Scuola Superiore Sant'Anna
di Studi Universitari e di Perfezionamento

Essays in the Economics of Aggregation
in Empirical Contexts

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics

Candidate:
Matias N. Iglesias

Supervisor:
Prof. Federico Tamagni

International Doctoral Program in Economics
Institute of Economics
Scuola Superiore Sant'Anna
Pisa, Italy

This research was possible thanks to key people along the way.

Specially, thanks to Johannes Fink, Andrea Roventini, Frank Neffke and Federico Tamagni.

And thanks to the support of many more of you.

Contents

Chapter 1

Introduction

This Thesis is the result of continued and careful inquiry on the foundations of certain accepted results in subfields of Economics.

My past experiences with experiments in Physics shaped how I attempt to understand an a-priori unknown system. Mostly, I learned about complex mechanisms by complementing theory with computer simulations and conducting careful observations to be confident about each small assertion. In a laboratory, a postulate needs empirical confirmation or it is not real, and possibly not worth developing it further. I was alarmed to see that this intimate connection between observation and theory was much more blurred in Economics. This is understandable partly because Economics involves thought and reflection on human and social phenomena. And it is also true that, especially during XXth century, the detailed observations and technical means to process such information were largely unavailable. In my opinion this constraint explains the a weak incorporation of observation into the theoretical path of the discipline. It started to recede once detailed evidence in digital formats became available and the use of computers became widespread, thereby allowing a slow recovery from this dissociation of observation and theory that is ongoing.

In this context, I believe much of the improvement to the field can come from spending time to carefully revisiting the initial stages of the processing of observations. Weak steps there can spoil any results.

Works that dedicate to study observations, as opposed to focusing on results are not unknown to Economics. The thesis of R. Gibrat (1931) is an example of a collection of multiple empirical observations accompanied by models and analytical developments to formalize them. He was precisely an engineer, and after these four years I come to perceive the habits of engineering as a north that at this moment can strengthen Economics as a science. More concretely: actual magnitudes need to take a more central role. Observations of numbers of people, amounts of value, time periods, geographical distances or areas need to be actively combined and contrasted among themselves. The range and ticks on horizontal and vertical axes of plots are not peripheral information, they are telling key truths about the values that are observed. We need to be at every step aware of observed values and seek to be correct in our interpretation of the problem to guide ourselves straight to the place where there is the needle in a haystack.

Taking formal aspects as a priority when approaching problems may come at the cost of devoting less attention to the discussion of specific results put forward in economic studies. The contribution here is in many senses technical, although it seeks to mention the connections to specific economic concepts when possible.

The Thesis comprises four chapters, reporting papers I have produced or contributed to.

Chapter 2 offers tools to understanding how the aggregation of nonlinear micro agents all the way up to a national aggregate comes about. It deals with log distributed and log fluctuating agents, typical of any economy. It shows how entry and exit (an essential feature of economic agents) should be accounted.

In Chapter 3 I test the grounds for unification of a variety of similarity (correlation) measures in settings where a total quantity is disaggregated by subunits of the type 'economic activity class' and 'geographical areas'. It works out the formalities of connecting similarity indices computed from administrative areas, to approaches based on continuous geographical space. It shows techniques to process correlation matrices. Good quality data from 2002-07 and robust techniques show the correlation structure of US employment clearly points to phenomena proposed by Marshall (1890) (Ch. XI).

Two other papers complete the Thesis. Next I introduce each of these four pieces in more detail.

The Chapter 2 tracks the aggregation of fluctuations of a population of economic agents. If the aggregate is exactly the sum of all micro contributions then a single formal framework valid for micro and macro may be possible. I show that indeed this is the case although there are formal obstacles to deal with. To constrain and guide the analytical developments I use information of French exports and imports at the firm level (Customs office).

The strategy I offer for dealing with the aggregation of micro shocks consists of separating the task in two stages: from firms into sectors (first stage), and from these sectors into an aggregate (second stage). The reason for this is that firm level fluctuations are highly non linear, so that for aggregating them we need a special technique (sum of powers) akin to considering $\int p(t)10^{kt}$, where $p(t)$ is the distribution of micro fluctuations ($k = 1, 2, ...$ is used for moments of $k - th$ order). Instead, once enough firms are grouped, the fluctuations showed by this sector are milder and allow linear combination of fluctuations. This second stage even if simpler, is not covered carefully in currently available references. I offer an approach to rigorously track the aggregation from micro to macro and possibly address a variety of specific questions in future studies.

Following this aggregation carefully in light of the empirical constrains already lets us revise some of the accepted conceptions regarding aggregate volatility. A relation between the idiosyncratic part of aggregate variance and parameters of the size distribution proposed in Gabaix, 2011 is seen to not hold empirically and happens to involve some formal gaps that need revision. I show that the increased volatility due to concentration stems from the basic properties of aggregate variance and does not require an economic argument (eg. Hulten's theorem and Domar weights). In addition, I show that the milder decay of variance with population size can be explained from comovement among firms if we account for micro fluctuations carefully, as is unrelated to the parameters of the size distribution.

The paper is treats every open aspect with the necessary care. I make a compromise between offering a robust formal treatment of the problem and leaving expositions simple when possible. The problem of aggregating micro fluctuations involves some complicated formal steps but the insights we can derive are worth the effort. In essence we are relating micro to macro accounts with precision.

After having studied aspects that matter to disaggregation under a single criteria, in Chapters 3 and 4 I focus on few special cases which involve a *double* disaggregation.

In Chapter 3, the empirical setting consists of counts of total employment (or number of establishments) disaggregated by industrial category and geographical units of the contiguous United States (US). Here we introduce so called *similarity* measures. First we study aspects that matter for the robustness of results, such as whether different similarity measures can be in practice equivalent, and how can continuous spatial accounts enter the picture. Then we compute and discuss the correlation structure derived from such cross sectional data of the US and thereby we show this technique can be useful to detect spatial patterns that matter to economic geographers. Next I describe these steps in greater detail.

We begin by testing the grounds for unification of a variety of tecniques used in the literature to estimate cooccurrence. More precisely, we consider raw data, log transformed levels, and binarized location quotients as possible pre processing steps. In addition, we consider similarity measures including simple joint coocurrence ($X^T X$) Pearson correlation, cosine similarity, covariance, and proximity as in Hidalgo et al. (2007). We see that, at least when applied on employment levels by 4 digit industry and US county, all combinations of this data processing and similarity measure lead to ranks of industry pairs (less similar to most similar) that are not plainly contradictory among themselves. This delimits the room for unifying the methods applied by different studies, with the aim of higher cross study comparability.

Secondly, given the side categories are geographical units we seek to formalize the connection between accounts in continuous space and the above mentioned measures of cooccurrence computed on cross sections by activity and administrative area. We determine how exactly cosine similarity can be an indicator of actual coexistence of locations in continuous space. We approach this analytically and confirm the validity, caveats and details of this connection through computer simulations.

After having contextualized a family of measures of coexistence and showing details of its interpretations in a continuous space framework, and showing how space would enter the picture formally, we look at the outcome of applying these measures on the counties of US. We argue that the correlation structures carry key information regarding the spatial distribution of activities, and that they are objective tools that can be applied in for studying spatial patterns in a general setting. When applied in the US we infer a network of economic activities, and the neighborhoods in this network are linked to types of spatial patterns. These latter can be organized in four themes: population, large cities, natural resources, and manufacturing, pointing directly and clearly to the phenomena discussed in Marshall (1890) and multiple subsequent works.

We are thus offering tools for understanding correlations in spatial cross sections and showing that they allow formal and relevant analysis for spatial phenomena.

Next, Chapter 4 is dedicated to the study of a specific type of measure called *Location Quotient* (LQ). Still in the context of double disaggregation (eg exports by country and product, number of patents by city and technology field) the location quotient is defined as the ratio between the observed probabilities (values/total) and the expectation from multiplying the marginal probabilities the pair of categories.

The location quotient has been adopted widely, but it shows effects that might be undesired.

Among them, an effect that has been partially acknowledged but never truly measured is a *size distortion* by which the scale of LQ tends to be effectively wider for observations from smaller entities of the datasets.

I propose to use the probability of an observation crossing the LQ = 1 threshold within a time period, conditional on the values of size parameters, as an a posteriori estimator of the effective scale of the LQ metric. In this way we can sense effective distortions of the scales of LQ. Apart from a direct estimation of this patterns, I offer an analytical model that explains the qualitative result. They depend crucially on the fact that larger entities are less volatile (H. R. Stanley et al., 1996).

Drawing upon these insights, future empirical works can have ideas as to how to control size effects of LQ indices, allowing more robust and consistent results.

The paper closing this thesis, presented in 5 emerged from a collaboration during a visiting research period at Harvard Kennedy School, where I worked with Prof. Neffke and colleagues. It sets the focus on the methodological choices used in studies of economic diversification.

In this paper the dialogue with strands of research in economic geography is more clearly present. Hypothesis regarding diversification involving concepts of related variety (Frenken et al., 2007), complexity, and product spaces (Hidalgo et al., 2007) may highlight the role of different mechanisms. For example, depending on the reference, diversity of related activities would lead to growth by a process of Schumpeterian learning (Frenken et al., 2007), while a coexistence of unrelated activities in a region would be a signal of wider sets of underlying capabilities and then a better economic performance (Hidalgo et al., 2007).

Formalize measures denoting diversity can help comparing these two approaches formally. Then, in this paper we introduce and discuss formal measures of diversity borrowed from an interdiscipline centered in ecological studies. One can quantify three main meanings of diversity (variety, balance and disparity), each of which has a formal definition that would allow us to quantify.

When it comes to the studies of complexity we propose to take a closer look at what the method of reflections as introduced by Hidalgo et al. (2007) implies in general. When applied on cross sections of exports by country and product it can lead to a ranking of countries by economic complexity index (ECI) that has highly developed, diversified countries on one end and less developed one on the other. ECI has then been interpreted as a measure of diversity. But does this method lead to a diversity measure in general? The answer is that the method of reflections is analogous to a spectral clustering algorithm that will arrange structurally dissimilar nodes in two opposite (similar within) ends, as best as it can. Then whether ECI can be a measure of diversity will depend on what cross section of the data we are applying it on. We see that it manages to rank US cities by income based on their industry information, but it fails to reach the same outcome in a cross section of industries by occupations. This latter disaggregation leads to thematic correlation structures and so a two pole clustering algorithm as ECI is likely to not work.

Same as the papers in previous chapters, this last one makes an effort to highlight the importance of revisiting methods before interpreting results further and offers formal tools for more careful quantification. It also features *product spaces*, which are essentially correlation structures like those covered in Chapter 3. But it finds that the characteristics of the entities into which the total is disaggregated are decisive for what are the outcomes. This paper however does a more committed address of the contemporaneous discussions involving the connection of economic growth and diversification.

Chapter 2

Aggregate accounts from populations of micro agents

Abstract

In this paper, we offer a technique that helps one trace the aggregation of firm level exports and imports into national aggregates.

On the one hand we clarify the linear relations that link sectoral fluctuations to aggregate fluctuations of an economy. Then, we study the aggregation of groups of fluctuating agents within a parts of the aggregate. In this way a precise reconstruction of the aggregation of fluctuations from the micro to the macro level is possible.

We use firm level export and import data from France Customs over the period 1997-2013 to constrain the theoretical derivations to parameter regions that matter in practice. In this way a problem that is generally difficult can be simplified. Empirically constrained computational tests let us know that each analytical expressions in this paper is true, either exactly or approximately.

I show that there is a 'postponement' of the law of large numbers by which idiosyncratic variance decays more slowly than $1/N$ *with the population size* N *caused by large, multiplicative micro fluctuations and can be accounted as a comovement among agents. I clarify that this is unrelated to another situation by which concentration allows groups of few large firms drive the aggregate towards their level of volatility.*

2.1 Introduction

Understanding the connection between the dynamics of a large number of economic agents and the aggregate characteristics observed on such a population is a clear and recurrent motivation of economic studies. Even if it is not a new problem, technical possibilities available to us today provide approaches that were out of reach even in recent times. So that new insights to this old question may definitely still come about.

In current days there is an increasing wealth of detailed evidence in digital formats and an ever growing adoption of computers in research. Especially, I am referring to the use of computers for heavy and sophisticated data processing and in some cases as programmable calculators. In this way, long standing technical limitations everywhere to be seen in older papers have faded. But not all questions have been reformulated in light of the new technical possibilities.

In the case that the aggregate variable (X) is plainly the sum of a quantity (s_i) observed in each agent of the population, the relation $X = \sum_i s_i$ offers a formal condition that can unequivocally guide the study of the problem. An aggregate variable X is characterized by its mean level $E[X]$ and the width of deviations from it $\mathrm{std}[X] = \sqrt{\mathrm{var}[X]}$, where E, std and var are *expected value*, *standard deviation* and *variance* operators respectively. I say *aggregate volatility* to refer to $\mathrm{var}[X]$. Are there insightful expressions of $var[X]$ in terms of characteristics of the population of agents? This is what we aim to study in this paper, both by reviewing some established ideas, and by completing undeveloped links in the aggregation of micro fluctuations.[1]

I said computers and detailed evidence mean a substantial opening of paths to study complicated economic settings. More concretely, what possibilities do we have now that did not exist before? First of all we can restrict theoretical developments to empirically relevant settings. A general problem may be too wide and demanding a variety of mathematical frameworks, but the data can let us simplify the task by directing us to understand the specific mechanisms that play a relevant role in practice (in our case, characteristics of size distributions of agents and their fluctuations determine much of what issues need to be studied). Parameters of an empirical system (in our case, population sizes, moments of micro fluctuations) can also constrain and guide the analysis and give meaning to the answers we find along the way.[2] A consequence of the new powerful technical means is that, as much as we avoid empirically irrelevant problems, we are also more confident to go through certain slightly uncomfortable formal paths if we see that we do need them to understand an empirical situation (see section 2.8). Another novelty is that we can *know* if the equations we arrive at are true or not. The importance of this single item cannot be underestimated. I can know that even the more complex equations in this paper are valid in the context of the problem I study with a confidence I could simply not count on if I had not tested them on a computer. A final possibility for

[1] The words 'agent' and 'entity' or 'firm' are used indistinctly, being them the atomistic agents of international trade.

[2] The word 'size' refers generically to the value of the variable which is basis of aggregation. Because we use firm level international trade as empirical benchmark, the size of an agent for us is the value an agent imported or exported in a time period.

hypothesis testing that I find quite helpful, especially when counting on rich micro data is that of using random samples with replacement, or *bootstrapping*. With this technique we can (among other exercises) measure macro results of changing micro characteristics of a system and explain these outcomes in light of theoretical expectations.

Our empirical benchmark are all French firms importing and exporting over the period 1997 - 2014. Even if this work is at every step tightly attached to the real dynamics of this reference population of economic agents, it does not follow the most popular approaches to understanding volatility aggregation found in the literature. Let us explain the main reasons for these departures.

Neoclassical models in the context of business cycle theory (Kydland & Prescott, 1982; Long & Plosser, 1983) have influenced the approaches to studying aggregate volatility in economics. Even up to this date it is not surprising to see a paper studying macro volatility begin by positing a system of optimizing agents. I do not recur to them because aggregate volatility does not forcefully need them. The benefit of leaving models aside for a moment is that the insights we derive are not dependent on particular choices. In addition, models may involve variables not easily observable, preventing solid empirical testing. Another usual approach to understanding volatility is by looking for factors that correlate with increased volatility (Stockman (1988) as an early example). Here however I seek to trace micro fluctuations from the bottom up and in this way understanding how micro characteristics determine the observed aggregate variance, as opposed to explaining variance by a factor. Finally, the decision of agents is often at the center of theoretical approaches in economics (Lucas (1977), as an example among countless others in a long tradition). If our goal is to account for fluctuations however, we can abstract from the subjective point of view of an agent. In fact, it is highly advisable to separate the problem in two complementary tasks. On the one hand, knowing how agent levels are aggregated and what non trivial mechanisms play a role there. This is a clean formal task and it is what I undertake in this work. On the other hand, understanding why and how the micro observations came to be what they are. This is out of the scope of this paper, and it involves the study of a wide range of particular economic situations resulting in the micro fluctuations observed. The former task does not need the latter.

When it comes to the aggregate volatility shown by a population of agents, the agents' size distribution as well as the distribution of their fluctuations are the two single features that decide the formal frameworks that will be needed. The size distribution of economic agents in a variety of contexts has been long accepted to be possibly log-normal or power law (Pareto) (Axtell, 2001; Gibrat, 1931; Hart & Prais, 1956). The distributon of fluctuations, although already suggested log distributed by Gibrat, became more clearly determined in more recent studies as H. R. Stanley et al. (1996) and related papers, as multiplicative (thereby non linear) and of large magnitude.

A population of thousands of non linearly fluctuating agents is not a simple system to understand fully and presents a variety of non intuitive mechanisms. For example, the concentration due to Pareto or log-normal size distributions, means that the few largest agents drive aggregate volatility towards their level of volatility. This feature is intuitive and easy to explain, but gained consensus only in the last decade.

A paper that focuses in the issue of aggregate variance and became the main reference in current studies is Gabaix (2011). This contribution, even if it was well received in the community has established a series of misconceptions regarding volatility aggregation that need to be urgently clarified. It claims a direct relation between size distribution parameters and the rate of decay of idiosyncratic variance. This relation has not been observed empirically, and this is expectable because the formal steps leading to this result do not hold. This is mostly because of ignoring that agents show large multiplicative fluctuations so that aggregate volatility cannot be expressed as a linear combination of agents volatility.[3]

In this paper instead I follow the aggregation with the care it requires, and tap into a variety of situations that interact among themselves to result in the levels of aggregate volatility that we observe in large populations of fluctuating agents.

I clarify that the basic properties of aggregate variance and the shape of the logarithmic curve implies value weighted contributions of parts to aggregate variance. Concentration due to log-normal or Pareto size distributions therefore allows groups of few large agents to drive the idiosyncratic part of aggregate volatility.

As a separate phenomenon, there is a departure from the law of large numbers (LLN) in the

[3] Additionally it assumes a proportionality between Herfindahl index and idiosyncratic volatility that does not hold in general.

decay of volatility with number of agents. It is not hard to see clearly this milder decay of variance by the LLN is accounted by highly non linear firm level shocks resulting in comovements among agents. The idea that the power law of size distributions explains the postponement of the law of large number as installed by Gabaix (2011) needs revision in light of its substantial formal gaps.

The next sections are organized as follow. In section 2.2 we review related strands of literature. In section 2.3 we introduce the data and methods. In section 2.4 we explore the size distribution and the distribution of growth rates in our data. Section 2.5 contains the generally useful mathematical definitions and properties. Section 2.6 discusses why acknowledging non linearities of firm level data is necessary to avoid incorrect outcomes. Section 2.7 does a concise formal review of the diversification argument as in Lucas (1977) and the contribution of Gabaix (2011). It then clarifies the framework for studying the decay of variance with population size. Once the sectoral to aggregate (linear) relations are clarified, I proceed to studying groups of agents with the goal of completing the non linear part of the aggregation (Section 2.8). The core contributions of this paper are explored formally in this section. Essentially, if we understand the variance of groups of agents, we can then aggregate them simply by linear equations to arrive at aggregate volatility, thereby having a connection between the micro parameters and the macro volatility observed.

Section 2.9 controls for results robustness when changing agents' size distribution. Section 2.10 shows how extensive margins can be accounted. It shows and discusses estimations of cross covariance elements. Section 2.11 is a summary of how all the mechanisms we found combine in our empirical benchmark system to let the aggregate show its observed variance.

The developments in Appendix are important, although out of the main body for brevity. Here there are estimations of uncertainty introduced by off diagonal covariances, clarification of accounts in the frequency domain, derivation of moments of log-normal and log-Laplace distributions, and most importantly, introduction of the codes that define the tests and estimation procedures used in the paper.

2.2 Related Works

There have been different approaches to the goal of explaining aggregate volatility, we review them in this section.

Many of the papers in economics studying aggregate volatility estimate the relevance of certain variables by estimating regressions with a variety of specifications (Canals et al., 2007; Castro et al., 2015; Foerster et al., 2011; Giovanni & Levchenko, 2014; Koren & Tenreyro, 2007; Stockman, 1988). Even if this can be useful for estimating the importance of certain factors, in this work we seek actual accounting of fluctuations based on the condition $X = \sum_i S_i$ constraining how micro and macro fluctuations relate. This is a different approach from using regressions to find 'variables that can help explain' aggregate volatility. Some of the mathematical framework in this paper can still be applied in exercises involving factor models.

The business cycle tradition in economics lent itself naturally to studying variance which in fact comes directly from the amplitude of wave components (Eq. 2.65 in Appendix). Impactful contributions proposing one time step specifications in real bussiness cycle logic (Kydland & Prescott, 1982; Long & Plosser, 1983) popularized the study of volatility under the framework of neo classical equilibrium models. Recent papers studying volatility in economics continue to start by asking new classical macroeconomic models (Carvalho & Grassi, 2019; Giovanni & Levchenko, 2014).

I do not adopt economic models in this paper. The reason is that problems strictly related to aggregate volatility are independent of characteristics of models. In fact an economic model can be an unnecessary blur complicating our understanding of inherent mathematical situations (which are already non trivial). Not to mention that key variables of models may not be well observable or unequivocally defined, resulting in fatal gaps between theory and empirics.[4] [5] This paper is long and detailed, and we only dedicate to study open problems present when aggregating agent sales, which fulfill $X = \sum_i S_i$ exactly. Many of our results extend to the case where we aggregate, e.g. production, although they precede the introduction of an

[4] By involving production models many papers deal with total factor productivity (TFP). However it is known that measurement of TFP is problematic and goes in hand with assumptions on labour, capital and possibly other unprecised factors (cf for example Felipe and Fisher (2003) for discussion).

[5] Some of our expressions for aggregate volatility are of course linked closely to analytical developments in the context of equilibrium models, the closest of such works is possibly D. Baqaee and Farhi (2018).

’aggregation of production’. [6]

In models following Kydland and Prescott (1982) and Long and Plosser (1983) the time step perspective is close to autoregresive model estimations (cf Bollerslev et al. (1994), esp. S4). They lend themselves to decompositions into frequency domain (as in Dupor (1999) and Horvath (1998)), and they are akin to the modern study of higher frequency financial time series (Jacod & Protter, 2012). A subtle but important departure from this paper with respect to this tradition is in the description of fluctuations by deviation from a mean, as opposed to one step time differences. This is the best choice when it comes to aggregation of micro fluctuations and it should be adopted for easier, more solid results. Even if an autorregresive framework is useful for studying other type of questions related to agents’ dynamics.

If we seek to relate time series of national aggregate sales to the dynamics of a population of agents there are two elements that are basic and indispensable: the levels of sales of the agents, and their evolution over time.

When it comes to empirical studies, evidence on size distribution of firms has shown that it is usually stable over time, possibly adapting to a lognormal (cf. for example Hart and Prais (1956)) or a power law, Pareto distribution (Axtell, 2001).

The evolution of firm sales has been subject of wider debate. Minimal quality evidence for sofisticated hypotesis testing did not appear at least until the 1950’s and high quality data not even available in the 2000’s. Hypotheses of Brownian drift of firm logarithmic levels of sales (wrongly attributed to Gibrat 1931) have been rejected. There is a significant autocorrelation pattern that leads to stability (Boeri, 1989; Chesher, 1979). That is, the diffusion expected from an iid sequence of shocks is not observed and indeed there is usually a negative autocorrelation between $\log(S_t)$ and $\log(S_{t-1})$. For the purposes in this paper, the precise shape of growth rate distributions and their self dependence is not crucial. The important fact is that their distribution can be naturally given as a distribution of log values in line with known results on domestic sales in countries other than France reported previously in Amaral et al. (1997), Bottazzi and Secchi (2006), and H. R. Stanley et al. (1996) among others.

The most debated issues stemming from variance aggregation have to do with certain non

[6] We will use the term ’sales’ to refer to the value of transactions between firms (or agents if we include individuals) be it exports (French sellers to foreign buyers) or imports (French buyers to foreign sellers). We are accounting flows accumulated in time periods.

intuitive aggregate results of the micro dynamics. The most clear example is the observation that the standard deviation observed in the total fluctuations from a population of N normally fluctuating contributions should be of the order of $1/\sqrt{N}$. And nevertheless national or world aggregates are known to fluctuate far more than this (the number of agents N would easily be of thousands or more).

The most frequent reference when it comes to answering this naive diversification intuition is the 'granularity' paper of Gabaix (2011) (from here on denoted Gb11). The typical mention of this paper is on the lines of the following: *"Gb11 used Hulten's theorem to argue that the existence of very large, or in his language granular, firms can be a possible source of aggregate volatility. If there exist very large firms, then shocks to those firms will not cancel out with shocks to much smaller firms, resulting in aggregate fluctuations."* (D. R. Baqaee & Farhi, 2019).

A digression to mention with respect to this quote are that if looking at the aggregation of sales and not of production, Hulten's theorem is not necessary and value weight in aggregate idiosyncratic volatility (which is the key here) stems from how the log curve and variance are (see Section 2.7).[7]

The issue that Gb11 is tapping into are the combination of concentration and the value weighted contribution to idiosyncratic variance. If few largest firms hold a significant part of the total, it is reasonable that shocks to them are an equally proportional part of idiosyncratic volatility. Then it makes sense to regress aggregate volatility with shocks of largest 100 firms as explanatory variable, and in some cases Gb11 is cited in relation to this specific exercise.

The paper of Gabaix (2011) however has a problem which is fatal when working with aggregate variance. The hypothesis of uncorrelated cross covariance among agents cannot be applied if we are describing a population of economic agents because their micro fluctuations are relatively large leading necessarily to comovements which are at the heart of what we seek to study.[8]

[7] As a smaller digression, it is not about shocks of large firms being cancelled by other shocks, but it is about shocks to large firms dwarfing those to small firms and so contributing more to idiosyncratic aggregate volatility.

[8] There has been another example where Farmer and Lillo (2004) failed to confirm the results of deductions of (Gabaix et al., 2003) because of correlation among agents existent in empirical settings has been overlooked. To be fair, the dismissal of cross correlations is frequent and may be a problem in other papers. Acemoglu et al. (2012) for example also throw the baby with the bathwater when they drop cross correlations (they look at the trace of the cross covariance matrix, first equation of page 1988) while working with network propagation of shocks, a problem where cross correlations are key.

In the beginning of his paper Gabaix (2011) implicitly assumes fluctuations to individual agents are small, linear and uncorrelated among themselves, which is not true for (and not reconcilable with-) empirical systems. The shocks $\sigma_i \epsilon_{it}$ do not follow the assumptions that are asked of them. These issues are already a big warning sign. And the paper continues with a step that is not valid in general and needs special care, which is taking the firm level width of deviations out of the sum where its multiplied by firms' value share. In this way we arrive at an automatic connection between the Herfindahl index (h) of firm value shares and idiosyncratic volatility (σ_ϵ^2), and relation $h \sim N^{-1+1/\zeta}$ deduced for h is taken to apply to σ_ϵ^2.[9]

Gabaix then postulates a relation between volatility and population size, which would then forcefully be an answer to the naive diversification debate. Indeed it has been taken as such, a few quotes are included in footnote. [10] These papers are just a small sample of the widespread confusion related to understanding micro effects on aggregate volatility.

In this paper I followed the aggregation of agents' fluctuations carefully from the micro up to the aggregate level and I am able to show that the departure from the naive $1/\sqrt{N}$ diversification rule has nothing to do with the agents' size distribution, contrary to the view installed by Gabaix (2011). Understanding the dependence of aggregate idiosyncratic volatility with population size is thus an issue still open. I show the departure from $\sigma_\epsilon \sim 1/\sqrt{N}$ is a consequence of slower convergence of averages because micro fluctuations are multiplicative. This is potentiated by fat tails in agents' fluctuations (Bottazzi & Secchi, 2006; H. R. Stanley et al., 1996). The reader is encouraged to not believe this and see it for themselves with the help of this paper.

It seems to me that before Gabaix (2011) there was a consensus on the intuition that con-

[9] here $h^2 = \sum_i (S_i/X)^2$ and ζ is the power law of the Pareto size distribution, such that the probability density function of firm sizes is $p(x) = \zeta x_m^\zeta / x^{\zeta+1}$ and so its counter cumulative density function is $c(x) = (x_m/x)^\zeta$

[10] Some of the quotes of Gabaix, 2011 as replying to the diversification argument are:

"Gb11 demonstrates that aggregate fluctuation decays much more slowly in an economy with a fat-tailed firm-size distribution, contradicting the diversification argument put forward by Lucas." (Nguyen et al., 2020)

"Gb11 shows how aggregate fluctuations can be generated by firm-specific shocks in an economy with a heavy-tailed distribution of firm size." (Kogan & Papanikolaou, 2012)

"Gb11 and Acemoglu et al. (2012) derive conditions under which these heterogeneties can generate aggregate fluctuations from idiosyncratic or sectoral real shocks invalidating the diversification argument of Lucas (1977). [...]. Gb11 and Acemoglu et al. (2012) show theoretically the network structure and the firm size distribution are potentially important propagation mechanisms for aggregate fluctuations originating from firm and industry shocks."(Pasten et al., 2019)

"Gb11 demonstrates that aggregate fluctuation decays much more slowly in an economy with a fat-tailed firm-size distribution, contradicting the diversification argument put forward by Lucas." (Guiso et al., 2016)

centration together with value share weights means that few agents can drive the aggregate time series regardless of how many agents there are in total. And the community found in Gb11 a paper that could symbolise this idea.

The answer of Gb11 to the diversification debate however has established important misconceptions. Mostly, a link between size distribution and the decay of volatility with population size which does not exist in reality. An important role in this story however, was played by the research community. Of the papers who cite Gb11 the large majority focus on issues unrelated to the problem of understanding aggregate variance. No paper was dedicated to confirming, replicating, revising, refining or rejecting (in short, studying any further) the proposed dependency of aggregate idiosyncratic volatility with population size which is the core of what the paper claims (apart from being a clear relation to measure, and an important one).[11][12]

In this work thus, we will follow the aggregation of agents fluctuations from scratch and decide for ourselves what elements of previous studies are pointing to actual mechanisms and what other elements should be reconsidered.

2.3 Data and methods

2.3.1 Data

Aggregate figures of international trade are composed of the export and import transactions undertaken by a large population of firms (apart from purchases by consumers, not considered in this analysis). Our main source of evidence are the records of French customs. Datasets from this source have been used widely in recent studies in international trade. For documentation regarding this dataset see eg. Bergounhon et al. (2018). The full data set covers all transactions that involve a French exporter or importer. The data spans along the 1997-2013 period monthly, although once I have confirmed that most volatility is of annual or lower frequency (cf Appendix, section 2.13), I use the annual time series as reference. Once the variance accounting vs. frequency has been clarified, using annual time series has the benefit of implying

[11] 320 papers tagged in the Scopus database up to November 2020.

[12] Counter examples I may have missed are welcome.

a smaller computational burden. [13]

Between the years 1997 and 2013, a total of 114000 firms have reported sales of about 8970 products (CN 8 digits), to 442000 buyers in up to 234 destinations. Notwithstanding this large amount of diverse transactions, only 5233 firms account for 90% of the exported value, the same way that 1845 products (CN 8 digits, 333 CN 4 digit products), or 11869 buyers or 42 partner countries do. Indeed high concentration is a known characteristic of the international trade landscape. On the side of imports, 9159 firms account for 90% of the traded value, the same way that 2107 products (CN 8 digits, 334 CN 4 digit products), or 36 partner countries do. Detailed custom records information have become available only recently. There is therefore some necessary novelty in the outcomes of this study.

Results computed from exports data are largely equivalent to those computed for imports data. In some cases I show them both. If instead only results from one of these flow directions is shown, it can be assumed that the opposite flow shows essentially equivalent outcomes that would be redundant to display.

2.3.2 Methods

This paper is centered on determining the formal relations between micro characteristics of sales time series of a large population of agents and the moments of the time series of aggregate sales that result of them. This goal could potentially be achieved through analytical developments only, although this would be practically impossible if we did not use empirical evidence to constrain our formal path.

I control the validity of every formal step drawing empirical evidence from our reference system, the population of French traders. Empirical data lets us decide on the validity of key conditions, and from there on the validity of certain equations.

I also complement some of the formal developments with computational experiments. One of these, for example, involves drawing random vectors $\boldsymbol{x}$ and studying the average of $10^{\boldsymbol{x}}$, as a function of a variety of parameters. This problem is at the heart of what we seek to

[13] As further details of the data preprocessing: before 2010 transactions below 1000 EUR did not need to be reported. To avoid distortions from the removal of this rule in 2011, we drop all transaction below 1000 EUR along the full timespan. For extended technical review of the dataset see Bergounhon et al. (2018).

understand in section 2.8, but its general study is not quite simple. In this case, empirical conditions constrain the relevant ranges of parameter values and simplify the problem. Then, the computational tests offer answers that should be achievable formally.

A final way in which empirical evidence is exploited is by the study of multiple counterfactuals. We count with time series of sales at the firm level for tens of thousands of firms. Then, for example, we can sample an increasing number of firms from this total population and measure aggregate moments as a function of population size. Or, we can let all firms be fixed at their average level when they are active, and so we would have an account of changes due to entry and exit events. We could measure the dependence of aggregate volatility with the magnitude of firm level deviations from their mean levels, or we could test the consequence of changing the shape of the firm level fluctuations of the shape of the size distribution. Of course, randomized steps can lets every configuration be repeated multiple times thereby offering a measure of uncertainty on any of these results to be derived.

This exercises however are best exploited if there is a formal framework that guides the details of how each experiment is performed and where to look for the evidence on a specific mechanism that one is seeking to measure. That is why, all in all this paper reaches its results informed by the empirical benchmark of French traders, and complementing analytical developments with computational exercises.

The codes for reproducing computational results are in the dedicated GitHub repository. Pseudo codes for all experiments are in Appendix (section 2.15).

2.4 Key empirical features: distribution of firm sizes and fluctuations

Characteristics of the distribution of firm sizes and their fluctuations decide what type of problem we have in hands and what path we will have to follow to understand it. This is why we begin by reviewing these two fundamental characteristics of the population of fluctuation French traders. These size distribution and the distribution of logarithmic growth rates of populations of economic agents have been observed in contexts such as measuring firm sizes by employment level, or by total sales, including domestic sales, in different regions and over time. The key feature observed across multiple populations of economic agents is that both their sizes and their multiplicative growth rates are distributed nicely on a logarithmic scale. This property, which so far is universal and robust, determine much of the steps that we will have to follow in this paper.

2.4.1 Size distribution of firms

The size distribution of firms in the French traders dataset is plotted in figure 2.1 (left: Exporters, right: Importers). The distributions of multiple years is superimposed. Blue and yellow dots draw the distribution of number of agents by size bin, and traded value by size bin. The scale of the plots is log-log, log-normal size distributions thus appear as a quadratic parabola.

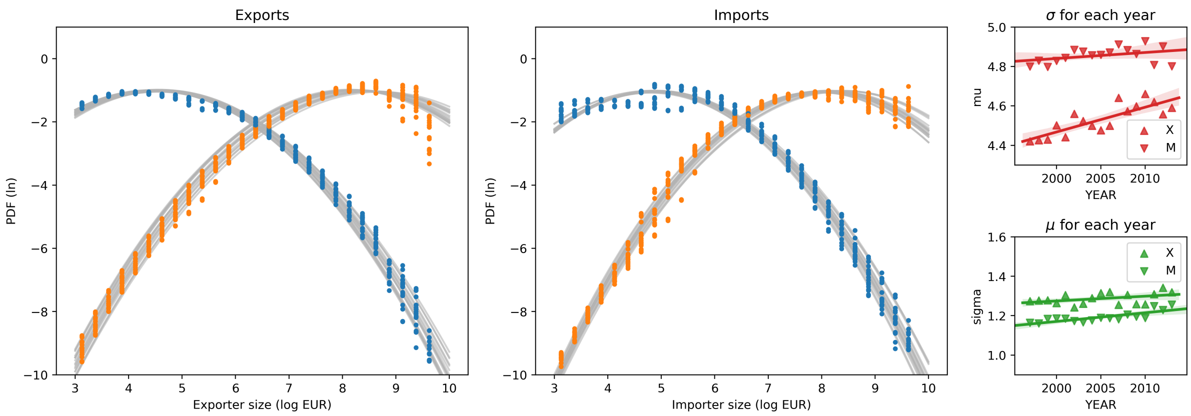


Figure 2.1: Distributions of agents sizes (blue) and value (yellow). Exports (left) and imports (right). Log log scale, parabolas stand for log-normal distributions. Data of years 1997 - 2013 superimposed. Right insets show the mean and standard deviation over time as implied by parabolas OLS fits. Population data is fitted and value data is deduced by computing a first moment.

It is evident that the size distribution of French traders is compatible with a log-normal distribution[14] and this is in line with previous evidence of populations of economic agents. [15]

[16]

[14] Truncated at the minimum value $x = 3$.

[15] This work does not dedicate to study the size distribution in detail, only to understand its main features.

[16] Strict interpretations of $p_{cnt}(s)$ as exactly log-normal forced debates (cf Prais 1973) asking for example whether a perfect lognormal can be a stationary state stemming from modeled micro shocks. However modern day evidence shows us that there is some drift and widening of the fitted lognormal parameters over the years (right inset in figure 2.1), so it is debatable that the size distribution stationary after all. Still this widening should be far larger if firms would do random log jumps as in a brownian motion, suggesting that negative autocorrelation of shocks plays a role as stabilizer of agents' time series.

A log-normal size distribution implies concentration of significant part of value among the few largest agents. Indeed, for French traders 90% of exports are concentrated by firms exporting more than $log(s) = x$, and 90% of imports are concentrated by firms importing more than $log(s) = x$.

If we restrict to these subsets of largest agents, a Pareto (power law) rule is also partly compatible with the observed size distribution. In this sense, a log-normal and a Pareto rule are not contradictory descriptions of the size distribution of firms. A Pareto model means a linear approximation to the upper tail (right end) of blue dots in plots of figure 2.1. The Pareto power law then can be a practically useful model for studying consequences of the shapes of size distributions, regardless of whether it is completely accurate empirically. [17].

Similar to a log-normal, a Pareto size distribution implies concentration. They both lead to the familiar near 80% - 20% concentration rule.

In the remainder of this subsection, let us formalize the description of size distributions a little further.

Consider a histogram telling how many agents one can find with total sales in each interval $[\bar{s}_b, \bar{s}_b + ds)$ of a partition of the real numbers. The total number of firms is the sum of population of all bins $N = \sum_{bins} n_j$. If ds is small enough, we can approximate the sizes of firms by $\bar{s}_b$. In such case the total value associated to the firms in the bin is $S_b = n_b \bar{s}_b$. The sum of total value disaggregated by firm size is:

$$X = \sum_{j=1}^{N} s_j = \sum_{bins} S_b \approx \sum_{bins} n_b \bar{s}_b \tag{2.1}$$

Let us denote firm sizes in linear scale as s and the normalized probability density function (PDF) of firm sizes as $p_{cnt}(s)$. Similarly, there is a PDF of value: $p_{val}(s)$. Equation 2.1 in the continuous limit is:

$$X = \int X \, p_{val}(s) \, ds = \int N \, p_{cnt}(s) \, s \, ds \tag{2.2}$$

Where value X and population N appear linked by the fact that the first moment of value of

[17] Broido and Clauset (2019) discuss a similar situation in applications to network theory whereby which scale free networks may be less frequent than log-normal degree distributed networks but are anyway interesting as benchmarks.

the population distribution is the zeroth moment of the value distribution. If one integrates this, note that $\bar{s} \equiv \mathrm{E}[s] = \int s\, p_{cnt}(s)\, ds$ are the average firm sales, and of course $X = N\bar{s}$. But also for each particular size bin: $S_b \approx n_b \bar{s}_b$.

The integrands in equation 2.2 tell us:

$$Xp_{val}(s) = Np_{cnt}(s)\ s \tag{2.3}$$

For a given value s, the amount $v_x = X\ .\ p_{val}(s) = N\ .\ p_{cnt}(s)\ .\ s$ is how much the holders of such sum hold altogether. For example if $n_2 = 10^3$ people hold 100 dollars (10^2) each, this group has $v_2 = 10^{3+2} = 10^5$ dollars. [18]

From equation 2.3 one can see the distribution of firm sizes $p_{cnt}(s)$ determines the distribution of value $p_{val}(s)$. For example, use the log variable t in place of the linear s, i.e. $s = 10^t = e^{\ln(10)t}$. As a function of the log levels t the size distribution is: $p_{cnt}(t) = \mathcal{N}(\mu, \sigma) \sim e^{-a(t-\mu)^2}$. The product of $p_{cnt}(t)$ and 10^t (as in eq. 2.3) means summing their exponents. And note that a parabola (from $\mathcal{N}(\mu, \sigma)$) plus a line (from t) lets the distribution of value $p_{val}(t)$ be a shifted parabola $\mathcal{N}(\mu + \ln(10)\sigma^2, \sigma)$ (see Appendix for details). This is why the distribution of value is log-normal if distribution of agents sizes is lognormal[19]. The fact that the variable in the horizontal scale is 10^t is what is causing the shift $ln(10)\sigma^2$ and thus the concentration of value in the hands of the large agents.

Indeed, one can say the single most crucial characteristic of empirical distributions of firm level sales is that they are defined most easily with a log scale on the horizontal axis. That is, as log variates easily expressable as $10^{C(\cdot)}$ with $C(.)$ a normal distribution, or an exponential decreasing distribution (Pareto rule). This has been observed consistently over time in a variety of populations of economic agents, and is a key feature for the problem we are studying.

World trade split by different categories (by partner country, by product categories, by partner foreign firm) also show log-normal distributions and log fluctuations, which suggests

[18] If two other groups of, for example $n_0 = 10^4$ people holding a dollar (10^0) each, and $n_4 = 1 = 10^0$ person holding 10 thousand dollars (10^4) these hold respectively $v_0 = 10^{4+0} = 10^4$ and $v_4 = 10^{0+4} = 10^4$ dollars. The numbers in the exponents have to do with the vertical level of the parabolas in figure 2.1. This feature has to do with the slopes in between the log-normal distributions of population and value and may be important as it is essentially depicting a scenario of concentration.

[19] In fact, in figure 2.1 the yellow lines come from this analytical expression and not from an OLS fit to the yellow points.

many of the tools we develop as applying to agents can actually be used in a wider scope.

2.4.2 Firm level fluctuations

If firm sales are expressed in logs, it is natural to describe their dynamics as changes in log levels. We show the observed distribution of year on year log differences in plot of figure 2.2. Three plots are used to show growth rates for the smallest firms concentrating up to 25% of the value, intermediate size firms accounting for the next 25% of sold value and largest firms concentrating 50% of the value. Results for exports and imports largely overlapping in each of the plots. The scale is Log log, so that a double exponential (Laplace) distribution would look as $\sim -|x|$, normally distributed log shocks would appear as $\sim -x^2$ and on the contrary, for large firms we observe a $\sim -|x|^c$ with $0 < c < 1$. We see nearly symmetric fluctuations.

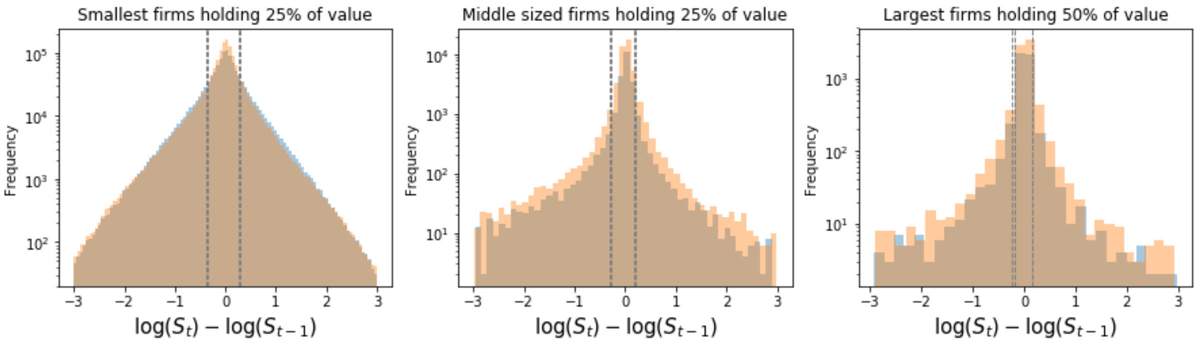


Figure 2.2: Fluctuations as log difference from previous period. Small (left) medium (mid) and large (right) firms. Exports and imports data are mostly overlapping. Vertical gray lines show smallest shocks accounting for half of the growth and half of the shrinkage. Fat tails are important but small log-normally distributed shocks are also present and account for a significant part of fluctuations.

Very importantly, note that for computing total sales by firms k at a time t, we would want to know the terms in $\sum S_{kt}$. This can be reconstructed by knowing the initial value S_{k0} and accumulating the log shocks up to t. This can be called the growth rates accumulation approach. Unfortunately, reconstructing the time series of levels of a firm solely from the general distribution of log differences would be a mistake because we would be not acknowledging autocorrelation structures that indeed exist. In most empirical settings there is a mild negative one-step auto correlation of log shocks, so that growth events are more likely followed by shrinkage events and vice versa.

To avoid this complications we can instead express the information of firm sales levels as deviations from a stationary (or average) value. In this case we do not need to accumulate a time series of log growth rates any more. We observe this directly from the empirical data (fig. 2.3). It would not be easy to derive analytical expressions for these distributions, but we do not need them.

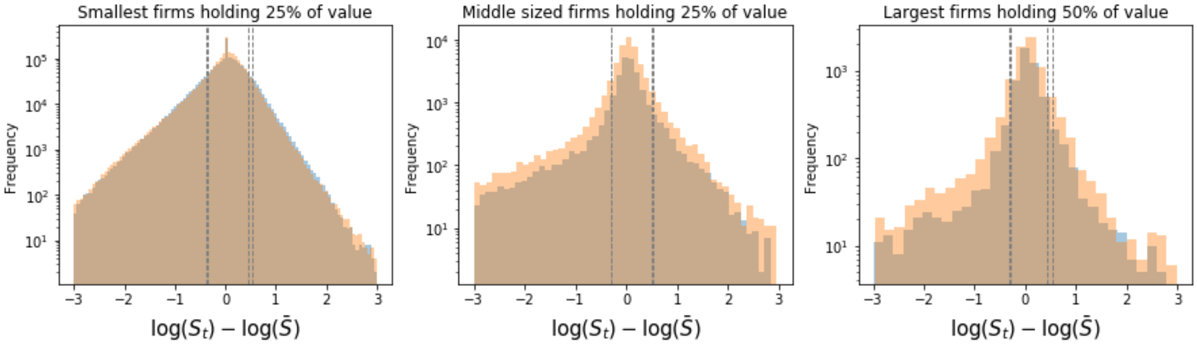


Figure 2.3: Fluctuations from mean $D(\cdot)$. Small (left) medium (mid) and large (right) firms. Exports and imports data are mostly overlapping. Vertical gray lines show smallest shocks accounting for half of the growth and half of the shrinkage. These fluctuations acknowledge accumulation of successive growth rates and they are to be used directly in accounting.

The values with distribution illustrated in Figure 2.3 are what we call *fluctuations*, to distinguish them from log jumps shown by agents (as in growth rates). Take these distributions (call it $D(.)$) as the large numbers limit distribution of deviations of a firm from its average value $\bar{s}$. If one wanted to sum the value sold by this firm over time, one would need to sum the values $\bar{s}10^{D(.)}$. Analogously, if $D(.)$ is the limit distribution of deviations for a population of firms at any given time step from their mean $\bar{s}$, then one would also need to sum $\bar{s}10^{D(.)}$ to know total sales at each time step.

All in all, expressing the information on firm levels over time as fluctuations has double advantage: we avoid the problem of accumulation of auto correlated time series, and also their powers are exactly what one needs to sum to obtain aggregates.

Same as with size distributions, the single most important feature of firm level fluctuations is that they adapt naturally to a description of the type $10^{D(\cdot)}$, where $D(\cdot)$ is usually a mixture of gaussians. In other words, they appear nicely when the horizontal axis is in log scale. This is the most important characteristic of agents' fluctuations and determines the formal path we can follow to aggregate them.

2.5 Mathematical framework: Aggregate volatility in log scale

The system we study in this work is a population of agents i that in time periods t contribute an amount s_{it} to total exports or imports $X_t = \sum_i s_{it}$. For practical purposes I call s_{it} and X_t as *sales* of firm i at time t, even if referring to exports or imports and not to domestic sales. The expected total sales from the population of agents is $E[X_t]$. This paper is dedicated to understanding how the variance of the time series of the total $var[X_T]$ can be expressed in terms of parameters describing the population of agents. This is a challenging but necessary task that so far has not been pursued completely.

The aggregate variance $var[X_t]$ can be used as measure of the width of fluctuations of aggregate sales, by taking its square root to arrive at the standard deviation of X_t . That is: $var^{1/2}[X_t] = std[X_t]$. In practice however this square root complicates aggregation, so that it is far more convenient to work with $var[X]$ and leave the square root for the very last step. This is why throughout this work I focus on var[X] as measure of volatility.

In the preceding section we have seen that agent sales and their fluctuations are best described in a log scale. In the remainder of this section I will review how fluctuations expressed in log levels should enter into expressions of variance accounting.

2.5.1 Easy facts about volatility

We start by having a time series X_t of length (dimension) T.[20] An estimator of its mean value is $\bar{X} \equiv \sum_t X_t/T$.

For accounting deviations of the elements of X_t we will look at their difference to the sample mean value. We will define this as $(\Delta X)_t \equiv \Delta X \equiv X_t - \bar{X}$.

The unbiased sample variance of X_t is:

$$var(X_t) = \frac{\sum_t (\Delta X)^2}{T-1} \tag{2.4}$$

and works as estimator of population variance of the time series. The biased sample variance

[20] In our case, $T = 2014 - 1997 = 17$. See section 2.3 for data details.

is $v\hat{a}r(X_t) = \sum_t (\Delta X)^2/T$. In a Normal iid distribution the variance we compute on a finite sample of length T is related to the variance at the large T limit by:

$$\lim_{T\to+\infty} var(X_t) = var(X_t)(T-1)/T = v\hat{a}r(X_t)$$

In this work we use sample variances computed on time series of length $T = 17$, unless otherwise stated. Then for us: $var(X_t) = 1.0625\ v\hat{a}r(X_t)$. [21]

The expression of total as linear combination ($X \equiv \sum_k S_k$) is very general. The components S_k are time series that add up to X but they can represent individual agents, or otherwise groups of agents (sectors). I use the name *parts* to refer to components that add up to X, but where it does not lead to confusion I call them *sectors*.

The aggregate variance is the sum of cross covariances among the time series of these parts. This is:

$$\mathrm{var}\left[\sum_k S_k\right] = cov\left(\sum_k S_k, \sum_k S_k\right) = \sum_{k_1,k_2} cov(S_{k_1}, S_{k_2}) \tag{2.5}$$

This property is very important as it is a general expression of aggregate sales variance in terms of the covariances among its parts. It is valid regardless of the details of parts' fluctuations and their cross correlations. In it we have introduced the sample covariance operator:

$$\begin{aligned} cov(S_{k_1}, S_{k_2}) =& E\left[\Delta S_{k_1,t} \cdot \Delta S_{k_2,t}\right] T/(T-1) \\ \equiv& \sum_t \left(\left(S_{k_1,t} - \bar{S_{k_1}}\right)\left(S_{k_2,t} - \bar{S_{k_2}}\right)\right)/(T-1) \end{aligned} \tag{2.6}$$

Aggregate variance is the total of all those elements. In the same way that the total sales need to include all agents' sales for an exact match.

2.5.2 More about partitions

If firms are grouped into a family of non overlapping subsets called *parts* (akin to sectors). Both aggregate sales, and their variance are defined in analogous way for firms and sectors (eg. change the index).

[21] This factor would be needed when comparing volatilities measured in different studies.

$$X_t = \sum_k S_{kt} = \sum_p S_{pt}$$

with indices k representing firms and p representing parts. Partitions could be given by any criteria: industry sectors, geographical regions, random allocation or others.[22] [23]

There are two types of partitions that will be specially useful due to their formal properties. These are *equal weight partitions* and *quantile partitions*.

Ideal equal weight partitions are those in which the value held by any P parts is the same and so it is $\bar{S}_p = \bar{X}/P$. In practice, this condition may not be achievable in precision but it is likely that we can separate the firms in parts with weights quite close to $\bar{X}/P$ for practical purposes. Apart from this condition, the assigment to parts can be randomized. We can use the name *random* partitions to refer to this case. Equal weight parts may simplify accounting.

A *size-sorted equal weight* partition (or quantile partition) is an equal weight partition into

[22] There is the possibility to define partitions on the sets of buyer firms as well as seller firms. This is amenable to a network for cross accounting of sales. If we use the indices p, r for referring to a pair of parts, then total sales expressed from exchanges between parts and exchanges between firms are given as:

$$X = \sum_{p,\, r} s_{pr} = \sum_{p,\, r} \sum_{k,\, l \in p,\, r} s_{kl}$$

All of the sales of firm k are associated to a firm l on the other end.

$$X = \sum_k S_k = \sum_{k,\, l} S_{kl}$$

Here S represent the value of sales, k and l represent a pair of firms.

[23] For the sake of completeness we can define an atomistic micro level that adds up to our firms 'micro' level. It is clear that the sales between a pair of firms during a given time period can in turn be disaggregated into transactions i, made of sales of items j in quantities q_j at prices p_j. The value of a transaction s_j is in units of currency which arises from $s = p.q$. Here, p and q area measured in terms of conceptual units that 'cancel out' at $s = p.q$. All aggregation is therefore in units of currency. In general the price can be assigned to each item j when observed. We do not need to assume that they belong to a product or that they are a function of time, or that they are the same for different pairs of agents exchanging the same product.

$$S_k = \sum_i t_i = \sum_i \sum_{j \in t_i} p_j q_j$$

So that prices and quantities in a collection of transactions from firm k determine the observed time series for sales from firm k, i.e. S_k.

In this work however we take S_k as given and study the aggregation up to the national level. In this picture, firm level sales and atomistic items exchanged still relate exactly in an ordinary sum.

$$X = \sum_k S_k = \sum_j p_j q_j$$

From here one would have an aggregation involving prices.

Q parts with the condition that firms have been sorted by size before cutting them into the Q groups. The total sales of each partition is near $\bar{S}_q = \bar{X}/Q$. With a large enough number of parts (typically Q = 10, 20) it is likely that firms in most parts are quite close to a mean size ($\bar{s}_q$).

We would expect that $\bar{S}_q = \bar{X}/Q = \bar{s}_q n_q$ where n_q is the number of firms in quantile (part) q. The mean size $\bar{s}_q$ grows monotonically in successive parts q, then the quantile population n_q must decrease monotonically (overall keeping the quantile value fixed near $\bar{S}_q = \bar{X}/Q$). In fact, both of them can be taken as a function of the percentile q. Therefore quantile parts tend to have a defined agent size and population size, this is the feature that makes them very useful. They also allow working with micro moments that can be functions of agent size, and as such become functions of quantile q.

2.5.3 Simple structure of sectoral sales

Assuming a structure for the time series of sectoral sales means defining it as a sum of time series that partially make it up. As an example, fitting sectoral sales with a generic s_{pt} leads to $S_p = \hat{\beta}_p s_t + r_{pt}$ and an expression of aggregate variance (sum of elements of the cross covariance matrix):

$$var(X) = \sum_{ij} cov(S_i, S_j) = \sum_{ij} cov(\hat{\beta}_i s_t + r_{it}, \hat{\beta}_j s_t + r_{jt}) \tag{2.7}$$

where $\hat{\beta}_p$ are elements of $\boldsymbol{\beta} = (s_t^T s_t)^{-1} s_t^T Y$, with $Y \in \mathbb{R}^{T \times P}$ containing the observed parts' time series and s_t is a column vector with the time series used to fit. It can be z-standardized so that $cov(s_t, s_t) = 1$ (see figure 4). [24]

The elements of the cross covariance matrix will look like:

$$\begin{aligned} cov(\hat{\beta}_i s_t + r_{it}, \hat{\beta}_j s_t + r_{jt}) =& cov(\hat{\beta}_i s_t, \hat{\beta}_j s_t) + cov(\hat{\beta}_i s_t, r_{jt}) \\ &+ cov(r_{it}, \hat{\beta}_j s_t) + cov(r_{it}, r_{jt}) \end{aligned} \tag{2.8}$$

so that each of the $P \times P$ elements is made of *four* components. In general, a structure of

[24] A specific example: if we fit (OLS) using a factor of linear time evolution $s_t = t$, then: $S_{pt} = \hat{\beta}_p t + r_{pt}$, the β_p is as described and r_{pt} are the residuals of the fit.

sectoral sales made from N terms leads to $N \times N$ terms making up each element of the cross covariance matrix.

The $\hat{\beta}$ coefficients are thus arranged into a column vector $\hat{\boldsymbol{\beta}}$ of length P, and the standard deviation of residuals σ in the column vector $\hat{\boldsymbol{\sigma}}$ of length P. The matrix product $\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T$ creates a $P \times P$ matrix (it is an outer product), and the cross covariance matrix among sectors can be written as:

$$\begin{aligned} C =& \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \cdot C(s_t, s_t) + \hat{\boldsymbol{\beta}}\boldsymbol{\sigma}^T \cdot C(s_t, \epsilon_t) \\ &+ \boldsymbol{\sigma}\hat{\boldsymbol{\beta}}^T \cdot C(\epsilon_t, s_t) + \boldsymbol{\sigma}\boldsymbol{\sigma}^T \cdot C(\epsilon_t, \epsilon_t) \end{aligned} \tag{2.9}$$

where $C(a, b)$ denotes the $P \times P$ cross covariance matrix computed from vectors a, b. They are multiplied to the outer product matrices element wise.

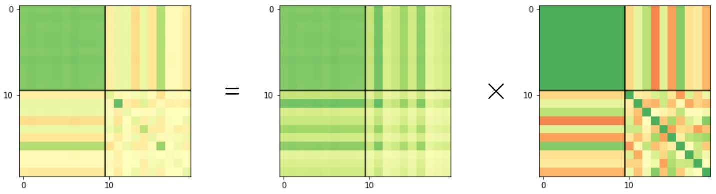


Figure 2.4: Structure of cross covariance matrix ($N \times N$ blocks with elements made of outer product times net covariance). Illustration from simplified toy example where sectoral time series are decomposed into dependence as mean time series (fitted by OLS) and residuals.

This structure of the cross covariance matrix as an elementwise product of an outer product matrix and a net cross covariance is illustrated in Figure 4.

Here we are seeing that aggregate variance is made of a *comovement* contribution and an *idiosyncrasies* contribution, apart from cross terms that often can be expected to not contribute significantly.

2.5.4 Non linearities

Fluctuations of the level of agents given in logarithmic scale imply nominal contributions from different agents given by an exponential function. When we talk about nonlinearities in this work we are essentially referring to the fact that the levels we need to add are along an exponential curve. Let us introduce general properties of a logarithmic transformation.

We know 100 extra EUR sold by some firm translate directly to aggregate exports, we also know that this is in some sense 'insignificant' given that total exports can easily be over 100 bn EUR. This insignificance can be formalized noting that the ratio between the two levels is 1.0000001, i.e. aggregate exports are largely the same after the 100 EUR. Indeed, computing ratio to a base level is a good way of abstracting changes from the nominal scale. As such, S_t/S_0 (in the last example $(10^9 + 10^2)/10^9 = 1 + 10^{-7}$) is a useful transformation.

Alternatively, pondering equal *multiplicative changes* to X as equally important (as opposed to considering their *additive* magnitude) is what we get by evaluating *additive changes* of a logarithmic transformation of X, denoted $\log(X)$.

For practical purposes we can treat $log(z) : \mathbb{R} \to \mathbb{R}$ simply as a nonlinear function that can be approached by a Taylor series. The base is 10, unless otherwise stated. Its Taylor series about z_0 tells us:

$$\log(z)\ \ln(10) = \log(z_0) + \sum_{n\geq 1} (-1)^{n+1}(z - z_0)^n/(n\ z_0)$$

where the factor $\ln(10) \approx 2.3026$ is the natural logarithm of 10. [25]

The expansion of $\log(z)$ (base 10) at $z_0 = 1$ up to the first order:

$$(z - 1) \approx \ln(10) \log(z)$$

This relation can be brought to practical use if we identify $z = X_t/\bar{X} \approx 1$, in which case we get to:

$$\frac{X_t - \bar{X}}{\bar{X}} \approx \ln(10) \log\left(\frac{X_t}{\bar{X}}\right) \Rightarrow \frac{\Delta X}{\bar{X}} \approx \ln(10)\Delta(\log(X)) \tag{2.10}$$

This relation is extremely important as it is telling us the connection between small nominal deviations from the stationary level $\bar{X}$, percentage change factors and logarithmic fluctuation. This connection is of course widely exploited, however less attention is placed to the equally important ability of knowing when eq. 2.10 does not apply, which indeed is fairly often as

[25] This factor appears repeatedly and originates from our use of the base 10 instead of natural logarithms. We use the base 10 because it allows simple interpretation of log levels (6 = 1 mn EUR, 9 = 1 bn EUR and so on). We have to consciously carry the $\ln(10)$ factor along calculations. It helps developing intuitions by seeing where the choice of the base 10 over the base e plays a role.

well.

Approximating the log curve at $z_0 = 1$ to the second order tells us: $\log(z)\ \ln(10) \approx (z-1) - \frac{1}{2}(z-1)^2$, we then have:

$$(z-1) \approx \ln(10)\log(z) + \frac{1}{2}(z-1)^2$$

Therefore for a rough condition for the first order approximation to be valid we should ask that this last second order term is not very large:

$$(\Delta X/\bar{X})^2/2 \ll 1$$

The second order error term is about 0.005 for deviations of about 10% and 0.1 for deviations of 50%. It means that for time series that fluctuate on the order of 10% or less it is safe to use equation 2.10, but it would not be recommended for more volatile time series.

Next, let us characterize the inverse of $log(z)$. Let us define logarithmic fluctuations as:

$$F_t \equiv \log\left(\frac{S_t}{\bar{S}}\right) \tag{2.11}$$

where because we can refer to sectoral sales (not only aggregate sales) we have introduced S in place of X. Inverting this relation we have:

$$S_t/\bar{S} = 10^{F_t} \tag{2.12}$$

From there, subtracting 1 on both sides, we can express additive deviations ΔS in terms of logarithmic fluctuations:

$$\Delta S_t/\bar{S} = 10^{F_t} - 1 \tag{2.13}$$

This is the exact relation between nominal additions and logarithmic fluctuations observed in a time series. A Taylor expansion of this exponential curve brings us back to the approximate expression in 2.10. Indeed: $10^{F_t} - 1 = \ln(10)(F_t) + \frac{1}{2}\ln^2(10)(F_t)^2 + O(F_t)^3$. So that up to first order:

$$\Delta S_t/\bar{S} \approx \ln(10)F_t \tag{2.14}$$

which needs at least that: $\frac{1}{2}\ln^2(10)F_t^2 \ll 1$. This means we would need these log fluctuations to be $F_t \approx .1$ or smaller. This condition is usually met by aggregate sector fluctuations, but not by firm level fluctuations.

2.5.5 Mean and variance of transformed random variables

If we knew the moments (expected value E[X] and variance var[X]) of aggregate sales, can we estimate the moments of the logarithm of aggregate sales? The answer is yes, and the method for this is straightforward. We can consider a Taylor expansion for the moments of $f(x) := \log(x)$, which is a function of a random variable X. Using that the first and second derivatives of $f(x)$ are: $f'(x) = 1/(\ln(10)x)$, $f''(x) = -1/(\ln(10)x^2)$, the expected value and variance of $\log(X)$ must be approximately:[26]

$$\mathrm{E}\left[\log(X)\right] \approx \log(\mathrm{E}\left[X\right]) + \frac{f''(\mathrm{E}\left[X\right])}{2}\operatorname{var}\left[X\right] = \log(\mathrm{E}\left[X\right]) - \frac{1}{2\,\mathrm{E}\left[X\right]^2\ln(10)}\operatorname{var}\left[X\right] \tag{2.15}$$

$$\operatorname{var}\left[\log(X)\right] \approx \left(f'(\mathrm{E}\left[X\right])\right)^2\operatorname{var}\left[X\right] = \frac{1}{\left(\ln(10)E[X]\right)^2}\operatorname{var}\left[X\right] \tag{2.16}$$

To know the order to which we should approximate a variable we need to consider the magnitude of its fluctuations. If X stands for gross exports (imports) of a large national economy it is safe to keep up to linear terms in the expansions of moments so that $\mathrm{E}\left[\log(X)\right] \approx \log(\mathrm{E}\left[X\right])$ and $\operatorname{var}\left[\log(X)\right]$ is expressed up to the terms of the series as in eq. 2.16.

In the situation of equation 2.16 $\operatorname{var}\left[\log(X)\right]$ and $\operatorname{var}\left[X\right]$ (*log* and *linear* variance) can be taken as proportional to each other. We can easily distinguish them because $\sigma^2(log(X)) \sim 1$ while $\sigma^2(X) \sim 10^{20}$ if $\bar{X} \sim 10^{11}$ in EUR.

Note also that:

[26] we have dropped the subindex of X_t for simplicity

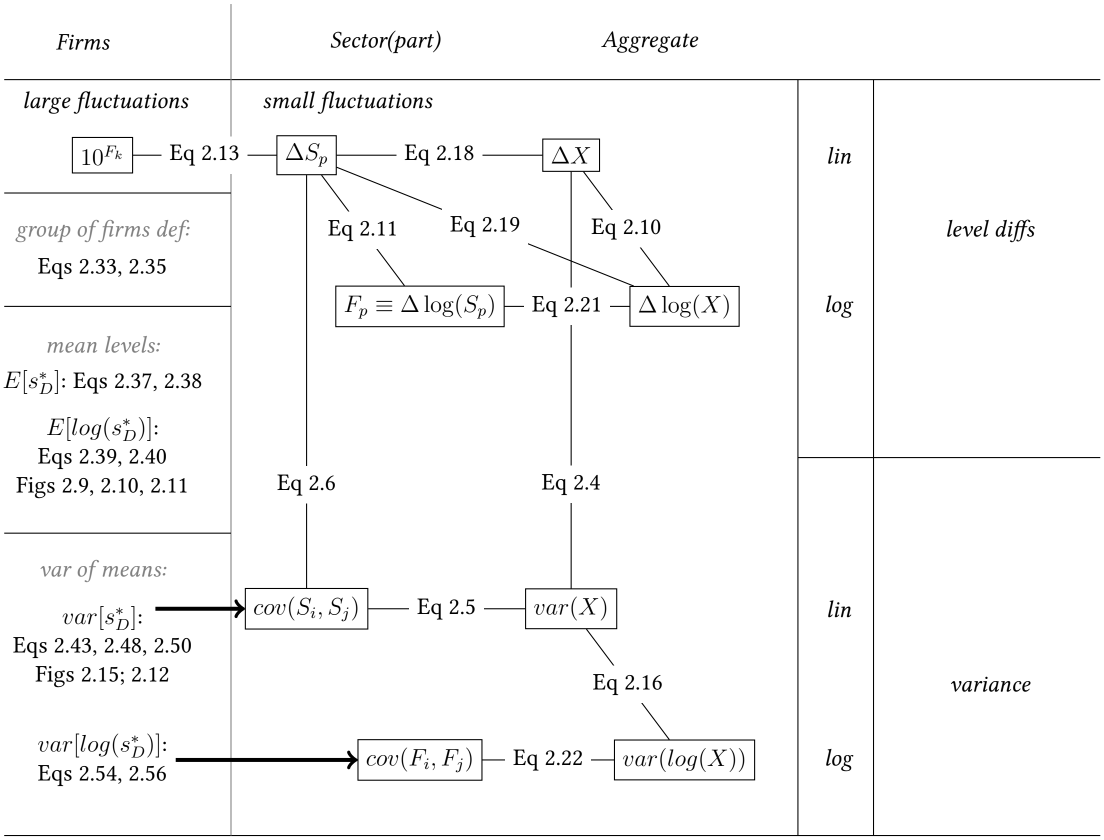


Figure 2.5: Scheme for the equations describing differences and variances of linear combinations, in linear and log scale.

$$\mathrm{var}\left[\frac{X}{\bar{X}}\right] = \frac{\mathrm{var}[X]}{\bar{X}^2} \approx \ln^2(10)\,\mathrm{var}[log(X)] \tag{2.17}$$

The variance of log levels is closer in magnitude to the variance of (X_t/X), but for a factor $ln^2(10) \approx 5.3$. If the higher order terms of equation 2.16 are small.

The scheme in figure 2.5 offers an overview of the relations laid out in this section.

2.6 Firms are not Sectors

As we have noted the total X_t can be expressed as sum of parts in the same way that it is expressed as sum of firm level sales:[27]

$$X_t = \sum_k S_{kt} = \sum_p S_{pt}$$

where k, p index firms or sectors respectively. Nominal fluctuations of the aggregate can of course be also expressed as aggregations of firms levels and sectoral levels.

$$\frac{X_t}{\bar{X}} - 1 = \frac{\Delta X_t}{\bar{X}} = \frac{\sum_k \Delta S_{kt}}{\bar{X}} = \frac{\sum_p \Delta S_{pt}}{\bar{X}} \tag{2.18}$$

Aggregate fluctuations are usually mild enough to allow a linearization of their log deviations (eq. 2.16).

$$\ln(10)\log\left(\frac{X_t}{\bar{X}}\right) \approx \frac{\Delta X_t}{\bar{X}} = \frac{\sum_k \Delta S_{kt}}{\bar{X}} = \frac{\sum_p \Delta S_{pt}}{\bar{X}} \tag{2.19}$$

Without loss of generality we can register the observed deviations from mean levels in log scale, that is, as F_t with $F_{pt} = \log(S_{pt}) - \log(\bar{S}_p)$ for part p, and in an analogous way $F_{kt} = \log(S_{kt}) - \log(\bar{S}_k)$ for firm k. Usually one would expect that F_t are small, near null fluctuations. For example it could be that $F_t = m_{kt} + \sigma_k \epsilon_{kt}$ with $m_{kt} \ll 1$ and ϵ_{kt} a time series of random shocks centered in zero and with $std(\epsilon_{kt}) = 1$.

This definition of fluctuations implies $S_t = \bar{S}10^{F_t}$, both for sectors and firms. And so the nominal fluctuations are $\Delta S_t = S_t - \bar{S} = \bar{S}(10^{F_t} - 1)$, as in equation 2.13. This step is the key to match log shocks that one observed to nominal shocks that one needs to account for. The relation between log aggregate sales and log micro shocks to firms or sectors is then:

$$\ln(10)\log\left(\frac{X_t}{\bar{X}}\right) \approx \frac{\Delta X_t}{\bar{X}} = \sum_k \frac{\bar{S}_k}{\bar{X}}(10^{F_{kt}} - 1) = \sum_p \frac{\bar{S}_p}{\bar{X}}(10^{F_{pt}} - 1) \tag{2.20}$$

We will get a substantial idea of what is going on, however, if we consider what are the actual

[27] In this section I may use the name *sector* or *parts* indistinctly to refer to groups of firms as introduced in section 2.5.2.

magnitudes that these F_t have in each of the cases. And for this, refer to figure 2.6 where on the left side we have the distribution of F_{kt} observed in firms, and on the right, the aggregate lo fluctuations (top) the F_{pt} fluctuations observed in a random partition into $P = 10$ parts (mid), and the F_{qt} fluctuations observed in a quantile partition into $Q = 10$ parts (bottom). The horizontal axis is for log fluctuations and the vertical axis is the magnitude of the nominal fluctuations that these F_t imply. The black lines are then an exponential curve (base 10) and I show on red the approximations to these curves by polinomials of increasing degree.

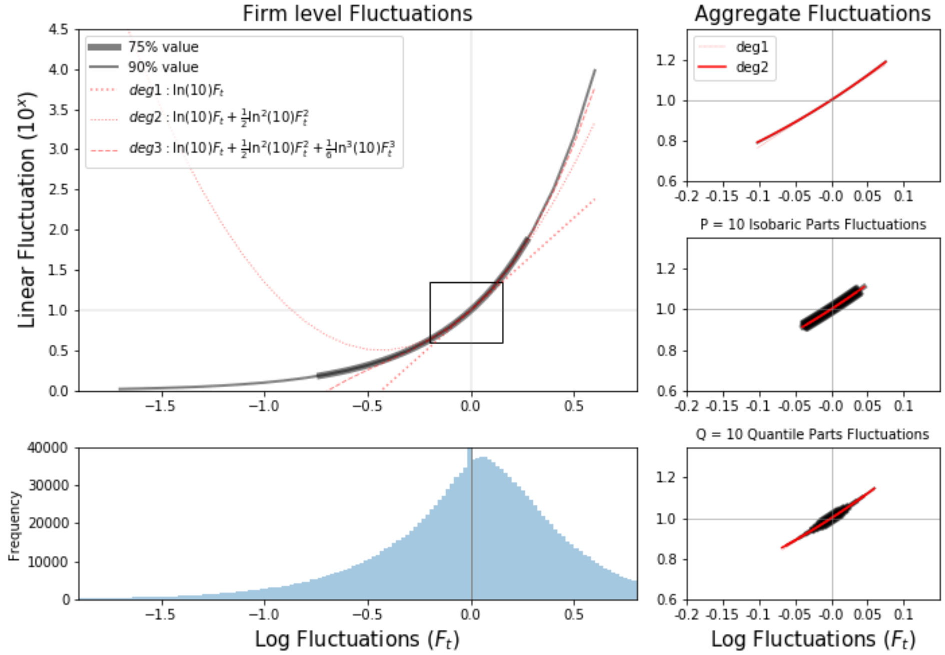


Figure 2.6: Left: distribution of log micro shocks and magnitude of nominal differences they imply. The curve is 10^{F_t}, series approximations in red. Right: log fluctuations and nominal differences for the aggregate (top) and groups of firms arranged into P = 10 random parts (mid) and Q = 10 quantile parts (bottom). The linear approximation for nominal differences can be used in these cases.

The information to take from figure 2.6 has to do with the magnitude of nominal differences that log fluctuations imply. The thicker curve is accumulating 75% of the total value, and the thinner ones accumulate up to 90%. Log fluctuations of firms are too wide to proxy the implied nominal fluctuations by means of a linear dependence.

Sectoral fluctuations (10 parts) on the contrary are mild enough to allow this linearization, and we can largely benefit from this. Sectors will adapt to the following rule:

$$\Delta \log(X_t) \approx \frac{1}{\ln(10)} \frac{\Delta X_t}{\bar{X}} = \frac{1}{\ln(10)} \frac{\sum_p \Delta S_{pt}}{\bar{X}} \approx \sum_p \frac{S_{pt}}{\bar{X}} F_{pt} \tag{2.21}$$

where the rightmost term is now a linear combination, as opposed to a sum of nonlinear functions as it used to be. [28]

At firm level, we have no option but to keep equation 2.20 and using equation 2.21 would be grossly incorrect. If the micro shocks are too large the approach of using a Taylor series for 10^F will not work unless we include too many orders, which is not practical anymore.

If we can express log aggregate deviations as a linear combination of sectoral log deviations as in 2.21 then by the properties of aggregate variance (especially 2.5) the following relation will apply:

$$\begin{aligned} var(\log(X)) \approx \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} cov(F_{it}, F_{jt}) &= \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} cov\left(\log\left(\frac{S_i}{\bar{S}_i}\right), \log\left(\frac{S_j}{\bar{S}_j}\right)\right) \\ &= \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} cov\left(\log(S_i), \log(S_j)\right) \end{aligned} \tag{2.22}$$

where indices i, j represent sectors. For the second line note that dividing by a fixed value $\bar{S}$ before taking log does not change variance.

In this equation we have clarified how small log fluctuation F_t of parts of a total contribute to variance of the log total (if these F_t are small enough). This is the 'log equivalent' of the linear sum of variance rule $var[X] \equiv \sum_{i,j \in P} cov(S_i, S_j)$.

Shocks to individual agents can easily reach a magnitude of $\sigma_k \approx .5$ or larger. This means in a given year a firm may sell a third (or three times) its value of average annual sales. In those conditions (and even with milder shocks $\sigma_k > .1$) expressions of aggregate variance as in 2.22 are not valid. Using them is essentially a mistake and they may easily lead to incorrect

[28] If one wanted to draw a connection to aggregation under a production function, and so, a connection to usual growth accounting notations, we could introduce $h(X) = X$ and consider an expression of log derivative at a finite time step Δt. There, $h'(X) = \Delta X/\Delta t$ and from $h'/h \approx d\ln(h)/\Delta t$ one has $\Delta X/(\Delta t\ \bar{X}) \approx (\ln(X_t) - \ln(\bar{X}))/\Delta t = \ln(10)\log(X_t/\bar{X})/\Delta t$, as in equation 2.21. We can 'cancel out' the Δt. This has no problems because we are determining how far a value is from a reference, independently of whether we frame it as time evolution.

In a growth accounting equation there is also a production function $Y(t) = F(\mathbf{x}, t)$ and a difference in log production is:

$$\frac{1}{Y}\frac{dY}{dt} = \sum_i \frac{1}{F}\left(\frac{dF}{dx_i}\frac{dx_i}{dt} + \frac{dF}{dt}\right)$$

but our total is defined in as a simple sum, not a production function. In our case then $dF/dx_i = 1\ \forall\ i$ and $dF/dt = 0$. We get directly to identity in eq. 2.18.

results.

Still, the type of relation in equation 2.22 is assumed at the firm level by Gabaix (2011) in its equation 3. There is no guarantee that the right hand side in this equation is aggregate idiosyncratic volatility because necessary conditions on ϵ_{it} are likely not met.

The fact that a linear relation as in equation 2.21 does not apply for firms is not caused by having proposed F_t log shocks and therefore 10^{F_t} nominal shocks. The problem is implicit in the distribution of the nominal shocks themselves. In real world settings if we say (as in Gabaix's eq. 1) that $\Delta S_{it} = \sigma_i \epsilon_{it}$, we can ask that $var(\epsilon_i) = 1$ but these ϵ_{it} will be not at all normally distributed. They will be highly asymmetric and not centered in zero. This results in that we will not know what $cov(\epsilon_{it}, \epsilon_{jt})$ are. Most certainly they will not be uncorrelated, so that the whole path followed in this paper has no guarantees of adapting to empirical realities.

Fortunately, at least the log differences to stationary levels do show distributions $D(\cdot)$ that one can work with. Then expressing firm level shocks as $10^{D(\cdot)}$ is then not a convention among others we could choose from. Instead, it is the best open path for accounting micro shocks in a correct way. Uniting $10^{D(\cdot)}$ micro shocks to aggregate volatility is viable although not exempt from certain difficulty. This path will be pursued in section 2.8.

2.7 Reviewing the diversification issues

If a national (or global) aggregate time series is made of the sum of contributions from (N) thousands or even millions of agents, why would not their idiosyncratic shocks cancel out?

This intuition can be formalized as an expectation that the volatility of the time series telling the total from a population of fluctuating agents should fall as $\sigma \sim 1/\sqrt{N}$. Such is the rate of decay observed in a population of agents showing additive gaussian fluctuations. This intuition has been invoked often. For an implicit reference, among many others, one could mention the paragraph: *"... in a complex modern economy, there will be a large number of such shifts in any given period, each small in importance relative to total output. There will be much 'averaging out' of such effects across markets."* in (Lucas, 1977)

But is the magnitude of aggregate standard deviation approximately $1/\sqrt{N}$? And if not why?

A quick answer is that it is not because all agents can be growing or shrinking as part of an 'aggregate shock' of a magnitude larger than $1/\sqrt{N}$. Still in Lucas (1977) we have: *"... there have been many instances of shocks to supply which affect all, or many, sectors of the economy simultaneously. Such shocks will not cancel in the way I have described, and they will induce output fluctuations in the aggregate."*

Then, still, the question can be reformulated as referring to the idiosyncratic part of aggregate volatility. Does this term follow a $1/\sqrt{N}$ rule? The answer is it falls more mildly, but let us approach this issue patiently because it is not as simple as it may appear.

Given a partition of the population of firms into sectors fluctuating mildly enough (from equations 2.5, 2.16, 2.21 and 2.22 in preceding sections) $\mathrm{var}[\log(X)]$ is approximately:

$$\mathrm{var}[\log(X)] \approx \frac{\mathrm{var}[X]}{(\ln(10)E[X])^2} \approx \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} \mathrm{cov}(F_i, F_j) \tag{2.23}$$

where i, j are used to denote a pair of parts p.

Without loss of generality, sectoral log fluctuations can be expressed as $F_{pt} = m_{pt} + \sigma_p \epsilon_{pt}$. In that case:

$$\begin{aligned} \mathrm{cov}(F_i, F_j) = cov(m_{it} + \sigma_i \epsilon_{it}, m_{jt} + \sigma_j \epsilon_{it}) =& cov(m_{it}, m_{jt}) + cov(m_{it}, \sigma_j \epsilon_{jt}) \\ &+ cov(\sigma_i \epsilon_{it}, m_{jt}) + cov(\sigma_i \epsilon_{it}, \sigma_j \epsilon_{jt}) \end{aligned} \tag{2.24}$$

In general, one should consider these as the P^2 elements to be summed to arrive at aggregate variance. There are special cases where this expression simplifies. These cases are relevant as guides for analysing real life settings in practice, although one must not forget the specific conditions that let them be derived from the general case.

On the one hand one can ask for uncorrelated shocks of unit variance, that is: $var(\epsilon_{it}, \epsilon_{jt}) = \delta_{ij}$ (Kronecker's delta, 1 if $i = j$, else 0). A unit variance may be asked without further problems, given we have the parameters σ_j to capture the magnitude of idiosyncratic fluctuations. The independence of cross sectoral idiosyncrasies is however a qualitative limit case that may not apply completely. The left hand side of Eq. 2.23 is so far:

$$\mathrm{var}[\log(X)] \approx \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} \left[cov(m_{it}, m_{jt}) + \sigma_j cov(m_{it}, \epsilon_{jt}) + \sigma_i cov(\epsilon_{it}, m_{jt}) + \sigma_i \sigma_j \delta_{ij}\right] \quad (2.25)$$

In addition, we could ask that the time series m_{pt} are uncorrelated from idiosyncratic sectoral shocks ($cov(m_{it}, \epsilon_{jt}) = cov(\epsilon_{it}, m_{jt}) = 0$). Not just any m_{pt} time series will fulfill this condition. This actually constrains the scope of what we can take as an m_{pt}. In practice, once we have m_{pt} candidates one can confirm whether the terms $cov(m_{it}, \epsilon_{jt})$ are small enough for the following equation to be valid:

$$\mathrm{var}[\log(X)] \approx \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} \left[cov(m_{it}, m_{jt}) + \sigma_i \sigma_j \delta_{ij}\right] \quad (2.26)$$

Here we have the aggregate shocks term adding to aggregate volatility. Both comovements (aggregate shocks) and idiosyncratic shocks contribute, so that any of them can dominate. As we mentioned, this possibility was contemplated all along. However, also note that the proposed early solutions invoking aggregate shocks still assumed the idiosyncratic variance is vanishing because of decaying at a rate $\sigma^2 \sim 1/N$, although empirical systems do show a milder decay.

If all parts p shared a single m_{pt}term:

$$\mathrm{var}[\log(X)] \approx \sum_{i,j} \frac{\bar{S}_i \bar{S}_j}{\bar{X}^2} \left[var(m_{pt}) + \sigma_i \sigma_j \delta_{ij}\right] \quad (2.27)$$

If all P parts are of near equal size: $\bar{S}_p/\bar{X} \approx 1/P$, then performing the sum in Eq. 2.27 we have:

$$\mathrm{var}[\log(X)] \approx \underbrace{\frac{\bar{S}_p^2}{\bar{X}^2} P^2 var(m_{pt})}_{\sigma_m^2} + \underbrace{\sum_p \frac{\bar{S}_p^2}{\bar{X}^2} \sigma_p^2}_{\sigma_\epsilon^2} = \underbrace{var(m_{pt})}_{\sigma_m^2} + \underbrace{\frac{1}{P^2} \sum_p \sigma_p^2}_{\sigma_\epsilon^2} \quad (2.28)$$

Linear variance is approximately $\mathrm{var}[X] \approx \ln^2(10)\bar{X}^2 \, \mathrm{var}[\log(X)]$ (Eq. 2.16), so that:

$$\mathrm{var}[X] \approx \ln^2(10) \left(\bar{X}^2 var(m_{pt}) + \sum_p \bar{S}_p^2 \sigma_p^2\right) \quad (2.29)$$

In equation 2.28, the terms $\sigma_m^2 \equiv \bar{S}_p^2 var(Pm_{pt})/\bar{X}^2$ and $\sigma_\epsilon^2 \equiv \sum_p \bar{S}_p^2\sigma_p^2/\bar{X}^2$ are respectively aggregate and idiosyncratic shocks variances.

A combination of the fact that that contributions from a part p to idiosyncratic volatility are proportional to the magnitude of this part ($\bar{S}_p$) when it comes to $var[X]$ (eq. 2.29) and to the share ($\bar{S}_p^2/\bar{X}^2$) when it comes to $var[log[X]]$ (eq. 2.28), together with the fact that typical size distributions (Pareto, lognormal) imply a concentration of a significant piece of the total in relatively few agents is what lets aggregate idiosyncratic volatility not be as low as $\sigma_\epsilon^2 \sim 1/N$.

These two features combine to let the sample variance of small groups of large agents drive the idiosyncratic volatility seen on the whole population. For example, consider 100 companies are responsible for one half of the exports, and 10 000 companies are responsible for the other half, we must not say naively that we have 10100 agents and so expect the population idiosyncratic volatility to be $\bar{\sigma}/\sqrt{10100} \approx \bar{\sigma}/100$. Instead the two groups lead to a larger standard deviation of the aggregate: $\bar{\sigma}(1/\sqrt{100} + 1/\sqrt{10000})/2 \approx \bar{\sigma}/18$. In this case, there is an effective $N_{eff} \approx 330$, smaller than $N = 10100$ when it comes to sample variance.

Note that this value weighted average leading to an apparently 'inflated' idiosyncratic variance can be conceptualized as effectively implying a smaller number of agents $N_{eff} < N$ in the population. The variance σ_ϵ^2 is higher because a small number of agents translate their log fluctuations to the aggregate. However, very importantly note that this is not in itself a milder decay rule $\sigma_\epsilon^2 \sim N^{-\alpha}$ with $\alpha < 1$.

Gabaix (2011) postulates a power law size distribution is responsible for a milder decay $\sigma_\epsilon^2 \sim N^{-\alpha}$ that may be a function of parameters of the size distribution. If we follow their paper carefully however, we see this postulate is in fact derived for the Herfindahl index $h^2 = \sum_i(\bar{S}_i/\bar{X})^2$. By means of relations between the Herfindahl and aggregate idiosyncratic variance (eqs. 4, 5 in op. cit.) the property derived for Herfindahl index is automatically assumed to apply to aggregate idiosyncratic variance. I have discussed in previous sections how it is a mistake to use linear expressions (Gabaix's equation (3)) to aggregate agents volatility. Even if we ignored this, equations (4) and (5) imply assuming that σ_k has a single value for all agents, which takes us away from real life settings in an uncontrolled way. In other words, if we have a sum of multiplied pairs of factors (sum of variance times value shares as in eq. 2.28) taking one of these factors out of the sum as if it was a constant will be misleading. It is possible that one can write $\sigma_\epsilon^2 = \bar{\sigma}^2 h^2$ but nothing guarantees that this $\bar{\sigma}^2$ is the variance of individual agents, and nothing guarantees that it does not depend with N which is precisely the parameter on which we are trying to determine how σ_ϵ^2 depends.

The Proposition 2 of Gabaix (2011) therefore needs to be restricted to Herfindahl index decay with population size and not to aggregate idiosyncratic variance. [29]

The important issue of knowing how idiosyncratic variance decays with population size (and understanding the reasons why) has therefore been open all along, even if wrongly believed to be sorted out. It is interesting to note that hundreds of papers have referred to results in Gabaix (2011), but so far I have found no citing paper dedicated to confirming or studying further the results in its Proposition 2, which are the main result of the paper. Is this an indication by omission that the $\sigma_\epsilon^2 \sim N^{-\alpha}$ relation proposed there is not observed empirically?

In the following sections I offer tools for studying the problem of $\sigma_\epsilon^2 \sim N^{-\alpha}$ dependence carefully. After that, I show that this milder decay with N is explained by nonlinearities adding to comovement terms among agents.

2.7.1 Dependence of var(X) with N

The framework in this section is useful for approaching the problem of dependence with population size (N). We denote a reference population size as N_0 and consider a different population size $N_1 = k\ N_0$.

Figure 2.7 introduces the variables involved in this problem. A power law of variance with N implies a linear relation between $log(\sigma^2)(= 2log(\sigma))$ and $log(N)$. Its slope is $-\alpha$, and this slope determines the ratio Dy/Dx. If we consider two populations, the second of which is a multiple k of the first one, then $Dx = log(k)$ and the difference between the variance that these two populations show is given by $-\alpha Dx = -\alpha log(k)$. In the following steps we use these type of relations, and in addition we will consider that they can apply to parts of the total, not just to the total population of the sample.

To link changes in total population to changes in parts' population consider: if we sample $N_1 = k\ N_0$ agents from a population, on average we expect that the population of each part p

[29] Still some other steps also need to be studied carefully. The condition in equation (13), for example is very sensible to the size distribution being exactly a Pareto, so that we may not count with it in many real cases.

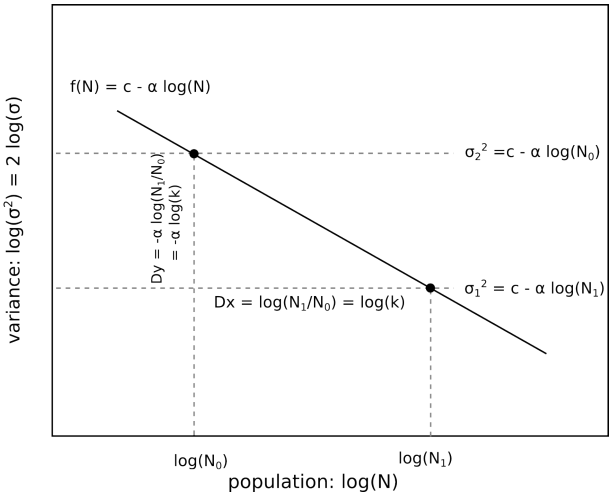

f(N) = c - α log(N)
σ$_2$^{2} =c - α log(N$_0$)
Dy = -α log(N$_1$/N$_0$)
= -α log(k)
σ$_1$^{2} = c - α log(N$_1$)
Dx = log(N$_1$/N$_0$) = log(k)
variance: log(σ^2) = 2 log(σ)
log(N$_0$)
log(N$_1$)
population: log(N)

Figure 2.7: Scheme for analysing the decay of volatility with population size implied by a power law $\sigma^2 = 10^c N^{-\alpha}$.

is $k\ n_p(N_0)$, where $n_p(N_0)$ is the population expected at part p when the total population size is N_0. In logarithmic scale, this means that if $log(N_1) = log(N_0) + log(k)$ then $log(n_p(N_1)) = log(n_p(N_0)) + log(k)$, for all parts $p \in P$. [30]

Empirically we see that the dependence of a part's log variance with changes in the part's log population can be approximated qualitatively by a line of slope $-\alpha$.

$$log(\sigma_p^2(n_p)) = c - \alpha_p \log(n_p) \Leftrightarrow \sigma_p^2(n_p) = \frac{10^c}{n_p^{\alpha_p}} \tag{2.30}$$

The accuracy of this model can be tested a posteriori. But what would it imply? When changing n_p for kn_p, the levels of $log(\sigma_p^2)$ change as:

$$log(\sigma_p^2(n_p(N_0))) - \alpha_p log(k) = log(\sigma_p^2(k\ n_p(N_0))) \quad \Leftrightarrow \quad \frac{1}{k^{\alpha_p}} \sigma_p^2(n_p(N_0)) = \sigma_p^2(k\ n_p(N_0))$$

If all parts p present a common $\alpha_p \equiv \alpha$ exponent, then when replacing this value in the expression of the idiosyncratic term of aggregate variance, the dependence with α comes out as common factor:

$$\frac{1}{k^\alpha} \frac{1}{P^2} \sum_p \sigma_p^2(n_p(N_0)) = \frac{1}{P^2} \sum_p \sigma_p^2(k\ n_p(N_0)) \tag{2.31}$$

[30] On average, when population numbers are large enough, eg. $n_p > 50$.

So that the relation shown by the parts is itself valid for the aggregate:

$$log(\sigma_\epsilon^2(N_0)) - \alpha\ log(k) = log(\sigma_\epsilon^2(k\ N_0)) \quad \Leftrightarrow \quad \frac{1}{k^\alpha}\sigma_\epsilon^2(N_0) = \sigma_\epsilon^2(k\ N_0)$$

That is:

$$log(\sigma_\epsilon^2(N)) = c' - \alpha \log(N) \Leftrightarrow \sigma_\epsilon^2(N) = C' N^{-\alpha} \tag{2.32}$$

This last equation is telling us that if we plot the idiosyncratic term of $var(X)$ as a function of population sampling size N in log-log they will show a slope $-\alpha$.

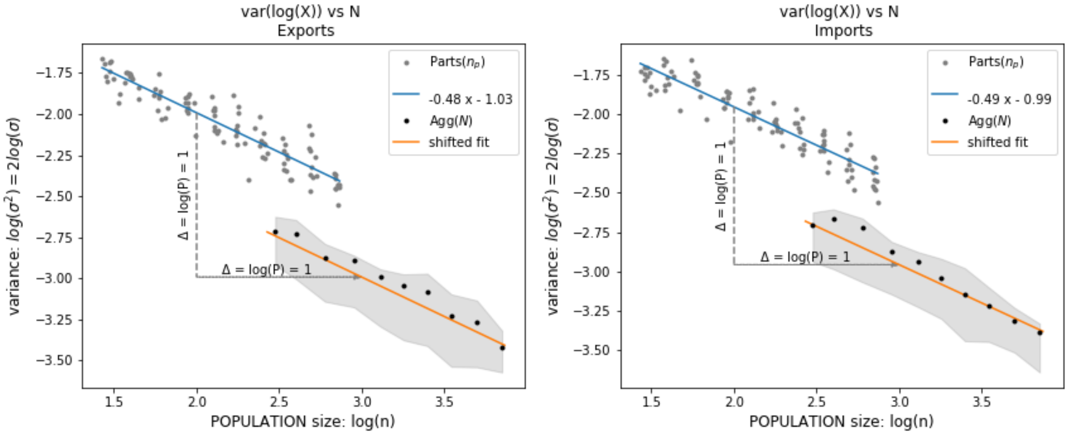


Figure 2.8: Decay of idiosyncratic volatility with population size. As explained by the equations in this section, if parts show a rate $-\alpha$ the aggregate must have this same decay rate. An OLS linear fit on the parts is shifted in $-1 = -\log(P)$ vertically because of the $1/P$ factor, and in $+1 = \log(P)$ horizontally because the aggregate has P times the parts population. The OLS on parts then fits the observations in the aggregate.

In the special case that all parts have the same variance σ_p^2, we also have that the idiosyncratic part of aggregate variance fulfills $\sigma_\epsilon^2 = P\sigma_p^2/P^2 = \sigma_p^2/P$. So that $\log(\sigma_\epsilon^2) = \log(\sigma_p^2) - \log(P)$. In our case $P = 10$, so that $\log(P) = 1$. This determines the -1 variance drop when comparing parts to aggregate in figure 2.8.

What we have done so far is expressing the idiosyncratic part of aggregate variance both as a function of total population N and as a function of parts' population n_p. We see that they both should show a common α. Empirically, the observed slope α of variance decay with population size is $\alpha_X = -0.48$ for exports data, and $\alpha_M = -0.49$ for imports data. These values were computed from parts' variances (blue lines, Figure 2.8) and can be extended to describe aggregate idiosyncratic variance by to equation 2.31 (yellow lines, Figure 2.8).

Then, so far we can measure the rate of decay of aggregate variance with population size (α). We know that the rate of decay of parts is related to the rate of decay in the aggregate. But we know nothing about *why* this slope has the value it has.

In order to reach that, we need to look at what happens in the parts themselves. This is what we will do in the following section.

2.8 Aggregating a group of agents. Sum of powers.

We have seen what happens between the parts and the total, and now its time to see what happens among the agents in the parts. So, we will take a closer look at groups of fluctuating agents.

One goal is to understand a possible relation $\sigma_p^2 = f(n_p)$ between the parts' variance and population. For example, we have seen empirically that idiosyncratic aggregate variance can be described by $\sigma_\epsilon^2 = C' N^{-\alpha}$, as in Eq. 2.32.

Another dependence to determine is that beween σ_p^2 and the moments of the log micro fluctuations (μ and $\hat{\sigma}$).

To reconstruct aggregate volatility from quantiles volatility, we will need additional information about their cross correlations (what is the same, we need to know if there are aggregate shocks). Equations 2.11 and 2.22 to 2.29 guide the path arriving at aggregate variance once we know sectoral volatilities.

For working out this problem, arrange agents into Q *quantile parts* (each denoted q), as opposed to P *random parts* (denoted p) that we have been using so far. The partition into quantile parts implies sorting agents by size before splitting them into Q parts each concentrating nearly equal values $\bar{S}_q = \bar{X}/Q$. As a consequence of sorting before splitting, agents in each quantile part would have similar sizes. This feature allows analysis to go further than if working with random parts, as will come clear soon.

If we want to know the variance that the time series of a group of firms can show, we need to first be clear on how the group's level is expressed in terms of its agents' contributions. Then we can compute the moments of the quantile parts' time series.

The difficulty is in that firm sales are given as exponential levels. That is, if we denote the sales of firm i in linear levels by s_i and in log levels by x_i, total sales of a quantile part q are:

$$S_{qt} = \sum_{i \in q} s_i = \sum_{i \in q} 10^{x_i} \tag{2.33}$$

The sales of firms s_i are observed at a specific time step t (i.e. $s_i = s_{it}$), and the same is true for the log levels x_i.[31]

Firms are taken to belong at their *zero fluctuation* level x_i^0 , and they have fluctuated to their x_i level that is actually observed. In linear scale, this zero level is $10^{x_i^0} \equiv s_i^0$. The quantile total when all its firms are at zero fluctuations is $S_q^0 = \sum_i 10^{x_i^0}$. The quantile has a single zero level for any time steps t.

To begin, consider S_{qt}/S_q^0, the ratio between the observed quantile level and the zero quantile level.

$$\frac{S_{qt}}{S_q^0} = \frac{\sum\limits_i^{n_q} 10^{x_i^0 + D(\cdot)}}{\sum\limits_i^{n_q} 10^{x_i^0}} \tag{2.34}$$

Where $D(.)$ is the distribution of log fluctuations defined as difference to mean values, as in figure 2.3. The log level of firm i observed at a time step can be denoted $x_i = x_i^0 + D(\cdot)$ (and here $D(\cdot)$ stands for a draw from such distribution).[32] Eq. 2.34 can refer to a time series or to a single observation. It is clear that it indicates the level sales of a quantile part q.

We are lucky to have had B. Mandelbrot discuss some features of sums of log-normal distributions in Mandelbrot (1997). Do not be discouraged by the name of the chapter *"A case against the lognormal distribution"*. Instead, let us look at the content. The first comment refers to the fact $10^{x_1} + 10^{x_2}$ is an ugly sum to work with: *"A more-than-counterbalancing drawback: the distributions of sums are unmanageably complicated. Dollars and firm sizes do not multiply; they add and subtract. But sums of lognormals are not lognormal and their analytic expressions are unmanageable."*. And in the comment itself there is also the reason why we will still work with

[31] We omit the index t on the firm level expressions to keep them less cluttered.

[32] It is important to not confuse them with the distribution of growth rates, as in figure 2.2. The real distributions $D(.)$ do not have a closed form and are acknowledging the accumulation of subsequent growth rates, implicitly acknowledging possible growth rates' auto correlation. The empirical $D(.)$ may usually be described through a mixture of normals. Remember if $D(\cdot)$ is eg. a normal, $10^{D(\cdot)}$ is a log-normal.

sums of log distributions in this chapter: if we accept multiplicative growth as in Bottazzi and Secchi (2006), Gibrat (1931), and H. R. Stanley et al. (1996) then we *have to* add and subtract the firm sizes (in dollars). Mandelbrot is warning us that it is not the most natural path, but the path we need is decided by the empirical system, and not chosen by us.

Note anyway that in this book Mandelbrot is referring to size distributions, and not to distributions of micro fluctuations. And even if the general problem is complicated, in the context of micro fluctuations distribution I show in this section that tracking the sums is not such a difficult problem. The so called sums of powers have also been studied for engineering applications which offer us some technical guides. Marlow (1967) shows that under general conditions (especially small coefficient of variation std/mean of the gaussian exponent) the sum of draws from a lognormal (thus also the mean) will be normally distributed.[33] And this opens the path for trying to estimate the moments of such a distribution.[34] More recent works dedicated to the situation of summing powers are Beaulieu and Xie (2004), Filho et al. (2005), and Schwartz and Yeh (1982). Their relevant range of parameters can differ from ours.

So, the strategy is to first determine the expected value of the ratio in 2.34. Then it will be easier to characterize its volatility. The variance of this ratio is approximately proportional to the variance of the log levels by a factor $\ln^2(10) \approx 5.30$ (see Section 5.5).

2.8.1 Expected level of parts' time series

To go further, note that from equation 2.34, if the total is split in enough (Q) parts all agents in each q are of about the same size s_q. This is the *narrow quantile* condition. In that limit equation 2.34 becomes:

$$\frac{S_{qt}}{S_q^0} = \frac{\sum_i^{n_q} s_q^0 10^{D(\cdot;\, n_q)}}{n_q s_q^0} = \frac{1}{n_q}\sum_i^{n_q} 10^{D(\cdot;\, n_q)} \equiv \tilde{M}_t \tag{2.35}$$

where $D(\cdot; n_q)$ denotes a draw of n_q elements from the distribution $D(\cdot)$. The sum represents a sum of all n_q elements of this draw. Here we have introduced $\tilde{M}_t$ as an estimator of the mean of $10^{D(\cdot)}$ computed from the n_q values in $10^{D(\cdot;n_q)}$.

[33] In fact the sum of draws from a lognormal can be log-normally distributed.

[34] The distribution of means of draws from a log-normal is denoted as s_N^* in the upcoming paragraphs.

Now there is a key idea that we have to consider. The value of the quantile, which in the occasion t turned out to be $\tilde{M_t}$ can be assumed to have been drawn from an underlying distribution that we can call s_D^*.

Such a distribution is assumed to have a 'true' mean value M, for which $\tilde{M_t}$is an estimator, and a variance Σ^2. Other than that, we do not know anything about s_D^*, it does not need to be normally distributed or have some closed form expression.

The first moment of s_D^* ($M \equiv E[s^*]$) is then a proxy for the levels S_{qt}/S_q^0 shown by quantile q. If we want to know what M is, we need to look at the limit:

$$M = \lim_{n\to\infty}\left(\frac{1}{n}\sum_i^n 10^{D(\cdot;\,n)}\right) = \frac{1}{n}\int n\,p(t)10^t dt = \int\ p(t)10^t dt \equiv E[10^{D(\cdot)}] \qquad (2.36)$$

where $p(t)$ is the probability density function of the distribution $D(.)$.[35] This suggests that s_q^* may be given as $s_q^* = 10^{D(\cdot)}$ for some $D(\cdot)$.[36]

Keep in mind that the expressions in eq. 2.36 are a large n limit. The $\tilde{M_t}$ we observe should converge to those levels progressively as n_q increases. Mandelbrot (1997) comments *"The population moments of a lognormal or approximate lognormal will eventually be approached, but how rapidly? The answer is: 'slowly'."*. This convergence can be evaluated graphically in Figure 2.9, (equivalent to Figure E9-2, op. cit.). Before discussing the convergence patterns in Figure 2.9, let us introduce some formal tools.

We have said that there may not be closed form expressions for the fluctuations $D(\cdot)$. Still, these tent shapes (figure 2.3) can naturally be expressed as mixtures of possibly assymetric log-normal, log-Laplace, or even fatter tail distributions. I will use the log-normal and log-Laplace distributions as clear cut benchmarks for other more general log-distributions that may appear empirically and for which we do not have an expression. In all experiments empirical distribution of micro shocks show results with features in between the log-Laplace and log-normal functions, thus justifying this choice. The conventions for defining the log-normal and log-Laplace distributions are in Appendix. They are defined such that the theoretical standard

[35] What is the same: $\lim_{n\to\infty}\left(\sum_i^n p_i 10^{D(\cdot;\,n)}\right) = \int\ p(t)10^t dt$.

[36] If firm level fluctuations are small enough, the underlying distribution s_D^* is approximately lognormal.

deviation of log shocks matches the parameter $\hat{\sigma}$.

In the derivations that follow, we may use the word *micro* to refer to agents' characteristics. The most likely value of log micro fluctuations is denoted μ and the width of log micro fluctuations are denoted $\hat{\sigma}$. These can be loosely called the micro moments. As an example, a log-Laplace distribution is $10^{L(\mu,\hat{\sigma})}$, with the definition of $L(\mu,\hat{\sigma})$ in Appendix (eq. 2.75).

Replacing $D(.) = N(\mu,\hat{\sigma})$, or $D(.) = L(\mu,\hat{\sigma})$, the moments (levels) $M_D = M_N$ or $M_D = M_L$ around which the levels of quantiles $\tilde{M}_t = S_{qt}/S_q^0$ are situated are, for log-normal micro shocks:

$$E[s^*_N] = M_N = E[10^{N(.)}] = 10^{\mu+\hat{\sigma}^2\ln(10)/2} \tag{2.37}$$

Equation 2.37 means that the level at which we observe the quantile part is $S_0 10^{\mu}$, and there is also an extra expansion of $10^{\hat{\sigma}^2\ln(10)/2}$. This expansion is quadratic on the standard deviation of firm fluctuations. Note that it will be present even when the population of agents has $\mu = 0$. That is, even if mean log growth is zero, the part will be expanded in a factor of $\frac{1}{2}\hat{\sigma}_q^2\ln(10)$ from its zero level. [37] This is one among a series of non intuitive features stemming from non linearities that we will come accross. Having introduced a zero fluctuation level S_0 different from a stationary level $\bar{S} = S_0 10^{\mu+\sigma^2 ln(10)/2}$ helps us be clear on these non linear contributions to the total and not leave any of them behind.

For log-Laplace micro shocks, following eqs. 2.75 to 2.80 of Appendix:

$$E[s^*_L] = M_L = E[10^{L(.)}] = \frac{10^{\mu}}{1-\frac{1}{2}\hat{\sigma}^2 ln^2(10)} \tag{2.38}$$

a guided derivation of these moments is in appendix. Their expressions (Eq. 2.37, 2.38) are well known but following their derivations may be helpful for understanding exactly what they mean and imply. Essentialy they are the mean of lognormal and log-Laplace distribution, but they also represent the level that comes out of averaging log-normal and log-Laplace

[37] For an intuitive approach to this effect think of the geometric effect by which multiplicative shocks (i.e. log shocks) make the mean of equal fluctuations show an expansion. For example, the mean between $1\cdot(1+\epsilon)$ and $1/(1+\epsilon)$ is $M' > 1$. Here ϵ is playing the role of $\hat{\sigma}$. Apply the example with $\epsilon = 0.05$. We have $[1.05+1/1.05]/2 = 1.00119 > 1$. Note that $M' = 1.00119 > 1$, and $ln(M') = 0.00119 \approx \epsilon^2/2 = 0.05^2/2 = 0.00125$.

This rule also means that twice as large firm level shocks lets the total go four times as far from $S_0 10^{\mu}$. In our previous example, duplicating the deviation to $2\epsilon = .1$ lets the mean between 1.1 and $1/1.1$ be $1.0045 > 1$. About four times as far from 1 than 1.0012.

nominal fluctuations.

Note that eq. 2.38 is not valid if (the denominator is zero) $\hat{\sigma} = \sqrt{2}\ln(10) \approx 0.61$. The average of log-Laplace shocks does not converge if $\hat{\sigma} > \sqrt{2}/ln(10) \approx 0.61$, when the potential expansion due to new agents overrides the averaging from law of large numbers (LLN) letting it diverge as n increases. At this point the probability of large shocks balances the $1/n$ factor in the large n limit of 2.36. Dynamics related to this will be present in almost every result we observe in this section, and it offers a nice benchmark for discussing and thinking about the problem of a population of fluctuating agents.

If we are interested in $log(S_q/S_0)$, we can use the equation 2.15 which relates the expectation of a random variable to the expectation of the log level of a random variable. If quantile part fluctuations are not excessively large $E[log(s^*_D)] \approx \log(E[s^*_D])$ so that replacing 2.37 we have

$$E[log(s^*_N)] \approx \log(M_N) = \mu + \hat{\sigma}^2 \ln(10)/2 \tag{2.39}$$

$$E[log(s^*_L)] \approx \log(M_L) = \mu + \log\left(\frac{1}{1 - \frac{1}{2}\hat{\sigma}^2 \ln^2(10)}\right) \tag{2.40}$$

and if n_q is sufficiently large $log(S_q/S_0) \approx log(M_D)$. In the limit of small $\hat{\sigma}$:

$$\log(M_L) \approx \mu + \frac{1}{2}\hat{\sigma}^2 \ln(10) + \frac{1}{8}\ln^3(10)\hat{\sigma}^4 + O(\hat{\sigma}^6) \tag{2.41}$$

so that in this limit both log-normal and log-Laplace fluctuations show a common dependence of the type: $log(S_{qt}/S_0) = \mu + \hat{\sigma}^2 \ln(10)/2$.

Note that expressions for quantile mean level (Eqs. 2.37, 2.38) and log level (Eqs. 2.39, 2.40) are in terms of the parameters of the log micro shocks distribution, μ, $\hat{\sigma}$. They determine limits at large n. To see how large n_q needs to be for $S_{qt}/S^0_q \approx M_D$, a graphical answer is in figure 2.9. If we had to summarize the situation, we can say that the mean of empirical fluctuations (average $\hat{\sigma} = 0.49$) does not diverge as it happens with the log-Laplace shocks with $\hat{\sigma} > 0.61$, although it does shows a convergence slower than mean of log-normal fluctuations with the same $\hat{\sigma}$. The benefit of the exercise in 2.9 is that we can evaluate the convergence of the means along the range of n_q parameters that are relevant to the problem.

The plots in figures 2.9, 2.10 and 2.11 summarize the dependence of $E[log(S_{qt}/S_q^0)]$ on the population size n_q, mean log micro fluctuation μ and width of log micro fluctuations $\hat{\sigma}$, superposing computations with the equations 2.39 and 2.40 (in red). Plots on the left and right show the results when micro shocks are log-normal and log-Laplace respectively and those in the middle show the result of using the empirically observed distribution of micro fluctuations. For details and guides regarding the computational exercise refer to the Appendix.

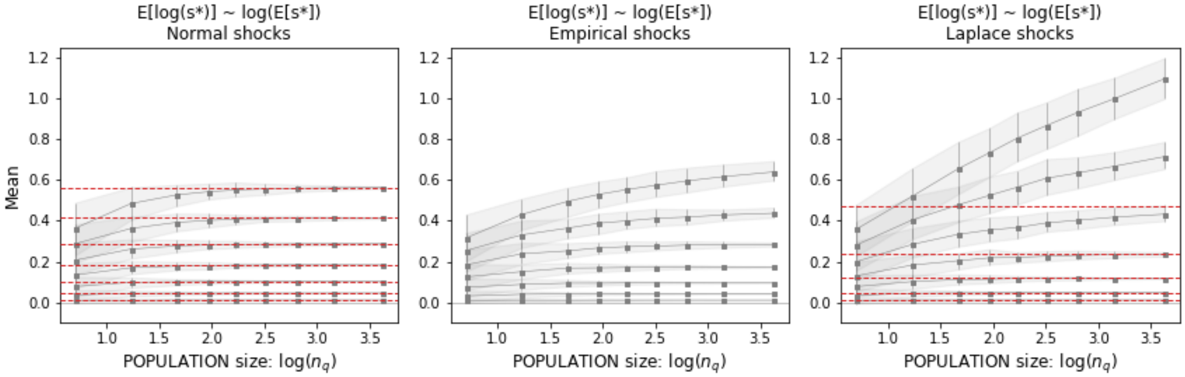


Figure 2.9: As a function of population n_q, expectation of the log of quantile levels for various widths of micro shocks $\hat{\sigma}$ and $\mu = 0$. log-normal (left) empirical (mid) and log-Laplace (right) micro shocks. With increasing n_q we see the convergence of mean to values of eqs. 2.39 and 2.40 (red), especially when micro log shocks are gaussian. The values of $\hat{\sigma}$ are in the range 0.1 to 0.7. log-Laplace is seen to not converge with increasing n_q if $\hat{\sigma} > 0.61$. Empirical shocks show an intermediate scenario in between gaussian and Laplace log shocks.

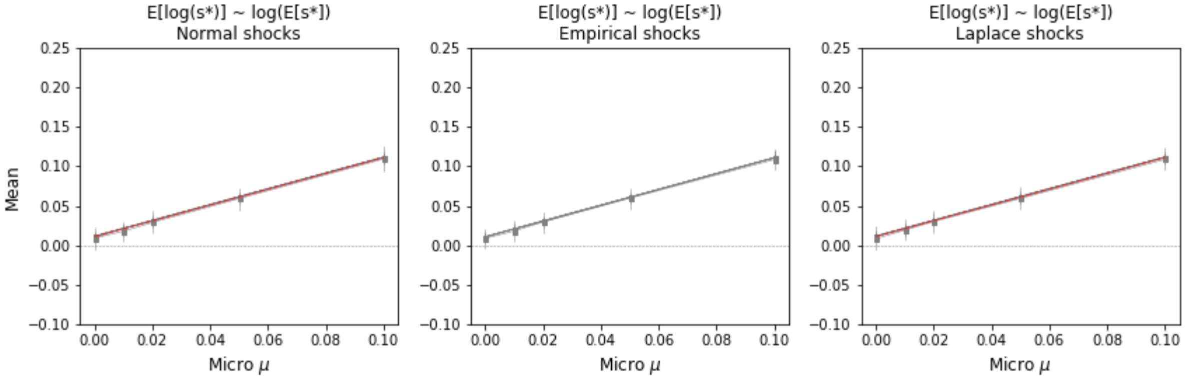


Figure 2.10: As a function of μ, expectation of the log of quantile levels for various n_q and $\hat{\sigma} = 0.1$. log-normal (left) empirical (mid) and log-Laplace (right) micro shocks. On red, equations 2.39 and 2.40 with a linear dependence of slope 1. μ is varying in the range 0.0 to 0.1.

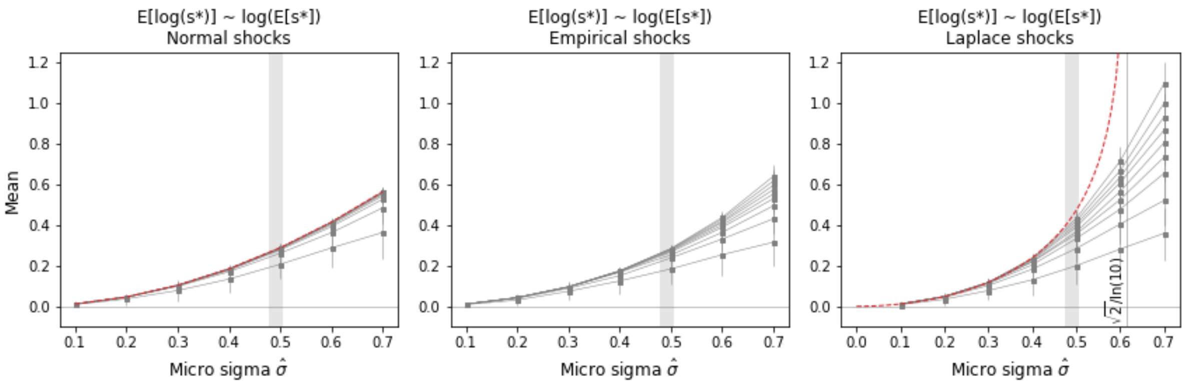


Figure 2.11: As a function of $\hat{\sigma}$, expectation of the log of quantile levels for various n_q and $\mu = 0$. log-normal (left) empirical (mid) and log-Laplace (right) micro shocks. On red, equations 2.39 (log-normal shocks) with a quadratic dependence, and 2.40 (log-Laplace shocks) with terms of order $o(\hat{\sigma}^4)$ and higher in addition to the terms already present for log-normal micro shocks. The expectation in this last case diverges if $\hat{\sigma} > 0.61$, and adding more agents we will not be able to average the quantile level. The average magnitude of micro fluctuations is denoted by the gray vertical band. $\hat{\sigma}$ is varying in the range 0.1 to 0.7.

2.8.2 Variance of parts' time series mean (the law of large numbers)

We have postulated that the levels $\tilde{M}_t = S_{qt}/S_0$ that a quantile q shows are drawn from a hypothetical distribution s^*_D. We do not know nothing of this distribution, although we could have an expression for its 'true' mean M_D to which we approached by averaging the levels shown by n_q agents.

What is the variance of this mean $\tilde{M}_t$? In the following step, take $\tilde{M}_t$ and $10^{D_{it}}$ respectively as time series of T realizations of the observed average $\tilde{M}_t$, and observed fluctuations D_{it} of the agent i at time step t.

$$var[\tilde{M}_t] = var[\frac{1}{n_q}\sum_i^{n_q} 10^{D_{it}}] = \frac{1}{n_q^2} var[\sum_i^{n_q} 10^{D_{it}}]$$

If the fluctuations belonging to each i are uncorrelated (i.e. if $cov[10^{D_{it}}, 10^{D_{jt}}] = \delta_{ij} var[10^{D_{it}}]$) and if the variance that each of these agents i are showing are of about the same magnitude $var[10^{D(\cdot)}]$, then $var[\sum_i^{n_q} 10^{D_{it}}] = n_q var[10^{D(\cdot)}]$, and the following applies:

$$var[\tilde{M}_t] \approx \frac{1}{n_q^2}\sum_i^{n_q} var[10^{D_{it}}] = \frac{1}{n_q^2} n_q var[10^{D(\cdot)}] = n_q^{-1} var[10^{D(\cdot)}] \tag{2.42}$$

This is a 'law of large numbers' situation.[38]

[38] To compare with the developments in Dupor (1999), this paper examines the volatility of aggregates from a

Computational tests show the variance of parts' mean instead follow the more general expression:

$$var[\tilde{M_t}] = n_q^{-\alpha} var[10^{D(\cdot)}] \tag{2.43}$$

with $\alpha \leq 1$. Which may be taken as a 'postponement' of such law of large numbers.

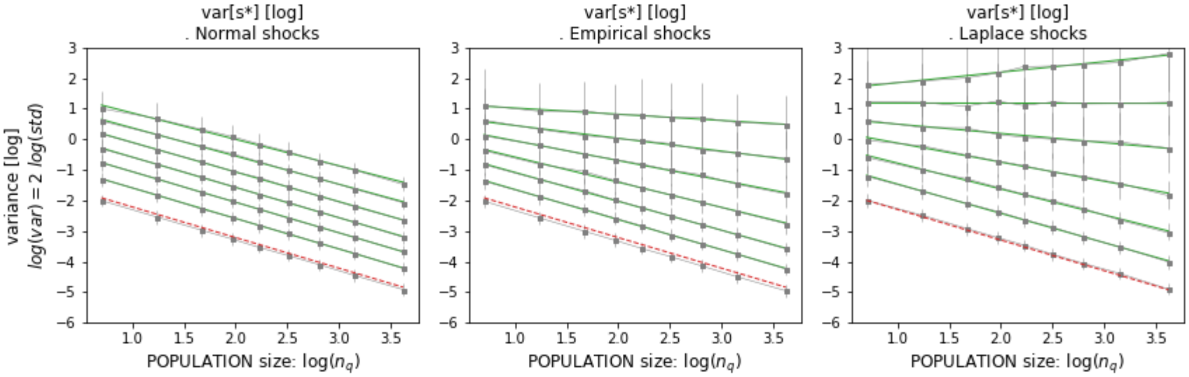


Figure 2.12: Variance of quantile levels as a function of n_q, for various levels of micro fluctuations $\hat{\sigma}$ and $\mu = 0$. log-normal (left) empirical (mid) and log-Laplace (right). In the limit of small fluctuations the LLN applies (red). As $\hat{\sigma}$ grows, variance decay with population size is milder, although still dominant. This is only broken by log-Laplace shocks with $\hat{\sigma} > \sqrt{2}/ln(10) \approx 0.61$, in which case the more agents one averages the more noisy the averages get.

This expression of the variance of mean of part q vs. part's population n_q as power law with exponent α is in line with the models of volatility vs population size of eqs. 2.30 and 2.32 in section 2.7.1. In that section we saw that if parts' follow such a power law, then the aggregate inherits an average of their rate of decay. A power law for parts as in eq. 2.43 translates to analogous 'large number postponement' power law for the idiosyncratic part of aggregate variance.

model as in Long and Plosser (1983). These aggregates are capital (k) and also output, with equivalent results for any of them. Autoregressive dynamics are managed by accounting variance in the frequency domain (see eqs 2.63, 2.64 and 2.65 in Appendix) and recurring to eigenvectors of functions of the input output matrix. Irrespective of all this, the relevant result in the paper is a $1/n$ variance decay.

In this paper, the number of sectors is n. The log level of a sector, is denoted k. There is an "aggregate statistic" $\bar{k} = \mathbf{1}^T k \sim n\ k = n\ \log(S_p)$. This statistic is inconvenient to interpret, as it is the sum of log levels of sectors. There is another aggregate statistic $\tilde{k} = \bar{k}/n$. Let us denote an average as $\langle \cdot \rangle$. Then

$$\bar{k}/n \equiv \langle \bar{k} \rangle$$

and $\tilde{k}$ should be identified with $\log(S_p)$. The variance of $\tilde{k}$ is postulated to depend as $var_\omega[\tilde{k}] = S_{\tilde{k}}(\omega) \propto n^{-1} f(\omega)$. This is equivalent to $var[S_{qt}/S_q^0]/ln^2(10) = var[\tilde{M_t}]/ln^2(10)$ (see eq. 2.42). Although note that he is referring to sectors and not the number of agents.

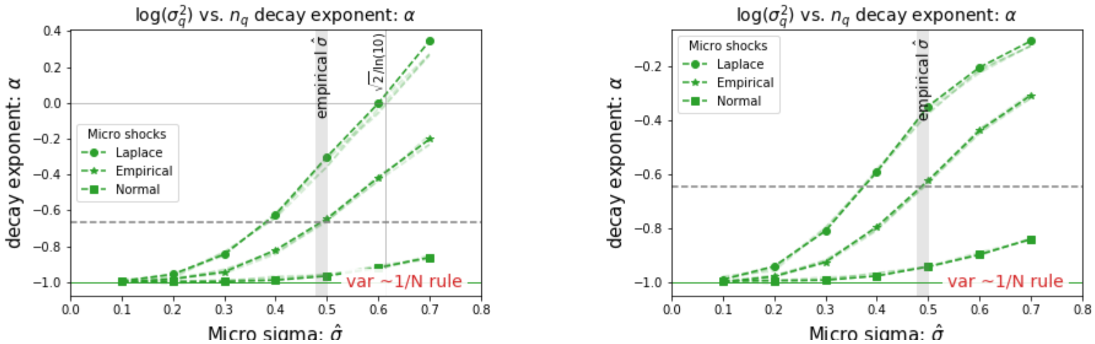


Figure 2.13: Decay rate of quantile variance with populations size as a function of width of micro shocks $\hat{\sigma}$. Left: $var[S_{qt}/S_q^0]$. Right: $var[log(S_{qt}/S_q^0)]$. The bottom level implies fast law of large number convergence. Empirical level of micro fluctuations are in vertical band. They suggest a rate $-\alpha \approx -0.64$ given the empirical distribution of micro shocks.

There is a problem that has not been clarified so far, what is the mechanism that lets α be smaller than 1? From results plotted in figure 2.13 we are in a position to understand the origin of this feature. In this figure the value of the parameter $-\alpha$ is plotted as a function of the width of micro fluctuations $\hat{\sigma}$. Curves stand for log-normal, empirical and log-Laplace micro fluctuations.

The line in the bottom is the $-\alpha = -1$ level, which essentially is the naive diversification rule by which variance falls as $1/n_q$ with the number of agents. As $\hat{\sigma}$ increases from small values $\hat{\sigma} \ll 1$ the rate of decay α starts to depart from the $-\alpha = -1$ level. These departures are stronger if micro shocks are fat tailed (eg. log-Laplace as opposed to log-normal) although a log-normal does generate them. In the empirical scenario, firm level shocks are large $\hat{\sigma} \approx 0.5$ and thus $-\alpha \approx -0.6$. Curves for various mean micro shocks μ are mostly overlapping suggesting little to null dependence on this parameter.

There is still an open front however. Why would this departure appear as a milder power law? A good answer to that will probably be developed in future studies, but there is a hint that we can already profit from.

Remember the condition for arriving at the 'large numbers' $-\alpha = -1$ expression of equation 2.42 was that there were no correlations among the time series $10^{D_{it}}$ of agents i belonging to part q. But should one allow non zero net cross covariances among the firms belonging to a part? If we do, eq. 2.43 should alternatively be:

$$var[\tilde{M_t}] = n_q^{-\alpha} var[10^{D(\cdot)}] = \underbrace{cov(10^D)}_{\sigma_M^2} + \frac{1}{n_q} var[10^{D_{it}}] \tag{2.44}$$

For some σ_M^2 of which we know nothing, but which we can see occupies the place of a comovement. In this equation I used $cov(10^D)$ to denote covariance terms among agents which would come as a function $f(\mu, \hat{\sigma})$ of the micro moments of log deviations the same way that $var[10^D]$ does.

The idiosyncratic variances add a contribution $var[10^{D(\cdot)}]$ for each agent. The comovement part adds a contribution $\tilde{\sigma}_M^2 \equiv n_q \sigma_M^2$ per agent.

$$var[\tilde{M_t}] = \frac{1}{n_q} (\underbrace{n_q cov(10^D)}_{\tilde{\sigma}_M^2} + \underbrace{var[10^{D_{it}}]}_{\sigma_E^2}) \tag{2.45}$$

Adopting this convention means comparing the magnitudes of the comovement and $var[10^D]$ contributions on an equivalent, per agent basis. This is convenient because it allows a neat relation between the two contributions:

$$n_q^{1-\alpha} = 1 + \frac{\tilde{\sigma}_M^2}{var[10^{D_{it}}]} \quad \Leftrightarrow \quad (1-\alpha)\log(n_q) = \log\left(1 + \frac{\tilde{\sigma}_M^2}{var[10^{D_{it}}]}\right) \tag{2.46}$$

This relation is plotted in figure 2.14. This figure lets us understand the balance between the self-variance per agent ($var[10^{D_{it}}]$) and the covariance to all other agents per agent ($n_q cov(10^D)$) as a function of the part's population n_q.

The slope of the lines is α, the parameter of decay of variance with population size. As we have said, this α is relevant especially when micro fluctuations are fat tailed and departs from 1 as a function of the width of micro fluctuations $\hat{\sigma}$. In horizontal gray lines I show when $\tilde{\sigma}_M^2 \ll var[10^{D_{it}}]$ (lower line), when $\tilde{\sigma}_M^2 \approx var[10^{D_{it}}]$ (middle line) and when $\tilde{\sigma}_M^2 \gg var[10^{D_{it}}]$ (upper line). So that the plot is essentially telling us, if a comovement as in 2.45 explains the departure from LLN, how large is the covariance of an agent to all others compared to the variance of an agent itself.

In groups of few largest firms, the convergence of the mean is limited by self variance

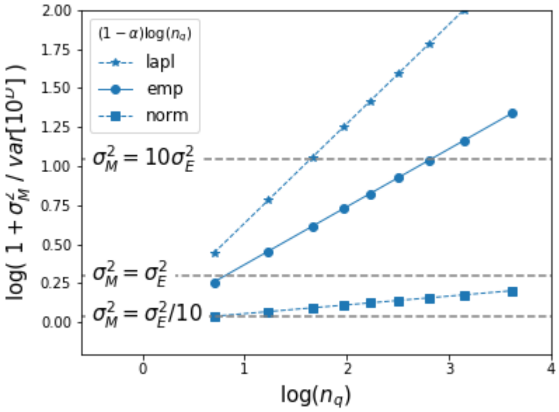


Figure 2.14: The balance between the terms that make and break the law of large numbers: variance $var[10^D]$ and covariance per agent $\tilde{\sigma}_M^2 = n_q cov(10^D)$. Their relation is plotted as a function of population size n_q. The slopes are the rate of variance decay with population α, given each of the micro shocks distribution, in this case at the level of micro fluctuations $\hat{\sigma} = 0.5$ (close to their empirical magnitude). This plot can tell us how important are the comovements among firms of a quantile given its population. For our problem, the quantile of largest firms has $n_q < 10$ and $\tilde{\sigma}_M^2 \approx \sigma_E^2$ in that case. In large groups of smaller firms instead the fact that more agents add more net contributions $\tilde{\sigma}_M^2 = n_q cov(10^D)$ is more important than the averaging of $var[10^D]$ itself with higher sample size.

as well as net contributions from comovement (smallest $\tilde{\sigma}_M^2 \equiv n_q cov(10^D) \approx var[10^D]$). In the parts made of many small agents, most of the contribution to $var[\tilde{M}_t]$ is on the form of comovement among agents. Still, the decay force of large number averaging $1/n_q$ does result in shrinking volatility of groups of many small agents and groups of few large agents are still more volatile.

2.8.3 $var[10^{D_{it}}]$ as a function of micro moments μ, $\hat{\sigma}$

To advance further we can use expressions for the moments of the distributions of log shocks into $var[10^{D(\cdot)}] = E[10^{2D(\cdot)}] - E^2[10^{D(\cdot)}]$ and by equation 2.43, express the variance of mean shown by a quantile part in terms of the moments of the micro distribution of shocks, μ, $\hat{\sigma}$. Equation 2.43 becomes:

$$var[\tilde{M}_t] = n_q^{-\alpha} var[10^{D(\cdot)}] = n_q^{-\alpha} 10^{2\mu} f(\hat{\sigma}) \approx n_q^{-\alpha} 10^{2\mu} \left(\hat{\sigma}^2 + o(\hat{\sigma}^4)\right) \tag{2.47}$$

The $f(\hat{\sigma})$ are functions of the moments of the distribution of micro deviations. This is developed in appendix for the ideal cases of log-normal and log-Laplace micro shocks. For log-normally distributed firm level shocks, $D(\cdot) = N(\mu, \hat{\sigma})$, and (see eqs. 2.67, 2.68, 2.69):

$$var[10^{N(\cdot)}] = 10^{2\mu + \hat{\sigma}^2 \ln(10)}(10^{\hat{\sigma}^2 \ln(10)} - 1) \tag{2.48}$$

In the limit of very small micro fluctuations:

$$var[10^{N(\cdot)}] \approx 10^{2\mu} \left(\hat{\sigma}^2 + \frac{3}{2}\hat{\sigma}^4 + o(\hat{\sigma}^6) \right) \tag{2.49}$$

For log-Laplace distributed firm level shocks, $D(\cdot) = L(\mu, \hat{\sigma})$. A calculation of the moments of the log-Laplace is in Appendix (Eqs. 2.76 to 2.83):

$$var[10^{L(\cdot)}] = 10^{2\mu} \left(\frac{1}{1 - 2\hat{\sigma}^2} - \frac{4}{(4 - \hat{\sigma}^2)^2} \right) \tag{2.50}$$

In the limit of small micro fluctuations:

$$var[10^{L(\cdot)}] \approx 10^{2\mu} \left(\hat{\sigma}^2 + \frac{13}{4}\hat{\sigma}^4 + o(\hat{\sigma}^6) \right) \tag{2.51}$$

So that $var[10^{D(\cdot)}]$ shows the same dependence with moments of the microshocks both for lognormal or log-Laplace distributions only for small fluctuations. That is why in a sense being expressable as 10^D is a key feature for micro fluctuations, regardless of whether D is a normal or a fat tails distribution.

These expressions for $var[10^{D_{it}}]$ (eqs. 2.48 to 2.51) would let us express $var[\tilde{M_t}]$ in function of the micro moments μ and σ. For the moment I have not worked the correct expression in terms of micro moments for a term like $cov(10^D)$. Still notice by looking at eq. 2.44 that we should expect it to be of the type: $cov(10^D) = 10^{2\mu} f(\hat{\sigma})$.

The σ^2 contribution (equations 2.49 and 2.51) is the one showing cancellation of opposite shocks and convergence of the mean as when we average a time series showing additive deviations from a level. Both log-normal and log-Laplace shocks contain this $\hat{\sigma}^2$ dependence, although these multiplicative micro shocks have additional higher order terms $o(\hat{\sigma}^4)$ that grow up from zero as micro fluctuations are turned on. These are the nonlinearities that make multiplicative shocks different from additive gaussian shocks. Note they are stronger if micro

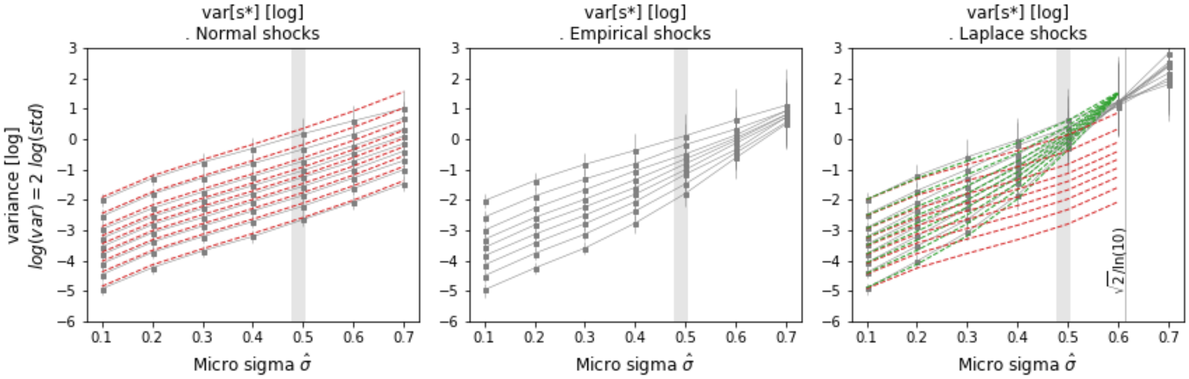


Figure 2.15: Variance of quantile levels as a function of width of micro fluctuations $\hat{\sigma}$, for various population sizes n_q and $\mu = 0$. log-normal (left) empirical (mid) and log-Laplace (right). The contribution from self variance following the $1/n_q$ rule (eqs. 2.48, 2.50, into 2.47 with $-\alpha = -1$, red). In green, for the log-Laplace case, acknowledgement of comovements as product of micromoments with population size n_q (2.50, into 2.47 with observed $-\alpha$). Magnitude of empirical $\hat{\sigma}$ is shown with vertical gray band.

shocks are fat tailed (log-Laplace).

Note however that non linearities adding to variance is one story, and the law of large numbers (and its 'postponement') is another story. They are different features that get combined. Note that in 2.44 we could replace each of equations 2.48, 2.49if we wanted to study the case of log-normal (eqs. 2.50, 2.51 for log-Laplace) micro shocks. For brevity, let me use the expressions in the case of small fluctuations which apply to any distribution of micro shocks if $\hat{\sigma}$ is small enough:

$$var[\tilde{M_t}] = n_q^{-\alpha} var[10^{D(\cdot)}] = 10^{2\mu}\left(f(\hat{\sigma}) + \frac{1}{n_q}\left(\hat{\sigma}^2 + o(\hat{\sigma}^4)\right)\right) \tag{2.52}$$

At the moment I did not develop the expression for $f(\hat{\sigma})$, but it is likely that its first term is of order $o(\hat{\sigma}^4)$.[39] If we kept only the $o(\hat{\sigma}^2)$ terms, we are in the linear setting. In such case: $var[\tilde{M_t}] = 10^{2\mu}\hat{\sigma}^2/n_q$. Note, that if it was not for the comovement term (eg. hypothetically set $f(\hat{\sigma}) = 0$), the nonlinearities by themselves would increase the variance that agents are having although still the law of large number would be there: $var[\tilde{M_t}] = 10^{2\mu}(\hat{\sigma}^2 + o(\hat{\sigma}^4))/n_q$. In

[39] For example, a covariance term can be of the type:
$cov = \frac{1}{4}10^{2\mu}\left((10^x - 1)^2 + 2(10^x - 1)(10^{-x} - 1) + (10^{-x} - 1)^2\right)$
With linear fluctuations $10^x - 1 \approx 1 + \epsilon - 1 = \epsilon$ and $10^{-x} - 1 \approx 1 - \epsilon - 1 = -\epsilon$ and the covariance above would be $\epsilon^2 - 2\epsilon^2 + (-\epsilon)^2 = 0$. Instead, in the nonlinear case this is:
$cov = 4\sinh^4\frac{x\ln(10)}{2} \approx \frac{1}{4}(x\ln(10))^4 + o(x^6)$
so that there is a leading quartic term. These are the terms that get turned on with large symmetric multiplicative fluctuations.

this sense, nonlinearities by themselves are not the explanation to milder convergence to the mean with larger numbers. The breaking of the large numbers rule is due to the comovements across agents and it is turned on by non linearities (that in the additive gaussian case are null) that wake up the comovement terms per agent which are n_q times stronger than self agent variance $var[10^D]$.

By coming down this path we were able to see the origin of the departure from LLN, and describe the situation formally in terms of the parameters of populations of agents. With this, there are already certain conceptions established in previous research that will need to be re analyzed and there are new things to think about.

The interpretation of 'postponement' in LLN as firm to firm comovement, and the neat relation that lets us balance it against the magnitude of self variance shown by agents are some connections worth exploring further in new studies.

This paper is now slowly coming to a closure, first by including some additional results which may be useful (moments of log levels of quantiles) and in subsequent sections by testing robustness to generalizations regarding size distributions, covering the accounting of extensive margins and exploring the elements of cross covariance matrices.

2.8.4 Moments of log quantile levels

So far we have worked with $\tilde{M_t} \equiv S_{qt}/S_q^0$, and there is a hypothetical distribution s_D^* from where S_{qt}/S_0 values are drawn. If we were interested in the $var[log(S_{qt}/S_q^0)]$ (which is also the $var[log(S_{qt})]$ because S_q^0 is a fixed level) as opposed to $var[S_{qt}/S_q^0]$ on the linear levels we have looked at so far we can use the relation in equation 2.16, which tells us:

$$var[log(\tilde{M_t})] \approx \frac{var[\tilde{M_t}]}{\ln^2(10)E^2[\tilde{M_t}]} \tag{2.53}$$

The variance of log levels is the one that we identify with σ_q^2 in previous sections (as in eqs. 2.21, 2.22, 2.23, 2.28, 2.29).

For the case of log-normal fluctuations, replacing the expressions for expected value and variance of equations 2.48 and 2.37 into equation 2.53, and incorporating eq. 2.44, we have:

$$var[log(\tilde{M_t})] \approx \frac{var[S_{qt}/S_q^0]}{\ln^2(10)E^2[S_{qt}/S_q^0]} = n_q^{-\alpha}\frac{10^{2\mu+\hat{\sigma}^2\ln(10)}(10^{\hat{\sigma}^2\ln(10)}-1)}{\ln^2(10)10^{2\mu+\hat{\sigma}^2\ln(10)}} = n_q^{-\alpha}\frac{10^{\hat{\sigma}^2\ln(10)}-1}{\ln^2(10)} \tag{2.54}$$

In the limit of small fluctuations this is:

$$var[log(\tilde{M_t})] \approx n_q^{-\alpha}\left(\hat{\sigma}^2 + \frac{1}{2}\hat{\sigma}^4 ln^2(10) + o(\hat{\sigma}^6)\right) \tag{2.55}$$

We can repeat analogous steps for an equivalent result applying to log-Laplace fluctuations. It may be helpful to use $var[x]/E^2[x] = (E[x^2]/E^2[x]) - 1$, and apply the expressions for kth moment of a log-Laplace (eqs. 2.81 in Appendix) already acknowledged in equations 2.38 and 2.50. Relations analogous to the ones in equations 2.54 and 2.55 are:

$$var[log(\tilde{M_t})] \approx \frac{var[S_{qt}/S_q^0]}{\ln^2(10)E^2[S_{qt}/S_q^0]} = \frac{n_q^{-\alpha}}{\ln^2(10)}\left(\frac{(1-\frac{1}{2}\hat{\sigma}^2\ln^2(10))^2}{1-2\hat{\sigma}^2\ln^2(10)} - 1\right) \tag{2.56}$$

In the limit of small fluctuations equation 2.56 becomes:

$$var[log(\tilde{M_t})] \approx n_q^{-\alpha}\left(\hat{\sigma}^2 + \frac{9}{4}\hat{\sigma}^4 ln^2(10) + \frac{9}{2}\hat{\sigma}^6 ln^4(10) + o(\hat{\sigma}^8)\right) \tag{2.57}$$

Where again we see that the dependence for small fluctuations goes as $\hat{\sigma}^2$ as with log-normal fluctuations, but the non linear terms are more than four times as large.

2.9 Acknowledging size distribution

Results in the preceding section were derived under the condition that all firms in a part are of the same size. This *narrow quantile* condition implies abstracting away from the sizes of agents. In that context we have seen that the mean value and variance of mean that a group of agents presents can be approximated by functions of the moments of the distribution of micro shocks, μ and $\hat{\sigma}$. The variance of mean has been seen to also follow a $n_q^{-\alpha}$ dependence with the number of agents in the quantile part. But how are these results changing if we acknowledge firms in a quantile part are not all of the same size?

A simple and decisive way to test the robustness of these results is to repeat the computations in a generalized setting. Therefore I repeat the experiments with firm sizes given by three ideal size distributions with parameters matching those observed in the population of French traders. We have seen that outcomes involving empirical growth rates are always in between the outcomes with log-Laplace shocks and log-normal shocks. For simplicity I apply fluctuations given by these two models.

Firm levels in the zero fluctuations data are given by levels of the cummulative density functions (CDF) of the following distributions:[40]

- x $\sim$ Lognormal($\mu = 4.54, \sigma = 1.28, N$), for $x > 3$

where the $x > 3$ is a clipping that imitates the censoring of actual data. [41]

Next size distribution we consider is the upper tail of the previous distribution, which contains 90% of the value. That is:

- x $\sim$ Lognormal($\mu = 4.54, \sigma = 1.28, N'$), for $x > x(q_{1/10}) = 6.67$

We also consider the Pareto distribution that fits this upper tail.

- x $\sim$ Pareto($z_0 = -1.10, s_m = 6.37, N'$), for $x > x(q_{1/10}) = 6.67 \equiv x_m$

The resulting dependence of mean of quantile levels, and variance of this mean is summarized in the plots of figure 2.16. Top plots show mean level of the quantiles as a function of quantile population $log(n_q)$, bottom plots are for variance of mean of the quantiles as a function of quantile population $log(n_q)$. Left side plots show results applying log-normal shocks and right side plots show results applying log-Laplace shocks. In all cases, mean micro fluctuation parameter is set to $\mu = 0$, and each of the many curves are for multiple possible values of the width of micro fluctuations $\hat{\sigma}$, in increasing levels from nearly zero up to their actual magnitude $\hat{\sigma} \approx 0.5$.

[40] The parameters of these distributions match the empirical distribution of exports by firm. The experiment is repeated on analogous synthetic zero fluctuation data matching parameters of imports at the firm level, with equivalent results.

[41] The empirical size distribution of French trading firms is bound from below at $x = 3$, i.e. $s = 1000 EUR$. Because most value is on the other end of the distribution, this is not critical for results. However, for examples of how this type of censoring feature could be treated in general, see Yamamoto (2014) where moments of a clipped log-normal are studied for an application on a practical problem. Computational results necessarily incorporate this feature.

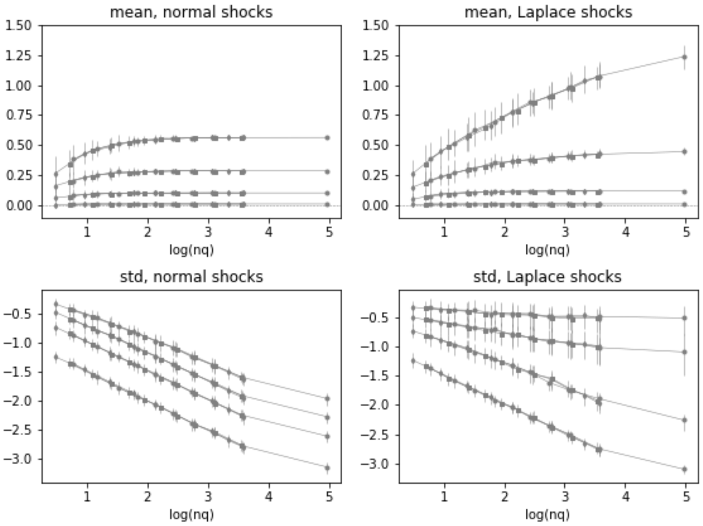


Figure 2.16: Mean and standard deviation of group of agents appear clearly as functions of $\hat{\sigma}$, n_q. The details of size distribution do not matter significantly. Top plots are for mean of log quantile levels for log-normal (top left, 2.39) and log-Laplace (top right, 2.40) micro shocks of many widths. Bottom plots for variance of quantile log levels as function of population size for log-normal (bottom left, 2.54) and log-Laplace (bottom right, 2.56) micro shocks of many widths.

The outcomes suggest that $E[s^*_D]$ and $var[s^*_D]$ are functions of the micro moments μ, and $\hat{\sigma}$, but most importantly, they are showing that these moments of the quantile time series are functions of the population n_q that apply regardless of what is the size distribution. Changing the size distribution changes the set of n_q values describing quantile population, but the functions derived in the previous section always apply in the same way.

The first of the size distributions used in this experiment includes the large number of smallest firms that accumulate 10% of the value. The results suggest that this sub population of firms qualitatively follows the same pattern as the remaining parts. The matching however is not complete, so that it may be advisable to follow this quantile (which in the end includes the majority of agents) in a separate account. In practical applications this small firms quantile can be left out because it weighs only a minority of total value, but this depends on the intended application.

2.10 Acknowledging entry and exit events

So far we have worked in settings where there is no entry or exit of firms. When considering the largest firms comprising 90% of the value this is reasonable because the ample majority of these firms are present throughout the time series, especially if the time span of these time series is not too long.

In general however, firms are entities that can become active or inactive over time and it is important to extend the formal framework to account for them.

Consider the following decomposition of sectoral sales time series S_{pt}:

$$S_{pt} = \bar{S}_p + \Delta S_{pt} = S_p^0 + B_{pt} + M_{pt} + E_{pt} \tag{2.58}$$

Here $\bar{S}_p$ is the observed mean of the time series and ΔS_{pt} are the observed deviations from such level. The term B_{pt} accounts for nominal fluctuations that exist still when all firm level fluctuations are zero. As such it can essentially be used to account for fluctuations due to entry and exit events. [42]. The terms M_{pt} and E_{pt} capture contributions related to comovements and sectoral idiosyncracies respectively. Note that the relation in 2.58 implies $S_{pt} - S_p^0 = B_{pt} + M_{pt} + E_{pt}$. So that contributions B, M, E set the difference between S_p^0 and S_{pt}. The zero level S_p^0 is that observed when removing all fluctuations, and note that in such case $S_p^0 = \bar{S}_p$ although that does not mean that $S_p^0 = \bar{S}_p$. This is because $\bar{S}_p$ includes the net mean of the B, M, E terms in it.

Because $X_t = \sum_p S_p$ we have $\sigma^2(X_t) \equiv \sum\limits_{i,j \in P} Cov(S_i, S_j)$ (equation 2.6). And analogous to equation 2.8 the element i, j of this sum will be made of 3×3 terms:

$$\begin{aligned} cov(B_{it} + M_{it} + E_{it}, B_{jt} + M_{jt} + E_{jt}) = & cov(B_{it}, B_{jt}) + cov(B_{it}, M_{jt}) + cov(B_{it}, E_{jt}) \\ & + cov(M_{it}, B_{jt}) + cov(M_{it}, M_{jt}) + cov(M_{it}, E_{jt}) \\ & + cov(E_{it}, B_{jt}) + cov(E_{it}, M_{jt}) + cov(E_{it}, E_{jt}) \end{aligned} \tag{2.59}$$

Without loss of generality log fluctuations of sales of some part p can be decomposed in an equivalent way:

[42] An analogous term can of course be used to account for other fluctuations not related to firms, for example a nominal drift which would appear as an exponential B_t component. I do not emphasize this possibility but the mathematical framework is useful to account for it.

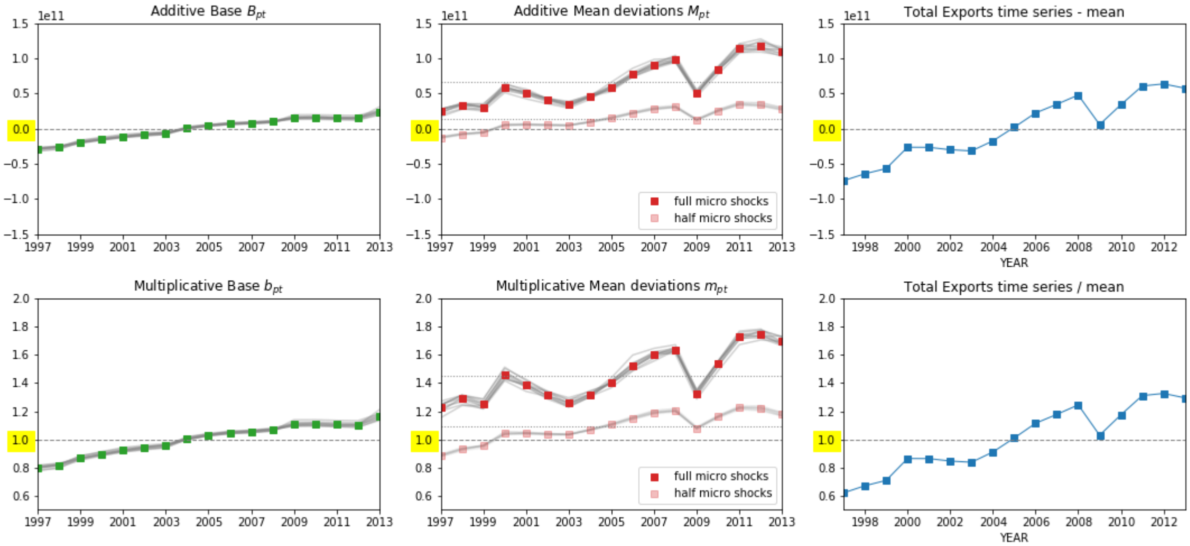


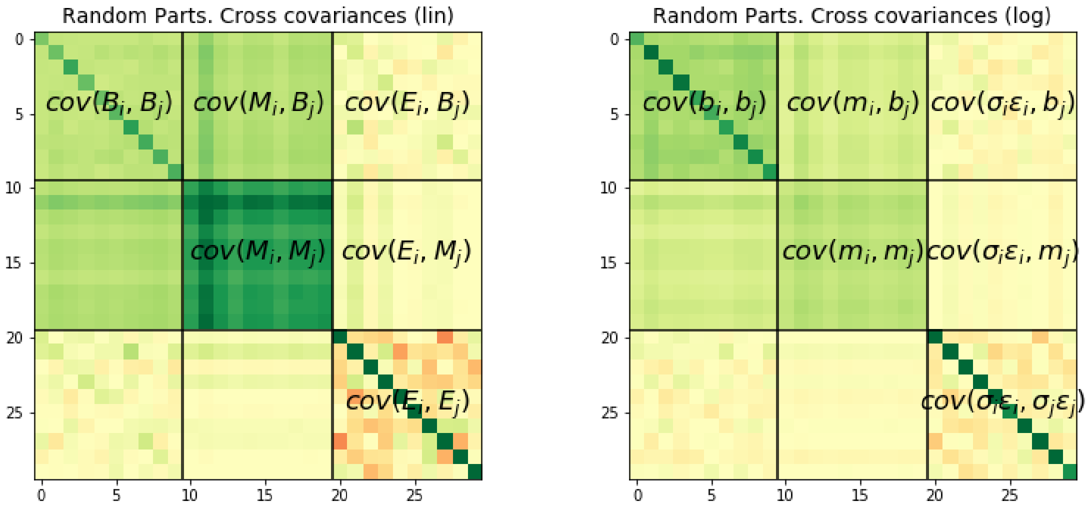
Figure 2.17: Base components B_{pt} (green, left), Mean fluctuation components M_{pt} (red, middle) estimated as the median levels across 100 bootstrap samples, and actual total exports time series. Top plots are for the additive BME decomposition (equation 2.58) and bottom for the multiplicative decomposition (equation 2.60). In the middle plots, results for micro shocks of half their empirical magnitude ($\hat{\sigma}$) lets parts time series be a quarter as far from the baseline level.

$$log(S_{pt}) = log(\bar{S}_p) + F_{pt} = log(S_p^0) + b_{pt} + m_{pt} + \sigma_p \epsilon_{pt} \tag{2.60}$$

where ϵ_{pt} as a random variable with mean zero and std = 1. This equation implies: $log(S_{pt}/S_p^0) = b_{pt} + m_{pt} + \sigma_p \epsilon_{pt}$. The relation to nominal accounts presented before therefore is: $S_{pt} = S_p^0 + B_{pt} + M_{pt} + E_{pt} = S_p^0 10^{b_{pt}+m_{pt}+\sigma_p \epsilon_{pt}}$. The lowercases are not simply the uppercases in log scale, although they do account for the same types of sources of volatility.

Consider a thought experiment where we 'turn on' fluctuations to individual agents from zero to their actual empirical magnitudes. At the start $log(S_{pt}/S_p^0) = b_{pt}$ as shocks are turned on we observe a different $log(S'_{pt}/S_p^0)$. These differences are the sum of all micro shocks and $log(S'_{pt}) - log(S_{pt}) = \delta_{pt} = m_{pt} + \sigma_p \epsilon_{pt}$. m_{pt} should be taken as a mean value around which the fluctuations of part p are observed, so that $cov(m_{pt}, \epsilon_{pt}) \approx 0$. The width of this part's time series is $\sigma = \sqrt{\sigma_\delta^2 + \sigma_b^2}$. If fluctuations are shut off then $\sigma \approx \sigma_b$. If instead the micro fluctuations are so large to dwarf b_{pt}, then $\sigma \approx \sigma_\delta$. So that depending on the magnitude of micro shocks, and the incidence of firms entry, exit or merger, any of these sources may be the explanation to observed aggregate volatility.

From the components we have proposed for $log(S_{pt})$, if sectoral fluctuations are mild


Random Parts. Cross covariances (lin)

Random Parts. Cross covariances (log)

(a)

(b)

Figure 2.18: Expected values of the cross covariance matrices as in equations 2.59 and 2.61. Aggregate variance is the sum of all elements of the linear cross covariance matrix (eq. 2.5) and the log matrix fulfills eqs. 2.22 and 2.23. Red-Yellow-Green colors denote negative, null, positive values. Left: log cross covariance. Right: linear cross covariance. Extensive margin (top left) was accounted separately and shows a constant plus diagonal (comovement plus idiosyncratic) structure. The intensive part is separated into a comovement components (mid) estimated by medians across bootstrap samples and idiosyncratic (bottom right) variance. Correlation between the idiosyncratic and B, M components components is near null (yelllow), as well as idiosyncratic covariance among different parts (off diagonal $cov(E_i, E_j)$ plots).

enough, the elements of cross covariance matrix that we need to sum in order to approximate aggregate variance are:

$$\begin{aligned} cov(log(S_{it}), log(S_{jt})) =& cov(b_{it} + m_{it} + \sigma_i \epsilon_{it}, b_{jt} + m_{jt} + \sigma_i \epsilon_{jt}) \\ =& cov(b_{it}, b_{jt}) + cov(b_{it}, m_{jt}) + cov(b_{it}, \sigma_j \epsilon_{jt}) \\ +& cov(m_{it}, b_{jt}) + cov(m_{it}, m_{jt}) + cov(m_{it}, \sigma_j \epsilon_{jt}) \\ +& cov(\sigma_i \epsilon_{it}, b_{jt}) + cov(\sigma_i \epsilon_{it}, m_{jt}) + cov(\sigma_i \epsilon_{it}, \sigma_j \epsilon_{jt}) \end{aligned} \tag{2.61}$$

In the remaining of this section I will show and discuss estimated dependence of such components B_{pt}, M_{pt}, and their counterparts b_{it}, m_{it}. Looking at the time dependence of these components is more interesting because they include net comovements, as opposed to the E_{pt} and $\sigma_p \epsilon_{pt}$ terms which are centered on zero.

Details regarding the procedure for estimation of this decompositions are important, and included in Appendix. Here I discuss the outcomes, see a brief description of technique for estimating elements of the covariance matrix in footnote. [43]

In figure 2.18 I illustrate the mean magnitude of elements of the cross covariance matrix of linear (left) and log (right) sectoral time series. Green elements are positive adn thus add a contribution to aggregate variance. Yellow elements are near null.

The top row and left column of blocks are for the extensive margin. The top left $cov(B_i, B_j)$ block tells the extensive part of covariance among parts. We can see a comovement plus idiosyncratic structure. Eventually one might want to disentangle it, but in general it confirms the general comovement plus idiosyncratic pattern we have observer throughout this work (see section 2.5.3. A partial correlation between the extensive margin and the intensive comovement is seen in the $cov(M_i, B_j)$ and $cov(B_i, M_j)$ blocks. This is to be expected, as we can see they both grow over time in our empirical benchmark (see Figure 2.17).

In our empirical case, the $cov(M_i, M_j)$ block is the one contributing most to aggregate variance. The idiosyncratic block $cov(E_i, E_j)$ is mostly diagonal as expected.

Matrix elements can be classified into their 3×3 blocks, as well as their diagonal or off diagonal status. Diagonal elements are those of a part with itself, and off diagonals are those among different parts. This classification is exploited in Figure 2.19 with the goal of summarizing the dependence of matrix elements with the magnitude of micro shocks $\hat{\sigma}$. Some comments are in order. Starting by the upper left, extensive margin block, note lack of dependence with increasing micro shocks as expected, and note the relatively higher magnitude of diagonal elements. This feature, also visible in figure 2.18 tells us there is room for separating a comovement from idiosyncratic changes of parts' extensive margin. When it comes to covariances of the intensive comovement with itself (mid column) we see elements grow with increasing microshocks. Diagonal and off diagonal elements are equivalent. Idiosyncratic

[43] For estimation of BME decompositions I exploit a bootstrapping approach in which I repeat the following steps multiple (eg. 100) times. For each run of the experiment, sample half of the firms randomly, design a partition into $P = 10$ random parts. First, force all firm sales levels to not deviate from their mean. These are the base time series, i.e. the fluctuations observed even when no firm fluctuates, because of changes in the population of firms.

Next, add firm level fluctuations by multiplying their actual magnitude slowly from zero to one. Measure the differences from the base level and they are $\delta_{pt} = m_{pt} + \sigma_p \epsilon_{pt}$. An estimation method needs to be introduced for disentagling these two terms. I use the median across bootstrap repetitions as estimation of m_{pt} and subtracting it from actual observations to arrive at the estimation of $\sigma_p \epsilon_{pt}$.

Once these decompositions have been done, one can look at all $9 \times P^2$ elements of the cross covariance matrix of equations 2.59 and 2.61.

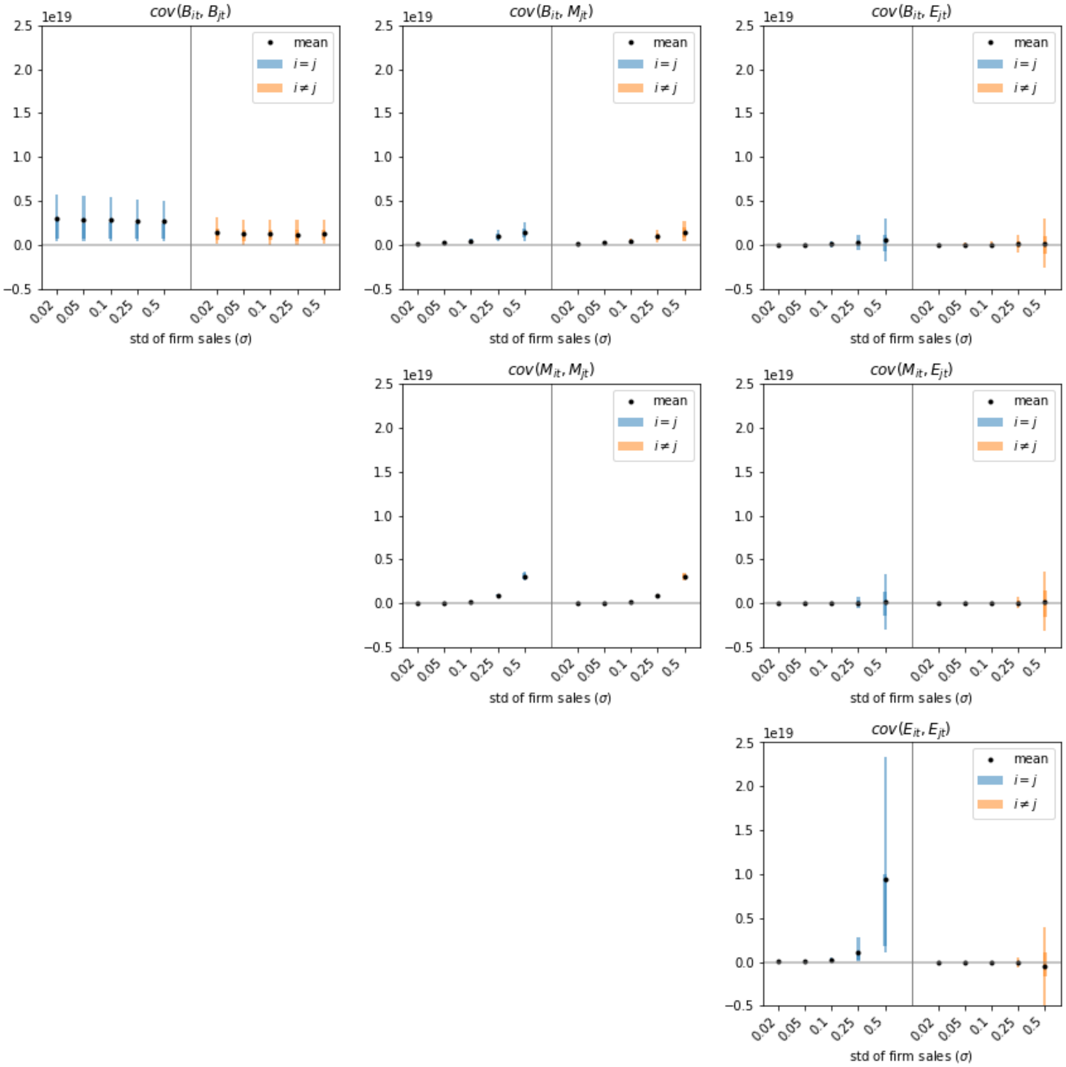
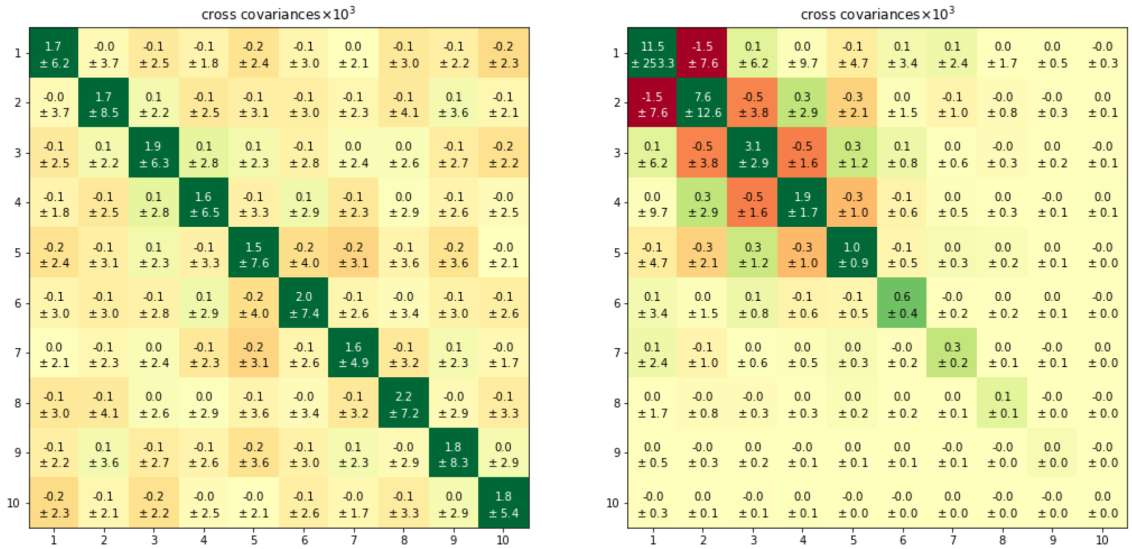


Figure 2.19: Cross covariance terms of equations 2.59 and 2.61. Expected value for increasing level of microshocks. Vertical scale shared across plots. The 3x3 boxes arrange combination of B, M, E components, the blue and yellow sides of each box distinguish diagonal ($i = j$, part variance) from off diagonal elements. Black dots denote mean values and lines show 25% to 75% quantiles estimated from bootstrap. Note the magnitude is that of single elements, which then need to be summed to arrive at the variance contribution of each block. Thus, for example, comovement elements (mid, mid) are smaller than idyosincratic ones (botom left), but they are P times as many. Discussion of the figure in text.

contributions to aggregate variance (left column) grow decidedly as microshocks are turned on from zero. Note that off diagonal terms, even if expected to be null (black point on the zero level) appear distributed around this value with a width that increases with the magnitude of micro shocks. The same happens with the correlation between idiosyncratic sectoral components and the comovement times series (left, top and mid). This means that in real examples, the off diagonal elements, are not null and can contribute positively or negatively to aggregate variance. A method for estimating the importance of these contributions is in section 2.12.



cross covariances$\times 10^3$

	1	2	3	4	5	6	7	8	9	10
1	1.7 ± 6.2	-0.0 ± 3.7	-0.1 ± 2.5	-0.1 ± 1.8	-0.2 ± 2.4	-0.1 ± 3.0	0.0 ± 2.1	-0.1 ± 3.0	-0.1 ± 2.2	-0.2 ± 2.3
2	-0.0 ± 3.7	1.7 ± 8.5	0.1 ± 2.2	-0.1 ± 2.5	-0.1 ± 3.1	-0.1 ± 3.0	-0.1 ± 2.3	-0.1 ± 4.1	0.1 ± 3.6	-0.1 ± 2.1
3	-0.1 ± 2.5	0.1 ± 2.2	1.9 ± 6.3	0.1 ± 2.8	0.1 ± 2.3	-0.1 ± 2.8	0.0 ± 2.4	0.0 ± 2.6	-0.1 ± 2.7	-0.2 ± 2.2
4	-0.1 ± 1.8	-0.1 ± 2.5	0.1 ± 2.8	1.6 ± 6.5	-0.1 ± 3.3	0.1 ± 2.9	-0.1 ± 2.3	0.0 ± 2.9	-0.1 ± 2.6	-0.0 ± 2.5
5	-0.2 ± 2.4	-0.1 ± 3.1	0.1 ± 2.3	-0.1 ± 3.3	1.5 ± 7.6	-0.2 ± 4.0	-0.2 ± 3.1	-0.1 ± 3.6	-0.2 ± 3.6	-0.0 ± 2.1
6	-0.1 ± 3.0	-0.1 ± 3.0	-0.1 ± 2.8	0.1 ± 2.9	-0.2 ± 4.0	2.0 ± 7.4	-0.1 ± 2.6	-0.0 ± 3.4	-0.1 ± 3.0	-0.1 ± 2.6
7	0.0 ± 2.1	-0.1 ± 2.3	0.0 ± 2.4	-0.1 ± 2.3	-0.2 ± 3.1	-0.1 ± 2.6	1.6 ± 4.9	-0.1 ± 3.2	0.1 ± 2.3	-0.0 ± 1.7
8	-0.1 ± 3.0	-0.1 ± 4.1	0.0 ± 2.6	0.0 ± 2.9	-0.1 ± 3.6	-0.0 ± 3.4	-0.1 ± 3.2	2.2 ± 7.2	-0.0 ± 2.9	-0.1 ± 3.3
9	-0.1 ± 2.2	0.1 ± 3.6	-0.1 ± 2.7	-0.1 ± 2.6	-0.2 ± 3.6	-0.1 ± 3.0	0.1 ± 2.3	-0.0 ± 2.9	1.8 ± 8.3	0.0 ± 2.9
10	-0.2 ± 2.3	-0.1 ± 2.1	-0.2 ± 2.2	-0.0 ± 2.5	-0.0 ± 2.1	-0.1 ± 2.6	-0.0 ± 1.7	-0.1 ± 3.3	0.0 ± 2.9	1.8 ± 5.4

cross covariances$\times 10^3$

	1	2	3	4	5	6	7	8	9	10
1	11.5 ± 253.3	-1.5 ± 7.6	0.1 ± 6.2	0.0 ± 9.7	-0.1 ± 4.7	0.1 ± 3.4	0.1 ± 2.4	0.0 ± 1.7	0.0 ± 0.5	-0.0 ± 0.3
2	-1.5 ± 7.6	7.6 ± 12.6	-0.5 ± 3.8	0.3 ± 2.9	-0.3 ± 2.1	0.0 ± 1.5	-0.1 ± 1.0	-0.0 ± 0.8	-0.0 ± 0.3	0.0 ± 0.1
3	0.1 ± 6.2	-0.5 ± 3.8	3.1 ± 2.9	-0.5 ± 1.6	0.3 ± 1.2	0.1 ± 0.8	0.0 ± 0.6	-0.0 ± 0.3	0.0 ± 0.2	-0.0 ± 0.1
4	0.0 ± 9.7	0.3 ± 2.9	-0.5 ± 1.6	1.9 ± 1.7	-0.3 ± 1.0	-0.1 ± 0.6	0.0 ± 0.5	0.0 ± 0.3	-0.0 ± 0.1	0.0 ± 0.1
5	-0.1 ± 4.7	-0.3 ± 2.1	0.3 ± 1.2	-0.3 ± 1.0	1.0 ± 0.9	-0.1 ± 0.5	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.1	-0.0 ± 0.0
6	0.1 ± 3.4	0.0 ± 1.5	0.1 ± 0.8	-0.1 ± 0.6	-0.1 ± 0.5	0.6 ± 0.4	-0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.1	-0.0 ± 0.0
7	0.1 ± 2.4	-0.1 ± 1.0	0.0 ± 0.6	0.0 ± 0.5	0.0 ± 0.3	-0.0 ± 0.2	0.3 ± 0.2	0.0 ± 0.1	-0.0 ± 0.1	0.0 ± 0.0
8	0.0 ± 1.7	-0.0 ± 0.8	-0.0 ± 0.3	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.1	0.1 ± 0.1	-0.0 ± 0.0	-0.0 ± 0.0
9	0.0 ± 0.5	-0.0 ± 0.3	0.0 ± 0.2	-0.0 ± 0.1	0.0 ± 0.1	0.0 ± 0.1	-0.0 ± 0.1	-0.0 ± 0.0	0.0 ± 0.0	-0.0 ± 0.0
10	-0.0 ± 0.3	0.0 ± 0.1	-0.0 ± 0.1	0.0 ± 0.1	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0

Figure 2.20: Idiosyncratic block of the log cross covariance matrix. Details of the mean values and standard deviations across bootstrap samples denoted by the $\pm$ simbol below mean values. Left: random partition into P = 10 parts. Right: quantile partition into Q = 10 parts.

Finally, figure 2.20 shows mean values of elements in the idiosyncratic block of the log cross covariance matrix and their standard deviation. The benefit of having applied equal weight partitions is that when we look at the matrix elements we can know they are contriuting equally to aggregate variance.

On the left side random partitions, and on the right side quantile partitions. The sum of these two matrices should be nearly similar, but they differ in how their elements contribute to this sum. The results with random partition are nearly uniform across sectors, as is to be expected, and most variance is on the diagonal elements. These diagonal terms are larger on average thanks to large firms' volatility. Notice however that deviations of all values are relatively large. Especially notice the amplitude of off diagonal terms and how they average to nearly zero across the $M = 200$ bootstrap repetitions.

The decomposition into quantile parts on the right is quite interesting to analyse. The structure is that of an outer product times a net covariance, as shown in the elementary examples of Figure 5 (section 2.5.3). Elements in the bottom and right side are associated to large groups of smaller firms. The sample variance decay with population size (milder due to non linearities and comovements but still present) results in a near null contribution of the large number of smallest firms accumulating up to 20% or 30% of value, as in the traditional picture where random shocks to thousands of agents cancel out. On the upper left end, on the contrary, we have the granularity situation where equal weight groups of fewer and fewer firms contribute their large magnitude sample variances. As in the outer product of of $\boldsymbol{\sigma}$ expression of section 2.5.3, multiplied by $cov(\epsilon_i, \epsilon_j)$ terms, we can see that the elements associated to these parts are able to contribute significant positive and negative terms to aggregate variance by means of their cross covariances (large off diagonal values in the upper left).

2.11 Conclusion

If we want to explain how the observed variance of an economic aggregate comes about, we need to acknowledge a variety of mechanisms that combine, potentiate or counteract among themselves. Let us make a summary of all the elements that come into play.

First, we need to know that variance is expected to be the sum of variance from aggregate shocks (comovement) and variance from the parts of the economy (idiosyncratic variance). Depending on their relative magnitude, aggregate variance can be identified with any of these two, or with a partial combination of both.

There are off diagonal variance components related to net cross sectoral correlations or to correlations between a sector's fluctuations and its aggregate shocks. These terms are expected to be null but in actual datasets they are never null, and they may need to be accounted for, together with the aggregate comovement and the idiosyncratic volatilities.

When it comes to the idiosyncratic contribution to aggregate volatility, there are many interesting mechanisms that combine to determine its magnitude.

Size distributions of economic agents, usually adapting to a log-normal distribution, or a power law distribution (Pareto) for the subset of largest firms imply a considerable concentration. This means that some parts of the value will be accounted for by few very large agents and others by a large number of small agents. The first ones will show a large variance of their observed sum (mean value), which dwarfs the volatility shown by a large number of small firms.

We need to know that the contributions of a group of agents to the idiosyncratic part of aggregate volatility $var[X]$ are weighted by the total value of those agents (and contributions to $var[log(X)]$ are weighted by the group's value share in the total). This means that the large sample variance observed in equal-weight parts that are little populated will drive the idiosyncratic aggregate volatility upwards. This outcome can be interpreted as if the population consisted of a smaller effective number of agents, and as such, as a breaking of the $\sigma_\epsilon^2 \sim 1/N$ rule.

This situation is nevertheless not describing any rule of decay of idiosyncratic aggregate volatility with number of agents, as such it should not be contrasted with a law of large numbers (LLN).

For understanding what happens with the LLN in empirical settings, we need to study groups of fluctuating agents. In general, there are two contributions to the volatility that such a group will present. These are a net comovement of each agent with all others in the group and a self variance of each agent. If fluctuations to agents were additive, then only the last term is non zero, and $var[S/S_q^0] \sim \hat{\sigma}_q^2/n_q$. In empirical settings, agent fluctuations are multiplicative, i.e. non linear. This means that agents' self variances will increase to $(\hat{\sigma}_q^2 + o(\hat{\sigma}_q^4))/n_q$. In addition, the comovement among agents will also grow as $o(\hat{\sigma}_q^4)$. Fat tail agent fluctuations exacerbate the onset of the comovement contributions to the group's volatility, which easily dominate when the group's population exceeds, in our case $n_q \gg 10$.

The effects we just described can be seen as a decay of variance of a group of agents with population that is milder than the $1/n_q$ rule, and instead adapts to a $\sigma^2 \sim n^{-\alpha} f(\mu, \hat{\sigma})$. The volatility of a large group of small agents is thus not so small because of this effect. Still, it is far smaller than the volatility shown by equal weight groups of large agents (because $\alpha \ll 0$). In fact, small agents contribute little to nothing to aggregate volatility in what can be taken as a milder but clear averaging out of their fluctuations.

Finally, the idiosyncratic aggregate variance inherits the variance rate of decay α from the parts that make up the total. Indeed we observe a decay idiosyncratic aggregate variance with total population given by $\sigma_\epsilon^2 = CN^{-\alpha}$, with $\alpha \approx 0.50$.

The combination of all these elements results in the observed values of aggregate variance in our benchmark empirical system. Part of their characterization and study has been absent in recent studies, possibly due to the complications of accounting for largely non linear fluctuations of economic agents. In any case, here I show that certain established conceptions need an open revision. Hopefully this papers' coverage of a large variety of ingredients intervening in volatility aggregation under empirical constrains can contribute to the understanding of non intuitive features of populations of economic agents.

2.12 Appendix I: Uncertainty introduced by off-diagonal elements.

Assume we estimated the σ_i^2 variance that each group of agents i present (equations 2.43, 2.48, 2.50, 2.54, 2.56). If apart from knowing sectoral variance we have information on aggregate shocks, equations 2.28 or 2.29 lead us to agrgegate variance. To arrive at those equations we assumed uncorrelated cross terms to go from equation 2.23 to eqs. 2.24 and 2.25. However, we have seen that in actual settings these terms are never null (see right column of plots in Figure 2.19 and off diagonal elements in matrices of plots in Figure 2.20).

This is why we will estimate the typical magnitudes of positive or negative contributions introduced by the terms of cross covariance, which are in expectation null, but actually never null. These contributions can be taken as an uncertainty on the value of aggregate variance expected. This uncertainty will be given in terms of the magnitudes of parts' fluctuations, σ_i.

We start from the idiosyncratic contributions to volatility, including cross covariances (Fig. 2.20):

$$\sigma^2_{parts} = \frac{1}{Q^2}\sum_{q_i,q_j}\sigma_i\sigma_j Cov(\epsilon_{it},\epsilon_{jt})$$

Here assume the σ_i values have been determined. The uncertainty of $E[\sigma^2(log(X))]$ will come from the terms $cov(\epsilon_{it},\epsilon_{jt})$. We do not know the ϵ_{it} draws, but the upcoming steps can be developed assuming some distribution for the $cov(\epsilon_{it},\epsilon_{jt})$ terms. For the sake of applying a simple example, let off diagonal values of $cov(\epsilon_{it},\epsilon_{jt})$ be a random variable drawn from a uniform distribution $U(-1,1)$. We will choose this description for the $cov(\epsilon_{it},\epsilon_{jt})$ terms and illustrate how one can estimate the uncertainty measure $std[\sigma^2(\log(X_t))]$ in this case. A generalization to the case of other possible distributions of $cov(\epsilon_{it},\epsilon_{jt})$ can be derived analogously.

Then, what is the expectation and variance of σ^2_{parts}? For this, let us separate the diagonal terms and express the rest as double the upper- (or lower-) diagonal terms, because $Cov(\epsilon_{it},\epsilon_{jt})$ is symmetric.

$$\sigma^2_{parts} = \frac{1}{Q^2}\sum_{q}^{Q}\sigma_q^2 + \frac{2}{Q^2}\sum_{i<j}\sigma_i\sigma_j Cov(\epsilon_{it},\epsilon_{jt})$$

$$\sigma^2_{parts} \overset{d}{=} \frac{1}{Q^2}\sum_q^Q \sigma_q^2 + \frac{2}{Q^2}\sum_{i<j}\sigma_i\sigma_j U(-1,1)$$

The expected value of the off diagonal terms is zero. This feeds the idea that the cross covariance terms could be dismissed in general. However these latter terms will actually most surely not be null. What can we do to estimate its importance? An option is to compute the variance coming from $Cov(\epsilon_{it}, \epsilon_{jt}) \overset{d}{=} U(-1,1)$ taking the σ_q terms as given. This is:

$var[Y] = E[Y^2] - E[Y]^2 = E[Y^2]$ with $Y = \frac{2}{Q^2}\sum_{i<j}\sigma_i\sigma_j U(-1,1)$ so that it reduces to computing the expectation of:

$$E[Y^2] = \frac{4}{Q^4}E\left[\left(\sum_{i<j}\sigma_i\sigma_j U(-1,1)\right)^2\right]$$

let us use the notation $\sigma_i\sigma_j \equiv \sigma^2_{ij} \equiv \sigma_p$ and $U_{ij} \equiv U_p$ to denote the random draw of $U(-1,1)$ that filled the entry i, j (the pair p) of the modelled cross covariance matrix. Then we have a sum of upper diagonal pairs to the square. Let us expand it.

$$\begin{aligned} E[Y^2] = \frac{4}{Q^4}E\left[\left(\sum_{p\equiv i,j}^{Q(Q-1)/2}\sigma_p U_p\right)^2\right] =& \frac{4}{Q^4}E\left[\sum_p^{Q(Q-1)/2}\sigma_p^2 U_p^2 + \sum_{p_1,\,p_2}^{Q(Q-1)/2} 2\sigma_{p_1}\sigma_{p_2}U_{p_1}U_{p_2}\right] \\ =& \frac{4}{Q^4}\frac{1}{3}\sum_p^{Q(Q-1)/2}\sigma_p^2 \end{aligned} \tag{2.62}$$

The latter term has null expected value because its distribution is given by the products of independent draws of $U(-1,1)$. The term U_p^2 is instead the distribution of the squares of $U(-1,1)$ which has $[0,1]$ as support and has a mean value of $\int_0^1 v^2 dv = 1/3$. Finally:

$$var[Y] = E[Y^2] = \frac{4}{3Q^4}\sum_{i<j}^{Q(Q-1)/2}\sigma_i^2\sigma_j^2$$

Grouping common base terms (eg: extensive margin terms plus time fixed effects as discussed in section 2.10) into a single base m_t and applying the same steps on the terms of cross covariance to their time series:

$$var\left[\frac{2}{Q^2}\sum_{q_i}\sigma_i Cov(m_t,\epsilon_{it})\right] = \frac{4}{3Q^4}\sigma_m^2\sum_i^Q\sigma_i^2$$

All this means that aggregate volatility has expected value and standard deviation:

$$E[var[\log(X_t)]] = \sigma_m^2 + \frac{1}{Q^2}\sum_q^Q\sigma_q^2$$

$$std[var[\log(X_t)]] = \frac{2}{Q^2}\sqrt{\frac{1}{3}\sum_{i<j}^{Q(Q-1)/2}\sigma_i^2\sigma_j^2 + \frac{1}{3}\sigma_m^2\sum_i^Q\sigma_i^2}$$

This is an uncertainty measure for aggregate variance estimations from using parts' volatility and aggregate shocks magnitude as in equation 2.28, based on the magnitude of cross covariance terms.

2.13 Appendix II: Accounting variance by frequency

The setting for our developments are time series of annual frequency. A question that may arise is whether measuring the same variables at a higher frequency does change the observed volatility substantially. The answer is that most of the variation of aggregate sales over time is explained by low frequency components, that is, cycles longer than a year well captured in annual data. This suggests measuring volatility on annual time series is a reasonable choice, although we will be missing certain additional volatility linked to seasonal changes.

To explore this issues we offer the expression for our problem in terms of Fourier series, which allows comparability with approaches as in Dupor (1999) and Horvath (1998).

The time series **x** we consider have monthly frequency over $T = 17$ years, so that their length is $L = 204 = 12T$. They can be expressed as:

$$\mathbf{x} = \sum_{k=0}^{L-1} y(k)\exp\left(i2\pi\frac{k\,\mathbf{x}_L}{L}\right) \tag{2.63}$$

with $k \in [0, 203]$ the wave number (minimum to maximum wave frequency), $\mathbf{x}_L = (1, ..., L)$ a vector of time steps, $y(k)$ the k-th element of Fourier transform of **x** divided by L, that is:

$$y(k) = \frac{1}{L} \sum_{t=0}^{L-1} \exp\left(-i2\pi \frac{k\,t}{L}\right) \mathbf{x}[t] \tag{2.64}$$

Alternative conventions to express these are as well valid. From here we can associate a magnitude of volatility to each frequency k by their amplitude in equation 2.63, i.e. by the magnitude of $y(k)$, excluding the zero frequency component. This is:

$$var(\mathbf{x}) = \sum_{k=1}^{L-1} |y(k)|^2 \tag{2.65}$$

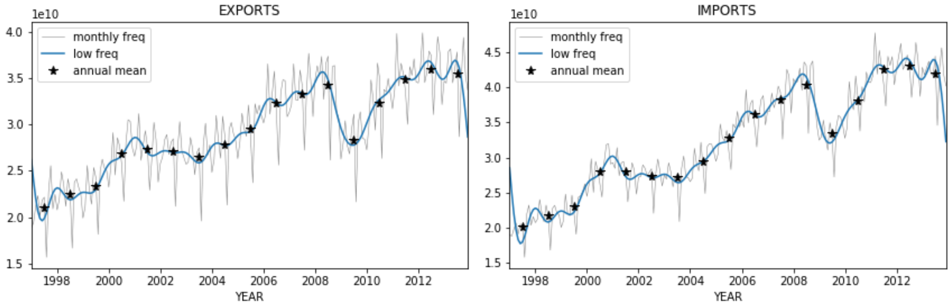


Figure 2.21: Annual or lower frequency (blue), full frequency (gray) and annual average levels for monthly aggregate French exports (left) and imports (right).

We have information on monthly disaggregation of exports and imports to assess whether most variation is explained by low or high frequency components. These are illustrated in the plots of figure 2.21. Annual and lower frequencies are called low, and higher than annual frequencies are called high. The only significant high frequency components are related to seasonal patterns (six month, quarters, two months, etc). For imports, annual or lower frequencies add up to 90% of variance, while for exports this figure is 75%.

This result encourages the study of variance in annual time series as a fair account of variance even in higher frequency measurements. Eventually we might want to add a contribution to aggregate volatility from seasonality. This is not problematic and we leave it out of the rest of the analysis.

2.14 Appendix III: log-normal and log-Laplace

The convention adopted for the formulas for the growth distributions are as follow:

Normal

$$N(\mu,\hat{\sigma},x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} exp\bigg(-\frac{(x-\mu)^2}{2\hat{\sigma}^2}\bigg)$$

Lognormal: $10^{N(\mu,\hat{\sigma},x)} = e^{N(m,\sqrt{V},\ln 10(x))}$, with $m = \mu\ln(10)$ and $V = \sigma^2\ln^2(10)$.

Also, $D(\mu,\hat{\sigma};n)$ denotes a draw of n elements from the distribution $D(\mu,\hat{\sigma})$.

2.14.1 Moments of a log-normal

The expected value (first moment) is $E[10^{N(\mu,\hat{\sigma},x)}] = e^{m+V/2} = 10^{\mu+\sigma^2\ln(10)/2}$. In general, the k-th moment is $E[10^{kN(\mu,\hat{\sigma},x)}] = \exp\left(\frac{k^2}{2}\sigma^2\ln^2(10) + k\mu\ln(10)\right) = 10^{\left(\frac{k^2}{2}\sigma^2\ln(10)+k\mu\right)}$

We can also calculate it from 2.36. If the microshocks pdf is Gaussian, i.e. $p(t) = exp(-(t-\mu)^2/(2\sigma^2))/\sqrt{2\pi}\sigma$, the expectation of 10^{kt} is:

$$\begin{aligned} E[10^{kt}] =& \int p(t)10^{kt}dt \\ =& \frac{1}{\sqrt{2\pi}\sigma}\int \exp\left(\frac{-(t-\mu)^2}{2\sigma^2} + \ln(10)kt\right)dt \\ =& \frac{1}{\sqrt{2\pi}\sigma}\int \exp\left(\frac{-((t-\mu)-\sigma^2\ln(10))^2}{2\sigma^2} + k^2\sigma^2\frac{\ln^2(10)}{2} + k\mu\ln(10)\right)dt \end{aligned} \tag{2.66}$$

where we completed squares in the exponential. Now we can integrate away the terms that represent a unit normalized Gaussian pdf, to be left with the terms:

$$E[10^{kt}] = \exp\left(k\mu\ln(10) + \frac{k^2}{2}\sigma^2\ln^2(10)\right) \tag{2.67}$$

From here, if t is given by a Normal distribution, then:

$$E[10^t] = 10^{\mu+\frac{1}{2}\hat{\sigma}^2 ln(10)} \approx 10^{\mu}\left(1 + \frac{1}{2}\hat{\sigma}^2\ln^2(10)\right) \tag{2.68}$$

where the expression on the left side are series approximations in the limit of small fluctuations $\hat{\sigma}$.

$$E[10^{2t}] = 10^{2\mu+2\hat{\sigma}^2 ln(10)} \tag{2.69}$$

Then variance, which is $var[10^{N(\cdot)}] = E[10^{2N(\cdot)}] - E^2[10^{N(\cdot)}]$ is:

$$E[10^{2t}] = 10^{2\mu+\hat{\sigma}^2 ln(10)}(10^{\hat{\sigma}^2 ln(10)} - 1) \approx 10^{2\mu}((\hat{\sigma}\ln(10))^2 + o(\hat{\sigma}^4)) \tag{2.70}$$

where the left side is a series approximation in the limit of small fluctuations $\hat{\sigma}$.

2.14.2 Moment of a log-normal in the context of size distribution

Apart from being relevant to knowing the average of small multiplicative micro shocks, the first moment of the lognormal can be used to derive the distribution of value from the distribution of population. Mandelbrot (1997) also looks this calculation in the paragraph titled "The lognormal's density and its population moments".

For this, apply a lognormal PDF as $p_{cnt}(s)$ in equation 2.3 and we will arrive at a PDF for $p_{val}(s)$. Recall (eq. 2.3):

$$Xp_{val}(s) = Np_{cnt}(s)\ s \tag{2.71}$$

Propose:

$$p_{cnt}(s) = \frac{1}{s}\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln s - \mu_c)^2}{2\sigma^2}\right). \tag{2.72}$$

Apply the change of variable $x \equiv \log(s)$. Thus when we apply eq. 2.72 in eq. 2.3, the lognormal multiplied by the value variable is:

$$p_{val}(s)ds = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln s - \mu_c)^2}{2\sigma^2}\right)ds$$

Because $ds/s = \ln(10)dx$ and already replacing $s = 10^x = e^{\ln(10)x}$

$$p_{val}(x)ds = \frac{\ln(10)}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x - \mu_c)^2}{2\sigma^2}\right)\exp(\ln(10)x)dx \tag{2.73}$$

the multiplication of counts per the corresponding value becomes the sum in the exponent, that is, it can be expressed as:

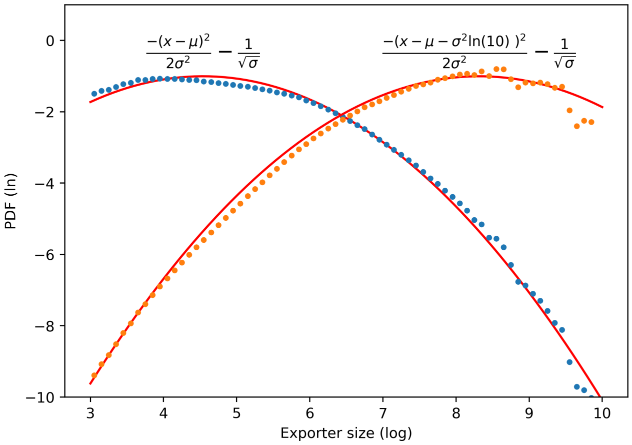


Figure 2.22: Size distribution in log log scale fitted by log-normal model (parabola). Data from all firm-years available are binned and fitted by OLS. Value distribution (yellow, right) is derived analytically from parameters of agents' size distribution (see expressions annotated and equation 2.74).

$$p_{val}(x)ds = \frac{\ln(10)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-(\mu_c+\sigma^2\ \ln(10)))^2}{2\sigma^2} + \frac{\sigma-2\mu_c}{2\sigma^2}\right) dx$$

that is:

$$p_{val}(x)ds = \exp\frac{\sigma-2\mu_c}{2\sigma}\frac{\ln(10)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_v)^2}{2\sigma^2}\right) dx$$

or, after normalization:

$$p_{val}(x)ds = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_v)^2}{2\sigma^2}\right) dx \tag{2.74}$$

which could be taken back to the linear horizontal scale if wanted. This means that the distribution of value is a lognormal with $\mu_v = \mu_c + \sigma^2\ \ln(10)$ and the same σ. These are the yellow curves to the right of the plots in figure 2.1.

We can also think of the parameters of the distribution as coefficients of a parabola (cf plots in figure 2.22, 2.1). The multiplication and normalization in the equations preceding 2.74 are the equivalent of adding a line to the parabola, and subtracting a constant to normalize. The result is a new parabola with the same quadratic coefficient but displaced $\sigma^2 \ln(10)$ to the right.

2.14.3 Moments of a log-Laplace

The Laplace is used as:

$$L(\mu, b, x) = \frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right) = \frac{1}{\sqrt{2}\hat{\sigma}} exp\left(-\frac{|x-\mu|}{\hat{\sigma}/\sqrt{2}}\right) \tag{2.75}$$

The parameters $\hat{\sigma} = \sqrt{2}b$ are proportional to each other and they are measures of the width of the distribution. Using b is a used convention but $\hat{\sigma}$ has the benefit of being the actual standard deviation observed.

The log-Laplace in base 10 is: $10^{L(\mu,\hat{\sigma},x)}$. A good reference is in Kozubowski and Podgorski (2003).

I show how one can derive the first moment in the simplified case of symmetric shocks distribution and mean zero. For that, we want to compute the expected value $E[10^t]$ when t is distributed as a Laplace.

To do this, first separate the negative and positive expression for the Laplace shocks pdf:

$$E[10^t] = \int_{-\infty}^{0} p(t)10^t dt + \int_{0}^{\infty} p(t)10^t dt \tag{2.76}$$

Knowing the expression of the indefinite integral of $10^t p(t)$ we only need to evaluate it at the integration limits (Barrow).

Use that $\int 10^t \exp(\pm t/b)/(2b)dt = 10^t \exp(\pm t/b)/(2(b\ln(10) \pm 1))$

$$E[10^t] = \frac{1}{2(b\ln(10)+1)} \left[\exp(t/b)10^t\right]_{-\infty}^{0} + \frac{1}{2(b\ln(10)-1)} \left[\exp(-t/b)10^t\right]_{0}^{\infty}$$

for evaluating this more easily, using again that $10^t = e^{\ln(10)t}$

$$E[10^t] = \frac{1}{2(b\ln(10)+1)} \left[\exp([1/b+\ln(10)]t)\right]_{-\infty}^{0} + \frac{1}{2(b\ln(10)-1)} \left[\exp([-1/b+\ln(10)]t)\right]_{0}^{\infty}$$

From here, the mean will diverge unless the exponentials are zero at the infinity, i.e. the exponent on the right needs to be negative, from where we need $\hat{\sigma} < \sqrt{2}/\ln(10) \approx 0.61$. The theoretical mean values do diverge upwards when approaching this $\hat{\sigma}$ level. Means from experiments when $\hat{\sigma}$ is above this level do show an 'explosion' upwards, although they are finite because there is a bounded number of agents (N).

Evaluating the primitive at the limits:

$$E[10^t] = \frac{1}{2(b\ln(10)+1)} - \frac{1}{2(b\ln(10)-1)} = \frac{1}{1-(b\ln(10))^2} \tag{2.77}$$

If the Laplace shocks were centered in $\mu \neq 0$, then the mean is just multiplied by 10^{μ} so that:

$$E[10^t] = \frac{10^{\mu}}{1-(b\ln(10))^2} = \frac{10^{\mu}}{1-\frac{1}{2}\sigma^2\ln^2(10))} \tag{2.78}$$

An expression for the moments of a log-Laplace, generalizing to possible asymmetries is in Kozubowski and Podgorski (2003) as:

$$E[10^{kt}] = \int p_L(t)10^{kt}dt = 10^{k\mu}\frac{\alpha\beta}{(\alpha - k\ln(10))(\beta + k\ln(10))} \tag{2.79}$$

where the α and β parameters stand for possibly asymmetric slopes on both sides of the mean. Here I incorporated the base 10 as we have done in the log-normal case, and departing from the natural base in Kozubowski and Podgorski (2003). The symmetric case implies using: $\alpha = \beta = 1/b$, which leads to:

$$E[10^{kt}] = 10^{k\mu}\frac{1}{1-(kb\ln(10))^2} \tag{2.80}$$

The parameter b relates to the standard deviation of a Laplace distribution ($\hat{\sigma}$) as: $b = \hat{\sigma}/\sqrt{2}$. Then the second moment is

$$E[10^{2t}] = 10^{2\mu}\frac{1}{1-(\hat{\sigma}\ln(10))^2/\sqrt{2}} \tag{2.81}$$

From here the variance $var[10^{L(\cdot)}] = E[10^{2L(\cdot)}] - E^2[10^{L(\cdot)}]$ would be:

$$var[10^{L(\cdot)}] = 10^{2\mu}\left(\frac{1}{1-4b^2} - \frac{1}{(1-b^2)^2}\right)$$

and in terms of the micro moment $\hat{\sigma} = \sqrt{2}b$, this is:

$$var[10^{L(\cdot)}] = 10^{2\mu}\left(\frac{1}{1-2\hat{\sigma}} - \frac{4}{(4-\hat{\sigma}^2)^2}\right) \tag{2.82}$$

In the limit of small micro fluctuations:

$$var[10^{L(\cdot)}] \approx 10^{2\mu}\left(\hat{\sigma}^2 + \frac{13}{4}\hat{\sigma}^4 + \frac{15}{2}\hat{\sigma}^6 + o(\hat{\sigma}^8)\right) \tag{2.83}$$

2.15 Appendix IV: Computational Experiments

The codes for reproducing all results in this paper will be available in a GitHub repository. The language used is python and files are jupyter notebooks. Next, I describe each of the experiments performed and I include a pseudo code.

2.15.1 Exp. 1. Dependence with N

This experiment has the goal of measuring the dependence of aggregate idiosyncratic variance (σ_ϵ^2) with population size (N).

It follows a sequence of steps:

- Prepare dataset
- Sample N agents with replacement
  - Compute aggregate statistics.
  - Apply random partition and compute parts' time series.
  - Apply quantile partition and compute parts' time series.
- Separate medians from idiosyncracies.
- Compute covariance matrices

For this exercise, we wish to have a dataset that is largely equivalent to the raw data, but for the condition that there is no entry and exit. To achieve this, the steps are to first, keep only firms that are active in at least 6 of the total 17 years available. Then, fill inactive years with mean value of the respective firms. Finally, keep only firms that are trading more than $1mEUR$ on average. These changes leave us with a resulting dataset with a size distribution that is largely equivalent to the upper tail (Pareto part) of the original data. Indeed, firms trading more than $1mEUR$ on average are present in most of the time steps so that the n/a filling is not really substantial. In addition, the largest firms concentrate most of the value, so that keeping only those above $1mEUR$ lets us still account for nearly the totality of french firms' international trade.

```
# Number of parts:
Q = 10

# Load data, sales by firm, year.
df = pd.read_csv(’./ID_Y.csv’)
sales = df.groupby([’IMPORT, ’ID’, ’YEAR’])[’VART’].sum().unstack()

for i in [0, 1]:

    # Choose firms with presence in most sample, to avoid high distortion
        filling exit gaps
    sales_filt = sales.loc[sales.count(1) > 6]
    filt_fm = sales_filt.copy()

    # Large firms pareto filled mean.
    for col in filt_fm:
        filt_fm[col] = filt_fm[col].fillna(sales_filt.mean(axis=1))

    # Hard cut for Pareto tail
    filt_fm = filt_fm.loc[filt_fm.mean(1) > 1e6]

    ## Sanity checks. What is the total after we filled non active gaps and
         kept large firms
    X = sales.sum().mean()
    X_actives = sales_filt.sum().mean()
    X_act_lrg = filt_fm.sum().mean()
    print(X_actives/X)
    print(X_act_lrg/X_actives)
    print(X_act_lrg/X)

# Save dataset
filt_fm.to_csv(’./firms_data.csv’)
```

The processed dataset is always within 10 percent of the total observed in the raw data.

For the experiment, we will sample N agents with replacement and compute aggregate statistics, as well as apply partitions and compute the cross covariance matrices these present.

```

# Population numbers
logn_vals = [2.5 , 2.65, ... , 3.7, 3.85] # log scale
n_vals = [ 300.,  400.,  ...,  5000., 7100.] # linear scale

# Repetitions
M = 150

data = pd.read_csv(’./firms_data.csv’)

for i in [0, 1]: # Exports / Imports
    for N in n_vals:
        for m in range(M):

            # Sample with replacement from agents’ time series
            n_sample = data.sample(int(N), replace = True)

            ## Calculate aggregate magnitudes: Total, firm sizes,
               Herfindahl.
            X_t = n_sample.sum()
            Si = n_sample.mean(1)
            herf2 = ((Si/Si.sum())**2).sum()
            agg_res += [[m, X_t.mean(), X_t.std(), np.log10(X_t).std(),
               herf2]]

            # Partition (random parts)
            n_sample_p = n_sample.copy()
            n_sample_p[’p’] = pd.cut(n_sample_p.sum(1).cumsum(), Q, labels
               = range(Q))

            # Aggregate to parts’ time series, and count parts’ population.
            n_m_p_out = n_sample_p.groupby(’p’).sum().reset_index()
            n_m_p_out[’n’] = n_sample_p.groupby(’p’).size().values




            # Partition (quantile parts)
            n_sample_q = n_sample.copy()
            n_sample_q = n_sample_q.loc[n_sample_q.sum(1).sort_values().
               index]  ## SORTING
            n_sample_q[’p’] = pd.cut(n_sample_q.sum(1).cumsum(), Q, labels
               = range(Q))

            # Aggregate to parts’ time series, and count parts’ population.
            n_m_q_out = n_sample_q.groupby(’p’).sum().reset_index()
            n_m_q_out[’n’] = n_sample_q.groupby(’p’).size().values

            <Store results>

<Concatenate results and save>
result_aggs
result_Sp
result_Sq
```

Now, we compute medians across the M repetitions and use it as proxy of the comovement time series of parts’.

```
# Store medians accross M repetitions
medians_p = result_Sp.groupby([’IMPORT’, ’N’, ’p’]).transform(’median’);
```

```
# Idiosyncracies are the actual values minus the medians
res_nmp = result_Sp.set_index([’IMPORT’,’N’, ’p’]) - medians_p;

# Store info
info_p = pd.concat([medians_p, res_nmp])

<Repeat for result_Sq in place of result_Sp>
```

Next, compute the values of the cross covariance matrix, in each setting and for each of the M repetitions.

```

cov_out_list = [] # List for outcoming cov values

for i in [0, 1]: # Exports / Imports
    for N in n_vals: # For all the sampling sizes (population N)
        for m in range(M): # For each of the repetitions
            for k, sorting in enumerate([False, True]): # For both random
               and sorted parts
                info = [info_p, info_q][k]
                df_ = info.loc[(info.IMPORT == i) & (info.m == m) & (info[’
                   N’] == N)]
                df_ = df_.set_index([’comp’, ’p’])[[str(y) for y in range
                   (1997, 2013)]]
                cov_m = df_.T.cov()

                cov_vals = cov_m.stack([0, 1])
                cov_vals.index.names = [’comp1’, ’p1’, ’comp2’, ’p2’];
                cov_vals.name = ’cov_ij’

                # Store
                cov_out_list += [cov_vals]

# Concatenate
cov_results = pd.concat(cov_out_list)
```

Now, we have M (eg. M = 150) realizations of each of the elements of the cross covariance matrix, separated into median (comovement) and residual (idiosyncratic) parts, for different population sizes N and both for imports and exports.

This information has multiple uses, for example making the plots in figure 2.8.

2.15.2 Exp. 2. Power sums

The aim of this experiment is to accompany the derivations in section 2.8. The essence of the program is quite simple. First generate a vectors of length n_q with realized values D_i drawn from a theoretical distribution $D(\cdot)$. Then, compute the sum of values in the vector $10^{D(\cdot)}$ and divide it by the sum of the zero levels ($10^{\mathbf{0}} = \mathbf{1}$) which is equal to n_q. The outcome is a time series of length T and we store its moments, as well as the moments of its log levels.

```
import pandas as pd
import numpy as np

# (names of) Log shocks distributions
dists = [’norm’, ’lapl’, ’empirical’]

# Values of micro standard deviation
ss = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]

# Values of micro mean fluctuations
mus = [0, 0.01, 0.02, 0.05, 0.1]

# Repetitions
M = 200

# Time steps
T = 17

## List of results
results = []

s0 = empirical_shocks.std()
```

```
for distribution in dists:
    # Second quantile and next contain most of the value with few agents
    for q in range(Q)[1:]:
        n = population[’n_q’][q] # number of agents in quantile.

        for s in ss: # micro sigma
            for mu in mus: # micro mu

                for m in range(M): # Repeat M times
                    if distribution == ’norm’:
                        shocks = np.random.normal(mu, s, (n, T))
                    elif distribution == ’lapl’:
                        shocks = np.random.laplace(mu, s/np.sqrt(2), (n, T)
                           )
                    elif distribution == ’empirical’:
                        shocks = (mu + np.random.choice(emp_shocks, n * T)
                           *(s/s0)).reshape(n, T)

                    ratio = np.power(10, shocks).sum(0)/n

                    log_ratio = np.log10(ratio)

                    results += [[distribution, s, mu, n, m, ratio.mean(),
                       ratio.std(), ratio.var(), log_ratio.mean(),
                       log_ratio.std(), log_ratio.var()]]


# Create dataframe with the computed information
result = pd.DataFrame(results, columns = [’distribution’, ’s’, ’mu’, ’nq’,
   ’repeat’, ’mean_ratio’, ’std_ratio’, ’var_ratio’, ’mean_log_ratio’, ’
   std_log_ratio’, ’var_log_ratio’])

# Save pandas DataFrame
result.to_csv(’./filename.csv’, index = False)
```

Chapter 3

Areal disaggregation: correlation leads to spatial patterns

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.

3.1 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; F. 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.

[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.

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 3.2 reviews works applying cooccurrence measures. Section 3.3 describes the data. Section 3.4 presents a overview of the methods used, clarifying notation and terminology. Section 3.5 shows the grounds for unifying a variety of discrete coexistence measures. Section 3.6 shows how discrete similarity measures match a continuous model of space. Section 3.7 discusses the correlation structures observed in US and we conclude in Section 3.8.

3.2 Related works

3.2.1 The use of similarity measures

Inner products such as $X^T X$ are basic measures of joint cooccurrence and as such they have been featured often. The elements of this matrix are $(X^T X)_{ii'} = \sum x_{ij}.x_{i'j}$. 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) and 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) and 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

Table 3.1: 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) and 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), and 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) and 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 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 (F. 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 3.1 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

[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).

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 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.

3.2.2 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) and 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) and 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) and 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 3.6 where we will probe the formal connection between areal and point data, offering a solution to the MAUP in studies of co-location.

3.3 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]

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.

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 3.1), 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.

3.4.1 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 $X^T X$. There are other measures that can fulfill this role, such as Pearson correlation, cosine similarity and covariance. If we have a pair of columns $X_j = (x_{1j}, \dots, x_{ij}, \dots, x_{nj})^T$ and $X_{j'}$, these similarity measures are defined as follows:

[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'.

Pearson correlation:

$$Corr(j,j') = r_{jj'} = \frac{\sum\limits_{i}^{n}(x_{ij} - \bar{X_j})(x_{ij'} - \bar{X_{j'}})}{||X_j - \bar{X_j}||||X_{j'} - \bar{X_{j'}}||} \tag{3.1}$$

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:

$$CosSim(j,j') = r_{ii'} = \frac{\sum\limits_{i}^{n} x_{ij}\ x_{ij'}}{||X_j||||X_{j'}||} \tag{3.2}$$

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

Sample covariance:

$$Cov(j,j') = \frac{1}{n}\sum_{i}^{n}(x_{ij} - \bar{X_j})(x_{ij'} - \bar{X_{j'}}) \tag{3.3}$$

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 $X^T X$ 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 $Cov(Y) = Y^T Y/(n-1)$, 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. $Corr(Z) = Z^T Z/(n-1)$ 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. $Cossim(V) = V^T V$ 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 $Corr(W) = W^T W = Cos(W)$.

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.

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.

3.5 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

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

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) $\approx$ corr(X)
- X^TX $\propto$ cov(X)

For the first item (cos(X) $\approx$ corr(X)), the correspondence is an identity. For the second one (X^TX $\propto$ 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 3.1 applied on employment level data. Analogous results are observed for number of establishments data.

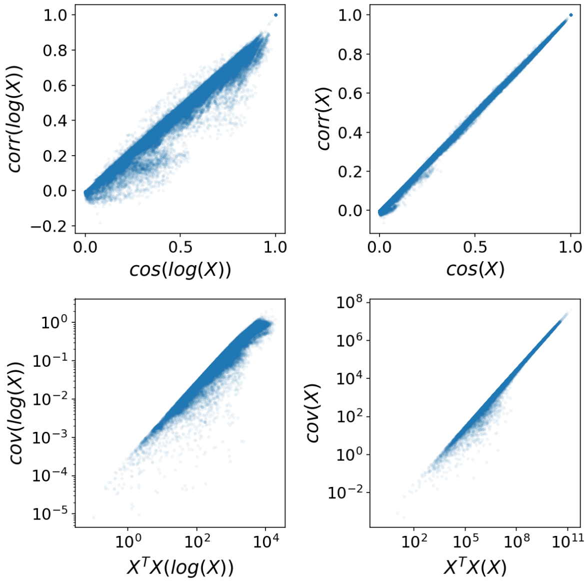


Figure 3.1: Scatterplots with direct comparison of selected industry pair similarity measures from US employment by county data. The notation is (top plots): $corr()$: Pearson correlation (eq. 3.1), $cos()$: cosine similarity (eq. 3.2) $cov()$; (bottom plots): covariance (eq. 3.3), $X^TX()$: simple joint 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. 3.3). These clear connections between similarity indices suggest paths for unification of methodologies applied in different studies.

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 3.2 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:

- 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.

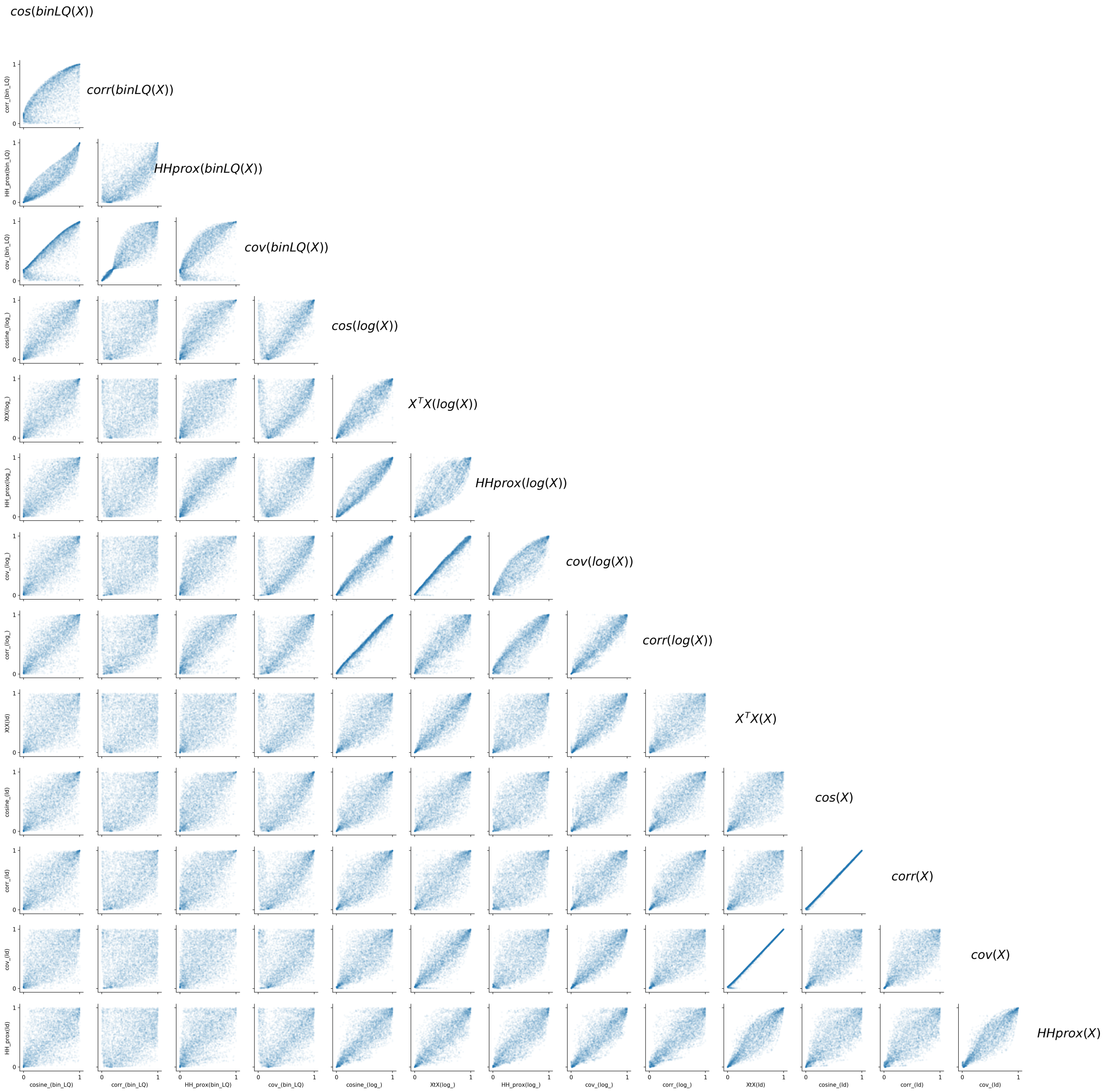


Figure 3.2: Comparison of rankings for multiple measures combining 5 similarity measures (cosine similarity, Pearson correlation, covariance, Hausmann Hidalgo proximity and dot product) applied on the cross section of employment by NAICS 4 digit industry and US county and two transformations thereof (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 3.1. Also, $HHprox()$ stands for proximity as in Hidalgo et al. (2007) (minimum conditional probability). The argument $binLQ(X)$ stands for binarized location quotients.

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 3.6 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.

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.

3.6 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 **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 **x** is given by the product of probabilities: $F_x(\mathbf{x})\ F_y(\mathbf{x})$.

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 3.3. If we wanted to add up all places across the country influenced by both industries x and y, we compute the integral:

$$\iint_R F_x F_y dR \tag{3.4}$$

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:

$$M_{x,y} = \sum_a E_{x,a}.E_{y,a} = \sum_a \left(\sum_{i=1}^{N_{x,a}} E_{xi} \sum_{j=1}^{N_{y,a}} E_{yj} \right) = \sum_a \sum_{\substack{i \in x,a \\ j \in y,a}} E_{xi} E_{yj} \tag{3.5}$$

For a graphical representation of $E_{x,a}$, $E_{y,a}$ and their product, see the right side of figure Figure 3.3. The lower plots is for the product of employment levels. The grid demarcates the

[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.

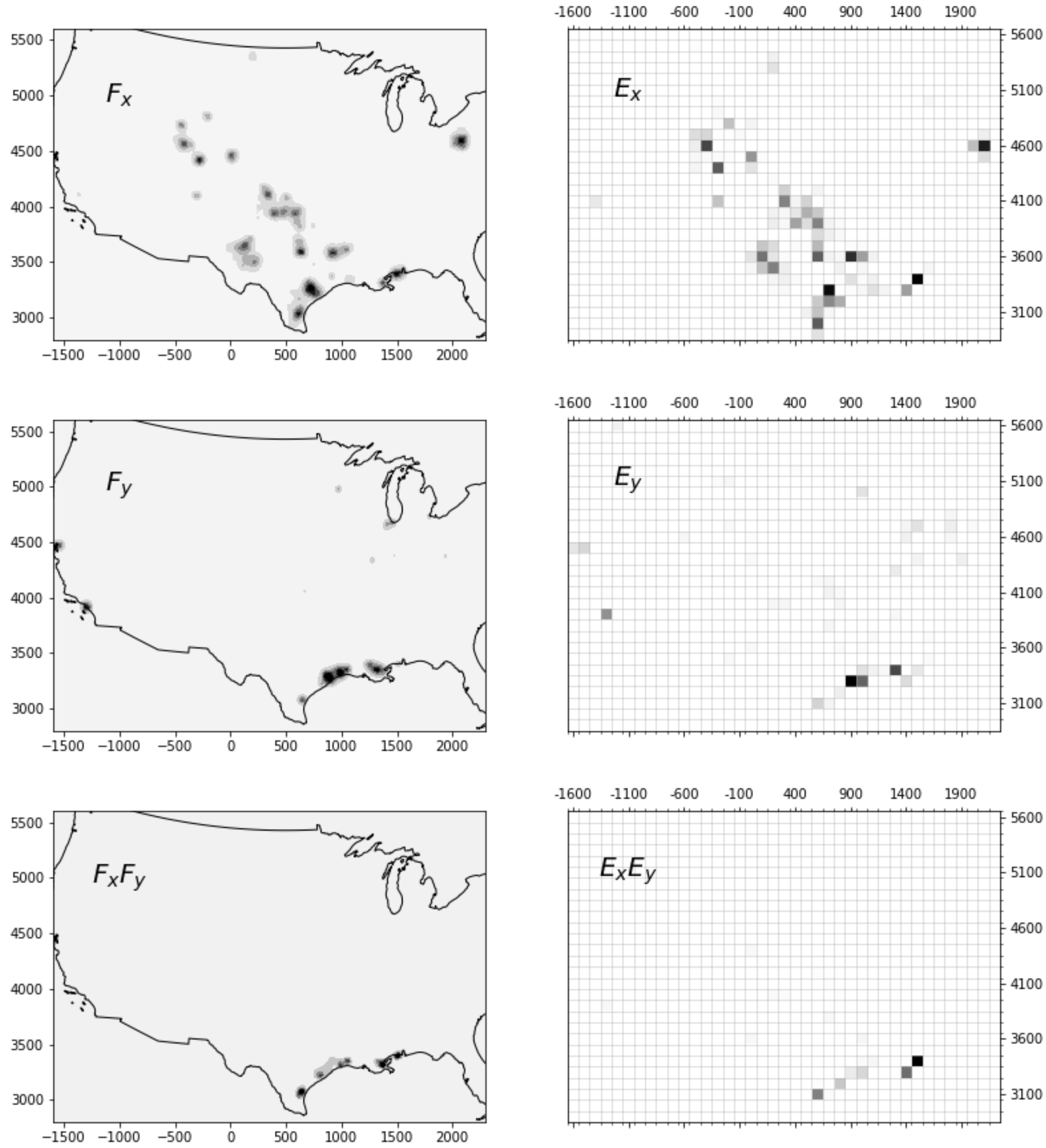


Figure 3.3: Demonstration of setup for continuous space (left) versus areal data (right) comparison. Top plots relate to locations of natural gas extraction fields (industry x). Middle plots relate to locations of oil refineries. Lower plots are the result of multiplying the upper plots. Grid lines depict 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.

modelled areas a. The exercise in this section is simply to compare a normalized volume under the $\mathbb{R}^2 \rightarrow \mathbb{R}$ 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:

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

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:

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

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:

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

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 $\iint\limits_R f_i f_j dR$ 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:

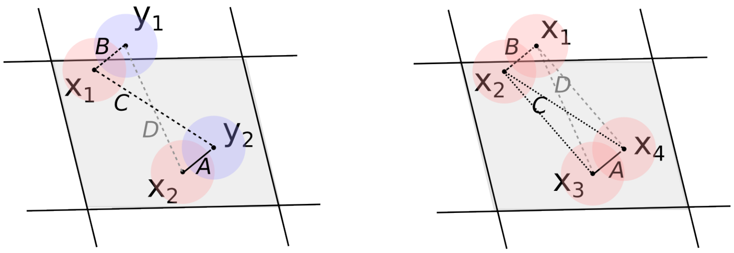


Figure 3.4: Micro accounting of coexistence between facilities belonging to a pair of industries (left) or a single industry (right). All links belong to 4 sets we named A, B, C, D, depending on whether they share the same administrative area and whether they actually are close to each other in the continuous space.

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 3.4.

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:

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

3.6.1 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 3.4 and 3.5). 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:

$$\frac{\iint\limits_{R} F_x F_y dR}{\sqrt{\iint\limits_{R} F_x^2 dR}\sqrt{\iint\limits_{R} F_y^2 dR}} \tag{3.6}$$

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:

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

$$\frac{\sum\limits_a E_{x,a}.E_{y,a}}{\sqrt{\sum\limits_a E_{x,a}^2}\sqrt{\sum\limits_a E_{y,a}^2}} \tag{3.7}$$

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:

$$\frac{E_{xi}E_{yj}}{\sqrt{\sum\limits_a E_{x,a}^2}\sqrt{\sum\limits_a E_{y,a}^2}} = \frac{\iint\limits_R f_{x,i}(\mathbf{x})f_{y,j}(\mathbf{x})dR}{\sqrt{\iint\limits_R F_x^2 dR}\sqrt{\iint\limits_R F_y^2 dR}}; \ \ \forall i,j \in A \tag{3.8}$$

3.6.2 Solution for industry self-overlap

Applied to some industry x on itself this will be:

$$\frac{E_{xi}^2}{\sum\limits_a E_{x,a}^2} = \frac{\iint\limits_R f_{x,i}^2(\mathbf{x})dR}{\iint\limits_R F_x^2 dR}; \ \ \forall i,j \in A \tag{3.9}$$

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:

$$\iint\limits_R g_{x,i}(\mathbf{x})g_{y,j}(\mathbf{x})dR = \frac{t_i t_j}{2\pi\left(\sigma_i^2+\sigma_j^2\right)}\exp\left(-\frac{\Delta^2}{2\left(\sigma_i^2+\sigma_j^2\right)}\right) \tag{3.10}$$

and we are asking that this is comparable to $E_{xi}E_{yj}$ (Eq. 3.8). Note that in Eq. 3.10 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 $\Delta \to 0$

$$\iint\limits_R g_{x,i}(\mathbf{x})g_{y,j}(\mathbf{x})dR \to \frac{t_i t_j}{2\pi\left(\sigma_i^2+\sigma_j^2\right)} \tag{3.11}$$

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

$$\iint\limits_R f_{x,i}(\mathbf{x})f_{y,j}(\mathbf{x})dR = \frac{t_i t_j}{2\pi\left(b_i+b_j\right)^2} \tag{3.12}$$

To summarize these two results, consider self overlap of an industry (then $\Delta = 0$, and $i = j$) and let eqs 3.11 and 3.12 be expressed as:

$$\iint\limits_R h_{x,i}^2(\mathbf{x})dR = \frac{t_i^2}{2\pi s_i^2} \tag{3.13}$$

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. 3.9 links overlaps in continuous space with the observed counts of employees by establishment. Replacing 3.13 into 3.9 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 3.9 to hold. Taking square root of 3.9:

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

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

$$t_i = \sqrt{2\pi} s_i \left(\frac{\sqrt{\int_A F_x^2 dR}}{||E_{x,a}||} \right) E_{x,i} \tag{3.14}$$

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]

3.6.3 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. 3.8 instead of Eq. 3.9) requires that instead of the simplified equation 3.13 (valid when $\Delta = 0$) we use an expression such as Eq. 3.10 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.

[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 3.14 would be preserved.

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]

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 3.5, applied on the location of oil refineries, with $100km^2$ square areas. Generalizations of the experiment applied to natural

[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.

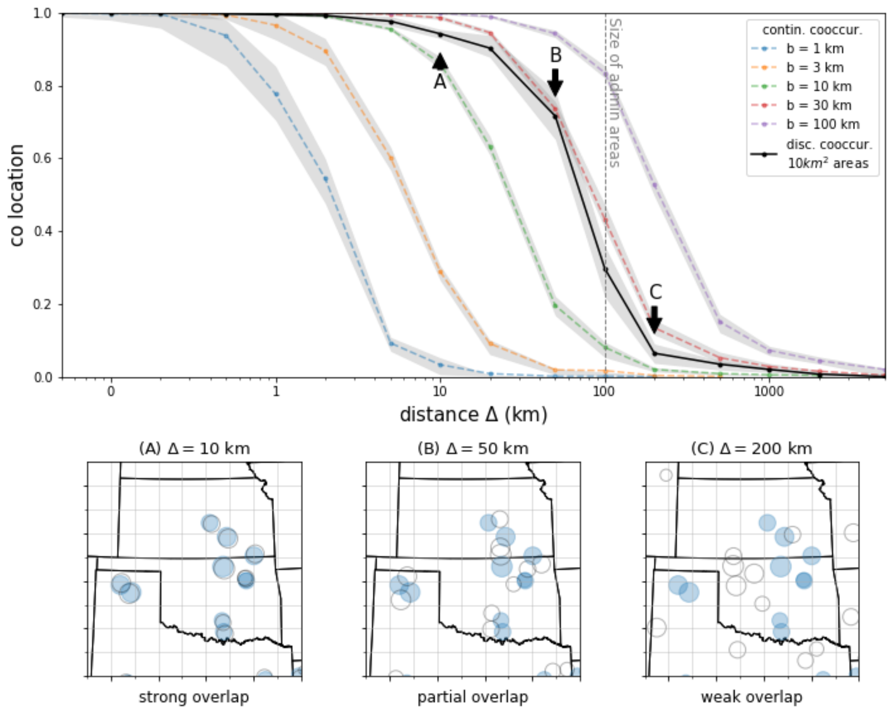


Figure 3.5: (Top) Decay of coexistence measures with distance Δ. Cosine similarity on admin. areas (black) and overlap of density functions in continuous space (colors). Density functions are radial exponential decay, computed for various width parameters (see legend). On the left end ($\Delta \to 0$) 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.

gas extraction locations and repeating the exercise for $10km^2$ areas are shown in Figure 3.6. 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 3.4). 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 equivalent. So that with all these developments we are finding the concrete meaning of coexistence measures when applied on US county data.

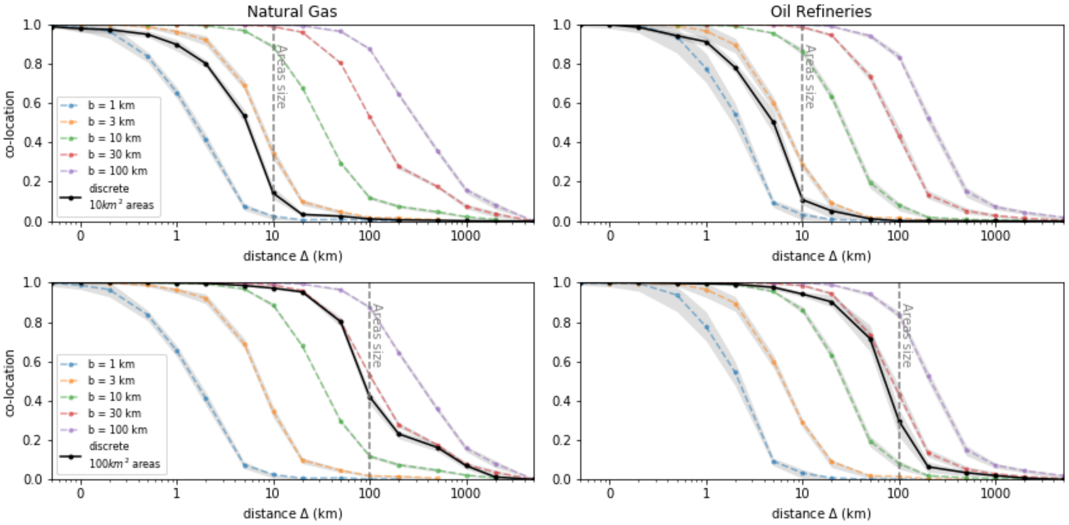


Figure 3.6: Decay of coexistence measures with distance Δ. Cosine similarity on admin. areas (black) and overlap of density functions in continuous space (colors, see legend). Results for two industries (left - right) and for two area sizes (10 km, top - 100 km, bottom). The decay of discrete area coexistence (black) appears linked to area size (gray vertical line), as they both shift by the same amounts.

In Figure 3.6 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).

3.7 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.

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 3.7.1 we introduce the methods applied to arrive at an industry space and geographical patterns and in subsections 3.7.2 and 3.7.3 I show and discuss the results.

3.7.1 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:

$$C = \sum_k \lambda_k u_k u_k^T = \sum_k \lambda_k V_k \tag{3.15}$$

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. 3.15 works similarly for cosine similarity and correlation of logs matrices. We illustrate it graphically in Figure 3.7 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]

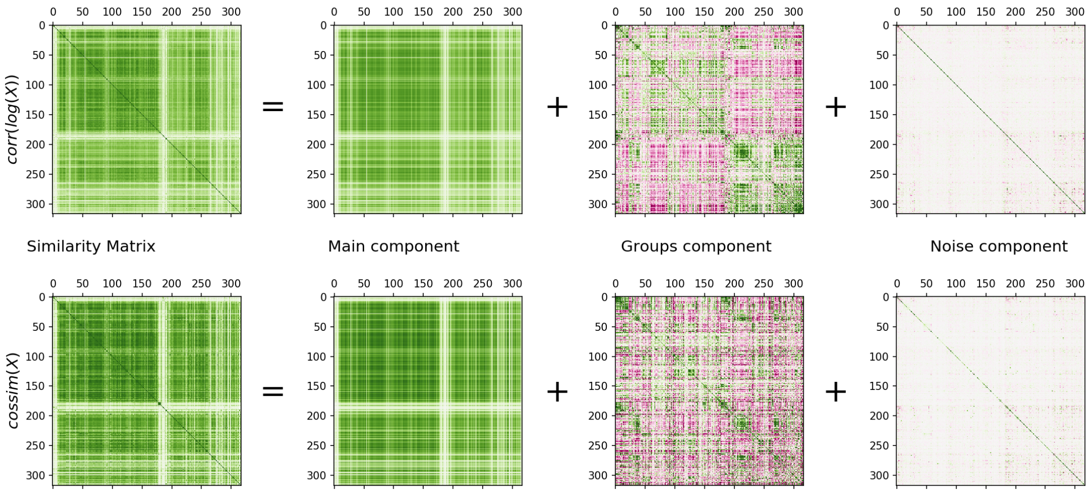


Figure 3.7: Decomposition of similarity by eigenvalue components (eq. 3.15)

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.

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

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. 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. 3.15) 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 preprocessing 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 -

[11] Documentation for sklearn.cluster.SpectralClustering

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 3.7 are the following:

- 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.

3.7.2 Results: Network of industries

Let us begin by presenting the network structure of industries. To begin understanding the outcome, we can look at Figure 3.8. 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 3.7.1 I detect 11 components. These are represented in colors in the networks of Figure 3.8. They are listed in Table 3.2.

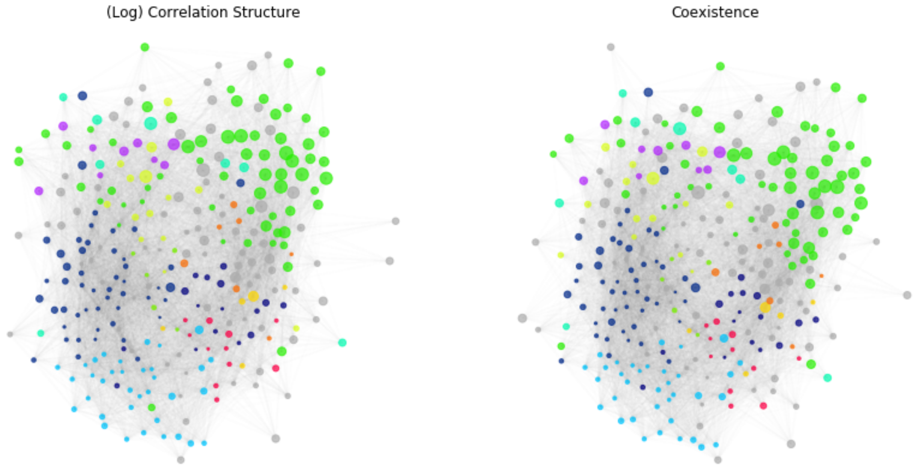


Figure 3.8: Networks of industries. Left: from groups component of correlation of log levels. Right: from groups component of cosine similarity. Edge weights computed from employment levels and number of establishments are averaged for each plot. The colors depict the components we built from clustering in each of the four variable - similarity combinations (cf Section 3.7.1). 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 3.9 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.

Finally on Figure 3.10 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

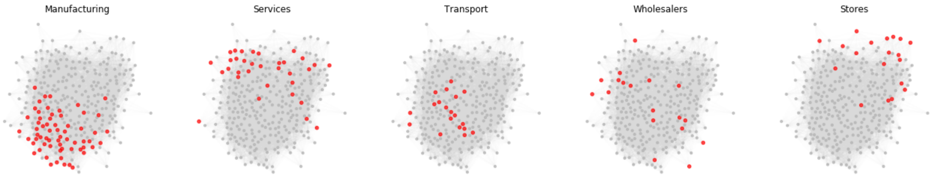


Figure 3.9: Network of industries. Highlight of frequent words in industry titles. The aim of this plot is to help understand the regions of the networks plotted in Figure 3.8.

leasing, 54 Professional and technical services.

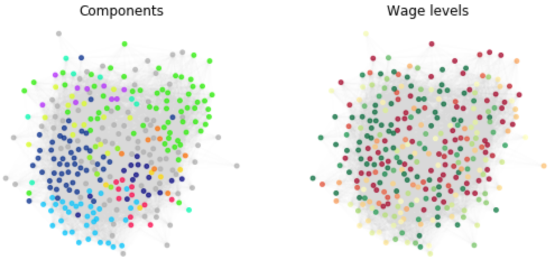


Figure 3.10: Network of industries. Clusters by community detection (left) and wage levels (right). Only clusters of urban activities are characterized by a (higher) average wage level. The rest of the clusters have mixed wage levels.

3.7.3 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 3.2). 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.

The following table summarizes the components we could detect:

Next we review them in further detail.

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 3.2: Summary of detected *Themes* and *Components*

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."*.

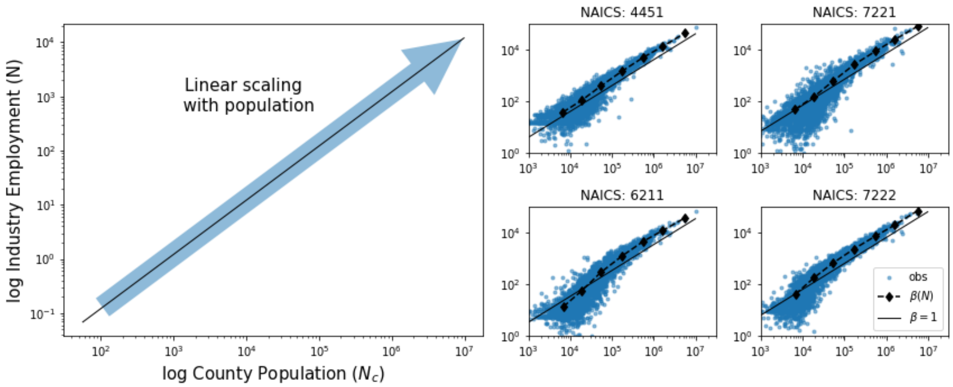


Figure 3.11: Population-linear scaling of activities in the 'population' theme. Right: scatterplots of data (NAICS 4451 Grocery stores, NAICS 6211 Offices of physicians, NAICS07 7221 Full-service restaurants, NAICS07 7222 Limited-service eating places). Left: qualitative scaling pattern.

Non tradables: stores and personal services.

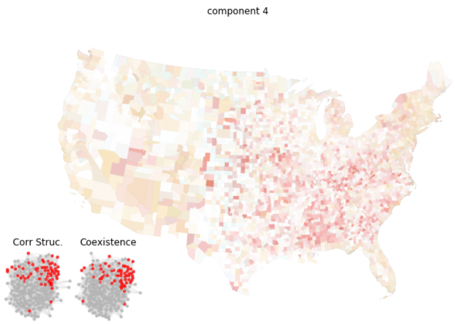
Distribution	Activities
	NAICS 238 Construction contractors NAICS 44-45 Retail trade NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services NAICS 62 Health care and social assistance NAICS 72 Accommodation and food services NAICS 81 Other services, except public administration

Table 3.3: Non tradables. LINK TO INTERACTIVE MAP

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 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.

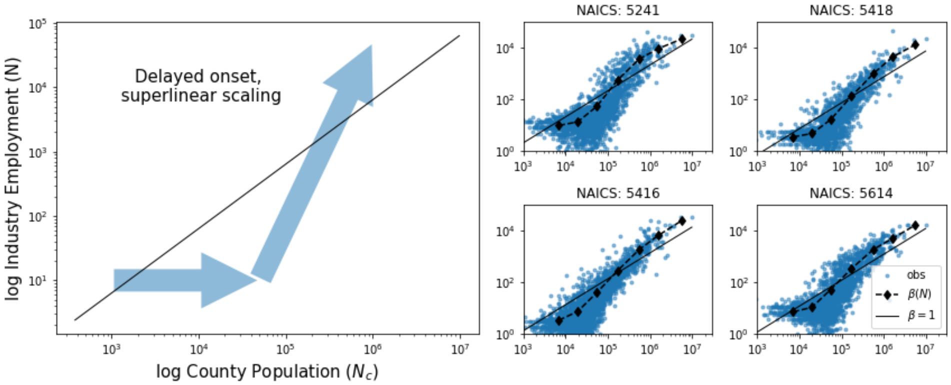


Figure 3.12: Delayed onset and superlinear scaling of activities in the 'cities' theme. Right: scatterplots of data (NAICS 5241 Insurance carriers, NAICS 5416 Management and technical consulting services, NAICS 5418 Advertising, PR, and related services, NAICS 5614 Business support services). Left: qualitative scaling pattern.

On figures 3.11 and 3.12 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

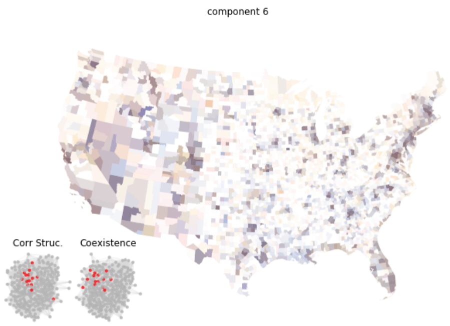
Distribution	Activities
	NAICS 5112 Software publishers NAICS02 5181 Isps and web search portals NAICS 5182 Data processing, hosting and related services NAICS 5415 Computer systems design and related services NAICS 5417 Scientific research and development services NAICS 5612 Facilities support services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

Table 3.4: Large city economies I. LINK TO INTERACTIVE MAP

Large city economies II

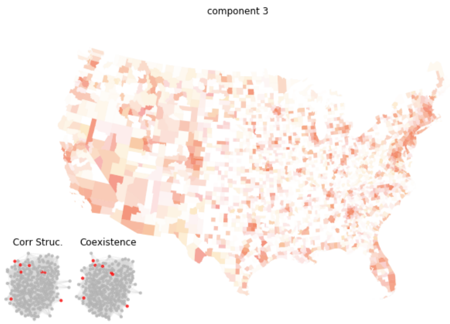
Distribution	Activities
	NAICS 51 Information NAICS 52 Finance and insurance NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services

Table 3.5: Large city economies II. LINK TO INTERACTIVE MAP

Other high wage activities

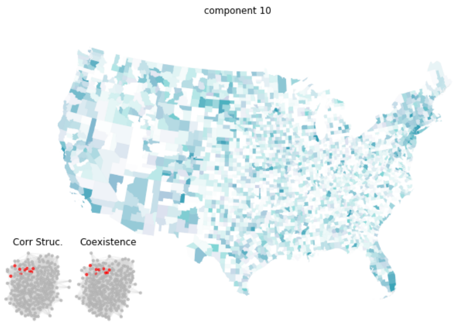
Distribution	Activities
	NAICS 51 Information NAICS 52 Finance and insurance NAICS 53 Real estate and rental and leasing NAICS 54 Professional and technical services

Table 3.6: 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

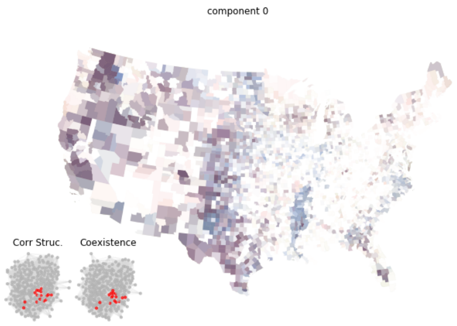
Distribution	Activities
	NAICS 111 Crop production NAICS 112 Animal production and aquaculture NAICS 311 Food manufacturing

Table 3.7: Ranching. 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).

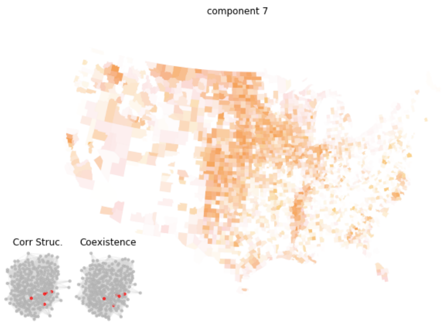
Agriculture and Food II: Corn Belt	
Distribution	Activities
	NAICS 1111 Oilseed and grain farming NAICS 1122 Hog and pig farming NAICS 311 Food manufacturing NAICS 4245 Farm product raw material merch. wholesalers

Table 3.8: Corn Belt. LINK TO INTERACTIVE MAP

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).

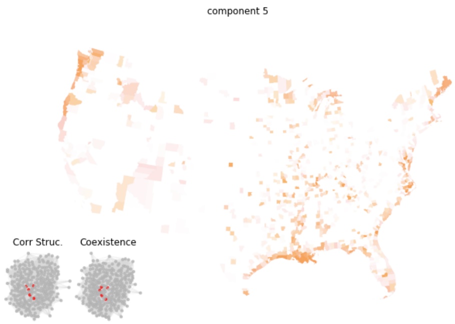
Water Economy	
Distribution	Activities
	NAICS 1141 Fishing NAICS 3117 Seafood product preparation and packaging NAICS 3366 Ship and boat building NAICS 4831 Sea, coastal, and great lakes transportation NAICS 4832 Inland water transportation NAICS 4872 Scenic and sightseeing transportation, water NAICS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

Table 3.9: 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 3.6 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

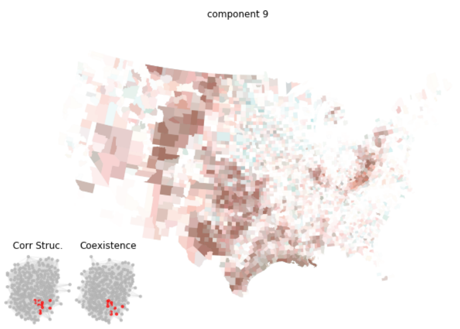
Distribution	Activities
	NAICS 21 Mining, quarrying, and oil and gas extraction NAICS 324 Petroleum and coal products manufacturing NAICS 3251 Basic chemical manufacturing NAICS 486 Pipeline transportation

Table 3.10: 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

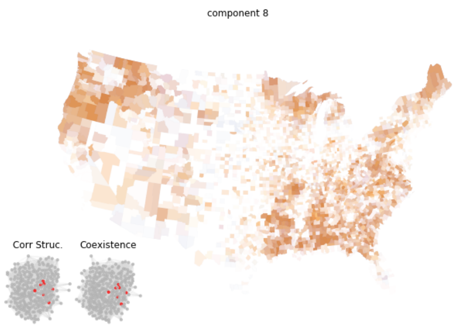
Distribution	Activities
	NAICS 1131 Timber tract operations NAICS 1132 Forest nursery and gathering forest products NAICS 1133 Logging NAICS 1153 Support activities for forestry NAICS 3211 Sawmills and wood preservation NAICS 3212 Plywood and engineered wood product mfg. NAICS 3371 Household and institutional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

Table 3.11: 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

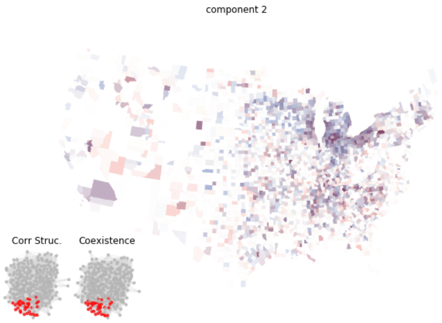
Distribution	Activities
	NAICS 325 Chemical manufacturing NAICS 326 Plastics and rubber products manufacturing NAICS 327 Nonmetallic mineral product manufacturing NAICS 331 Primary metal manufacturing NAICS 332 Fabricated metal product manufacturing NAICS 333 Machinery manufacturing NAICS 335 Electrical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

Table 3.12: Manufacturing. LINK TO INTERACTIVE MAP

3.8 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 par-

Other manufacturing and other activities

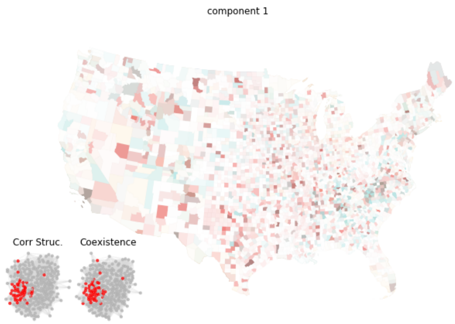
Distribution	Activities
	NAICS 31-33 Manufacturing NAICS 48 Transportation NAICS 51 Information NAICS 52 Finance and insurance

Table 3.13: Other than steel belt manufacturing and other activities. LINK TO INTERACTIVE MAP

tially equivalent among themselves (Section 3.5). I then explore the connection to coexistence accounts computed from continuous space (i.e. based on establishments' point locations and employment levels) (Section 3.6). 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.

Chapter 4

Overlooked features of Location Quotients $LQ = s_{cp}\, S/(S_c S_p)$

Abstract

Contingency tables appear frequently in empirical studies in Economics. Location quotient (LQ) indices have been widely adopted as relative measure of intensity. They are defined as ratio between observed values and the expectation from uncorrelated marginal distributions (size factor). If $LQ = 1$, the observation matches expectation. Higher values of LQ denote a relative high intensity. The values resulting from this transformation are known to not be comparable across rows (columns) of the table, mostly because they are sensitive to the total sum of the row (column) in question. Such effects however have not been formally studied.

In this paper I propose a line of reasoning for defining curves of fixed distance to LQ = 1. For this, I first define two dimensional coordinates that describe all pairs of observed values and size factors. Then I compute the probabilities that $LQ_{t+1} > 1$ given the parameters observed at t. This works as an a posteriori estimator of the chances of surpassing $LQ = 1$ after a time step conditional on the coordinates of the starting point.

This technique allows effective measurement of distortions dependent on the size of the rows (columns) on the scale of the location quotient. For example, in the context of trade empirics (the observed values are exports by country and product) I show and explain how, starting from an equal level of $LQ < 1$ observations from smaller countries are more likely to surpass the $LQ = 1$ threshold than those from large countries. I find that the decay of volatility of log fluctuations of exports with increasing size can explain the pattern of size effects in LQ levels.

Acknowledging these effects allows them to be controlled, in studies where location quotients are used as dependent variable. Among other uses, such controlled level curves allow defining intermediate LQ values, so that studies that used binary mappings before can now exploit a richer variety of LQ categories consistent across all observations in a dataset.

4.1 Introduction

The context for the developments in this paper are contingency tables where a total volume (S_W) is disaggregated independently by two sets of categories (*C*, *P*). That is, $S_W = \sum_{cp} s_{cp}$

where s_{cp} is the volume (s) accounted into categories c and p simultaneously. Concrete examples of this setting in economics can be the study of exports disaggregated by countries and products or accounting patents disaggregated by city and technology classes, as well as countless other scenarios that are mathematically equivalent.

Within that context, this paper is dedicated to studying a particular transformation of the observed volumes s_{cp} which has been called the 'location quotient' (LQ), and given by:

$$LQ_{cp} = \frac{s_{cp}/S_c}{S_p/S_W} = \frac{s_{cp}/S_p}{S_c/S_W} \tag{4.1}$$

where $S_c = \sum_p s_{cp}$ and $S_p = \sum_c s_{cp}$ are the total volumes of categories c, p (we will also call them the *sizes*). From eq. 4.1 one can see that LQ is the ratio between relative size of element s_{cp} on column p and relative size of row c in the matrix total. A transposed version of this statement is valid as well.

LQ has been introduced in studies involving international trade flows mostly after the antecedent of Balassa (1965), who coined the name of *revealed comparative advantage* index (RCA). The LQ index is mostly meant to be compared to a threshold value to provide a binary variable $LQ > th$. The value $th = 1$ is special because when $LQ_{cp} = 1$ from equation 4.2 we have:[1].

$$\frac{\hat{s}_{cp}}{S_W} = \frac{S_c}{S_W}\frac{S_p}{S_W}$$

which is the condition that distribution of values is independent across countries and products (regions and technologies) in determining expected $\hat{s}_{cp}$. Therefore when considering LQ we are comparing observed values to a model of row and columns as uncorrelated probabilities.

The LQ can essentially be taken as an outcome of combining two quantities: the observed levels and some expected levels. As such, the problem is two dimensional. For good expositions in this line of reasoning see Kunimoto (1977), and sections I, II in Bowen (1983), section IV in

[1] I use the notation $f(x) \equiv x > x_0$ for denoting the function $f : \mathbb{R} \rightarrow 0, 1$ that takes value 1 if $x > x_0$ and 0 otherwise. It is an expression that produces a boolean value.

Vollrath (1991).

In logarithmic scale LQ can be taken as a difference between two factors:[2]

$$\log(LQ_{cp}) = \log(s_{cp}) - \log\left(S_c S_p / S_W\right)$$

The goal of this work is to characterize *size effects* in LQ. That is, we want to measure whether LQ levels should be interpreted differently when computed for small entities as opposed to large entities.[3] The key innovation in this work, which has not been undertaken before, is using the probability $P(\log(LQ_{t+1}) > 0|\boldsymbol{x}_t)$ (chances of being above the threshold $LQ = 1$ at time $t+1$, conditional on the *state* $\boldsymbol{x}_t$ at time t) as an estimator of *distance to* $\log(LQ_{t+1}) = 0$. The so called state $\boldsymbol{x}_t$ defines any observation by its value in two coordinates, for example $\boldsymbol{x} = (\log(s_{cp}) = 6, \log\left(S_c S_p / S_W\right) = 5.5)$ means s_{cp} is higher than the expectation $S_c S_p / S_W$, in particular, this pair implies $log(LQ) = .5 > 0$.

With these tools we offer options for systematically studying open issues on the interpretation of LQ values other than 1 and understanding the nature of *size effects* by which there are distortions in the scale of LQ which depend on s_{cp}, S_c, S_p. This is an open issue that needs to be addressed if researchers want to arrive at trustworthy results when using LQ indices. In addition, understanding the comparability of LQ levels across entities can allow use of new techniques so far forbidden, such as robust replacement of the binary $LQ > 1$ by a set of LQ level categories, once size effects are controlled for.

The following sections are organized as follows. Section 4.2 reviews literature on location quotient indices. Section 4.4 presents a convenient mathematical framework for studying LQ indices and introduces the probabilistic LQ index, pLQ. Section 4.4.3 shows a possible method for estimating pLQ and section 4.4.5 shows a model that can explain the size effect patterns qualitatively. Section 4.5 concludes.

[2] This transformation is justified mostly because the levels s_{cp} themselves are well distributed in log scale. Apart from that, $\log(LQ_{cp})$ has multiple useful properties that LQ lacks, as will become evident throughout the paper.

[3] by entities I refer to any of the categories c, p

4.2 Literature Review

Apart from its use in international trade at least since Balassa (1965), LQ transformations have been used in classified patent counts at least since de Solla Price (1981) and Soete and Wyatt (1983). The exercises in Hidalgo and Hausmann (2009), Hidalgo et al. (2007), and Tacchella et al. (2012) use LQ in an early stage of data processing (exports data) and have revitalized its use on subfields of Economic Geography dealing with regional development and technological innovations, such as in Balland and Rigby (2016) relating patent categories to US cities. Of course there is a variety of options for categories to fill the roles of sets C, P used to classify the total, generalizations to more dimensions can be incorporated, and the choice of classifying criteria will depend the focus intended by each study. We can always opt between different variables as s_{cp}. Works such as F. Neffke et al. (2011) and Hausmann and Neffke (2016) describe the prominence of industries in regions using the size of workforce as indicator variable and relates industries with each other through the observed flow of workers between them. The LQ has been applied in studies of firm portfolios (Teece et al., 1994), plants in subnational regions (F. Neffke et al., 2014).

Hundreds of studies discuss outcomes of applying LQ on specific datasets, although only a minority of papers are dedicated to studying the index itself.

Within this latter group, most of the papers are embedded in the study of international trade. In this context, the name given to the index is 'revealed comparative advantage' (RCA). Interestingly a large part of these works incline towards discouraging the use of LQ. The two main arguments for this are that (i) the LQ index as is would not 'reveal comparative advantage' in light of existing trade models or theory, or (ii) certain characteristics of the observed LQ values (such as an asymmetry about its median values, difficulties in comparability across datasets) undermine its safe or consistent usability (for a clear and concise review of results in this line of reasoning see the section I in Liu and Gao (2019)).

Bowen (1983) and Leromain and Orefice (2014) are examples of studies that prioritize trade theories over the study of the LQ index itself. Unfortunately sometimes involving mathematically awkward expressions in the search for an index [4] or depending on specific trade models

[4] cf Cai and Leung (2008) replacing s_{cp} by nominal changes Δs_{cp} in the middle of the LQ ratio, or Hoen and Oosterhaven (2006) proposing to look at the difference of ratios $s_{cp}/S_c - S_p/S_W$, or Redding and Proudman (1998)

[5]. However, see French (2017) for a recent, fairly complete review of alternative RCA indices suited to a range of special applications.

This paper is centered on the index itself independent of the setting where it is applied. This is why I simultaneously consider LQ in the context of patent data and exports data, and why I do not intervene on the compatibility between LQ indices and international trade models. With respect to the second mentioned argumentative line often used to motivate departures from Balassa's LQ expression, I will argue that most of the alleged *problems* with the LQ index are not critical, and the decisions to abandon it have been likely premature. On top of that, the proposed alternative indices present important weaknesses.

The location quotient *as is*, is no more and no less than a ratio from observed data in a contingency table to the expectation from uncorrelated marginal probabilities. As such it is worth studying, and it will continue to appear in different contexts over the years. It has been systematically observed in empirical settings that the sizes of parts of a disaggregated total (also, the sizes shown by a population of agents) are well explained by a 'log' distribution (Axtell, 2001), such as lognormal, or power law (Pareto). This is also valid for the volumes of exports disaggregated by country, by product, or by both simultaneously. The same property is true for the number of patents classified by region and technology class. If the values $\log(s_{cp})$, $\log(S_c)$, $\log(S_p)$ are near normally distributed, one should expect $log(LQ) = \log(s_{cp}) - \log(S_c) - \log(S_p) + \log(S_W)$ to be near normally distributed too. In fact, in this paper I will work with the $log(LQ)$. Essentially it would not be a separate index from LQ. It is simply the same index in a different scale.

In the remaining of this section I will review the main perceived problems with the LQ index according to the literature. The first series of problems exist if we stick strictly to the values of LQ in linear scale. In that context, Hinloopen and Van Marrewijk (2001) and Hoen and Oosterhaven (2006) look at country mean and see that this moment is usually far above 1 (the neutral value), unstable over time and across countries (Leromain & Orefice, 2014; Yu et al., 2009). Other appreciations are that the shape of the distribution of the LQ is highly sensitive to extreme values (De_Benedictis & Tamberi, 2004; Hoen & Oosterhaven, 2006). Many of the

computing the ratio of observed shares to average shares (questioned already by De_Benedictis and Tamberi (2001)).

[5] For example Leromain and Orefice (2014) depending on Costinot 2012 trade model

comments and choices in these papers suggest that they are not embracing the idea that it is more reasonable to study the distribution of LQ levels in a logarithmic scale.

The strong attachment to values in linear scale is clear from phrases such as *"Hinloopen [...] empirically observe that the mean of the sectoral MRCAs (LQs) is well above 1. This seems strange as it suggests that each country has a comparative advantage in its 'average sector', whereas one would expect the 'average sector' to be neutral in terms of its MRCA. (LQ)"* (Hoen & Oosterhaven, 2006). Although it is perfectly natural that the nominal *average* of a log (normally) distributed variable never points to *an average sector.*

Another frequent observation from studies that stick to the linear levels of LQ and do not consider log levels is the "high skewness" (De_Benedictis & Tamberi, 2004; Hinloopen & Van Marrewijk, 2001; Laursen, 2015). And the asymmetry: *"the Ballasa Index (LQ) has a strongly asymmetric distribution with a fat right tail"* (Laursen, 2015). These issues are regarded mostly a problem for fitting regressions using LQ levels as dependent variable. As mentioned before, they motivate the creation of alternative indices, such as the *regression* or *symmetric* RCA, with the formula $SRCA = (LQ + 1)/(LQ - 1)$ (Dalum et al., 1998; Laursen, 1998) which is a near log transformation for non extreme values. This index may have improved properties with respect to LQ in linear scale, but its interpretation is not as straightforward (De_Benedictis & Tamberi, 2001) as it detaches further from actual observed values. In addition it complicates analytical treatment. Further suggestions of non acknowledgement of the possibility that LQ values are best expressed in logs are plots as in Hinloopen and Van Marrewijk (2001) where the curves of CDF of LQ values in linear scale are all accumulated on the left side of plots non allowing an appreciation of a possibly bell shaped distribution of $log(LQ)$ values.

The single most significant obstacle for using $log(LQ)$ is the presence of multiple null entries in any large contingency table (*"Unfortunately, LRCA (log(LQ)) becomes minus infinity when the export is zero."*, Liu and Gao (2019)). Rather than *minus infinity*, log of null values are *not defined.* To that issue, we can say that if the null values are few, they can simply be ignored for most uses. If instead the sparsity of the data is significant, when taking logs the zero values can be held on a separate account. Almost any desired study can still be performed by including a dummy for null values or by allowing extensive margins where there is the possibility that an observation becomes null, or jumps from being null into some positive level. In many practical cases, most of the value is concentrated on the largest agents, and so not only null entries, but also a large part of the smaller entries may not hold a significant share of total value, meaning that both small and null entries may not influence aggregate outcomes significantly anyway.

In short using log values of LQ does not imply a fatal problem, even if some of the entries were null. Actually log(LQ) has been used naturally and without further questioning in papers such as Vollrath (1991) (although quickly dismissed due to the null values issue) and also Bahar et al. (2014), Boschma et al. (2016), De_Benedictis and Tamberi (2001), Hidalgo and Hausmann (2009), Liu and Gao (2019), and Soete and Wyatt (1983) among others. In fact $log(LQ)$ fulfills some of the desirable conditions for a revealed comparative advantage index as asked by Hoen and Oosterhaven (2006), i.e. *the index has a stable mean or median*, *the index is symmetric around the mean or median* and *the index has a stable distribution.* Good performance of $log(LQ)$ in light of trade models has been verified in Deb and Hauk (2015).

Log levels are also the convention in information theory. For instance, the pointwise mutual information (PMI) measure of association is given by $log(\rho_{cp}/(\rho_c\rho_p))$ where ρ stand for probabilities. If we identify them with the probabilities observed in the data: $\rho_{cp} = s_{cp}/S_W$, $\rho_c = S_c/S_W$ and $\rho_p = S_p/S_W$, then we have $log(LQ) = PMI$.

The parallelisms continue, it is known, for instance that the maximum value of PMI is given by $max\{log(1/\rho_p), log(1/\rho_c)\} = log(1/\rho_{cp})$, which is attained either when activity p exists only in location c, or when activity p is the only one in location c(van Dam et al., 2020). This parallels the observation that LQ has an upper bound from country size and product size (exports, region size and technology class size in patents counts) (De_Benedictis & Tamberi, 2001). Consider that if $s_{cp} = S_c$ (the product is as large as it can get in the country) then $\bar{LQ} = (S_p/S_W)^{-1}$ and it should be lower than this in all other situations. Same can be derived in the transposed case $c \leftrightarrow p$. It means that for smaller countries and products LQ values have higher upper bounds. In practice, small countries can achieve LQ levels that large countries cannot.

This type of observations suggested that in general the comparability of LQ levels across entitites had open issues to be resolved. There is a consensus that values of LQ different from one do not have an absolute interpretation. These *size effects*, or distortions in the scale of LQ are sometimes acknowledged informally and sometimes mentioned explicitly, as when

(Leromain & Orefice, 2014) say *"[Balassa's LQ definition] might imply huge values of Balassa Index for very small countries".* Although there has not been an approach for a systematic study so far. Size effects have not been exactly defined, and their dependence with parameters of the problem has not been determined. Many papers in the literature simply apply cuts of LQ levels at fixed values across countries and products, implicitly ignoring the issue Boschma et al. (2016), De_Benedictis and Tamberi (2001, 2004), and Hinloopen and Van Marrewijk (2001). While it is possible that the outcomes of these studies are robust to such size effects, it is still important to understand their origin and characterize their magnitude so that they can be controlled for in future studies, or exploited in favor of specific research goals.

4.3 Data

We will use empirical information in two settings: value of exports by country and product, and number of patents by country and technological category. In the first case, if we say the matrix X describes the export volume s_{cp} of country c and product p. In the latter case, we can say a matrix N describes the number of patents n_{cp} of region c and technology p, and a ratio analogous to the one in 4.1 has been called index of Revealed Technological Advantage. In the next sections, we will use the name 'Location Quotient' to refer in general to these ratios, regardless of the context in which they are defined. Also, we will denote the matrix values as s_{cp}, and the marginal totals as S_c and S_p referring to any of the two mentioned settings indistinctly.

Empirical information on export flows (in US$) accounted to 235 'country' categories and 1244 HS02, 4 digit 'product' categories, for the 12 years within 2003 and 2014 (more than 1.7 million non zero entries) is sourced from UN COMTRADE, with files openly available (here) through the Atlas of Economic Complexity.

Information on number of patents is sourced from the OECD REGPAT database which collects records at the European Patent Office (EPO) and the Patent Cooperation Treaty (PCT) starting in 1978 up to 2015. We aggregate patents at 639 first level subnational units ('regions') and a total of 124 technological classes ('technologies'). We count with a total of 480 thousand observations.

4.4 Analytic steps

4.4.1 The 2D space for log location quotients

In this section we will first introduce convenient representations for the 2D space that characterizes all possible LQ values. Then, we will introduce the notion of a *probabilistic location quotient* (pLQ), which is simply defined as the chances that $LQ_{t+1} > 1$ conditional on the situation at t, which is described by point in the mentioned 2D space. The section continues discussing results and applications.

We have seen that LQ can be written as:

$$LQ_{cp} = s_{cp} / \left(\frac{S_c S_p}{S_W} \right) \tag{4.2}$$

where subindices c, p refer to a country and product category, and S_W is the total sum of the matrix, that is the total registered. I will call S_c and S_p the *size* of the categories c and p. To be precise we should call it the total volume of exports in US$ (total number of patents) assigned to category c, p, respectively. The matrix X is typically measured annually. I do not write time index unless it is necessary. [6]

The logarithm of the definition in 4.2 converts the ratio into a difference:

$$\log(LQ_{cp}) = \log(s_{cp}) - \log\left(S_c S_p / S_W\right) \tag{4.3}$$

And it turns out that the empirical distribution of these three terms is nicely bounded in log scale, as can be seen from Figure 4.1.

Expressing location quotient as difference hints to a two dimensional space for the problem. Indeed, equation 4.3 is one condition involving three independent terms, and so the system is 2 dimensional. Denote: $y \equiv log(LQ_{cp})$, $x_1 \equiv log(s_{cp})$, and $x_2 \equiv \log\left(S_c S_p / S_W\right) = \log(S_p) + \log(S_c) - \log(S_W)$ which I call the *size factor.*

[6] All observations necessarily belong to some year or time period and if there is no time index, all variables in the equation belong to the same time period.

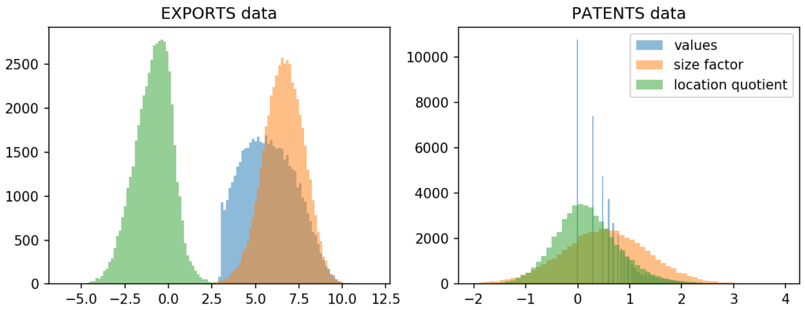


Figure 4.1: Distribution of $log(s)$, T and $log(LQ) \equiv log(s) - T$. The cut at $log(s) = 3$ reveals the lower bound in $s_{cp} = 1000US\$$ in the dataset. These distributions of log variables are much 'better behaved' than those of the original LQ, x, or $S_c S_p / S_W$. The distribution of number of patents is largely constrained to the natural numbers, and it is frequent to observe small values. This feature is not very problematic. It explains the fragmented appearance of its histogram and scatter plots in Figure 4.2.

$$\begin{aligned} log(LQ) &= log(s) - [\log(S_p) + \log(S_c) - \log(S_W)] \\ y &= x_1 - x_2 \end{aligned} \tag{4.4}$$

Depending on the use one wants to pursue, we may choose to describe observations in the axes LQ vs observed value (y, x_1), LQ vs size factor (y, x_2), observed value and size factor (x_1, x_2) (right plots in Figure 4.2), or difference and mean ($y = x_1 - x_2$, $x = (x_1 + x_2)/2$) (left plots in Figure 4.2). The take away is that the *state* of country-product (region-technology) at year t is fully determined by knowing the values for any independent pair of these variables. I call the variable $(x_1 + x_2)/2$ *mean size factor*. On the mid of the transition where $log(LQ) = 0$, we have $x_1 = x_2$ and so the mean size factor summarizes both the magnitude of x_1 and x_2. It comes in handy as direction independent from $log(LQ) = x_1 - x_2$.

4.4.2 Probability that $LQ_{cp} > 1$ at time $t + 1$ conditional on the state at t.

Many works seek to study how likely it is that the state of a country-product (region-technology) has $LQ > 1$, or whether there are factors that increase the chances that the state will pass from $LQ < 1$ to $LQ > 1$ or vice versa (eg. Boschma et al. (2014), Hausmann and Klinger (2007), and F. Neffke et al. (2014), among many others). The factors considered are usually motivated

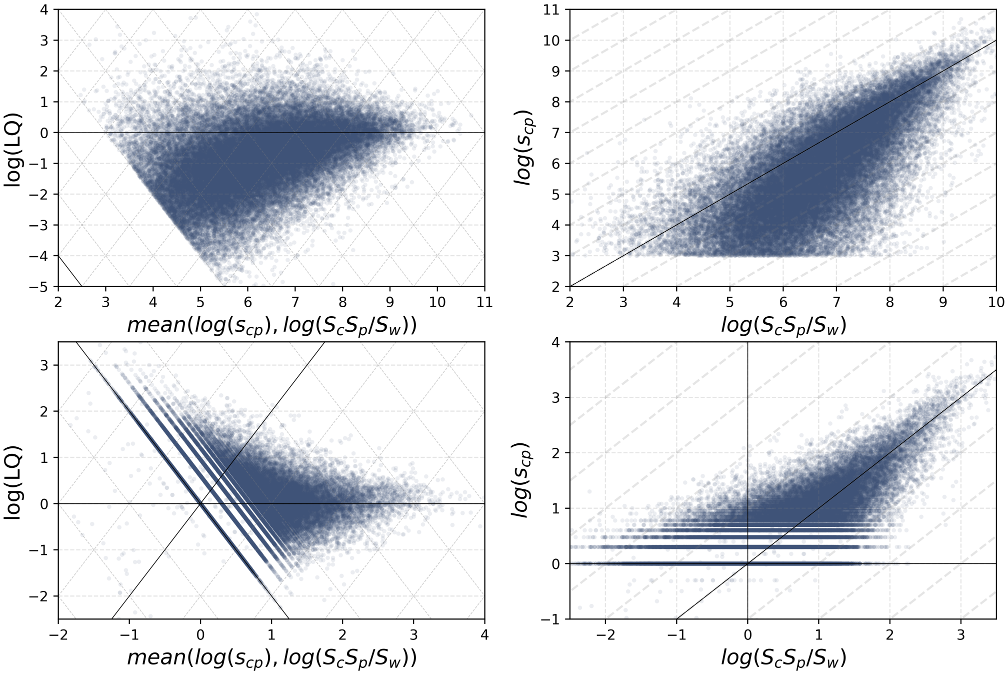


Figure 4.2: Distributions of observed points in log scale. Here the dots represent actual data points in our dataset. Plots on the top are for trade data. Plots on the bottom for patent counts data. Left: points in $((x_1 + x_2)/2, log(LQ) = x_1 - x_2)$ coordinates. Right, points in (x_2, x_1) coordinates, with $x_1 = los(s_{cp})$ and $x_2 = log(S_cS_p/S_W)$. Left plots can be transformed to right plots by applying a linear transformation. Dark lines indicate axes $log(LQ)$, x_1, x_2. A log transformation lets all observations form a bounded cloud.

by theory or specific research questions. The very variable LQ, (or $\log(LQ)$ or $LQ_t > 1$) is often included as predictor, and this is completely reasonable given that if observations are relatively stable LQ_{t+1} is likely to be near LQ_t and the condition $LQ > 1$ can be expected to persist over time.

If we use LQ_t as the sole regressor for estimating the probability that $LQ_{t+1} > 1$ we confirm this intuition. This is shown in Figure 4.3 where we can see that being above or below the threshold $LQ = 1$ is an important factor in determining whether $LQ_{t+1} > 1$.

Note that the probability $P(log(LQ_{t+1}) > 0| \log(LQ)_t)$ approaches 0.5 when $\log(LQ) \approx 0$ and in the limits of very low and very high $log(LQ)$, this probability tends to approach zero and one respectively. This initial approach points to an extremely useful feature: an interpolation between the 0, 1 values, that complements the discrete threshold by adding some *structure* to this jump. Through $P(log(LQ_{t+1}) > 0| \log(LQ)_t)$ we obtain information about the width of

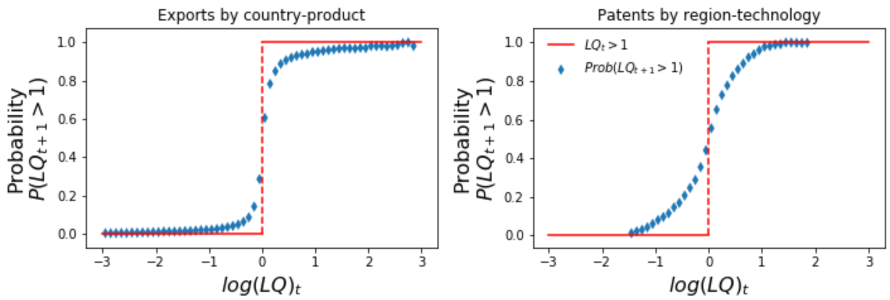


Figure 4.3: The discrete variable $LQ > 1$ (red), and the probability that $P(log(LQ_{t+1}) > 0| \log(LQ)_t)$ (blue). Plotted as a function of $\log(LQ)_t$.In the extremes both coincide, but near the threshold of $log(LQ)_t \approx 0$, the latter one provides a natural interpolation between the two values. Each dataset shows a certain transition width suggesting that the scale of LQ is not unique for all datasets.

this jump measured from the chances that an observation will jump over the threshold in one time period.

Another possible interpretation of this interpolation is as an effective distance to the $LQ = 1$ threshold. This is one of the key points we are putting forward in this work. Usually, we have no information as to whether a value of, say $LQ = 0.8$ is close enough to $LQ = 1$ and there are open debates as to how to treat values $LQ \neq 1$. However, in this way we can precise that such a value meant an 18% chance of $LQ_{t+1} > 1$ in the dataset. I suggest an interpretation of the gap $0.8 \to 1$ as the chances of seeing $LQ_{t+1} > 1$ given $LQ_t = 0.8$.

For the next step, recall that the observations in the context of location quotients are determined by *two* independent variables. So that we are compelled to computing the probabilities that an observation will be $LQ_{t+1} > 1$ conditional not only on LQ_t but on an additional coordinate simultaneously. Actually any pair of independent coordinates works, and I choose to use $(S_pS_c/S_W)_t$, s_cp.

$$\begin{aligned} pLQ &= P(LQ_{t+1} > 1 \mid (S_pS_c/S_W)_t, s_t) \\ &= P(s_{t+1} > (S_pS_c/S_w)_{t+1} \mid (S_pS_c/S_W)_t, s_t) \end{aligned} \tag{4.5}$$

The indices cp have been omitted. In practice, it is more convenient to use the log levels, profiting from the amenable the characteristics of their distribution:

$$\boxed{\begin{aligned} pLQ &= P(log(LQ)_{t+1} > 0 \mid x_{2,t}, log(s_t)) \\ &= P(log(s)_{t+1} > x_{2,t+1} \mid x_{2,t}, log(s_t)) \end{aligned} \quad (4.6)}$$

where $x_2 = \log(S_p S_c / S_w)$ is the size factor.

A precise characterization of pLQ should be a baseline to be studied before a variety of external factors are introduced to explain $P(log(LQ)_{t+1} > 0)$. To the best of my knowledge however, the approach presented here has not been pursued elsewhere, or at least it does not seem to have been published. As a by product we will be able to characterize the effect that the size of countries and products (regions, technologies, etc) have on the time evolution of LQ, and so in the LQ index itself. As discussed previously, such distortions have been partially acknowledged for long but not measured satisfactorily. We therefore offer a step towards understanding the effects that complicate comparison of LQ across countries, products and time periods.

Indeed, it is interesting to consider how we should interpret level curves of pLQ and how they compare to levels of LQ. For example, consider two countries that for some product and year show the same level of LQ (say $LQ_0 = 0.5$) but different levels of pLQ (such that we expect, say 1% vs 10% chances of surpassing the threshold in the next year). Are they equally distant to the $LQ > 1$ situation? On the same line: is it a desirable property for an index of comparative advantage to be not dependent on the size of the country (product, etc) involved? If the answer is yes, then we would want to acknowledge effects of the type discussed here and be able to counter them out.

In the next section (4.4.3) I offer a method for estimation of pLQ, and in section 4.4.5 I discuss numerical reconstructions of pLQ derived from models of the growth rates of $log(s)$ and $log(S_c S_p / S_W)$.

4.4.3 Estimation of pLQ

We estimate $Prob(LQ_{cp,t+1} > 1)$ given $log(s_{cp})$ and $log(S_c S_p / S_W)$ by exploiting the information from all points observed empirically. The problem is analogous to computing a density function, that is, what fraction of the points in a certain small region of the ($log(s)$, $log(S_cS_p/S_W)$) plane fulfil the condition ($LQ_{cp,t+1} > 1$). There are a few methods available for this task, I decide for a k-nearest neighbor algorithm (knn), even if there are multiple alternatives (i.e. partitioning the space in bins and computing their fraction of ($LQ_{t+1} > 1$) in them). In the knn method, we use the k nearest points in the ($log(s)$, $log(S_cS_p/S_W)$) plane (called *feature space* in the context of knn regression) are used to compute the ($LQ_{t+1} > 1$) fraction, and assign such value to the probed point. A key advantage of using knn is that the set of neighbors has a fixed size (I use $k = 200$ in both datasets) so that sparse regions do not loose statistical robustness, and the high density of other regions leads to modelling at finer grained resolution. [7]

The outcomes are plotted in Figure 4.6, there I plot the estimated pLQ values as a function of the mean size factor and $log(LQ)$ (left) and as a function of the observed values $log(s)$ and $log(LQ)$ (right). The coordinates used are those demonstrated in Figure 4.2. Top plots refer to trade data bottom plots to patent data. We can take both an ideal continuous probability function and its estimations from empirical data as probabilistic location quotients: *pLQ*. [8]

First and foremost, as seen before in Figure 4.3 it is clear that the space is split in two regions (red, green). These largely correspond to the condition $LQ > 0 \leftrightarrow x_1 > x_2$ (green if true, red if false).

Secondly, the most interesting feature is the transition zone ($0 \to 1$) in between the red, green regions. It shows a width, and this width changes along with the sizes of the involved observations. If we say that the width of the transition between probabilities 0 and 1 is an indicator of the inherent scale of the LQ values involved, then we are observing how size of the c, p entities involved affect the scale of the LQ index. These are precisely the effects we are seeking to characterize in this study.

More concretely, take the trade example: the plot of Figure 4.6 shows how for countries

[7] See Wu et al. (2008) for a review of the knn method in the context of other regressor algorithms. See also Loftsgaarden and Quesenberry (1965) for an early discussion of the knn concept.

[8] A minimal snippet of code (python), which anyone can use to estimate pLQ given a dataset of observations is in the Appendix. Refer to it for essential precisions on what the knn regressor is doing. Given any two independent combinations of $log(LQ)$, $log(s_{cp})$ or $log(S_pS_c/S_W)$ the knn regressor can be used to predict pLQ.

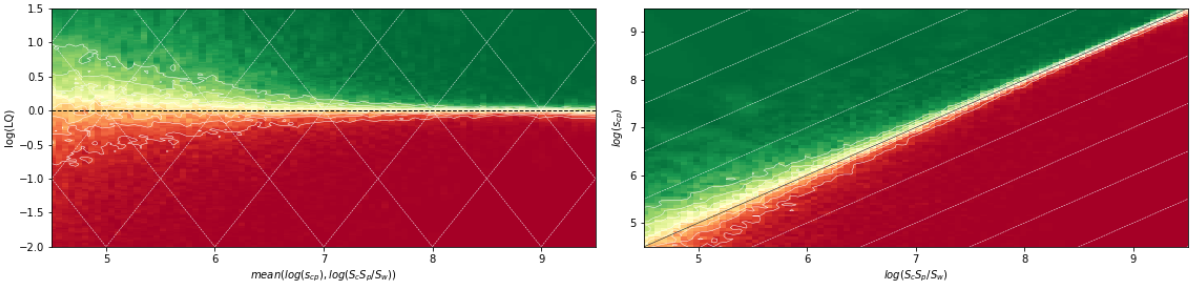


Figure 4.4: pLQ. Trade data.

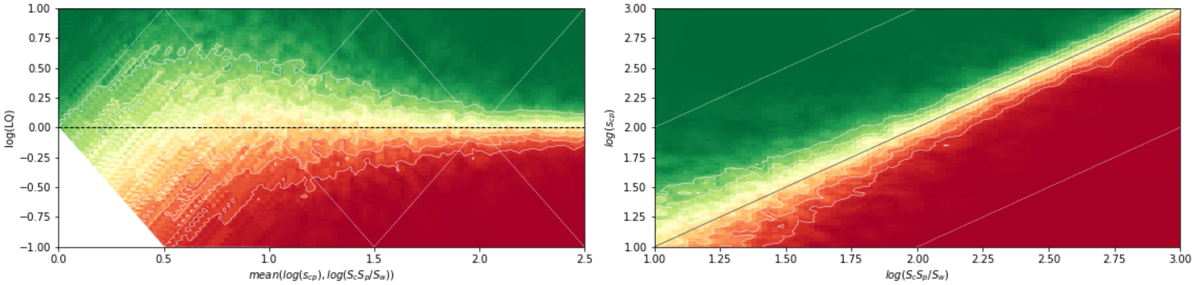


Figure 4.5: pLQ. Patents data.

Figure 4.6: Plots of probability $pLQ = P(log(LQ)_{cp,t+1} > 0|log(s_{cp}), log(S_pS_c/S_W))$ as a function of $log(S_pS_c/S_W)$ and $log(LQ)$ (left) and $log(s_{cp})$ and $log(LQ)$ (right), computed on the trade dataset (top) and the patents dataset (bottom).
These probability densities are computed applying k-nearest neighbor algorithm (k = 200) to a training set with at least a few hundred thousand observations of consecutive years. The red - yellow - green scale is used for low medium - high probabilities.
The condition $log(LQ)_t > 0$ arises as the main determinant for whether $log(LQ)_{t+1} > 0$. However, we can also see the role played by the sizes of the observed values: the transition $pLQ = 0$ to $pLQ = 1$ becomes narrower as the sizes of the involved observations increase.

and products with the same level of $LQ < 1$, the ones of smaller size have higher chances of achieving $LQ > 1$ the following year. This analysis therefore serves to probe how country-products (region-technologies) of equal LQ and different sizes may imply different situations regarding their possibility to surpass the LQ = 1 threshold. It can be used to decide *how* LQ values are not comparable across countries or products.

To have an impression of the relevance of these effects, from the trade dataset, an LQ = 0.5 ($log(LQ) \approx -0.3$) for a product of average size this LQ means more than 15% chance of getting to $LQ > 1$ for small economies, a 5% chance for mid-sized countries, and between 1% and 2% for large economies.

The level curves of pLQ are important as an a-posteriori estimation of starting points equally distant to the $LQ = 1$ threshold. As such they offer a path for controlling size effects on the LQ metric. Additionally, using them as border between categories allows mapping the observations to intervals of qualitative intensity which controls by entities' sizes.

Studies that use levels of $LQ \neq 1$ without acknowledging size effects need to be revised following these observations. Levels of pLQ can be taken as levels of LQ corrected by size effects. In section 4.4.4 I test the persistence of pLQ levels by means of a Markovian transitions analysis to find that pLQ levels near 0, 1 are stable and that the jump $0 \to 1$ (and viceversa) usually involves stepping on intermediate values.

4.4.4 Dynamics

Figure 4.7 plots pLQ_t (horizontal axis) vs pLQ_{t+1} (vertical axis). This can be thought of as an empirical Markov matrix. Some features are clearly distinguishable. The extremes of very low and high pLQ are stable: Most points in there continue to have similar values when time evolves one step. Very different dynamics characterize the situations of intermediate pLQ levels. Volatility of pLQ is higher and there are significant chances of having any pLQ at the end of the period.

If we partition pLQ in three categories:

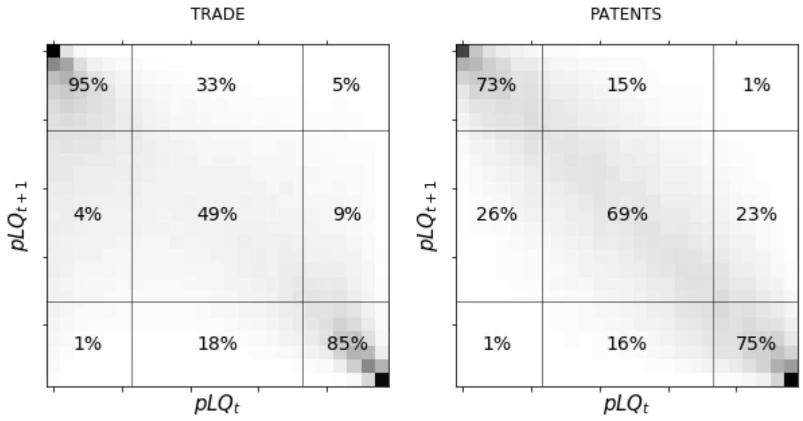


Figure 4.7: Plot of an empirical Markov matrix: pLQ next year as a function of pLQ this year. Both extremes show persistence, while the values in the middle show a high volatility. This result suggests a split of the range in three categories and here I choose the thresholds .2 and .8 for illustration. The percentages (sum 100% vertically) indicate transition probabilities from categories at time t to those at time $t+1$. They suggest that points are likely to transit the middle category to jump between the extremes.

1 low values ($0 <$ pLQ < 0.25)

2 medium or transition values ($0.25 <$ pLQ < 0.75)

3 high values ($0.75 <$ pLQ < 1)

then we can compute the 3×3 Markov matrix and we see a further interesting feature (numbers annotated on the blocks of matrix, Figure 4.7). There is significant probability to jump between categories 1 and 2, as well as between categories 2 and 3. However the probabilities of direct jumps between categories 1 and 3 without passing through 2 are reduced. Is this possibly revealing an ordering in LQ values? In the sense that the transition category, which maps to a region of the (LQ, size factors) plane, seems to be a stage that country-products typically go through in their drift from no-advantage to advantage status, and viceversa.

Studies that use $LQ > 1$ for defining a binary matrix and corresponding bipartite network, can use pLQ for a weighted bipartite network. Also, from this Markov analysis we confirm that pLQ levels allow to define categorical {low, mid, high} values in place of the binary {0, 1}. This is discussed in section 4.4.5.

4.4.5 pLQ as continuous generalization of $LQ > 1$

The $LQ > 1$ binary variable is usually taken as a proxy for telling which entries in the contingency table are significant. The variable pLQ has (by definition) a value between zero and one, and each country-product pair is assigned one. We have seen that it matches the binary 0, 1 in most cases, but takes intermediate values when $LQ \approx 1$ (see fig 4.3)

If the variable $LQ > 1$ is arranged in a rectangular matrix $M \in \mathbb{R}^{n_C \times n_P}$, where columns correspond to the n_P products and rows correspond to the n_C countries, it can be interpreted as the adjacency matrix of a bipartite network (as in Hidalgo et al. (2007)). Such bipartite network is undirected and its two sets of nodes are 'countries' and 'products'. If we replace $LQ > 1$ for pLQ, the only difference is that possible values are not only 0, 1, but all those in the interval [0, 1] and hence it should instead be interpreted as a *weighted* binary matrix. The analysis based on binary bipartite networks as in Hidalgo et al. (2007) can be easily adapted to this more general case.

We can use this consistent criteria for turning the binary LQ variable into a categorical. This is illustrated in Figure 4.8, which plots the matrix M_{cp} (where rows represent countries and columns represent products) both in a binary fashion and in categorical.

Estimates of pLQ by integrating growth distributions

So far pLQ has been estimated from a nearest neighbors (knn) approach. Next, in section 4.4.5 I show that it matches estimation by assuming independent growth distributions of $x_1 = \log(s_{cp})$, $x_2 = \log\left(S_c S_p / S_W\right)$, estimating them and integrating them numerically to find the chances $Prob(LQ_{t+1} > 1|x_1, x_2)$. In this way we have elements to qualitatively explain the patterns of size effects observed.

There are formal ways to model the outcomes observed from knn regressors. By characterizing the typical fluctuations of observed points one can compute the chances that they are above $log(LQ) = 0$ after a time period.

As was done in section 4.4.2 I first present the setting in one dimension, using the variable $\log(LQ)$ before extending the discussion to the general 2 dimensions of the $\log(LQ)$ system.

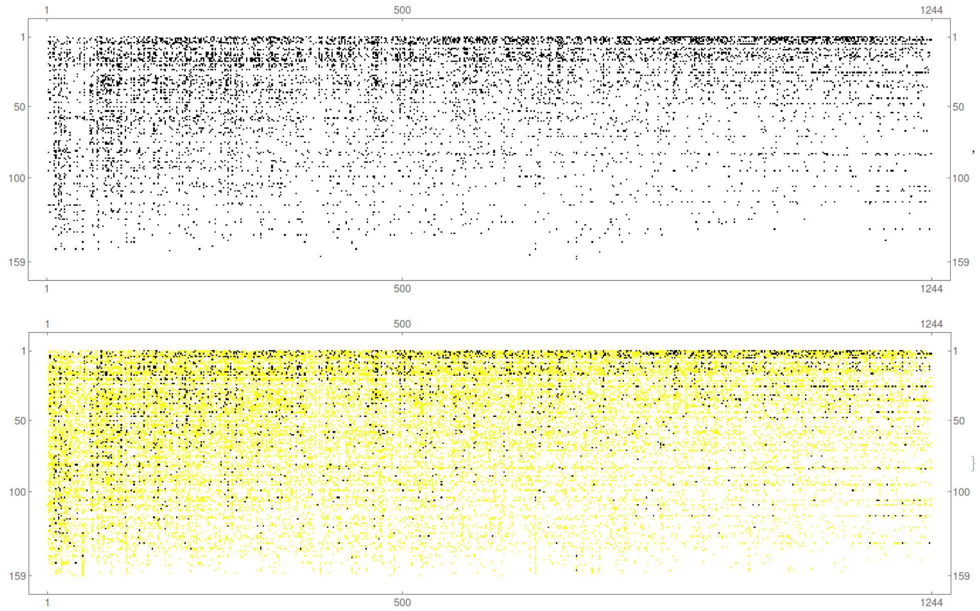


Figure 4.8: Matrix plot demonstrating the result of including an intermediate category in the binary country-product matrix M_{cp} of year 2014. Columns represent products, rows represent countries, ordered by ubiquity and diversity respectively. LQ = 0 (1) is represented in white (black) and values in the transition category of pLQ are in yellow.

We will consider differences in log levels. A condition for satisfactory results of the technique we will use is that the distribution of such differences has a nearly continuous support. More concretely: in the case of patent data, a major part of observations are below $n_{cp} = 10$ (fig 4.1, 4.2) and log differences among these levels are constrained to very few values strongly influenced by the levels of the first few natural numbers. Trade data has much higher values (minimum of 10^3) an so it is perfectly suited to the analysis of this section.

The definition of pLQ as the probability that $log(LQ) > 0$ within a time period links it directly with the growth distributions of $\log(LQ)$ (or for the same matter, those of LQ). To see this: if we determine that after a time period the points in the neighborhood of LQ_0 present a shift $\Delta \log(LQ) = \log(LQ) - \log(LQ_0)$ distributed according to the probability density function:

$$g_0(\Delta \log(LQ))$$

Then the condition $log(LQ)_{t+1} > 0$ will be fulfilled if:

$$log(LQ)_t + \Delta log(LQ) > 0 \qquad \Longleftrightarrow \qquad \Delta log(LQ) > -log(LQ)_t$$

so that for estimating pLQ we need to sum all the chances that $\Delta\log(LQ)$ is larger than the gap $-log(LQ)_t$ from $\log(LQ)_t$ to zero. Formally, this is the following integral:

$$pLQ(LQ_0) = \int_{-log(LQ_0)}^{\infty} g_0(\Delta\log(LQ))\, d\log(LQ) \tag{4.7}$$

This is illustrated in Figure 4.9.

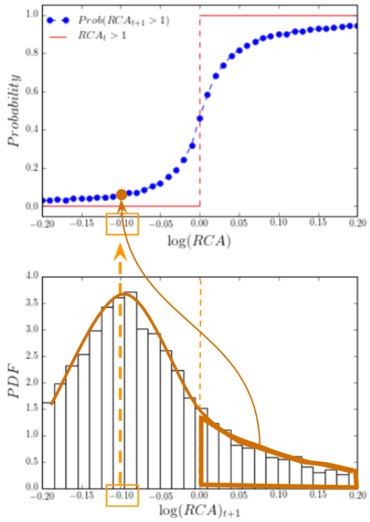


Figure 4.9: Growth distribution of $\log(LQ)$ and pLQ in empirical and analytical form. Top: probabilities that $LQ > 1$ in the next period as a function of $log(LQ)$. This is the same plot of Figure 4.3 but in a thinner horizontal range. The variable T is aggregated out for illustration purposes. Bottom: the histogram shows where all points of the vicinity of $\log(LQ) = -0.1$ ended up in the following period. This histogram is an empirical version of the growth distribution g_{p_0}. Highlighted are those points which surpassed the $\log(LQ) = 0$ threshold. Their area is equal to the height of the corresponding dot in the plot above, and it corresponds to the integral in equation 4.7.

This is how integration of growth distributions should work as estimation of pLQ. Now we can extend the reasoning to the pair of variables $x_1 = \log(s_{cp})$, $x_2 = \log\left(S_c S_p / S_W\right)$ that fully describe the 2D space for the LQ problem. Denote points in this coordinates as $\mathbf{x} = (x_2, x_1)$, denote $G_0(\Delta x_2, \Delta x_1)$ as the two dimensional distribution of growth rates of these coordinates, and denote $(x_1 > x_2)$ as the two dimensional region of the plane (R) where $\log(s_{cp}) > \log\left(S_c S_p / S_W\right) \Rightarrow s_{cp} > S_c S_p / S_W$, i.e. $LQ > 1$. The cases of growth resulting in $LQ > 1$ condition are, in calculus notation:

$$pLQ(\mathbf{x}_0) = \iint_R G_0(\Delta x_2, \Delta x_1)\,(x_1 > x_2)\, dR \tag{4.8}$$

This integration would be approximated numerically if we had the 2D distribution of growth out of the **x** coordinates stored in an array `G` and the condition $x_1 > x_2$ stored in another array `C` of the same shape and relating to the same (large enough) rectangular region. In this notation the numerical integration in equation 4.8 can simply be:[9]

$$pLQ(\mathbf{x}_0) = \texttt{(G * C).sum() / G.sum()} \tag{4.9}$$

The particular approach I choose for estimating this is to assume the growth rate distribution to be separable, i.e. express it as product of two one dimensional marginal growth distributions: $G_0(\Delta x_2, \Delta x_1) = g_{x_2}(\Delta x_2) g_{x_1}(\Delta x_1)$ and integrate it numerically in the $x_1 > x_2$ region.

Here it is key to observe that the typical width of log level fluctuation decays with increasing size. This applies both to the actual $x_1 = log(s_c p)$ values and to the size factor $x_2 = log(S_c) + log(S_p) + log(S_W)$ (Figure 4.10). This observation is in line with the results in H. R. Stanley et al. (1996) and other studies looking into volatility decay with agent size.

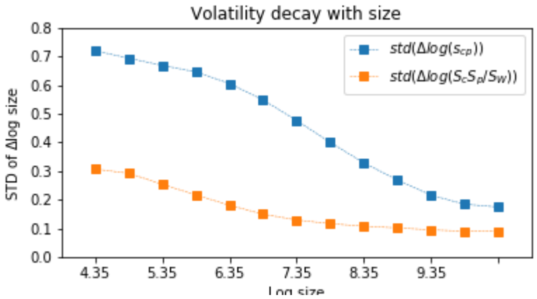


Figure 4.10: Volatility versus size. Larger observations fluctuate less and as such they are less likely to traverse a given gap in LQ levels. The standard deviation in these plots is the width of the axis of ellipses in figures 4.11 and 4.12.

For the problem of location quotients, the decay of volatility with size implies that large observations and observations from large countries and products are less volatile. This single qualitative feature results on (eg.) less likelihood for a large observation s_1 to surpass $LQ = 1$, compared to a smaller observation s_2, considering they start at the same $LQ < 1$ level.

The plots of Figure 4.11 illustrate the model of uncorrelated growth rates in the variables $x_1 = log(s_{cp})$ and $x_2 = log(S_c S_p / S_W)$. The ellipses show the standard deviation of these variables. We can qualitatively explain changes in the transition width of pLQ as a consequence

[9] Python language. Arrays are numpy arrays. Product `*` is element wise and `.sum()` is the sum of all array elements.

of the dependence of the moments $std(x_1)$ and $std(x_2)$ with x_1, x_2.

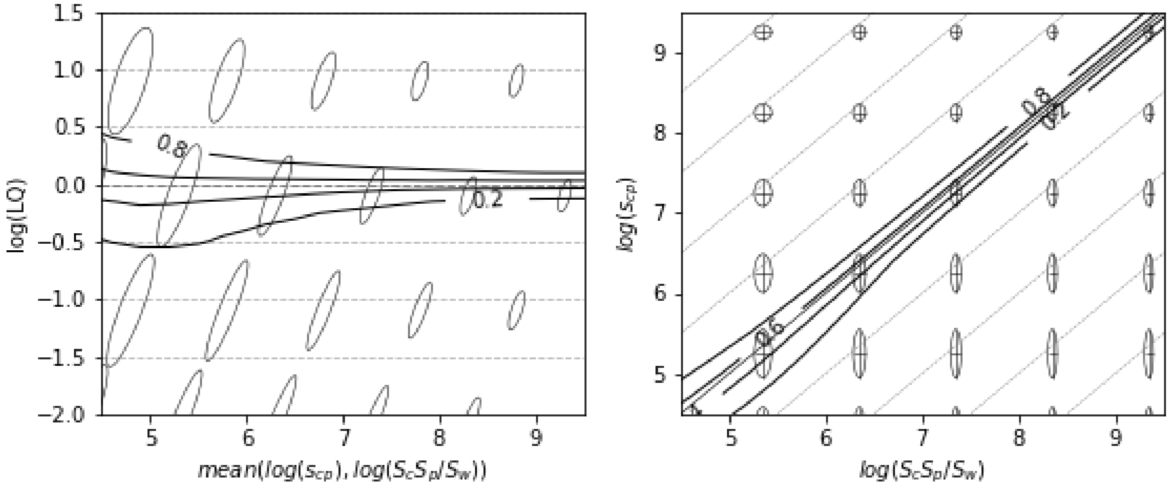


Figure 4.11: Magnitude of growth rates (ellipses) and level curves of numerically integrated pLQ as in eq. 4.8. Left: plot in coordinates $((x_1+x_2)/2, log(LQ) = x_1-x_2)$. Right: plot in coordinates (x_2, x_1). Growth rate distribution is modelled as independent product of marginal growth rates in coordinates (x_2, x_1), i.e. $G_0(\Delta x_2, \Delta x_1) = g_{x_2}(\Delta x_2) g_{x_1}(\Delta x_1)$. Widths of Δx_1, Δx_2 decay monotonously with x_1, x_2 and result in narrower transition pLQ = 0 to pLQ = 1 for larger entities.

For the numerical integration of these 2D growth rate probabilities I apply the following procedure. I bin observations into intervals which have bin centers $\{x_1\}$. Comparing observations at consecutive time periods, for each of these bins we can observe histograms of Δx_1. These histograms, normalized by the bin population serve as estimate of an assumed $g_{x_1}(\Delta x_1)$. From these I create a 2 dimensional continuous interpolator $g_{x_1}(\Delta x_1)$ that will tell the chances of any x_1 to become $x_1 + \Delta x_1$ after one time period. The procedure is applied on the x_2 variable as well, and we finally estimate G_0 as $g_{x_2} g_{x_1}$for chances of jumping in the 2D LQ plane.

The numerical integration is performed by evaluating growth rates interpolators in a fine grid covering a large rectangle about point $\mathbf{x} = (x_2, x_1)$ and storing it in a 2D numpy array `G`. Then, the condition $x_1 > x_2$ is stored in another array `C` of the same shape and relating to the same rectangular region. The numerical integration in eq. 4.8 is performed simply by computing `(G * C).sum() / G.sum()` as in eq. 4.9.

In Figure 4.11 I show the level curves of the pLQ from integration of growth rates, together with indicators of the width of changes in x_1, x_2 in ellipses.

To conclude on the validity of this model for reconstructing pLQ, I compare with level

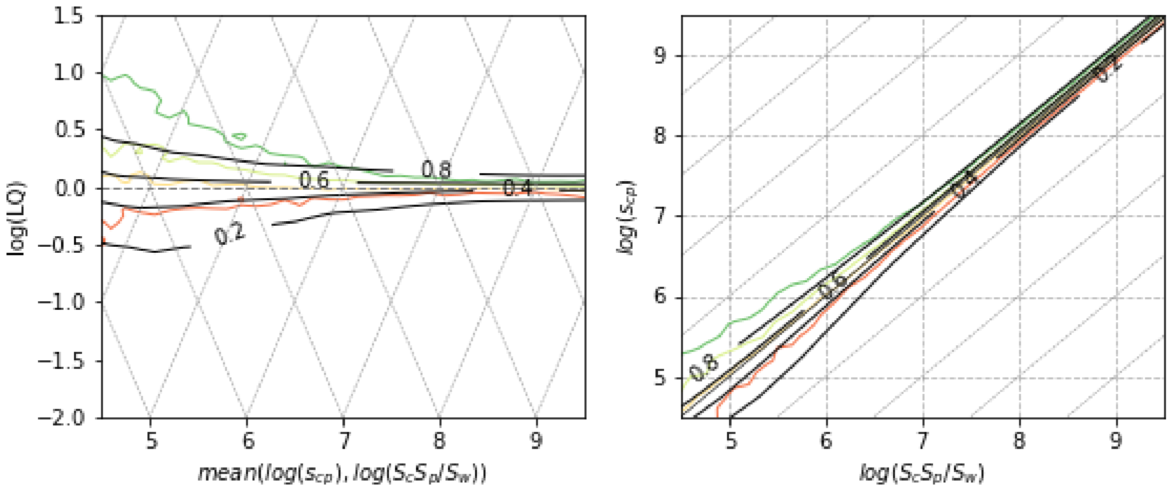


Figure 4.12: Direct comparison of level curves of pLQ estimated from knn algorithm (colors) and from integrated growth rates (black) (eq. 4.8). Left: plot in coordinates $((x_1 + x_2)/2,\ log(LQ) = x_1 - x_2)$. Right: plot in coordinates $(x_2,\ x_1)$. We find a qualitative match, confirming that size effects of LQ can be explained by the monotonous dependence of $std(\Delta x_1)$, $std(\Delta x_2)$ with $x1$, $x2$.

curves of pLQ from the knn estimator. From this we can observe that there is a qualitative match in the patterns of widening of the effective LQ metric with decreasing x_1, x_2 values. On the transition $x_1 = x_2 = (x_1 + x_2)/2$ and size effects can be evaluated by measuring $std(x_1)$ and $std(x_2)$ along this line.

4.5 Conclusion

In certain empirical contexts where a total can be disaggregated by two independent sets of categories (c, p), the location quotient (LQ) arises as a natural metric that compares the observed magnitudes (s_{cp}) relative to the expectation from independent marginal probability of its categories (S_c, S_p).

This paper studies to what extent the values of LQ indices from different observations (c, p) can be compared to each other.

To deal with this problem, we first suggest to work with the log transformation of LQ and express it as the difference $log(LQ) = log(s_{cp}) - log(S_c S_p / S_W)$, where S_W is the dataset total. This constrain involving three variables implies that all possible configurations of the problem are fully determined by knowing two (independent combinations) of these variables.

Secondly, I propose to use the observed probabilities that $LQ_{t+1} > 1$, conditional on the state at time t ($log(LQ_t)$, $log(s_{cpt})$) as effective measures of distance of LQ_t to the $LQ = 1$ threshold. Therefore, possibly different LQ values from two observations are considered equivalent if given their $log(s_{cpt})$ magnitudes (alternatively given their $log(S_c S_p / S_W)$ magnitudes) they show the same chances that $LQ_{t+1} > 1$. I name this variable as probabilistic LQ (pLQ).

For estimating pLQ I offer two alternative methods. Fitting a k nearest neighbors model, and using estimations of the growth rates of $log(s)$ and $log(S_c S_p / S_W)$. The first one is recommended for the goal of estimating level curves of pLQ and from there transforming to categorical levels. The second method is offered as a feasible explanation for the patterns observed in the level curves of pLQ.

The usefulness of the approach we present are multiple. On the formal side it allows measurements of such distortions, or *size effects*, that have otherwise been elusive. On the practical side, level curves of pLQ suggest effectively equivalent distance to LQ = 1, and allow conversion to consistent categorical LQ classes (eg. low LQ, intermediate LQ, high LQ). By applying this logic we can translate LQ measurements computed from different entities into a single consistent scale, countering size effects in a controlled way. Indeed there are multiple published results which use LQ as dependent or independent variable in regressions without controlling for size effects. We offer insights for attempting their revision. Eventually, exploiting our approach LQ values computed at different points in time or from different datasets can as well be cast into a single universal scale. This paper hopes to bring about easier comparability of published research by having offered formal tools to approach concerns on the nature of location quotient indices.

4.6 Appendix: pLQ regressor

The following few lines can be used to compute the pLQ probabilities in customized settings.

```
import pandas as pd
from sklearn import neighbors


def pLQ_regressor(df, n):
    """Estimate a pLQ regressor
    Arguments:
    df -- (pandas dataframe)
        Input data. must contain columns for:
        - The log observed values at t ('log_s')
        - The log size factor Sc Sp / Sw at t ('log_size_factor')
        - Whether LQ > 1 at t + 1 ('LQ_t+1')
    n -- number of neighbors
    """

    # Prepare X, y data for knn
    M = df[['log_s','log_size_factor','LQ_t+1']].as_matrix()
    X, y = M[:,:2], M[:, 2]

    # Fit
    knn = neighbors.KNeighborsRegressor(n_neighbors = n, weights = 'uniform
       ').fit(X, y)

    return knn
```

And it is used as in the following example:

```
# Load the data as pandas DataFrame
df = pd.read_csv('./data.csv')

# Fit the regressor
knn = pLQ_regressor(df)

# Ask pLQ at a point i
pLQ_i = knn.predict([log_s_i, log_size_factor_i])
```

Chapter 5

Related variety, economic complexity and the product space

[1]

[1] This chapter is being prepared for submission as F. Neffke, A. van Dam, C. Bottai, M. Iglesias, S. Orazbayeb, R. Hausmann and K. Frenken. The concept of diversity in economic geography: related variety, economic complexity and the product space.

Abstract

The last fifteen years have witnessed a renewed interest in the role of diversity in local economies. Here, we discuss three contributions to this literature: the notion of related and unrelated variety, economic complexity and the path dependent diversification patterns described in the work on product and industry spaces. Although these three different lines of research share many commonalities, we describe how they differ fundamentally in some of their ontological starting points. Moreover, we argue that there is substantial distance between some of the conceptual considerations in these approaches and their empirical implementation. Finally, building on work in ecology, we describe how to quantify and decompose diversity into three components: the variety of industries in a city, the balance of employment across these industries and the disparity among them. Armed with these tools, we show how more or less equally defensible modeling approaches yield different answers to the main hypotheses put forward in the research on diversity, diversification and growth in US cities.

5.1 Introduction

One of the most remarkable features of successful cities is the myriad ways in which their inhabitants can earn a living. To some urbanists like Jane Jacobs, their diversity is precisely the defining quality of cities. This economic diversification is both an outcome of and a prerequisite for urban growth: cities grow by diversifying their economies at the same time that a diversified economy allows cities to grow more productive and innovate. Recently, this relation between economic growth and diversification has been scrutinized in two connected, yet distinct bodies of research: Evolutionary Economic Geography (EEG) and Complexity Economics. In this paper, we discuss the treatment of economic diversity in these two strands of research. We focus our discussion on three concepts: related variety, economic complexity and industrial relatedness. First, we argue that the relation between these concepts and economic diversity is less straightforward than it may seem. Second, we highlight some important, yet often overlooked differences in ontological convictions on which they are based. Moreover, in an application to US cities, we show that there is some distance between the original narratives underpinning these concepts and their empirical measurement.

The notion that urban diversity matters finds widespread support among economic geographers and urban economists. The latter have stressed, for instance, that economic diversity improves production and consumption in a city, as formalized in "love-of-variety" utility and production functions (Dixit & Stiglitz, 1977; Krugman, 1991). Accordingly, diversity allows suppliers to specialize and customize products and services to the needs of specific customers (Duranton & Puga, 2004). A related argument posits that the wide variety of intermediate products and services offered in large and diversified cities lowers the barriers for new firms to enter new markets. Accordingly, diversity offers relevant building blocks – or *capabilities* – required for the successful operation of economic activities that are shared across industries. Jacobs' (1969) iconic New York City brazier maker serves as a colorful illustration of this logic.

Others have instead stressed that local diversity affords opportunities for *learning*. Accordingly, new technologies often emerge as new combinations of existing technologies. By facilitating the sharing of knowledge and ideas across industries, diverse cities spur innovation through Schumpeterian "new combinations".[2]

The latter, Schumpeterian, argument was further refined by Frenken et al. (2007). These authors stress that learning is most effective when the parties involved are at an optimal cognitive distance (Nooteboom et al., 2007). Frenken et al. therefore distinguish between *related* and *unrelated variety*, each of which play different roles in a city.

Like Frenken et al. (2007), Hidalgo and Hausmann (2009) argue that diversity spurs growth. However, like Jacobs (1969), Hidalgo and Hausmann's reasoning relies not so much on benefits for learning as for the overall operations of economic activities. They argue that different products require different capabilities. What matters for urban growth is therefore not superficial industrial diversity, but rather the diversity in capabilities that sustain a city's industry mix (see also F. Neffke et al. (2017) on this distinction), or, as the authors refer to this, a city's *complexity*. Industrial diversity is merely an imperfect reflection of this complexity. Ultimately,

[2] Jane Jacobs is often credited with the notion that diversity in cities facilitates such new combinations. However, Jacobs' original argument does not refer to technological spillovers, but is based on the idea that a deeper division of labor allows firms to outsource non-critical elements of their production processes, which lowers entry barriers for new firms and industries.

what determines a local economy's development potential is the fundamental breadth (i.e, diversity) of capabilities it can mobilize.

Finally, diversity is not only an input into, but also an output of, local economic development. This insight also goes back to Jacobs (1969), who proposed that cities grow by diversifying into new activities. More recently, Hidalgo et al. (2007) have provided empirical corroboration for this conjecture at the level of national economies by showing that the process of diversification is not random but follows predictable paths. Countries, regions and cities tend to diversify into activities that are closely *related* to the ones they already host, where relatedness is expressed in *product* or *industry spaces*. The idea of related diversification has been embraced by evolutionary economic geography, where it was transferred from a country-level to a region-level phenomenon (F. Neffke et al., 2011). Since then, processes of related diversification have been identified across a wide range of contexts (Hidalgo et al., 2018).

Interestingly, the EEG literature that emerged from Hidalgo et al.'s 2007 pioneering work seems somewhat agnostic about whether the path dependent nature of related diversification should be attributed to benefits in local learning or in local production. However, whereas Hidalgo et al.'s original contribution emphasized that relatedness and product spaces should be considered as constraints to the feasibility of growth paths, the subsequent literature has often embraced related diversification as a desirable growth strategy. This suggests an implicit embrace of the learning model: if related diversification maximizes knowledge spillovers, such diversification paths would not just be more feasible, but also dynamically efficient.

A complication in both lines of research is that, in spite of its appearance, diversity is not a monolithic concept. First, there is the aforementioned difference between superficial diversity in *industries* and the more substantive diversity in *underlying capabilities*. Economic complexity attempts to capture this latter fundamental diversity in its economic complexity index (ECI). However, recent work has cast doubt on whether the ECI can indeed be interpreted as a diversity measure. Second, diversity alludes to the notion that there are some primitive objects that are fundamentally distinct from one another. For instance, manufacturing cars is obviously different from running a restaurant. However, things are not always as easy. For instance, are fast-food chains and family restaurants different activities? Or are they different instances of the same activity? As definitions of economic activities become more fine-grained, it becomes harder to decide which activities are fundamentally different.[3] Third, there are at least three aspects to diversity. Diversity depends on (1) the number of distinct activities in a city, (2) how spread out employment or output is across these activities and (3) how dissimilar these activities are to one another (Stirling, 2007).

We will discuss all of these issues in greater detail. Our aim is to highlight the commonalities and differences in philosophical starting points that underlie related variety on the one hand and economic complexity and the product space on the other hand. These differences mirror the differences in intellectual antecedents: whereas the literature on related variety is firmly grounded in innovation theory, economic complexity and the product space emerged from combining trade theory with concepts of complex networks and combinatorial growth found in the complexity sciences. Furthermore, we discuss how the different conceptual starting points lead to different measurement strategies. To bridge the two frameworks, we build on a decomposition of diversity that separates the aforementioned aspects of diversity: the *variety* of different industries in a city, the *balance* of employment distribution across these industries and the *disparity* or (un)relatedness of the city's industries.

We illustrate our argument with data on US cities. The goal of this exercise is modest. We do not aim to provide definitive answers to the question of what role diversity plays in the growth and development of these cities. Instead, we use these data to explore how different empirical strategies yield different conclusions on the same core hypotheses put forward in prior literature.

The main lessons from our analysis are:

1. Related variety and economic complexity are based on fundamentally different beliefs about why diversity matters.

2. Economic complexity is no measure of generalized diversity and will only reveal an economy's complexity under specific circumstances.

3. The effects of related variety are sensitive to *ad hoc* empirical choices.

[3] Note that this aggregation problem is precisely what the measurement of relatedness aims to overcome: relatedness captures the how distinct different activities are.

4. Path dependent related diversification may not reflect the effects of a large diversity, but of a large mass of related activities.

In the remainder of the paper, we will elaborate on these lessons. We start by introducing the concepts of related variety, economic complexity and the industry space, paying special attention to the implicit stances they take on diversity. In Section 5.3 we describe the empirical implementation of these concepts. Next, we introduce the data and discuss our empirical exercise in 5.4. Finally, as a companion to this paper, we provide structured Python Notebooks that allow easy replication of our analyses. Our hope is that, by providing transparent access to the measures and calculations in this paper, we allow others to test hypotheses across datasets and applications and hopefully arrive at a scientific consensus about what roles diversity plays in local economic development.

5.2 The role of diversity in local economies

In both evolutionary economic geography and complexity economics, scholars have studied the role of diversity in local economic development. However both strands of the literature have done so using different concepts and empirical tools.

5.2.1 Related Variety

Economic geographers have long recognized that cities benefit from having a diversified economy. Since Glaeser et al. (1992), these benefits are known as Jacobs' externalities. Frenken et al. (2007), however, argue that regional diversity affects economic development in more than one way. First, a greater variety of economic activities in a city facilitates knowledge spillovers between industries. Second, like diversified financial portfolios lower investment risks, regional diversity reduces a city's exposure to idiosyncratic demand or supply shocks.

The main insight of Frenken and his colleagues is that these two effects build on different types of diversity. Whereas spillovers associated with Jacobs' externalities are most likely to materialize between "complementary sectors" Frenken et al., 2007, p. 686, risk diversification is maximized when industries differ in their exposure to market forces. Therefore, it is not just the variety of industries that a region hosts that matters, but also the extent to which these industries are related to one another. *Unrelated variety* reduces the region's exposure to adverse shocks, which should translate into less unemployment. *Related variety* instead benefits a local economy through the inter-industry learning associated with Jacobs' externalities. However, learning is most fruitful when it happens at an optimal cognitive distance (Nooteboom et al., 2007): to learn from one another, economic actors should neither be too similar nor too different from one another. By facilitating Schumpeter's "new combinations", related variety should therefore spur innovation and accelerate productivity growth.

5.2.2 Economic Complexity

Scholars in complexity economics have put forward different metrics of diversity to capture an economy's latent growth potential. The earliest metric was the Economic Complexity Index, introduced by Hidalgo and Hausmann (2009) as a measure of an economy's complexity. It builds on Hausmann and Klinger (2007) insight that "what [a country] export[s] matters." Accordingly, rich countries are rich because they produce products that require a broad capability base. Because the full list of factors that could count as capabilities is unknown – ranging from physical infrastructure and an educated labor force to efficient institutional arrangements and a capable state – identifying the precise capability requirements for each product is nigh impossible. Therefore, Hausmann and colleagues instead propose to infer the implicit productivity a product requires from the kind of countries that are able to export it. If only high-productivity countries – proxied as countries with high per-capita incomes – manage to export a product, the product is likely to require complex capabilities. The authors thus define a product's implicit productivity requirement, PRODY, as the average per-capita Gross Domestic Product (GDP) of countries that export the product. Next, the implicit productivity of country c, EXPY$_c$, can be calculated as the (export-value weighted) average productivity implied by the products it exports. This implicit productivity proves to predict a country's future income growth remarkably well.

The so-called method of reflections Hidalgo and Hausmann (2009) generalizes this notion of "implicit productivity" by ranking the complexity of products and countries without requiring information on countries' per-capita incomes. Instead, it defines the complexity of a country (the Economic Complexity Index, or ECI) and the sophistication of a product (the Product Complexity Index, or PCI) iteratively. In each iteration, the ECI of a country is the average of the (previous iteration's) PCI of all products that the country produces. Similarly, the PCI of a product is the average ECI of all countries that produce the product, where "producing" refers to exporting a product with revealed comparative advantage. As the iteration progresses, it updates its guesses of industry and country complexities. To seed the iterations Hidalgo and Hausmann use the number of products that a country produces as an initial guess of its complexity and the number of countries that are able to produce a product as the initial guess of the product's lack of sophistication (complexity). Iteratively updating these initial guesses yields an ECI for each country and a PCI for each product.

The authors interpret these indices as measures of the number of capabilities that a country has or that a product requires. That is, the ECI is supposed to reflect a fundamental, capability-based notion of diversity.

In later work, Hausmann and Hidalgo (2011) and Caldarelli et al. (2012) discovered that the method of reflections simplifies to an eigenanalysis, in which the ECI and PCI can be expressed as eigenvectors. However, this same insight ultimately cast doubt on the interpretation of the ECI and PCI as measures of capability endowments and capability requirements. Mealy et al. (2019) and Gomez-Lievano (2018) describe the close relation between ECI and spectral clustering:[4] the ECI splits countries into two groups such that the export baskets of countries in one group are similar to one another and different to those of countries in the other group.

The close relation between ECI and graph partitioning helps explain a number of known conundrums. First, the direction of the ranking of the ECI is undetermined: it can rank countries in ascending or descending order of complexity. Consequently, researchers need to determine the right direction in an ad hoc way.[5] The reason is now clear: because the ECI and PCI are eigenvectors, their sign is undetermined. Second, Hidalgo and Hausmann (2009) claim that the ECI is a generalized measure of diversity has been questioned by Tacchella et al. (2012)

[4] In fact, the method of reflections is exactly equivalent to an ordination in Ecology called 'reciprocal averaging' (Hill, 1973), which is in turn equivalent to the method of Correspondence Analysis, a technique for analyzing associations in high-dimensional categorical data (Greenacre, 1984). The complexity indices can thus be seen as the 'principal component' in a dimensionality reduction technique analogous to principal components analysis.

[5] For instance, the ECI should correlate positively with countries' GDP per capita, or Germany, Japan and the U.S. should be ranked as complex economies.

and Kemp-Benedict (2014). The latter showed that the ECI is in fact *orthogonal* to a country's export diversity.[6] Third and finally, the rankings produced by the ECI and, in particular, by the PCI can be strikingly counterintuitive. We will show some examples of this in Section 5.4.3.

5.2.3 The Product Space

Like the ECI, Hidalgo et al.'s (2007) product space builds on the notion that products differ in their underlying capability requirements. However, instead of trying to assess how many different capabilities one product requires, the product space attempts to measure to what extent two products share the same capability requirements. Once again, measurement is indirect: Hausmann and Klinger (2006) and later Hidalgo and co-authors posit that two products require similar capabilities if they are often co-exported by the same countries. By counting co-occurrences of products in countries' export baskets, the authors build a network that connects co-exported products. This network is referred to as the *Product Space*.

The product space has been shown to map a country's likely diversification paths. To predict future diversification, Hidalgo et al. (2007) create a variable they call "density". Density measures the proximity of a product to a country's overall export basket. The higher a product's density, the more likely it is that the country will start exporting it. This empirical regularity has been replicated across various data sets and contexts and was dubbed the Principle of Relatedness by Hidalgo et al. (2018). In Section 5.3.3, we will see that this density is, in fact, a measure of the *variety of related products*.

5.3 Measurement

The research reviewed above has yielded three quantities of interest: related variety, economic complexity and inter-industry proximity or relatedness. Below, we describe how each of these quantities can be measured. Herein, we stay close to the original papers, while simplifying some elements.

[6] Note that diversity is here measured as the number of products that are exported with revealed comparative advantage.

5.3.1 Related variety

Related variety as defined by Frenken et al. (2007) is based on the entropy of a city's employment distribution across industries. For a given city, the entropy is given by

$$S(\mathbf{p}_c) = -\sum_{i \in I} p_{ic} \log p_{ic}, \tag{5.1}$$

where $\mathbf{p}_c$ is the vector of employment shares $p_{ic} = \frac{E_{ic}}{E_{.c}}$ of industry i in city c (the "." in $E_{.c}$ indicates a summation over the omitted category, in this case industries).

A city has maximum entropy if all of its industries are equally large. In this case $S(\mathbf{p}_c) = \log N_c$, where N_c is the number of industries with nonzero employment share in city c. If all employment is concentrated in a single industry, $S(\mathbf{p}_c)$ reaches its minimum of $S(\mathbf{p}_c) = 0$.

If industries belong to broader sectors $\sigma \in \Sigma$, entropy can be decomposed into two components:

$$\begin{aligned} S(\mathbf{p}_c) &= -p_{\sigma c} \sum_{\sigma \in \Sigma} \log p_{\sigma c} - \sum_{\sigma \in \Sigma} p_{\sigma c} \sum_{i \in \sigma} \frac{p_{ic}}{p_{\sigma c}} \log \frac{p_{ic}}{p_{\sigma c}} && (5.2) \\ &= \mathrm{UV}_c + \mathrm{RV}_c, && (5.3) \end{aligned}$$

where $p_{\sigma c} = \frac{E_{\sigma c}}{E_{.c}}$ is the sectoral employment share in city c.

The first term is the city's *sectoral employment entropy*. It measures how equally spread out a city's employment is across sectors. Frenken et al. (2007) refer to this term as the city's *unrelated variety*. The second term is the city's *related variety*: a weighted average of industry-level employment entropies within each sector, where weights represent a sector's employment share. Related and unrelated variety thus quantify a city's degree of diversification at two different levels of aggregation: across sectors, and across industries within sectors.

5.3.2 Economic Complexity

To calculate the ECI and PCI, we first need to determine the activity mix of a local economy. That is, we need to decide whether or not an industry has a substantial presence in a city. To do so, we calculate a quantity known in economic geography as the location quotient (LQ).[7] Let E_{ic} be the employment of an industry i in a city c and omitted indices mark a summation over the corresponding dimension. We say that industry i is *present* in city c, whenever the industry is overrepresented in the city:

$$P_{ic} = \begin{cases} 1 & \text{if } \frac{E_{ic}/E_{i.}}{E_{.c}/E_{..}} > 1 \\ 0 & \text{elsewhere} \end{cases} \tag{5.4}$$

We collect the industry mixes of all cities in the matrix P. The entries of this matrix consist of zeros and ones, $P_{ic} \in \{0, 1\}$, that mark which industries (listed in rows) are present in which cities (listed in columns). Next, we calculate the ECI of each city and the PCI of each industry using the eigenvector implementation of the method of reflections. For details, we refer to Hausmann and Hidalgo (2011). A step-by-step description with embedded Python code is provided in the companion Jupyter Notebook.

5.3.3 Product Space

Inter-industry relatedness can be measured in a variety of ways (see, for instance, F. Neffke and Henning (2013) for an overview). In what follows, we largely follow the approach in Hidalgo et al. (2007). That is, we infer the relatedness between industries from how often industry i and i' co-occur in the same cities:

$$C_{ii'} = \sum_{c \in C} P_{ic} P_{i'c} \tag{5.5}$$

where C represents the set of cities in the dataset. The number $C_{ii'}$ is simply a count of the number of times that i and i' are present in the same city. The proximity of activity i to i', $\phi_{ii'}$, is now defined as:[8]

[7] When applied to export volumes, this quantity is known as revealed comparative advantage (RCA) in the trade literature.

[8] Note that this measure is similar to the one proposed by Hidalgo et al. (2007), but, unlike their metric, $\phi_{ii'}$ is symmetric. Given that co-occurrences are undirected, we see no advantage in artificially creating asymmetries in this measure.

$$\phi_{ii'} = \begin{cases} \frac{C_{ii'}/C_{.i'}}{C_{i.}/C_{..}} & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{5.6}$$

That is, to calculate proximity, we compare how often i co-occurs with industry i' to a benchmark that tells us how often we would have expected them to co-occur, had the industries been randomly distributed across cities.[9] Furthermore, we set the proximity of industry i to itself equal to zero. This will allow us to separate the effect of the presence of related industries from the effect of the industry's own presence in a city. Given that the metric defined in eq. (5.6) tends to have a highly skewed distribution, we map $\phi_{ii'}$ onto the interval $[0, 1)$ using:[10]

$$\tilde{\phi}_{ii'} = \frac{\phi_{ii'}}{\phi_{ii'} + 1}. \tag{5.7}$$

$\tilde{\phi}_{ii'}$ defines a network of related industries, the industry space.[11] We can use $\tilde{\phi}_{ii'}$ to calculate how close an industry is to a city's entire portfolio of industries. Following Hidalgo et al. (2007), we call this measure an industry's density in the city:

$$D_c^i = \sum_{i' \neq i} \frac{\tilde{\phi}_{ii'}}{\tilde{\phi}_{i.}} P_{i'c} \tag{5.8}$$

where the sum is taken over all industries in the classification system, excluding industry i itself. D_c^i counts the weighted number of different industries with $LQ > 1$ in city c *relevant to industry* i. The superscript i signals that the weights reflect how related each industry is to industry i.

In the empirical section, we will also introduce a close cousin of density, namely the *mass* of industries in city c relative to industry i:

$$E_c^i = \sum_{i' \neq i} \frac{\tilde{\phi}_{ii'}}{\tilde{\phi}_{i.}} E_{i'c}. \tag{5.9}$$

[9] Note that this normalization is essentially the same as in the LQ.

[10] For a detailed justification of this approach, see F. M. Neffke et al. (2017). An alternative, information-theory based normalization is proposed in van Dam et al. (2020).

[11] To increase visual clarity, we will require minimum thresholds for these edges when drawing the networks – but not when calculating densities – using the method laid out in M. Coscia and Neffke (2017).

Whereas density represents a proximity-weighted *count* of industries in a city – and is therewith essentially a measure of industrial variety – mass represents the proximity-weighted *size* of all industries. The difference is that, for mass, the *variety* of industries is unimportant: all related industries are perfect substitutes for one another, whether employment is distributed across many or few (equally related) industries.

In Section 5.4.3, we will use several alternative relatedness measures, all but one of which follow the same measurement approach. First, we estimate the proximity between cities, $\tilde{\phi}_{cc'}$, to produce a *city space* that expresses how similar cities are in terms of their industry mix:

$$\phi_{cc'} = \begin{cases} \frac{C_{cc'}/C_{.c'}}{C_{c.}/C_{..}} & \text{if } c \neq c' \\ 0 & \text{if } c = c' \end{cases} \tag{5.10}$$

Second, we estimate an *occupation space*, $\phi_{oo'}$ by looking at how often two occupations co-occur in the same cities:

$$\phi_{oo'} = \begin{cases} \frac{C_{oo'}/C_{.o'}}{C_{o.}/C_{..}} & \text{if } o \neq o' \\ 0 & \text{if } o = o' \end{cases} \tag{5.11}$$

In eqs (5.10) and (5.11), $C_{cc'}$ and $C_{oo'}$ are constructed analogously to $C_{ii'}$, counting the number of industries that are co-hosted by cities c and c' or the number of cities in which occupations o and o' co-occur. Furthermore, we map $\phi_{cc'}$ and $\phi_{oo'}$ onto the interval $[0, 1)$ to yield $\tilde{\phi}_{cc'}$ and $\tilde{\phi}_{oo'}$, using the transformation of eq. (5.7).

Third, we estimate a measure of *cognitive proximity* between industries:

$$\psi_{ii'} = \begin{cases} \frac{C^{\text{occ}}_{ii'}/C^{\text{occ}}_{.i'}}{C^{\text{occ}}_{i.}/C^{\text{occ}}_{..}} & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{5.12}$$

where $C_{ii'}^{occ}$ counts the number of occupations that are simultaneously present in industry i and i', using the definition of "presence" of eq. (5.4). Once again, we map this metric onto the interval $[0, 1)$, using the transformation in eq. (5.7).

Fourth, we calculate the relatedness, or similarity, of two industries' growth patterns as the correlation between the industries' growth rates:[12]

$$\rho_{ii'} = \begin{cases} corr\left(\frac{E_{it+1}}{E_{ict}}, \frac{E_{i't+1}}{E_{i't}}\right) & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{5.13}$$

This metric captures the extent to which industries are exposed to correlated economic shocks. The higher the correlations in industrial growth rates in a city are, the less well the city managed to diversify its portfolio risks.

5.3.4 Decomposing diversity: variety, balance and disparity

All three concepts discussed above, related variety, economic complexity and the product space pertain to the notion of diversity, but they do so in different ways. To compare these concepts and their relation to diversity, it will be helpful to explore more carefully what we actually mean by diversity.

Figure 5.1 shows three cities and their employment distribution.[13] In principle, each of these cities could claim to be equally diverse, as each contains two industries. However, city B has a more evenly distributed employment across these industries, making it arguably more diverse as its employment is not dominated by one industry. City C, in turn, has a similar composition as city B, but hosts industries that are most distinct from one another, making it more diverse than city B.

Industrial diversity is thus a compound concept that consists of three components (Stirling, 2007):[14]

1. How many different industries exist in the city? This is known as a city's industrial *variety*.

2. How equally is employment distributed among these industries? This is known as the industrial *balance* in a city.

[12] We check the significance level of the correlations; if p-value$_{ii'}$ < 0.05 then $\rho_{ii'} = 0$.

[13] The figure is inspired by Figure 1 in Rafols and Meyer (2010).

[14] Work on incorporating disparity into measures of diversity measures goes back to the seventies (Rao, 1982). More recent work applies these ideas in Scientometrics (Rafols & Meyer, 2010) and economics (van Dam, 2019).

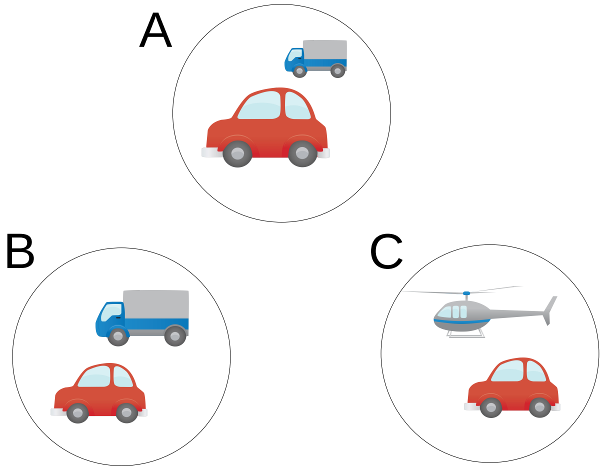


Figure 5.1: Three cities (A, B, and C) with a different employment structure. City A contains two industries of uneven size. City B contains two equally sized industries. City C also contains two equally sized industries, but they are more dissimilar than those in A and B.

3. How dissimilar are the industries in a city? This is known as a city's industrial *disparity*.

Using this framework, the distinction between related and unrelated variety can be understood as an interaction between the combination of (1) and (2) with (3). That is, related variety is high in cities with high industrial variety and/or balance, but low industrial disparity, whereas unrelated variety is high in cities with high industrial variety and/or balance, and high industrial disparity. Similarly, the density metric in Hidalgo et al. (2007) combines elements of (1) with (3): an industry's density is high in cities with many different industries that are strongly related to i.

Generalized diversity and Hill numbers

We will quantify diversity using the notion of Hill numbers (Hill, 1973). Unlike commonly used diversity indices such as the entropy or the Herfindahl-Hirschman index (HHI), Hill numbers express diversity in units of 'effective numbers' (Jost, 2006). The effective number of industries in a city is the number of equally large industries that would be needed to obtain the same diversity as the city under consideration. To be precise, Hill numbers answer the question: If we wanted to find a city with the same diversity, but where all industries are equally large, how many different industries would that city need? Hence, for an equally sized industries, the Hill number returns the number of industries in a city. For industries with unequal size, the Hill number returns the number of industries in the city, discounted for the inequality in the industry distribution.

Jost (2006) shows how a number of diversity indices can be transformed into effective numbers. For the Herfindahl-Hirschman index (HHI) for example, the effective number is given by the reciprocal of the index. To see this, consider how many different, yet equally large, industries a city would need to attain an HHI of $1/A$. The answer is A. In this case, each of A industries employs a share of $p_i = 1/A$ of the city's population. The HHI of this imaginary city is HHI $= \sum_{i=1}^{A} \frac{1}{A^2} = \frac{1}{A}$. Therefore, $\frac{1}{\text{HHI}}$ yields a Hill number. Similarly, the entropy can be converted to effective numbers by taking its exponential (Jost, 2006).[15]

Hill numbers provide a measure of diversity that takes into account variety and balance, but can be further extended to incorporate disparity. These generalized Hill numbers measure diversity in units that answer the question: How many *equally large and maximally distinct industries* would a city need to attain the same industrial diversity score as the city at hand? Let matrix Z represent a measure of industry relatedness. Leinster and Cobbold (2012) shows that an augmented Hill number of generalized diversity can now be defined as:

$$D_Z(\mathbf{p}_c) = -e^{\sum_i p_{ic} \log((\mathbf{Zp_c})_{ic})}. \tag{5.14}$$

This is a measure of diversity that takes into account variety, balance, and disparity, and can be interpreted in terms of effective numbers. When the proximity matrix is the identity matrix, $Z = I$, representing a situation where all industries are maximally dissimilar, eq (5.14) simplifies to the standard Hill number:

$$D_I(\mathbf{p}_c) = -e^{\sum_i p_{ic} \log(p_{ic})}.$$

[15] That is, a city's industrial diversity can be expressed as the exponential of the Shannon entropy: $e^{-\sum_i p_i \log(p_i)}$, where p_i represents the employment share of industry i in the city. For a city with N_c equally large industries, we then have $p_i = \frac{1}{N_c}$, so that $e^{-\sum_{i=1}^{N_c} \frac{1}{N_c} \log(\frac{1}{N_c})} = N_c$.

Decomposing diversity

The generalized Hill number of eq. (5.14) can be decomposed into separate components that measure variety, balance and disparity (van Dam, 2019). The decomposition is based on the fact that variety simply counts the number of industries in a city with nonzero employment share, N_c:

$$N_c = \sum_{i \in I} 1(E_{ic} > 0). \tag{5.15}$$

where $1(.)$ is an indicator function that evaluates to 1 if its argument is true and 0 otherwise.

Assuming that the standard Hill number is the product of variety and balance, we can then express balance as

$$bal_c = \frac{D_I(\mathbf{p}_c)}{N_c}. \tag{5.16}$$

Likewise, assuming that the generalized Hill numbers is the product of variety, balance and disparity, we obtain disparity as

$$disp_c = \frac{D_Z(\mathbf{p}_c)}{D_I(\mathbf{p}_c)}. \tag{5.17}$$

The intuition behind this decomposition is as follows. Balance and disparity are essentially factors between 0 and 1 that correct variety (the number of different industries found in a city) for the unevenness of the distribution of employment and the differential relatedness between industries. We can furthermore normalize variety itself such that it lies between 0 and 1 as well, by dividing variety by the total number of industries in the classification $|I|$, so normalized variety is expressed as:

$$var_c = \frac{N_c}{|I|}. \tag{5.18}$$

Relative Hill numbers

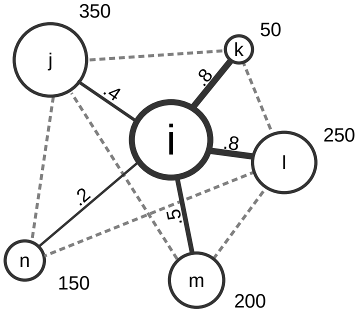
So far, we have discussed the aggregate diversity of an entire local economy. However, in the research on product spaces, the focus is not as much on cities as a whole as on individual industries within a city. Therefore, it is useful to extend the notion of general Hill numbers such that they relate to the diversity within a city in the neighborhood of a specific industry.

We can do so as follows. Imagine standing on a node in the industry space and looking around at all neighbors. We are interested in the amount of employment observed in each neighboring node, where we weight related nodes more heavily than unrelated. We can define a proximity-weighted employment of i' relative to i as follows:

$$E^i_{i'c} = \frac{Z_{ii'}}{\sum_{i' \neq i} Z_{ii'}} E_{i'c}$$

$E^i_{i'c}$ captures an industry's importance to the focal industry i, assuming that industries matter more the larger and more related they are. This idea is shown schematically in Figure 5.2. Note, furthermore, that if we sum $E^i_{i'c}$ across all neighboring industries of i, we get the quantity of mass as defined in eq. (5.9).

Let $p^i_{i'c}$ be the share of each if i's neighbor's relative employment to i, $p^i_{i'c} = \frac{E^i_{i'c}}{E^i_{.c}}$. Using these shares instead of p_{ic} in eqs (5.15) to (5.18) yields the amount of generalized diversity that exists in the immediate neighborhood of industry i. We will call this quantity the relative Hill number with respect to i. As before, we can decompose this relative diversity into its constituent components: *relative variety*, *relative balance* and *relative disparity*.



Node	$E_{i'c}$	$\frac{Z_{ii'}}{\sum_{i'} Z_{ii'}}$	$E^i_{i'c}$
j	350	.15	52.5
k	50	.3	15
l	250	.3	75
m	200	.18	36
n	150	.07	10.5

Figure 5.2: Schematic section of the industry space containing a focal industry (i) and its neighbors (i' in general). The size of a node indicates the industry's employment level, given by the number next to it and shown in the second column of the table. The edge labels represent the proximity $Z_{ii'}$ between the nodes, leading to the weights in the third column of the table. The product of the employment and the weights give the employment level relative to the focal industry, given in the fourth column of the table. The diversity relative to the focal industry is computed based on this relative employment. It consists of the relative variety (here 5), relative balance (the evenness of the distribution of the proximity weighted employment) and the relative disparity (the proximity among the neighbors, indicated here by grey dashed lines).

5.4 Empirical tests

5.4.1 Data

To illustrate the approaches discussed in the previous sections, we use data on US cities. The dataset contains information on the industrial composition of the economies of 369 Metropolitan Statistical Areas (MSAs) between 1990 and 2006. It records employment and average wages for each city-industry pair, as well as the unemployment rate for each city. We limit the analysis to 278 non-resource based, private-sector industries. Furthermore, We add two additional datasets that contain information on employment and wages for all occupation-city and occupation-industry pairs.[16]

5.4.2 Related variety

Frenken et al. (2007) test their related variety framework using data on Dutch labor market areas. Here, we will explore two of their main hypotheses: (a) because related variety facilitates product innovations through new technological combinations, related variety spurs employment growth; and (b) because unrelated variety reduces an urban economy's exposure to industry-specific, idiosyncratic shocks, unrelated variety protects a city against unemployment.

Frenken et al. (2007) find empirical support for both hypotheses. Some later studies replicate these results for different countries, time periods and sectors. Others, however, fail to corroborate them or report contradictory results (Content & Frenken, 2016).

This divergence in findings may be due to methodological shortcomings in Frenken et al.'s original study (see also Content and Frenken (2016)). First, it is unclear how related two industries must be to contribute to related variety instead of to unrelated variety. In Frenken et al. (2007), this threshold is arbitrarily set to whether or not two industries belong to the same 2-digit sector.

To illustrate this issue, we explore how the exact delineation between related and unrelated industries affects the estimated association between related or unrelated variety and employ-

[16] Appendix 5.6 provides details on the original data sources and our data cleaning. Appendix 5.7 contains an overview of the variables used in this section and their descriptive statistics.

ment growth. To do so, we estimate Ordinary Least Squares (OLS) regression models of the following kind:

$$\log\left(E_{cT}\Big/E_{ct}\right) = \beta_0 + \beta_1 \log E_{ct} + \boldsymbol{X}_{ct}\boldsymbol{\beta} + \varepsilon_{ct} \tag{5.19}$$

where E_{ct} is employment in city c in the base year t, and E_{cT} employment in city c in some later year T. The term $\log E_{ct}$ captures mean reversion effects, whereas the vector $\boldsymbol{X}_{ct}$ contains variables that describe an urban economy: its related variety, unrelated variety and size.[17]

Table 5.1 shows results. The models in each column differ by when two industries are considered related. In column (1), related industries are industries that belong to the same 1-digit sector, in column (2), the industries must belong to the same 2-digit sector and in column (3) to the same 3-digit sector. Unrelated variety is thus taken over 1- 2-, and 3-digit sectors, respectively.

Table 5.1: Employment growth in cities. Models differ by when two industries are considered related: column (1) same 1-digit sector, column (2): same 2-digit sector, column (3): same 3-digit sector.

	(1)	(2)	(3)
RV$_c$	0.0845	-0.0155	0.3147***
	(0.0701)	(0.0814)	(0.1082)
UV$_c$	-0.6318***	-0.1214	-0.2417***
	(0.1721)	(0.1077)	(0.0928)
ln E_c	-0.0954***	-0.0791***	-0.0865***
	(0.0163)	(0.0164)	(0.0161)
Intercept	0.4434***	0.4434***	0.4434***
	(0.0105)	(0.0108)	(0.0106)
R2	0.32	0.28	0.30
R2 adj.	0.31	0.27	0.29
N.obs.	369	369	369
Note:	* p < 0.1; ** p < 0.05; *** p < 0.01		

Although the exact relatedness cut-off is arguably an *ad hoc* choice, it does affect our findings. Whereas in models (1) and (2), related variety has no statistically significant effect on

[17] One could add further control variables for a city's human capital, infrastructure and so on. However, such variables risk being endogenous: they may be a consequence of a city's industrial diversity. Note that our aim is not to conclusively determine how diversity affects growth, but rather to explore whether arbitrary modeling choices affect our findings. We emphatically do not presume that we chose an optimal regression specification.

Table 5.2: Average wage growth in cities. Models differ by when two industries are considered related: column (1) same 1¯/digit sector, column (2): same 2¯/digit sector, column (3): same 3¯/digit sector.

	(1)	(2)	(3)
RV$_c$	0.0660**	0.0692**	0.2866***
	(0.0281)	(0.0308)	(0.0540)
UV$_c$	-0.0487	0.0084	-0.0713
	(0.0615)	(0.0422)	(0.0435)
$\ln w_c$	-0.1646***	-0.1694***	-0.1841***
	(0.0459)	(0.0454)	(0.0432)
$\ln E_c$	0.0164**	0.0189***	0.0152**
	(0.0069)	(0.0068)	(0.0065)
Intercept	0.5798***	0.5798***	0.5798***
	(0.0046)	(0.0046)	(0.0045)
R2	0.10	0.09	0.16
R2 adj.	0.09	0.08	0.15
N.obs.	369	369	369

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.3: Unemployment level in cities. Models differ by when two industries are considered related: column (1) same 1¯/digit sector, column (2): same 2¯/digit sector, column (3): same 3¯/digit sector.

	(1)	(2)	(3)
RV$_c$	-0.7249***	-0.5645***	-0.4927*
	(0.1963)	(0.2053)	(0.2577)
UV$_c$	-1.1732**	-1.1563***	-0.9717***
	(0.4640)	(0.3107)	(0.2699)
$\ln E_c$	0.9084***	0.9129***	0.9120***
	(0.0439)	(0.0452)	(0.0454)
Intercept	8.8857***	8.8857***	8.8857***
	(0.0222)	(0.0220)	(0.0222)
R2	0.84	0.85	0.84
R2 adj.	0.84	0.85	0.84
N.obs.	369	369	369

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

employment growth, we find a substantial and positive effect in model (3). Similarly, the effect of unrelated variety, which is negative in each model, is numerically unstable. Although these findings are roughly in line with Frenken et al. (2007), the dispersion of parameter estimates is worrisome.

Results are somewhat more robust if we repeat the analysis using two alternative dependent variables in Tables 5.2 and 5.3: growth in average wages and end-of-period unemployment levels.[18] Wage growth is positively associated with related variety, but not significantly associated with unrelated variety.[19] Unemployment levels, by contrast, are negatively associated with both related and unrelated variety, but more so with the latter than with the former.

A second concern about Frenken et al. (2007) approach is that the theoretical considerations put forward for why related and unrelated variety matter implicitly build on two different notions of relatedness. Whereas the growth benefits associated with inter-industry learning require that relatedness acts as a measure of cognitive proximity, the unemployment-averting portfolio benefits require a measure of similarities in exposure to idiosyncratic shocks. F. M. Neffke et al. (2017) find that these two concepts of relatedness are, in fact, close to uncorrelated.

Using the generalized Hill numbers of section 5.3.4, we can resolve both issues at once. First, we can choose any type of relatedness to measure the degree of disparity between a city's industries. Second, because disparity enters the generalized Hill number, in principle, as a continuous variable, there is no hard dichotomy between related and unrelated variety. Instead, the related versus unrelated variety hypotheses can be tested using interactions between continuous variables.

Starting with the latter, we follow Frenken et al. and use the classification hierarchy to decide how related two industries are. However, instead of distinguishing between related and unrelated industries, we define *classification-based relatedness* as the number of leading digits two industry codes have in common. If we normalize this relatedness to lie between 0 and 1, for a classification system with four digits, classification-based relatedness can attain one of five values: $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. Consequently, industries in, for instance, the same 3-digit sector have a relatedness score of $\frac{3}{4}$.

Table 5.4 runs similar OLS regressions to Table 5.1 above. However, instead of related and unrelated variety, it uses the generalized Hill-number based diversity metric that incorporates classification-based relatedness into its disparity component. Column (1) shows that generalized diversity displays a statistically significant and positive association with employment growth. When we decompose this generalized diversity in columns (2)–(5), we find that this

[18] Frenken et al. (2007) studied the effect on unemployment *growth*. However, unemployment rates essentially follow the business cycle. Changes in unemployment rates between 1990 and 2006 therefore are mostly driven by how far these years are from the closest troughs and peaks in the local business cycle and do not capture some characteristic city-specific unemployment dynamic.

[19] To capture mean-reversion effects, these analyses also control for the wage level in 1990.

association is mostly driven by the disparity between, and, to a lesser extent, the balance in the employment distribution across, a city's industries.

Table 5.4: Employment growth in cities (classification-based relatedness).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\ln D_Z(\mathbf{p}_c)$	0.3042***						
	(0.0738)						
$\ln \text{var}_c$		-0.1684			0.1167	0.3218***	
		(0.1122)			(0.0829)	(0.1009)	
$\ln \text{bal}_c$			0.0492		0.3272***		0.1295
			(0.0829)		(0.1011)		(0.1227)
$\ln \text{disp}_c$				0.1881***	0.3358***	0.0493	0.2425***
				(0.0673)	(0.0694)	(0.0676)	(0.0695)
$\ln \text{var}_c \times \ln \text{disp}_c$						-0.2996***	
						(0.0850)	
$\ln \text{bal}_c \times \ln \text{disp}_c$							0.2795**
							(0.1400)
$\ln E_c$	-0.1627***	-0.0434	-0.0911***	-0.0885***	-0.1055***	-0.1638***	-0.0795***
	(0.0234)	(0.0292)	(0.0106)	(0.0087)	(0.0245)	(0.0281)	(0.0093)
Intercept	0.4434***	0.4434***	0.4434***	0.4434***	0.4434***	0.4322***	0.4486***
	(0.0105)	(0.0107)	(0.0108)	(0.0105)	(0.0103)	(0.0107)	(0.0111)
R2	0.32	0.28	0.27	0.31	0.34	0.36	0.34
R2 adj.	0.31	0.28	0.27	0.31	0.33	0.35	0.33
N.obs.	369	369	369	369	369	369	369

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Columns (6) and (7) provide an alternative way to test the hypotheses in Frenken et al. (2007). To do so, we interact a city's industrial variety (column 6) or balance (column 7) with its industrial disparity. To facilitate the interpretation of these interaction effects, all variables have been mean-centered.

The answer to Frenken et al. (2007) question about related and unrelated variety turns out to depend on whether we think of industrial diversity as the number of different industries in a city or of how balanced the employment distribution across these industries is. Disparity moderates the effect of variety downwards, but that of balance upwards. Since disparity is the opposite of relatedness, this means that the effect of the variety component of diversity increases with increasing relatedness, whereas the effect of the balance component decreases with increasing relatedness.

The coefficient[20] of $+0.32$ for variety in column (6) of Table 5.4 means that the association

[20] Given that all variables are expressed in natural logs, coefficients should be interpreted as elasticities.

between variety and employment growth is $+0.32$ at an average level of disparity, but varies from $+0.44$ (at the minimum disparity, or maximum relatedness, in the sample) to -0.15 (for the maximum disparity in the sample). In contrast, the association with balance is 0.13 at average disparity levels, but varies from 0.01 to 0.56 between the minimum and maximum disparity in the sample. The finding of positive effects of related variety and negative effects of unrelated variety in Table 5.1 thus only holds if we measure diversity in terms of a city's variety (i.e. number of industries), not in terms of its employment balance across industries.

What happens if we change our measure of disparity to more closely reflect the theoretical considerations behind the hypotheses in Frenken et al. (2007)? To do so, we repeat the analysis of Table 5.4 twice with some slight modifications. First, Table 5.5 measures disparity using the (transformed) metric $\tilde{\psi}_{ii'}$ proposed in eq. (5.12), based on the number of occupations that industries share, instead of classification-based relatedness. This way, the relatedness between industries more accurately measures the cognitive proximity that would lead to inter-industry spillovers. Second, in Table 5.6 we change the dependent variable to the end-of-period unemployment rate in a city and use the growth-similarity based metric $\chi_{ii'}$ of eq. (5.13) to more accurately capture portfolio diversification effects. Note that $\tilde{\psi}_{ii'}$ and $\chi_{ii'}$ define relatedness as continuous variables. To allow for a fair comparison with Frenken et al. (2007), we convert $\tilde{\psi}_{ii'}$ and $\chi_{ii'}$ into categorical (or better, ordinal) variables in such a way that each class contains the same number of industry pairs as its counterpart in the classification-based relatedness matrix.

Table 5.5 shows that results when disparity is based on cognitive proximity are very similar to the ones when disparity is based on classification-based relatedness. Once again, results corroborate Frenken et al. (2007) hypothesis in the interaction between disparity and variety, but not in the interaction between disparity and balance. Moreover, the interaction effects are somewhat stronger than when using classification-based disparity.

Table 5.6 shows that general diversity, and in particular, a more balanced employment distribution offer some protection against high unemployment rates. Moreover, disparity in growth correlation-based relatedness weakly strengthens the benefits of employment balance. In line with the theoretical considerations put forward by Frenken et al. (2007), this suggests that the greater the difference in growth patterns of industries in a city are, the more a balanced

Table 5.5: Employment growth in cities (cognitive-proximity-based relatedness).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\ln D_Z(\mathbf{p}_c)$	0.0217						
	(0.2031)						
$\ln \text{var}_c$		-0.1684			0.1120	0.4824***	
		(0.1122)			(0.2286)	(0.1320)	
$\ln \text{bal}_c$			0.0492		0.3460		-0.1283
			(0.0829)		(0.2282)		(0.1619)
$\ln \text{disp}_c$				0.0998	0.3734	0.0265	0.1446
				(0.0885)	(0.2705)	(0.0940)	(0.1173)
$\ln \text{var}_c \times \ln \text{disp}_c$						-0.4517***	
						(0.0835)	
$\ln \text{bal}_c \times \ln \text{disp}_c$							0.6796***
							(0.2017)
$\ln E_c$	-0.0955***	-0.0434	-0.0911***	-0.0785***	-0.0566*	-0.2116***	-0.0741***
	(0.0202)	(0.0292)	(0.0106)	(0.0129)	(0.0320)	(0.0332)	(0.0209)
Intercept	0.4434***	0.4434***	0.4434***	0.4434***	0.4434***	0.3976***	0.4490***
	(0.0108)	(0.0107)	(0.0108)	(0.0108)	(0.0107)	(0.0123)	(0.0109)
R2	0.27	0.28	0.27	0.28	0.29	0.38	0.32
R2 adj.	0.27	0.28	0.27	0.27	0.28	0.37	0.32
N.obs.	369	369	369	369	369	369	369

Note: * p < 0.1; ** p < 0.05; *** p < 0.01

employment distribution across these industries can shield the city from high unemployment rates. In contrast, a greater variety of industries is associated with higher unemployment rates, especially if their growth rates are uncorrelated (or anti-correlated).

These results show that Frenken et al. (2007) theoretical framework can be brought to the data in a more principled way using the generalized Hill number approach to measuring diversity. In general, our findings suggest that there is support for benefits in inter-industry learning at an optimal cognitive distance if we focus on the variety component of diversity. That is, cities that host many related industries, regardless of their size, create more opportunities for learning. Similarly, a balanced industrial portfolio seems to be associated with less unemployment, especially of industries that exhibit different growth patterns.

5.4.3 Economic complexity

Hidalgo and Hausmann (2009) motivate the ECI as a measure that aims to capture a city's fundamental diversity in terms of the number (or variety) of capabilities a city makes available to

Table 5.6: Unemployment level in cities (growth-similarity-based relatedness).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\ln D_Z(\mathbf{p}_c)$	-0.2815**						
	(0.1332)						
$\ln \text{var}_c$		0.1250			0.0063	0.6108***	
		(0.0780)			(0.1759)	(0.1878)	
$\ln \text{bal}_c$			-0.2406***		-0.3347**		-0.5076***
			(0.0925)		(0.1367)		(0.1570)
$\ln \text{disp}_c$				-0.0110	-0.1254	0.2074*	-0.1627**
				(0.0610)	(0.1523)	(0.1073)	(0.0698)
$\ln \text{var}_c \times \ln \text{disp}_c$						-0.1974***	
						(0.0698)	
$\ln \text{bal}_c \times \ln \text{disp}_c$							0.3596**
							(0.1732)
$\ln E_c$	0.0717***	-0.0033	0.0235**	0.0324**	0.0028	-0.1073**	-0.0041
	(0.0225)	(0.0257)	(0.0117)	(0.0132)	(0.0370)	(0.0434)	(0.0155)
Intercept	0.0032	0.0032	0.0032	0.0032	0.0032	-0.0152	0.0064
	(0.0149)	(0.0150)	(0.0149)	(0.0150)	(0.0149)	(0.0160)	(0.0151)
R2	0.04	0.03	0.04	0.02	0.05	0.05	0.06
R2 adj.	0.03	0.02	0.04	0.02	0.04	0.04	0.04
N.obs.	369	369	369	369	369	369	369

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

its firms. How does the ECI compare to the generalized diversity described above as a measure of fundamental diversity? Figure 5.3 shows a scatter plot between the two metrics. ECI and generalized diversity are strongly correlated, with $\rho = 0.47$. Table 5.7 documents three additional facts about the relation between ECI and generalized diversity. First, both ECI and generalized diversity are strong predictors of a city's average wage level (columns 1 and 2). Second, however, when the two variables enter the model jointly, only the ECI is significantly associated with a city's wage level, regardless of whether we control for the city's size or not (columns 3 and 4). Third, the correlation between ECI and generalized diversity in Figure 5.3 seems to be fully mediated through both variables' association with city size. Controlling for city size, the statistical association between ECI and generalized diversity disappears (column 5). This suggests that the ECI may indeed measure a more fundamental complexity of a city than generalized diversity. In the remainder of this section, we scrutinize this claim by studying three use scenarios of the ECI.

The first scenario is close to the original paper by Hidalgo and Hausmann (2009). It fol-

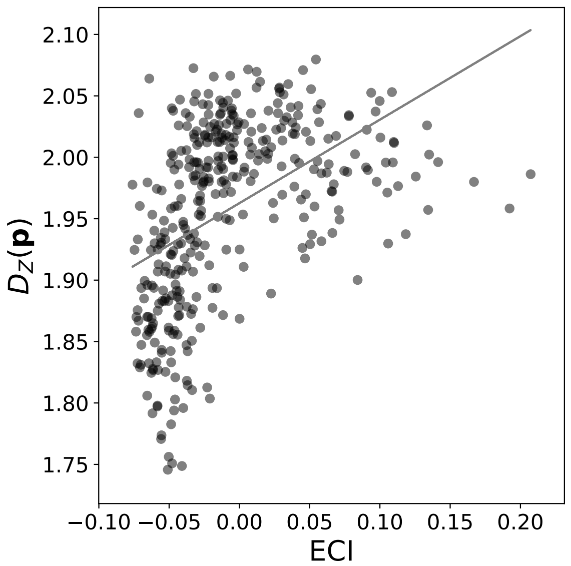


Figure 5.3: Generalized diversity and ECI.

Table 5.7: ECI, generalized diversity and urban wages. OLS regressions with dependent variables in the first row.

	(1)	(2)	(3)	(4)	(5)
dep. var.	ln avg. wage	ln avg. wage	ln avg. wage	ln avg. wage	$\ln D_Z(\mathbf{p}_c)$
ECI$_c$	3.0234***		2.8931***	1.3963***	-0.0046
	(0.1741)		(0.2003)	(0.3942)	(0.1288)
$\ln D_Z(\mathbf{p}_c)$		1.1953***	0.1787	-0.0747	
		(0.1326)	(0.1330)	(0.1469)	
$\ln E_c$				0.0784***	0.0342***
				(0.0190)	(0.0055)
Intercept	10.2801***	7.9397***	9.9330***	9.5239***	1.5497***
	(0.0072)	(0.2579)	(0.2606)	(0.2556)	(0.0636)
R2	0.55	0.18	0.56	0.60	0.34
R2 adj.	0.55	0.18	0.55	0.60	0.33
N.obs.	369	369	369	369	369

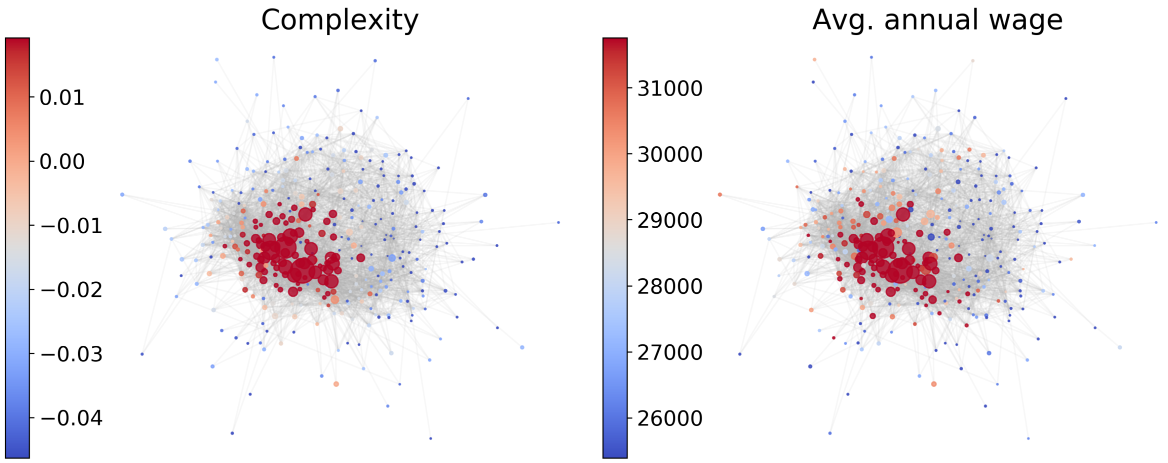
Note: * p < 0.1; ** p < 0.05; *** p < 0.01

lows the analysis of Figure 5.3 and Table 5.7 above and quantifies the complexity of US cities using the economic complexity index based on city-industry employment information. The second repeats this exercise, but focuses on the occupational mix of US cities. In analogy to the city-industry application, having many different occupations is assumed to be a sign of a city's complexity and being found in few cities (being "non-ubiquitous") is taken as a sign of the occupation's sophistication. In the final application, we turn to data that describe the occupational mix used by different industries. Note that, although it is easy to mechanically apply the method of reflections in occupation-industry data, the intuition for why this would be meaningful is less convincing: Although industries that use many different occupations may be complex, it is hard to see why the using occupations that are not used by many other industries would make industries sophisticated.

City-industry analysis

Figure 5.4 shows the city-space network constructed from city-industry employment data. The nodes in this network represent US cities. These nodes are connected by edges that express how similar two cities are in terms of the industries they host. In the first panel, we color these nodes by a city's ECI. In the second, colors instead show the average wage in each city.

Figure 5.4: ECI and wages in the city space (industry-city analysis)



High-ECI areas in the network (colored dark red in the left panel) tend to coincide with high-wage areas (right panel). The scatter plot in Figure 5.5 corroborates this impression: the regression of average wages on ECI has an R^2 of 0.554. This offers a visual confirmation of the relation described in model (1) of Table 5.7. Moreover, if we regard the average wage level in a city as a reflection of its productivity, these findings would also offer support for the notion that the ECI captures a city's complexity.

Table 5.8 lends further credence to this interpretation. It shows the top 10 most complex cities, which consists exclusively of high-income cities with plausibly complex economies, such as Los Angeles, San Francisco, Chicago and Boston.

Figure 5.5: ECI versus average wage in a city (industry-city analysis)

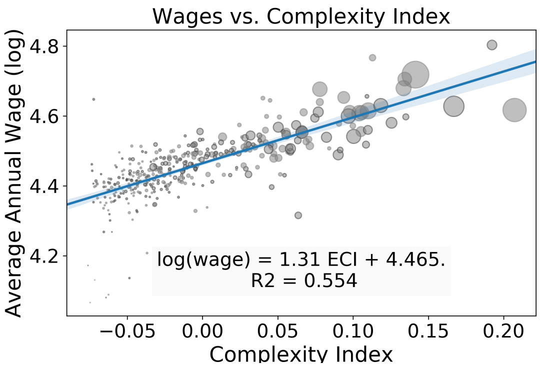


Table 5.8: Top 10 of most complex cities (city-industry analysis)

City	ECI	Avg. Wage
Los Angeles-Long Beach-Santa Ana, CA	0.207	41400
San Jose-Sunnyvale-Santa Clara, CA	0.192	63500
Chicago-Naperville-Elgin, IL-IN-WI	0.167	42400
New York-Newark-Jersey City, NY-NJ-PA	0.141	52300
New Haven-Milford, CT	0.135	39600
San Francisco-Oakland-Hayward, CA	0.134	50700
Boston-Cambridge-Newton, MA-NH	0.134	47800
San Diego-Carlsbad, CA	0.126	38000
Detroit-Warren-Dearborn, MI	0.118	42600
Bridgeport-Stamford-Norwalk, CT	0.113	58400

However, results become less convincing when we turn to the PCI. Figure 5.6 shows analogous panels to Figure 5.4, but now using industries as nodes in an industry space network. There is no clear relation between PCI and average wages, both when comparing the two network graphs and in terms of the correlation between PCI and wages in Figure 5.7. With an R^2 of 0.21, the PCI has no predictive power for industry-level wages. Moreover, some high-PCI industries in Table 5.9, such as urban transit systems, seem poor examples of complex economic activities.

City-occupation analysis

What happens when base the ECI on city-occupation instead of city-industry employment data? Figures 5.8 and 5.9 show the city space and a scatter plot of log(wage) against a city's ECI, using data on occupational employment in cities. Once again, the ECI is strong predictor of a

Figure 5.6: PCI and wages in the industry space (industry-city analysis)

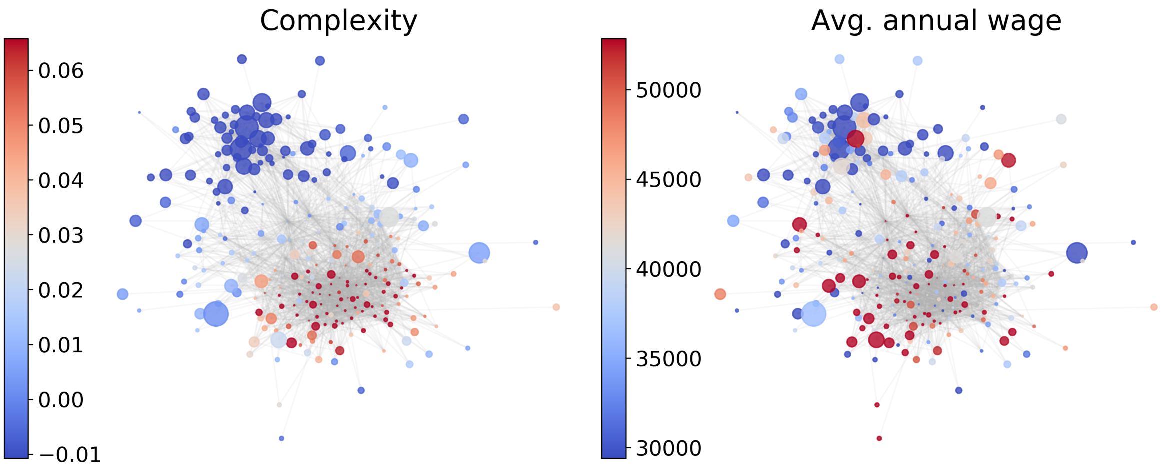


Figure 5.7: PCI versus average wage in an industry (industry-city analysis)

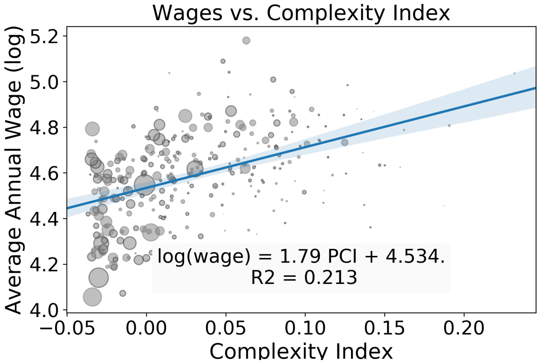


Table 5.9: Top 10 of most complex industries (city-industry analysis). We limit this list to industries that employ at least 25,000 workers in the US.

Industry	PCI	Avg. Wage
Motor vehicle manufacturing	0.191	64,111
Urban transit systems	0.147	43,015
Scheduled air transportation	0.125	54,095
Electric lighting equipment manufacturing	0.117	38,908
Steel product mfg. from purchased steel	0.114	46,611
Iron and steel mills and ferroalloy mfg.	0.109	55,467
Pharmaceutical and medicine manufacturing	0.104	75,532
Motion picture and video industries	0.101	53,333
Junior colleges	0.099	32,752
Other nonferrous metal production	0.097	52,113

city's wage levels: high-ECI cities tend to exhibit high average wages. In contrast, the PCI fails to accurately predict occupational wages. Figures 5.10 and 5.11 show that some occupations with high PCI levels pay very high wages, but others do not. In fact, the list of most complex occupations contains a number of high-skill occupations, such as computer software engineers and financial analysts, as well as low-skill jobs, such as parking lot attendants.

Figure 5.8: ECI and wages in the city space (occupation-city analysis)

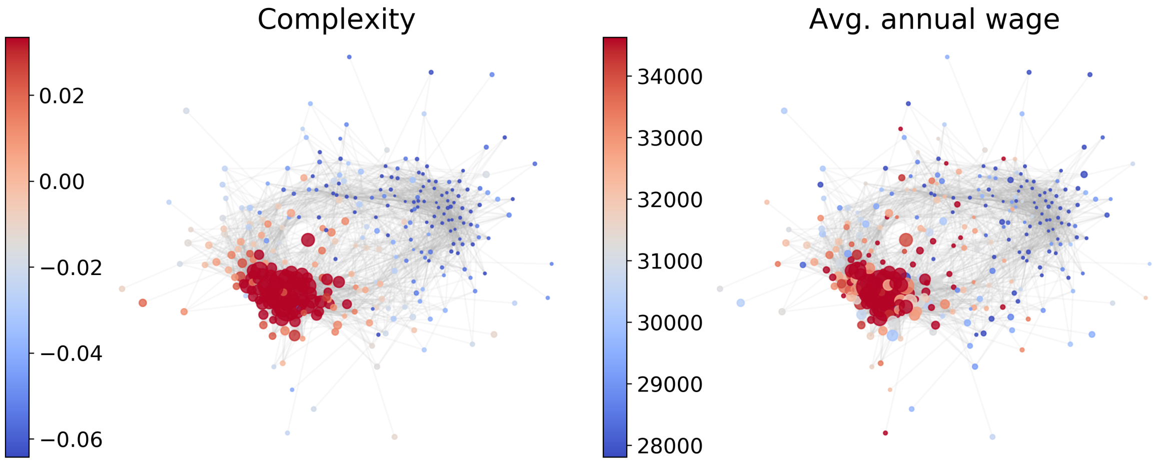


Figure 5.9: ECI versus average wage in a city (occupation-city analysis)

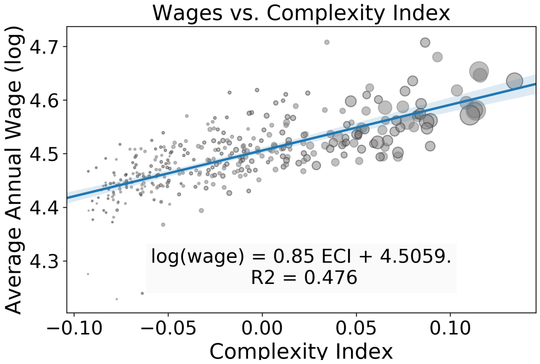


Table 5.10: Top 10 of most complex cities (city-occupation analysis)

City	ECI	Avg. Wage
Washington, DC-MD-VA-WV	0.134	43200
Boston, MA-NH	0.116	44300
New York, NY	0.115	45100
Chicago, IL	0.114	38100
Philadelphia, PA-NJ	0.113	38100
Los Angeles-Long Beach, CA	0.110	37300
Minneapolis-St. Paul, MN-WI	0.109	39300
Seattle-Bellevue-Everett, WA	0.104	41500
San Francisco, CA	0.093	47900
Dallas, TX	0.089	36500

This raises an interesting question: Why does the ECI seem a plausible measure of a city's complexity, regardless of whether we use cities' occupational or industrial composi-

Figure 5.10: PCI and wages in the occupation space (occupation-city analysis)

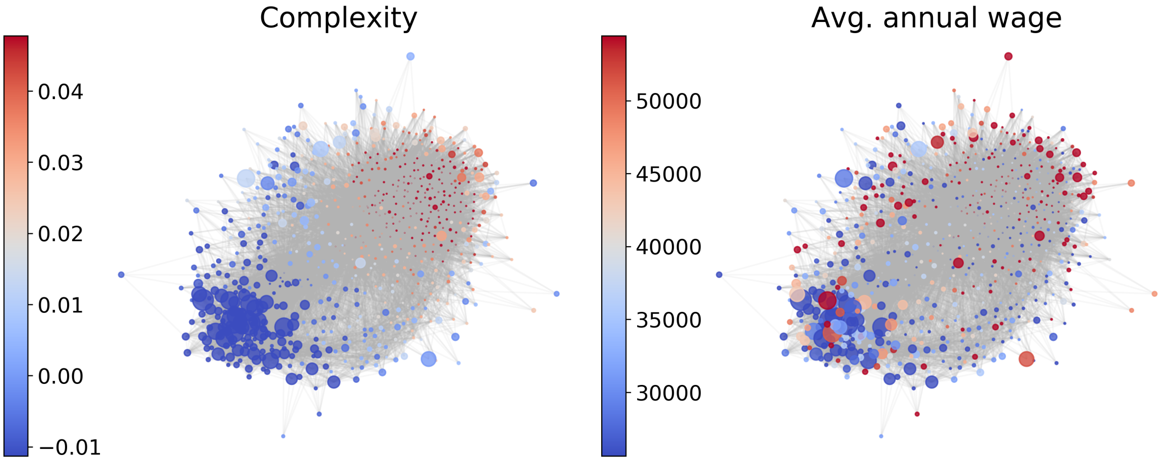


Figure 5.11: PCI versus average wage in an occupation (occupation-city analysis)

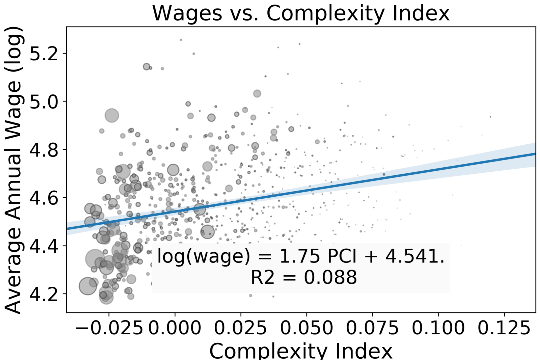


Table 5.11: Top 10 of most complex occupations (occupation-city analysis). We limit this list to occupations with at least 25,000 across all cities.

Occupation	PCI	Avg. Wage
Actors	0.079	49,648
Parking Lot Attendants	0.058	17,277
Financial Analysts	0.054	67,811
Musicians and Singers	0.048	53,474
Computer Software Engineers, Systems Software	0.044	76,574
Operations Research Analysts	0.043	61,426
Market Research Analysts	0.043	60,539
Brokerage Clerks	0.041	36,258
Multi-Media Artists and Animators	0.04	52,902
Computer Hardware Engineers	0.039	78,306

tions, whereas the PCI fails to provide an equally intuitively appealing ranking of industries or occupations?

The problem is not necessarily that the method of reflections does not work for industries and occupations. However, to understand the algorithm's outcomes, we must interpret them through a graph partitioning lens: the ECI does not count capabilities. Instead, it aims to split the city space network into two sets of nodes (Gomez-Lievano, 2018; Mealy et al., 2019). In each set, cities tend to have similar industries or occupations. The real question, therefore, is: Why does the ECI still manage to predict wage-levels in cities, whereas the PCI does not predict wage-levels in industries or occupations?

A possible answer to this conundrum lies in the fact that not all industries base their location choices predominantly on the availability of local capabilities. Although industries will preferably locate where they can access the right mix of skills, specialized suppliers, infrastructure and institutions, some industries produce goods and services that need to be consumed where they are produced. Such nontradable goods and services, like fresh bread, theater productions or daycare provision, need to be produced close to consumers. Some of these goods and services will found everywhere. Others can only be profitable provided in places with a large and affluent population.

A complex city, therefore, attracts two different types of industries and their occupations. First, it attracts complex industries from the tradable sector, which seek out the city to access its large capability base. These industries typically hire well-educated workers, who earn high incomes. These incomes, in turn, attract a second set of industries: industries from the nontradable sector that cater to the needs of a wealthy population. These industries provide goods and services, such as fine dining and childcare. Moreover, because high-income cities tend to be large, they may also offer services that can only be sustained in large population centers, like public transportation. These industries in the nontraded sector may not draw much from the city's capability base and, instead, employ low-skill workers with relatively low wages.

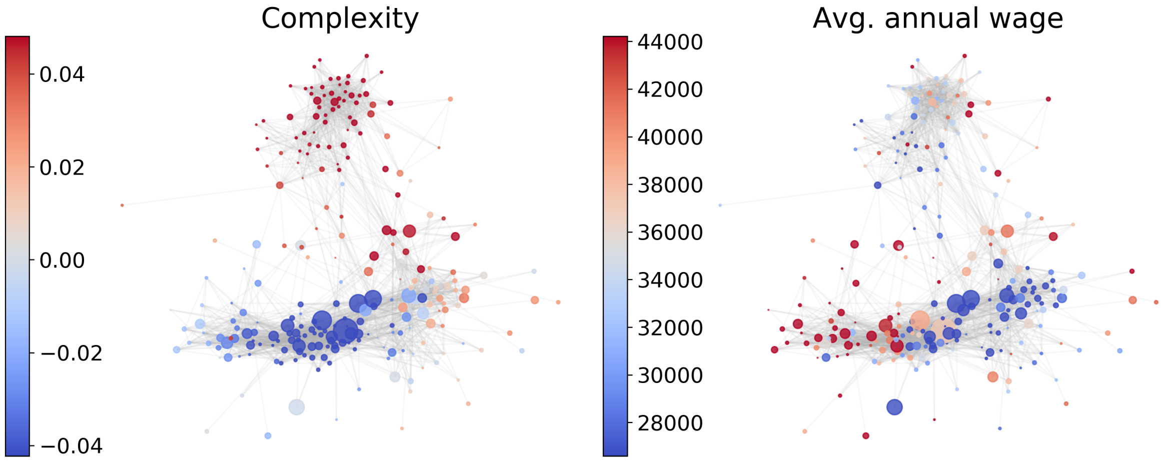
If accurate, the account predicts that the similarities described by the ECI will not just group cities with similar capability requirements, but also with similar consumption patterns. This dual logic divides cities neatly into high and low income cities, because income earned in the tradable, capability-seeking sector is spent in the local nontradable sector. In contrast, the PCI, which captures which industries locate in similar cities, would group a mix of two different types of industries. It would first distinguish between low- and high-complexity industries in the tradable sector. However, it would then augment the set of high-complexity industries with a set of, often low-skill, industries that cater to the needs of a wealthy population. As a consequence, the ECI would be a reliable predictor of wages, but the PCI would not be.[21]

Industry-occupation analysis

To more forcefully show that the ECI and PCI should not be uncritically considered as indices of economic complexity, we now turn to an application that uses industry-occupation employment data. Figure 5.12 shows the results from the industry perspective, Figure 5.13 from the occupation's perspective. Unlike the core-periphery patterns of Figures 5.4 and 5.8, the industry space now consists of various weakly connected areas. Moreover, the relation between ECI or PCI and wages has vanished completely: the R^2 of both regressions is below $R^2 = 0.03$.

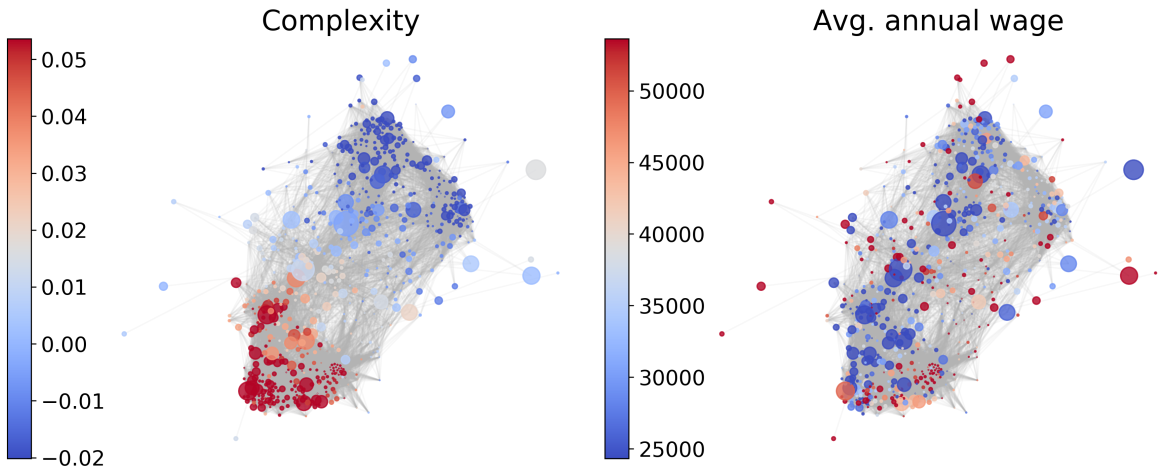
In spite of the fact that the ECI is a better predictor of a city's productivity (proxied by its wage level) than generalized diversity, it is unclear to what extent the ECI measures a city's fundamental diversity, i.e., the breadth of its capability base. Because of this, Mealy et al. (2019) conclude that the ECI and PCI offer a dimension-reduction technique, with no clear link to complexity as fundamental diversity. Providing a more positive evaluation, Schetter (2019) derives a set of sufficient conditions under which the ECI reliably ranks economies in terms of their complexity. Overall, however, the true meaning of the ECI and its role in economic development remains an active area of research.

Figure 5.12: ECI and wages in the industry space (occupation-industry analysis)



[21] Note that this issue does not arise in Hidalgo and Hausmann (2009). Because these authors base the ECI on a country's exports, by definition, their data reflect production that is not meant for local markets.

Figure 5.13: PCI and wages in the occupation space (occupation-industry analysis)



5.4.4 The product space

The product space was originally used to predict how countries will diversify their trade baskets Hidalgo et al. (2007). Since then, many authors have not just predicted the emergence of new products (or industries) in an economy – so-called growth at the extensive margin – but also how *existing* products and industries have grown. In this section, we will focus on this growth at the intensive margin and estimate models based on the following regression equation:

$$\log\left(E_{icT}\middle/E_{ict}\right) = \beta_0 + \beta_1 \log E_{ict} + \beta \log X_{ict} + \log E_{it} + \log E_{ct} + \varepsilon_{ict}$$

In other words, our dependent variable is the logarithm of industry i's growth factor in city c. As explanatory variables, we include a mean reversion term, $\log E_{ict}$, as well as the size of the industry ($\log E_{it}$) and of the city ($\log E_{ct}$) in the base year. The main variables of interest are collected in the vector $\boldsymbol{X}_{ict}$.

Table 5.12 shows the results. In column (1), apart from industry and city size variables, we only add the mean reversion term and the product space density, using $\tilde{\phi}_{ii'}$ of eq. (5.7), as explanatory variables. As expected, and in line with the literature's prior consensus, the mean reversion term shows a negative, and the product space density a positive association with employment growth.

Table 5.12: Product space regression. Dependent variable: Employment growth in city-industry pairs. Regressors use the industry space as defined by $\tilde{\phi}_{ii'}$ of eq. (5.7).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\ln D_c^i$	0.2705^{***}		0.1945^{***}					
	(0.0181)		(0.0184)					
$\ln E_c^i$		0.7910^{***}	0.6619^{***}	0.7873^{***}	0.7934^{***}	0.7754^{***}	0.7938^{***}	0.7982^{***}
		(0.0456)	(0.0466)	(0.0456)	(0.0457)	(0.0459)	(0.0458)	(0.0462)
$\ln D_Z(\mathbf{p}_c^i)$				-0.3161^{**}				
				(0.1237)				
$\ln \mathrm{var}_c^i$					0.1378^{***}			
					(0.0294)			
$\ln \mathrm{bal}_c^i$						-0.0765^{***}		
						(0.0229)		
$\ln \mathrm{disp}_c^i$							-0.0135	-1.0519^{***}
							(0.0156)	(0.1822)
$\ln D_I(\mathbf{p}_c^i)$								-0.9886^{***}
								(0.1926)
$\ln D_I(\mathbf{p}_c^i) \times \ln \mathrm{disp}_c^i$								-0.1561^{***}
								(0.0197)
$\ln E_{ic}$	-0.3896^{***}	-0.3977^{***}	-0.4020^{***}	-0.3984^{***}	-0.3966^{***}	-0.3968^{***}	-0.3978^{***}	-0.4018^{***}
	(0.0054)	(0.0055)	(0.0056)	(0.0055)	(0.0055)	(0.0055)	(0.0055)	(0.0056)
$\ln E_c$	0.2517^{***}	-0.4784^{***}	-0.3755^{***}	-0.4732^{***}	-0.5158^{***}	-0.4667^{***}	-0.4839^{***}	-0.4905^{***}
	(0.0054)	(0.0435)	(0.0442)	(0.0435)	(0.0444)	(0.0437)	(0.0441)	(0.0455)
$\ln E_i$	0.3141^{***}	0.3115^{***}	0.3162^{***}	0.3121^{***}	0.3112^{***}	0.3108^{***}	0.3116^{***}	0.3151^{***}
	(0.0058)	(0.0059)	(0.0059)	(0.0059)	(0.0059)	(0.0059)	(0.0059)	(0.0059)
Intercept	0.3450^{***}	0.3450^{***}	0.3450^{***}	0.3450^{***}	0.3450^{***}	0.3450^{***}	0.3450^{***}	0.3241^{***}
	(0.0032)	(0.0032)	(0.0032)	(0.0032)	(0.0032)	(0.0032)	(0.0032)	(0.0042)
R2	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
R2 adj.	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
N.obs.	43322	43322	43322	43322	43322	43322	43322	43322

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Hidalgo et al. (2007) interpret this finding as evidence that a large variety (counted as the number of industries with $LQ > 1$) of relevant (i.e., related) industries in a city enhances the focal industry's growth potential. However, is this really the case? An alternative explanation is that density is a proxy for having a large *quantity* of related activity in the city. In columns (2) and (3), we test this hypothesis by adding the relative mass (the total employment in related industries, as defined in eq. (5.9)) to the regression model.

The mass of related activity turns out to be more important than its variety: mass displays a stronger statistical association with employment growth than density does. Moreover, when adding both variables simultaneously, the association between density and employment growth weakens substantially.

In the remaining columns, we investigate the relation between the growth of local industry and the diversity of related industries in more detail. To do so, we replace density by the relative Hill numbers proposed in Section 5.3.4. These variables offer an alternative, more disaggregated way to look at diversity in a neighborhood of related industries.

Outcomes are shown in columns (4) to (7). The association of employment growth with overall relative diversity in column (4) is negative. When decomposing relative diversity into relative variety, relative balance and relative disparity, this negative association turns out to be driven by the relative balance component. That is: the more equally proximity-weighted employment is distributed across related industries, the more slowly the focal industry grows.

Learning versus producing

Hidalgo et al. (2007) interpret a large diversity of related industries as a sign that a city offers many capabilities that are relevant to the focal industry. In the introduction, we referred to this as a production-based logic: industries can only get established in places where they can mobilize all capabilities they require. The EEG literature has typically stressed another reason why diversity of related industries would be beneficial: the existence of opportunities for local learning. Both rationales can explain why density is positively associated with a local industry's growth rate. So how can we decide which of these interpretations is correct?

To answer this question, note that the two narratives differ in their interpretation of the edges in the product space. In the EEG literature, such connections are often interpreted as estimates of how easily knowledge can flow within an economy. In this reading, a large number of related industries provides greater scope for local knowledge sharing and local learning. A production-based interpretation, in contrast, regards the industry space as a reflection of shared capability requirements. From this perspective, industry spaces capture economies of scope between industries.

Although both perspectives suggest that a greater variety or balance of related industries is beneficial, they give different predictions with respect to relative disparity. In a shared-capability world, related activities would ideally be *unrelated* to one another. That way, each activity offers non-redundant capabilities to the focal industry. In contrast, in a learning world, related activities are ideally also related among each other. This way, all related industries can exchange knowledge, setting in motion a virtuous cycle of local learning.

Column (8) explores which of the two hypotheses finds most support in the data. It does so by interacting relative disparity with a compound measure of relative variety and relative balance. This interaction term is negative: the smaller the disparity among related industries is – i.e., the more the focal industry's related industries are also related to one another – the faster the focal industry will grow. This supports the local learning hypothesis, not the capability-sharing hypothesis. Note, however, that although the effect of the relative variety-balance compound variable increases as relative disparity drops, it remains negative for the entire range of relative disparity values observed in the sample. Such negative effects contradict both the learning and the capability-sharing hypothesis. However, this conclusion depends on the econometric specification, relatedness matrix and dependent variable we choose.[22] A definitive conclusion would thus require a more careful analysis and ideally a replication of these findings.

5.5 Discussion and conclusion

Recent years have seen a renewed interest in, and debates about, the importance of diversification in local economies. These debates were fuelled by three different lines of research: research on related variety, on economic complexity and on product and industry spaces. Although these lines of research emerged more or less contemporaneously and share many commonalities, they trace their origins to different intellectual traditions. As a consequence, they depart from different ontological starting points and measurement philosophies. Whereas related variety research is rooted in evolutionary economic geography, complexity and product space research is rooted in the economics of trade and growth on the one hand and the complexity sciences on the other.

As a result, the role of diversity differs across these approaches. Related variety research

[22] The reader can explore this in the companion Python Notebook.

attributes the benefits of a diversified economy first and foremost to greater opportunities for inter-industry learning. As such, it stresses the dynamic efficiency of diversified economies – and in particular of economies in which different industries are related to one another. The complexity approach, in contrast, regards industrial diversity as a sign of a broad capability base. In the economic complexity framework, an industry can only emerge in places that offer all the capabilities it requires. This idea has been illustrated with the metaphor of the game of Scrabble. In Scrabble, players hold letters that allow them to put together words. However, a word can only be written once a player owns every single letter it requires. In analogy, cities can only develop industries if they can offer each and every capability the industry requires. More diversified economies therefore typically dispose of a wider variety of capabilities and more complex industries will only be able to locate in few, highly complex cities. Moreover, diversification will be path-dependent, branching into nearby activities in the industry space. However, this related diversification is not considered to be *optimal*. Rather, industry spaces *constrain* economies to incremental change and may prevent them from moving immediately into industries that are most productive or that pay the highest wages. Unlike the Schumpeterian learning dynamics that underlie the concept of related variety, the Scrabble logic thus reflects static efficiency: it explains why certain cities can host industries that other cities cannot.

In the paper, we aimed to describe these and other differences and commonalities between the different lines of research, as well as critically assess some of the theoretical and empirical claims they make. Doing so, we pointed out a number of inconsistencies between the underlying conceptual frameworks and the empirical strategies that have been developed.

Furthermore, we proposed a measurement methodology that allows bridging the different research lines. This methodology first builds on existing work in ecology to quantify what we have called *generalized diversity*. We showed how this generalized diversity can be decomposed into three components: variety, balance and disparity. Furthermore, we showed how this generalized diversity can be used to calculate a *relative* diversity, i.e., the diversity a local industry finds in a city among a set of closely related neighbors. Armed with these new tools, we showed how to scrutinize – in a principled and unified way – some of the main theoretical claims in the newly emerging literature on the importance of diversity in local economic development.

This exercise yielded a set of preliminary, yet interesting results. First, we documented that findings that build on the notions of related and unrelated variety are sensitive to *ad hoc* choices about how to measure relatedness and the thresholds at which two activities are considered to be related or not. Second, we discussed why the ECI cannot immediately be interpreted as a measure of the fundamental diversity of a city's capability base. Yet, we also found that it does correlate fairly well with generalized diversity and that it is a strong predictor of a city's average wage level. Third, we showed how the empirical regularity of related diversification documented in the product space literature is not necessarily due to a large diversity in related activities, but due to the importance of the (correlated) mass of related activities in a region.

There are a number of important caveats to our study. First, the debate on diversity is both older and larger than what we cover in this paper. However, the limited focus allowed us to focus on the recent contributions to this debate and to provide some nuance on the different intellectual positions these contributions assume. Yet, even within this narrower scope, we had to leave out many contributions. For instance, several proposals have been made to improve the related and unrelated variety framework (e.g. Kogler et al. (2013)). Similarly, alternatives to the ECI and PCI have been proposed (e.g. Tacchella et al. (2012)).

Second, although the generalized and relative diversity measures and their decomposition are helpful tools to study different aspects of urban diversity, we do not claim that they are optimal. Alternatives exist – even within the Hill number approach we followed – and should be explored. Moreover, the fact that, in spite of the difficulties in interpretation, the ECI outperforms generalized diversity in predicting urban wage levels suggests that there is still much we do not understand about the relation between a city's industry mix and its growth potential.

Third, the aim of our empirical analyses was not to prove or disprove specific hypotheses, but rather to show that empirical findings can depend crucially on modelling choices. Therefore, we left a number of important issues unexplored. Importantly, we did not make any attempts to deal with issues of miss-specification or endogeneity in our statistical models. We also did not explore to what extent findings differ across contexts. For instance, the relation between diversity and growth may be different in different sectors or across the urban hierarchy.

In spite of this, we believe that our paper clarifies some important conceptual distinctions in the literature that have so far remained somewhat implicit. Moreover, we offer new empirical tools to explore the empirical importance of these distinctions. We hope that this has created a solid starting point for future research that not only addresses the aforementioned shortcomings, but also other concerns and research questions. The companion *python* code in the format of Jupyter notebooks codifies the construction of variables, as well as of our regression models and should therefore facilitate others to build on this work. Ultimately we hope that this will help arrive at a better understanding of, and, possibly even a scientific consensus about, the role of diversity in the growth and development of local economies.

5.6 Data

The data are taken from the Bureau of Labor Statistics (BLS) and come from three main sources: the Quarterly Census of Employment and Wages (QCEW) for employment (E) and wages (w);[23] the Local Area Unemployment Statistics (LAUS) for unemployment (U);[24] and the Occupational Employment Statistics (OES) for occupations.[25]

5.6.1 Employment, Unemployment and Wages

The data cover the time window 1990–2006, with 369 Metropolitan Statistical Areas (MSA) and 278 industries in each year.[26]

To aggregate the data at the MSA level from data at the County level we use a crosswalk from the US Census Bureau (2004 MSA definition;[27] i.e., the same used by BLS).[28] In this way, the MSAs considered in the paper are consistent over time, in terms of their composition of counties.

The industries are classified according to the NAICS 2002 system.[29] We consider industries at the 4-digit level, so the data consist of 278 industry groups at the 4 digit level, 78 sub-sectors at the 3-digit level, and 20 sectors at the 2-digit level.[30]

Wages refer to the average annual wage per-employee. For employment and wages, "undisclosed" information is dropped; i.e., there are no employees and the variable *avg. wage* is zero for these city-industry pairs.[31]

[23] https://www.bls.gov/cew/downloadable-data-files.htm

[24] https://download.bls.gov/pub/time.series/la/la.data.64.County

[25] https://www.bls.gov/oes/

[26] The following NAICS codes are excluded: 11 (Agriculture, forestry, fishing and hunting); 21 (Mining, quarrying, and oil and gas extraction); 49 (Postal service, delivery services, warehousing); 92 (Public administration); 99 (Unclassified); 482 (Rail transportation); 814 (Private households); 5211 (Monetary authorities - central bank).

[27] https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2003/historical-delineation-files/0312cbsas-csas.xls

[28] https://www.bls.gov/cew/questions-and-answers.htm

[29] Data from 1990-2000 were originally coded in the 1987 SIC classification. In a NAICS reconstruction project, the data had been reclassified to the NAICS 2002 classification.

[30] https://www.bls.gov/sae/additional-resources/what-is-naics.htm
https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2002

[31] https://www.bls.gov/cew/overview.htm#confidentiality

5.6.2 Occupations

We further use occupation-MSA and occupation-industry tables for the year 2002, which cover the same set of 278 industries. However, in the occupation tables, there are 337 MSAs and there is no exact correspondence to the MSAs used in the industry tables. Indeed, while for CEW data we use the 2004 MSA definition, the BLS provides the OES database already aggregated at the MSA level and in accordance with the 1999 MSA definition. Since the latter uses NECTA areas for the New England states (i.e., an aggregation of towns and not counties), it is impossible to make the two sources consistent.

To obtain consistent occupation labels, the data have been harmonized by taking the intersection of occupations across the MSA and industry tables. This resulted in 688 occupations at the 6-digit "detailed" level (SOC 2010 classification) after excluding the following SOC codes: 11-1031 (Legislators); 11-9131 (Postmasters and mail superintendents); 13-2081 (Tax examiners, collectors, and revenue agents); 23-1021 (Administrative law judges, adjudicators, and hearing officers); 23-1023 (Judges, magistrate judges, and magistrates); 33-3011 (Bailiffs); 33-3031 (Fish and game wardens); 39-6031 (Flight Attendants); 43-5051 (Postal service clerks); 43-5052 (Postal service mail carriers); 43-5053 (Postal service mail sorters, processors, and processing machine operators); 47-5061 (Roof bolters, mining); 49-9097 (Signal and Track Switch Repairers); 51-8011 (Nuclear Power Reactor Operators); 53-2011 (Airline Pilots, Copilots, and Flight Engineers); 53-4011 (Locomotive Engineers); 53-4021 (Railroad Brake, Signal, and Switch Operators); 53-4031 (Railroad Conductors and Yardmasters); 53-6011 (Bridge and lock tenders).

5.7 List of variables

Table 5.13: Overview of variables and descriptive statistics

Basic variables	
P_{ic}	industry-city matrix ($LQ > 1$)
E_{ic}	industry employment matrix
p_{ic}	employment share of industry i in city c
$E^{i}_{i'c}$	proximity-weighted employment of i' relative to i in city c
$p^{i}_{i'c}$	proximity-weighted employment share of i' relative to i in city c
Proximity matrices	
$\tilde{\phi}_{ii'}$	co-occurrence based industry proximity matrix
$\tilde{\psi}_{ii'}$	occupation based industry proximity matrix
$\tilde{\rho}_{ii'}$	growth correlation based industry proximity matrix
$\tilde{\phi}_{cc'}$	industry based city proximity matrix
$\tilde{\phi}_{oo'}$	co-occurence based occupation proximity matrix
City level variables	
$\mathbf{p}_c$	vector of industry employment shares
E_c	total employment in city c
$S(\mathbf{p}_c)$	entropy of industry employment in city c
UV_c	unrelated variety in city c
RV_c	related variety in city c
$D_I(\mathbf{p}_c)$	‘effective number’ of industries in city c
$D_Z(\mathbf{p}_c)$	(disparity-weighted) diversity of industries in city c
var_c	(normalized) variety of industries in city c
bal_c	balance of industries in city c
$disp_c$	disparity of industries in city c
City-industry level variables	
$\mathbf{p}^{i}_{c}$	vector of industry employment shares relative to i in city c
D^{i}_{c}	density of industries relative to i in city c
E^{i}_{c}	total employment (mass) of industries relative to i in city c
$D_I(\mathbf{p}^{i}_{c})$	‘effective number’ of industries relative to i in city c
$D_Z(\mathbf{p}^{i}_{c})$	(disparity-weighted) diversity of industries relative to i in city c
var^{i}_{c}	(normalized) variety of industries relative to i in city c
bal^{i}_{c}	balance of industries relative to i in city c
$disp^{i}_{c}$	disparity of industries relative to i in city c

Table 5.14: Descriptive statistics for city level data.

	count	mean	std	min	25%	50%	75%	max
$\ln E_c$	369	10.94	1.36	6.67	9.98	10.67	11.68	15.61
$\ln w_c$	369	9.81	0.20	8.88	9.71	9.81	9.93	10.43
$\ln U_c$	369	8.88	1.05	6.63	8.14	8.63	9.38	13.03
$\ln E_{cT}/E_{ct}$	369	0.44	0.24	-0.08	0.29	0.41	0.57	1.77
$\ln w_{cT}/w_{ct}$	369	0.58	0.09	0.22	0.52	0.57	0.63	1.21
$\ln U_{cT}/U_{ct}$	369	0.00	0.29	-1.34	-0.17	-0.00	0.19	0.89
Related-Unrelated variety								
RV_c 1-dig.	369	2.35	0.37	0.74	2.14	2.35	2.59	3.13
RV_c 2-dig.	369	1.74	0.27	0.54	1.59	1.76	1.94	2.37
RV_c 3-dig.	369	0.73	0.17	0.06	0.62	0.74	0.86	1.12
UV_c 1-dig.	369	1.73	0.09	1.12	1.70	1.75	1.79	1.87
UV_c 2-dig.	369	2.33	0.20	1.13	2.25	2.36	2.46	2.63
UV_c 3-dig.	369	3.34	0.27	2.02	3.21	3.35	3.51	3.84
Diversity using the classification-based proximity								
$\ln D_Z(\mathbf{p}_c)$	369	1.14	0.35	0.58	0.79	1.12	1.49	1.72
$\ln \mathrm{var}_c$	369	-0.91	0.43	-3.14	-1.16	-0.89	-0.62	-0.04
$\ln \mathrm{bal}_c$	369	-0.64	0.17	-1.84	-0.70	-0.62	-0.54	-0.20
$\ln \mathrm{disp}_c$	369	-2.93	0.26	-3.35	-3.08	-2.96	-2.83	-1.37
Diversity using the co-occurrence-based proximity								
$\ln D_Z(\mathbf{p}_c)$	369	1.51	0.15	0.94	1.40	1.49	1.65	1.79
$\ln \mathrm{var}_c$	369	-0.91	0.43	-3.14	-1.16	-0.89	-0.62	-0.04
$\ln \mathrm{bal}_c$	369	-0.64	0.17	-1.84	-0.70	-0.62	-0.54	-0.20
$\ln \mathrm{disp}_c$	369	-2.56	0.32	-3.18	-2.75	-2.59	-2.44	-0.88
Diversity using the cognitive-proximity-based proximity								
$\ln D_Z(\mathbf{p}_c)$	369	1.38	0.16	0.66	1.27	1.37	1.49	1.70
$\ln \mathrm{var}_c$	369	-0.91	0.43	-3.14	-1.16	-0.89	-0.62	-0.04
$\ln \mathrm{bal}_c$	369	-0.64	0.17	-1.84	-0.70	-0.62	-0.54	-0.20
$\ln \mathrm{disp}_c$	369	-2.69	0.27	-3.23	-2.88	-2.72	-2.58	-1.42
Diversity using the growth-similarity-based proximity								
$\ln D_Z(\mathbf{p}_c)$	369	1.36	0.21	0.76	1.26	1.31	1.38	2.21
$\ln \mathrm{var}_c$	369	-0.91	0.43	-3.14	-1.16	-0.89	-0.62	-0.04
$\ln \mathrm{bal}_c$	369	-0.64	0.17	-1.84	-0.70	-0.62	-0.54	-0.20
$\ln \mathrm{disp}_c$	369	-2.71	0.28	-3.15	-2.92	-2.74	-2.59	-1.26

Note: Wherever not necessary, the subscript t is omitted for brevity.

Table 5.15: Descriptive statistics for city-industry level data, using the classification-based proximity.

	count	mean	std	min	25%	50%	75%	max
$\ln E_{ic}$	43315	5.90	1.59	0.00	4.79	5.80	6.92	12.43
$\ln w_{ic}$	43315	9.79	0.49	7.68	9.44	9.83	10.13	13.14
$\ln E_{icT}/E_{ict}$	43315	0.34	0.76	-5.02	-0.06	0.31	0.71	5.84
$\ln w_{icT}/w_{ict}$	43315	0.57	0.26	-2.53	0.43	0.56	0.70	3.91
$\ln E_c^i$	43315	5.91	1.42	0.02	4.82	5.76	6.80	10.35
$\ln D_c^i$	43315	-1.34	0.44	-5.25	-1.49	-1.25	-1.06	-0.35
$\ln D_I(\mathbf{p}_c^i)$	43315	3.71	0.61	0.54	3.55	3.86	4.13	4.54
$\ln D_Z(\mathbf{p}_c^i)$	43315	1.22	0.14	0.70	1.23	1.28	1.30	2.98
$\ln \text{var}_c^i$	43315	-1.25	0.65	-4.93	-1.42	-1.06	-0.81	-0.49
$\ln \text{bal}_c^i$	43315	-0.66	0.19	-2.82	-0.74	-0.65	-0.57	-0.04
$\ln \text{disp}_c^i$	43315	-2.49	0.55	-3.24	-2.85	-2.63	-2.32	2.33
$\ln E_c$	43315	11.45	1.43	6.58	10.31	11.25	12.42	15.61
$\ln E_i$	43315	12.42	1.11	4.79	11.63	12.52	13.23	14.74

Note: Wherever not necessary, the subscript t is omitted for brevity.

Table 5.16: Descriptive statistics for city-industry level data, using the industry space as defined by $\tilde{\psi}_{ii'}$ of eq. (5.12).

	count	mean	std	min	25%	50%	75%	max
$\ln E_{ic}$	43322	5.90	1.59	0.00	4.79	5.79	6.92	12.43
$\ln w_{ic}$	43322	9.79	0.49	7.68	9.44	9.83	10.13	13.14
$\ln E_{icT}/E_{ict}$	43322	0.34	0.76	-5.02	-0.06	0.31	0.71	5.84
$\ln w_{icT}/w_{ict}$	43322	0.57	0.26	-2.53	0.43	0.56	0.70	3.91
$\ln E_c^i$	43322	5.84	1.42	0.61	4.71	5.66	6.80	10.42
$\ln D_c^i$	43322	-1.39	0.30	-3.86	-1.56	-1.37	-1.19	-0.51
$\ln D_I(\mathbf{p}_c^i)$	43322	4.07	0.44	1.29	3.82	4.12	4.39	4.97
$\ln D_Z(\mathbf{p}_c^i)$	43322	0.78	0.07	0.32	0.75	0.78	0.82	1.61
$\ln \text{var}_c^i$	43322	-0.85	0.43	-4.02	-1.15	-0.82	-0.54	-0.05
$\ln \text{bal}_c^i$	43322	-0.70	0.19	-2.51	-0.79	-0.68	-0.58	-0.02
$\ln \text{disp}_c^i$	43322	-3.29	0.43	-4.16	-3.60	-3.34	-3.04	0.11
$\ln E_c$	43322	11.45	1.43	6.58	10.30	11.25	12.42	15.61
$\ln E_i$	43322	12.42	1.11	4.79	11.63	12.52	13.23	14.74

Note: Wherever not necessary, the subscript t is omitted for brevity.

Table 5.17: Descriptive statistics for city-industry level data, using the industry space as defined by $\rho_{ii'}$ of eq. (5.13).

	count	mean	std	min	25%	50%	75%	max
$\ln E_{ic}$	38484	5.92	1.58	0.00	4.83	5.82	6.93	12.43
$\ln w_{ic}$	38484	9.79	0.48	7.68	9.46	9.84	10.13	13.05
$\ln E_{icT} \big/ E_{ict}$	38484	0.32	0.75	-5.02	-0.07	0.29	0.68	5.84
$\ln w_{icT} \big/ w_{ict}$	38484	0.57	0.26	-2.53	0.43	0.56	0.70	3.91
$\ln E_c^i$	38484	5.97	1.41	0.33	4.91	5.81	6.88	10.76
$\ln D_c^i$	38484	-1.23	0.40	-3.93	-1.42	-1.17	-0.98	0.00
$\ln D_I(\mathbf{p}_c^i)$	38484	3.14	0.73	0.01	2.82	3.33	3.64	4.51
$\ln D_Z(\mathbf{p}_c^i)$	38484	0.84	0.39	0.32	0.61	0.73	0.98	6.75
$\ln \text{var}_c^i$	38484	-1.90	0.74	-4.93	-2.19	-1.72	-1.39	-0.60
$\ln \text{bal}_c^i$	38484	-0.58	0.22	-2.59	-0.65	-0.54	-0.46	-0.00
$\ln \text{disp}_c^i$	38484	-2.30	0.98	-3.63	-2.95	-2.60	-1.88	6.74
$\ln E_c$	38484	11.34	1.45	3.56	10.16	11.19	12.39	15.45
$\ln E_i$	38484	12.37	1.10	4.06	11.59	12.51	13.15	14.71

Note: Wherever not necessary, the subscript t is omitted for brevity.

Bibliography

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., & Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, *80*(5), 1977–2016.

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.

Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Maass, P., Salinger, M. A., Stanley, H. E., & M.H.R., S. (1997). Scaling behavior in economics: The problem of quantifying company growth. *Physica A*, *244*(1), 1–24.

Axtell, R. L. (2001). Zipf distribution of u.s. firm sizes. *Science*, *293*(5536), 1818–1820.

Bahar, D., Hausmann, R., & Hidalgo, C. A. (2014). Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion? *Journal of International Economics*, *92*(1), 111–123.

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.

Balland, P.-A., & Rigby, D. (2016). The geography of complex knowledge. *Economic Geography*, *93*(1), 1–23.

Baqaee, D., & Farhi, E. (2018). *The microeconomic foundations of aggregate production functions* (tech. rep.). National Bureau of Economic Research.

Baqaee, D. R., & Farhi, E. (2019). The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem. *Econometrica*, *87*(4), 1155–1203.

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

Beaulieu, N., & Xie, Q. (2004). An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology*, *53*(2), 479–489.

Bergounhon, F., Lenoir, C., & Mejean, I. (2018). *A guideline to french firm level trade data.*

Berman, B. (2010). *Retail management : A strategic approach.* 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.

Boeri, T. (1989). Does firm size matter? *Giornale degli Economisti e Annali di Economia*, *48*(9/10), 477–495.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). Chapter 49 arch models. *Handbook of econometrics* (pp. 2959–3038). Elsevier.

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., Martın, V., & Minondo, A. (2016). Neighbour regions as the source of new industries. *Papers in Regional Science*, *96*(2), 227–245.

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.

Bottazzi, G., & Secchi, A. (2006). Explaining the distribution of firm growth rates. *The RAND Journal of Economics*, *37*(2), 235–256.

Bowen, H. P. (1983). On the theoretical interpretation of indices of trade intensity and revealed comparative advantage. *Weltwirtschaftliches Archiv*, *119*(3), 464–472.

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

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*(1).

Cai, J., & Leung, P. (2008). Towards a more general measure of revealed comparative advantage variation. *Applied Economics Letters*, *15*(9), 723–726.

Caldarelli, G., Cristelli, M., Gabrielli, A., Pietronero, L., Scala, A., & Tacchella, A. (2012). A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. *PLoS ONE*, *7*(10), 1–11.

Canals, C., Gabaix, X., Vilarrubia, J. M., & Weinstein, D. E. (2007). Trade patterns, trade balances and idiosyncratic shocks. *SSRN Electronic Journal*.

Carvalho, V. M., & Grassi, B. (2019). Large Firm Dynamics and the Business Cycle. *American Economic Review*, *109*(4), 1375–1425.

Castro, R., Clementi, G. L., & Lee, Y. (2015). Cross sectoral variation in the volatility of plant level idiosyncratic shocks. *The Journal of Industrial Economics*, *63*(1), 1–29.

Chesher, A. (1979). Testing the law of proportionate effect. *Journal of Industrial Economics*, *27*(4), 403–11.

Content, J., & Frenken, K. (2016). Related variety and economic development: A literature review. *European Planning Studies*, *24*(12), 2097–2112.

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

Coscia, M., & Neffke, F. M. (2017). Network backboning with noisy data. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 425–436.

Dalum, B., Laursen, K., & Villumsen, G. (1998). Structural change in OECD export specialisation patterns: De-specialisation and 'stickiness'. *International Review of Applied Economics*, *12*(3), 423–443.

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.

Deb, K., & Hauk, W. R. (2015). RCA indices, multinational production and the ricardian trade model. *International Economics and Economic Policy*, *14*(1), 1–25.

De_Benedictis, L., & Tamberi, M. (2001). A note on the balassa index of revealed comparative advantage. *SSRN Electronic Journal*.

De_Benedictis, L., & Tamberi, M. (2004). Overall specialization empirics: Techniques and applications. *Open Economies Review*, *15*(4), 323–346.

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.

de Solla Price, D. (1981). The analysis of scientometric matrices for policy implications. *Scientometrics*, *3*(1), 47–53.

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.

Dixit, A. K., & Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, *67*(3), 297–308.

Dupor, B. (1999). Aggregation and irrelevance in multi-sector models. *Journal of Monetary Economics*, *43*(2), 391–409.

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

Duranton, G., & Puga, D. (2004). Micro-foundations of urban agglomeration economies. *Handbook of regional and urban economics* (pp. 2063–2117). Elsevier.

Ellison, G., & Glaeser, E. L. (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.

Farmer, J. D., & Lillo, F. (2004). On the origin of power-law tails in price fluctuations. *Quantitative Finance*, *4*(1), 7–11.

Felipe, J., & Fisher, F. M. (2003). Aggregation in production functions: What applied economists should know. *Metroeconomica*, *54*(2-3), 208–262.

Filho, J. S., Cardieri, P., & Yacoub, M. (2005). Simple accurate lognormal approximation to lognormal sums. *Electronics Letters*, *41*(18), 1016.

Foerster, A. T., Sarte, P.-D. G., & Watson, M. W. (2011). Sectoral versus aggregate shocks: A structural factor analysis of industrial production. *Journal of Political Economy*, *119*(1), 1–38.

French, S. (2017). Revealed comparative advantage: What is it good for? *Journal of International Economics*, *106*, 83–103.

Frenken, K., Oort, F., & Verburg, T. (2007). Related variety, unrelated variety and regional economic growth. *Regional Studies*, *41*(5), 685–697.

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

Gabaix, X. (2011). The Granular Origins of Aggregate Fluctuations. *Econometrica*, *79*(3), 733–772.

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, *423*(6937), 267–270.

Gibrat, R. (1931). *Les inègalitès èconomiques*. Librairie du Recuil Sirey, Paris.

Giovanni, D., & Levchenko. (2014). Firms, destinations, and aggregate fluctuations. *Econometrica*, *82*(4), 1303–1340.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., & Shleifer, A. (1992). Growth in cities. *Journal of Political Economy*, *100*(6), 1126–1152.

Gomez-Lievano, A. (2018). Methods and Concepts in Economic Complexity. *arxiv.org/abs/1809.10781*.

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.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis* (3rd ed.). Academic Press.

Guiso, L., Lai, C., & Nirei, M. (2016). An empirical study of interaction-based aggregate investment fluctuations. *The Japanese Economic Review*, *68*(2), 137–157.

H. R. Stanley, M., Amaral, L., Buldyrev, S., Havlin, S., Leschhorn, H., Maass, P., Salinger, M., & Stanley, H. (1996). Scaling behavior in the growth of companies. *Nature*, *379*.

Hart, P. E., & Prais, S. J. (1956). The analysis of business concentration: A statistical approach. *Journal of the Royal Statistical Society. Series A (General)*, *119*(2), 150–191.

Hausmann, R., & Hidalgo, C. A. (2011). The network structure of economic output. *Journal of Economic Growth*, *16*(4), 309–342.

Hausmann, R., & Klinger, B. (2006). *Structural transformation and patterns of comparative advantage in the product space* (tech. rep. No. 128). CID Research Fellow & Graduate Student Working Paper, Harvard University.

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. *Unifying themes in complex systems IX* (pp. 451–457). 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.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Hinloopen, J., & Van Marrewijk, C. (2001). On the empirical distribution of the balassa index. *Weltwirtschaftliches Archiv*, *137*(1), 1–35.

Hoen, A., & Oosterhaven, J. (2006). On the measurement of comparative advantage. *The Annals of Regional Science*, *40*(3), 677–691.

Horvath, M. (1998). Cyclicality and sectoral linkages: Aggregate fluctuations from independent sectoral shocks. *Review of Economic Dynamics*, *1*(4), 781–808.

Jacobs, J. (1969). *The economy of cities* (tech. rep.). Vintage Books: A Division of Random House.

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

Jacod, J., & Protter, P. (2012). *Discretization of processes*. Springer Berlin Heidelberg.

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.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375.

Kemp-Benedict, E. (2014). An interpretation and critique of the Method of Reflections. *MPRA Paper No. 60705*.

Kogan, L., & Papanikolaou, D. (2012). Economic activity of firms and asset prices. *Annual Review of Financial Economics*, *4*(1), 361–384.

Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping Knowledge Space and Technological Relatedness in US Cities. *European Planning Studies*, *21*(9), 1374–1391.

Koren, M., & Tenreyro, S. (2007). Volatility and Development. *The Quarterly Journal of Economics*, *122*(1), 243–287.

Kozubowski, T., & Podgorski, K. (2003). Log-laplace distributions. *Int. Math. J.*, *3*, 467–495.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, *99*(3), 483–99.

Kunimoto, K. (1977). Typology of trade intensity indices. *Hitotsubashi Journal of Economics*, *17*(2), 15–32.

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica*, *50*(6), 1345.

Laursen, K. (1998). *Revealed comparative advantage and the alternatives as measures of international specialisation* (DRUID Working Papers No. 98-30). DRUID, Copenhagen Business

School, Department of Industrial Economics and Strategy/Aalborg University, Department of Business Studies.

Laursen, K. (2015). Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian Business Review*, *5*(1), 99–115.

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3), 477–489.

Leromain, E., & Orefice, G. (2014). New revealed comparative advantage index: Dataset and empirical distribution. *International Economics*, *139*, 48–70.

Liu, B., & Gao, J. (2019). Understanding the non-gaussian distribution of revealed comparative advantage index and its alternatives. *International Economics*, *158*, 1–11.

Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, *36*(3), 1049–1051.

Long, J. B., & Plosser, C. I. (1983). Real business cycles. *Journal of Political Economy*, *91*(1), 39–69.

Lucas, R. E. (1977). Understanding business cycles. *Carnegie-Rochester Conference Series on Public Policy*, *5*, 7–29.

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

Mandelbrot, B. B. (1997). A case against the lognormal distribution. *Fractals and scaling in finance* (pp. 252–269). Springer New York.

Marlow, N. A. (1967). A normal limit theorem for power sums of independent random variables. *Bell System Technical Journal*, *46*(9), 2081–2089.

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.

Mealy, P., Farmer, J. D., & Teytelboym, A. (2019). Interpreting economic complexity. *Science Advances*, *5*(1), eaau1705.

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., Hartog, M., Boschma, R., & Henning, M. (2014). *Agents of structural change. the role of firms and entrepreneurs in regional diversification* (Papers in Evolutionary Economic Geography (PEEG) No. 1410). Utrecht University, Section of Economic Geography.

Neffke, F., Hartog, M., Boschma, R., & Henning, M. (2017). Agents of structural change: The role of firms and entrepreneurs in regional diversification. *Economic Geography*, *94*(1), 23–48.

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, *34*(3), 297–316.

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.

Neffke, F. M., Otto, A., & Weyh, A. (2017). Inter-industry labor flows. *Journal of Economic Behavior & Organization*, *142*, 275–292.

Nguyen, L. X. D., Mateut, S., & Chevapatrakul, T. (2020). Business-linkage volatility spillovers between US industries. *Journal of Banking & Finance*, *111*, 105699.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, *36*, 1016–1034.

Pasten, E., Schoenle, R., & Weber, M. (2019). The propagation of monetary policy shocks in a heterogeneous production economy. *Journal of Monetary Economics*.

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*. 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.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, *21*(1), 24–43.

Redding, S. J., & Proudman, J. (1998). Productivity convergence and international openness. *SSRN Electronic Journal*.

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.

Schetter, U. (2019). *A structural ranking of economic complexity* (tech. rep. No. 119). CID Research Fellow & Graduate Student Working Paper, Harvard University.

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

Schwartz, S. C., & Yeh, Y. S. (1982). On the distribution function and moments of power sums with log-normal components. *Bell System Technical Journal*, *61*(7), 1441–1462.

Soete, L. G., & Wyatt, S. M. E. (1983). The use of foreign patenting as an internationally comparable science and technology output indicator. *Scientometrics*, *5*(1), 31–54.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, *4*(15), 707–719.

Stockman, A. C. (1988). Sectoral and national aggregate disturbances to industrial output in seven european countries. *Journal of Monetary Economics*, *21*(2-3), 387–409.

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. (2019). Diversity and its decomposition into variety, balance and disparity. *Royal Society Open Science*, *6*(7), 190452.

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.

Vollrath, T. L. (1991). A theoretical evaluation of alternative trade intensity measures of revealed comparative advantage. *Weltwirtschaftliches Archiv*, *127*(2), 265–280.

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.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Yamamoto, M. (2014). A moment method of the log-normal size distribution with the critical size limit in the free-molecular regime. *Aerosol Science and Technology*, *48*(7), 725–737.

Yu, R., Cai, J., & Leung, P. (2009). The normalized revealed comparative advantage index. *The Annals of Regional Science*, *43*(1), 267–282.

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