 0.10 and 0.11 with the three Marshallan and consumption externalities all positively and significantly explaining colocalization. For model 3, knowledge sharing has the largest coefficient estimate at 0.11 with consumption externalities generating a coefficient estimate of 0.04. Jointly, a one standard deviation increase in all the variables would generate a 0.47 standard deviation increase in colocalization, which represents an increase in $Coloc_{j,k}$ from its mean value of 0.43 to 0.59.

We test some additional models in Table 6. We extend results to only estimate industry ordered pairs outside the same two digit sector in column 2. The exclusion of same sector industry pairs test two concerns with our results. First, excluding same sector pairs alleviates any concern that the use of two or three digit aggregate data assigned to four digit industries for input-output linkages or occupation classifications is influencing our main results. Second, the finding of consistent results even across industry sectors highlights that our results are not just a function of four digit industry pairs in the same sector doing fundamentally similar economic activity. Results for this model highlight similar results across all of our variables except knowledge sharing and labor similarity. We find point estimates for knowledge sharing to increase to 0.15 while point estimates for labor similarity become insignificant. These results are somewhat encouraging given that we expect knowledge spillovers to be a stronger factor at the urban area scale while labor market benefits would be less pronounced given the ability of workers to commute across our study area. Model 3 allows industry pairs with more establishments to have greater weight than those pair with few establishments. These

results, which are weighted by the number of establishments in industry j plus industry k, generate similar results to column 2. Columns 4 through 6 restrict analysis to only service industry pairs. Results are slightly weaker across most variables with the main difference being that service industries have no impact from knowledge sharing, but a significant and relatively large impact from labor similarity. The lack of impacts from knowledge sharing and larger impacts from labor similarity may be a result of the large presence of a number of low-skilled service industries in our dataset.

6.1 Sensitivity of Results to Colocalization Indices

Given the prominence of the Duranton-Overman (D-O) and Ellison-Glaeser (E-G) indices in the literature, we test the sensitivity of results to the use of different indices for colocalization. Table 7 provides results for $Coloc_{j,k}$ in comparison to the E-G coagglomeration index at CBG geographies and the D-O colocalization index using a distance criteria of 10km. The choice of CBG and 10km is done to maintain a spatial scale similar to the average CBG size in our dataset. Comparisons across indices do highlight consistent findings that access to transportation infrastructure, input-output linkages, consumption externalities and knowledge spillovers positively influence colocalization. In the case of the E-G index, we find large effects for both access to transportation infrastructure as well as consumers.[34] The fact that access to consumers is three times as large is likely due to the fact that we measure access to consumers based on CBG data. Since $Coloc_{j,k}$ and D-O both use point data, their measure of colocalization is not defined by a specific geography which likely weakens the explanatory power of natural advantage variables based on aggregate data. We find stronger effects for $Coloc_{j,k}$ in terms of consumption externalities, knowledge sharing and labor similarity and stronger effects for E-G in terms of input-output linkages.[35] In general, D-O generates weaker effects than both $Coloc_{j,k}$ and E-G across all determinants of agglomeration. Results across indices indicate that natural advantage, Marshallan factors and consumption externalities all positively explain agglomeration at the urban area scale.

Table 8 attempts to isolate some of the factors that generate differences across indices. In this table, we generate different spatially defined versions of the $Coloc_{j,k}$, E-G and D-O

[34]One thing to note is that for all three indices, we derive our measures of natural advantage based on CBG data and then estimate a measures of natural advantage based on the dependent variable specific colocalization index.

[35]Even though our measure of input-output linkages is directional, we find minimal differences across indices for estimated coefficients on Input-Output. Furthermore, the use of a unidirectional version of Input-Output which adopted the maxim value of $Coloc_{j,k}$ or $Coloc_{k,j}$ provided similar results as our directional variable.

indices from Table 7. Our new version of $Coloc_{j,k}$ is based on assigning point data to its corresponding CBG centroid and implementing our algorithm for constructing $Coloc_{j,k}$. This modification provide a means to test the effect of data aggregation on our results. Comparing columns 1 and 2 do provide some noticeable differences with stronger NA variable coefficients for $Coloc_{j,k}$ based on CBG aggregate data. This result is consistent with the fact that we measure most of our NA variables at the CBG level. Results for input-output, knowledge sharing and labor similarity are stronger and consumption externalities is weaker using the original disaggregate $Coloc_{j,k}$. This lends support to the fact that data aggregation does have a non-trivial impact on our results. We can test for the role of MAUP by comparing E-G using CBGs and zip codes. Our study area contains 234 CBGs and 213 zip codes that contain at least one business with the major difference being that CBG boundaries are typically based on residential population and zip code boundaries are drawn based on businesses. The choice of geographical units generates noticeably stronger results for zip codes with every variable being positive and significant for E-G at the zip code level. The fact that coefficients more than double for a number of variables indicates that beyond simple data aggregation, how geographic units are defined is a legitimate concern in estimating determinants of agglomeration within an urban area.

We test the role of distance criteria by comparing D-O using a median distance based criteria (30km) and our original 10km distance criteria. In general, we find some variation across these two models with weaker effects for natural advantage and consumption externalities in D-O at 30km. The results in Table 8 support the importance of aggregation and geographical units in measuring industrial agglomeration. Across these six models, a one standard deviation increase in all seven variables generates between 0.12 and 1.05 standard deviations increase in $Coloc_{j,k}$.

Across all indices and a variety of models, which are reported in Appendix Tables E.2, E.3, E.4, E.5, E.6, we find a number of consistent determinants of agglomeration. First, access to transportation infrastructure typically generates the largest effect due to natural advantage and natural advantage based on industrial diversity (NA herf) does not explain agglomeration. Input-output linkages positively influence indices across geographical scales. The magnitude of this effect ranges from 0.01 to 0.13 and indicates that input-output linkages do matter at the urban area scale, but to a lesser degree than national estimates by Ellison et al. (2010) of between 0.11 and 0.23 standard deviations. Knowledge spillovers consistently have a positive relationship with colocalization and generate some of the largest coefficients for non-natural advantage determinants of agglomeration. Coefficients range

from 0.05 to 0.18 and highlight the importance of knowledge spillovers within an urban area. These coefficient estimates are larger than Ellison et al. (2010) who finds point estimates of between 0.03 and 0.11 standard deviations. Our new measure of consumption externalities generates positive coefficients of between 0.01 and 0.06 standard deviations and highlights that consumer shopping behavior does help explain agglomeration within an urban area.

7 Conclusions

A number of scholars have contributed to our understanding of the determinants of industrial concentration using manufacturing industries and national datasets and most find a role for both natural advantage as well as Marshallan factors. Our results extend this literature by examining agglomeration within a single urban area and across a range of manufacturing and service industries. Specifically, our results agree with Rosenthal and Strange (2001), who find that knowledge spillovers as well as labor factors impact agglomeration at small geographies. Similar to Greenstone et al. (2010), we find support for labor pooling/sharing and knowledge spillovers, but we also find evidence of input-output relationships. Our methodology for testing the determinants of agglomeration closely match Ellison et al. (2010) and support a similar positive role of knowledge spillovers, labor pooling/sharing and input sharing. We differ from Ellison et al. (2010) in that our strongest determinant is due to natural advantage and knowledge sharing instead of input sharing and labor market pooling. This difference is consistent with our scale of analysis which focuses on smaller geographic distances and a single labor market. Our results are informative to policies that influence industrial location such as land-use regulation, industrial park subsidies and transportation infrastructure, which are often implemented at the sub-metropolitan scale.

The sensitivity of these results to specific indices as well as the distance criteria and geographical unit used to measure colocalization highlights the importance of improving our measurement of industrial agglomeration. Our new index of colocalization has a number of desirable properties in terms of addressing the MAUP, reporting statistical significance and controlling for general industry concentration. This new index provides a different approach to measuring across industry spatial relationships which complement existing indices and offers an additional tool for applied agglomeration research.

References

Anas, A.: 2007, A unified theory of consumption, travel and trip-chaining, *Journal of Urban Economics* **62**, 162–186.

Anderson, S. P. and Engers, M.: 1994, Spatial competition with price-taking firms, *Economica* **61**(242), 125–136.

Arzaghi, M. and Henderson, J. V.: 2008, Networking off Madison Avenue, *Review of Economic Studies* **75**(4), 1011–1038.

Audretsch, D. B. and Feldman, M. P.: 1996, Innovation clusters and the industry life cycle, *Review of Industrial Organization* **11**(2).

Barlet, M., Briant, A. and Crusson, L.: 2012, Location patterns of services in france: A distance-based approach.

Bayer, P., Ross, S. L. and Topa, G.: 2008, Place of work and place of residence: Informal hiring networks and labor market outcomes, *Journal of Political Economy* **116**(6), 1150–1196.

Billings, S. B. and Johnson, E. B.: 2012, A nonparametric test for industrial concentration, *Journal of Urban Economics* **71**(3), 312–331.

Billings, S. B. and Johnson, E. B.: 2013, Localization in polycentric urban areas, *Working Paper* .

Cassey, A. S. and Smith, B. O.: 2012, Simulating confidence for the ellison-glaeser index. Washington State University Working Paper 2012-8.

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

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

Eaton, B. C. and Lipsey, R. G.: 1979, Comparison shopping and the clustering of homogeneous firms, *Journal of Regional Science* **19**(4), 421–435.

Ellison, G. D., Glaeser, E. L. and Kerr, W. R.: 2010, What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns, *The American Economic Review* **100**(3), 1195–1213.

Ellison, G. and Glaeser, E.: 1999, The determinants of geographic concentration, *American Economic Review Papers and Proceedings* **89**(2), 311–316.

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

Fu, S.: 2007, Smart caf cities: Testing human capital externalities in the boston metropolitan area, *Journal of Urban Economics* **61**(1), 86–111.

Glaeser, Edward (Ed.): 2010, *Agglomeration Economics*, University of Chicago Press, NBER.

Gould, E. D., Pashigian, B. P. and Prendergast, C.: 2002, Contracts, externalities and incentives in shopping malls. SO: C.E.P.R. Discussion Papers, CEPR Discussion Papers: 3598, 2002.

Greenstone, M., Hornbeck, R. and Moretti, E.: 2010, Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings, *The Journal of Political Economy* **118**(3), 536–598.

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

Ingene, C. A. and Ghosh, A.: 1990, Consumer and producer behavior in a multipurpose shopping environment, *Geographical Analysis* **22**(1), 71–93.

Jacobs, J.: 1969, *The Economy of Cities*, Vintage.

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

Jofre-Monseny, J., Marin-Lopez, R. and Viladecans-Marsal, E.: 2011, The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms, *Journal of Urban Economics* **70**(1), 61–74.

Kantorovich, L.: 1940, On one effective method of solving certain classes of extremal problems, *Dokl. Akad. Nauk USSR* **28**, 212–215.

Kantorovich, L. and Rubinstein, G.: 1958, On the space of completely additive functions, *Vestnik Leningrad Univ., Ser. Mat. Mekh. i Astron* **13**(7), 52–59.

Kerr, W. R. and Kominers, S. D.: 2010, Agglomerative forces and cluster shapes, *NBER Working Paper 16639* .

Kolko, J.: 2010, *Agglomeration Economics*, University of Chicago Press, chapter Urbanization, Agglomeration, and Co-Agglomeration of Service Industries, pp. 151–180.

Konishi, H.: 2005, Concentration of competing retail stores, *Journal of Urban Economics* **58**, 488–512.

Krugman, P.: 1999, *The Spatial Economy: Cities, Regions, and International Trade*, MIT Press.

Leslie, T. F. and Kronenfeld, B. J.: 2011, The colocation quotient: A new measure of spatial association between categorical subsets of points, *Geographical Analysis* **43**, 306–326.

Marshall, A.: 1920, *Principles of Economies*, London: MacMillan.

Martin, P., Mayer, T. and Mayneris, F.: 2011, Spatial concentration and plant-level productivity, *Journal of Urban Economics* **69**(1), 182–195.

McCann, B. T. and Folta, T. B.: 2009, Demand- and Supply-Side Agglomerations: Distinguishing between Fundamentally Different Manifestations of Geographic Concentration, *Journal of Management Studies* **46**(3), 362–392.

Mulligan, G. F. and Fik, T. J.: 1994, Price and location conjectures in medium- and long-run spatial competition models, *Journal of Regional Science* **34**(2), 179–198.

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

Rosenthal, S. S. and Strange, W. C.: 2003, *Evidence on the Nature and Sources of Agglomeration Economies*, V. Henderson and J.F. Thisse (Ed.) Handbook of Regional and Urban Economics IV, Elsevier, pp. 2120–2149.

Rosenthal, S. S. and Strange, W. C.: 2008, The attenuation of human capital spillovers, *Journal of Urban Economics* **64**(2), 373–389.

Saxenian, A.: 1996, *Regional Advantage: Culture and Competition in Silicon Valley*, Harvard University Press. Cambridge, MA.

Sheppard, E., Haining, R. P. and Plummer, P.: 1992, Spatial pricing in interdependent markets, *Journal of Regional Science* **32**(1), 55–75.

Silverman, B.: 1986, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Thill, J.-C. and Thomas, I.: 1987, Toward conceptualizing trip-chaining behavior: A review, *Geographical Analysis* **19**(1), 1–17.

Wasserstein, L.: 1969, Markov processes over denumerable products of spaces describing large systems of automata, *Probl. Inform. Transmission* **5**, 47–52.

Table 1: Colocalization within Major Industry Sectors

	Manufacturing$_k$	Business Services$_k$	Non-Business Services$_k$
Manufacturing$_j$	18.3%	12.2%	11.4%
Business Services$_j$	6.5%	9.5%	5.5%
Non-Business Services$_j$	2.2%	4.4%	6.4%

We classify NAICS 3111 through NAICS 3399 as Manufacturing; NAICS 4231 through NAICS 4251 as well as NAICS 4811 through NAICS 6244 as Business Services; and NAICS 4411 through NAICS 4543 as well as NAICS 7111 through NAICS 8139 as Non-Business Services. Each cell contains the portion of four digit industries that contain $Coloc(j,k)$ greater than or equal to 0.95 for row and heading sectors.

Table 2: Colocalization by Industry Sector

Two Digit Sector (Industry j)		Portion of 4 digit industry ordered pairs with statistically significant colocalization
31-33	Manufacturing (22%)	13.2%
42	Wholesale Trade (8%)	7.9%
44-45	Retail Trade (13%)	3.9%
48-49	Transportation and Warehousing (6%)	14.3%
51	Information (6%)	8.2%
52	Finance and Insurance (4%)	8.0%
53	Real Estate and Rental and Leasing (3%)	6.7%
54	Professional Scientific and Technical Services (4%)	4.9%
55-56	Administrative, Management and Remediation Services (5%)	7.2%
61	Educational Services (3%)	5.2%
62	Health Care and Social Assistance (7%)	5.7%
71	Arts Entertainment and Recreation (3%)	5.4%
72	Accommodation and Food Services (2%)	3.3%
81	Other Services (6%)	5.5%
	Total (100%)	8.1%

Statistically significant industries indicates the portion of industries in a given two digit sector that contain Coloc(j,k) greater than or equal to 0.95. Parenthesis indicate the portion of all industry ordered pairs in a given sector.

Table 3: Correlation between Colocalization Indices - All Industries

		Duranton & Overman (ψ)		Ellison & Glaeser (γ)	
	$Coloc_{j,k}$	*10 km*	*30 km*	*CBG*	*Zip*
$Coloc_{j,k}$	1.00				
ψ - *10 km*	0.194	1.00			
ψ - *30 km*	0.085	0.850	1.00		
γ - *CBG*	0.235	0.219	0.122	1.00	
γ - *Zip*	0.305	0.092	0.032	0.552	1.00

Table 4: Correlation between Natural Advantage Indices

	$Coloc_{j,k}$	$NATrans_{j,k}$	$NAPop_{j,k}$	$NAHerf_{j,k}$
$Coloc_{j,k}$	1.00			
$NATrans_{j,k}$	0.134	1.00		
$NAPop_{j,k}$	0.124	0.244	1.00	
$NAHerf_{j,k}$	0.088	0.236	0.262	1.00

Table 5: Base OLS Models - Coloc(j,k)

	Coloc(j,k)				
	Base Estimation	Exclude Natural Advantage	Base Estimation	Exclude Natural Advantage	Only Trans Natural Advantage
	(1)	(2)	(3)	(4)	(5)
NA trans	0.093***		0.096***		0.103***
	(0.009)		(0.012)		(0.012)
NA pop	0.090***		0.111***		
	(0.010)		(0.016)		
NA herf	0.031***		-0.010		
	(0.007)		(0.008)		
Input-Output	0.025*	0.027*	0.065***	0.069***	0.067***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Consumption Ext.	-0.000	-0.005	0.038***	0.037***	0.038***
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Knowledge Spillovers	-0.001	-0.021	0.106***	0.133***	0.126***
	(0.021)	(0.021)	(0.036)	(0.038)	(0.037)
Labor Similarity	0.150***	0.182***	0.052***	0.047**	0.045**
	(0.018)	(0.018)	(0.020)	(0.021)	(0.020)
Fixed Effects Industry j and k			X	X	X
R-squared	0.077	0.055	0.344	0.332	0.338
Obs.	40,200	40,200	40,200	40,200	40,200

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients. All standard errors are clustered by industry j. Dep var = Coloc(j,k). All models include major sector j by major sector k fixed effects, where major sector dummies are defined to be manufacturing, business services or non-business services industries. NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table 6: Regression Models - Coloc(j,k)

	$\frac{Coloc(j,k)}{}$					
	Base	Exclude pairs in same NAICS2	Weight by pairwise Number of Establishments	Only Service	Only Service no NA	Only Service Trans NA
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	0.096***	0.094***	0.096***	0.089***		0.097***
	(0.012)	(0.012)	(0.014)	(0.013)		(0.013)
NA pop	0.111***	0.104***	0.118***	0.096***		
	(0.016)	(0.016)	(0.017)	(0.016)		
NA herf	-0.010	-0.012	-0.009	-0.006		
	(0.008)	(0.008)	(0.009)	(0.010)		
Input-Output	0.065***	0.072***	0.071***	0.051***	0.055***	0.054***
	(0.014)	(0.015)	(0.017)	(0.015)	(0.016)	(0.015)
Consumption Ext.	0.038***	0.039***	0.031***	0.030***	0.030***	0.031***
	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Knowledge Spillovers	0.106***	0.167***	0.166***	-0.028	-0.012	-0.014
	(0.036)	(0.037)	(0.054)	(0.039)	(0.042)	(0.041)
Labor Similarity	0.052***	0.015	0.039	0.176***	0.176***	0.171***
	(0.020)	(0.020)	(0.025)	(0.026)	(0.027)	(0.026)
R-squared	0.344	0.349	0.328	0.346	0.335	0.340
Obs.	40,200	37,808	40,200	24,806	24,806	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.
All standard errors are clustered by industry j. Dep var = Coloc(j,k). All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table 7: Does How We Measure Colocalization Matter?

	Coloc(j,k)	Coloc(j,k)	E-G CBG	E-G CBG	D-O 10km	D-O 10km
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans [DV specific]	0.096***		0.166***		0.042***	
	(0.012)		(0.015)		(0.008)	
NA pop [DV specific]	0.111***		0.392***		-0.002	
	(0.016)		(0.025)		(0.007)	
NA herf [DV specific]	-0.010		0.015		-0.002	
	(0.008)		(0.012)		(0.010)	
Input-Output	0.065***	0.069***	0.080***	0.118***	0.037***	0.038***
	(0.014)	(0.014)	(0.014)	(0.015)	(0.010)	(0.010)
Consumption Ext.	0.038***	0.037***	0.018***	0.012*	0.013**	0.015**
	(0.008)	(0.008)	(0.006)	(0.007)	(0.006)	(0.006)
Knowledge Spillovers	0.106***	0.133***	0.012	0.139***	0.074**	0.074**
	(0.036)	(0.038)	(0.029)	(0.042)	(0.029)	(0.030)
Labor Similarity	0.052***	0.047**	0.027	0.007	0.011	0.014
	(0.020)	(0.021)	(0.020)	(0.025)	(0.012)	(0.013)
R-squared	0.344	0.332	0.319	0.129	0.519	0.519
Obs.	40,200	40,200	40,200	40,200	40,200	40,200

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients. All standard errors are clustered by industry j. Dep var = column heading. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects. NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table 8: Comparing Measures of Colocalization

	Coloc(j,k)	Coloc(j,k) CBG	E-G CBG	E-G Zip	D-O 10km	D-O 30km
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans [DV specific]	0.096***	0.281***	0.166***	0.391***	0.042***	0.022*
	(0.012)	(0.018)	(0.015)	(0.065)	(0.008)	(0.011)
NA pop [DV specific]	0.111***	0.364***	0.392***	0.149***	-0.002	-0.002
	(0.016)	(0.020)	(0.025)	(0.018)	(0.007)	(0.011)
NA herf [DV specific]	-0.010	-0.006	0.015	0.044***	-0.002	0.003
	(0.008)	(0.008)	(0.012)	(0.013)	(0.010)	(0.009)
Input-Output	0.065***	0.012	0.080***	0.115***	0.037***	0.017*
	(0.014)	(0.008)	(0.014)	(0.016)	(0.010)	(0.010)
Consumption Ext.	0.038***	0.052***	0.018***	0.036***	0.013**	0.024***
	(0.008)	(0.011)	(0.006)	(0.009)	(0.006)	(0.007)
Knowledge Spillovers	0.106***	0.075**	0.012	0.113**	0.074**	0.055**
	(0.036)	(0.031)	(0.029)	(0.043)	(0.029)	(0.028)
Labor Similarity	0.052***	0.015	0.027	0.084***	0.011	-0.007
	(0.020)	(0.015)	(0.020)	(0.023)	(0.012)	(0.017)
R-squared	0.344	0.404	0.319	0.391	0.519	0.555
Obs.	40,200	40,200	40,200	40,200	40,200	40,200

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients. All standard errors are clustered by industry j. Dep var = column heading. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects. NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Figure 1: Wasserstein Distance

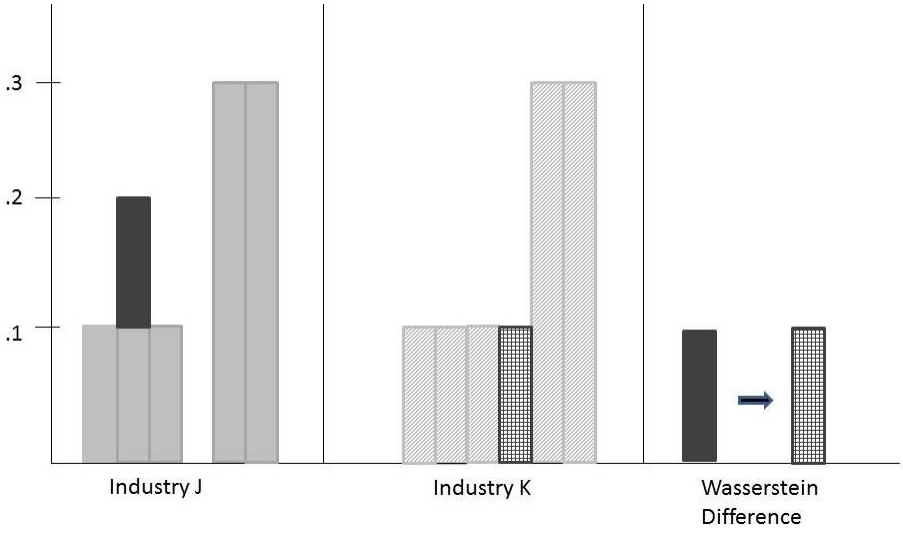


Figure 2: Colocalization Index

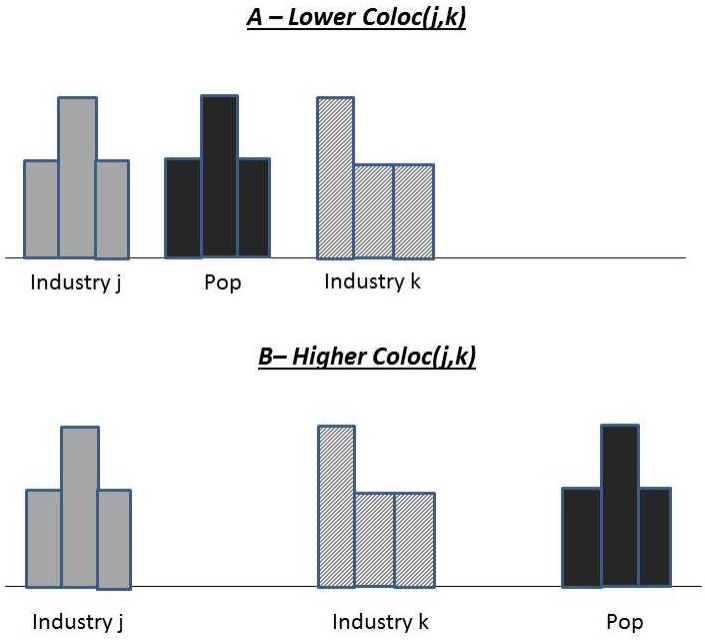


Figure 3: Distribution of Coloc(j,k)

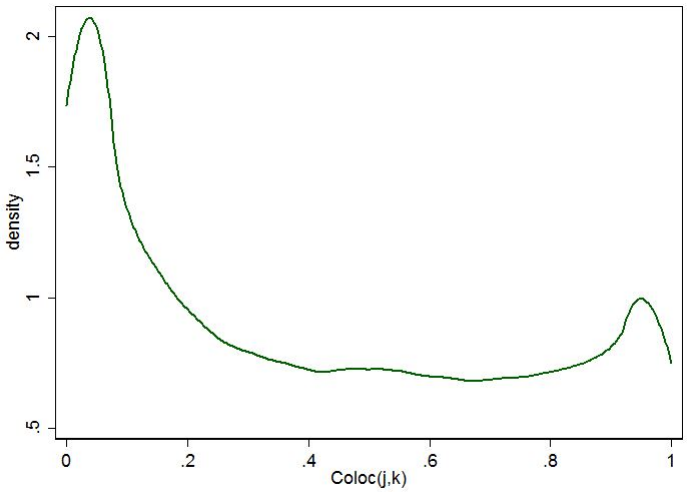
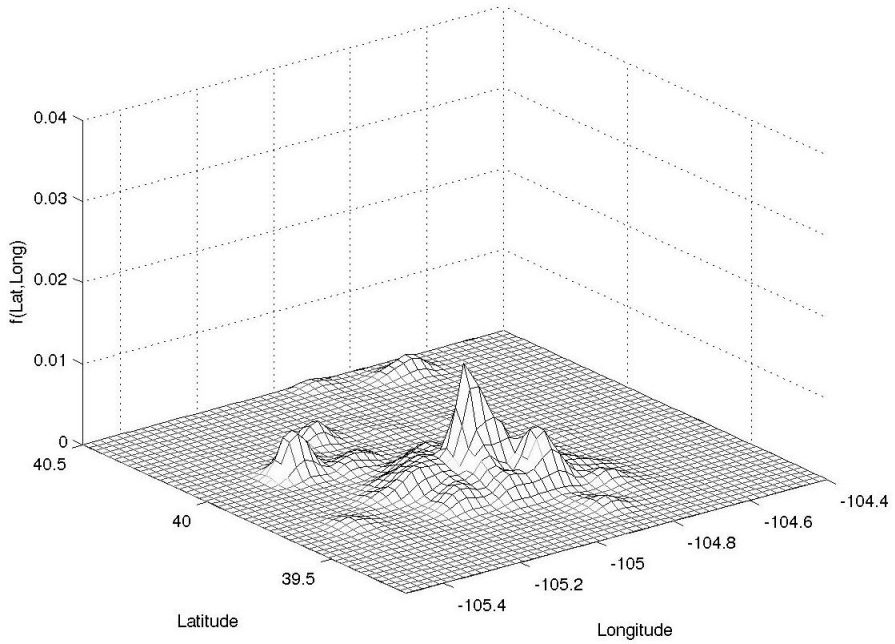
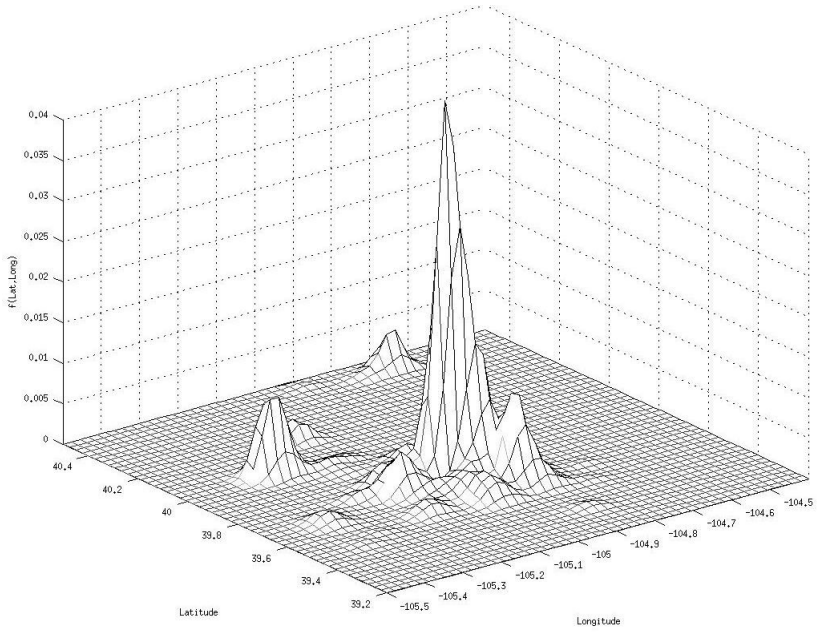


Figure 4: Industrial Concentration



(a) All Industries



(b) NAICS 5411 Legal Services

Appendix

A Ellison-Glaeser Coagglomeration Index

We provide a number of descriptive and regression based results using the Ellison-Glaeser index for coagglomeration. This index is less commonly used then the more prominent agglomeration index of Ellison and Glaeser (1999) (E-G). More recently, Ellison et al. (2010) derive a theoretical model consistent with coagglomeration and we leave discussion of the theoretical model to this literature. In our application, we implement the original E-G coagglomeration index with a few modifications outlined here. We calculate the E-G coagglomeration index as

$$\gamma_{j,k} = \frac{\sum_{m=1}^{M}(s_{mj} - x_m)(s_{mk} - x_m)}{1 - \sum_{m=1}^{M} {x_m}^2} \tag{A.1}$$

where m indexes geographic regions. $s_{1j}, s_{2j}, ...s_{mj}$ are the share of industry j in each of these regions; $s_{1k}, s_{2k}, ...s_{mk}$ are the share of industry k in each of these regions and $x_1, x_2.....x_M$ indicate the size of region m. We measure size as the number of establishments in any industry in region m divided by the total number of establishments in our dataset. We assign m to be 213 five digit zip codes or 234 Census 2000 block groups. In essence, $\gamma_{j,k}$ provides a measure of the spatial correlation between two industries across all geographies M.

The simplicity of computing the E-G index has made it one of the most commonly used indices of spatial relationships in the industrial agglomeration literature with its main criticism being its inability to account for neighboring or adjacent industry concentration thus making it sensitive to MAUP. We modify the original E-G measure in two ways. First, we incorporate different spatial units than typically adopted for this index. Namely, we use small spatial units including zip code areas based on 2000 postal codes and Census tracts based on Census 2000 TIGER files. Second, our measure of the E-G index is based on establishment counts and not employment. We present basic descriptive statistics in Table E.1 and find a mean E-G coagglomration index is 0.0015 for zip codes and 0.0005 for CBGs.

B Duranton-Overman Colocalization Index

The Duranton and Overman (2005) (D-O) colocalization index begins by calculating the Euclidean distance between each pairwise permutation of establishments in industry j and industry k. These $\frac{n_j*(n_k-1)}{2}$ unique pairwise distances, based on n_j establishments in industry j and n_k establishments in industry k, represent the distribution of pairwise distances for a given industry pair. The main idea behind this measure is that industries with greater density at smaller pairwise distances will represent industry pairs with a greater degree of colocalization. In order to generate a smoothed distribution of pairwise distances, one implements a univariate kernel density estimator ($\hat{K}$) across all pairwise distances. This kernel estimator may be defined for areas where the pairwise distance is less than zero, so data reflection is done following the Silverman (1986) technique. Kernel bandwidths are set along one dimension, the pairwise distance, using the Silverman (1986) 'rule of thumb' procedure. The counterfactual of randomly located industries is based on randomly sampling $n_j = N_j$ and $n_k = N_k$ establishments from the population of all establishments.[36] We simulate a full empirical null distribution of kernel smoothed pairwise distances using $10,000$ replications. Finally, local critical values are determined from the empirical null distribution for all possible pairwise distances.

$$\hat{K}_{j,k}(d) = \frac{1}{N_j N_k h} \sum_{r=1}^{N_j} \sum_{s=1}^{N_j} f\left(\frac{d - d_{j,k}}{h}\right) \tag{B.1}$$

To create global confidence bands, we sort kernels at each pairwise distance such that 95% of the kernels lie entirely below the upper confidence band. The envelope of kernel density values that satisfy these criteria provide the global confidence band for each pairwise distance ($\hat{K}_{jk}^{UC}(d)$). We conclude colocalization when an industry specific kernel exceeds the global upper confidence band for any distance less than or equal to our distance threshold. The actual index value is based on the area between a industry's kernel density and the upper global confidence band for pairwise distances between zero and the distance threshold (C) given by Equation B.2. We incorporate distance thresholds of $10km$ and $30km$ in our results to account for our scale of analysis. The median pairwise distance is 25.6 km, but this

[36] There is some debate over the appropriate counterfactual of randomly located industries with the original Duranton and Overman (2005) colocalization index based on a counterfactual of sampling from the joint distribution of establishments in only industry j and industry k, while the implementation of the Duranton-Overman index by Ellison et al. (2010) involves sampling their counterfactual from the full distribution of all establishments. We adopt this later approach simply to be consistent across how we and Ellison-Glaeser model the counterfactual of general industry concentration.

distance is greater than the distance between the two major commercial centers of Downtown Denver and the Denver Technology Center of around 18km. Therefore, we incorporate smaller distance thresholds to avoid concluding colocalization based on the pairwise distance between establishments in these two commercial centers. For the D-O index, the number of colocalized industries increases and correlation with other indices decreases for larger distance thresholds.

$$\Gamma_{j,k} = \sum_{d=0}^{C} max\left[\hat{K}_{jk}(d) - \hat{K}_{jk}^{UC}(d)\right] \tag{B.2}$$

Given the computational burden of computing $\frac{201*200}{2} = 20,100$ industry pairs, we do implement some simplifying assumptions and computational shortcuts. First, we randomly draw a subset of establishments equal to 200 for any industry or counterfactual with more than 200 establishments. Second, we construct global confidence bands for various industry sizes by interpolating global critical values using a subset of values for N_j and N_k. For example, if industry j contained 72 establishments, we use global critical values estimated from the interpolation of actual global critical values for industries of size 70 and 80. We do some sensitivity test on these assumptions in order to verify that these computational shortcuts produce negligible differences in results. We find a mean D-O colocalization index of 0.006 with 27.6% of industries greater than zero for a 10 km distance threshold and 0.011 with 57.1% of industries containing indices greater than zero for a 30 km distance threshold.

C Input-Output Relationships

Here we provide additional detail on some of our determinants of localization and colocalization. For more details on any of these data sources, data documentation is available from the Bureau of Economic Analysis, Bureau of Labor Statistics or US Census Bureau websites. The 2002 Input-Output (IO) accounts from the Bureau of Economic Analysis provides information on customer and supplier relationships across industries as well as by final users (e.g. consumers, government). We incorporate IO relationships to construct our measure of *ColocIO* as a determinant of colocalization. The IO accounts provide estimates of the value of commodity flows between pairs of industries. Commodity flows are developed and estimated by the Census Bureau and include the full range of both manufacturing and service goods.

We only incorporate the direct requirement table of the 2002 IO accounts that links

industry pairs based on commodity based industry definitions. Matching IO industry definitions to NAICS 4 digit industry classification is provided from the BEA for most industries. In some cases, IO industries were linked to more than one NAICS 4 digit category because the correspondence between IO and NAICS industry classifications was given only for two or three digit NAICS classifications.[37] In these cases, IO data is identical between 4 digit industries within the same aggregated industry correspondence.

D Natural Advantage

In order to implement our measure of natural advantage that is industry specific, yet allows us to identify coefficients for a number of natural advantage variables with industries that are limited in location to only a few CBGs, we estimate a simple binomial regression model for a limited set of variables. This specification is empirical simpler than the non-linear least squares model estimated by Ellison et al. (2010) , but is not theoretically linked to a model of firm location. The use of binomial regression models does overcome a number of estimation problems encountered in trying to implement models akin to Ellison et al. (2010) for our dataset that has a number of smaller industries.

Specifically, we include four variables to predict the role of natural advantage due to transportation infrastructure. We construct a measure of distance to highways based on the Euclidean distance between the centroid of a CBG and Interstates 25 or 70, whichever is closest. We include this distance as well as squared distance. Our other two transportation variables include Euclidean distance to heavy rail tracks (non-commuter) as well as squared distance. These four variables are used in a binomial regression model to predict establishment counts for each industry across 234 CBGs. We implement our colocalization algorithm on these predicted points to construct $NATrans_{j,k}$. We implement a similar procedure for access to consumers and construct two variables: population density of a CBG and aggregate income of a CBG (models using these variables for within 5 miles of a CBG centroid produced identical results). Population density is the number of people per square mile given by the 2000 Census and aggregate income is per capita income times the population given by Census 2000. These two variables are used to predict establishments for each industry and lead to the creation of $NAPop_{j,k}$. Our final measure of natural advantage is due to industrial diversity, which we measure by a herfindahl index of industry concentration. We

[37] The correspondence to two digit aggregate IO categories is limited to Wholesale Trade (42) and Retail Trade (44,45)

construct the herfindahl as $H = \sum_j (\frac{N_{ij}}{N_i})^2$ where j indicates industry and i indicates CBG.[38] We use H to predict establishment locations for each industry and then create $NAHerf_{j,k}$.

E Census Expenditure Survey

In order to construct our measure of trip-chaining shopping behavior, we incorporate data from the Consumer Expenditure (CE) Survey program. The CE program is composed of two surveys, the Quarterly Interview Survey and the Diary Survey, which provide detailed information on the consumption habits as well as detailed expenditures by consumer units (families and single consumers). This survey data is collected by the Bureau of Labor Statistics. Our analysis focuses on a detailed extract of the 2002 Consumer Expenditure Interview Survey, which contains annual mean expenditures for hundreds of consumer expenditures categories as well as reports the percent of respondents who purchased an item from each expenditure category in the previous three months. We focus on the percent of respondents who purchased an item in the last three months and use this reported percentage as our measure of consumption frequency ($Freq(\%)_j$)

In order to assign consumption frequency to four digit industries, we have to match final product consumption goods to industry descriptions. This procedure involved taking the NAICS four digit industry description and comparing it to the description of each consumption good. In some cases this matching is relatively easy. For example, frequency of gasoline purchases can be confidently linked to NAICS 4471 Gasoline Stations. In other cases the linkage is more complicated. For example, linking the percentage of respondents who report eating out at a restaurant to a specific industry required deciding between 7221 Full Service Restaurants and 7222 Limited Service Eating Places. In most of these cases, we chose the industry category which is the better fit with a given product. For the remaining cases, we would assign both industries the same frequency. In a few cases, multiple expenditure categories corresponded to a single industry classification. For example, the National Survey of Consumer Expenditures provides frequencies for over thirty different grocery store products and we use the maximum frequency for an individual purchasing any of these 30 products since the purchase of any of these items would require a visit to a grocery store.

We only obtained non-zero values for $Freq_j$ in those industries that provide direct goods or services to consumers (This excludes manufacturing industries, but includes all service

[38] We did consider a number of other natural advantage variables including distance to the central business district, latitude, longitude as well as distance to airports. These variables provide almost no additional explanatory power when combined with the above natural advantage variables.

industries except NAICS 48-49 Transportation and Warehousing and NAICS 55-56 Administrative Services). For example, 13.4% of consumers purchased furniture in the previous three months. This frequency ($Freq_j$) would be assigned to NAICS 4421 - Furniture Stores. For the highest frequency industry of NAICS 4451 - Grocery Stores, ($Freq_k = 0.988$). In this example, our measure of $ConsumptionExt_{j,k}$ for NAICS 4421 and NAICS 4451 would be $\frac{0.13+0.988}{2} * [1 - |0.13 - 0.988|] = 0.086$. Overall, our measure of consumption externalities would give a value close to one for high frequency industry pairs with identical purchasing frequency and a value of zero for the cases when either industry does not provide direct goods or services to consumers.

Table E.1: Colocalization Variables

Variables	**Description**	**Mean**	**(Std Dev)**	**Min**	**Max**
$Coloc_{j,k}$	Portion of industry j and randomly located industry k (j,$\hat{k}$) pairs that are more spatially dissimilar than industry j and k (j,k).	0.427	(0.337)	0	1
$NATrans_{j,k}$	Measure of $Coloc_{j,k}$ if the location of establishments based solely on distance to transportation infrastructure (e.g. Interstates, rail).	0.678	(0.372)	0	1
$NAPop_{j,k}$	Measure of $Coloc_{j,k}$ if the location of establishments based solely on access to consumers.	0.756	(0.368)	0	1
$NAHerf_{j,k}$	Measure of $Coloc_{j,k}$ if the location of establishments based solely on industrial diversity.	0.854	(0.308)	0	1
$\psi(10km)$	Duranton and Overman index for distance criteria = 10 km	0.006	(0.018)	0	0.25
$\psi(30km)$	Duranton and Overman index for distance criteria = 30km	0.011	(0.022)	0	0.25
γ_{CBG}	Ellison-Glaeser index for census block groups	0.0013	(0.013)	-0.098	0.142
γ_{Zip}	Ellison-Glaeser index for zip codes	0.0008	(0.046)	-0.016	0.048
$CoLocIO_{j,k}$	Portion of industry k's commodity input value attributed to the output of industry j	0.009	(0.026)	0	0.84
$ConsumptionExt_{j,k}$	The similarity in consumption frequency between consumer goods produced by industries j and k	0.011	(0.051)	0	0.88
$LaborSim_{j,k}$	Portion of labor in industry j with the same occupational classification as k	0.567	(0.110)	0.247	1
$KnowSpill_{j,k}$	Mean of Grad(%) for industry j and k times $LaborSim_{j,k}$	0.169	(0.068)	0.022	0.602

Table E.2: Regression Models with Coloc(j,k) based on CBG measure of Establishments

	Coloc(j,k) based on CBG					
	Base	Exclude Natural Advantage	Include Only Nat Adv. Trans	Exclude same NAICS2	Weight by Number of Establishments	Only Service Industries
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	-0.008		0.053***	-0.005	0.031	0.002
	(0.014)		(0.016)	(0.014)	(0.020)	(0.020)
NA pop	0.218***			0.212***	0.269***	0.200***
	(0.023)			(0.024)	(0.033)	(0.027)
NA herf	-0.050***			-0.054***	-0.038***	-0.057***
	(0.009)			(0.009)	(0.011)	(0.011)
Input-Output	0.022**	0.024***	0.025***	0.015	0.013	0.019*
	(0.009)	(0.009)	(0.009)	(0.014)	(0.011)	(0.010)
Consumption Ext.	0.052***	0.051***	0.050***	0.053***	0.069***	0.058***
	(0.014)	(0.013)	(0.013)	(0.014)	(0.015)	(0.015)
Knowledge Spillovers	0.144***	0.161***	0.164***	0.208***	0.252***	0.067
	(0.040)	(0.041)	(0.041)	(0.035)	(0.051)	(0.042)
Labor Similarity	-0.002	-0.001	-0.004	-0.037**	-0.044**	0.056**
	(0.018)	(0.018)	(0.018)	(0.016)	(0.021)	(0.022)
R-squared	0.306	0.288	0.289	0.308	0.291	0.297
Obs.	40,200	40,200	40,200	37,808	40,200	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.

All standard errors are clustered by industry j. Dep var = Coloc(j,k) based on establishments assigned to CBG centroid. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table E.3: Regression Models with E-G Index based on CBG

	E-G Index based on CBG					
	Base	Exclude Natural Advantage	Include Only Nat Adv. Trans	Exclude same NAICS2	Weight by Number of Establishments	Only Service Industries
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	0.166***		0.238***	0.167***	0.192***	0.161***
	(0.015)		(0.024)	(0.016)	(0.019)	(0.020)
NA pop	0.392***			0.389***	0.452***	0.403***
	(0.025)			(0.024)	(0.043)	(0.029)
NA herf	0.015			0.014	0.002	0.053**
	(0.012)			(0.012)	(0.012)	(0.024)
Input-Output	0.080***	0.118***	0.104***	0.090***	0.060***	0.057***
	(0.014)	(0.015)	(0.015)	(0.016)	(0.014)	(0.014)
Consumption Ext.	0.018***	0.012*	0.011	0.023***	0.011**	0.002
	(0.006)	(0.007)	(0.007)	(0.006)	(0.005)	(0.005)
Knowledge Spillovers	0.012	0.139***	0.091**	-0.026	-0.002	-0.123***
	(0.029)	(0.042)	(0.036)	(0.033)	(0.038)	(0.032)
Labor Similarity	0.027	0.007	0.002	0.028	0.022	0.141***
	(0.020)	(0.025)	(0.023)	(0.021)	(0.021)	(0.026)
R-squared	0.319	0.129	0.175	0.303	0.246	0.298
Obs.	40,200	40,200	40,200	37,808	40,200	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.
All standard errors are clustered by industry j. Dep var = EG Index for CBGs. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table E.4: Regression Models with E-G Index based on Zip Codes

	E-G Index based on Zip Codes					
	Base	Exclude Natural Advantage	Include Only Nat Adv. Trans	Exclude same NAICS2	Weight by Number of Establishments	Only Service Industries
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	0.391***		0.401***	0.385***	0.130***	0.187***
	(0.065)		(0.066)	(0.064)	(0.026)	(0.042)
NA pop	0.149***			0.143***	0.125***	0.144***
	(0.018)			(0.019)	(0.012)	(0.017)
NA herf	0.044***			0.044***	0.089***	0.044***
	(0.013)			(0.012)	(0.015)	(0.014)
Input-Output	0.115***	0.129***	0.121***	0.114***	0.098***	0.086***
	(0.016)	(0.017)	(0.016)	(0.020)	(0.017)	(0.013)
Consumption Ext.	0.036***	0.059***	0.050***	0.037***	0.031***	0.029***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.007)	(0.006)
Knowledge Spillovers	0.113**	0.112**	0.116***	0.147***	0.163***	-0.032
	(0.043)	(0.045)	(0.042)	(0.039)	(0.053)	(0.051)
Labor Similarity	0.084***	0.099***	0.079***	0.047**	0.035	0.211***
	(0.023)	(0.025)	(0.023)	(0.023)	(0.027)	(0.032)
R-squared	0.391	0.314	0.381	0.359	0.254	0.195
Obs.	40,200	40,200	40,200	37,808	40,200	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.
All standard errors are clustered by industry j. Dep var = EG index for zip code areas. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per zip code area with point coordinates assigned to the centroid of each zip code area. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table E.5: Regression Models with D-O Index at 10km

	D-O Index based on 10km					
	Base	Exclude Natural Advantage	Include Only Nat Adv. Trans	Exclude same NAICS2	Weight by Number of Establishments	Only Service Industries
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	0.042***		0.041***	0.045***	0.055***	0.054***
	(0.008)		(0.008)	(0.008)	(0.012)	(0.012)
NA pop	-0.002			-0.008	-0.009	0.010
	(0.007)			(0.006)	(0.008)	(0.010)
NA herf	-0.002			-0.006	-0.026***	-0.001
	(0.010)			(0.011)	(0.008)	(0.016)
Input-Output	0.037***	0.038***	0.037***	0.033***	0.030**	0.035***
	(0.010)	(0.010)	(0.010)	(0.011)	(0.012)	(0.011)
Consumption Ext.	0.013**	0.015**	0.013**	0.017**	0.009*	0.014**
	(0.006)	(0.006)	(0.006)	(0.007)	(0.005)	(0.007)
Knowledge Spillovers	0.074**	0.074**	0.074**	0.085***	0.089**	0.080**
	(0.029)	(0.030)	(0.030)	(0.031)	(0.038)	(0.036)
Labor Similarity	0.011	0.014	0.011	0.004	-0.003	0.022
	(0.012)	(0.013)	(0.013)	(0.013)	(0.015)	(0.015)
R-squared	0.519	0.519	0.519	0.516	0.572	0.562
Obs.	40,200	40,200	40,200	37,808	40,200	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.
All standard errors are clustered by industry j. Dep var = D-O Index up to 10km. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.

Table E.6: Regression Models with D-O Index at 30km

	D-O Index based on 30km					
	Base	Exclude Natural Advantage	Include Only Nat Adv. Trans	Exclude same NAICS2	Weight by Number of Establishments	Only Service Industries
	(1)	(2)	(3)	(4)	(5)	(6)
NA trans	0.022*		0.022**	0.025**	0.010	0.011
	(0.011)		(0.009)	(0.011)	(0.013)	(0.010)
NA pop	-0.002			-0.012	0.015	0.018
	(0.011)			(0.010)	(0.012)	(0.013)
NA herf	0.003			-0.004	-0.008	0.010
	(0.009)			(0.009)	(0.008)	(0.010)
Input-Output	0.017*	0.017*	0.017*	0.011	0.010	0.016
	(0.010)	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)
Consumption Ext.	0.024***	0.024***	0.024***	0.023***	0.008	0.018***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
Knowledge Spillovers	0.055**	0.054*	0.055**	0.060**	0.052	0.112***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.036)	(0.041)
Labor Similarity	-0.007	-0.005	-0.007	-0.011	-0.007	-0.030
	(0.017)	(0.017)	(0.017)	(0.016)	(0.019)	(0.026)
R-squared	0.555	0.554	0.555	0.556	0.593	0.590
Obs.	40,200	40,200	40,200	37,808	40,200	24,806

All variables transformed to have one unit standard deviation and can be interpreted as standardized coefficients.
All standard errors are clustered by industry j. Dep var = D-O Index up to 30km. All models include industry j and industry k fixed effects as well as major sector j by major sector k fixed effects . NA (natural advantage) variables are dependent variable specific and represent the portion of Coloc(j,k) predicted uniquely for access to transportation infrastructure, consumers and industrial diversity. All NA variables based on predicted number of establishments per CBG with point coordinates assigned to the centroid of each CBG. The unit of observation is all ordered pairs of 201 four digit NAICS industries.