Papers in Evolutionary Economic Geography

21.05

The Overlooked Insights from Correlation Structures in Economic Geography

Matias Nehuen Iglesias

Utrecht University

Human Geography and Planning

The Overlooked Insights from Correlation Structures in Economic Geography

Matias Nehuen Iglesias[1,*]

[a]*Scuola Superiore Sant'Anna, Univ. di Pisa*
(33 Piazza Martiri della Liberta, Pisa, PI, Italy)

*Corresponding author
Email address: matuteiglesias@gmail.com (Matias Nehuen Iglesias)

Abstract

Measures of cooccurrence computed from cross sectional data are used to rationalize connections among economic activities. In this work we show the grounds for unifying a multiplicity of similarity techniques applied in the literature and we precise the identification of cooccurrence to actual coexistence in space, when one side of the cross section are small administrative areas. All the similarity techniques studied here are akin to a correlation structure computed from spatial intensity, also known as locational correlation. We argue that these correlations offer objective tools to detect spatial patterns. Indeed we show that when applied to data of employment by industry and county in United States (from 2002-7) the communities of networks derived from locational correlations detect spatial patterns long acknowledged in economic geography. By addressing critical open issues on the interpretation of cooccurrence indices, this work offers technical guides for their exploitation in Economic Geography studies.

Keywords: Economic geography, co-location, spatial analysis, areal data, point data, correlation structures, distribution of economic activities.

1. Introduction

The study of a wide range of questions in Economic Geography is based on characterizing the spatial distribution of activities, their employment, facilities, suppliers or customers. These questions can be related to agglomeration externalities, diffusion of knowledge or regional development, to name some examples.

Researchers usually seek to condense the full spatial information related to some economic activity into indices that can express special features of interest. There are measures that aim to capture spatial concentration, for instance those in Duranton and Overman (2005) and M. Porter (2003) (under the name 'locational correlations'). The first ones compute all pairwise distances among establishments of an industry and compare their distribution with expectations from a null model to determine if certain industries have their establishments more frequently located at certain distances. The latter proposes to compute the correlation matrix from cross sectional data of employment by US state suggesting that high correlation across space signals 'locational linkages' between a pair of activities.

In other cases we have so called *cooccurence measures*, as in Hidalgo et al. (2007). They apply a proximity measure on cross sectional data of exports by country to estimate a network of products (product space). This method has inspired a very active strand of literature that studies inferred networks of economic activities, technologies or regions (Boschma et al., 2014; Delgado et al., 2015; Neffke et al., 2011) and has put forward the idea of *relatedness* as a central concept (Hidalgo et al., 2018).

The product space of Hidalgo et al. (2007) appears to be a technique unrelated to the ones mentioned before. In fact, however, the proximity derived in Hidalgo et al. (2007) can be taken as a correlation structure like that in M. Porter (2003).

In this paper we suggest that technical efforts devoted to understanding correlation structures would solidify the foundations of recent research papers in various strands within Economic Geography. We focus on correlation structures computed on cross sections where one of the sides are geographical units. In that particular case pairwise similarities must have spatial interpretations.

In our view, two issues are among the most critical. Firstly, there seems to be no unified criteria in the transformation of raw data, and the computation of similarity measures. Different works adopt slight variations of the same processing steps rendering their results incomparable. In addition, some of the most popular methodological decisions are approximately equivalent to comfortable mathematical tools but depart slightly from them. This complicates the formal study of the indices used, even if possibly not changing the published results significantly.

A second clear open issue that applies to this type of studies has to do with the formal treatment of space. Physical distance plays key roles in almost any phenomena studied in Economic Geography. But (back to the connection between Duranton and Overman (2005) and M. Porter (2003)) when computing locational correlations, how do distances enter the picture? We aim to tackle and overcome this problem and reconcile correlations computed from data of administrative areas to accounts in continuous space.

To address the first issue, exploiting data on number of employees and number of establishments by industry (4 digit North American Classification System, NAICS) and county in the United States (US). We first compare similarities presented by all pairs of industries, testing alternative combinations of raw data processing (no

transformation, log transformation, binarized location quotient (LQ)) and similarity measures (cosine similarity, Pearson correlation, proximity as in Hidalgo et al. (2007), covariance, and dot product of the cross sectional matrix). These are the *discrete* similarity measures, so called because they are computed from areal data. We find that all these transformations and similarity measures lead to partially equivalent rankings of similar - dissimilar industry pairs. [1]

To address the second issue, we compare expressions of overlap in continuous space to these discrete measures. Analytical developments suggest a close relation between cosine similarity measures and coexistence in continuous space. Computational experiments confirm this connection inequivocally and help understanding the implications of certain characteristics of geographical areas. In a nutshell, computing cosine similarity of employment levels in counties is equivalent to superposition in continuous space of exponential decay density around establishments. As long as the decay width is about one third of the typical area size (diameter).

After addressing these open issues with similarity measures, we explore the co-occurrence inferred from data of employment by industry and US county. Because one side of the cross section are small areas, communities detected from correlation structures are associated to a spatial pattern (neighboring activities in the network have a similar distribution across counties). Indeed, correlation structures allow us to classify industries by their spatial distribution, and the classes that we find point clearly to long theorized economic phenomena. More precisely, we distinguish large cities, distribution of population, presence of natural resources (forests, coastal regions, agriculture or minerals/fuels) and activities that predominate in each of them. A last group comprises most manufacturing activities. One can say this technique is a dimensionality reduction, as instead of more than 3000 counties we can describe spatial distribution of industries by means of few patterns. It is interesting to note that this classification, while clearly pointing to concepts studied in Economic Geography, is achieved without any informed intervention from the researcher. The information is encoded in the raw data and thus in the correlation matrices.

Overall, results of this work help to make the case for the use of correlation structures as an objective tool in the study of spatial patterns.

The paper is organized as follows. Section 2 reviews works applying cooccurrence measures. Section 3 describes the data. Section 4 presents a overview of the methods used, clarifying notation and terminology. Section 5 shows the grounds for unifying a variety of discrete coexistence measures. Section 6 shows how discrete similarity measures match a continuous model of space. Section 7 discusses the correlation structures observed in US and we conclude in Section 8.

[1] In the literature the names proximity, co-occurrence or coexistence measures, correlation structures or locational correlations (M. Porter, 2003) refer to similarity measures of this family. Sometimes referring to measures sharing a definition (formula) or differing in their definitions.

2. Related works

2.1. The use of similarity measures

Inner products such as X^TX are basic measures of joint cooccurrence and as such they have been featured often. The elements of this matrix are $(X^TX)_{ii'} = \sum x_{ij}.x_{i'j}$. Antecedents of studies that applied this framework may be found outside Economic Geography. Applications to counts of patents appear at least as early as in Jaffe (1986) where a cosine similarity between vectors of firms patents by technological categories is called proximity and used to weight investments in related firms. Basic joint cooccurrence and cosine similarity is also applied on patent data in Breschi et al. (2003), Engelsman and van Raan (1994). In fact, these and other types of similarity measures (co-authorship, joint thematic classification of published works) have been naturally welcomed in scientometric research (cf van Eck and Waltman (2009) for a review). Much earlier appearance of such similarity methods is likely, although the lack of good quality data and computational availability may have discouraged this type of analysis. Counts of joint occurrences of products in the portfolio are used by Teece et al. (1994) to evaluate the coherence of firms portfolios. Some more recent examples which prompted a revitalization of the approach are in Hausmann and Klinger (2007), Hidalgo et al. (2007), where they call a minimum conditional probability as 'proximity' (ϕ). That is $\phi_{ii'} = \sum x_{ij}.x_{i'j}/max(\Sigma\ x_{ij}, \Sigma\ x_{i'j})$, applied on a transformed matrix of exports by country.

These contributions had strong influence in making clear that a network structure derived from similarity measures offers a quantitative tool to estimate how industry or technology categories relate to each other. Then, it became useful to branches of Economic Geography studying capabilities of labour (Neffke et al., 2011), knowledge diffusion and technological evolution (Balland et al., 2015; Boschma et al., 2014). It helped mapping landscapes of technological (Alstott et al., 2017) or productive capabilities (Hausmann & Neffke, 2016), grouping regions based on what happens inside them, among other applications interesting to other branches of economic geography. [2]

Table 1 shows the similarity methods used in these and other contributions. The content of its columns highlight the specific features that make each work be different to the rest, but they also represent factors that unite these works under a single framework.

Analogous rationales for relating entities appear often in works out of geography, as is natural to expect. And they can be interpreted from the points of view of bipartite networks, correlation structures, dimensionality reduction techniques, and other equivalents.

Empirical data that fits rectangular matrices happens often across scientific fields. In financial analysis of time series, the side is usually made of time intervals and the structure of so called cross-correlations have been widely studied in a rich strand of literature mainly featured in the journal Physica A with a kick starter contribution in Plerou et al. (1999), among others. This strand has thoroughly studied the spectra (i.e. eigenvalue distribution) of correlation matrices from financial time series. It is clear by now that it is useful to express correlation matrices as the sum of a 'modal' matrix, a groups structure matrix and a noise matrix, all obtained directly

[2] Further possibilities for applying similarity analysis with an interesting variety of configurations can be found in Nedelkoska et al. (2018) and Farinha et al. (2019).

Table 1: Non extensive list of works applying similarity analysis. (*) This paper

	Variable (unit)	Transform.	Main cat	Side cat	Proximity measure
Jaffe (1986)	Patents		Firms	Technological fields	Cosine
Teece et al. (1994)	ownership of plants in industries		Firms	Industries	
M. Porter (2003)	Employment (#)		Industries	US states	Pearson corr
Breschi et al. (2003), Engelsman and van Raan (1994)	Patents (#)		Patent Id	Technological fields	$X^T X$, cosine
Zhang and Horvath (2005)	Gene Expression		Gene	Locus	Pearson corr
M. A. Porter et al. (2005)	vote (nay = -1, yea = +1, else = 0)		Roll-call votes	Representatives	$X^T X$, $X X^T$
Hausmann and Klinger (2007), Hidalgo and Hausmann (2009), Hidalgo et al. (2007), Tacchella et al. (2012)	Exports (USD)	LQ >1	Product (HS / SIC)	Country	min cond. Prob. (proximity)
J. Wang and Yang (2009)	mean daily temperature		Chinese cities	Time periods	
Coscia et al. (2013)	joint appearance in online documents ('hits') (#)		Countries — organizations — Issues (keywords)	(idem)	LQ >1 of hits
Boschma et al. (2012)	Exports (USD)	LQ >1	Product (HS / SIC)	Spanish region (NUTS 3)	min cond. Prob. (proximity)
Boschma et al. (2014), Santoalha and Boschma (2020)	Patents (#)	LQ >1	Firms	Technological fields	min cond. Prob. (proximity)
Hausmann and Neffke (2016)	Labor flow (#)	(LQ - 1) / (LQ + 1)	Industry	Industry	
Petralia et al. (2017)	Patents (#)	LQ >1	Country	Technological fields	Cosine
Iglesias (*)	Employment, Firms (#)	No transformation, log, $LQ > 1$	Industries (NAICS)	counties	Pearson corr, cosine, cov, $X^T X$, min cond. Prob. (proximity)

from the eigenvalues and eigenvectors of the empirical correlation matrix. A recent work dealing patiently with the caveats of computing some clustering in a network derived from a correlation structure is MacMahon and Garlaschelli (2015). Results from this strand of literature can be helpful for approaching the community detection in correlation structures.

Another discipline in which this data type is widespread is in genomics. In that context, gene expression data is naturally displayed in a rectangular matrix where columns stand for different genes and rows indicate expression levels under various conditions (Y. R. Wang & Huang, 2014). A squared similarity matrix is usually built. Research in an interdiscipline involving genomics and computational statistics delves further into the details, choices and implications of this type of analysis (eg. Zhang and Horvath (2005)).

If the strands of Economic Geography working with similarity measures placed more importance on the mathematical identity of the indices it wants to use (ie. discouraging continual creation of new independent indices, and keeping track of the implications of each transformation of raw data and how different indices can be formally related), it could benefit largely by borrowing from powerful technical developments arising in these other disciplines. In addition, results within the field would be more easily comparable to each other.

2.2. Focus of this paper: Areas are side categories

The focus of this paper is on the specificities derived from having geographical units as one side of the cross section. In such a setting, cooccurrence techniques must be related to other techniques of spatial analysis. This connection has however not been formally addressed to the best of our knowledge.

In M. Porter (2003), *'locational linkages'* among industrial activities are inferred from Pearson correlations of employment disaggregated by (4 digit) SIC industry categories and US State. A more recent work that makes use of such locational correlations is Diodato et al. (2018).

Another index of industry to industry coagglomeration is proposed in Ellison and Glaeser (1999). It is defined as the covariance of employment shares (normalized by one minus a Herfindahl index). If we work at a single level of disaggregation this last normalization does not play a role. On top of that, in practical cases it will be very close to one. A simplified version of the index would then be taking just the covariance of employment shares. This index is 'similar in spirit' to the Pearson correlations that Porter uses. The shares covariance of Ellison and Glaeser (1999) are unfortunately hard to match analytically with the similarity measures computed on absolute values. This is why they are excluded from the analysis of this paper, even if they would be worth including in follow up studies.

Many papers have countries as side categories, eg Hausmann and Klinger (2007), Hidalgo et al. (2007). Even if these are geographical areas, one should acknowledge that they are relatively few units with disparate sizes of about 4 orders of magnitude between extremes in terms of surface area, population or gross product. Instead, if we take a single country or region, and split it into a large enough number of small areas of about the same size we are closer to bridging point based pictures to small comparable areas arranged in a kind of lattice, to larger regions that contain a bunch of these areas. Indeed, we will first apply our analysis on the contiguous United States of America split into (nearly 3200) counties of about $(40km)^2$ average size. They offer some of the few cases of a large region split into uniform small areas of comparable size (even with a few exceptions), in addition to good quality data, strong and varied economic activities throughout the country and compiled quite harmonically in central agencies. Successful tests on US counties would be a first step before applying the methods on evidence from other parts of the world. This analysis in thus a substantial improvement over the very coarse picture one can get from the 50 states as in M. Porter (2003). Smaller geographical units allow finer resolution of spatial patterns.

The issue of how to interpret a high correlation of spatial distribution is usually not addressed formally. Multiple reasons can lead to such observation. This issue is of course not simple to approach, but it is nevertheless needed before outcomes of studies can be safely interpreted.

Finally, a promising alternative approach to pairwise similarities of industries based on their distribution over areas has been put forward in van Dam et al. (2020), who introduce the use of pointwise mutual information. This index has reasonable foundations and understanding its exact relation to the rest of correlation measures may be a useful exercise. This however would demand a dedicated study that we have to leave for the future.

Much of the information on regional economics have administrative areas as the basic unit of analysis. An issue that has been discussed is the effects of arbitrariness in administrative divisions' size and shape. The fact that firms can be in the border of an area and show no co-location with firms just across the border, while they would co-locate with distant firms within the same district may influence the results. This issue has been acknowledged for long, often as the Modifiable Unit Area Problem (MAUP). See for example Hennerdal and Nielsen (2017), Menon (2009) for review and further discussion.

Even if the arbitrariness of administrative borders is a factor that will unavoidably alter results, if there is only one version of the underlying facts, then continuous and discrete measures of it should not contradict each other. That is, on average two points *close* to each other are likely to lie in the same area, and two points *far* from each other are likely to lie in different areas. Irregularities of areas would introduce a certain distortion but it cannot mess up with this principle completely.

The idea of having *solutions* to the MAUP is for example discussed in Dark and Bram (2007). Some works such as Duranton and Overman (2005), Scholl and Brenner (2016) present it as a reason for choosing point based measures instead of areal measures. Instead, I would like us to see they can all be interpretations of a single observation of a given spatial pattern. This will be developed in Section 6 where we will probe the formal connection between areal and point data, offering a solution to the MAUP in studies of co-location.

3. Data and Methods

We test the methods on the contiguous United States, both due to their intrinsic weight as a major economy where a wide variety of economic activities take place, and because it is known to present multiple known geographies and spatial patterns in its vast territory. The source of information for this study are recent editions of the County Bussiness Patterns (CBP) datasets, produced by the Bureau of Labor and Statistics (BLS). Among other possibilities, the CBP data offers a dissagregation of the variables 'average annual employment', 'number of establishments' and 'total annual wages' into more than 3200 counties and 300 NAICS 4 digit industries.

We leave activities that show a dependency based on administrative decisions out of the analysis. These includes mostly non productive activities registered more or less intensely depending on the conventions adopted within each US State. [3]

4. Review of the formal framework

The arbitrariness in the design of any classification of activities and their interpretation at the stage of data creation, as well as the researchers' use of chains of transformations can altogether heavily influence the outcome of any study. This constitutes the gap between actual phenomena and the data which is finally used (eg for a regression). The methodological stages that make up this gap however fall out of the focus of most papers, as the attention is placed on an answer offered to some question. Unfortunately, these answers may loose force if there are open issues regarding the methods. For this reason we would like to review the steps where many studies depart from each other partially undermining comparability of results.

[3] Namely: 'NAICS 2213 Water, sewage and other systems', 'NAICS 4854 School and employee bus transportation', 'NAICS 4911 Postal service' 'NAICS 6111 Elementary and secondary schools', 'NAICS 6113 Colleges and universities', 'NAICS 6241 Individual and family services', 'NAICS 7132 Gambling industries' 'NAICS 8131 Religious organizations', 'NAICS 8141 Private households', 'NAICS 9211 Executive, legislative and general government', 'NAICS 9221 Justice, public order, and safety activities', 'NAICS 9231 Administration of human resource programs', 'NAICS 9241 Administration of environmental programs', 'NAICS 9261 Administration of economic programs'.

Therefore, in these next sections we will briefly review the formalisms that let us view the methods in multiple papers as variants of a single 'similarity approach' (cf. Table 1), and then review the choices at the stage of data processing, and the particularities that may let datasets from different studies be inherently different from each other.

4.1. The similarity measures

Given a matrix $X_{(n\times p)}$ we may want to know whether its columns or rows have some relations among them. For this question, answers can come from multiple association coefficients such as the matrix product X^TX. There are other measures that can fulfill this role, such as Pearson correlation, cosine similarity and covariance. If we have a pair of columns $X_j = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})^T$ and $X_{j'}$, these similarity measures are defined as follows:

Pearson correlation:

$$Corr(j, j') = r_{jj'} = \frac{\sum_i^n (x_{ij} - \bar{X_j})(x_{ij'} - \bar{X_{j'}})}{||X_j - \bar{X_j}||||X_{j'} - \bar{X_{j'}}||} \tag{1}$$

where j, j' represent a pair (e.g. a pair of industries) $\bar{X_j}$ denotes the mean of column j and the square norm is naturally defined as $||X_j - \bar{X_j}|| = \sqrt{\sum_i^n (x_{ij} - \bar{X_j})^2}$ and the same for column j' in place of j.

Cosine similarity:

$$CosSim(j, j') = r_{ii'} = \frac{\sum_i^n x_{ij}\ x_{ij'}}{||X_j||||X_{j'}||} \tag{2}$$

we can again see that $Corr(X_j, X_{j'}) = CosSim(X_j - \bar{X_j}, X_{j'} - \bar{X_j'})$.

Sample covariance:

$$Cov(j, j') = \frac{1}{n}\sum_i^n (x_{ij} - \bar{X_j})(x_{ij'} - \bar{X_{j'}}) \tag{3}$$

where n is the number of counties. These measures are partially related to each other as can be seen from their formulas. In certain special cases, a X^TX product, covariance matrix, cosine similarity or Pearson correlation becomes identical to some of the other measures.

If the column variables are centered (their mean is zero) the covariance matrix is $Cov(Y) = Y^TY/(n-1)$, with $Y = X - \bar{X}$. If we z-standardize the columns (demean and divide them by the standard deviation) Pearson correlation will match the covariance, i.e. $Corr(Z) = Z^TZ/(n-1)$ with $Z = (X - \bar{X})/std(X)$. If instead we unit scale the columns of X, that is, we scale the columns so that their sum of squares is 1 (their norm is 1) then we can have the cosine similarity matrix. $Cossim(V) = V^TV$ with $V = X/||X||$. If we had centered the matrix before unit scaling, i.e. with a matrix $W = (X - \bar{X})/||X - \bar{X}||$ then we again obtain the Pearson correlation matrix, this time equal to the cosine similarity matrix as is the case for centered matrices. This is $Corr(W) = W^TW = Cos(W)$.

This discussion emphasises that if the matrix fulfills some properties the expressions for covariance, Pearson correlation or cosine similarity can be compacted in an inner (i.e. matrix dot-) product. In general, however, our empirical data (X) would not fulfill those special conditions on their rows or columns. Then, these measures will partially differ from each other. If we are counting populations or total nominal values of output or one directional trade then the X matrix will not be centered. In general, empirical data will not be normalized or standardized even if we could allow this transformation in some cases. We may however not have a strong justification for applying these transformations, so that it is best to not transform the raw data and confirm whether some of the similarity measures coincide or not when applied on our particular empirical case.

The possible sets of categories

Even if mathematically it would not make difference to transpose our rectangular data and exchange the role of rows for that of columns and vice versa, we will adopt the convention to call the columns the *main* categorization, and call the rows the *side* one. This means that the covariance and other similarities will be defined for pairs of the main variables based on the values they take on the side variables.

When dealing with empirical data we may rely on classifications, e.g. for political entities, time periods, industries, occupations of workers, technological categories of patents, traded products or services, research fields and disciplines, etc. These classifications have multiple possible levels of aggregations, often hierarchical but not necessarily. Higher levels of disagreggation can allow detection of more specific phenomena but at the same time increase noisy values from little populated categories, possibly exacerbating distortions from arbitrariness at the step of data collection.

In this work we use counts of formal employment classified by administrative regions (US counties) and industry (NAICS).

Transformations of the observed data

Transformations of the original data are very frequent. They influence the outcomes of any study in a sensible way but often not enough attention is placed on them. The most frequent transformations are *logarithmic transformation*, and the *Location quotient* (LQ) usually followed by a *binarization*. Expressing raw data in logarithmic scale can help arrive at a more natural distribution of the matrix values. For example, nominal monetary values or counts of people are often better expressed after a log transformation that can let matrix entries' values follow a bell shaped distribution afterwards. The so called 'Location Quotient' (often called 'Revealed Comparative Advantage' (RCA) in the context of international trade as in Balassa (1965) or Hidalgo et al. (2007)) involves dividing entries of the rectangular matrix by the partial margins and implies comparing the observed values to those expected if marginal distributions were independent. [4] 'Binarization' (often applied after computing LQ) transforms the original matrix elements into a boolean (0, 1) telling where the variable was higher than a threshold. Depending on the application, it is possible that we want to know just *where* something happens and not *to which extent* it happens, which is what a binarization achieves.

[4] If the raw data is well distributed in logs it is advisable to use the log of the location quotient.

Units of measurement

Depending on the specific application, the observations may refer to numbers of people, nominal value in some currency, number of patents, among multiple other possibilities.

Naturally, when all the data are consistent in the choice of unit of measurement (for example values in USD) mathematical tools can be applied more powerfully. When we mix different kinds of variables into a single rectangular matrix we may have problems at the transformation stage. Eg. if one column has values in [0, 1] and the rest are population numbers in the thousands, an LQ or a row-wise z-score will be 'broken' for the first column. This needs to be contemplated in each particular application.

5. Unifying a whole family of discrete coexistence measures

As we have discussed, similarity measures given by different definitions may match each other in special cases. In the cross sections of employment or number of establishments by county and industry the conditions of centered, normalized data are not fulfilled. Still, it is worth exploring to which extent these discrete similarity measures can still match each other in our setting.

After plotting all values of pairwise similarity according to the multiple discrete similarity measures we detect a clear correspondence only in the following cases:

- $\cos(X) \approx \text{corr}(X)$
- $X^T X \propto \text{cov}(X)$

For the first item ($\cos(X) \approx \text{corr}(X)$), the correspondence is an identity. For the second one ($X^T X \propto \text{cov}(X)$) it is a proportionality. These relations are also observed if the raw data X was transformed to $log(X)$, both for measures of employment by county and industry, and number of establishments by county and industry. These are illustrated in the plots of Figure 1 applied on employment level data. Analogous results are observed for number of establishments data.

If we widen the choices of possible measures of similarities and transformations of the original data (X) we can uncover a whole family of similarity measures that agree on which are the most and least similar pairs of activities. In that sense, we can argue they are all imperfect measures of a single property of industry pairs that we should call their *'similarity by US counties'*. This family includes at least all measures that apply a log transformation, or a binarized location quotient, or possibly do not transform the original data at all, followed by applying a similarity among: cosine, Pearson correlation, covariance, dot product ($X^T X$) or Hausmann Hidalgo proximity. All 15 possible combinations thereof are partially equivalent, at least in our setting of employment and number of establishments by US county and 4 digit NAICS industry. The specific correspondence between each pair of these measures can be appreciated in the plots of Figure 2 which compares ranks directly. The closest the points are to the diagonal, the closest the ranking of similar pairs of activities according to a pair of measures match each other.

Among all the explored similarity measures there are two which we will use further in the remaining of the paper. We take them as references for the whole family of US county based similarity measures. These are:

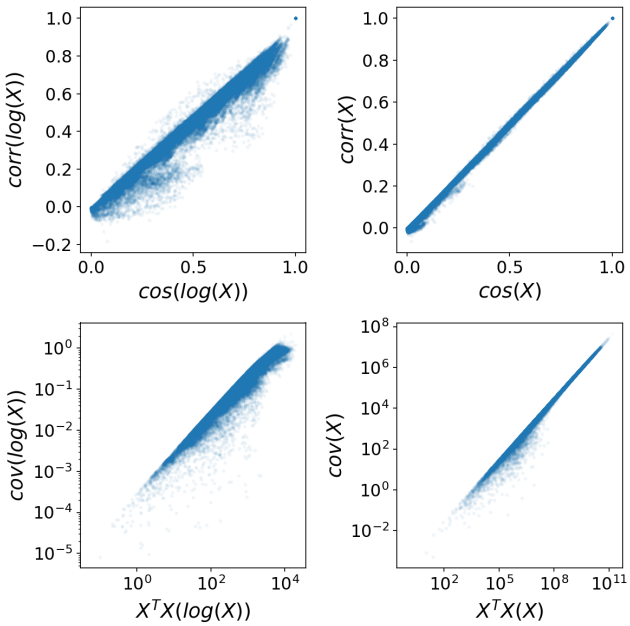


Figure 1: Scatterplots with direct comparison of selected industry pair similarity measures from US employment by county data. The notation is (top plots): *corr*(): Pearson correlation (eq. 1), *cos*(): cosine similarity (eq. 2) *cov*(); (bottom plots): covariance (eq. 3), $X^T X()$: simple joint coocurrence. The arguments can be raw data (X) or log transformed data ($\log(X)$). Top plots are depicting a near identity. Bottom plots (log log scale) show a proportionality. The proportionality factor is related to the number of counties (denominator in eq. 3). These clear connections between similarity indices suggest paths for unification of methodologies applied in different studies.

- Pearson correlation of log(X)
- cosine similarity of X

with X being the observed employment levels or alternatively the number of establishments, by US county and 4 digit NAICS activity.

The first measure is justified in that the distribution of values in rows and columns of X acquire near gaussian or other well defined distributions when transformed by log(X). It makes sense to compute Pearson correlation once the matrix values show a distribution closer to a normal. In our case, where does a high correlation of log variables lead to? To see this assume two industries X, Y such that their employment levels fulfill $Corr(\log E_x, \log E_y) \approx 1$. Then $\log E_y \sim a \ \log(E_x) + b$, with a, b real coeficients of a line. From there $E_y \sim e^b \ E_x^a$. In the cases of high correlation (all pairs with correlation higher than 0.85), we are able to fit this linear regressions and find that $a \approx 1$ in all cases, and $b \approx 0$ with a standard deviation of 0.35. All in all this tells us that in our case, a high correlation of log variables indicates that the employment levels of the pair of industries are roughly proportional to each other.

The focus on the second measure (cosine similarity) comes from a first principles approach to the problem of coexistence of industry facilities. We will show in Section 6 how cosine similarity can be used as a measure of actual coexistence (within a typical distance) of the locations that belong to a pair of industries.

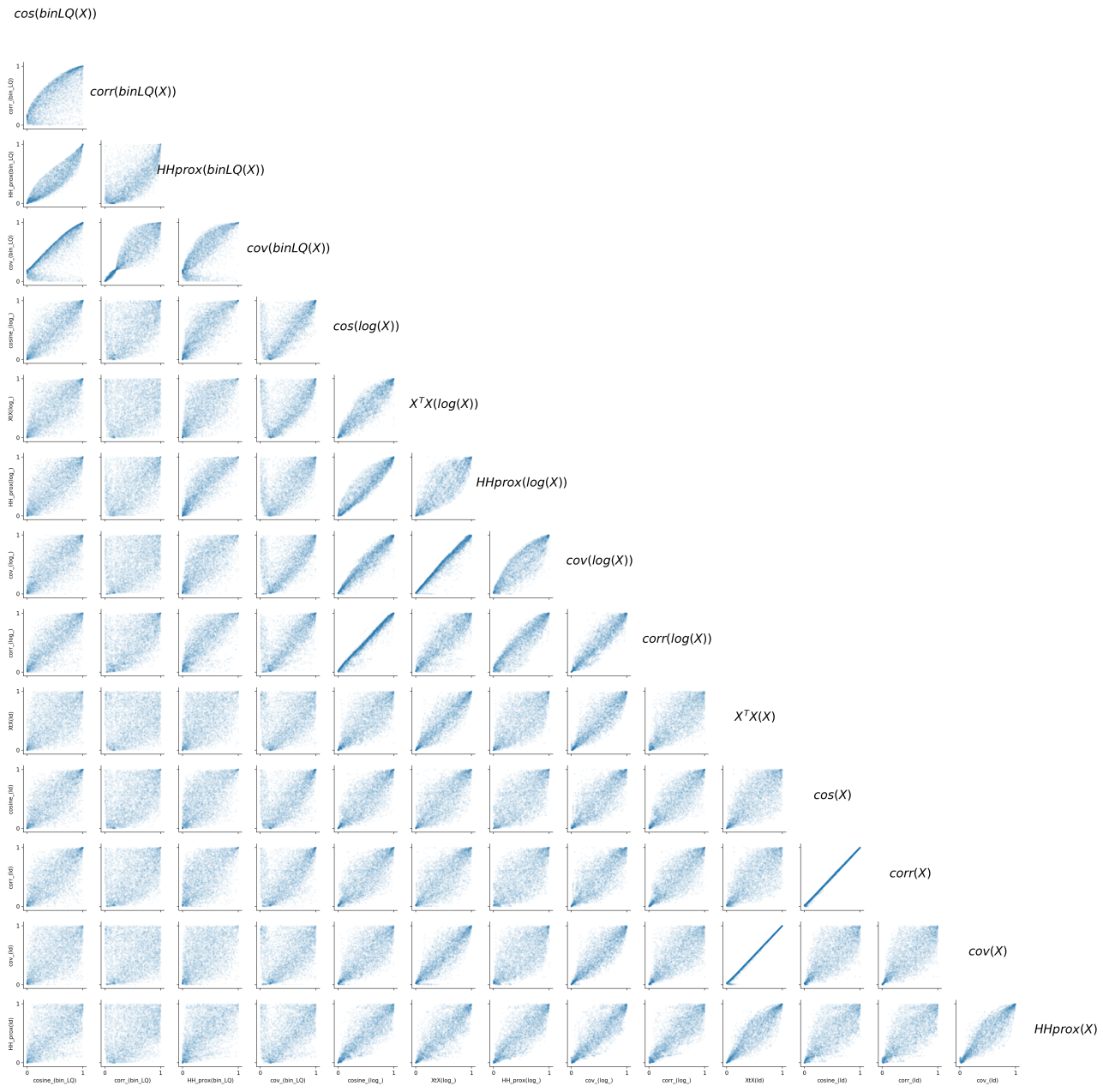


Figure 2: Comparison of rankings for multiple measures combining 5 similarity measures (cosine similarity, Pearson correlation, covariance, Hausmann Hidalgo proximity and dot product) applied on the cross section of employment by NAICS 4 digit industry and US county and two transformations thereof (logarithm and binarization of location quotient). Results applied to data of number of establishments are analogous. In some cases it is hard to asses their exact relationship analytically. Nevertheless these rank plots show that in most cases there is not a sharp contradiction on which pairs of activities are (dis-) similar to each other. The accumulation on diagonal corners, together with empty (0, 1) and (1, 0) corners show that they all agree in the extreme cases, suggesting that we can take them as alternative measures capturing a single underlying similarity. Notation: see caption of Figure 1. Also, $HHprox()$ stands for proximity as in Hidalgo et al. (2007) (minimum conditional probability). The argument $binLQ(X)$ stands for binarized location quotients.

Now, we have two indicators of similarity that can be linked to models involving employment levels or to spatial micro foundations. Furthermore, even if we do not explore a direct link between Pearson coefficient of the log variables and cosine similarity, we do see that these measures do not contradict each other. They generally agree on which pairs of industries show high similarity and they also agree with a larger family of measures that capture the same underlying characteristic of a pair of industries: their similarity by spatial distribution.

In the rest of the analysis we will use both these measures, computed for the variables 'employment level' and 'number of facilities'. The four outcomes thereof are not exactly equivalent but we will see they depict a coherent account of spatial patterns by which economic activities are distributed across the US. Results change when changing the similarity measure relatively more than they do when changing the observed variable.

6. Matching discrete to continuous coexistence measures

In this section we look for conditions under which measures of coexistence in continuous space match the outcomes of cooccurrence in administrative areas. Here we are also offering tools to evaluate caveats in the use of discrete areal data for cooccurrence, often framed under the title of 'Modifiable Area Unit Problem' (MAUP). The MAUP argument is brought by Duranton and Overman (2005) to motivate avoiding using an index like that of Ellison and Glaeser (1997). Instead, we choose to find out the conditions under which continuous and discrete coexistence indices should agree on their outcomes. In particular, we find a connection between continuous accounts of coexistence and cosine similarity on county based levels.

Works in spatial analysis have repeatedly pointed to issues when using administrative districts as the basic unit of analysis. These type of areas can have different surface areas, population or economic relevance, they can have irregular shapes and the distance that separates each pair of districts may be unacknowledged in some analyses.

To study these potential issues methodically, let us introduce a model of continuous space. Assume any establishments has an influence around it that is a function of distance to the establishment location. This influence is formalised as a probability density function.[5] An industry will be described by the collection of facilities that belong to it. And so, the influence of an industry in continuous space is the sum of probability density functions describing all plant locations:

$$F_x(\mathbf{x}) = \sum_{i}^{N_x} f_{x,i}(\mathbf{x})$$

where the subscript x refers to industry x, the vector $\mathbf{x}$ refers to position in a 2D plane, the subscript i is for each plant belonging to industry x, and N_x is the total number of plants that make up industry x.

If taken as probability distributions, the joint probability that two industries are influencing a place $\mathbf{x}$ is given by the product of probabilities: $F_x(\mathbf{x})\ F_y(\mathbf{x})$.

[5] This probability density function can have a shape designed to proxy transport costs, probability of interaction with workers of the establishment, potential demand, fits of gravity models, etc. It can typically be an exponential radial decay (Laplace), a 2D Gaussian decay, or any other reasonable bounded PDF.

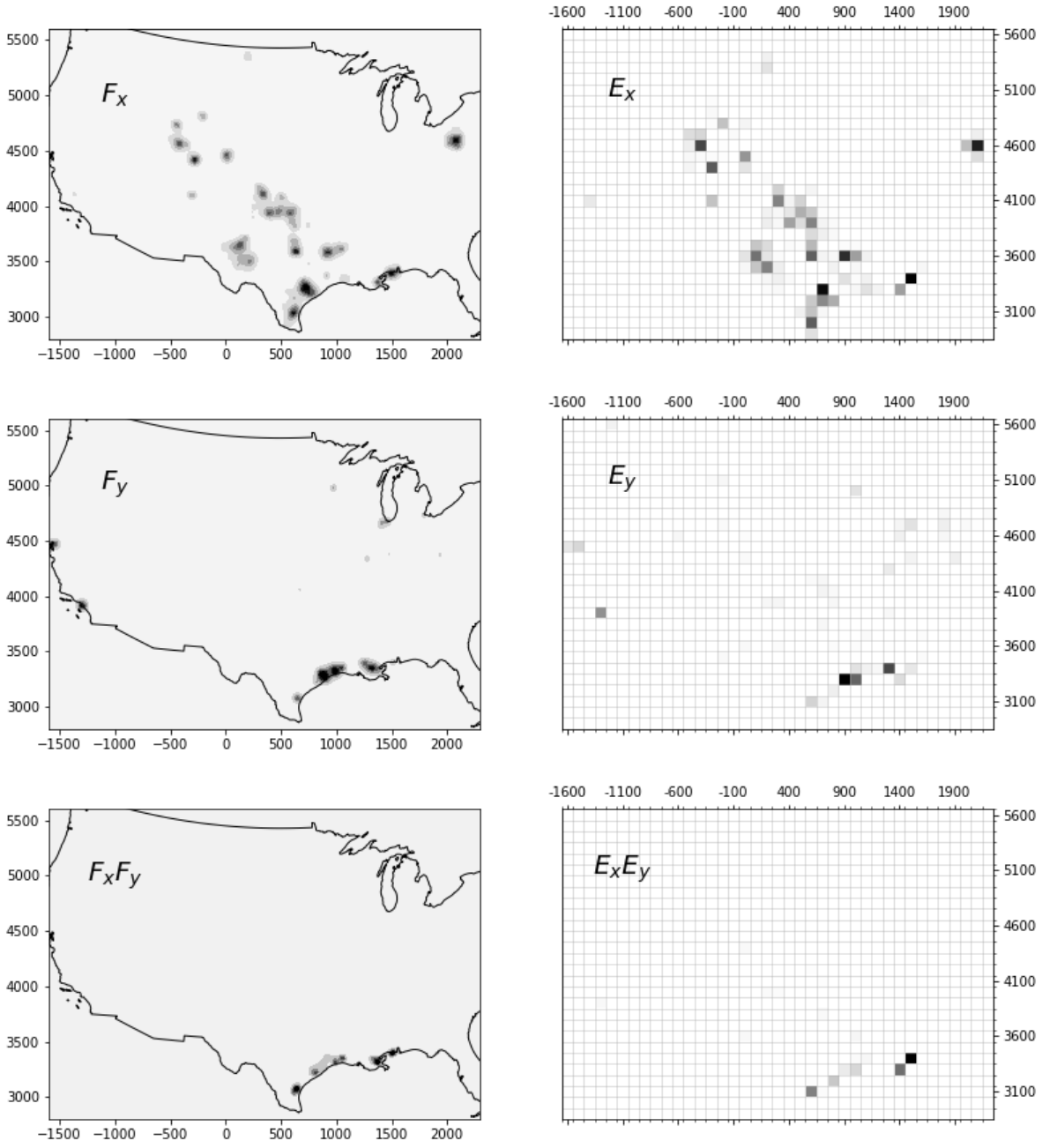


Figure 3: Demonstration of setup for continuous space (left) versus areal data (right) comparison. Top plots relate to locations of natural gas extraction fields (industry x). Middle plots relate to locations of oil refineries. Lower plots are the result of multiplying the upper plots. Grid lines depict artificial square areas of 100km width (map coordinates are UTM 14S). In these particular plots the probability function of the point locations has width $b = 100km$. Lower left are products of density functions and the lower right are coocurrence measures.

For a graphical representation of such $F_x(\mathbf{x})$, $F_y(\mathbf{x})$ and $F_x(\mathbf{x})$ $F_y(\mathbf{x})$ see the left side of figure Figure 3. If we wanted to add up all places across the country influenced by both industries x and y, we compute the integral:

$$\iint_R F_x F_y dR \tag{4}$$

where R represents the whole area of integration (the whole country).

A cosine similarity between a pair of industries is a normalized dot product. The dot product of the vector of areal employment for industry x and industry y is the x-th, y-th element of the matrix $M = E^T \cdot E$ where E is the matrix of employment by area. This is:

$$M_{x,y} = \sum_a E_{x,a}.E_{y,a} = \sum_a \left(\sum_{i=1}^{N_{x,a}} E_{xi} \sum_{j=1}^{N_{y,a}} E_{yj} \right) = \sum_a \sum_{\substack{i \in x,a \\ j \in y,a}} E_{xi} E_{yj} \tag{5}$$

For a graphical representation of $E_{x,a}$, $E_{y,a}$ and their product, see the right side of figure Figure 3. The lower plots is for the product of employment levels. The grid demarcates the modelled areas a. The exercise in this section is simply to compare a normalized volume under the $\mathbb{R}^2 \to \mathbb{R}$ function in the lower left plot to the normalized area based product in the lower right plot.

Can the dot product between two industries expressed in their areal values be compared to the overlap of their density functions? In the continuous case, in principle the density function of each firm has an overlap with all others.

Expressed from the density functions of individual plants:

$$\iint_R F_x F_y dR = \iint_R \left(\sum_i^{N_x} f_{x,i}(\mathbf{x}) \sum_j^{N_y} f_{y,j}(\mathbf{x}) \right) dR$$

This sum will potentially consist of $N_x.\ N_y$ terms, as the density function around each location can have a non negative overlap to all other locations. Distributing the product of these sums and because of the additivity of integrals:

$$\sum_{\substack{i \in x \\ j \in y}} \left(\iint_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR \right) = \sum_a \sum_{\substack{i \in x,a \\ j \in y}} \left(\iint_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR \right)$$

which can be separated into sums for each area, where the terms involving a firm x_i in area a are assigned to such area.

Now let us compare the contribution of the areal terms, both in the discrete and in the continuous case. That is, how we can draw a relation of the type:

$$\sum_{\substack{i \in x,a \\ j \in y}} \left(\iint_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR \right) \sim \sum_{\substack{i \in x,a \\ j \in y,a}} E_{xi} E_{yj}$$

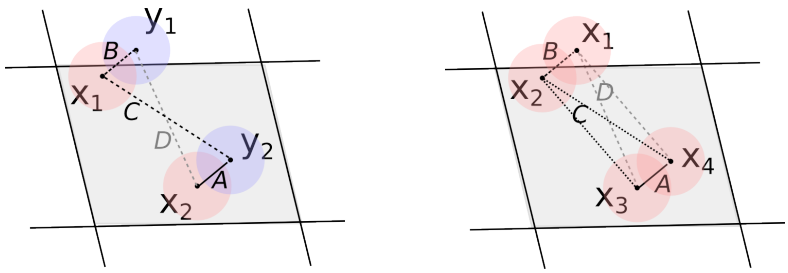


Figure 4: Micro accounting of coexistence between facilities belonging to a pair of industries (left) or a single industry (right). All links belong to 4 sets we named A, B, C, D, depending on whether they share the same administrative area and whether they actually are close to each other in the continuous space.

For managing this, we will distinguish four possible situations that apply to each of these pairs of x, y locations. To make this description easier we will say that two locations i, j *overlap* or that they are *close to each other* if $\iint_R f_i f_j dR$ is significantly larger than zero, or non negligible. There are two conditions here, firms may overlap or not in the continuous space, and firms location may lie within a single area, or not. The combination of these two conditions gives us four situations to consider. We will call these:

A The pair overlaps and shares the area.

B The pair overlaps while belonging to different areas

C The pair does not overlap, but they belong to the same area.

D The pair does not overlap and they belong to different areas.

This is illustrated schematically in Figure 4.

Splitting the pairwise relations like this will allow relating the individual terms of pairs, for pairs falling into the condition A letting us move further. The cases in B and C will introduce differences between the continuous and discrete accounts. These are the situations sometimes raised in a criticism to the use of areal data and in the discussion of the MAUP problem. Namely, points can be close to each other and lie in different areas, and points can lie in the same area while in practice being far from each other. Separating these terms allows us to find out in which cases they will become small enough for the terms in A to dominate the relation. The pairs in D contribute to the agreement between the continuous and discrete accounts [6]. Expressing the relation split according to these cases we have:

$$LHS = \sum_{i,j \in A} \iint_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR + \sum_{i,j \in B} \iint_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR$$

$$RHS = \sum_{i,j \in A} E_{xi} E_{yj} + \sum_{i,j \in C} E_{xi} E_{yj}$$

[6] Mostly, these terms will describe the pair which are definitely far from each other. In the continuous case, depending on the shape of the density functions we will have a non negative term for any pair, however they will be negligible, as it happens for the area below two gaussians separated by several standard deviations from each other

these expression will match each other if the terms in the first sum match for each i, j and the sums over cases B and C are relatively small.

For reducing the terms from pairs in C we need that areas are not much larger than the radius of influence of a location. For the pairs in B, we need that locations from a given area do not overlap with locations from neighboring areas, which will be the case if the radius of influence is not much larger than the area itself. Therefore these differences between the diuscrete and the continuous account will be relatively smaller if the area of influence we model around the locations is about the size of the typical administrastive area, not much smaller, not much larger.

As for the terms in A, the sums will be equal if each of the terms in them are equal. That is we ask that:

$$E_{xi}E_{yj} = \iint_R f_{x,i}(\mathbf{x})f_{y,j}(\mathbf{x})dR;\ \forall i,j \in A$$

6.1. Normalizations

It is useful in practice to let the coexistence of an industry wit itself be equal to 1. For this a normalization needs to be introduced in the definition of the dot product and the joint probability (equations 4 and 5). We rescale the joint probability, so that when computed for a function on itself the result is 1 and we let the normalized joint probability to be independent of a proportional scaling of the density function of some of the industries (for example by changing F_y for $2F_y$). [7] The expression for the normalized joint probability would read:

$$\frac{\iint_R F_x F_y dR}{\sqrt{\iint_R F_x^2 dR}\sqrt{\iint_R F_y^2 dR}} \tag{6}$$

An analogous requirement, but applied in the dot product of areal vectors from the last section actually leads us to an expression of cosine similarity, that is:

$$\frac{\sum_a E_{x,a}.E_{y,a}}{\sqrt{\sum_a E_{x,a}^2}\sqrt{\sum_a E_{y,a}^2}} \tag{7}$$

The separations into terms of the previous paragraphs can be kept unaltered, so that we will still want the summations B and C to be small, and we will have an expression where the amplitudes in the continuous and discrete cases are link to each other. That is:

$$\frac{E_{xi}E_{yj}}{\sqrt{\sum_a E_{x,a}^2}\sqrt{\sum_a E_{y,a}^2}} = \frac{\iint_R f_{x,i}(\mathbf{x})f_{y,j}(\mathbf{x})dR}{\sqrt{\iint_R F_x^2 dR}\sqrt{\iint_R F_y^2 dR}};\quad \forall i,j \in A \tag{8}$$

[7] This will also let it fulfill the condition that an arbitrary splitting of an industry category does not alter results significantly

6.2. Solution for industry self-overlap

Applied to some industry x on itself this will be:

$$\frac{E_{xi}^2}{\sum\limits_a E_{x,a}^2} = \frac{\iint\limits_R f_{x,i}^2(\mathbf{x})dR}{\iint\limits_R F_x^2 dR}; \quad \forall i,j \in A \tag{9}$$

We could now introduce some possible expressions for $f(\mathbf{x})$ in order to have a specific relation between these density functions and the magnitude of employment.

We can consider the following cases:

- Gaussian

$$g_{x,i}(\mathbf{x}) = \frac{t_i}{2\pi\sigma^2} e^{-(\mathbf{x}-\mu_i)^2/(2\sigma^2)}$$

- or Laplace (exponential decay)

$$f_{x,i}(\mathbf{x}) = \frac{t_i}{2b^2} e^{-|\mathbf{x}-\mu_i|/b}$$

These two functions are characterized by three parameters. An amplitude, here represented in t (the density functions for an individual plant are not normalized (the volume under them is not 1) unless t=1). There is a width parameter, given by σ and b respectively, and a position parameter given by the 2D vector μ.

The area integral of the product of two 2D Gaussian bells separated by a distance Δ is:

$$\iint\limits_R g_{x,i}(\mathbf{x}) g_{y,j}(\mathbf{x}) dR = \frac{t_i t_j}{2\pi\left(\sigma_i^2 + \sigma_j^2\right)} \exp\left(-\frac{\Delta^2}{2\left(\sigma_i^2 + \sigma_j^2\right)}\right) \tag{10}$$

and we are asking that this is comparable to $E_{xi}E_{yj}$ (Eq. 8). Note that in Eq. 10 there is a dependence with the distance Δ. While this is natural to expect, it means that the integral joint density depends not just on the magnitude of the points but also on their relative position, captured in the term $E_{xi}E_{yj}$ only in a binary fashion, either they share the same district or they do not. To deal with this difficulty we will proceed as follows: in the remaining of this section I consider the case of self cooccurrence, where $\Delta \to 0$, and in the following section I study the general case of any Δ through computational simulations.

In the limit that $\Delta \to 0$

$$\iint\limits_R g_{x,i}(\mathbf{x}) g_{y,j}(\mathbf{x}) dR \to \frac{t_i t_j}{2\pi\left(\sigma_i^2 + \sigma_j^2\right)} \tag{11}$$

Density functions of exponential decay may not have an easy expression for the volume under their product. But when $\Delta = 0$ we have:

$$\iint\limits_R f_{x,i}(\mathbf{x}) f_{y,j}(\mathbf{x}) dR = \frac{t_i t_j}{2\pi\left(b_i + b_j\right)^2} \tag{12}$$

To summarize these two results, consider self overlap of an industry (then $\Delta = 0$, and $i = j$) and let eqs 11 and 12 be expressed as:

$$\iint_R h_{x,i}^2(\mathbf{x})dR = \frac{t_i^2}{2\pi s_i^2} \tag{13}$$

where $s_i \equiv 2\sigma_i^2$ if assuming Gaussian influence around point locations, and $s_i = 4b_i^2$ id assuming an exponential decay influence (Laplace).

In the case of similarity of an industry with itself Eq. 9 links overlaps in continuous space with the observed counts of employees by establishment. Replacing 13 into 9 we can find out the intensity of the density function of an establishment in terms of the observed employment of the establishment. This tells us how to normalize the density functions for the discrete to continuous equivalence in 9 to hold. Taking square root of 9:

$$\frac{E_{x,i}}{||E_{x,a}||} = \frac{t_i}{\sqrt{2\pi}s_i}\frac{1}{\sqrt{\int_A F_x^2 dR}}$$

Where $||E_{x,a}||$ is simply the euclidean norm of the employment by area vector. From there we find we need:

$$t_i = \sqrt{2\pi}s_i \left(\frac{\sqrt{\int_A F_x^2 dR}}{||E_{x,a}||} \right) E_{x,i} \tag{14}$$

for the discrete and continuous accounts to match each other.

This last equation is telling us that the framework we devised is consistent as long as the intensity of the probability density function of an establishment is proportional to its number of employees. The proportionality factor is given by two factors: the ratio of the norms in discrete and in continuous space, and a normalization by the width of the influence (wider s_i would be met by by a smaller t_i that balances out the width effect). [8]

6.3. Solution for cross industry overlap

The generalization of the results of last section to spatial coexistence between a pair of industries (i.e. continuous and discrete accounts described by Eq. 8 instead of Eq. 9) requires that instead of the simplified equation 13 (valid when $\Delta = 0$) we use an expression such as Eq. 10 valid for any establishments distance Δ.

There are, however, important obstacles when trying to express the coexistence of establishemnts from a pair of industries in continuous space. First, the volume under the product of two (bell shaped) density functions may not have closed form expressions. This happens already when considering radial the exponential decay. Even the expression for the area of the intersection between two circles is non trivial. We sort this out by integrating numerically.

In the computational experiments that will be described next, we use square, equal size areas. Then, results are clean from irregularity of area shapes, although we do test the role played by area sizes. [9]

[8] The norm in continuous space does itself depend on t_i. To sort out this conundrum think that (once s_i are fixed) the condition of proportionality to $E_{x,i}$ implies the relative magnitudes among all t_i are fixed, and so a change in t_i implies a change of equal proportion in all t_j, $\forall j \neq .i$. This means a change in equal proportion in F_x and then the relation in 14 would be preserved.

[9] The vast majority of US counties in the contiguous US states are of similar size, making this procedure reasonable. Results might not apply if administrative areas are of extremely different sizes.

Even if we could have an approximate expression for joint probabilities at any Δ and even if we assumed square, fixed size areas, there are further difficulties that cannot be easily treated analytically. On the one hand, each pair of 'overlapping' establishments $(i, j \in A, B)$ in general need that we consider their own separating distance Δ_{ij}. In the computational experiments all separations Δ are the same, and I sweep across a wide range of Δ. Then, I am estimating the dependence with Δ if these were all the same. In an actual empirical setting, there would be an effective Δ that is representative of the distance between an average overlapping pair of establishments.

On the other hand, whether a pair of establishments is in the same area or not depends on the relative location of the i establishment within its area, and the magnitude and angle of distance to the j establishment. An analytic treatment is possible only on probabilistic grounds.

In short, the best path for comparing discrete area vs. continuous accounts in general is by computational experiments. The experiment I present here is intuitive and consists of the following procedure. Define hypothetical administrative areas by a square grid. Load the actual spatial distribution of establishments of an industry. Generate copies of this distribution, but let all establishment positions of a copy be shifted a distance Δ in random angles. Then, compute discrete cooccurrence (cosine similarity) and integrate numerically the product of continuous density functions between the original data and each of the copies. From there we will have estimates of expected discrete and continuous cooccurence, as a function of Δ and for various administrative area sizes. In this way, we will first be able to study the continuous/discrete correspondence suggested in the preceding sections as a function of the parameters of the problem.

The outcome of this experiment is first illustrated on Figure 5, applied on the location of oil refineries, with $100km^2$ square areas. Generalizations of the experiment applied to natural gas extraction locations and repeating the exercise for $10km^2$ areas are shown in Figure 6. These generalizations allow us to abstract results from the distribution each specific industry studied, and probing the role played by area sizes.

This exercise tests the decay of coexistence when we move slowly from a full coexistence situation (co location of establishments with themselves, left) to a zero coexistence situation (right). We see that the parameter describing the influence of establishments (b) governs the onset of the decay of coexistence measured in continuous space. This is to be expected. Additionally, we observe that the discrete account (where we have computed cosine similarity of total employment by areas as a measure of similarity) also presents a decay of similar shape. Given that the decay of coexistence in continuous space is shifted when increasing b, there has to be an intermediate b for which the continuous and discrete accounts match each other. In the preceding discussions, we said that for having few pairs of establishments matching conditions B ad C the width b has to be not much larger than areas, and not much smaller than areas respectively (cf 4). From the computational exercises we see that continuous and discrete accounts match each other best if $b \approx 0.3d$ (denoting area size as d).

US counties are $\sim 160km^2$ size on average. A square county of this area is $40km$ wide. The result of simulations are telling us that if we use cosine similarity as cooccurrence measure for employment by county we are testing coexistence assuming influence of establishments decaying radially with a parameter $b \approx 13km$.

In the previous section we have seen that a whole family of discrete coexistence measures are partially

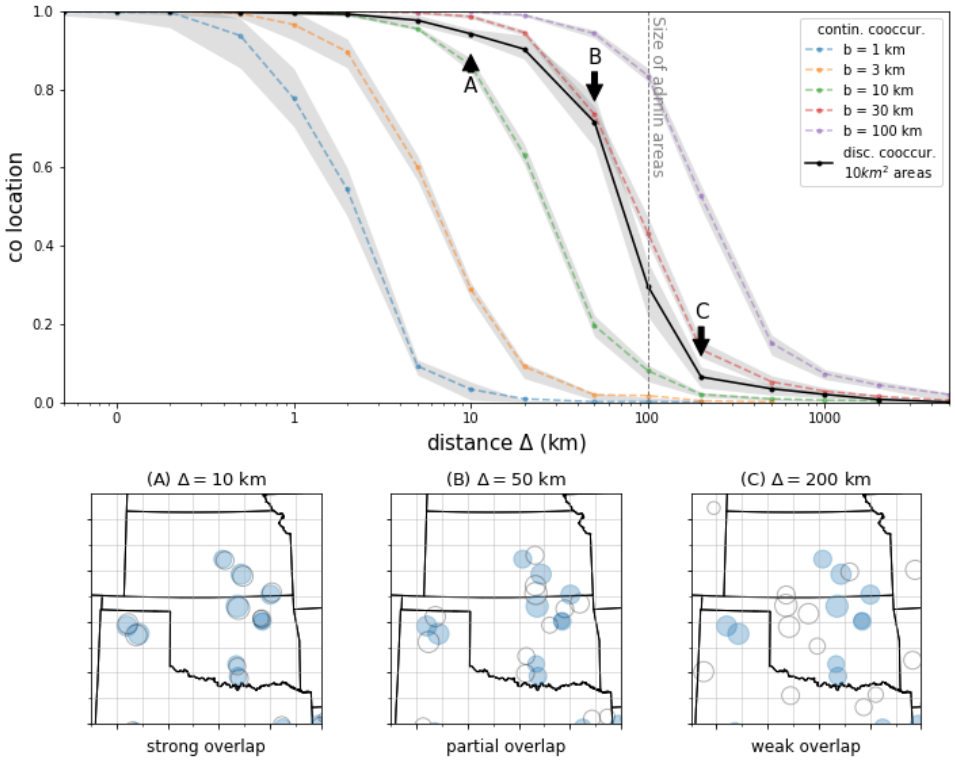


Figure 5: (Top) Decay of coexistence measures with distance Δ. Cosine similarity on admin. areas (black) and overlap of density functions in continuous space (colors). Density functions are radial exponential decay, computed for various width parameters (see legend). On the left end ($\Delta \to 0$) there is full overlap and coexistence is near 1. On the other end ($\Delta \gg 1$) there is no overlap and coexistence is near zero. The interesting feature is the transition between these extremes. We can see that discrete coexistence matches the continuous account of coexistence only when the width b of establishments density functions is slightly less than $\hat{b} = 30km$. The dashed vertical line shows the area size.
(Bottom) Maps with circles around establishment locations (blue) and Δ shifted locations, for three values $\Delta = 10km, 50km, 200km$, denoted as A, B, C on the upper plot.

equivalent. So that with all these developments we are finding the concrete meaning of coexistence measures when applied on US county data.

In Figure 6 I replicate the decay test for two area sizes and two different industries. From here we can see that results just discussed are largely equivalent on both industries tested. Also, we confirm that the decay of area based measures is directly related to the size of areas. Larger areas mean considering coexistence at a larger distance (the relative location of the black curve and the vertical gray line is preserved when changing area size).

7. Application: what correlation structure tells about industries and regions of the United States.

So far we have seen that many similarity techniques are partially equivalent to each other and can be interpreted as coexistence in continuous space. We have also seen that area size determines the distance at which coexistence is detected. In the remaining sections of the paper I show the actual correlation structure we observe in our data and discuss it briefly.

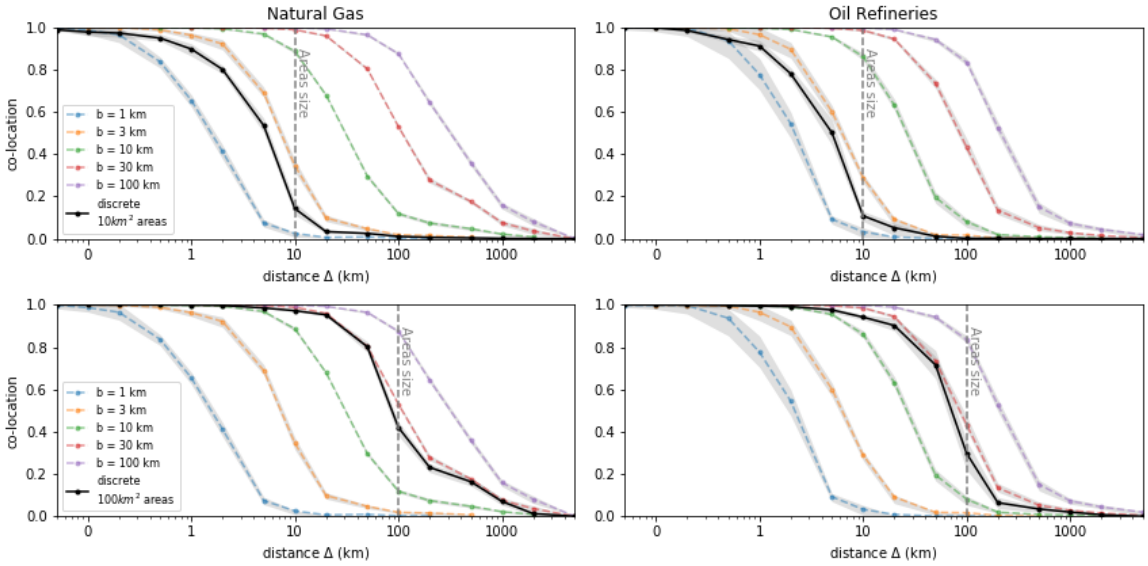


Figure 6: Decay of coexistence measures with distance Δ. Cosine similarity on admin. areas (black) and overlap of density functions in continuous space (colors, see legend). Results for two industries (left - right) and for two area sizes (10 km, top - 100 km, bottom). The decay of discrete area coexistence (black) appears linked to area size (gray vertical line), as they both shift by the same amounts.

The square matrix encoding the correlation structure can be translated into an adjacency matrix, i.e. a matrix that defines a network. In such networks each industry is a node, it can be taken as an *industry space*. Each node has a spatial distribution across counties. Nodes within a community, or cluster of tightly connected nodes, approximately share a common distribution across space. Then, because of having geographical units on one side of the cross section, the correlation structures will also lead to *geographical patterns*.

In the next subsection 7.1 we introduce the methods applied to arrive at an industry space and geographical patterns and in subsections 7.2 and 7.3 I show and discuss the results.

7.1. Methods for analyzing correlation matrices

There are techniques particularly adapted to processing similarity matrices. The eigenvalues of random matrices are studied theoretically and have known distributions. Correlation matrices however, tend to have a single large eigenvalue linked to the main mode of the matrix. Subsequent eigenvalues are much smaller but can still be larger than the largest expected eigenvalue of the random matrix, therefore suggesting they are linked to non-random structure of the correlation matrix. The remaining majority of eigenvalues match the eigenvalues of the random matrix.

It turns out, that a correlation matrix can be expressed as a sum of components related to each eigenvalue and their eigenvectors. Indeed, because of being a real symmetric matrix, $C_{(p\times p)}$ fulfills $C = U\ \Lambda\ U^{-1}$ with U an orthogonal matrix (i.e. $U^{-1} = U^T$) so that $C = U\ \Lambda\ U^T$. The similarity (real symmetric) matrices can be

decomposed as:

$$C = \sum_k \lambda_k u_k u_k^T = \sum_k \lambda_k V_k \tag{15}$$

where λ_k, u_k denote the $k-th$ eigenvalue and eigenvector, and so that $V_k \in \mathbb{R}^{p \times p}$.

This connection is useful for 'cleaning out' the correlated background (main eigenvalue) and letting us capture (slightly) far from average values of the correlation matrix that suggest (positive, null or negative) association between industries.

The decomposition in Eq. 15 works similarly for cosine similarity and correlation of logs matrices. We illustrate it graphically in Figure 7 where we can grasp the conceptual idea of what we achieve with this decomposition: removing the main component leaves us with an underlying structure which we call groups structure. Further components only contain small fluctuations.[10]

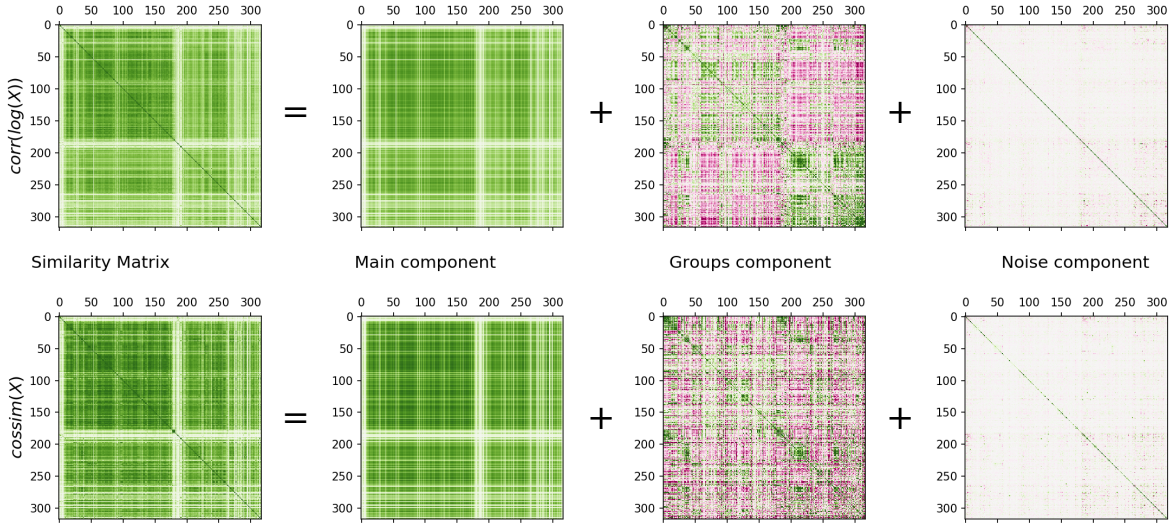


Figure 7: Decomposition of similarity by eigenvalue components (eq. 15)

Even if it would seem natural to take a correlation matrix, or cosine similarity matrix directly as adjacency matrix of a network, it is better to do this extra processing first. It is not uncommon that the majority of industries follow a common trend (e.g. they are nearly proportional to all-industries totals), which is reflected by a degree of correlation among most industry pairs, and therefore a 'complete' network structure with a single community.

At this point, we are close to the framework of a principal component analysis (PCA). For applying this technique we need to first *center* the dataset by subtracting industry means. Use X to denote the centered data. The covariance matrix is $C = X^T X/(n-1)$ and we have to diagonalize it to arrive at the principal components.

[10] One can take 20 or 30 components without much difference in outcomes.

This diagonalization leads to $V^{-1}CV = D$, where D is the diagonal matrix with eigenvalues of C, and U has the eigenvectors of C in columns. Then, for concluding the PCA decomposition, we would look at the eigenvectors of the first few largest eigenvalues.

On the one hand, PCA can relate to the processing of correlation matrices I am proposing, because in both cases we are diagonalizing the similarity matrices. Nevertheless, what I am proposing is to express the similarity matrix itself as a sum of a first eigenvector (modal) component plus subsequent few eigenvector groups components (Eq. 15) as in Plerou et al. (1999) and later plotting communities of this network on the map. As opposed to taking principal components that can be plotted on the map.

The two techniques can be seen as complementary analyses. Studying their connections fully can be certainly interesting. Some of the difficulties though, have to do with the data centering. We may take logs as a pre-processing step, still there are key difference between Pearson correlation and covariance that would need to be addressed. If we did the analysis as in Plerou et al. (1999) but using the covariance matrix, the gap between this technique and PCA would be: what is the difference between spatial patterns from the principal eigenvectors, and spatial patterns shown by communities of the 'groups' contribution to the covariance matrix. This is an open question for future research.

We apply Scikit Learn (Python module) spectral clustering algorithm with all its options in default values[11]. We repeat the fitting with 10 (or 15) different random seeds and obtain groups of industries that are grouped together in all these optimizations. This way we find 'cores' of comunitites that are strongly similar among each other and weed out activities that can jump in between communities because they link weakly to more than one core.

As we explained in the previous section, we explore the outcome of applying Pearson correlation of logs and cosine similarity to both employment levels and number of establishments. These constitute four criteria that we label: A - corr(log(establishments)), B - cos(establishments), C - corr(log(employment)), D - cos(employment). We apply the discussed community detection process on each of these four situations and see that communities from these four outcomes partially overlap. For this specific step, the algorithm we apply is to see in which clusters (computed in one of the four combinations) more then 50% of the activities of any given cluster are contained. Reciprocally, we ask that it cluster represents at least 10% of the cluster it is potentially contained in. In that way, if for instance the activities of two clusters computed by D - cos(employment) and B - cos(establishments) are 10 in each and there is an intersection of 6, we associate these two into a single *component*. We study these components. As another example, if a cluster of 5 activities from C - corr(log(employment)) is contained into a very large cluster of 60 activities from D - cos(employment) we keep it separate. The idea is to not merge all small, possibly interesting clusters of activities into very large overarching components.

The goal of this process is to reassure that outcomes are robust enough to not fade away when changing choices of similarity matrix or the specific measure of economic activity.

To sum up, the processing steps for results in Section 7 are the following:

[11] Documentation for sklearn.cluster.SpectralClustering

- Averaging yearly values in 2002-2007. This can be stored as a rectangular table X of shape (3272, 320), with counties as rows and industries as columns.

- Computing cosine similarity and Pearson correlation of the log values between industries from this cross-section.

- Decompose the similarity matrices by their eigenvalues and study the structure of groups, which is non random and independent from the general trend.

- Apply spectral clustering to detect cores of activities that link strongly to each other, see what is the geographic pattern that they depict and discuss these outcomes.

7.2. Results: Network of industries

Let us begin by presenting the network structure of industries. To begin understanding the outcome, we can look at Figure 8. Here each node corresponds to a NAICS 4-digit industry. The plot on the left is derived from correlation of log levels and the one on the right from cosine similarity. Given that the results are largely analogous when changing between number of establishments or employment level, we extract the groups component of the similarity matrices and average the similarity computed from each of these variables to arrive at a single network plot.

Following the community detection methods described in detail in Section 7.1 I detect 11 components. These are represented in colors in the networks of Figure 8. They are listed in Table 2.

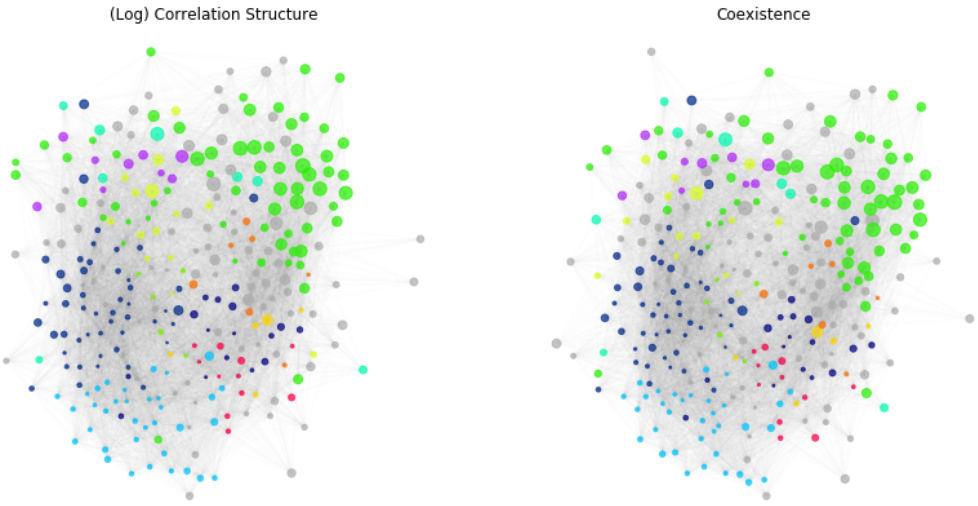


Figure 8: Networks of industries. Left: from groups component of correlation of log levels. Right: from groups component of cosine similarity. Edge weights computed from employment levels and number of establishments are averaged for each plot. The colors depict the components we built from clustering in each of the four variable - similarity combinations (cf Section 7.1). LINK TO INTERACTIVE PLOT

An online version of this plot (link in figure caption) allows exploring the network interactively. To gain further intuition into the regions of the plotted network, the plots of Figure 9 successively highlight some of the most common words in industry titles: *manufacturing, services, transport, wholesalers, stores*. We use the coexistence network, although the outcome of this exercise is largely analogous if one used the correlation structure.

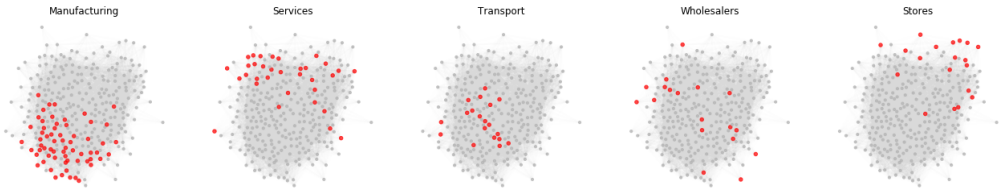


Figure 9: Network of industries. Highlight of frequent words in industry titles. The aim of this plot is to help understand the regions of the networks plotted in Figure 8.

Finally on Figure 10 we paint nodes according to wage levels. Even if it is not clearly distinguishable in the plot, certain components of the network are characterized by a higher than average wage level. These are the components related to urban activities, including in NAICS categories: 51 Information, 52 Finance and insurance, 53 Real estate and rental and leasing, 54 Professional and technical services.

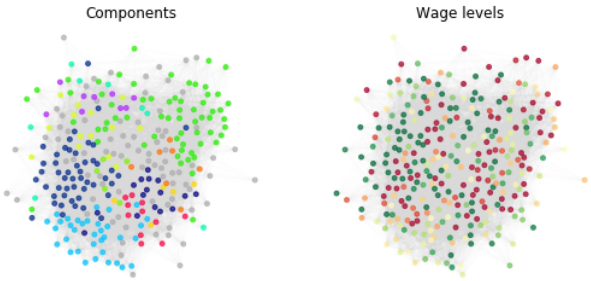


Figure 10: Network of industries. Clusters by community detection (left) and wage levels (right). Only clusters of urban activities are characterized by a (higher) average wage level. The rest of the clusters have mixed wage levels.

7.3. Results: Geographical patterns

The so called *components* we just discovered are neighborhoods of the network of industries. Neighbors in this network show a high locational correlation, they share a common distribution over space. Neighborhoods of the correlation structure can thus be identified to spatial patterns. In this section we explore the patterns coming out of this analysis.

I group the components into four *themes*, or types of spatial distribution (cf table 2). These are *population*, *cities*, *land uses* and *manufacturing*. There is a different factor dominating the location of industries in each of these themes. Respectively these are consumer demand, urban agglomeration externalities, availability of a natural resource, and manufacturing externalities.

Theme	Component
Population	Non tradables: stores and personal services
Cities	Large city economies I
	Large city economies II
	Other high wage activities
Land Uses	Agriculture and Food I: Ranching
	Agriculture and Food II: Corn Belt
	Water Economy
	Fuels and Mining
	Forests and Timber
Manufacturing	Manufacturing I: Steel Belt
	Other manufacturing and other activities

Table 2: Summary of detected *Themes* and *Components*

The following table summarizes the components we could detect:

Next we review them in further detail.

Population

The activities in this theme are those that most closely match the distribution of population. Even if these activities may not follow it exactly, the distribution of population is a reference to many accounting considerations and as such its acknowledgement is useful and justified.

In practice, the activities that fall in this category are mostly retail shops and personal services (such as restaurants), in other words, consumer goods. Two factors combine for the location of shops to show this pattern. These businesses have people as customers, and proximity to customers is central in their strategy (Berman, 2010; Runyan & Droge, 2008). In these industries, demand appears as a decisive factor for location. References discussing these facts are multiple. For example, referring to Los Angeles, Fujita et al. (1999) distinguish *"on one side of film studios, arms manufacturers, and so on who produce for the U.S. or world market, on the other side of restaurants, supermarkets, dentists, and so on who sell only locally."* (p 27). These latter are precisely the types of activities that fall under our 'population' theme. M. Porter (1980) also draws a connection between dependence on demand, and intensity proportional to population: *"In consumer goods, demographic changes are one key determinant of the size of the buyer pool for a product and thereby the rate of growth in demand. The potential customer group for a product may be as broad as all households, but it usually consists of buyers characterized by particular age groups, income levels, educational levels, or geographic locations."*.

Cities

Cities are of course a notorious singular feature of our society. There is an abundance of discussions about what is the magic of cities, with questions approached from a variety of literature strands. When it comes

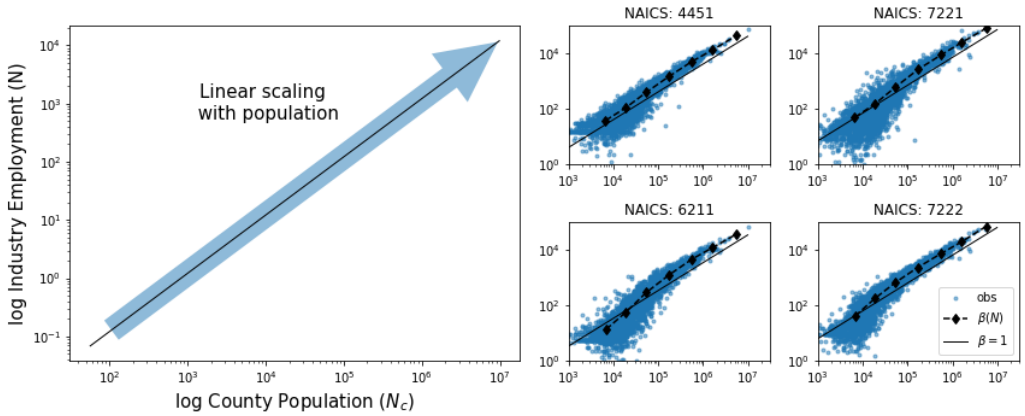


Figure 11: Population-linear scaling of activities in the 'population' theme. Right: scatterplots of data (NAICS 4451 Grocery stores, NAICS 6211 Offices of physicians, NAICS07 7221 Full-service restaurants, NAICS07 7222 Limited-service eating places). Left: qualitative scaling pattern.

Non tradables: stores and personal services.

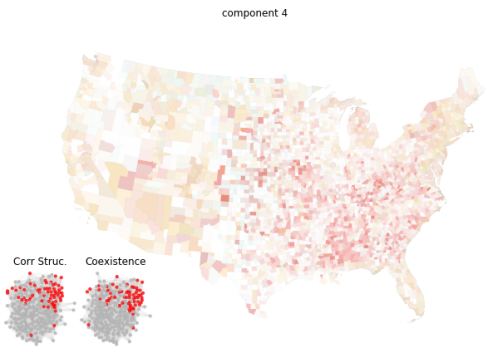
Distribution	Activities
	NAICS 238 Construction contractors NAICS 44-45 Retail trade NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services NAICS 62 Health care and social assistance NAICS 72 Accommodation and food services NAICS 81 Other services, except public administration

Table 3: Non tradables. LINK TO INTERACTIVE MAP

to quantification, a tool that appears promising and convenient is that of scaling, given it quantifies apparent externalities related to city size.

The *cities* theme comprises activities such as NAICS 5112 software publishers, NAICS 5418 advertising, NAICS02 5161 Internet publishing and broadcasting, NAICS 5416 management and technical consulting services, NAICS 4251 electronic markets and agents and brokers, NAICS 5415 computer systems design, NAICS 5616 investigation and security services, NAICS 5614 business support services, NAICS 5414 specialized design services, NAICS 5511 management of companies and enterprises.

Let us first show that these activities present particular features of scaling, which distinguish them from activities in the 'population' theme. In this way we also offer a possible path for connecting our results with some formal accounts. Then we will briefly mention strands of literature studying the phenomena of cities. Discussing in depth these formal and conceptual approaches to cities is however out of the scope of this paper.

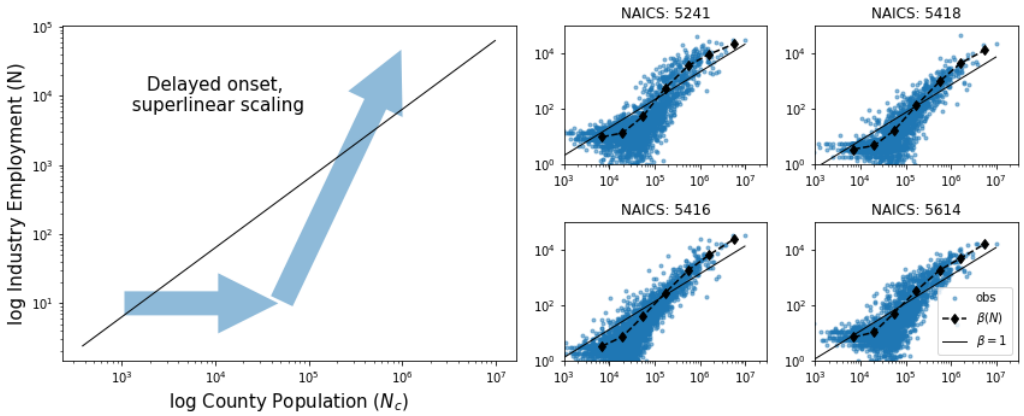


Figure 12: Delayed onset and superlinear scaling of activities in the 'cities' theme. Right: scatterplots of data (NAICS 5241 Insurance carriers, NAICS 5416 Management and technical consulting services, NAICS 5418 Advertising, PR, and related services, NAICS 5614 Business support services). Left: qualitative scaling pattern.

On figures 11 and 12 we show the scaling patterns of industries in the 'population' and 'cities' themes respectively. The schemes on the left show our interpretation of such scaling patterns, exaggerated for clarity. The horizontal axes stand for county population, and the vertical ones stand for population in each of the industries. There is a point for each county with non zero employment in the industry. Activities which abound proportionally to population would show all points on the diagonal line. Instead, we find that activities in the 'cities' theme are less than proportionally represented in small town and cities, but catch up to be more than proportionally represented in larger cities. Actually, the distinction between these two groups is somewhat blurry. All activities have a mixture of the two patterns, although it is clear that activities in each of the themes lean clearly closer to one of the two limiting cases.

The activities in the 'cities' theme would typically be deemed as *complex* in the sense that they did not exist decades ago and even today they are missing in poorer, less developed regions. They can then be conceived as activities near a technological frontier. It is expectable that this type of activities arise in large cities (as opposed to small towns or rural areas) although formalizing this intuition is challenging. The framework of scaling (Bettencourt et al., 2007) may be helpful for a goal of eventually quantifying correctly. The superlinear scaling (Bettencourt et al., 2007; Gomez-Lievano et al., 2012) would suggest that largest cities have scale advantages over mid size cities. A superlinear scaling of a complex (knowledge or technology demanding) activities would be consistent with *most of this activity appearing in large cities* and *most of the activity of a large city being complex*, but the details of this relation need to be worked out carefully.

When it comes to the singularity of cities conceptually there are of course studies in many strands of literature which would be hard to review comprehensively. Cities are relatively more productive and show higher average educational attainment. Marshall (op cit) dedicates lines to externalities involving skilled labor, although he typically refers to towns or certain industrial districts, more than to large agglomerations as we know them today. Instead, Jacobs (1970) centers her thesis on the fact that innovations near a technological frontier tend to be engendered in large cities before possibly finding ordinary longer term adoption in other types of geographies.

Indeed the activities we classify in the 'cities' theme are near the technological frontier and are clearly knowledge intensive. There is a richness of recent works studying learning and diffusion of specialized knowledge (Puga, 2010) and the development of knowledge intensive, complex activities (Balland et al., 2015; Balland et al., 2020; Boschma et al., 2014) to name only a few of these.

Large city economies I

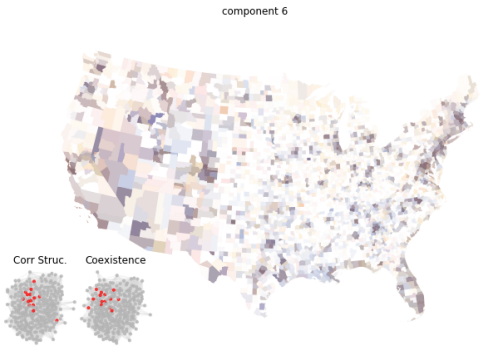
Distribution	Activities
	NAICS 5112 Software publishers NAICS02 5181 Isps and web search portals NAICS 5182 Data processing, hosting and related services NAICS 5415 Computer systems design and related services NAICS 5417 Scientific research and development services NAICS 5612 Facilities support services NAICS 5619 Other support services NAICS 6114 Business, computer and management training

Table 4: Large city economies I. LINK TO INTERACTIVE MAP

Large city economies II

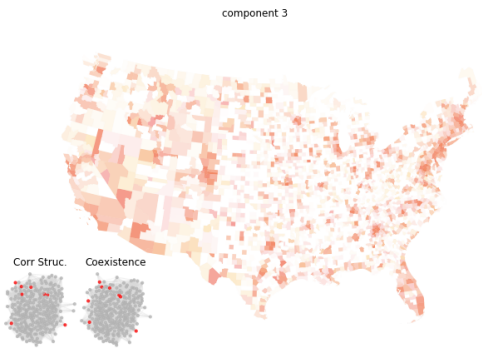
Distribution	Activities
	NAICS 51 Information NAICS 52 Finance and insurance NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services

Table 5: Large city economies II. LINK TO INTERACTIVE MAP

Other high wage activities

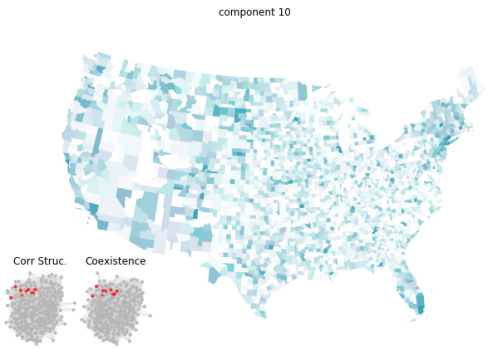
Distribution	Activities
	NAICS 51 Information NAICS 52 Finance and insurance NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services

Table 6: Other high wage activities. LINK TO INTERACTIVE MAP

Land Uses

In words of Ellison et al. (2010) *"Natural advantages, such as the presence of natural inputs, differ spatially, and firms may choose locations to gain access to those inputs."*. This third theme includes activities that have a natural resource as important input. Or else, those that are near the upstream end of the supply chain and choose to locate their operations near the primary establishments to save on transport costs. In these theme we find five components each characterized by spatial patterns that point inequivocally to a type of natural resource.

Two components are related to agriculture, one including the grazing lands of Texas' west and other fertile areas for the production of crops and fruits in Washington state and in the Central Valley of California, and the other one centered on the Midwest corn belt region and Mississippi Valley.

Agriculture and Food I: Ranching

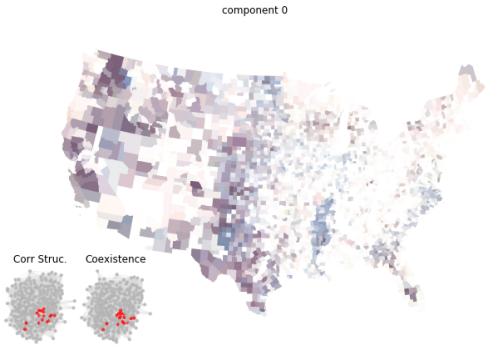
Distribution	Activities
	NAICS 111 Crop production NAICS 112 Animal production and aquaculture NAICS 311 Food manufacturing

Table 7: Ranching. LINK TO INTERACTIVE MAP

Agriculture and Food II: Corn Belt

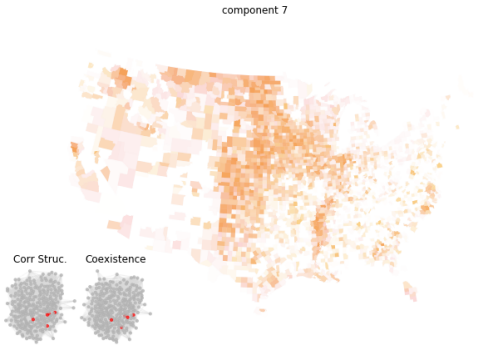
Distribution	Activities
	NAICS 1111 Oilseed and grain farming NAICS 1122 Hog and pig farming NAICS 311 Food manufacturing NAICS 4245 Farm product raw material merch. wholesalers

Table 8: Corn Belt. LINK TO INTERACTIVE MAP

One component comprises all fishing activities and touristic and transportation activities that take place in rivers, lakes and coasts. Ellison and Glaeser (1997) and Ellison et al. (2010) discuss repeatedly about the

importance of natural resource endowments for this type of activities. In a more formal passage *"the effects of natural advantages on profits are captured by the random variables* $\{\pi_i\}$*, which are chosen by nature at the start of the process when it assigns resource endowments to each area [...] these variances might be high in the shipbuilding industry because the profitability of a state will depend greatly on whether nature has put that state on the coast."* (actually the level of such π_i would be high in coastal states, not just their variance).

Indeed we are detecting patterns that seem to point at natural resources and determining what are the industries in them. With this exercise we are able to detect those activities to which Ellison and Glaeser are referring. In the case of coastal activities, the counties endowed with access to water form one of these spatial patterns. From that point of view, the components we are showing would be telling the counties endowed with a specific natural resourse (fertile lands, forests, water access or minerals).

Water Economy

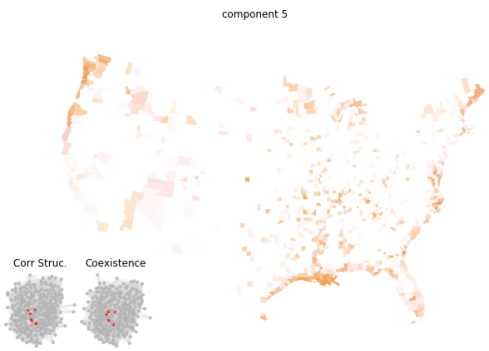
Distribution	Activities
	NAICS 1141 Fishing NAICS 3117 Seafood product preparation and packaging NAICS 3366 Ship and boat building NAICS 4831 Sea, coastal, and great lakes transportation NAICS 4832 Inland water transportation NAICS 4872 Scenic and sightseeing transportation, water NAICS 4883 Support activities for water transportation

Table 9: Water economy.

The next component in the natural resource theme includes activities of oil and gas extraction, as well as extraction of other minerals. In addition, some of their first downstream activities, such as manufacturing of petroleum and coal products (NAICS 324) fall into this component.

Ellison and Glaeser (1997) notes *"plants in the cane sugar refining and shipbuilding industries might be coagglomerated because coastal locations provide higher profits both for shipyards and for importers of bulky commodities"*. An additional quote on the same idea: (Ellison et al., 2010) *"Agglomeration and coagglomeration can also appear empirically even if there are no gains from locational proximity. [...] For example, the ship building and oil refining industries might be coagglomerated simply because both prefer coastal locations."*.

As a way to rationalize these ideas, consider first that if two activities overlap fully then they essentially share a single distribution. Otherwise it can happen that a pair of activities of a different kind coincide in some context. Indeed it is true that many oil refineries lie on the coast (Texas, Louisiana) and then share space with coastal activities. The volume under their joint density functions as in Section 6 will be non null along this coast and will contribute to certain overlap in continuous space. Also, counties on this coast will have employment in both industries and so they will add to measures of co-occurrence. The technique we are applying, however, is

made for distinguishing these two factors and classifying industries accordingly.

Fuels and Mining

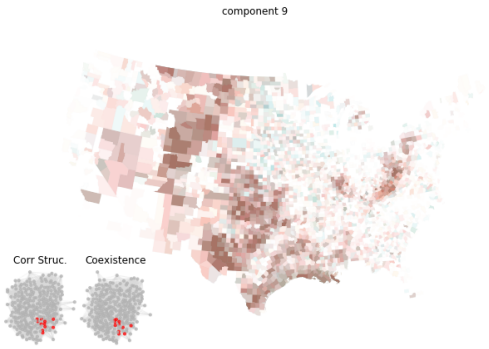
Distribution	Activities
	NAICS 21 Mining, quarrying, and oil and gas extraction NAICS 324 Petroleum and coal products manufacturing NAICS 3251 Basic chemical manufacturing NAICS 486 Pipeline transportation

Table 10: Oil and gas. LINK TO INTERACTIVE MAP

The last component we find in the natural resource theme are forest products industries. The pattern presented by this component matches closely the distribution of natural forests. The large majority of forest area in the US is non industrial privately owned. If the fraction of industrial timberland is approximately uniformly distributed it is expected that the primary stages of wood processing industries will follow the overall distribution of natural forests. At the upstream there is supply of raw materials including fuelwood and industrial roundwood which depend directly on the forest area and forest stock. This needs to be supplied to processing facilities (eg. mills) and it is convenient for these industries to be near the resource. In between is the transformation of wood into products, and at the other end is the demand for end products (sawnwood, wood-based panels, paper and paperboard) (Alig et al., 2003). The industries in this other end are grouped among the manufacturing activities and the logistics of their value chain may play a more important role to explain their spatial distribution.

Forests and Timber

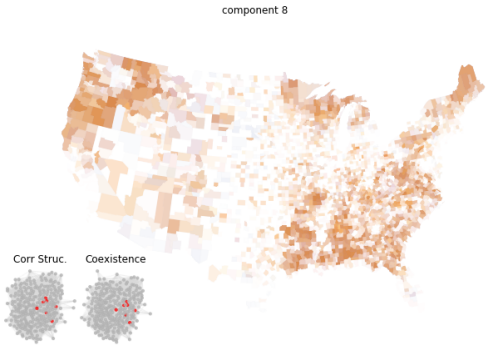
Distribution	Activities
	NAICS 1131 Timber tract operations NAICS 1132 Forest nursery and gathering forest products NAICS 1133 Logging NAICS 1153 Support activities for forestry NAICS 3211 Sawmills and wood preservation NAICS 3212 Plywood and engineered wood product mfg. NAICS 3371 Household and institutional furniture mfg.

Table 11: Forests. LINK TO INTERACTIVE MAP

Manufacturing

The fourth and last theme is manufacturing. It comprises activities in the NAICS categories 31 to 33. The distribution of these activities does not point clearly to population, natural resources or cities. The factors then left to explain the location decisions of industrial establishments are externalities of different kinds, built on historical paths of arbitrary or reasonable origin. Such externalities have been the focus of extensive research. As an early antecedent there is the proposed organizing criteria of Marshall (1890), who directed attention to a few mechanisms simplified as *transport cost externalities* (mainly the availability of intermediate goods), *availability of labor* (labor market pooling being the typical example) and *'ideas'*, meaning specialised and technical knowledge. These have been joined over time by other mechanisms such as proximity to a natural resource, pooling of demand, costs of distribution, competition forces, among others (Beaudry & Schiffauerova, 2009; de Groot et al., 2016; McCann & Folta, 2008). All these might influence firms decision to base their plants. However, each of these mechanisms is qualitatively different and may combine in special ways to determine each of the specific plant location choices that happened over time. Heterogeneities are expectable and have been the subject of recent studies (Diodato et al., 2018; Ellison et al., 2010).

The main industry sectors I identify are linked to the steel value chain, including the automotive and autoparts industry and their suppliers. This single example presents most, if not all of the mentioned externality channels across a network of thousands of heterogeneous businesses located throughout the US (with higher density in the Midwest region south of the Great Lakes). Other sectors in this theme, such as the textile industry are examples of activities that have developed in regional clusters. North Carolina has the largest textile mill industry and is the leading US state in textile exports. This industry existed for more than a century in the region. It is an example of path dependency in economic development and it also suggests an important role played by industry related tacit knowledge and possibly the existence of externalities leading to the formation of the cluster. All this would help explain why the industry did not continue to grow in regions other than North Carolina.

Manufacturing I: Steel Belt

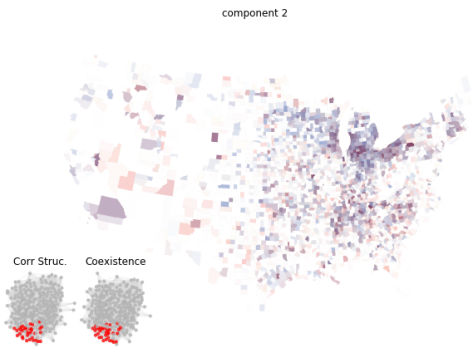
Distribution	Activities
	NAICS 325 Chemical manufacturing NAICS 326 Plastics and rubber products manufacturing NAICS 327 Nonmetallic mineral product manufacturing NAICS 331 Primary metal manufacturing NAICS 332 Fabricated metal product manufacturing NAICS 333 Machinery manufacturing NAICS 335 Electrical equipment and appliance mfg. NAICS 336 Transportation equipment manufacturing

Table 12: Manufacturing. LINK TO INTERACTIVE MAP

Other manufacturing and other activities

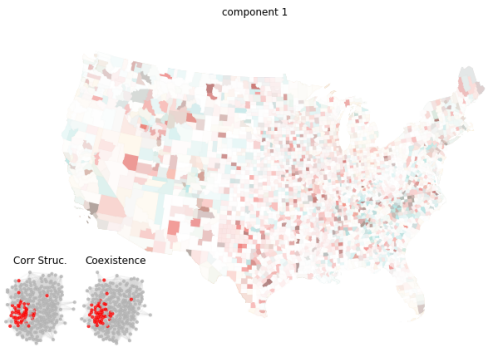
Distribution	Activities
	NAICS 31-33 Manufacturing NAICS 48 Transportation NAICS 51 Information NAICS 52 Finance and insurance

Table 13: Other than steel belt manufacturing and other activities. LINK TO INTERACTIVE MAP

8. Conclusion

This paper is centered on understanding the correlation structures derived from cross sectional data of intensity of economic activities by (a large number of small) geographical units. First I show how a variety of techniques for detecting coexistence from this type of data are partially equivalent among themselves (Section 5). I then explore the connection to coexistence accounts computed from continuous space (i.e. based on establishments' point locations and employment levels) (Section 6). Finally from these similarity measures I compute a network of industries (industry space) and I show that communities in this network stand for clear geographical patterns linked to specific drivers of estabishments' location.

More specifically I show that, both on employment and in number of establishment data, both using data in linear levels and in log levels, cosine similarity tends to match Pearson correlation, and covariance is proportional to simple joint cooccurrence $X^T X$. These are the clearest relations among similarity measures in our data, but in fact I show that among all techniques that apply cosine similarity, Pearson correlation, covariance, joint cooccurrence, or Hidalgo et al. (2007) proximity as similarity measure on raw data, log transformed data, or binarized location quotient data, there is a rank correlation. In other words, any of these fifteen slightly different techniques lead to partially equivalent rankings of industry pairs by similarity. In the remaining sections we use Pearson correlation of log levels and cosine similarity of linear levels as proxy for the whole family of similarity measures. These two are chosen because they are closest to having theoretical and practical interpretations, unlike some of the other similarity measures.

We also see that cosine similarity of the vectors of intensity by area can be linked to actual overlap of point locations. The basis of this continuous-discrete identity is deduced by using calculus. The conclusion though is reached thanks to computational simulations that acknowledge the arbitrariness of actual distributions of point locations of establishments, a task that is quite challenging to complete analytically. We find that for square shaped administrative areas, assuming an exponential decay (of typical distance b) of the influence of a point location with distance, cosine similarity matches actual coexistence of facilities within a radius b about

30% as large as the area width. In this way we offer a way out of the conundrum of the modifiable area unit problem, at least when it comes to the computation of correlation structures. At the same time we discover that cosine similarity of employment levels by area has a relevant micro interpretation. Co-location from areal data (cosine similarity) is tuned to measure interactions acting at a distance proportional to the average size of areas. Correlation structures can then be a lens focusable at different distances. This may allow studying heterogeneities across industries by sensing at which distances a pair of industries coexist with each other.

Once the interpretation of these similarity measures is clear, I look at the 'industry spaces' they imply and I map the neighborhoods of these networks. The goal of this last exercise is to validating the techniques by analyzing the outcomes. We determine several distinct patterns that explain the spatial distribution of most activities. The data driven approach of looking at the correlation structures leads directly to concepts often theorised in Economic Geography. The detected patterns (and drivers) for the location of most industries are among the following: population (consumer demand); agriculture, fuels and minerals, forest and timber, coastal and water economies (presence of natural resource); manufacturing (agglomeration forces) and large cities (urban externalities). These themes and components of the correlation structure are illustrated and discussed briefly.

With this exercise, we have used empirical data and objective mathematical tools (correlation matrices, its eigenvalue decomposition and spectral clustering to detect communities) and arrived at a classification of activities. This analysis was prohibitive only some decades ago due to its computational and data demands. And yet, it is quite remarkable that its outcome aligns clearly with reflections by Marshall (1890), (ch. XI) where he states: *The characteristic of manufacturing industries which makes them offer generally the best illustrations of the advantages of production on a large scale, is their power of choosing freely the locality in which they will do their work. They are thus contrasted on the one hand with agriculture and other extractive industries (mining, quarrying, fishing, etc.), the geographical distribution of which is determined by nature; and on the other hand with industries that make or repair things to suit the special needs of individual consumers, from whom they cannot be far removed, at all events without great loss.*.

In our interpretation this a sign of the validity of Marshall's analyses, as much as a suggestion that correlation structures computed from areal data are a relevant objective tool of analysis in Economic Geography. In this paper we have explored part of the technical context surrounding the computation of correlation structures, with the hope that future studies can safely and robustly use them to approach a variety of interesting questions.

References

Alig, R. J., Plantinga, A. J., Ahn, S., & Kline, J. D. (2003). *Land use changes involving forestry in the united states: 1952 to 1997, with projections to 2050.* (tech. rep.). U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Alstott, J., Triulzi, G., Yan, B., & Luo, J. (2017). Mapping technology space by normalizing patent networks. *Scientometrics*, *110*(1), 443–479.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage1. *The Manchester School*, *33*(2), 99–123.

Balland, P.-A., Boschma, R., & Rigby, D. (2015). The technological resilience of US cities. *Cambridge Journal of Regions, Economy and Society*, *8*(2), 167–184.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, *4*(3), 248–254.

Beaudry, C., & Schiffauerova, A. (2009). Who's right, marshall or jacobs? the localization versus urbanization debate. *Research Policy*, *38*(2), 318–337.

Berman, B. (2010). *Retail management : A strategic approach*. Upper Saddle River, N.J, Prentice Hall.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, *104*(17), 7301–7306.

Boschma, R., Balland, P.-A., & Kogler, D. F. (2014). Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, *24*(1), 223–250.

Boschma, R., Minondo, A., & Navarro, M. (2012). The emergence of new industries at the regional level in spain: A proximity approach based on product relatedness. *Economic Geography*, *89*(1), 29–51.

Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, *32*(1), 69–87.

Coscia, Hausmann, & Hidalgo. (2013). The Structure and Dynamics of International Development Assistance. *Journal of Globalization and Development*, *3*(2), 1–42.

Dark, S. J., & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography: Earth and Environment*, *31*(5), 471–479.

de Groot, H. L., Poot, J., & Smit, M. J. (2016). Which agglomeration externalities matter most and why? *Journal of Economic Surveys*, *30*(4), 756–782.

Delgado, M., Porter, M. E., & Stern, S. (2015). Defining clusters of related industries. *Journal of Economic Geography*, *16*(1), 1–38.

Diodato, D., Neffke, F., & O'Clery, N. (2018). Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, *106*, 1–26.

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

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

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

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

Engelsman, E., & van Raan, A. (1994). A patent-based cartography of technology. *Research Policy*, *23*(1), 1–26.

Farinha, T., Balland, P.-A., Morrison, A., & Boschma, R. (2019). What drives the geography of jobs in the US? unpacking relatedness. *Industry and Innovation*, *26*(9), 988–1022.

Fujita, M., Krugman, P., & Venables, A. (1999). *The spatial economy : Cities, regions and international trade*. Cambridge, Mass, MIT Press.

Gomez-Lievano, A., Youn, H., & Bettencourt, L. M. A. (2012). The statistics of urban scaling and their connection to zipf's law. *PLOS ONE*, *7*(7), 1–11.

Hausmann, R., & Klinger, B. (2007). *The structure of the product space and the evolution of comparative advantage* (tech. rep.). Cambridge, Mass., Center for International Development, Harvard University.

Hausmann, R., & Neffke, F. (2016). The workforce of pioneer plants. *SSRN Electronic Journal*.

Hennerdal, P., & Nielsen, M. M. (2017). A multiscalar approach for identifying clusters and segregation patterns that avoids the modifiable areal unit problem. *Annals of the American Association of Geographers*, *107*(3), 555–574.

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., & Zhu, S. (2018). The principle of relatedness, In *Unifying themes in complex systems IX*. Springer International Publishing.

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, *317*(5837), 482–487.

Jacobs, J. (1970). *The economy of cities*. New York, Vintage Books.

Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value. *American Economic Review*, *76*(5), 984–1001.

MacMahon, M., & Garlaschelli, D. (2015). Community detection for correlation matrices. *Physical Review X*, *5*(2).

Marshall, A. (1890). *The principles of economics*. McMaster University Archive for the History of Economic Thought.

McCann, B. T., & Folta, T. B. (2008). Location matters: Where we have been and where we might go in agglomeration research. *Journal of Management*, *34*(3), 532–565.

Menon, C. (2009). The bright side of maup: Defining new measures of industrial agglomeration*. *Papers in Regional Science*, *91*(1), 3–28.

Nedelkoska, L., Diodato, D., & Neffke, F. (2018). *Is Our Human Capital General Enough to Withstand the Current Wave of Technological Change?* (CID Working Papers 93a). Center for International Development at Harvard University.

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic Geography*, *87*(3), 237–265.

Petralia, S., Balland, P.-A., & Morrison, A. (2017). Climbing the ladder of technological development. *Research Policy*, *46*(5), 956–969.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (1999). Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, *83*(7), 1471–1474.

Porter, M. A., Mucha, P. J., Newman, M. E. J., & Warmbrand, C. M. (2005). A network analysis of committees in the u.s. house of representatives. *Proceedings of the National Academy of Sciences*, *102*(20), 7057–7062.

Porter, M. (1980). *Competitive strategy : Techniques for analyzing industries and competitors*. New York, Free Press.

Porter, M. (2003). The economic performance of regions. *Regional Studies*, *37*(6-7), 549–578.

Puga, D. (2010). The magnitude and causes of agglomeration economies. *Journal of Regional Science*, *50*(1), 203–219.

Runyan, R., & Droge, C. (2008). A categorization of small retailer research streams: What does it portend for future research? *Journal of Retailing*, *84*(1), 77–94.

Santoalha, A., & Boschma, R. (2020). Diversifying in green technologies in european regions: Does political support matter? *Regional Studies*, 1–14.

Scholl, T., & Brenner, T. (2016). Detecting spatial clustering using a firm-level cluster index. *Regional Studies*, *50*(6), 1054–1068.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific Reports*, *2*(1).

Teece, D. J., Rumelt, R., Dosi, G., & Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, *23*(1), 1–30.

van Dam, A., Gomez-Lievano, A., Neffke, F., & Frenken, K. (2020). *An information-theoretic approach to the analysis of location and co-location patterns* (Papers in Evolutionary Economic Geography (PEEG) No. 2036). Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography.

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651.

Wang, J., & Yang, H. (2009). Complex network-based analysis of air temperature data in china, *23*, 1781–1789.

Wang, Y. R., & Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, *362*, 53–61.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).