THE CAPABILITIES APPROACH TO ECONOMIC DEVELOPMENT: ON DIVERSITY, COMPLEXITY AND RELATEDNESS

ALJE VAN DAM

The capabilities approach to economic development: on diversity, complexity and relatedness

Economische ontwikkeling in termen van competenties: diversiteit, complexiteit en gerelateerdheid

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 13 november 2020 des middags te 4.15 uur

door

Alje van Dam

geboren op 3 november 1991 te Lilongwe, Malawi

Promotor: Prof. dr. K. Frenken

Copromotoren: Dr. A. Gomez-Lievano

Dr. F. Neffke

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under the Vici scheme, number 453-14-014.

Contents

Chapter 1

Introduction

Traditional models of long-term economic development typically consider how aggregate input like labor and capital get translated into aggregate economic output. This output is often quantified by some measure of size of an economy such as Gross Domestic Product. The issues pertaining to economic development then become the accumulation of aggregate inputs, the improvement of efficiency of inputs through technological innovation, and the optimal proportions of those inputs, all to achieve an increase in the size of the economy.

Recently, there has been a growing interest in the *structure* of economies as opposed to their size, considering the portfolio of specific economic activities in an economy. Differentiating between different types of activities does not only provide a more accurate description of an economy's state, but also gives insight in its future development. The export portfolio of a country, for example, is predictive of its future growth. That is, it is not how much you export, but *what* you export that matters (Hausmann et al., 2007).

Following the shift in focus from the aggregate economy to its structure, one can also shift the focus from aggregate inputs such as labor and capital to the specific inputs that are necessary to make each particular product. These inputs, referred to as *capabilities*, include not only the physical resources and assets that are required to make a product, but also the know-how required to make them, along with even more abstract requirements such as institutions and regulations needed for a production process (Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009). In this view, missing any of the 'ingredients' to produce a particular product will prevent an economy from producing that product. And, as capabilities are thought of to consist for large part of tacit knowledge that is not transferred easily, missing capabilities are hard to imitate from other countries.

In this view, economic development is a process of collective learning. Individuals specialize in mastering only one or few capabilities. Production, however, necessitates the coming together of many capabilities. As a consequence, countries need to coordinate the capabilities that are distributed among people, and they do so typically within formal organizations like firms. economics development can then be considered to occur through individuals who acquire new capabilities, and recombine them with existing ones, allowing an economy to make new products. This view does not only underlie the capabilities model (Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009), but is also in line with models of recombinant growth (Weitzman, 1998) and cultural evolution (Muthukrishna and Henrich, 2016). All these theories emphasize the increasing complexity of societies as they develop over time, due to the accumulation and recombination of capabilities.

This line of thinking can be thought of as a first step towards a 'theory of economic complexity' (Gomez-Lievano, 2018). Such a theory may help explain the diverging rates of development between rich and poor economies: economies that already have many capabilities can easily develop many new activities by recombining one new capability with the many existing ones. Economies with few capabilities face a much harder challenge, since for such countries a new capability enables only few new activities, as there are fewer existing capabilities to recombine a new capability with. Having few capabilities thus implies limited growth opportunities (Hausmann and Hidalgo, 2011). The capability model further sheds light on the extreme concentration of economic activities in cities, as they act as the typical places where capabilities accumulate and get recombined (Gomez-Lievano et al., 2016; Gomez-Lievano and Patterson-Lomba, 2018). In particular, the more complex -and valuable- products tend to concentrate in the largest cities (Balland et al., 2020).

Against the background of the capabilities framework, a large body of empirical literature has emerged that addresses questions of economic development by studying the structure of economies. These studies build on three central concepts. The first concept is *diversity*, where the main interest is to understand to what extent economic development is accompanied by an increase in the variety of products being produced. This is to be expected as the accumulation of capabilities over time allows an economy to produce an increasing variety of products. The second concept is *complexity*, which also follows from the recombinant logic of economic development. The acquisition of new capabilities over time does not only allow an economy to produce a larger variety of products but also more complex products, where complexity refers to the number of capabilities required to produce a product. The final concept is *relatedness.* As new products emerge by recombining new capabilities with existing ones, the new products will be related to the existing ones in the sense that they will rely for large part on the same set of capabilities as existing products.

1.1 Variety, complexity and relatedness

1.1.1 Variety

A key implication of the capabilities framework is that development is characterized by diversification as opposed to specialization, since the number of possible recombinations grows exponentially with the number of capabilities (Inoua, 2016). Following the assumption that countries produce all products that their capabilities allow them to produce, the variety of products that a country produces increases exponentially with the acquisition of new capabilities. This predicted pattern is consistent with the empirical observation that, at least for low-income countries, variety increase as these countries develop (Imbs and Wacziarg, 2003; Cadot et al., 2011). Moreover, recent evidence suggests that diversification is the main driver of economic development (Brummitt et al., 2020).

It should be further noted that variety can be seen not only as an outcome of the diversification process alone, but also as a source of economic growth (Weitzman, 1998; Saviotti and Frenken, 2008). This goes back to the ideas of Jacobs (1969), who argued that the higher the variety of economic activities, the more scope for new activities through recombination, leading to an ever greater variety of activities. While she did not theorize this endogenous dynamic in terms of underlying capabilities, Jacobs' early view can be considered as being consistent with the capabilities framework. Later, Frenken et al. (2007) qualified the benefits of variety in that recombinant innovation is more likely if the building blocks are technologically related rather than unrelated, as the inventors in question will find it easier to combine the underlying knowledge into useful new products and activities. An optimal portfolio thus consists of a high variety of activities that are also related, leading to the concept of *related* variety.

In their study, Frenken et al. (2007) constructed a measure of related variety by inferring relatedness from existing hierarchical Standard Industrial Classification (SIC), which translates into discrete levels of relatedness (ranging from related to semi-related to unrelated). This approach has resulted in a sizeable empirical literature that studies the effect of related and unrelated variety on economic growth, as reviewed by Content and Frenken (2016). However, the measurement of related variety suffers from taking the industrial classifications for granted. Instead, one would wish to measure relatedness explicitly, also allowing relatedness to change over time. An alternative measure for diversity has been proposed by Stirling (2007). His measure accounts both for variety (number of items), balance (relative frequencies) and disparity (the inverse of relatedness). Though the measure combines all three relevant aspect of diversity, it suffers from the need to weigh, arbitrarily, the relative contribution of variety and balance on the one hand, and disparity on the other.

1.1.2 Complexity

While the structure of the economy can be described by its (related) variety, it does not necessarily say much about the number of capabilities that are used as inputs in a country's products. Some products are highly sophisticated and will require many different capabilities (like producing a satellite), while others (like growing cotton) require only few. The number of capabilities required for a specific economic activity can be thought of as its complexity. Only countries with many capabilities will be able to engage in highly complex activities, while economies with few capabilities will be stuck in producing simple products. Complexity is considered economically relevant as complex products are thought to contribute to a greater extent to a country's GDP than simple products, for example because they have high value-added or because they are limited in supply since only few economies are able to produce them (Hausmann and Hidalgo, 2011).

The average complexity of products in a country can in turn be interpreted as a measure of the 'economic complexity' of a country. However, it is hard to establish the complexity of products empirically in terms of their capabilities, as we currently lack ways to observe and measure capabilities at the level of each individual product. Instead, Hidalgo and Hausmann (2009) proposed an index of economic complexity that aims to proxy the complexity of countries by examining their export portfolios. The authors define the complexity of an economy as the average complexity of the products it exports. The complexity of a product is in turn defined as the average complexity of the countries producing it. This leads to an iterative procedure called the ’method of reflections’ that assigns to every country and product a score known as the economic complexity index (ECI), and to every product a score known as the product complexity index (PCI), respectively. Where the ECI is meant to be an estimate for the number of capabilities in a country, the PCI is meant to be a measure of the number of capabilities required for its production. The ECI was shown to predict differences in GDP per capita across countries, as well as future growth (Hidalgo and Hausmann, 2009; Hausmann et al., 2011). Subsequent work proposed a variation of the the original ECI and PCI indices by incorporating a different weighting in the iterative algorithm, leading to a ’fitness-complexity’ algorithm (Caldarelli et al., 2012; Tacchella et al., 2012), yielding more accurate growth forecasts (Zaccaria et al., 2015; Tacchella et al., 2018). These new measures nevertheless continue to rely on an iterative procedure to infer complexity from export portfolios in an indirect manner, and lack a theoretical foundation. There have been limited attempts to provide a theoretical foundation for the complexity indices (noteable exceptions being (Schetter, 2019; Gomez-Lievano and Patterson-Lomba, 2018; Bustos and Yildirim, 2019)).

1.1.3 Relatedness

Relatedness is another central notion that has emerged in research on economic development over the past decade. This research has been summarized by Boschma (2017) and Hidalgo et al. (2018). The relatedness term obviously links to the notion of related variety, yet goes beyond the standard industrial classifications on which the related variety measure relies. Instead, authors attempt to measure the relatedness between each pair of economic activity. The ’relatedness’ literature also differs from the related-variety literature in that the latter focuses on economic growth, while the former focuses on economic diversification. Following the capabilities framework, diversification happens once new capabilities recombined with existing ones, leading to new products and services in an economy. Thus, new products will be related to the existing ones to the extent that new products share capabilities with existing products. This implies that diversification patterns are highly path dependent: the set of capabilities determines what an economy can make, but also sets the stage for future development by determining which new capabilities may lead to new economic activities through recombination. Hence, economies are constrained to develop new activities that are similar to the ones they were already engaged in, since they build on mostly on the same set of capabilities (Hidalgo et al., 2007).

One way to gain insight into the path-dependence of economic development is to study what activities enter and exit economies. If new activities build on new and exisiting capabilities, then new activities should be related to those that are already present, a process known as *related diversification*. This has been tested empirically by mapping out the relatedness between products (known as a 'product space') and showing that activities that enter the economy are close to existing activities in this space (Hidalgo et al., 2007; Neffke et al., 2011; Boschma et al., 2013) and activities that exit an economy are farther away from existing activities (Neffke et al., 2011). The probability of entry and exit of a new activities can thus be estimated by how 'far' it is in the product space from the current portfolio of an economy. Using this approach, many studies have shown related diversification to take place across a wide range of settings (Boschma, 2017; Hidalgo et al., 2018). The product space approach is particularly powerful as it, next to testing for related diversification empirically, allows visualization of the product space as a network, visualizing the relatedness structure underlying economic development (Hidalgo and Hausmann, 2008). The position of an economy in the product space determines its diversification opportunities: economies located in dense parts of the product space with plenty of growth opportunities have better growth prospects than economies that are 'stuck' in sparse parts of the network. It thus matters in which direction diversification takes place for future development, and the structure of the product space may suggest optimal paths of diversification (Alshamsi et al., 2018).

Methodologically, however, challenges remain. As for complexity measures, relatedness measures face the difficulty that the capabilities underlying economic activities are unobservable. One approach to measure relatedness is to use data that reflect economically meaningful relations between economic activities, for example through input-output relations between industries (Essletzbichler, 2015) or inter-industry labor flows (Neffke and Henning, 2013). Another widely applied approach, which is less demanding in terms of data availability, is to infer relatedness from the co-location of activities, based on the assumption that activities that often co-occur in countries or regions, are likely to build on the same set of capabilities (Hidalgo et al., 2007; Boschma et al., 2013).

1.2 This thesis

The focus on the structure of the economy has inspired a new and extensive body of empirical research on related variety, economic complexity and relatedness. Yet, while empirical research is blossoming in three parallel tracks, one can find very few attempts to link the concepts of variety, complexity and relatedness – be it theoretically, empirically or methodologically. Accordingly, the motivation underlying this PhD thesis is to elaborate the variety-complexity-relatedness framework as based on the capabilities model in ways that render the framework theoretically and methodologically more coherent. Having a clear framework is not only essential in order to drive questions in empirical research, but also to guide in deriving policy implications from empirical results.

The capabilities model (Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009; Inoua, 2016) provides an elementary framework to theorise about diversity, complexity and relatedness jointly, as well as about their interrelations. In this model, economic development stems from the acquisition of new capabilities. Combined with existing capabilities, a new capability will lead to a larger variety of products as well as, on average, more complex products. And, as new products emerge by recombining new capabilities with existing ones, the new products will be related to the existing ones. Thus, the capabilities model is capable to integrate all three key concepts in a relatively simple framework.

Another way of adding coherence to the literature is by proposing a consistent methodological framework. While measures of diversity, complexity and relatedness exist, they are often ad-hoc in nature and developed separately from each other. While often based on the same data, it is not well understood how these measure relate to each other methodologically. For example, what is the connection between relatedness and diversity? And should we expect more complex economies to be inherently more diverse, based on how we measure these things? Taking a more formal approach in constructing these measures and being more explicit about the assumptions underlying them may lead to a consistent methodological framework that allows a better understanding of how each of these concepts relate to each other. The objective is thus not only to obtain better measures, but also to gain better understanding of the nature of these concepts themselves. On top of that, a principled methodological approach may help to better understand the properties and implicit assumptions underlying currently used measures, which in turn may lead to new insights about the empirical regularities that have been established in the literature.

This thesis is built up in four parts divided into seven chapters. The first part serves as an introductory study, providing a review on the role of diversity in economic development and a number of empirical tests. In particular, it explores the subtle differences and interrelations in how the conceptualisation and measurement of diversity in the literature on related variety, economic complexity and related diversification. The second part is methodological and attempts to advance the mathematical and statistical foundations of measures of diversity, complexity and relatedness. The third part extends the basic capabilities model by relaxing the core assumption that countries produce all products that could be made with a set of capabilities, addressing the relation between diversification and economic development and providing a model to study different types of policy within the capabilities framework. The final part concludes. The following subsections provide a brief review of the contents of each part of this thesis.

1.2.1 Part 1: Setting the scene

Chapter 2 discusses the role of diversity in relation to economic complexity Hidalgo and Hausmann (2009), related variety (Frenken et al., 2007), and related diversification (Hidalgo et al., 2007; Neffke et al., 2011), and showcases the methodological approaches in each of these literatures. As explained, the related variety notion regards relatedness as cognitive proximity between an economy's products or industries, so that diversity with low disparity (a high relatedness between the activities considered) is beneficial as this allows for easy recombination. In the capability framework, however, it is not the diversity of products but of capabilities that is relevant, as a higher number capabilities implies that an economy can produce more, and more complex, products. Hence, having diversity with high disparity would be best for economic development, as it implies having many capabilities. To deal with these questions, I propose a methodology that attempts to express relevant measures in terms of diversities. This allows empirically testing the different hypotheses following from the literature by regressing the various diversity measures on economic growth.

In doing so, this chapter sets the scene for the following three methodological chapters that attempt to resolve some of the critical methodological issues that remain in measuring diversity, complexity and relatedness. Each of the three chapters deals with one of the three concepts separately and does so from a purely methodological point of view as to improve the foundations of our measurements.

1.2.2 Part 2: Methodological advances

Chapter 3 deals with the measurement of diversity whilst taking into account its three dimensions: variety (the number of types), balance (the distribution over types), and disparity (how dissimilar are the types) (Stirling, 2007). The explicit role of each of the three dimensions of diversity in economic development is yet to be considered, in particular in the context of the capabilities framework. Digging deeper into these questions requires a methodological framework for the measurement of diversity and that takes into account the relative contributions of variety, balance and disparity in a principled way.

In this chapter, I address the general issue of measuring diversity of a set of economic activities that have an overlap in the number of capabilities they require. The measurement of diversity has been discussed extensively, mainly in ecology (Hill, 1973a; Jost, 2006; Purvis and Hector, 2000; Tuomisto, 2010). This has resulted in a framework for the measurement of diversity based on 'Hill numbers', which shows the mathematical relation between many widely used diversity indices, and provides a formal way of measuring diversity. Furthermore, it has been extended to include similarities (viz. relatedness) between the elements of the communities under consideration (Chao et al., 2014; Leinster and Cobbold, 2012). This approach is limited, however, in that it considers only pairwise relatedness between the elements considered, and does not distinguish in which way the pairs are different. The main contribution of the chapter is to extend this framework beyond pairwise relatedness, and to propose a decomposition of diversity into separate components of variety, balance and disparity.

Chapter 4 engages with the ongoing debate about the measurement of complexity, and in particular with the complexity indices introduced in (Hidalgo and Hausmann, 2009). Despite the explanatory power and appealing rankings produced by the indices, their exact interpretation has remained a topic of discussion. The complexity indices were first critiqued by Caldarelli et al. (2012) and Tacchella et al. (2012), who proposed an alternative in which they incorporate a different weighting in the iterative algorithm. This lead to the 'fitness-complexity' algorithm that was argued to yield better results for growth forecasts (Zaccaria et al., 2015; Tacchella et al., 2018), but has in turn been confronted with its own methodological issues (Morrison et al., 2017). The two sets of complexity indices have lead to (heated) debates on how the measures should be weighted (Albeaik et al., 2017a; Gabrielli et al., 2017; Albeaik et al., 2017b; Pietronero et al., 2017). These discussions mostly concern different arguments regarding weighting schemes, and how well the resulting country rankings align with priors and growth rates. Yet, the debate to date has provided little insight in what these complexity measures are capturing exactly.

Recently, the meaning of the complexity indices was uncovered by Mealy et al. (2019) and Gomez-Lievano (2018), who note that the ECI is mathematically equivalent to methods for clustering and dimensionality reduction. In fact, the 'method of reflections' was noted to be an exact reinvention of a method called 'reciprocal averaging', which is a way of deriving a statistical method called correspondence analysis (Hill, 1973a). These recent insights suggest that the ECI and PCI are to be interpreted as measures of similarity rather than complexity. Nonetheless, the empirical relation between ECI and measures of productivity such as GDP per capita is robust, and found consistently both on national and sub-national scale (Hidalgo and Hausmann, 2009; Chávez et al., 2017; Gao and Zhou, 2018; Mealy and Coyle, 2019). This suggests that there is nevertheless a close relation between the specific activities that take place in economies and per capita income, going back to the original finding that 'what you exports matters' (Hausmann et al., 2007) from which most works originated.

In this chapter, I explore the economic complexity index (ECI) as an application of correspondence analysis (CA). CA has been widely applied in ecology as a method of ordination, in which the objective is to rank for example species by some latent variable, based on their occurrence patterns. The mathematics behind CA provides a rich set of tools and interpretations that lead to new insights in what the economic complexity indices actually capture, and provide new avenues for further research. We describe the different interpretations of CA as a method of identifying latent variables, clustering, and dimensionality reduction. We explore these interpretations by applying them to datasets from both ecology and economics, and propose a way to deal with the challenge of applying these ordination methods to clustered data.

Chapter 5 concerns the ways in which co-location patterns are used to derive a measure of relatedness. Following the first study by Hidalgo et al. (2007), relatedness measures are often derived from the empirical co-location patterns of products, industries or professions. There are, however, many different ways to devise such measures (Seung-Seok et al., 2010). Methodological frameworks evaluating different measures of co-location in the current context are lacking, with one notable exception (van Eck and Waltman, 2009).

Some of the currently used measure have problematic properties. For example, the relatedness measure used by Hidalgo et al. (2007) is biased towards ubiquitous products. Since the measures is based on the minimum of two conditional probabilities, it assigns high relatedness to ubiquitous products (with high marginal probabilities), even when they are indepdendently distributed (Muneepeerakul et al., 2013). Such biases may greatly affect further analyses.

Another issue is the binarizing procedure that is often applied using Ballassa's index of revealed comparative advantage (Balassa, 1965). It may transform the data in unpredictable ways and discards potentially relevant information. Alternative measures of relatedness derived from co-location have made some improvements in this respect, but, like the original measure by Hidalgo et al. (2007), they are not derived from first principles. The literature thus lacks a framework that allows analysis of co-location data in a principled way.

In this chapter, I propose an information-theoretic framework that quantifies the association between economic activities and the locations they occur in, as well as the association between activities pairs. The framework shows the mathematical and conceptual links between widely used indices that describe the distribution of economic activities, such as Ballassa's index of revealed comparative advantage (Balassa, 1965), co-agglomeration measures (Ellison et al., 2010), and measures of localization of economic activities (Mori et al., 2005). By estimating these quantities in a Bayesian framework, we also obtain measures for the uncertainty of the estimates.

1.2.3 Part 3: Extending the capabilities model

While the previous three chapters all focus on the methodological challenges in the measurement of variety, complexity and relatedness separately, the two chapters that make up the third part of the thesis are theoretical, and integrate the three concepts within an extended capability model (Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009; Inoua, 2016). The theoretical model that underlies much of the reasoning is the 'binomial model' presented in (Hausmann and Hidalgo, 2011). This model assumes that products have random capability requirements, and that countries differ in the number of capabilities they have. Under the assumption that countries then make all products they can given the capabilities they have, the model is able to replicate the distribution of the variety of products produced in different countries, as well as the relation between countries' variety and the average ubiquity of the products they produce. An extension of the model has been used to study the spatial distribution of social activities across cities, and successfully explains urban scaling laws (Gomez-Lievano et al., 2016).

The binomial model explains the distribution of activities across different economies, but does not model the process of development within a single economy. A model of the economic development of a single economy based on the capability framework was proposed by (Inoua, 2016). It assumes that products require a given capability with a fixed probability, and shows that acquiring new capabilities then leads to an exponential increase in the variety of products. This leads to proposing the logarithm of variety as a measure of the number of capabilities. Since the model is based on recombination of capabilities, it is also consistent with the idea of related diversification.

Chapter 6 takes issue with the key assumption that economies will make every possible product given the capabilities they have. Following that assumption, the accumulation of capabilities leads to an ever-increasing variety of products. Although this explains the increase of variety as countries develop, it does not match with the finding that at some point in the development process, the variety of products may decrease again, a phenomenon known as 'the hump' (Imbs and Wacziarg, 2003; Cadot et al., 2011). That is, the model only considers the entry of new products in a country portfolio, but does not cover products that *exit* a portfolio.

The model developed in this chapter poses instead that as countries develop, the least complex activities are dropped from their portfolio as they can no longer be competitive due to increased wages. As a consequence, variety starts decreasing at a certain point in the development process, while the average complexity of the products it makes, continues to increase. This leads to the identification of three stages of development consistent with the 'hump' phenomenon.

Chapter 7 extends the model presented in the previous chapter and argues that countries (or regions and cities for that matter) may well be constrained in the complexity of products that they are able to produce. It argues that putting together many capabilities into a single, complex product not only requires that all capabilities are present in a country, but also that all actors involved are able to effectively coordinate them (be it within or across organizations). Put differently, for a country to be able to produce complex products, it requires that the actors coordinate their activities through networks and institutions present in a country. Assuming that countries face such limits to coordination, policy makers are then faced with a choice to direct their efforts to obtaining additional capabilities (the only policy option in the original capabilities model), or to improve the coordination of capabilities in order to be able to produce more complex products with the capabilities already in place.

In this chapter, I further introduce a policy maker who, in each time step, computes the return to a capability policy (adding one capability to its capability set) and the return to a coordination policy (extending the maximum complexity of products it can make by one capability), and chooses the policy with the highest return. The contribution of this chapter, then, is to show that low-income countries – countries preceding the hump – should balance their policy efforts on acquiring new capabilities and improving their ability to coordinate capabilities, while high-income countries – going through the hump – should focus their policy on improving their ability to coordinate the many capabilities they already have. The model thus suggests that high-income countries should focus their policy on improving their ability to coordinate the (many) capabilities they already have. This latter results questions, albeit on theoretical grounds only, the rising popularity of industrial policy in European countries during the last decade (Mazzucato, 2011).

Extending theoretical models to examine the effects of relaxing some of the strong assumptions underlying the capabilities model provides a way forward in understanding how the key concepts of variety, relatedness and complexity are related, both theoretically and in a policy context. Only by dropping the assumption that countries make every product they could possibly make, one can generate patterns in which variety increases and decreases over time, consistent with the 'hump'-phenomenon. The model explains why variety and wealth get decoupled over time, where the original capabilities framework would predict a monotonic relationship instead (Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009; Inoua, 2016).

1.2.4 Part 4: Concluding remarks

Chapter 8, as the final chapter of the thesis, concludes and reflects on all chapters in the thesis. It will particularly focus on the methodological findings and connections between them. From this reflection, the chapter discusses some of the limitations encountered in the PhD thesis. It also puts forward open questions and challenges to the capabilities framework that remain.

Chapter 2

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 theoretical 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.

[*] This chapter is being prepared for submission as F. Neffke, A. van Dam, C. Bottai, M. Iglesias, S. Orazbayev, R. Hausmann and K. Frenken. The concept of diversity in economic geography: related variety, economic complexity and the product space.

2.1 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 theories 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 and Stiglitz, 1977; Krugman, 1991a). Accordingly, diversity allows suppliers to specialize and customize products and services to the needs of specific customers (Duranton and 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".[1]

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 Neffke et al. (2018) 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, 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 (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.

[1] 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.

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.[2] 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 theoretical 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

[2] Note that this aggregation problem is precisely what the measurement of relatedness aims to overcome: relatedness captures the how distinct different activities are.

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.
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 2.3 we describe the empirical implementation of these concepts. Next, we introduce the data and discuss our empirical exercise in 2.4. 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.

2.2 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.

2.2.1 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.

2.2.2 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 et al.'s (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 a wider range of 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 using 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 product 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[3]: 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.[4] The reason is now clear: because the ECI and PCI are eigenvectors, their sign is undetermined. Second, Hidalgo and Hausmann's (2009) claim that the ECI is a generalized measure of diversity has been questioned by Tacchella et al. (2012) and Kemp-Benedict (2014). The latter showed that the ECI is in fact *orthogonal* to a country's export diversity.[5] 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 2.4.3.

2.2.3 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

[3] In fact, the method of reflections is exactly equivalent to an ordination in Ecology called 'reciprocal averaging' (Hill, 1973b), 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.

[4] 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.

[5] Note that diversity is here measured as the number of products that are exported with revealed comparative advantage.

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 2.3.3, we will see that this density is, in fact, a measure of the *variety of related products*.

2.3 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.

2.3.1 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

$$S(\mathbf{p}_c) = -\sum_{i \in I} p_{ic} \log p_{ic}, \tag{2.1}$$

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:

$$\begin{aligned} S(\mathbf{p}_c) &= -p_{\sigma c} \sum_{\sigma \in \Sigma} \log p_{\sigma c} - \sum_{\sigma \in \Sigma} p_{\sigma c} \sum_{i \in \sigma} \frac{p_{ic}}{p_{\sigma c}} \log \frac{p_{ic}}{p_{\sigma c}} && (2.2) \\ &= \mathrm{UV}_c + \mathrm{RV}_c, && (2.3) \end{aligned}$$

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. Unrelated and related variety thus quantify a city's degree of diversification at two different levels of aggregation: across sectors, and across industries within sectors.

2.3.2 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).[6] 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:

$$P_{ic} = \begin{cases} 1 & \text{if } \frac{E_{ic}/E_{i.}}{E_{.c}/E_{..}} > 1 \\ 0 & \text{elsewhere} \end{cases} \qquad (2.4)$$

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

[6] When applied to export volumes, this quantity is known as revealed comparative advantage (RCA) in the trade literature.

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).

2.3.3 Product Space

Inter-industry relatedness can be measured in a variety of ways (see, for instance, 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:

$$C_{ii'} = \sum_{c \in C} P_{ic} P_{i'c} \tag{2.5}$$

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:[7]

$$\phi_{ii'} = \begin{cases} \frac{C_{ii'}/C_{.i'}}{C_{i.}/C_{..}} & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{2.6}$$

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.[8] Furthermore, we set the proximity of industry i to itself equal to one. Given that the metric defined in eq. (2.6) tends to have a highly skewed distribution, we map $\phi_{ii'}$ onto the interval $[0, 1)$ using:[9]

$$\tilde{\phi}_{ii'} = \frac{\phi_{ii'}}{\phi_{ii'} + 1}. \tag{2.7}$$

[7] 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.

[8] Note that this normalization is essentially the same as in the LQ.

[9] For a detailed justification of this approach, see Neffke et al. (2017). An alternative, information-theory based normalization is proposed in van Dam et al. (2020).

$\tilde{\phi}_{ii'}$ defines a network of related industries, the industry space.[10] 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:

$$D_c^i = \sum_{i' \neq i} \frac{\tilde{\phi}_{ii'}}{\tilde{\phi}_{i.}} P_{i'c} \tag{2.8}$$

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:

$$E_c^i = \sum_{i' \neq i} \frac{\tilde{\phi}_{ii'}}{\tilde{\phi}_{i.}} E_{i'c}. \tag{2.9}$$

Whereas density represents a proximity-weighted *count* of industries in a city – and is therewith essentially a measure of industrial variety (albeit the variety of industries with $LQ > 1$) – mass represents the proximity-weighted *size* of all industries, in terms of employment. Hence, mass does not distinguish between the different industries, but simply considers their total employment weighted by their proximity to the focal industry. All related industries are thus perfect substitutes for one another, whether employment is distributed across many or few (equally related) industries.

In Section 2.4.3, 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:

$$\phi_{cc'} = \begin{cases} \frac{C_{cc'}/C_{.c'}}{C_{c.}/C_{..}} & \text{if } c \neq c' \\ 0 & \text{if } c = c' \end{cases} \tag{2.10}$$

[10] 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 Coscia and Neffke (2017).

Second, we estimate an *occupation space*, $\phi_{oo'}$ by looking at how often two occupations co-occur in the same cities:

$$\phi_{oo'} = \begin{cases} \frac{C_{oo'}/C_{.o'}}{C_{o.}/C_{..}} & \text{if } o \neq o' \\ 0 & \text{if } o = o' \end{cases} \tag{2.11}$$

In eqs (2.10) and (2.11), $C_{cc'}$ 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. (2.7).

Third, we estimate a measure of *cognitive proximity* between industries:

$$\psi_{ii'} = \begin{cases} \frac{C^{\text{occ}}_{ii'}/C^{\text{occ}}_{.i'}}{C^{\text{occ}}_{i.}/C^{\text{occ}}_{..}} & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{2.12}$$

where $C^{occ}_{ii'}$ counts the number of occupations that are simultaneously present in industry i and i', using the definition of "presence" of eq. (2.4). Once again, we map this metric onto the interval $[0, 1)$, using the transformation in eq. (2.7).

Fourth, we calculate the relatedness, or similarity, of two industries' growth patterns as the correlation between the industries' growth rates:[11]

$$\rho_{ii'} = \begin{cases} corr\left(\frac{E_{it+1}}{E_{ict}}, \frac{E_{i't+1}}{E_{i't}}\right) & \text{if } i \neq i' \\ 0 & \text{if } i = i' \end{cases} \tag{2.13}$$

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.

2.3.4 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

[11] We check the significance level of the correlations; if p-value$_{ii'} < 0.05$ then $\rho_{ii'} = 0$.

compare these concepts and their relation to diversity, it will be helpful to explore more carefully what we actually mean by diversity.

Figure 2.1 shows three cities and their employment distribution.[12] 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.

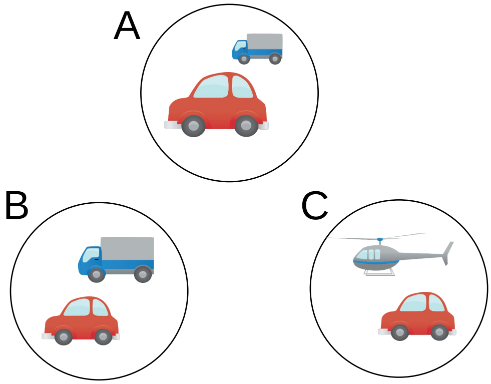


Figure 2.1: Three cities (A, B, and C) with a different employment structure. City A contains two industries of uneven size. City B contains two equally sized industries. City C also contains two equally sized industries, but they are more dissimilar than those in A and B.

Industrial diversity is thus a compound concept that consists of three components (Stirling, 2007):[13]

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] The figure is inspired by Figure 1 in Rafols and Meyer (2010).

[13] Work on incorporating disparity into measures of diversity measures goes back to the 1970s (Rao, 1982). More recent work applies these ideas in scientometrics (Rafols and Meyer, 2010) and economics (van Dam, 2019).

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.

2.3.4.1 Generalized diversity and Hill numbers

We will quantify diversity using the notion of Hill numbers (Hill, 1973a). 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 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.

The entropy in eq. (2.1), used there as an index for diversity, can be converted to a Hill number by taking its exponential (Jost, 2006).[14] 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:

[14] This yields a diversity of $e^{-\sum_i p_i \log(p_i)}$, 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 $e^{-\sum_{i=1}^{N_c} \frac{1}{N_c} \log(\frac{1}{N_c})} = N_c$.

$$D_Z(\mathbf{p}_c) = -e^{\sum_i p_{ic} \log((\mathbf{Zp_c})_{ic})}. \tag{2.14}$$

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 (2.14) simplifies to the standard Hill number:

$$D_I(\mathbf{p}_c) = -e^{\sum_i p_{ic} \log(p_{ic})}.$$

2.3.4.2 Decomposing diversity

The generalized Hill number of eq. (2.14) 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:

$$N_c = \sum_{i \in I} 1(E_{ic} > 0). \tag{2.15}$$

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

$$bal_c = \frac{D_I(\mathbf{p}_c)}{N_c}. \tag{2.16}$$

Likewise, assuming that the generalized Hill numbers is the product of variety, balance and disparity, we obtain disparity as

$$disp_c = \frac{D_Z(\mathbf{p}_c)}{D_I(\mathbf{p}_c)}. \tag{2.17}$$

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:

$$var_c = \frac{N_c}{|I|}. \tag{2.18}$$

2.3.4.3 Relative Hill numbers

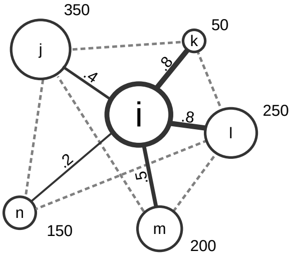
So far, we have discussed the aggregate diversity of an entire local economy. However, in the research on product spaces, the focus is not as much on cities as a whole as on individual industries within a city. Therefore, it is useful to extend the notion of general Hill numbers such that they relate 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 2.2. 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. (2.9).

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 (2.15) to (2.18) 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 2.2: 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).

2.4 Empirical tests

2.4.1 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.[15]

[15] Appendix A provides details on the original data sources and our data cleaning. Appendix B contains an overview of the variables used in this section and their descriptive statistics.

2.4.2 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 and 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 employment growth. To do so, we estimate Ordinary Least Squares (OLS) regression models of the following kind:

$$\log\left(E_{cT}\Big/E_{ct}\right) = \beta_0 + \beta_1 \log E_{ct} + \boldsymbol{X}_{ct}\boldsymbol{\beta} + \varepsilon_{ct}, \tag{2.19}$$

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.[16]

[16] 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 2.1 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 2.1: 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 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 2.2 and 2.3: growth in average wages and end-of-period unemployment levels.[17] Wage growth is positively associated with related variety, but not significantly associated with unrelated variety.[18] Unemployment levels, by

[17] 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.

[18] To capture mean-reversion effects, these analyses also control for the wage level in 1990.

Table 2.2: 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 2.3: 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		

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.'s (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. Neffke et al. (2017) find that these two concepts of relatedness are, in fact, close to uncorrelated.

Using the generalized Hill numbers of section 2.3.4, 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 2.4 runs similar OLS regressions to Table 2.1 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 association is mostly driven by the disparity between, and, to a lesser extent, the balance in the employment distribution across, a city's industries.

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.

Table 2.4: 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

The answer to Frenken et al.'s (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 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[19] of 0.32 for variety in column (6) of Table 2.4 means that the association 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

[19] Given that all variables are expressed in natural logs, coefficients should be interpreted as elasticities.

of related variety and negative effects of unrelated variety in Table 2.1 are thus driven mostly by a cities' variety (i.e. number of industries), and not by the balance of industries in the city.

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 2.4 twice with some slight modifications. First, Table 2.5 measures disparity using the (transformed) metric $\tilde{\psi}_{ii'}$ proposed in eq. (2.12), 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 2.6 we change the dependent variable to the end-of-period unemployment rate in a city and use the growth-similarity based metric $\rho_{ii'}$ of eq. (2.13) to more accurately capture portfolio diversification effects. Note that $\tilde{\psi}_{ii'}$ and $\rho_{ii'}$ define relatedness as continuous variables. To allow for a fair comparison with Frenken et al. (2007), we convert $\tilde{\psi}_{ii'}$ and $\rho_{ii'}$ 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 2.5 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.'s (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 2.6 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 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).

Table 2.5: 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)						
$\ln \mathrm{var}_c$		-0.1684			0.1120	0.4824***	
		(0.1122)			(0.2286)	(0.1320)	
$\ln \mathrm{bal}_c$			0.0492		0.3460		-0.1283
			(0.0829)		(0.2282)		(0.1619)
$\ln \mathrm{disp}_c$				0.0998	0.3734	0.0265	0.1446
				(0.0885)	(0.2705)	(0.0940)	(0.1173)
$\ln \mathrm{var}_c \times \ln \mathrm{disp}_c$						-0.4517***	
						(0.0835)	
$\ln \mathrm{bal}_c \times \ln \mathrm{disp}_c$							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: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

These results show that Frenken et al.'s (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.

2.4.3 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 its firms. How does the ECI compare to the generalized diversity described above as a measure of fundamental diversity? Figure 2.3 shows a scatter plot between the two metrics. ECI and generalized diversity are strongly correlated,

Table 2.6: 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: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

with $\rho = 0.47$. Table 2.7 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 2.3 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 follows the analysis of Figure 2.3 and Table 2.7 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

Figure 2.3: Generalized diversity and ECI.

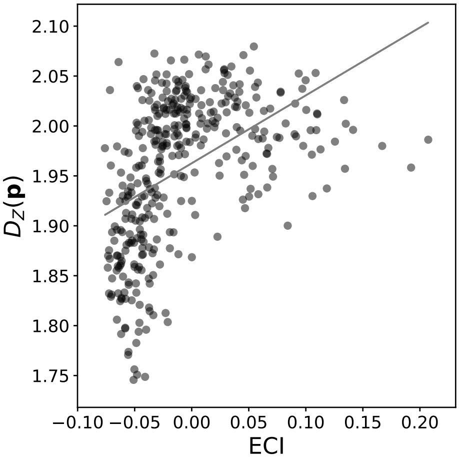


Table 2.7: ECI, generalized diversity and urban wages. OLS regressions with dependent variables in the first row.

	(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

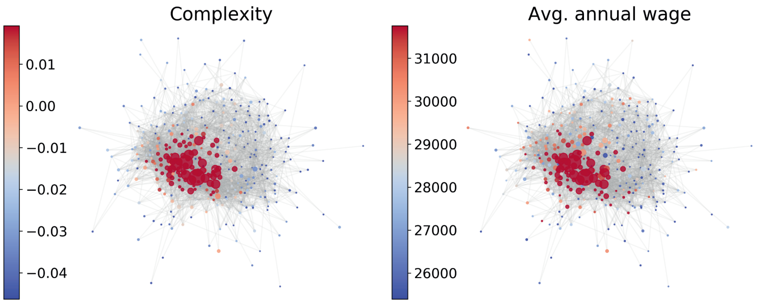
Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

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.

2.4.3.1 City-industry analysis

Figure 2.4 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 2.4: ECI and wages in the city space (industry-city analysis). The sizes of the dots reflect total employment.



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 2.5 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 2.7. 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 2.8 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 2.5: ECI versus average wage in a city (industry-city analysis).

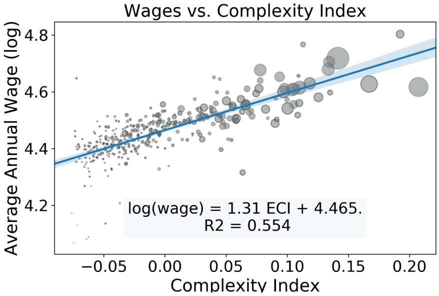


Table 2.8: Top 10 of most complex cities (city-industry analysis)

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 2.6 shows analogous panels to Figure 2.4, 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 2.7. With an R^2 of 0.21, the PCI has weak predictive power for industry-level wages. Moreover, some high-PCI industries in Table 2.9, such as urban transit systems, seem poor examples of complex economic activities.

2.4.3.2 City-occupation analysis

What happens when base the ECI on city-occupation instead of city-industry employment data? Figures 2.8 and 2.9 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 a strong predictor of a city's wage levels: high-ECI cities tend to exhibit high average wages. In contrast, the PCI fails to accurately predict occupational

Figure 2.6: PCI and wages in the industry space (industry-city analysis). The sizes of the dots reflect total employment.

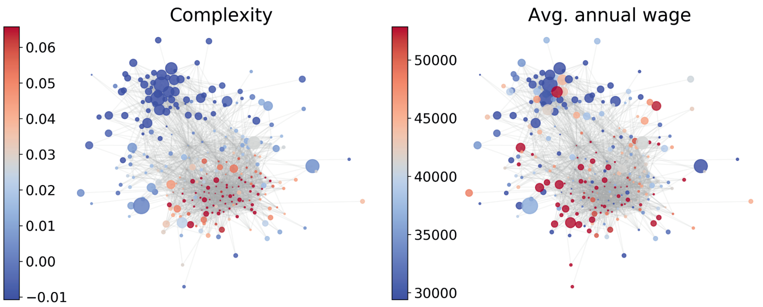


Figure 2.7: PCI versus average wage in an industry (industry-city analysis)

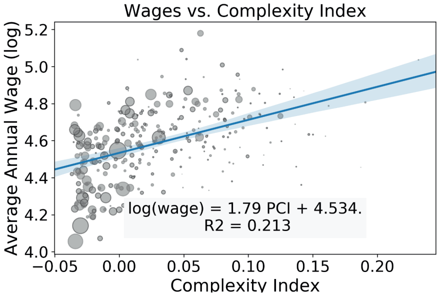


Table 2.9: Top 10 of most complex industries (city-industry analysis). We limit this list to industries that employ at least 25,000 workers in the US.

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

wages. Figures 2.10 and 2.11 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 2.8: ECI and wages in the city space (occupation-city analysis). The sizes of the dots reflect total employment.

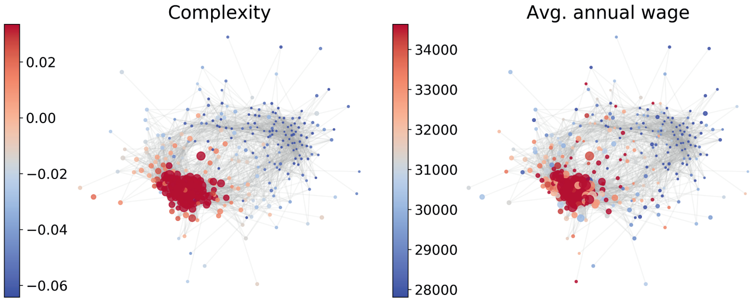


Figure 2.9: ECI versus average wage in a city (occupation-city analysis)

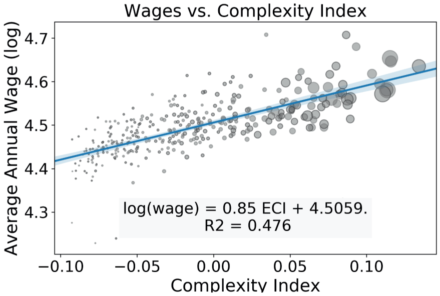


Table 2.10: Top 10 of most complex cities (city-occupation analysis)

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

Figure 2.10: PCI and wages in the occupation space (occupation-city analysis). The sizes of the dots reflect total employment.

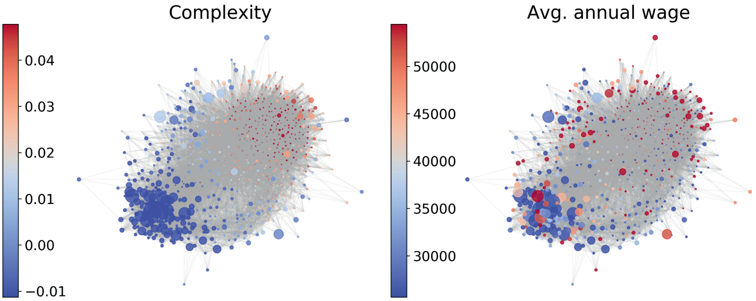


Figure 2.11: PCI versus average wage in an occupation (occupation-city analysis)

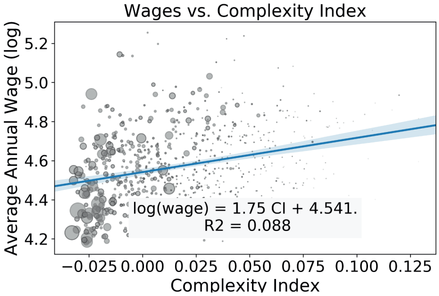


Table 2.11: Top 10 of most complex occupations (occupation-city analysis). We limit this list to occupations with at least 25,000 across all cities.

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

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 compositions, 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 (Mealy et al., 2019; Gomez-Lievano, 2018). 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.[20]

2.4.3.3 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 2.12 shows the results from the industry perspective, Figure 2.13 from the occupation's perspective. Unlike the core-periphery patterns of Figures 2.4 and 2.8, 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.

2.4.4 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

[20] 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 2.12: ECI and wages in the industry space (occupation-industry analysis). The sizes of the dots reflect total employment.

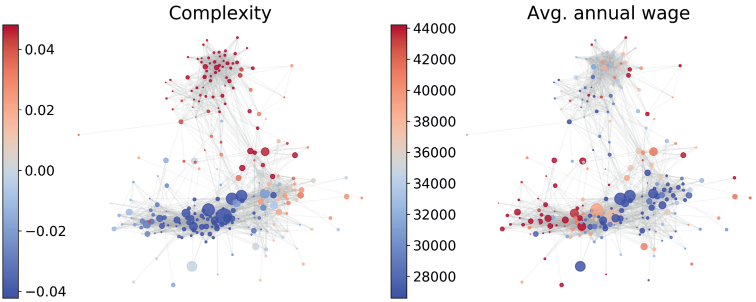
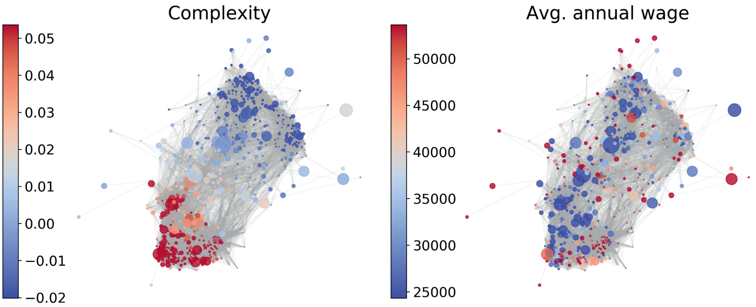


Figure 2.13: PCI and wages in the occupation space (occupation-industry analysis). The sizes of the dots reflect total employment.



this section, we will focus on this growth at the intensive margin and estimate models based on the following regression equation:

$$\log\left(E_{icT}\Big/E_{ict}\right) = \beta_0 + \beta_1 \log E_{ict} + \beta \log X_{ict} + \log E_{it} + \log E_{ct} + \varepsilon_{ict}$$

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 2.12 shows the results. In column (1), apart from industry and city size variables, we only add the mean reversion term and the industry space density, using $\tilde{\phi}_{ii'}$ of (2.7), as explanatory variables. As expected, and in line with the literature's prior consensus, the mean reversion term shows a negative, and the industry space density a positive association with employment growth.

Table 2.12: Product space regression. Dependent variable: Employment growth in city-industry pairs. Regressors use the industry space as defined by $\tilde{\phi}_{ii'}$ of eq. (2.7).

	(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. (2.9)) to the regression model.

The mass of related activity turns out to be more important than its density: 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 2.3.4.3. 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 number 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 industry 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. A definitive conclusion would thus require a more careful analysis and ideally a replication of these findings.

2.5 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 theoretical starting points. 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 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 to decide 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 this 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.

Chapter 3

Diversity and its decomposition into variety, balance and disparity[*]

Abstract

Diversity is a central concept in many fields. Despite its importance, there is no unified methodological framework to measure diversity and its three components of variety, balance and disparity. Current approaches take into account disparity of the types by considering their pairwise similarities. Pairwise similarities between types may not adequately capture total disparity, since they do not take into account in which way pairs are similar. Hence, pairwise similarities do not discriminate between similarities of types in terms of the same feature and similarities in which all pairs share different features. This paper presents an alternative approach which is based on the overlap of features over the whole set of types. This results in a measure of diversity that takes into account the aspects of variety, balance and disparity. Based on this measure, the 'ABC decomposition' is introduced, which provides separate measures for the variety, balance and disparity, allowing them to enter analysis separately. The method is illustrated by analyzing the industrial diversity from 1850 to present while taking into account the overlap in occupations they employ. Finally, the framework is extended to take into account disparity considering multiple features, providing a helpful tool in analysis of high-dimensional data.

[*] This chapter was published as: van Dam, A. Diversity and its decomposition into variety, balance and disparity. *Royal Society Open Science*, 6(7):190452, 2019. doi:10.1098/rsos.190452

3.1 Introduction

Diversity is a central concept in a wide range of scientific fields. In the natural sciences, it is often associated with the functional properties of a system, like the stability of ecosystems (MacArthur, 1955; Tilman et al., 2014). In the social sciences, the concept of diversity is key to theories regarding recombinant innovation (van den Bergh, 2008; Weitzman, 1998), regional development (Frenken et al., 2007), cultural evolution (Foley and Mirazon Lahr, 2011), and the science of science (Rafols and Meyer, 2010; Wang et al., 2015; Zhang et al., 2016). But what exactly is diversity and how can it be measured? Recent frameworks emphasize that diversity consists of three dimensions (Daly et al., 2018; Page, 2011; Purvis and Hector, 2000; Stirling, 2007). First, the *variety* describes the number of different types, species or categories present.[1] The variety is bounded by the total number of types in the classification or taxonomy that is used. Second, the *balance* describes how individuals or elements are distributed across these types. When elements are concentrated in few types the balance is low, whilst a high balance indicates a more even distribution. Last, the *disparity* takes into account to what extent the types considered differ from each other in terms of some given features or characteristics. If the types considered are very similar, they have low disparity. An increase along any of these three dimensions corresponds to an increase in overall diversity. A proper measure of diversity should therefore take into account all three dimensions.

Despite the importance of diversity as a concept, there is no unified methodological framework to measure and analyze the three dimensions of diversity. In the past, the disparity was not even considered by most diversity indices. There have been multiple attempts to incorporate disparity into a measure of diversity by including some measure of the pairwise distances or similarities between the types considered. An example is Rao's quadratic entropy (Rao, 1982), introduced into the social sciences in (Stirling, 2007) where it is known as the Rao-Stirling diversity. It expresses diversity as the average distance between types, weighed by their relative frequencies.

More recently, it has been shown in (Leinster and Cobbold, 2012) that Rao's quadratic entropy follows as a special case from a more general framework that generalizes the

[1] I follow the terminology used in (Stirling, 2007), but these concepts are known by different names in different fields, for example as 'richness', 'evenness' and 'similarity' in ecology.

so-called Hill numbers (Hill, 1973a; Jost, 2007) to include pairwise similarities between types. A similar approach was taken in (Chiu et al., 2014), who generalize Hill numbers to include phylogenetic or functional similarities. Both approaches compute diversity in terms of 'effective numbers' (Jost, 2007), based on a distribution of types and a given matrix that contains the pairwise similarities between those types. Other approaches to quantify diversity while taking into account disparity have been to compose a measure of diversity by separately measuring variety, balance, and disparity, and combining them into a single index of diversity (Leydesdorff et al., 2019). A more data-driven approach is taken in (Wang et al., 2015), who apply factor analysis to a range of different diversity indices to infer three variables that correspond to variety, balance and disparity.

What all approaches described above have in common is that disparity is quantified using pairwise similarities. The use of pairwise similarities however may lead to both practical and conceptual problems. One practical problem is that there are many different ways in which pairwise similarities can be inferred from given data (van Eck and Waltman, 2009; Yildirim and Coscia, 2014), so any diversity measure based on pairwise similarities is subject to an ad-hoc choice of a particular similarity measure. In addition, it is unclear how heavily such an index should weigh disparity versus variety and balance (Stirling, 2007).

More importantly, considering only pairwise similarities between types may not adequately capture total disparity, since pairwise similarities do not take into account *in which way* pairs are similar. Pairwise similarities are typically inferred by using some measure of how many features two types share from a pre-defined set of features. Types in a collection may then all be similar because each pair shares *the same feature*, or because each pair shares a *different* feature. Both situations could have different diversities, but have an identical similarity structure.

This paper presents a framework to measure diversity that does not rely on pairwise similarities between types. Instead, disparity is taken into account by looking at the overlap of features between types over the whole set. This is done by drawing on the concepts of alpha, beta and gamma diversity from ecology (Whittaker, 1972) and the corresponding decomposition of diversity as introduced in (Jost, 2007), which is based on Hill numbers (Hill, 1973a). The result is a measure of diversity that incorporates variety, balance and disparity simultaneously, and has a natural interpretation as the 'number of compositional units' (Tuomisto, 2010).

Building on this measure, I introduce the 'ABC decomposition' that decomposes diversity into separate measures of variety, balance and disparity. This enables the study of the distinct role each of these dimensions has in different systems[2]. The proposed framework is closely related to information-theoretic measures of uncertainty, and the use of multivariate information theory shows how the measure can be extended to take into account disparity along multiple dimensions. This leads to two results regarding the diversity of types given multiple feature sets, depending on the dependence structure of the variables involved. First, diversity considering multiple feature sets becomes multiplicative when different feature sets are independent. Second, additional feature sets may be neglected in measuring diversity when one feature set is conditionally independent of the types, given another feature set.

I proceed as follows. Section 3.2 starts with an example of a situation where using pairwise similarities fails to quantify disparity correctly. Subsequently the concepts of beta diversity are introduced along with the main result, namely a measure of diversity that takes into disparity as the overlap over a set of features. Section 3.3 then introduces a decomposition of diversity into separate measures of variety, balance and disparity. As an illustration, I apply the proposed measures to historical data in order to characterize the change in diversity of industries in the US, taking into account disparity in terms of the occupations that industries employ. Section 3.4 shows how the framework can be extended to take into account multiple sets of features. I conclude with a brief discussion of the results.

3.2 Decomposing diversity

3.2.1 An example

Consider a region in which certain economic activities take place in the form of industries. These industries can be thought of to consist of a certain set of inputs or features (Hidalgo and Hausmann, 2009). We will represent these features with letters in a set S, and the industries as words in set S'. For example, one might think of

[2] For example, the separate effects of variety, balance and disparity on scientific impact was studied in (Wang et al., 2015).

the letters as occupations required by a firm to engage in a particular industry, represented as a word. The diversity of words is determined by the number of different words (variety), their relative frequency (balance) and their similarity in terms of the letters they consist of (disparity). Adding words with similar composition of letters does not affect the diversity much, whereas adding words consisting of many new letters may greatly increase diversity.

The composition of words and letters in a region can be represented as a bipartite network as in Figure 3.1. In the three cases shown, the variety equals 3 (there are three unique words) and the balance is maximal (the relative frequency $p_i = \frac{1}{3}$ for each word). The disparity of words is different for each of the three cases, and is determined by how the words are composed from the letters.

A common approach to quantify diversity whilst taking into account disparity is by considering the pairwise similarity between types (Chiu et al., 2014; Leinster and Cobbold, 2012; Rao, 1982; Stirling, 2007). Computing the pairwise similarities can be interpreted as 'projecting' the bipartite network onto a weighted network in which the nodes are the types, and the weighted edges represent the pairwise similarities in terms of the overlap in features (see Figure 3.1). Here we consider the Jaccard similarity s_{ij}, which gives the similarity as the number of shared features divided by the total number of features used by both types.

An example of such a measure is the Rao-Stirling diversity, which is computed as[3] (Rao, 1982; Stirling, 2007)

$$\Delta = \sum_{ij} (1 - s_{ij}) p_i p_j .$$

This measure incorporates the variety by summing over all types, and the balance by taking into account the relative frequencies p_i. Disparity is then taken into account by weighing every pair of types by the distance between the types. This way, pairs with low similarity contribute more to the diversity than pairs with high similarity.

In the first case in Figure 3.1 the disparity is maximal (there is no overlap of letters between words), and the Rao-Stirling diversity reduces to $\Delta = \sum_{ij} \frac{1}{3}\frac{1}{3} = \frac{1}{3}$ since $s_{ij} = 0$ for all pairs. For the other two cases, the Jaccard similarities are given by

[3] $1 - s_{ij}$ gives the 'Jaccard distance' or dissimilarity between a pair of words.

$s_{ij} = \frac{1}{5}$ for all pairs. Since the pairwise similarities are identical in both cases, any diversity measure based on these pairwise similarities will give the same diversity for both cases. Indeed, computation of Rao-Stirling diversity shows a diversity of $\Delta = \sum_{ij}(1 - \frac{1}{5})\frac{1}{3}\frac{1}{3} = \frac{4}{15}$ for both cases.

However, note that the underlying network structure in the latter two cases in Figure 3.1 is different. While both have an identical variety and balance of words, the distribution of letters is different. In the middle case in Figure 3.1, all words share the same letter so that every word pair is similar *in the same way.* In the latter case, every word pair shares a *different* letter, so they are similar in different ways. This leads to a different distribution of features in both cases. Out of two collections of types with the same variety and balance, the collection that represents a higher diversity of features is arguably more diverse when taking into account the disparity between types. Hence, we expect the case with a higher diversity of features in Figure 3.1 to have a higher diversity. Since the projected networks for the middle and last case are identical however, such a difference cannot captured by diversity measures that are based on those pairwise similarities. This paper proposes a measure that takes into account the overlap in features over the whole collection of types, as opposed to pairwise similarities, leading to a measure that reflects the difference in composition between the two cases.

3.2.2 Hill numbers

In our measurement of diversity, we build on the framework of Hill numbers, which provides a unifying mathematical framework for the measurement of diversity when disparity is not taken into account (Hill, 1973a; Jost, 2006). Hill numbers define diversity as the inverse of a generalized weighted average of the relative frequencies of the types. In this definition, a collection is diverse if the types are on average rare, i.e. the average share of the types is low.

Hill numbers satisfy a number of axiomatic requirements for a measure of diversity, including symmetry, continuity, and monotonicity in the number of species (Daly et al., 2018). Another key property is the *replication principle*, which states that pooling together two collections that do not share any types but have equal distributions, should give a new collection with double the diversity of the original collections (Hill, 1973a).

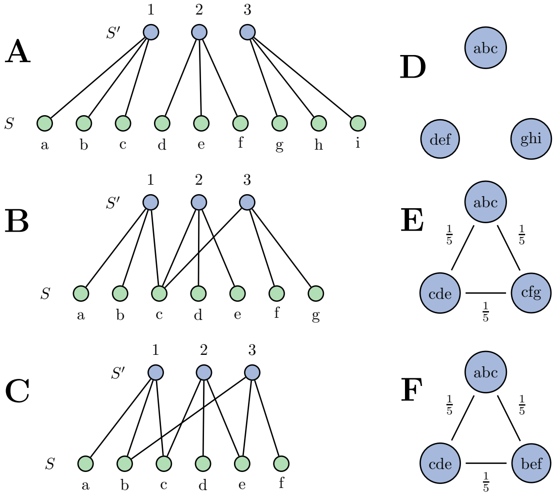


Figure 3.1: **A**, **B** and **C** show the bipartite networks as discussed in the main text. One can think of the blue nodes representing three industries (words), and the green nodes representing nine occupations (letters) that characterize the industries. **D**, **E** and **F** show the corresponding projected industry networks, in which the edge weights are given by the Jaccard similarity between industries. In **A** and **D** there is no overlap in occupations, and pairwise similarities between industries are 0, as shown by the absence of edges in **D**. The Rao-Stirling diversity is given by $\Delta = \frac{1}{3}$. **B** and **E** show a situation where the industries use a total of seven occupations, and the similarity between each industry equals $s_{ij} = \frac{1}{5}$. **C** and **F** show a situation where only six occupations are present, and where all pairwise similarities are again $\frac{1}{5}$. Although **B** has a different distribution of occupations than **C**, their projections **E** and **F** are identical, and therefore any diversity measure based on those pairwise similarities will assign identical diversities to both cases. The Rao-Stirling diversity is given by $\Delta = \frac{4}{15}$ for both cases.

Hill numbers give rise to a parametric family of diversity measures, in which a parameter q determines how heavily one weighs the rarity of types in a measure of diversity. For $q = 1$, rare and common species are weighed equally heavy and the Hill number equals the exponential of the Shannon entropy:

$$D(S) = e^{H(X)} = e^{-\sum_i p_i \log p_i}. \tag{3.1}$$

Here, S is a collection of elements with types i and relative frequencies p_i, and X is a random variable that represents the type i of a randomly drawn element from S.

A more elaborate discussion on Hill numbers and their relation to Shannon entropy can be found in the supplementary material A. It was shown in (Jost, 2007) that the Hill number with $q = 1$, i.e. the exponential of the Shannon entropy, is the unique measure that satisfies all axiomatic requirements and allows for a decomposition of independent within- and between components in the presence of groups.

Hill numbers have also been referred to as the 'true diversity' as opposed to an index, as many existing diversity indices in ecology and economics that were originally introduced based on heuristics have been shown to be a transformation of a Hill number (Jost, 2006). In particular, equation (3.1) shows how the Shannon entropy, a popular *index* of diversity but which is actually a measure of uncertainty (it has units in 'bits' or 'nats'), can be transformed into a measure of diversity (Jost, 2006).

Furthermore, the Hill number of a collection has a clear interpretation as the 'effective number' of types, meaning that the Hill number of a collection S can be interpreted as the number of types that would be present in a virtual collection $\tilde{S}$ that has maximal balance (i.e. a uniform distribution over types) and has the same diversity as S. In particular, for a uniform distribution, i.e. $p_i = \frac{1}{n}$ for all i, we have $D(S) = n$ so that the diversity equals the number of types. For any other distribution over types, the Hill number represents the equivalent number of types in a maximally balanced collection.

3.2.3 Alpha and beta diversity and the number of compositional units

The Hill number $D(S)$ quantifies both the variety and balance of types but not their disparity, and thus implicitly assumes that all types i are maximally disparate. Here, we aim to extend this framework to include the overlap of features between types. To this end, we build on the concepts of alpha, beta and gamma diversity from ecology.

Hill numbers provide a decomposition of diversity into its α and β components (Jost, 2007), which are used in ecology to describe the average within-sample diversity and the between-sample diversity, respectively (Whittaker, 1972). For example, consider a forest in which the distribution of species is sampled in different plots. The diversity of the collection of species that consists of all plots pooled together is called the total diversity or γ-diversity. The α-diversity represents the average diversity *within* each plot. The β-diversity represents the diversity *between* each plot, reflecting the diversity that is the result of the differences in species composition between each plot.

The γ-diversity of the forest can be multiplicatively decomposed into independent α and β components, i.e. $D_\gamma = D_\alpha D_\beta$ (Jost, 2007). In a homogeneous forest, where all plots have approximately the same species composition, the average within-plot diversity D_α is close to the diversity of all plots pooled together, D_γ. Hence, the between-plot diversity D_β will be close to 1. In a heterogeneous forest on the other hand, every plot has a very different species composition and contains only a small part of the total diversity, so D_α is much smaller than D_γ, leading to a higher value of D_β. D_β reflects the number of different plots needed, each with diversity D_α, to obtain a pooled diversity of D_γ. The maximum value of D_β is given by D_γ, corresponding to the case where every plot consists of a unique species ($D_\alpha = 1$). The β-diversity is thus bound from below by 1 and from above by D_γ.

Note that the situation described above corresponds with the example in Figure 3.1, in which the plots represent the types of interest (words) and the species represent some characterizing features of those types (letters). In this setting, the γ-diversity gives the total diversity of features, and the α-diversity the average diversity of features within a type. The β-diversity then represents the 'between-type' diversity based on the heterogeneity of the composition of types.

The values of each diversity for the example in Figure 3.1 are given in Table 3.1. Since for every case each of the three words contains three letters, the α-diversity is three for all cases. The diversity of letters as measured by the Hill number is different for each case however, as shown by the γ-diversity. The diversity of letters is lowest for the last example. This is reflected by the β-diversity, which gives a lower number of compositional units for the case with a lower diversity of features.

The β-diversity gives the number of types with average diversity D_α that are needed to obtain a total diversity of features D_γ when there would be no overlap of features between the types. It is obtained by dividing the total diversity of features, as given by the Hill number of order 1, by the average diversity of features within a type, so that

$$D_\beta(S') = \frac{D\gamma(S)}{D_\alpha(S)}. \tag{3.2}$$

It can be interpreted as a measure of the ’number of compositional units’, giving the effective number of types that would be present when the types do not share any features and would be equally abundant (Tuomisto, 2010). Framing beta diversity in terms of types and features provides a measure of diversity of types that takes into account variety, balance *and* disparity as given by the overlap of features between types. As a measure of diversity, the number of compositional units satisfies all of the mathematical properties that were proposed by (Leinster and Cobbold, 2012) that reflect a ’basic scientific intuition’ about diversity. The nine properties are divided into three categories: partitioning properties, elementary properties, and similarity properties (see the supplementary material B). How to compute the number of compositional units from data will be discussed using an empirical example in the following section.

	α-diversity $D_\alpha(S')$	β-diversity $D_\beta(S')$	γ-diversity $D_\gamma(S)$	eff. number $D(S')$
A	3	3	9	3
B	3	2.08	6.24	3
C	3	1.89	5.67	3

Table 3.1: Values of the α, β and γ diversities for the three examples depicted in Figure 3.1. The average diversity of occupations within an industry, $D_\alpha(S')$ is equal in all three examples, as every industry employs three different occupations with equal weight and all industries have an equal share. The total diversity of features $D_\gamma(S)$ is given by the effective number of occupations in all industries pooled together, and differs in each case. This also leads to different values of $D_\beta(S')$. For completeness, the effective number of industries $D(S')$, representing the diversity of industries when one assumes that they are totally disparate, is also included.

3.2.4 Measuring diversity of industries

Here, the general application of the proposed diversity measure is presented using an empirical example. The aim is to quantify industrial diversity in the US, where the distinguishing features of industries are considered to be the different occupations they employ. US census data was extracted from IPUMS-USA (Ruggles et al., 2018), providing a 1% sample[4] of total population in the US for every decade from 1850 to 2010. The data contains for every person their occupation $i \in S$ and industry $j \in S'$. The used classifications consist of 269 occupation types and 147 industry types.

[4] For 1980 and 1990 a 5% sample was given. Note that the analysis presented is for illustrative purposes, so further data cleaning and consistency issues are not considered here.

The data is interpreted as a weighted bipartite network as in Figure 3.1, with nodes i in the occupation layer S and nodes j in the industry layer S'. The edge weight between nodes i and j is given by the number of people q_{ij} working in occupation i and industry j. The strength of node i is given by $q_i = \sum_j q_{ij}$ and represents the total employment in occupation i, and similarly q_j denotes total employment in industry j. Normalizing the quantities q_i, q_j and q_{ij} by the total number of people $Q = \sum_{ij} q_{ij}$ gives the relative frequencies $p_{ij} = \frac{q_{ij}}{Q}$, $p_i = \frac{q_i}{Q}$ and $p_j = \frac{q_j}{Q}$ respectively. Each of the relative frequencies may in turn be interpreted as the probability distribution of a random variable that represents the occupation or industry type of a randomly sampled person, i.e. $p_i = P(X = i)$, $p_j = P(Y = j)$ and $p_{ij} = P(X = i, Y = j)$.

Using Hill numbers, the effective number of industries and occupations can be expressed as $D(S') = e^{H(Y)}$ and $D(S) = e^{H(X)}$, respectively. To obtain the effective number of occupations *within* an industry, consider the relative frequencies $p_{i|j} = \frac{q_{ij}}{q_j}$ of occupation i in industry j. The occupational diversity of an industry j is then given by

$$D(S_j) = e^{H(X|j)} = e^{-\sum_i p_{i|j} \log(p_{i|j})}.$$

The *average* within-industry diversity is then given by (Jost, 2007)

$$D_\alpha(S) = e^{H(X|Y)} = e^{-\sum_j p_j \sum_i p_{i|j} \log(p_{i|j})},$$

where $H(X|Y)$ is the conditional entropy of X given Y. Finally, the *within* industry diversity D_β follows from multiplicatively decomposing the total occupational diversity $D_\gamma(S)$ into its α and β components, leading to equation 3.2 (Jost, 2007). $D_\beta(S')$ can be interpreted as the effective number of industries, *discounted for the overlap in their occupational distributions.* Its units correspond to the number of industries that would be present in the case of equally-distributed, non-overlapping industries, and where the D_α and D_γ are the same.

Figure 3.2 shows the time evolution of variety, the effective number of industries $D(S')$ (taking into account variety and balance) and the number of compositional units $D_\beta(S')$ (which takes into account balance, variety and disparity) of industries in the US. The variety of industries, i.e. the number of different industry types that have at least one employee, is slightly increasing and then decreasing after 1950, with values ranging between 120 and 140 throughout the whole period. The sudden dip in variety in 1940 is unexplained, and most likely due to data inconsistencies.

The effective number of industries $D(S')$ starts at a much lower level of an effective number of around 10 industries, showing that total employment in the 120 industry types is initially heavily concentrated in a few industries. It shows a more pronounced hump-shaped pattern, with a period of diversification in $1850 - 1960$ in which the effective number of industries grows to around 80 industries as employment becomes more equally spread across industries, followed by a period of re-concentration after 1960. These findings are in line with work in economics that shows that countries first go through a 'diversification phase' as they develop, and then start specializing again at a later point in the development process (Imbs and Wacziarg, 2003).

In contrast, the number of compositional units $D_\beta(S')$ shows a pattern of steady decline since 1900, with values ranging from 10 to 4 compositional units. This means that although employment becomes more equally spread over an increasing number of industries during the diversification phase, industries become increasingly similar in terms of the occupations they employ, leading to a decreasing disparity between industries. In other words, using the notion of related variety (Frenken et al., 2007), these results suggest that variety has become more related over time. This results in a decreasing number of compositional units.

It becomes clear that taking into account the different dimensions of diversity can lead to very different representations of the same data. Considering only the variety for example may lead to an overestimation of diversity, as a the distribution over industries may be concentrated in only a few industries. Furthermore, when industries with similar occupational distributions are considered to be the same, the effective number of industries is an overestimation of the diversity, as some industries may be almost identical in terms of the occupations they employ.

The effect of taking into account disparity of course depends on which features are considered. That is, there might have been an increase in diversity when some other feature of industries was taken into account instead of the occupations they employ. Hence, application of these measures must be driven by the research question at hand. The interesting aspect of these measures however is that each dimension of diversity may show a distinct dynamic, which is not visible when considering diversity as a whole. As Figure 3.2 shows, the number of compositional units may decrease while both the balance and variety increase. Therefore, we consider a decomposition of the number of compositional units into separate measures of variety, balance and disparity in the next section.

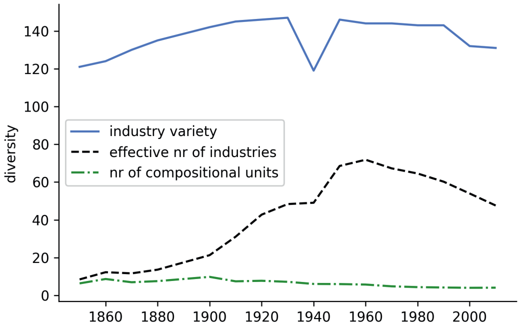


Figure 3.2: Variety, effective number and effective number of compositional units for industries. The variety of industries is approximately constant over time. The effective number of industries takes into account variety and balance and shows a hump-shaped pattern, where initially the distribution of people over industries becomes more equal reaching a diversity of 80 effective industries in 1960, where a re-concentration starts to take place. The number of compositional units takes into account the occupational overlap between industries. In 1850 the industrial diversity was equivalent to approximately ten non-overlapping industries, which declined to approximately four compositional units in 2000.

3.3 The 'ABC' decomposition

In order investigate the role of variety, balance, and disparity in practice, separate measures are required for each. To this end, I introduce the 'ABC decomposition', which decomposes diversity into its separate dimensions. Since $D_\beta(S')$ is a measure of diversity incorporating all three dimensions, a multiplicative decomposition into the variety (A), balance (B) and disparity (C) may be obtained as:

$$D_\beta(S') = D_A(S') \cdot D_B(S') \cdot D_C(S'). \tag{3.3}$$

The variety D_A is given by a simple count of the number of types in S', or equivalently by the Hill number of order $q = 0$ (see supplementary material A). The balance D_B

is computed by dividing the effective number of types in S' (which takes into account both balance and variety) by the variety, leading to (Hill, 1973a)

$$D_B(S') = \frac{D(S')}{D_A(S')} = \frac{D(S')}{n}.$$

$D_B(S')$ measures the evenness in the distribution of relative frequencies of the types. It takes values in $(\frac{1}{n}, 1)$, with a maximum of 1 that is attained when all relative frequencies are equal, i.e. $p_i = \frac{1}{n}$ for all types i in S. The minimum $\frac{1}{n}$ is achieved when the proportion of all but one type is vanishingly small.

Note that the obtained components of variety and balance are not independent, since a higher variety allows for a lower balance. For example, if nearly all employment is concentrated in one out of two industries, this gives a higher balance than a situation in which nearly all employment is concentrated in one out of 100 industries. Hence $D_B(S')$ is an 'absolute' measure of balance, as opposed to a 'relative' measure that characterizes the balance *given* a certain variety (Jost, 2010). An in-depth study concerning measures of balance and their (in)dependence with variety is given in (Jost, 2010).

Since $D_\beta(S')$, $D_A(S')$ and $D_B(S')$ are then determined, the disparity $D_C(S')$ can be obtained by dividing the number of compositional units $D_\beta(S')$ (which takes into account all three dimensions) by the effective number as

$$D_C(S') = \frac{D_\beta(S')}{D_A(S')D_B(S')} = \frac{D_\beta(S')}{D(S')} = e^{-H(Y|X)}.$$

$D_C(S')$ can be considered as the number of compositional units normalized for variety and balance, leaving a measure of disparity. It takes values in $(0, 1)$, attaining the maximum value when none of the types have overlap in their features. The minimum is attained when all types have identical features.

It is easily verified that (3.3) holds with these definitions of $D_A(S')$, $D_B(S')$ and $D_C(S')$. The decomposition allows to study the three dimensions of diversity separately. The diversity $D_\beta(S')$ can be seen as the variety $D_A(S')$, corrected by the factors $D_B(S')$ and $D_C(S')$ which are both between 0 and 1. The variety can in turn be normalized by the total number of types in the classification considered to make it have values between 0 and 1 so that it is comparable to the balance and the disparity as a fraction of its maximum value.

Applying the ABC decomposition to the example in Figure 3.1 leads to the results given in Table 3.2. The results show, as expected, a decreasing disparity as the overlap between words increases. The decrease in disparity as the total number of letters decreases is accurately captured by the proposed measure.

	eff. number $D(S')$	(β-)diversity $D_\beta(S')$	variety $D_A(S')$	balance $D_B(S')$	disparity $D_C(S')$
A	3	3	3	1	1
B	3	2.08	3	1	0.69
C	3	1.89	3	1	0.63

Table 3.2: Values for the effective number of industries, the number of compositional units and the variety, balance and disparity as given by the ABC decomposition for the three examples depicted in Figure 3.1. Variety and balance are equal in all three examples, as every industry employs three different occupations with equal weight. The disparity differs in all three cases, and is maximal for **A**, in which there is no overlap of occupations between industries. For **B** and **C**, the measures show a lower disparity and hence a lower diversity for **C**, in which the industries are composed of less occupations.

Figure 3.3 shows the *ABC* decomposition applied to the empirical example of industries in the US. It contains the same information as Figure 3.2, but shows the dynamics of variety, balance and disparity separately. Since all dimensions of diversity may move independently from each other, the ABC decomposition can help in analyzing the specific role of each dimension in different systems.

3.4 Multivariate extensions

From the framework of Hill numbers, an interesting relation follows between diversity and the information-theoretic notion of uncertainty. In particular, the beta diversity is given by

$$D_\beta(S') = \frac{D_\gamma(S)}{D_\alpha(S)} = e^{H(X)-H(X|Y)} = e^{MI(X,Y)}, \tag{3.4}$$

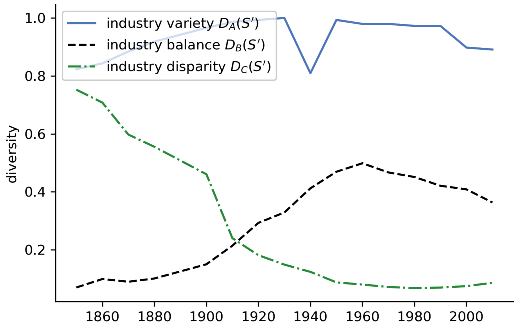


Figure 3.3: Variety, balance and disparity for industries given their occupational distribution. Variety is normalized by the total number of possible industries in the classification, which equals 147. The variety of industries remains approximately constant over the whole period, and the balance shows diversification of industries up to 1960, followed by a period of re-concentration. The disparity shows a decline over nearly the whole period, with a slight increase since 1980.

where $MI(X, Y)$ denotes the mutual information between the random variables X and Y[5]. Taking the exponential of the mutual information between random variables X and Y translates it into a measure of diversity of the corresponding collection S', discounted for the overlap in features given by S. Furthermore, the additive decomposition of information-theoretic measures corresponds to a multiplicative decomposition of diversities.

3.4.1 Taking into account multiple feature sets

Here, we generalize the diversity D_β to take into account multiple feature sets by exploiting the additive relations between multivariate information measures. For instance, returning to the example described in Figure 3.1, we could add a feature set to each word by color coding each letter, so that each word can be distinguished along two dimensions: its colors and its letters. Describing letters with random variable X, colors with random variable Y and words with random variable Z, one can consider

[5] The mutual information is a measure of dependence between two random variables X and Y, given by $MI(XY) = \sum_{ij} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right)$. It is nonnegative and symmetric, and can be interpreted as the average reduction in uncertainty about the outcome of one random variable, given knowledge about the outcome of the other.

the joint probabilities $p_{ijk} = P(X = i, Y = j, Z = k)$ that a randomly sampled element is letter i, has color j and is used in word k. In a network representation, the joint probabilities p_{ijk} can be considered as the relative frequencies of hyperlinks between nodes i, j and k in a hypergraph that connects colors, letters and words to each other.

Following equation (3.4), the diversity of words given their overlap in letters and colors is then given by

$$D_{\beta}^{XY}(S') = e^{H(XY)-H(XY|Z)} = e^{MI(XY,Z)},$$

where $H(XY) = -\sum_{ij} p_{ij}\log(p_{ij})$ is the Shannon entropy of the joint distribution p_{ij}, and the superscript in $D^{XY}(S')$ is used to indicate that diversity is taken with respect to the overlap in feature *pairs* given by XY. Hence, every color-letter pair is interpreted as a distinct feature of a word.

The effect of taking into account an additional feature set on the diversity depends on the information contained by these features. In the current example, taking into account color as a second feature set will not affect diversity much if colors and letters are highly correlated. On the other hand, diversity of words may be very high when colors and letters are independent of each other, thus capturing complementary information. Words that consist of the same letters may contain very different colors, and still add to the overall diversity. Mathematically, this can be seen by re-writing the beta diversity as (see supplementary material C)

$$D_{\beta}^{XY}(S') = e^{MI(X,Z)+MI(Y,Z)-MI(X,Y)+MI(XY|Z)},$$

from which it is clear that diversity decreases when the dependence between features, given by $MI(X,Y)$, increases. In the extreme case that letters i and colors j are independent, we have $MI(X,Y) = 0$ so that (see supplementary material C)

$$D_{\beta}^{XY}(S') = e^{MI(X,Z)+MI(Y,Z)} = D_{\beta}^{X}(S')D_{\beta}^{Y}(S'),$$

where $D_{\beta}^{X}(S')$ and $D_{\beta}^{Y}(S')$ denote the diversity of words with respect to the features described by random variables X and Y, respectively. Thus, when feature sets are independent the diversity that takes into account feature pairs can be obtained through multiplication of the diversities that take into account each of the feature sets separately.

Results like this may be useful and relevant when estimating diversity from high-dimensional datasets containing multiple feature sets. For example, one could consider the diversity of industries by not only taking into account the occupation but also the educational profile of people employed by an industry as a distinguishing characteristic. If educational profiles and occupations are uncorrelated, this diversity equals the product of the diversities that take into account occupations and educational profiles separately.

3.4.2 Aggregation

Another interesting interpretation is to consider the types in S' to be an aggregation of the features S. In this setting the words are thus considered to be a specific way of aggregating letters. These words can in turn be further aggregated into sentences, effectively 'adding a layer' on top of the bipartite network depicted in Figure 3.1. In such a setting, it follows from the current framework that the diversity of sentences depends only on the composition of words, and not on the composition of letters. The key assumption is that the two steps of aggregation are independent of each other, i.e. how words are aggregated into sentences is independent of how letter are aggregated into words.

In the situation described above, the links between letters and words are given by the joint distribution p_{ij} and the links between words and sentences by p_{jk}, where k is the index for sentences, i is the index for letters and j the index for words. The probability of a letter-word-sentence triplet is then given by $p_{ijk} = p_{ij}p_{k|j}$. In other words, sentences and letters are conditionally independent on knowing the word, and their joint probability is given by

$$p_{ik} = \sum_j p_{i|j}p_{k|j}p_j,$$

which implies that $MI(X, Z|Y) = 0$. The diversity of sentences given the overlap in words and letters is then equal to the diversity when considering the overlap in words only (see supplementary material D):

$$D^{Z}_{\beta}(S') = e^{MI(Z,XY)} = e^{MI(Z,Y)}. \tag{3.5}$$

Hence, when the composition of types described by Z in terms of features described by Y is independent of how the features Y themselves are composed of other features X, the features X are irrelevant for considering diversity of Z.

As an example, consider the diversity of industries in Figure 3.2, where occupations are taken to be features of the industries. Hence, we can consider the distribution over industries as a particular way of aggregating over occupations. Similarly, occupations can be considered to be a collection of particular skills and tasks, and hence industries are, indirectly, also an aggregation of those skills and tasks. Equation (3.5) shows that as long as the composition of occupations in terms of skills and tasks is independent of the industry they are employed in, the diversity of industries is fully captured by occupations alone, and there is no need to consider skills and tasks.

3.5 Discussion

This paper presented a framework to measure diversity while taking into account variety, balance and disparity of types. The framework builds on Hill numbers (Hill, 1973a; Jost, 2006) and the corresponding decomposition of diversity into independent alpha and beta components (Jost, 2007). It has a clear interpretation in terms of the 'number of compositional units' (Tuomisto, 2010), and satisfies the a set of basic intuitive properties for diversity measures as formulated in (Leinster and Cobbold, 2012). Contrary to current approaches (Daly et al., 2018; Leinster and Cobbold, 2012; Chiu et al., 2014), the measure does not rely on pairwise similarities but instead takes into account overlap of features between types over the whole set.

I have also proposed the 'ABC' decomposition of diversity that provides a way to capture variety, balance and disparity in separate measures. Such measures may help disentangle the distinct dynamics and functional properties that different dimensions of diversity may have in different systems. In the context of economics for example, economic development is often associated with an increase of the diversity of economic activities (Hidalgo and Hausmann, 2009; Saviotti, 1996; Imbs and Wacziarg, 2003). It is however an open question what the role of the individual components of diversity is in the process of economic development - as the preliminary results in the current paper show, economic development may actually go hand in hand with decreasing disparity, when disparity is measured in terms of industries and the occupations they employ.

The proposed framework reveals close connections between measuring diversity and information-theoretic measures of uncertainty. The simple additive properties of information-theoretic measures correspond to multiplicative properties of diversity measures, and enables derivation of special properties when considering multiple feature sets. These properties may provide useful tools in the analysis of high-dimensional datasets. Furthermore, the diversity measures presented here can be interpreted as centrality measures on bipartite networks, or hypergraphs in the multivariate case. In this sense, the beta diversity captures structural properties of the network. Application of these measure may also be extended to directed networks (e.g. input-output tables in economics (Leontief, 1966)), as any directed network may be interpreted as a bipartite network.

The current paper also leaves some open challenges that have not been addressed. First, there is the issue of the estimation of the proposed diversity measures from data, and finding a measure of precision of this estimate.

A promising way forward is to use a Bayesian framework as in (Wolpert and Wolf, 1995; Hutter and Zaffalon, 2005). These works provide closed-form solutions for the moments of the posterior distribution of information-theoretic quantities like the Shannon entropy en mutual information, given the data and a prior distribution for the probabilities p_{ij}. Such an approach should extend in a straightforward way to the exponentials of those quantities (which are our measures of diversity). In this way, an estimate can be obtained along with a 'Bayesian error bar' that shows the precision of that estimate given the data and a prior distribution, for example showing a lower precision estimate when the number of observations is low (Wolpert and Wolf, 1995). A major challenge in applying such an approach for the estimation of diversity is to find a suitable prior for the joint distribution of types and features in the situation at hand. Implementing this Bayesian approach in order to provide unbiased estimates of diversity with their corresponding error bars is a topic for future research.

A second line for future investigation is to further examine the relation between the two alternative approaches to include disparity into diversity using Hill numbers:

including pairwise similarities directly into Hill numbers as in (Leinster and Cobbold, 2012; Chiu et al., 2014), or -as in the current paper- using alpha and beta diversity instead. On the one hand, pairwise similarities between types may not adequately capture total disparity since they do not take into account in which way pairs are similar. On the other hand, taking into account the full distribution of features requires more data in order to estimate the joint distribution of types and features. Furthermore, it may be challenging to find variables that are readily interpretable as features of the types of interest. In situations where data is limited, an approach based on pairwise similarities may be preferable.

It is worth noting however that both approaches are not mutually exclusive. In particular, Chiu et al. (2014) show that their measure of diversity, which generalizes Hill numbers to include pairwise similarities, allows for a decomposition into alpha and beta components. Their beta component then gives the number of compositional units whilst taking into account a given set of pairwise simlarities. This highlights the fact that both approaches provide alternative operationalizations of the concept of disparity. In practice, selection of a method to measure diversity requires theory-driven justification, and should be guided by data availability and the research question at hand.

Chapter 4

A network view of correspondence analysis: applications to ecology and economic complexity[*]

Abstract

Research in natural and social sciences often requires the identification of the structure underlying high dimensional data, which is often represented as a network. We revisit a statistical method known as Correspondence analysis from a network perspective, emphasizing its close relation to spectral clustering and graph embedding techniques. This leads to a number of interpretations of the results generated by the method, which may guide practitioners in its application. We show how results generated by CA relate to the structure of the underlying networks through a set of stylized examples, and discuss two empirical examples from ecology and economics. In the first example, we analyze the global distribution of Carnivora species and show how clustering and ordination can be combined to find gradients in clustered data. In the second example, we revisit the economic complexity index as correspondence analysis, and we use the different interpretations of the method to shed new light on the empirical results within this literature.

[*] This is being prepared for submission as A. van Dam, M. Dekker, I. Morales-Castilla, M.-A. Rodriguez, D. Wichmann and M. Baudena. A network view of correspondence analysis: applications to ecology and economic complexity.

4.1 Introduction

Many systems in natural and social sciences are characterized by high dimensional data sets describing the interactions between the objects of study. Such data can be analyzed by using statistical methods that reduce their complexity by identifying the low-dimensional structures that define the systems' main features. Identifying these structures enables visualization of the data in two or three dimensions, and can be used in further analysis to gain insight and understanding of the dynamics underlying the system.

A frequently used method to describe the interactions between components in a system starts with collecting data on the joint occurrence of two types of variables in the system. Such data is commonly represented by a contingency table, reporting the frequency with which an outcome of certain variable is observed in association with the outcome of another variable. A contingency table can be represented as a bipartite network, i.e. a network that connects two sets of nodes (the possible outcomes of each variable), where the edges represent joint occurrences of the outcomes. A typical example of such data sets in Ecology are the records of presence (or absence) of species in sampling sites. In Economics, a representative data set could be the presence of different types of economic activity (in terms of money or employment) in different regions.

Data represented in that way can be used to infer the associations (or similarities), between nodes of the same type, by considering for example how often species occur together in the same site. In network terms, this entails 'projecting' the bipartite network onto one of its node sets, leading to a similarity network (Fouss et al., 2016, chapter 9).

The bipartite network and the inferred similarity network hold information on the underlying dynamics of the system. For example, ecologists have been investigating the existence of latent variables that determine which species occur in which sites, a practice known as gradient analysis or ordination. These latent variables can be related to environmental variations along a gradient, for example due to latitude or temperature (Whittaker, 1967; Legendre and Legendre, 1998; ter Braak, 1995). Furthermore, analyzing the similarity networks of species or sites may reveal the existence of multiple subsystems or clusters, such as distinct communities of species or regions with distinct species compositions (Rueda et al., 2013; Holt et al., 2013; Daru et al., 2017). More recently, economists have also started to study the latent structures underlying countries' economies by leveraging network analysis to infer measures of economic complexity based on the geographical co-occurrence of products (Hidalgo and Hausmann, 2009; Hausmann et al., 2011; Mealy et al., 2019).

The analysis of contingency tables is the domain of a statistical method called Correspondence analysis (CA). Dating back to the seminal statistical work of Hirschfeld in the 1930s (Hirschfeld, 1935a), CA became widespread thanks to a classical paper by Hill, which promoted its use especially for ecological data (Hill, 1973a), while at the same time the method was developed in France by Benzécri (Benzécri and Coll., 1973). CA has been used extensively since the 80's, being as common in ecological papers as principal components analysis (PCA). Since then, CA has been reinvented many times across different fields, leading to a plethora of different names, interpretations and applications of the method (Hill, 1974; Beh, 2004; Greenacre, 1984). Given a contingency table, CA returns a set of 'axes', which, analogously to the components in PCA, are used to represent the data in a lower-dimensional space, such that the distances between the data points represent the associations between them (Greenacre, 1984).

The representation of a contingency table as a bipartite network shows that CA can also be used for network analysis. In fact, it can be shown that CA is mathematically equivalent to network methods such as clustering and graph embedding techniques (Zha et al., 2001; Yen et al., 2011). The equivalence between CA and network methods is not a simple matter of reinventing the wheel. Since each of the methods is derived with different underlying motivations (ordination, clustering or dimensionality reduction), it has the important added value of introducing different interpretations for the same data set. In this paper, we aim to raise awareness about the fact that the outcome of CA can be interpreted at the same time as latent variables, as cluster labels and as coordinates in a low-dimensional Euclidean space. By clarifying the relations between these three interpretations, we aim to aid practitioners in the interpretation of both CA and network analysis. We do so by revisiting CA from a network perspective and by providing guidance and examples to illustrate how these methods can be applied in practice. We will discuss the three alternative derivations of CA.

First, we discuss CA as a form of 'canonical correlation analysis' (Hotelling, 1936), motivated here as a way to find latent variables that drive the connections in a bipartite network (Fouss et al., 2016, chapter 9). Second, we discuss the interpretation of CA as a spectral clustering algorithm applied to the network of similarities derived from the bipartite network (Shi and Malik, 2000). Third, we discuss CA as a method of graph embedding applied to the similarity network (Yen et al., 2011). Each approach leads to a complementary interpretation of the same set of eigenvectors and eigenvalues that result from applying CA.

We illustrate the different interpretations of CA by applying it to a number of stylized networks, showing how the eigenvalues and eigenvectors that result from CA relate to their structure. When the similarity network inferred from a bipartite network consists of a single cluster, the axes resulting from CA can be interpreted as a gradient underlying the data, leading to an ordination of the nodes. However, when the network consists of multiple weakly connected clusters, the CA axes hold information on the clustering structure of the underlying network, showing for each node to which cluster it belongs. Based on these examples, we propose to use CA to cluster the data first before applying it as an ordination method within each cluster, when performing gradient analysis on data containing multiple clusters. We illustrate these ideas by analysing two empirical examples, drawn from Ecology and Economics. The proposed methodological approach is available as an R package which can be retrieved at `https://github.com/UtrechtUniversity/SCCA`.

In the first example, we apply CA to an ecological dataset describing the global geographical distribution of Carnivora species, with the objective of finding gradients that reflect drivers of the species distributions. Interpretation of CA as a clustering algorithm motivates dividing the data into subsets, leading to the identification of meaningful bioregions. Applying CA to each bioregion separately results in identification of ecological gradients within those regions.

The second example is drawn from Economics, where CA was recently reinvented as a way to analyze bipartite networks under the name of the 'Economic complexity index' (ECI). The ECI is used to infer a ranking of countries based on the products they export which is associated to their economic productivity (Hidalgo and Hausmann, 2009; Mealy et al., 2019). Here we review the ECI from the perspective of CA, and show how the different interpretations of the mathematics behind CA may help in interpreting economic complexity. We focus in particular on the interpretation of higher order eigenvectors and eigenvalues, which were hitherto not considered in the context of economic complexity.

4.2 Interpreting CA

Let us first describe the setting and introduce some notation. The main object of analysis is a matrix A (a contingency table with n_r rows and n_c columns) that contains the counts of two variables. A common example from Ecology is that A_{ij} contains some measure of abundance of species i (rows) in sampling site j (columns). The matrix A can also be a binary incidence matrix, containing the "presence-absence" of species in sites. The matrix A can be interpreted as the bi-adjacency matrix of a bipartite network that connect species to sites. The network contains n_r nodes on one side (the species, given by the rows of A), indexed by i, and n_c nodes on the other side (the sites, given by the columns of A) indexed by j. In general, we will refer to the two sets of nodes as row nodes and column nodes, respectively. The degree of a row node i is defined by $r_i = \sum_j A_{ij}$, i.e. it is given by the total abundance of a species. Likewise, the degree of a column node j is defined as $c_j = \sum_i A_{ij}$, i.e. the total abundance of species in a site. The degrees of the row and column nodes are given by the vectors $\mathbf{r} = (r_1, r_2, \ldots, r_{n_r})^T$ and $\mathbf{c} = (c_1, c_2, \ldots, c_{n_c})^T$. We further define two square matrices, D_r (n_r by n_r) and D_c (n_c by n_c) as the diagonal matrices that have $\mathbf{r}$ and $\mathbf{c}$ on the diagonal, respectively. The sum $n = \sum_{ij} A_{ij}$ gives the total number of occurrences in the table (in the case of a species-sites example, the total abundance of species).

4.2.1 CA as canonical correlation analysis

One of the first derivations of CA was obtained by applying canonical correlation analysis to categorical variables (Hotelling, 1936; Hirschfeld, 1935b; Fisher, 1940). Here we follow the derivation in Fouss et al. (2016, chapter 9), where CA is derived as an application of canonical correlation analysis applied to a bipartite network. For ease of explanation, we will assume the network is defined by a binary presence-absence matrix (i.e. the network is unweighted), but the result generalizes to any contingency table (i.e. weighted bipartite networks). The aim is to assign a 'score' to each row and column node of the bipartite network described by A, under the assumption that edges in the network arise between nodes with similar scores. The scores can thus be seen as a variable that drives the structure of the network.

The scores can be inferred from the edges of the bipartite network. Recall that for a presence-absence matrix, the total number of edges in the bipartite network is given by $n = \sum_{ij} A_{ij}$. Let us construct a vector $\mathbf{y}_r$ of length n that contains, for each edge, the scores of the row node it connects to, and a vector $\mathbf{y}_c$ of length n that contains, again for each edge, the score of the column node it connects to. Given the assumption that edges connect row nodes and column nodes with similar scores, the node scores can be found by maximizing the correlation between $\mathbf{y}_r$ and $\mathbf{y}_c$, so that the row- and column scores for each edge are as similar as possible. Denoting the vector of length n_r containing the row scores by $\mathbf{v}$ and the vector of length n_c containing the column scores by $\mathbf{u}$, this leads to the optimization problem

$$\max_{\mathbf{v},\mathbf{u}} \operatorname{corr}(\mathbf{y}_r, \mathbf{y}_c). \tag{4.1}$$

In order to obtain standardized scores, the constraints that $\mathbf{y}_r$ and $\mathbf{y}_c$ have zero mean and unit variance are added. Solving this problem using Lagrangian optimization, the solution is given by

$$\begin{aligned} D_r^{-1} A D_c^{-1} A^T \mathbf{v} &= \lambda \mathbf{v} \\ D_c^{-1} A^T D_r^{-1} A \mathbf{u} &= \lambda \mathbf{u}. \end{aligned} \tag{4.2}$$

The score vectors $\mathbf{v}$ and $\mathbf{u}$ can thus be found by solving an eigenvector problem. Both matrices on the left-hand side of Eq. (4.2) are row-stochastic and positive definite, and have identical eigenvalues that are real and take values between 0 and 1. Assuming that we have a connected network, sorting the eigenvalues in decreasing order leads to $1 = \lambda_1 > \lambda_2 \cdots \geq 0$.

It can be shown that the correlation between $\mathbf{y}_r$ and $\mathbf{y}_c$ for a given set of eigenvectors $\mathbf{v}$ and $\mathbf{u}$ is given by their corresponding eigenvalue, so that $\lambda = \operatorname{corr}^2(\mathbf{y}_r, \mathbf{y}_c)$. The node scores leading to the highest correlation are thus given by the eigenvectors associated with the largest eigenvalue. However, the eigenvectors corresponding to λ_1 have all constant values and represent the trivial solution in which all row nodes and all column nodes have equal scores (leading to a perfect correlation). The solution to Eq. (4.1) is thus given by the eigenvectors $\mathbf{v}_2$ and $\mathbf{u}_2$, corresponding to the second largest eigenvalue λ_2, which corresponds to the square root of the (maximized) correlation. For a full derivation we refer to Fouss et al. (2016, chapter 9).

The second eigenvectors $\mathbf{v}_2$ and $\mathbf{u}_2$ hold the unique scores such that row- and column nodes with similar scores connect to each other. The second eigenvalue λ_2 indicates to what extent the row- and column scores can be 'matched', where the maximal value of 1 implies that links only occur between nodes with identical scores. For high correlations, the obtained scores can be thought of as a *latent variable* that drive the formation of links in the network. In ecology, such latent variables are referred to as *gradients* (Whittaker, 1967; Legendre and Legendre, 1998). In the case of sites and species for example, CA can be applied to obtain scores that may reflect some environmental gradient determining where species locate, such as the temperature of a site and the temperature preference of a species. Such relations can be investigated by comparing the obtained gradients with known environmental variables.

The higher order eigenvectors in Eq. (4.2) and their eigenvalues are solutions to Eq. (4.1) with the additional constraint that $\mathbf{y}_r$ and $\mathbf{y}_c$ are orthogonal to the other solutions. The vectors $\mathbf{v}_3$ and $\mathbf{u}_3$ for example may represent other gradients that may drive the formation of links (e.g. precipitation, primary productivity, etc.) on top of the gradients described by $\mathbf{v}_2$ and $\mathbf{u}_2$.

4.2.2 CA as a clustering algorithm

A completely different approach shows that the eigenvectors $\mathbf{v}_2$ and $\mathbf{u}_2$ (i.e. the second eigenvectors in Eq. (4.2)) can also be interpreted as approximate cluster labels, obtained when identifying clusters in the network of similarities that is derived from the bipartite network.

A similarity network can be constructed from a bipartite network by 'projecting' the bipartite network onto one of its layers (either the row nodes or the column nodes) through stochastic complementation (Yen et al., 2011). Projecting the bipartite network defined by A onto its row layer leads to the similarity matrix $S_r = AD_c^{-1}A^T$. The entries of S_r represent pairwise similarities between row nodes of A, based on how many links they share with the same column node, weighted for the degree of each column node. Similarly, $S_c = A^T D_r^{-1} A$ defines the pairwise similarities between the column nodes of A.

Identifying clusters in the similarity network can be done by minimizing the so-called 'normalized cut' (Shi and Malik, 2000). The normalized cut assigns, for a given partition of a network into K clusters, a score that represents the strength of the connections between the clusters for that partition. A partition can be described by assigning a discrete cluster label to each node. Hence, minimizing the normalized cut is equivalent to assigning a cluster label to each node in the network in such a way that the clusters are minimally connected. Finding the discrete cluster labels that minimize the normalized cut in large networks is in general not possible (Shi and Malik, 2000). However, a solution of a related problem can be obtained when the cluster labels are allowed to take continuous values as opposed to discrete values. Solutions of this 'relaxed' problem can be interpreted as continuous approximations of the discrete cluster labels.

Minimizing the normalized cut in S_r leads to the generalized eigensystem (Shi and Malik, 2000)

$$(D_r - S_r)\mathbf{v} = \tilde{\lambda} D_r \mathbf{v}, \tag{4.3}$$

where the entries of the generalized eigenvector $\mathbf{v}_2$ corresponding to the second smallest eigenvalue $\tilde{\lambda}_2$ of Eq. 4.3 hold the approximate cluster labels of the optimal partition into two clusters. It is easily shown that generalized eigenvectors in Eq. (4.3) are exactly the eigenvectors of Eq. (4.2), where the eigenvalues are related by $\tilde{\lambda}_k = 1 - \lambda_k$, where $k = 1, 2, \ldots, n_r$ (see Appendix B).

The matrix $D_r - S_r$ is known as the *Laplacian* matrix of the similarity network defined by S_r, and is well known in spectral graph theory (Chung, 1997). The number of eigenvalues $\tilde{\lambda} = 0$ (or equivalently $\lambda = 1$) denotes the number of disconnected clusters in the network. The corresponding generalized eigenvectors of these 'trivial' eigenvalues will have constant values for nodes in each cluster, indicating cluster membership.

The situation changes when the clusters are weakly connected. The optimal solution for partitioning the similarity network into two clusters is given by the eigenvector $\mathbf{v}_2$ associated to eigenvalue λ_2. The entries of $\mathbf{v}_2$ can be interpreted as approximations to the cluster labels that indicate for each row node to which cluster it belongs. The corresponding eigenvalue λ_2 represents the quality of the partitioning as determined by the normalized cut criterion. High values indicate nearly disconnected clusters (two totally disconnected clusters would yield eigenvalues $\lambda_1 = \lambda_2 = 1$), whereas lower values correspond to a partitioning into clusters that less well distinguished (i.e. they are more interconnected). A discrete partition can be obtained from the approximate (continuous) cluster labels by discretizing them, for example by assigning all negative values to one cluster and all positive values to the other (Newman, 2013).

Finding a partitioning into *multiple*, say K, clusters is more involved. Minimizing the normalized cut for K clusters yields a trace minimization problem of which the relaxed solution is given by the first K eigenvectors in (4.2) (Yu and Shi, 2003). The discrete cluster labels can then be obtained, for example, by running a k-Means algorithm on the matrix consisting of those K eigenvectors, a technique that is also known as *spectral clustering* (Ng et al., 2002; Von Luxburg, 2007). How well the network can be partitioned into K clusters is given by the average value of the first K eigenvalues, i.e. $\frac{1}{K}\sum_{k=1}^{K}\lambda_k$ (Yu and Shi, 2003).

The clustering approach thus brings an alternative interpretation to CA results. A key observation is that the eigenvalues and eigenvectors in Eq. (4.2) are directly related to the generalized eigenvectors of the Laplacian of the similarity matrix S_r, and thus hold information on the structure of the similarity network. The entries of the second eigenvector $\mathbf{v}_2$ can be interpreted as the approximate cluster labels of a two-way partitioning of the similarity network defined by S_r. Although at first sight the interpretation of CA scores as cluster labels may seem different from the interpretation as a latent variable as described in Section 4.2.1, note that cluster labels can be seen as latent variables, albeit a discrete rather than a continuous variable.

4.2.3 CA as a graph embedding technique

A third interpretation of the eigenvectors and eigenvalues in Eq. (4.2) arise from a so-called *graph embedding* of the similarity matrix S_r (or S_c). A graph embedding represents the nodes of a graph as *node vectors* in a Euclidean space, such that nodes that are ‘close’ in the network are also close in terms of their Euclidean distance in the embedding. A key feature of these embeddings is that their dimensionality can be reduced in order to obtain a low-dimensional representation of the data, while retaining its most important structural properties (see Fouss et al. (2016, chapter 10) for an overview of graph embedding techniques). This can be used for example for graph drawing, as it provides a way to obtain a two-dimensional representation of a high-dimensional network.

Several authors have shown the equivalence of CA to graph embedding in the case of a similarity matrix obtained through stochastic complementation. For example, computing a 1-step diffusion map of the similarity matrix S_r leads exactly to the eigenvectors of Eq. (4.2) (Coifman and Lafon, 2006; Yen et al., 2011). Belkin and Niyogi (2003) show the equivalence between constructing an embedding using the Laplacian eigenmap and clustering using the normalized cut, which in turn is equivalent to CA.

Embedding the similarity network S_r in a $(K-1)$-dimensional space yields an 'embedding matrix' $X_r \in \mathbb{R}^{n_r \times K-1}$. Each row of X_r represents a node of S_r as a 'node vector' in the embedding. The rows of X_r can be seen as components of $(K-1)$-dimensional basis vectors that span the embedding, and are identical to what is referred to as the 'axes' in CA. Every entry $X_{i,k}$ represents the coordinate of row node i on the k'th basis vector, and can be seen as the 'score' of i on the k'th CA axis. An embedding matrix of S_r can defined as $X_r = [\sqrt{\lambda_2}\mathbf{v}_2, \ldots, \sqrt{\lambda_K}\mathbf{v}_K]$, where the vectors $\mathbf{v}_k$ are the eigenvectors defined in (4.2), and each of them is weighted by the square root of their corresponding eigenvalue. We will refer to columns of the embedding matrix as 'CA-axes', given by $\mathbf{x}_k = \sqrt{\lambda_k}\mathbf{v}_k$.

The axes are constructed in such a way that they capture the largest amount of 'variation' or 'inertia' in the data, which is given by their corresponding eigenvalue (Greenacre, 1984). The sum of all the eigenvalues gives the total variation in the data (in CA, this is referred to as the *total inertia*). CA decomposes the total variation in such a way that the first axis captures a maximal part of the variation, the second a maximal part of the remaining variation, and so on. A low-dimensional embedding that preserves the maximal amount of variation can thus be obtained by discarding the eigenvectors corresponding to smaller eigenvalues. The 'quality' of the embedding can then be expressed as the share of the total variation that is preserved in the embedding.

A typical way of presenting CA results is by showing the first two coordinates of each row (or column) node, i.e. plotting $\mathbf{x}_2$ against $\mathbf{x}_3$, which is usually referred to as a biplot (Greenacre, 1984). Since the first two axes capture a maximal amount of inertia, such a plot is in a way the optimal two-dimensional representation of the data that captures the relations between the rows (or columns) of A. The distances between points in the biplot approximate the similarities between nodes. How well the biplot represents the similarities is given by the percentage of variation explained by the first two axes.

Each axis can be interpreted as a latent variable that accounts for part of the total variation in the data. Since the axes in the embedding are given by a scaled version of the eigenvectors discussed in Section 4.2.1, the interpretation of the eigenvalues as the amount of variation explained is complementary to the interpretation as the correlation between row and column scores which we introduced above in Section 4.2.1. Furthermore, the axes spanning the K-dimensional embedding are exactly the generalized eigenvectors that follow from minimizing the normalized cut for K clusters (Belkin and Niyogi, 2003). Indeed, when there are clear clusters in the similarity network, they will show up in the embedding space as separate groups of points.

Summarizing, we find three interpretations of CA axes and their corresponding eigenvalues: as latent variables that drive the formation of links in the bipartite network, as approximate clusters labels of a bi-partition of the similarity network, and as coordinates of an embedding of the similarity network. The different derivations of CA and their interpretations are summarized in Table 4.1.

Name	**Interpretation eigenvectors**	**Interpretation eigenvalues**
canonical correlation analysis	latent variables	strength of correlation between row an column scores
graph partitioning using the normalized cut	approximate cluster labels	quality of the normalized cut
graph embedding	coordinates in the embedding space	variation explained

Table 4.1: Different interpretations of the eigenvectors and eigenvalues resulting from CA.

4.3 Stylized examples

In the following, we illustrate the interpretations found above by applying CA to a set of simple stylized networks: a random bipartite network (Figure 4.1 a), a network with a band-diagonal structure (Figure 4.1 b), networks with two or three weakly connected clusters (Figure 4.1 c and d), and a network with two clusters that each have a band-diagonal structure (Figure 4.1 e).

Figure 4.1 shows from left to right the bi-adjacency matrices A of the bipartite networks (where the rows and columns are sorted according the their scores in $\mathbf{v}_2$ and $\mathbf{u}_2$), the similarity matrices of the row nodes S_r, the spectrum of eigenvalues of the row-normalized similarity matrices, the (scaled) eigenvector $\mathbf{x}_2$ corresponding to the second largest eigenvalue (the first CA axis) and the biplot, which shows the two-dimensional embedding spanned by the first two CA axes $\mathbf{x}_2$ and $\mathbf{x}_3$.

For the random bipartite network (Figure 4.1 a1), the single trivial eigenvalue $\lambda_1 = 1$ indicates that similarity network S consists of a single connected component (Figure 4.1 a3). Its corresponding eigenvector has constant values (not shown). The second eigenvector shows the node scores that maximize the correlation between row and column nodes (Figure 4.1 a4). Since the network is random, this correlation is low ($\sqrt{\lambda_2} = \sqrt{0.02} = 0.14$), indicating the absence of a clear underlying structure to the network that can be captured in a single variable. Accordingly, the biplot (Figure 4.1 a5) does not show any particular structure, and each axis explains a limited amount (approximately 2%) of the total variation in the data .

Different patterns are observed in a network with a clear band-diagonal pattern (Figure 4.1 b1). This pattern is indicative of a gradient underlying the structure of the bipartite network, since high-score row nodes (on the right-hand side of the matrix) connect to high-score column nodes (on the bottom of the matrix) and vice versa. Indeed, the spectrum contains, next to the trivial eigenvalue ($\lambda_1 = 1$) two eigenvalues that are larger than the others (Figure 4.1 H). The strength of the correlation between row nodes and column nodes is given by $\sqrt{\lambda_2} = \sqrt{0.7} = 0.84$, and the gradient for the row nodes is given by the axis $\mathbf{x}_2$ shown in Figure 4.1 b4). The third eigenvalue λ_3 is much smaller than the second, but slightly larger than the subsequent eigenvalues. The biplot (Figure 4.1 b5) shows that the corresponding axis $\mathbf{x}_3$ is approximately a quadratic function of the first. This is a statistical artefact known as the ‘arch effect’ (these type of axes were referred to as ‘polynomial axes’ by (Hill, 1974)). Such solutions arise because a quadratic function of the ‘true’ gradient also leads to positive correlation, and is orthogonal to the solution given by $\mathbf{x}_2$ (Gauch et al., 1977). The solution thus contains little extra information on top of what is already reported by the second eigenvector and can thus be ignored in practice (ter Braak, 1995).

In the subsequent example, the network is constituted by two weakly connected random bipartite networks, so that the similarity network presents two clusters (Figure 4.1 c1). This represents a slightly perturbed case of the situation in which the clusters are totally disconnected. Hence, the second eigenvalue (Figure 4.1 c3) is close to one, and, unlike the situation in Figure 4.1 b4, its corresponding eigenvector shows a clear separation between two sets of approximately constant values. These entries may be interpreted as approximation to cluster labels identifying the two clusters. The subsequent eigenvectors identify axes that show variation within each of the two clusters. For example, the biplot (Figure 4.1 c5) shows that the second axis $\mathbf{x}_3$ only varies in one of the clusters (as identified by $\mathbf{x}_2$) and is approximately constant for nodes in the other cluster.

A similar situation is found for a network constituted of three weakly connected random bipartite networks (Figure 4.1 d1). Since there are now three clusters, both eigenvectors $\mathbf{v}_2$ and $\mathbf{v}_3$ are associated to eigenvalues that are close to 1 (Figure 4.1 d3). While $\mathbf{x}_2$ only identifies a two-way partition, both axes together clearly identify the three clusters, as shown in the biplot (Figure 4.1 d5).

Finally, we consider what happens when a network consists of two weakly connected clusters that each have a clear gradient (as in Figure 4.1 F), shown by their band-diagonal pattern in the bi-adjacency matrix (Figure 4.1 e1). The spectrum shows four eigenvalues that are significantly nonzero (Figure 4.1 e3). The first two eigenvalues correspond to presence of two well-defined clusters, and the second eigenvector clearly separates the two clusters (Figure 4.1 e4). The third and fourth eigenvalues correspond to the gradients within each of the clusters. The biplot shows variation in one of the clusters (Figure 4.1 e5). The gradient of the other cluster is contained in the fourth axis (not shown).

4.4 Clustering versus ordination

The different interpretations of CA axes and their eigenvalues can help make sense of results when applying CA to data. In particular, two types of CA axes can be

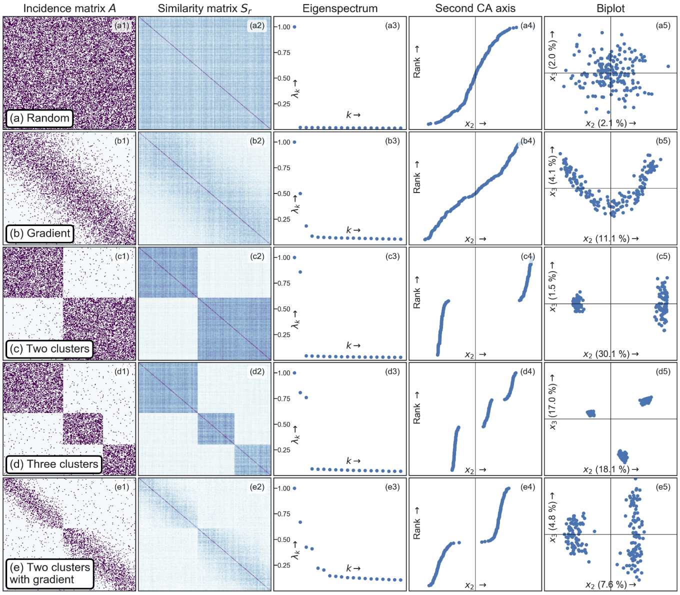


Figure 4.1: Stylized examples of five network structures and their corresponding spectra and eigenvalues. From top to bottom: a random bipartite network (a), a network with a clear gradient (b), a network consisting of two weakly connected clusters (c), a network consisting of three weakly connected clusters (d), and a network consisting of two weakly connected clusters containing a gradient (e). From left to right: the bi-adjacency matrix of the network, the similarity matrix S_r describing pairwise similarities between the rows, the spectrum of eigenvalues, the principal CA axis $\mathbf{x}_2$, and the biplot of the first and second CA axes $\mathbf{x}_2$ and $\mathbf{x}_3$, respectively, with the variation explained for both axes.

distinguished: those related to ordination and those related to clustering[1]. Axes describing a gradient are characterized by smoothly varying values within them, and are typically used to construct an ordination of the nodes under consideration[2]. Axes related to clustering are characterized by groups of approximately constant values that indicating cluster membership. These values can be discretized, for example by using k-Means, to obtain a set of discrete clusters. The eigenvalues corresponding to each axis either indicate the strength of the gradient, or how well the similarity network can be partitioned.

[1] Hill (1974) termed these two types of axes as 'seriation' axes and 'nodal' axes, respectively

[2] In Shi and Malik (2000) approach, these axes are discarded when finding a discrete partitioning of the data

While in the stylized networks presented above the number of clusters and/or the presence of a gradient was imposed and thus known beforehand, this is typically not the case in practice. Real data may consist of weakly connected clusters that may or may not have underlying gradients. This can make results hard to interpret, especially especially when using noisy data. When the objective of applying CA is to find gradients underlying the data, the CA axes with their continuously varying values are the subject of interest. In datasets consisting of multiple clusters, such gradients will be present only within each cluster, and thus found in the higher order axes (as in Figure 4.1 e). However, the higher order axes will be increasingly affected by noise (Shi and Malik, 2000). Therefore, we propose to separate the clusters in the data prior to finding gradients, so that each cluster can be analyzed separately, leading to better identification of the within-cluster gradients.

The clusters can be identified with CA by taking the spectral clustering approach, meaning that the clusters are identified by applying k-Means clustering to the embedding of the similarity network (Ng et al., 2002; Von Luxburg, 2007). This requires estimating the number of clusters beforehand. Getting this number right is important, since imposing too many clusters (i.e. including axes representing gradients) may lead to artificial cuts in the data.

A common approach for determining the number of clusters K is the 'eigengap heuristic' (Von Luxburg, 2007), which is based on the fact that the number of connected components is given by the number of eigenvalues equal to one. Since the case of a network with weakly connected clusters is similar that in which the clusters are disconnected, the spectrum of eigenvalues will also be similar, and the number of clusters can be estimated by counting the number of eigenvalues that remain close to one. This can be done by counting the number of eigenvalues that precede the largest difference between two subsequent eigenvalues (the 'gap') in the eigenspectrum. For example, the eigenspectrum of the example network with two and three clusters (Figure 4.1 c3 and d3) clearly shows how the position of the eigengap marks the number of clusters present.

A possible procedure to find gradients in data is then as follows: given a dataset, we apply CA to it and consider the results. Based on the results, we determine whether the data contains multiple clusters, and if so how many. This can be done, for example, by using the eigengap heuristic. We then separate the clusters, for example by applying K-means. The procedure can then be repeated for each individual cluster (note that the clusters themselves may also consist of multiple clusters), until each cluster represents a 'homogenous' similarity network, of which the CA axes represent continuous gradients. The procedure described in this paragraph is built into an R pacakge, which is available for general use at `https://github.com/UtrechtUniversity/SCCA`.

In the following, we explore the preceding considerations empirically by applying CA to two different datasets. The first example consists of an ecological dataset that contains clear clusters, and we aim to identify gradients within those clusters applying the procedure described above. We use the eigengap heuristic to obtain a clustering of the data and subsequently to analyze the clusters using CA. In the second example, we apply CA to data on international trade, and illustrate how the alternative interpretations of CA axes and their eigenvalues may help elucidate what is known in the literature as the economic complexity index (ECI).

4.5 Applying CA

4.5.1 Example I: Carnivore biogeography

To illustrate the application of CA as both a method of both clustering and ordination, we applied it to a common-use dataset in Macroecology –i.e. the branch of Ecology studying ecological patterns and processes at broad geographical scales, also known (*sensu lato*) as biogeography– referred to as the global geographical distributions of the species of the mammalian order Carnivora (Diniz-Filho et al., 2009). The dataset is comprised by an incidence matrix —i.e. presence-absence matrix— with 288 extant terrestrial and marine species (rows) and 41,580 non-empty sites (columns). The sites represent grid-cells rasterized at a resolution of 0.78 latitudinal degrees. The distributional data were extracted from the mammal range map database Phylacine v1.2 (Faurby et al., 2018), which we downloaded (last accessed in November 2019; `https://datadryad.org/stash/dataset/doi:10.5061/dryad.bp26v20`) and pruned to only include extant carnivorans. Data were processed in R (R Core Development Team 2014) and mapped in QGIS v2.18.16 (QGIS Development Team 2015).

In this example, we primarily focused analyzing the sites based on their species composition. Yet, considering that CA simultaneously uncovers patterns for both the rows and columns of the incidence matrix, we could have as well aimed at analyzing species based on their geographic distributions by applying the same procedure to columns instead of rows (see e.g. (Morales-Castilla et al., 2017)). Application of CA reveals the fact that we are dealing with a dataset consisting of multiple clusters. As discussed above, we first split the dataset into separate subsets that each consist of a single cluster. These clusters represent sites with similar species composition, termed bioregions in biogeography. Once we defined the bioregions, we asked whether the sites within them showed any distributional gradient, which would be reflected by the arrangement of sites along "meaningful" CA axes, i.e. axes explaining a considerable amount of variation as given by the corresponding eigenvalue. If so, it is then possible to ask whether this gradient would relate to some environmental factor or other spatially patterned process (e.g. distance to main areas of species interchange with other regions (Morales-Castilla et al., 2012)). These kind of questions are common in CA-based ecological investigations and oftentimes are addressed through correlational analyses between meaningful CA-axes and explanatory factors. Since this is well-known practice among ecologists, we did not include this part in the present analysis.

Applying CA to the complete dataset yields two eigenvalues equal to 1, showing that the network consists of two completely disconnected components. The eigenvectors corresponding to these two trivial eigenvalues indicate the membership of each node to one of the two disconnected components, which correspond to Madagascar and to the rest of the world, respectively (see Fig. 4.2 (B-D)). Malagasy carnivorans belong to the family Eupleplidae, comprised by ten species all endemic to Madagascar (i.e. they occur nowhere else), which accounts for this primary differentiation.

We proceed by analyzing the spectrum of each of the identified components separately. Looking at the spectrum of Madagascar, we find the eigengap in the first position, indicating a network consisting of a single cluster. The spectrum of the rest of the world, on the other hand, shows the largest decrease in the eigenvalues (the eigengap) between the 16th and 17th eigenvalue, indicating the existence of at least 16 other bioregions with faunas overlapping to some extent. To obtain these clusters, we follow the spectral clustering approach and apply a k-Means clustering on the embedding matrix defined by the 16 axes corresponding to these eigenvalues.

Applying CA again to each of the resulting 16 clusters, we found that 15 of them were recognized as integral bioregions (regions 4 to 20 in Figure 4.2 (A), with the exception of regions 12-13) and required no further partitioning based on the eigengap heursitic (i.e. the eigengap was found between the first and second eigenvalue). One of the clusters was split once more based on its eigenspectrum, yielding regions 12 and 13, which comprise the North American Pacific (Fig. 4.2 (A)). While the regions resulting from applying the eigengap heuristic match closely with well-established bioregions (Rueda et al., 2013), their ordering does not reflect hierarchical structure that can be linked to biogeographic history. This is probably due to most regions being defined already by the second iteration.

Although this analysis is only based on the mammalian order Carnivora, and even though it includes both terrestrial and marine species, the bioregions obtained correspond remarkably well to the bioregions first defined by Wallace and supported by recent work on bioregionalization (Rueda et al., 2013). Beyond finding the above described clustering in bioregions, we find informative results on potential gradients of site distribution within the obtained bioregions. In the Nearctic, for example, the first CA axis shows a marked north to south gradient (see Fig. 4.2 (E-G)) that would likely reflect a spatial structuring of species distributions as a response to latitudinal variation in climate.

The first CA axis has been recurrently used in Ecology (Greenacre, 2010) to conduct either cluster or gradient analyses, which has allowed ecologists and biogeographers to identify regions with similar species composition or to understand how species distributions structure along environmental gradients. However, gradient analyses may only yield interpretable results when the network considered does not consist of multiple clusters. A ‘naive’ application of CA as an ordination method in the example above would have yielded uninterpretable results, as the principle axis would have been a one-dimensional representation of a high dimensional data set. Recognizing the clustered nature of the data by considering the whole spectrum and multiple axes allows for splitting the data into homogeneous subsets using the clustering approach, before applying CA as a method for gradient analysis.

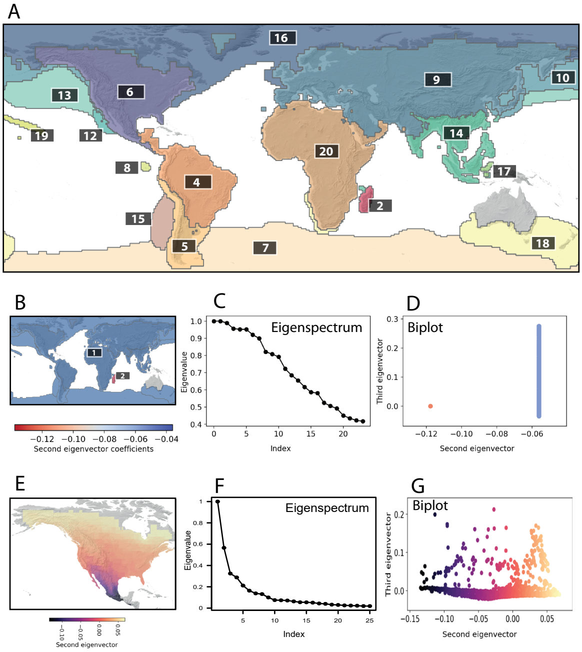


Figure 4.2: Results of applying CA to ecological data. Panel (A): Map showing the spatial distribution of clusters–i.e. bioregions–resulting from application of our heuristic approach to CA to the Carnivora dataset. The numbers indicate the resulting clusters: the first split separates Madagascar (2) from the rest of the world (1) (B). The rest of the world cluster is subsequently split into non-nested clusters, of which one is further split into clusters 12 and 13. Panel (B): Spatial distribution of the coefficients in the first CA axis when all data is considered. Madagascar is colored in red and the rest of the world in blue. Panel (C): Sorted eigenspectrum for all data showing two eigenvectors with value equal 1, indicating that there are two fully disconnected clusters. Panel (D): Biplot of first and second CA axis eigenvector. Colors indicate the first axis, corresponding to colors in panel (B). The first axis separates the two clusters but does not show variation within clusters. Panel (E): Spatial distribution of the coefficients in the second eigenvector for the subcluster 6 corresponding to the Nearctic bioregion. The eigengap is found between the first and second eigenvalues, suggesting that the network is does not contain any subclusters (F). Panel (G): The biplot for the first and second CA axes, colored according to panel (E), showing how the first axis separates sites along a clear latitudinal gradient.

4.5.2 Example II: Economic Complexity

In the second example, we apply CA to data on international trade, obtained from Harvard's Growth Lab.[3] From the data we construct a 'presence-absence' matrix with 234 countries (rows) and 1239 products (columns), in which a 'presence' indicates that a country was a significant exporter of a product in the year 2016 (see Appendix A for an exact description of this procedure). This matrix has been analyzed extensively in relation to economic development (Hidalgo et al., 2007; Hidalgo and Hausmann, 2009; Hausmann et al., 2011; Tacchella et al., 2012). In the literature on economic complexity, the first CA axes of countries and products (the row and column nodes) are known as the economic complexity index (ECI) and product complexity index (PCI), respectively (Mealy et al., 2019). The ECI has been used as a method of ranking countries by the complexity of their economy, and has become known for its ability to predict the cross-country differences and future growth of countries' GDP per capita (Hidalgo and Hausmann, 2009; Hausmann et al., 2011). The ECI has since been applied to a variety of datasets in economics (Balland and Rigby, 2017; Chávez et al., 2017; Gao et al., 2016), and beyond (Baudena et al., 2015).[4]

Applying CA to the country-product matrix results in an embedding of the country-country similarity network, where similarities are based on the countries' export portfolios (alternatively, we could have analyzed how products relate to each other in terms of the countries that export them, leading to a variation of the 'product space' introduced in Hidalgo et al. (2007)). Figure 4.3 (A) shows the country-product matrix, where countries and products are sorted by the first CA axis (the ECI and PCI, respectively). The correlation associated with the country and product scores is given by $\sqrt{\lambda_2} = 0.52$. This correlation is an indication to what extent countries with high ECI export products with high PCI and vice versa. The moderate correlation and the triangular shape of the matrix however show that this statement is only partially true, as the lower right side of the matrix shows that typically countries with high ECI also export products with low PCI. A high correlation would imply a more pronounced band-diagonal structure of the country-product matrix.

[3] The Growth Lab at Harvard University. International Trade Data (HS, 92), 2019

[4] See also `https://oec.world/en/resources/library` for a collection of work related to economic complexity.

Moving beyond the principal axes, we examine the spectrum of eigenvalues, of which the first 25 are shown in Figure 4.3 (B). The spectrum shows no significant gaps in its decay, suggesting that the country-country similarity network does not consist of multiple weakly connected clusters according to the eigengap heuristic. Figure 4.3 (B) shows the biplot of the first and second CA axes for the country-country similarity network. Expressed as a percentage of total variation, the first axis (the ECI) accounts for 3.5% of the total variation, and the second axis for 2.5%, so the biplot captures 6% of the total variation in the country-product matrix. The distances between countries in the biplot reflect similarities between countries in terms of their export baskets. As expected from the lack of a clear gap in the spectrum, the biplot shows no clearly delineated clusters, suggesting an interpretation of each axis as a continuous gradient.

The first axis differentiates between low-income countries that mostly export crude oil such as Chad, Iraq and South-Sudan on the left-hand side of the plot, and wealthy countries involved in high-tech manufacturing such as Japan, Taiwan and Switzerland on the right-hand side. The second axis assigns the highest scores to countries like Equatorial Guinea, Qatar and Venezuela, which are also major oil producers, and assigns low score to countries such as Bangladesh, Cambodia and Haiti, which are specialized in textiles and garments.

The separation of low-income and high-income countries by the first CA axis becomes clear when used as a predictor of GDP per capita (GDPpc; Figure 4.3 C), with a linear relationship explaining about 47% of the variance.[5] This relation is interpreted in Hidalgo and Hausmann (2009) as more complex countries being able to achieve higher levels of GDP per capita. Also, countries that are located below the regression line are expected to have high growth rates, as they are less rich than expected given their ‘complexity’, while countries above the regression line are richer than expected by their ‘complexity’. The typical interpretation of CA however leads to a more agnostic take on the meaning of ECI and its relationship with GDPpc. The ECI reflects a gradient that captures the maximal amount of variation in the data (around 3.5%), which in turn can be intepreted as a one-dimensional embedding of the country-country similarity network. In particular, ECI is a measure of similarity rather than ‘complexity’ (Mealy et al., 2019). The fact that ECI is associated with

[5] The data on GDPpc in 2016 is given in PPP constant 2017 international dollars, and taken from the World Bank databank `https://databank.worldbank.org`.

the GDPpc thus shows that countries with similar export baskets have also similar wealth.

Finally, we explore what can be learned from the higher order CA axes of the country-product matrix. Even though the eigengap heuristic does not suggest clearly delineated clusters, the higher order axes may still distinguish (groups of) countries that differ from the other countries in a particular way. To explore this possibility, we attempted to identify clusters in a high-dimensional embedding. We thus ran a k-Means algorithm on the 20-dimensional embedding of the country similarity matrix. The number of dimensions was motivated by the slight gap between the twenty-first and twenty-second eigenvalues. Choosing $K = 3$ (again motivated by a small gap in the spectrum) leads to the identification of the three clusters shown in color in Figure 4.3 (C and D) (see Appendix C for an overview of the clusters). The clusters clearly separate: i) countries who's exports consist almost entirely out of oil, ii) a number of small island economies, and iii) the rest of the world. The positions of the countries in each cluster in Figure 4.3 D suggest that the obtained clustering is able to explain some of the deviations from the relation between GDPpc and the ECI, attributing it to the presence of natural resources or to their unique geographical locations. In this sense, identifying clusters defined by higher order axes may provide a way to remove outliers in the data and reduce noise in the data before identifying a gradient. Recomputing the CA axes within the rest-of-the world cluster yields a noticeable increase of the R^2 for the linear relation between GDPpc and the first CA axis up to 0.65 .

The insight that ECI is equal to the first axis of CA questions its interpretation as a measure of complexity. Rather, analyses involving the ECI can be seen as a form of gradient analysis, revealing an underlying latent variable that is associated with GDPpc. The different interpretations presented in this paper may shed new light on the empirical results found within the literature on economic complexity, for example by providing an interpretation for the higher order axes and their eigenvalues. Furthermore, the observation that the complexity indices are a one-dimensional embedding of a similarity network unifies the complexity indices with the so-called ‘product space’ that is well known in the economic complexity literature (Hidalgo et al., 2007). That is, the product complexity index is simply a one-dimensional representation of the product space.

The literature on CA provides a rich set of tools (Greenacre, 2007) that can be used in the context of economic complexity, including analysis of the contribution of countries or products to the position of each axis ('which countries are the main contributors to ECI?'), measures for how well specific countries or products are represented by an axis ('which countries are well represented by ECI'?), and ways to visualize additional points in the embedding space ('how will a country move along the complexity rankings when adding some particular products?').

4.6 Discussion

In this paper, we provided an overview of different mathematical derivations that all lead to results that are equivalent to Correspondence analysis (CA). We showed that CA is closely related to the spectral analysis of the similarity network inferred from the bipartite network defined by a contingency table, providing a framework in which ordination, clustering and dimensionality reduction are three sides of the same coin. Better understanding of these relations and of their interpretation may guide practitioners in the application of CA to different datasets.

When performing CA, the eigenvalues corresponding to each axis are indicative of the correlation between row and column scores, as well as the variation explained by each CA axis. Axes corresponding to large eigenvalues may represent either a gradient underlying the data, or hold information on clustering structure in the similarity network. The distribution of the eigenvector components within an axis are suggestive of the appropriate interpretation: a continuous distribution suggests that an axis reflects a gradient underlying the data, whereas eigenvectors with a limited number of approximately constant values suggest that an axis holds information about the clusters in the similarity network.

The full spectrum of eigenvalues further provides information about the structure of the data. A spectrum containing multiple eigenvalues close to one, followed by a clear drop in the eigenvalues, may be indicative of a similarity network that contains multiple (weakly connected) clusters. In such a case, we propose to cluster the data prior to performing gradient analysis, and analyze potential gradients within each cluster separately. In Section 4.5.1 this was done using the eigengap heuristic. In Section 4.5.2, we manually determined the number of clusters based on the spectrum of eigenvalues, and showed that removing clusters that are weakly connected in the similarity network along some higher order axis may lead to clearer gradients, effectively removing outliers.

However, a formal way of distinguishing axes that represent a continuous gradient from axes that describe clustering structure is still lacking. This problem is very closely related to the question of how to determine the number of clusters in the spectral clustering approach. Although use of the eigengap heuristic is common practice, it is based on intuition, and formalizing approaches like the eigengap heuristic is an ongoing topic of research (Tibshirani et al., 2001). Another possible way forward is to take into account the distribution of values within the axes in addition to the eigenvalues, to determine whether an axis represents a continuous gradient or approximately discrete value indicating clustering structure (Zelnik-Manor and Perona, 2004).

Furthermore, we note that the usage of k-Means as a way to cluster the spectral representation of a network might be problematic, as it determines cluster labels by assuming spherical cluster shapes. As there is no underlying basis for assuming any cluster shape given the abstract networks derived from ecological or economic data, further research is needed to understand the performance of other clustering algorithms in the context of CA. In particular, density-based clustering techniques such as DBSCAN (Ester et al., 1996), which emphasize the similarity between nodes instead of partitioning a network, might be a promising step forward. In addition, exploring other dimensionality reduction techniques to obtain simplified representations of a data set different from the spectral embedding discussed here might be a promising way forward in the applications discussed in this paper.

It must be noticed also that within the framework of CA however, there is no clear-cut distinction between clustering and gradient analysis. As we have shown, both clustering and ordination can be seen as a way of identifying latent variables underlying the data. Especially in noisy data, a CA axis may be somewhere at an intermediate point between identifying clusters and representing a gradient. A principled way of distinguishing the different functions of CA axes would require an underlying theory or null model specific to the research question at hand, against which the results can be compared in order to select relevant axes. Lacking such models, the distinction between the two remains partly a matter of heuristics.

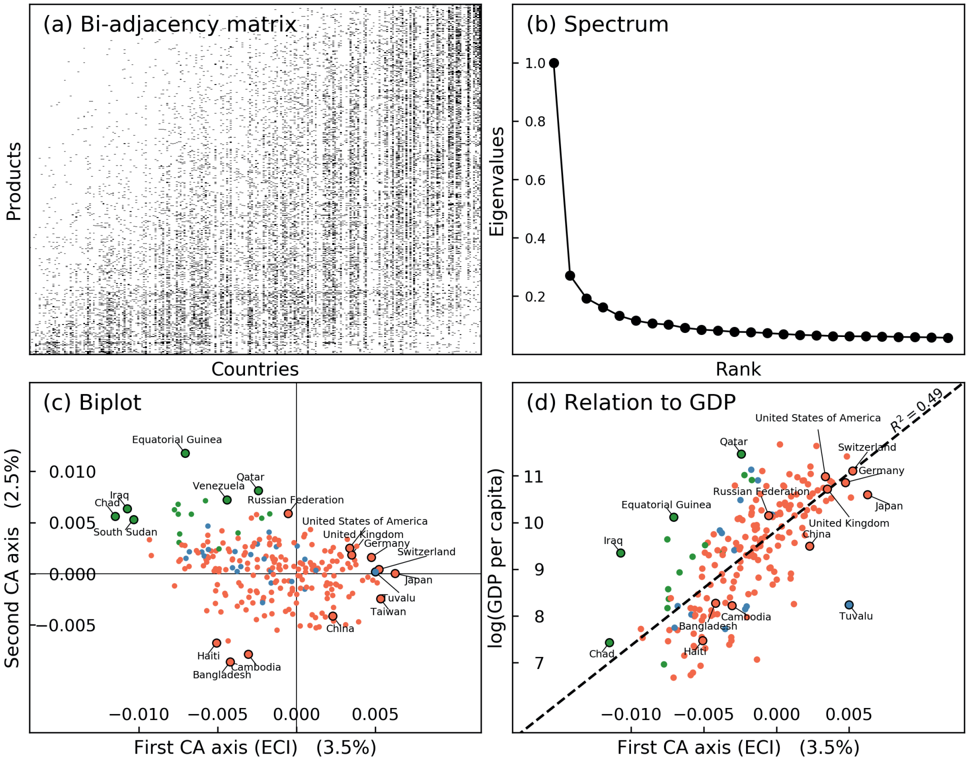


Figure 4.3: Results of applying CA to the country-product matrix. Panel (A): The country-product matrix, where column and rows are sorted by the principal CA axes, known as te ECI and PCI. The matrix shows a triangular structure. Panel (B): Sorted eigenspectrum for the country-product matrix. The slow decay of the spectrum and lack of clear gaps in the spectrum suggests a high-dimensional, homogeneous dataset. Panel (C): Biplot showing the first and second CA axis for countries. The first CA axis (horizontal) is known as the ECI, and explains 3.5% for the total variation. The second axis explains 2.5% of total variation and seems to distinguish countries specializing in garments and textiles from other countries. Colors indicate the obtained clusters when running k-Means with $K = 3$ on the embedding spanned by the first 20 CA axes. Panel (D): The relation between the first CA axis (ECI) and GDP per capita. The regression line of $GDPpc$ against ECI and the corresponding R^2 are also shown. Colors again indicate the obtained clusters.

Chapter 5

An information-theoretic approach to the analysis of location and co-location patterns*

Abstract

We propose a statistical framework to quantify location and co-location associations of economic activities using information-theoretic measures. We relate the resulting measures to existing measures of revealed comparative advantage, localization and specialization and show that they can all be seen as part of the same framework. Using a Bayesian approach, we provide measures of uncertainty of the estimated quantities. Furthermore, the information-theoretic approach can be readily extended to move beyond pairwise co-locations and instead capture multivariate associations. To illustrate the framework, we apply our measures to the co-location of occupations in US cities, showing the associations between different groups of occupations.

*This chapter is available online as a working paper as: van Dam, A., Gomez-Lievano, A., Neffke, F., and Frenken, K. An information-theoretic approach to the analysis of location and co-location patterns. *arxiv.org/abs/2004.10548*, 2020.

5.1 Introduction

The recognition of differential specialization patterns lies at the heart of economics since the works of Adam Smith and David Ricardo. Economists studying task assignments (Roy, 1951; Sattinger, 1993), urban economies (Ellison and Glaeser, 1997; Ellison et al., 2010), or international trade (Balassa, 1965; Krugman, 1991b), all stress the fact that different economic entities specialize in different activities. Scholars in each of these fields have relied on indices that quantify, for example, the revealed comparative advantage of exports, the specialization of regions, and the extent of localization and (co-)agglomeration of industries. However, these indices are often used ad hoc and lack a clear statistical foundation. In this paper, we propose a statistical framework from which such measures can be derived. Although the methodology generalizes immediately to other contexts, to fix ideas, we focus on economic geography and derive measures of (co-)location, specialization and localization from a single statistical framework, revealing the internal connections between these concepts.

We treat (co-)location as the realizations of two categorical random variables: the location and the type of an economic activity. We use the Pointwise Mutual Information (PMI) to express the association between a location and the type of an activity in terms of the information that the type of a unit of activity (e.g. a person's occupation) gives about the unit's location (e.g. the city where that person works). Next, we show how the PMI can be used to quantify the association between two activity types in terms of how much information observing a particular activity type in a location gives about observing another activity type in the same location. That is: if we observe a pair of people from the same city, how much information does the occupation of one of them provide about the likely occupation of the other?

The information-theoretic basis that underlies the PMI ensures that the framework is explicit about the null models, priors and data-generating processes we assume. This puts the measurement of location and co-location on a rigorous statistical footing. Furthermore, we show how the PMI can be estimated from data on the counts of activities across locations. To do so, we use a Bayesian framework that assumes that the data on the presence of units of economic activities across locations are generated from a multinomial distribution. This Bayesian estimation framework resolves some well-known measurement issues and provides a measure of uncertainty for the estimated quantities.

Metrics based on Information Theory such as the PMI have found various applications in economics (Theil, 1967), and are uniquely derived from axioms about how information can be gained from probability distributions (Shannon, 2001; Cover and Thomas, 2005). One of their key properties is that they can be aggregated and decomposed to form well-defined measures that have an interpretation in terms of information, by taking expectations. This allows the use of the PMI as a building block of information-theoretic measures that describe properties at the location, activity, or even system level.

We show how the resulting measures can be related to well-known existing indices of localization and specialization. In particular, at the level of location-activity pairs – as exemplified in country-product or city-industry data – our metric of association, the PMI, is conceptually similar to the logarithm of the widely used index of revealed comparative advantage (RCA) (Balassa, 1965).[1] This provides an information-theoretic motivation for considering the logarithm of the RCA index, which has the practical advantage that it overcomes the RCA index's problem of distributional skew. Moreover, the Bayesian estimation procedure ensures that the measure always attains finite values, and suggests a natural measure of uncertainty for the estimates.

Building on the location-activity PMI, we can furthermore derive a measure for the localization of economic activities, that is, for the degree to which economic activities are spatially constrained. We do so by calculating an activity's expected PMI (i.e., the expected association of the activity with a given location) over all locations. This yields the Kullback-Leibler divergence, which has been proposed as a measure of localization before (Mori et al., 2005). Likewise, we can calculate the expected location-activity PMI of a particular location across all activity types. This average association of a location with given activities provides a measure of specialization that is conceptually similar to Krugman's specialization index (Krugman, 1991b).

Finally, we apply the PMI to the distribution of co-located pairs of economic activity, which gives the probabilities that pairs of activities are located in the same geographic unit. This provides a measure of spatial association between economic activities. Such measures may reveal positive or negative co-location forces, and are conceptually similar to widely used (co-)agglomeration measures (Ellison and Glaeser, 1997; Ellison

[1] The RCA is also known as the Location Quotient in the regional science literature (Isard, 1960).

et al., 2010). Here we derive such measures from first principles, which clarifies their underlying assumptions and statistical properties.

As in the case of location-activity pairs, marginalizing the PMIs of co-located activity-activity pairs yields meaningful aggregate quantities. Accordingly, the expected spatial association of an activity with all other activities gives a measure of the spatial 'co-dependence' of an activity. This measure reveals how 'picky' activities are in their tendencies to co-locate with other activities. This spatial co-dependence is low for activities that locate independently of other activities, whereas co-dependence is high for activities that are preferentially found in the presence of specific other activities. As an empirical illustration, we calculate the associations between pairs of occupations groups, along with the aggregate spatial co-dependence of each occupation group, using US city-occupation employment data. The associations between occupation groups reveal three clear clusters. The first consists of occupations related to knowledge intensive services, the second to occupations related to non-traded services and the third to occupations related to manufacturing.

5.2 Information-theoretic measures of (co-)location

5.2.1 Notation

Consider data on the location of economic activities in the form of an $N_c \times N_i$ dimensional matrix $\mathbf{Q}$, where N_c and N_i are the number of locations and economic activities in the classifications of the data, respectively. We call $\mathbf{Q}$ the 'prevalence matrix' as its entries q_{ci} denote the number of occurrences of activity i in location c. This can be for example the number of people employed in a particular occupation i in a city c, the number of establishments of industry i in region c or the number of dollars of product i exported by country c. The total amount of activity of type i and the total activity in location c are given by the row sums $q_c = \sum_i q_{ci}$ and column sums $q_i = \sum_c q_{ci}$, respectively. Total economic activity is given by $q = \sum_{c,i} q_{ci}$.

We will consider the prevalence matrix $\mathbf{Q}$ to be the outcome of a sampling process from the underlying distribution $\mathbf{p}$ with probabilities

$$p_{ci} = P(X = i, C = c) \tag{5.1}$$

that a randomly sampled unit (i.e. an employee, an establishment, a dollar) is part of activity i in location c. Here, the categorical random variables X and C denote the activity and location of a randomly sampled unit, respectively. Their marginal probabilities are given by $p_i = \sum_c p_{ci} = P(X = i)$ and $p_c = \sum_i p_{ci} = P(C = c)$.

The location-activity probabilities p_{ci} will be the main object of interest as they hold information on the associations between locations and activities (Section 5.2.2). From these probabilities it is also possible to construct the probabilities p_{ij} that a pair of economic activities i and j are present in the same location, which is used to analyze the co-location association (Section 5.2.3). Both p_{ci} and p_{ij} are estimated from $\mathbf{Q}$ using a Bayesian framework as described in Section 5.3.

5.2.2 Location association

As noted, we will use the dependencies hidden in the joint probabilities p_{ci} to measure the association between an activity and a location. Information theory provides a framework in which these associations can be quantified explicitly in units of information. The association between the two events $X = i$ and $C = c$ is given by their pointwise mutual information $PMI(p_{ci})$ (Fano, 1961). Intuitively, it answers the question 'how much information does observing c provide about the presence of i?' PMI has been used in several fields, including economics (Theil, 1967), administrative sciences (Theil, 1972), and linguistics (Church and Hanks, 1989). Here, we use it in the context of economic geography to measure the association between economic activities and locations (location association) and within pairs of economic activities (co-location association).

The PMI measures the association between two outcomes by assessing the information content of the realization $(C = c, X = i)$ given the information content in case of a null model in which c and i are independent, i.e. $p_{ci} = p_c p_i$. This is given by the logarithm of ratio of both probabilities:[2]

$$PMI(p_{ci}) = \log\left(\frac{p_{ci}}{p_c p_i}\right). \tag{5.2}$$

$PMI(p_{ci})$ will be positive when it is more likely to observe c and i together than expected under independence, i.e. $p_{ci} > p_c p_i$, whereas $PMI(p_{ci})$ takes negative values when c and i are less likely to occur together than expected under the null model of independence, i.e. $p_{ci} < p_c p_i$. $PMI(p_{ci}) = 0$ if and only if $p_{ci} = p_c p_i$, indicating that c and i are independent (i.e., the incidence of an activity is independent of the place). The maximum value of $PMI(p_{ci})$ is given by $\max\{\log\left(\frac{1}{p_i}\right), \log\left(\frac{1}{p_c}\right)\} = \log\left(\frac{1}{p_{ci}}\right)$, which is attained either when activity i always occurs in location c, or when activity i is the only activity in location c.[3] $PMI(p_{ci})$ is not bounded from below, as it tends to $-\infty$ as the joint probability p_{ci} tends to 0.

5.2.3 Co-location association

We can also use this information-theoretic framework to obtain a measure of association between pairs of economic activities. To do so, we expand (5.1) to include two units of activity:

$$p_{cij} = P(X_1 = i, X_2 = j, C = c), \tag{5.3}$$

where X_1 and X_2 describe randomly sampled units of activity from the same location C.

The measure of co-location will come from integrating across places to get the joint distribution of economic activities

$$p_{ij} = P(X_1 = i, X_2 = j). \tag{5.4}$$

[2] In information theory, the information content or 'surprise' of an outcome i is defined as $\log(\frac{1}{p_i})$. Observing an event that occurs with small probability leads to a high information content or surprise, whereas highly likely events contain little information. The difference between the information contents of p_{ci} and $p_c p_i$ gives a measure of the surprise of observing p_{ci} while expecting $p_c p_i$. Depending on the base of the logarithm, PMI measures association in units of bits (base 2) or nats (natural logarithm).

[3] Notice that then $p_{ci} = p_i$ or $p_{ci} = p_c$ respectively.

The probability p_{ij} thus represents the joint probability that two units of economic activity that are randomly picked from the same (random) location are of type i and j. It can be obtained by exploiting the fact that, conditional on knowing the location c, the occurrence of types i and j are independent, i.e. $p_{ij|c} = p_{i|c}p_{j|c}$, since the full distribution of economic activities for every location is known. By the law of total probability, one then obtains

$$p_{ij} = \sum_{c} p_{i|c}p_{j|c}p_c. \tag{5.5}$$

This defines the probability that two randomly sampled units from the same (random) location have activity types i and j.

As with the location-activity associations, the association between activity types can be quantified with the PMI. The association between two activities is then defined as

$$PMI(p_{ij}) = \log\left(\frac{p_{ij}}{p_i p_j}\right), \tag{5.6}$$

where $p_i p_j$ is the null model that describes a situation where i and j are distributed independently of each other. What $PMI(p_{ij})$ captures is that the presence of some activities may increase or decrease the probability that other activities are present in the same location. Hence, observing a particular type of economic activity holds information about the likelihood of observing other types of activities in the same location. Economic activities that are more likely to occur together than expected under independence will have a positive association, whereas activities that are less likely to occur together than expected under independence will have a negative association.[4] The $PMI(p_{ij})$ is inherently symmetric, since $p_{ij} = p_{ji}$. Computing this measure for all pairs of activity types thus leads to a symmetric, square matrix that has as entries the co-location association $PMI(p_{ij})$.

The diagonal entries of this matrix hold 'self-associations' $PMI(p_{ii})$. Self-association is high when observing an activity of type i in a particular region increases the likelihood that a second randomly sampled unit in that location is also of type i. This

[4] Another way of seeing this, is by noting that $PMI(p_{ij})$ is positive when observing type i increases the probability of observing type j when sampling units of activity from the same location, i.e. $p_{j|i} > p_j$. Likewise, negative associations indicate that conditional on observing i, the probability of sampling a unit of activity j in the same location decreases.

is the case when the probability of observing i is above average in a few locations, and below average in others. The self-association can thus be interpreted as a measure of geographical concentration. Note that the self-association is always positive, i.e. $PMI(p_{ii}) \geq 0$, since observing a unit of activity of type i can never lower the probability of finding another unit of activity of type i (we sample with replacement). The matrix of co-location associations thus provides a joint estimate of geographic concentration and co-location.

5.3 Bayesian estimation

In order to compute the quantities above, an estimate of the probabilities p_{ci} is needed. A straightforward way to estimate these probabilities is to consider the share of every location-activity pair, corresponding to the maximum likelihood estimate $\hat{p}_{ci} = \frac{q_{ci}}{q}$. Here we estimate p_{ci} using a Bayesian framework, which has two major advantages over the maximum likelihood approach. First, the Bayesian approach always returns nonzero probability estimates, so that computing the PMI will always return finite values. Second, the Bayesian framework yields a full posterior distribution for the estimated probabilities as opposed to a point estimate. The posterior distribution provides a natural description of the uncertainty in the estimated parameter values, which can be used to construct a Bayesian error bar for the information-theoretic quantities based on those estimates (Wolpert and Wolf, 1995).

Assuming that $\mathbf{Q}$ is generated by an independent sampling process, the probability of its realization is given by a multinomial distribution

$$P(\mathbf{Q}|\mathbf{p}) = \frac{\Gamma(q+1)}{\prod_{c,i}\Gamma(q_{ci}+1)} \prod_{c,i} p_{ci}^{q_{ci}},$$

where $\mathbf{p}$ is the matrix containing probabilities p_{ci}, $\sum_{c,i} p_{ci} = 1$.

Applying Bayes' rule, the posterior distribution for the matrix of probabilities $\mathbf{p}$ is then given by

$$P(\mathbf{p}|\mathbf{Q}) \propto P(\mathbf{Q}|\mathbf{p})P(\mathbf{p}),$$

where $P(\mathbf{p})$ represents the prior distribution. A conjugate prior for the multinomial distribution is the Dirichlet distribution

$$P(\mathbf{p}|\boldsymbol{\alpha}) \sim Dir(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{c,i}\Gamma(\alpha_{ci})}\prod_{c,i} p_{ci}^{\alpha_{ci}-1},$$

where $\alpha = \sum_{c,i}\alpha_{ci}$. This gives the distribution of $\mathbf{p}$ given hyperparameter $\boldsymbol{\alpha}$. The posterior distribution for $\mathbf{p}$ given the data $\mathbf{Q}$ and hyperparameter $\boldsymbol{\alpha}$ is then given by

$$P(\mathbf{p}|\mathbf{Q},\boldsymbol{\alpha}) \sim Dir(\mathbf{Q}+\boldsymbol{\alpha}) \propto \prod_{ci} p_{ci}^{q_c i+\alpha_{ci}-1}.$$

The hyperparameter $\boldsymbol{\alpha}$ can be interpreted as a matrix of ‘pseudocounts’, giving the assumed number of observed units of activity for every c, i pair prior to seeing the data $\mathbf{Q}$. The total number of pseudocounts α determines the strength of the prior relative to the data. An estimate for the parameters p_{ci} is then given by the expectation of the marginals of the posterior distribution, so that

$$\hat{p}_{ci} = \mathbb{E}[p_{ci}|\mathbf{Q},\boldsymbol{\alpha}] = \frac{q_{ci}+\alpha_{ci}}{q+\alpha} = \frac{\tilde{q}_{ci}}{\tilde{q}},$$

where we write $\tilde{q}_{ci} = q_{ci} + \alpha_{ci}$ and $\tilde{q} = \alpha + q$. When the pseudocounts α_{ci} are nonzero for all c, i, then $\hat{p}_{ci} > 0$ will also be nonzero. This has the practical advantage that it prevents difficulties when computing logarithms of the estimated probabilities, as when calculating $PMI(p_{ci})$.[5]

A measure for the uncertainty of the estimate $\hat{p}_{ci}$ is given by the variance of the marginals of the posterior distribution, leading to

$$\begin{aligned}\mathrm{Var}[p_{ci}|\mathbf{Q},\boldsymbol{\alpha}] &= \frac{\tilde{q}_{ci}(\tilde{q}-\tilde{q}_{ci})}{\tilde{q}^2(\tilde{q}+1)} \\ &= \frac{\tilde{q}_{ci}/\tilde{q}(1-\tilde{q}_{ci}/\tilde{q})}{\tilde{q}+1}.\end{aligned}$$

Note that this implies that the variance is dependent on the granularity of the data in $\mathbf{Q}$. To see this, suppose we alter the units leading to a new matrix $\mathbf{Q}' = k\mathbf{Q}$, so

[5] In the context of information retrieval in text analysis, adding the pseudocounts α_{ci} to categorical data is known as ‘Laplace smoothing’ or ‘additive smoothing’ (Manning et al., 2008).

that for large q

$$\begin{aligned}\mathrm{Var}[p_{ci}|\mathbf{Q}',\boldsymbol{\alpha}] &= \frac{k\tilde{q}_{ci}/k\tilde{q}(1-k\tilde{q}_{ci}/k\tilde{q})}{k\tilde{q}+1}\\ &\approx \frac{1}{k}\mathrm{Var}[p_{ci}|\mathbf{Q},\boldsymbol{\alpha}].\end{aligned}$$

The variance thus decreases as the counts become more fine-grained. The reason is that the data generating process is assumed to create the data at the level of the counts, so that more-fine grained units represent more observations. The variance of the estimates is thus directly related to the units in which the underlying data generating process is assumed to generate the data.[6] However, the variance is affected by the granularity of the data in the same way across activities and locations, so that the *relative* uncertainty of estimates $\hat{p}_{ci}$ is independent of the units of $\mathbf{Q}$.

One could use the estimate $\hat{p}_{ci} = \mathbb{E}[p_{ci}]$ directly to compute $PMI(\hat{p}_{ci})$ and $PMI(\hat{p}_{ij})$. However, this will induce a systematic bias which comes from Jensen's inequality $\mathbb{E}[PMI(p_{ci})] \lesseqgtr PMI(\mathbb{E}[p_{ci}])$ depending on whether $PMI(p_{ci})$ is concave or convex.[7] One needs instead an estimate of $PMI(p_{ci})$, which in itself is a random variable whose distribution is determined by the posterior distribution of p_{ci}. Thus, we use the uncertainty for the estimates $\hat{p}_{ci}$ to determine the uncertainty of estimates for $PMI(p_{ci})$ and $PMI(p_{ij})$.

5.3.1 Estimation of the posterior mean and variance of $PMI(p_{c,i})$

Here, we approximate the mean and variance of the posterior distribution of $PMI(p_{ci})$, which will serve as estimates of the posterior distribution of the location-activity association. Our approach is based on Wolpert and Wolf (1995) and Hutter and Zaffalon (2005), in which the estimation of information-theoretic quantities using a Bayesian approach is discussed in depth.

To obtain an approximation for the posterior distribution of $PMI(p_{ci})$, we compute its Taylor expansion around the mean $\hat{p}_{ci}$. Writing $\Delta_{ci} = p_{ci} - \hat{p}_{ci}$, and noting the

[6] In the context of (co-)agglomeration of industries for example, the relevant unit of analysis is the one at which location decisions are made, which could be be assumed to be the plant level, suggesting an analysis of data containing the counts of plants of a specific industry for a given location.

[7] $PMI(p_{ci})$ is concave when $\partial^2 PMI(p_{ci})/\partial p_{ci}^2 = -1/p_{ci}^2 + 1/p_c^2 + 1/p_i^2 < 0$, and convex when $\partial^2 PMI(p_{ci})/\partial p_{ci}^2 > 0$.

fact that $|\Delta_{ci}| < 1$, this gives

$$\begin{aligned} PMI(p_{ci}) &= PMI(\hat{p}_{ci}) + \Delta_{ci}\left(\frac{1}{\hat{p}_{ci}} - \frac{1}{\hat{p}_c} - \frac{1}{\hat{p}_i}\right) \\ &+ \frac{\Delta_{ci}^2}{2}\left(-\frac{1}{\hat{p}_{ci}^2} + \frac{1}{\hat{p}_c^2} + \frac{1}{\hat{p}_i^2}\right) + \mathcal{O}(\Delta_{ci}^3). \end{aligned}$$

Note that $\mathbb{E}[\Delta_{ci}] = 0$ and thus $\mathbb{E}[\Delta_{ci}^2] = \text{Var}[p_{ci}]$, where expectations are taken with respect to the posterior distribution of p_{ci}. It follows that

$$\mathbb{E}[PMI(p_{ci})] \approx PMI(\hat{p}_{ci}) + \frac{\text{Var}[p_{ci}]}{2}\left(-\frac{1}{\hat{p}_{ci}^2} + \frac{1}{\hat{p}_c^2} + \frac{1}{\hat{p}_i^2}\right). \tag{5.7}$$

The second term accounts for systematic bias in the estimate of $PMI(p_{ci})$, in which the sign of the factor multiplying the variance is indicative of whether $PMI(p_{ci})$ is concave or convex, and thus determines whether the bias is positive or negative.

Using the Delta method, we then obtain for the variance of $PMI(p_{ci})$:

$$\begin{aligned} \text{Var}[PMI(p_{ci})] &\approx \text{Var}[p_{ci}]\frac{\partial PMI(p_{ci})}{\partial p_{ci}}|_{\hat{p}_{ci}} \\ &= \text{Var}[p_{ci}]\left(\frac{1}{\hat{p}_{ci}} - \frac{1}{\hat{p}_c} - \frac{1}{\hat{p}_i}\right)^2. \end{aligned} \tag{5.8}$$

This is a measure for the uncertainty around the point estimate $\mathbb{E}[PMI(p_{ci})]$. In particular, it can be used to determine whether the estimate for $PMI(p_{ci})$ is significantly nonzero, i.e. if there is a significant association between i and c.

5.3.2 Estimation of posterior mean and variance of $PMI(p_{ij})$

Approximations of $\mathbb{E}[PMI(p_{ij})]$ and $\text{Var}[p_{ij}]$ are obtained in a similar fashion, replacing p_{ci} with p_{ij} in equations (5.7) and (5.8), although the computation of $\text{Var}[p_{ij}]$ is more involved. Appendix A provides a discussion of how $\text{Var}[p_{ij}]$ is obtained. Appendix B provides comparisons to numerical simulations to justify the approximations made.

5.4 Location and co-location

5.4.1 Revealed Comparative Advantage

One of the most commonly used indices to study location patterns of economic activities originates from the trade literature, where it is known as Balassa's index of Revealed Comparative Advantage (RCA) (Balassa, 1965). The RCA of a location-activity pair is given by the ratio of the share of activity i within location c compared to the share of activity i in the overall economy:

$$RCA(c, i) = \frac{q_{ci}}{q_c} / \frac{q_i}{q}. \tag{5.9}$$

It compares the observed share of activity i within location c in the numerator to the total share of i as given by the denominator. Since q_i and q_c are exchangeable in (5.9), $RCA(c, i)$ can be interpreted in two ways: as a measure of 'localization' of activity i in location c, or as a measure of 'specialization' of location c in activity i. The neutral value is given by $RCA(c, i) = 1$, where the share of activity i in location c is equal to the total share of activity i over all locations.

A theoretical derivation of the RCA index is given by Kunimoto (1977), who uses a probabilistic approach that comes close to the approach presented in this paper. Properties of the RCA and related indices have since been discussed extensively (Yeats, 1985; Ballance et al., 1987; Vollrath, 1991), some of which are problematic when applying the index in empirical analysis. One of the issues of the RCA index is that it is heavily skewed and asymmetric around its neutral value. A possible solution that has been presented is taking the logarithm of the index, making it symmetric around a neutral value of 0 (Vollrath, 1991). This, however, leads to the problem that the index becomes undefined in the cases where $q_{ci} = 0$, since the logarithm of zero is undefined.

The approach presented in the current paper provides an information-theoretic derivation of the logarithm of the RCA index. Consider the maximum likelihood estimate for the multinomial probabilities $\hat{p}_{ci} = \frac{q_{ci}}{q}$.[8] We then have that

$$\begin{aligned} PMI(p_{ci}) &= \log\left(\frac{\hat{p}_{ci}}{\hat{p}_c \hat{p}_i}\right) \\ &= \log\left(\frac{\tilde{q}_{ci}}{\tilde{q}_c} \Big/ \frac{\tilde{q}_i}{\tilde{q}}\right) \\ &= \log(RCA(c,i)), \end{aligned}$$

showing that conceptually the PMI is equal to the logarithm of the RCA index.

Our approach stands therefore as a generalization of the RCA index. This shows that there is an information-theoretic notion of association underlying the RCA. Seen in this light, the practical problem of having to take the logarithm of zero when $q_{ci} = 0$ is in fact a problem related to miss-estimating p_{ci}. In our Bayesian approach, the estimates of probabilities p_{ci} are always strictly positive.

5.4.2 Measures of localization

Many questions are better answered at more aggregate levels of analysis than the level of location-activity pairs. Typical questions at these levels of aggregations rely on quantifying which activities are most localized in space, or which locations are most specialized in terms of their economic activities.

Localization of an activity can be defined as the degree of dissimilarity between the activity's own geographical distribution and the distribution of the population or of total economic activity across all locations (Hoover, 1936; Mori et al., 2005). Highly localized activities will be distributed across locations in a very different way than what one would expect from locations' sizes. Activities with a low degree of localization will be distributed proportionally to the relative (population) size of locations. This can be quantified by comparing how much, on average, the probability that a unit of activity of type i is located in a location differs from the probability that any unit of activity is located there.

[8] Note that here we write q_{ci} and not $\tilde{q}_{ci} = q_{ci} + \alpha_{ci}$, since the maximum likelihood estimate uses directly the observed counts, without adding the pseudocounts that where a consequence of incorporating a prior distribution of counts in the Bayesian estimate.

Let $p_{c|i} = p_{ci}/p_i$ be the probability that a unit of activity is located in c given that its activity type is i, and recall that the probability that a unit of economic activity is located in c regardless of its type is given by p_c. Considering the average deviations between $p_{c|i}$ and p_c leads to a measure of localization that is given by

$$\begin{aligned} KL(p_{c|i}|p_c) &= \sum_c p_{c|i} \log(p_{c|i}/p_c) \\ &= \sum_c p_{c|i} PMI(p_{ci}), \end{aligned}$$

where we used that $p_{c|i}/p_c = p_{ci}/p_c p_i$. Here, KL denotes the Kullback-Leibler divergence (Kullback and Leibler, 1951), and measures the deviation between the distribution across all locations of a specific activity, given by probabilities $p_{c|i}$, and the overall distribution of locations, given by the probabilities p_c. Hence, the proposed information-theoretic framework naturally suggests a localization measure by aggregating $PMI(p_{ci})$ to the activity level. The resulting metric can be interpreted as the activity type's expected locational dependence.

This measure has the exact same functional form as the measure of industrial localization put forward by Mori et al. (2005), although the null model implicit in their metric is based on a location's area. That is, they take the probability p_c to be proportional to the area of that location as opposed to its population size.[9] Here we show that their measure can be retrieved as the expected PMI values of a particular industry.[10] Ignoring differences in how these distributions are estimated, the functional of this measure is equal to $\mathbb{E}_{p_{c|i}}[\log(RCA(c,i))]$, showing that it can be understood as the expected value of the logarithm of the RCA of an activity over all locations it occurs in.

5.4.3 Measures of specialization

Similarly, the aggregate level of specialization of a location as a whole can be analyzed by quantifying the difference of the distribution of activities within the location, $p_{i|c}$,

[9] Furthermore, they obtain an error bar for this statistic based on a normal approximation. In the Bayesian framework, an estimate for the standard deviation of the KL can be obtained in a similar way as for the PMI, as shown in Appendix C.

[10] This holds regardless of the 'null model' considered. Hence, one could follow Mori et al. (2005) and use their area based null model to define a measure on the location-activity level that is analogous to the RCA index.

to the overall distribution of activities p_i. Such a measure of specialization is obtained by aggregating the $PMI(p_{ci})$ to the location level, thus considering the expected association of the activity with particular locations, leading to

$$KL(p_{i|c}|p_i) = \sum_c p_{i|c} PMI(p_{ci}).$$

Again, this can be interpreted as the expected value of the logarithm of the RCA, but now over industries within a given location: $\mathbb{E}_{p_{i|c}}[\log(RCA(c,i))]$. The measure is akin to Krugman's specialization index (Krugman, 1991b).[11] However, in our framework, the localization of activities and specialization of locations are essentially the same measures, defined for different units of analysis.

5.4.4 Overall specialization

Aggregating even further, a measure for the overall specialization at the system level can be obtained by taking the expectation over both locations and activities, leading to the expected association of a location-activity pair, or equivalently as either the expected localization of an activity or the expected specialization of a location. The resulting quantity is known as the Mutual Information (MI) (Cover and Thomas, 2005) and quantifies the dependence between two random variables. In this case, it measures the dependence between the random variables X and C, which describe the type and location of a randomly sampled unit of activity. It is given by

$$\begin{aligned} MI(C,X) &= \sum_{c,i} p_{ci} PMI(p_{ci}) && (5.10) \\ &= \sum_i p_i KL(p_{c|i}|p_c) && (5.11) \\ &= \sum_c p_c KL(p_{i|c}|p_i). && (5.12) \end{aligned}$$

When $MI(C,X) = 0$, the location of a randomly sampled unit is independent of its activity type, which implies that all economic activity is distributed proportionally to location size, or equivalently that every location has an identical distribution of

[11] The Krugman specialization index is given by $K(c) = \sum_i |p_{i|c} - p_i|$. Like $KL(p_{c|i}|p_i)$, it considers an 'average deviation' of $p_{i|c}$ to p_i, where the measure of deviation is taken to be the absolute difference.

unit of analysis	measure	formula
location-activity	association	$PMI(p_{ci})$
activity	localization	$KL(p_{c\|i}\|p_c) = \mathbb{E}_{pc\|i}[PMI(p_{ci})]$
location	specialization	$KL(p_{i\|c}\|p_i) = \mathbb{E}_{pi\|c}[PMI(p_{ci})]$
system	overall specialization	$MI(C, X) = \mathbb{E}_{pci}[PMI(p_{ci})]$

Table 5.1

activities. In this situation, there is no specialization in the system in the sense that all locations are identical. The maximum value of $MI(C, X)$ is reached when each location has its own unique activity, so that each location is maximally specialized and each activity is maximally localized. In the current context, the mutual information is a system-level measure of overall specialization that can be used to compare across different systems (e.g. comparing the degree of overall specialization across countries), or to track the changes over time (e.g. comparing the degree of overall specialization before and after the establishment of a trade union). Table 5.1 summarizes each of the measures derived thus far and the relation between them.

5.5 Co-location

5.5.1 Co-location association

So far, we have studied the matrix $\mathbf{Q}$, which summarizes location patterns of economic activity. Our framework can however readily be extended to study more complex patterns. Here we will discuss co-location patterns of pairs of activities, i.e., of the dependencies between activities that are located in the same region. Such co-location patterns have received increasing attention in studies on international trade (Hidalgo et al., 2007) and urban economies (Ellison et al., 2010). In the latter field, authors have used co-location patterns to test theories on Marshallian externalities (Marshall, 1920). In this literature, the co-agglomeration index of Ellison et al. (2010) has become a de facto standard (Faggio et al., 2017; Diodato et al., 2018). Here, we show how information theory can be used to derive an alternative measure based on the co-location association, $PMI(p_{ij})$.

Before presenting our co-location metrics in detail, it is useful to first discuss how Ellison et al. (2010) construct their co-agglomeration index. These authors present a location choice model for profit-maximizing plants (Ellison and Glaeser, 1997; Ellison et al., 2010) in which the (combined) effects of natural advantage and spillovers between activity types determine co-agglomeration patterns. They propose the following pairwise co-agglomeration index:[12]

$$\gamma_{ij} = \frac{\sum_c (p_{c|i} - p_c)(p_{c|j} - p_c)}{1 - \sum_c p_c^2}. \tag{5.13}$$

The co-agglomeration of all pairs can be collected in a matrix with entries γ_{ij}, completely analogous to the $PMI(p_{ij})$ in Section 5.2.3. The diagonal entries γ_{ii} contain the agglomeration index of a single activity (Ellison and Glaeser, 1997), when neglecting effects of the plant size distribution.[13]

Comparing the co-agglomeration index given in Eq. (5.13) to our co-location association metric rewritten as

$$\begin{aligned} PMI(p_{ij}) &= \log\left(\frac{\sum_c p_{i|c} p_{j|c} p_c}{p_i p_j}\right) \\ &= \log\left(\sum_c \left(\frac{p_{c|i}}{p_c}\right)\left(\frac{p_{c|j}}{p_c}\right) p_c\right). \end{aligned}$$

clarifies the conceptual similarity between the two. Both capture how different activities co-vary in space. In either case, the intensity of spatial co-location may be generated by a location choice model akin to the one by Ellison and Glaeser (1997).

The difference lies, however, in the functional form used to measure the deviation from the reference distribution. The co-location association compares probabilities by taking ratios $p_{i|c}/p_c$, whereas the co-agglomeration index considers differences

[12] Note that, in our notation, activity shares $\frac{q_{ci}}{q_i}$ and $\frac{q_c}{q}$ are replaced by probabilities $p_{c|i}$ and p_c. This makes specific that we regard the former shares as maximum likelihood estimates of the latter probabilities. For now, however, we leave the issue of estimating these probabilities open.

[13] Mori et al. (2005) show that the agglomeration index of (Ellison and Glaeser, 1997) can be written as $\gamma_i = a_i G_i - b_i \approx \frac{\sum_c (p_{c|i} - p_c)^2}{1 - \sum_c p_c^2}$. This approximation is valid when plants are reasonably uniformly distributed, in which case the plant size effect is negligible. The plant size distribution determines the size of the chunks in which the counts are generated in the data generating process. Quantifying the dependencies that arise from such a data generating process is an interesting direction for future research, but for now we focus on the simpler case in which information on the chunk sizes (e.g. the plant size distribution) is unavailable. Further note that unlike Mori et al. (2005), we compare the agglomeration index to the self-association $PMI(p_{ii})$ as opposed to the localization $KL(p_{c|i}|p_c)$.

$p_{i|c} - p_c$. Furthermore, the co-location association weights each of the differences by p_c.

Although the co-agglomeration index is derived from an economic model, the measure of concentration that lies at its heart enters the derivation as an assumption. Our framework provides a principled way to quantify these deviations, by leveraging information theory. The advantage of such an approach is that it gives insight into the underlying assumptions on the data generating process, the used reference distribution[14], and the estimation procedure with its corresponding uncertainties. Furthermore, as before, our statistical framework allows constructing measures of co-dependence at higher levels of aggregation, such as at the level of the activity or of the economic system as a whole.

5.5.2 Co-dependence

As in Section 5.4.2, the co-location associations can be aggregated by taking the expectation across all activities j, leading to a measure of the average association of activity i with all other activities, given by

$$KL(p_{j|i}|p_j) = \sum_j p_{j|i} PMI(p_{ij}). \tag{5.14}$$

We call this measure the co-dependence of a particular activity. It quantifies the deviation of the distribution of activity types conditional on having observed activity type i, $p_{j|i}$, with respect to the unconditional distribution of probabilities p_j. When activity type i has, on average, strong associations with other activity types it co-locates with, this deviation will be large. In other words, activity i 'cares' about the type of activity it co-locates with. A low value of $KL(p_{j|i}|p_j)$ on the other hand implies that the distribution of probabilities $p_{j|i}$ does not differ much from the distribution of p_j, meaning that activity i is uninformative for the type of activities it co-locates with. This implies that activity i co-locates with the 'average' distribution of activity types, suggesting it is indifferent of the other activities in the same location.

[14] In fact, the literature is not entirely consistent in the choice of the reference distribution that is used in the (co-)agglomeration indices. In some work the reference distribution is taken to be the share of total employment in location c, which we denote by p_c (Ellison and Glaeser, 1997, 1999; Faggio et al., 2017). In other work, the reference distribution is given by the average share of employment in industry i in a location, given by $\hat{p}_{c|i} = \frac{1}{N_i} \sum_i p_{c|i}$ (Ellison et al., 2010; Diodato et al., 2018).

unit of analysis	measure	formula
activity-activity	co-location association	$PMI(p_{ij})$
activity-activity	geographic concentration	$PMI(p_{ii})$
activity	co-dependence	$KL(p_{j\|i}\|p_j) = \mathbb{E}_{p_{j\|i}}[PMI(p_{ij})]$
system	overall co-dependence	$MI(X_1, X_2) = \mathbb{E}_{p_{ij}}[PMI(p_{ij})]$

Table 5.2

Note that activities that are heavily concentrated geographically, have by definition a high co-dependence, as $PMI(p_{ii})$ is part of the sum in (5.14). In that case, activity of type i typically co-locates with other activity of type i.

5.5.3 Overall pairwise dependence

Taking the expectation of the co-dependence over all activity types, or equivalently taking the expectation of the co-location association over all activity pairs leads to the mutual information

$$\begin{aligned} MI(X_1, X_2) &= \sum_i p_i KL(p_{j|i}|p_j) \\ &= \sum_{ij} p_{ij} PMI(p_{ij}). \end{aligned}$$

This is a measure of dependence between the random variables X_1 and X_2, which each describe the activity type of a randomly sampled unit of activity, both sampled from the same location (see (5.4)). The overall co-dependence is thus a system-level variable that describes how much two units of activity are on average (spatially) associated. This may, for instance, help understand how the overall strength of co-agglomeration externalities differs across economies or changes over time. Table 5.2 gives a summary of the measures that follow from analysis of the co-location distribution p_{ij}. Both Tables 5.1 and 5.2 construct similar sets of measures. Both sets of measures take averages across rows, columns, or both, of a matrix that summarizes associations between two variables. However, whereas the measures in Table 5.1 are based on the location-activity information of a matrix that collects elements $PMI(p_{ci})$, the measures in Table 5.2 are based on the spatial co-location information collected in a matrix with elements $PMI(p_{ij})$.

5.6 Empirical example

As an example, we apply the PMI to show the co-location associations of occupation groups in US employment data in 2016 provided by the Bureau of Labor Statistics.[15] The data consists of a matrix $\mathbf{Q}$ that gives for every city c the number of employees q_{ci} in a particular occupation group i. In this example, we choose a uniform prior, setting $\alpha_{ci} = 1$ for all c, i. This represents a single observation for every location-activity pair. Since the total number of pseudocounts $\alpha = N_r N_c << q$, the resulting estimates will be determined much more by the data than by than the prior.

The inferred $PMI(i,j)$ matrix is shown in Figure 5.1, showing the co-location associations between the occupation groups. The right hand side shows the co-dependence of each occupation group with respect to all other groups, corresponding to the expected value of a row in the PMI matrix. The error bars show one standard deviation in the posterior distribution, as derived in Appendix C. Red indicates positive associations, and blue negative ones.

The matrix delineates three clusters of occupations groups. The upper left block shows a cluster of positively associated occupations that seem to be related to knowledge-intensive services. The positive associations lead to a relatively high co-dependence for these occupations, suggesting that the presence of these occupations depends largely on which other occupations are present in the same city.

The lower right block of the matrix shows a smaller cluster of occupations related to production, transportation and repair. These occupations have a negative association with the knowledge-intensive occupations, and thus typically co-locate with a different set of occupations. The ‘Production’ occupations group also has a high co-dependence, which is mostly driven by a high self-association.

The ’Farming, fishing and forestry’ group is highly isolated, with mostly negative associations with other groups. The diagonal entry in the matrix shows the self-association is very high, which is also reflected in a high co-dependence, which is orders of magnitude larger than that of the other occupations (note the broken axis).

In the middle band of the matrix, occupation groups have a neutral association with most other occupations, and have a low co-dependence. These groups seem to be

[15] These data are available at `https://www.bls.gov/oes/special.requests/oesm16ma.zip`

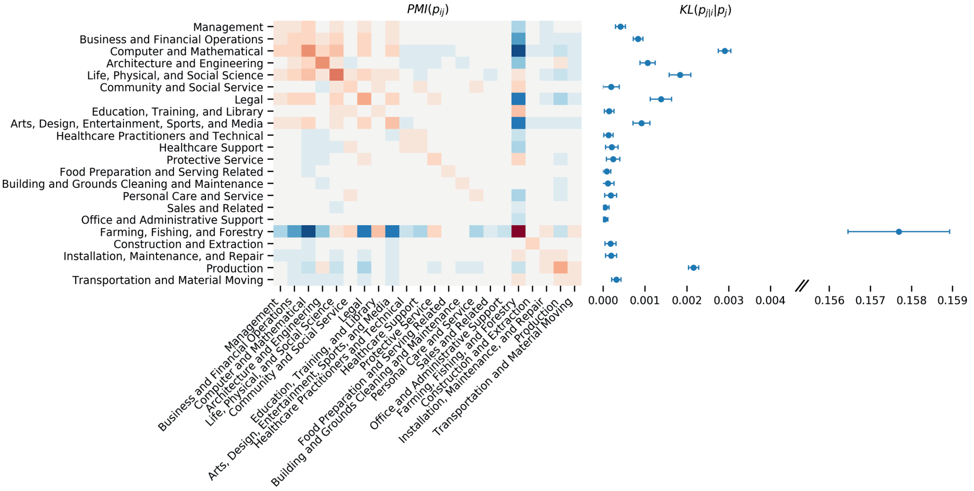


Figure 5.1: Values of the estimated $PMI(p_{ij})$ for major occupations groups. Red indicates positive associations, blue indicates negative associations, and grey indicates neutral ($PMI(ij) = 0$) associations. All pairwise associations are between -0.5 and 0.5 with the exception of the self-association of the 'Farming, fishing and forestry' occupations, which has a value of 3.15. The right hand side shows the co-dependence $KL(p_{j|i}|p_j)$ of every occupation group, given by the expected value of a row of the PMI matrix. The error bars depict one standard deviation of the posterior distribution as a measure of uncertainty for the estimate. Note the broken axis, showing the extreme dependence of the 'Farming, fishing and forestry' occupations.

related to non-traded services, including 'Protective service', 'Food preparation and serving' and 'Personal care and service'. The low co-dependence implies that these occupations are distributed approximately proportional to the total population, independently of which other occupation groups are present in a city.

5.7 Discussion

Information theory offers a unified way to estimate location and co-location associations using PMI. This yields measures that are similar to the well-known RCA index Balassa (1965) and the co-agglomeration index (Ellison et al., 2010). However, our metrics based in information theory have important advantages over these existing measures.

First, by deriving these metrics from a unified framework, we were able to show the intrinsic connections between hitherto disparate measures. This is not only satisfying from a methodological point of view, but allows exploring the relations between concepts like revealed comparative advantage, specialization, localization, concentration and co-location.

Second, the proposed measures are derived from a formal framework (information theory) in a way that is explicit in the assumed data generating process, the chosen null models and the estimation procedures. Different choices for these assumptions leads to different results. However, the afforded transparency allows to construct arguments against and in favor of such alternatives that take into consideration aspects of the specific context at hand. Such a discussion can be framed in terms of an underlying model, rather than of ad hoc specificities of a particular index. For instance, we used a null model based on the assumption that neutral associations imply a distribution of location-activity pairs that is proportional to the sizes of locations and activities (Hoover, 1936). Alternative null models could follow from the assumption that activities are distributed proportional to the area of a location (Mori et al., 2005). Another possibility is to determine the expected number of (co-)occurrences on the basis of external factors that could drive the distribution of activities over locations, using for instance a regression model (Neffke et al., 2011; Jara-Figueroa et al., 2018).

Third, the framework provides uncertainty estimates for all the information-theoretic quantities involved. Most currently used indices are applied without any notion of uncertainty. Using these uncertainties in practice however may present some challenges. For instance, the Bayesian estimation procedure leaves room for the selection of different priors. Here, for reasons of practicality, we applied a simple uniform prior. However, in some contexts, alternative priors may be natural choices. Many of these priors would still result in Dirichlet priors, but with different uniform values for α_{ci} to adapt the strength of the prior to the data at hand (Hutter and Zaffalon, 2005). In other contexts, non-uniform priors, such as the maximum entropy prior (Wolpert and Wolf, 1995), may be preferable. Furthermore, the absolute magnitude of the uncertainty will depend on the granularity of the data. This simply reiterates that inferences should always be made with an underlying data generating process in mind. In spite of this, we can still make statements about the relative magnitudes of uncertainties, which are independent of the granularity of the data generating process.

Fourth and finally, it is important to note that the information-theoretic approach can be readily extended to move beyond the analysis of pairwise co-locations, as it also allows analyzing multivariate associations. For instance, one could analyze associations between multiple variables (e.g. occupations, cities and industries) or multi-way co-locations (such as the co-location of triplets instead of pairs of activities).[16] Such higher-order associations could be further analyzed using the information-theoretic concepts of redundancy and synergy (Finn and Lizier, 2018). This may help disentangle different types of associations, capturing different economic interactions. The association between a pair of economic activities could be conditional on the presence of (a specific combination of) other activities, or be driven by the mutual dependence on a (combination of) other economic activities or on some external variable such as the presence of a natural resource. Further development of this analytical framework could reveal such higher-order relations among economic activities.

[16] The PMI between three economic activities i, j, k is given by $PMI(p_{ijk}) = \log\left(\frac{p_{ijk}}{p_i p_j p_k}\right)$.

Chapter 6

Variety, complexity and economic development*

Abstract

We propose a combinatorial model of economic development. An economy develops by acquiring new capabilities allowing for the production of an ever greater variety of products with an increasing complexity. Taking into account that economies abandon the least complex products as they develop over time, we show that variety first increases and then decreases in the course of economic development. This is consistent with the empirical pattern known as 'the hump'. Our results question the common association of variety with complexity. We further discuss the implications of our model for future research.

*This chapter was published as: van Dam, A. and Frenken, K. Variety, complexity and economic development. *Research Policy*, 2020a. doi:10.1016/j.respol.2020.103949

6.1 Introduction

Our understanding of economic growth has long been guided by the notion of a production function that specifies how inputs such as capital and labor translate into the total output of an economy. Theoretical models of economic growth typically abstracted from the exact products that an economy produces, describing economic growth instead as an increase in aggregate output. Only recently, more attention has been given to the specific products an economy produces, in particular, the products that a country exports (Hausmann et al., 2007).

At the level of products, inputs can be considered to be strictly complementary (Kremer, 1993; Hausmann and Hidalgo, 2011; Brummitt et al., 2017). This assumption is based on the idea that the production of any product or service requires a particular combination of complementary inputs. Missing one of those inputs renders the others useless in the production process. Inputs required to produce a product can be many, and include physical resources and assets as well as knowledge, skills, and even regulations. All these inputs are often referred to in an abstract and generic sense as 'capabilities' (Hidalgo and Hausmann, 2009; Hausmann and Hidalgo, 2011). Products can then be represented as strings of capabilities. The ability of an economy to produce products (including services) then depends on the number of capabilities present in a country, as well as the ways in which capabilities complement each other.

Developing new products consists of recombining old and new inputs into configurations that have economic value (Inoua, 2016). Since these new recombinations will consist largely of capabilities that were already present, new products will be similar, or 'related', to existing ones. The process of development can thus be described as one in which a country acquires one new capability at the time, and uses this new capability in combination with existing capabilities to start producing new products. This implies that economic development is a highly path-dependent process (Lall, 2000) characterized by a logic of related diversification (Hidalgo et al., 2007).

The acquisition of new capabilities does not only allow an economy to increase its variety of products, but also more complex products in terms of the number of capabilities used in products. Combining more capabilities implies a more intricate production process leading to products that are arguably more sophisticated than products combining only few capabilities. This line of thinking is consistent with the notion of economic complexity, i.e., the idea that the most complex products are produced in well-developed economics with many capabilities (Lall et al., 2006; Hausmann et al., 2007; Hidalgo and Hausmann, 2009; Sutton and Trefler, 2016).

Two streams of research have followed from this combinatorial framework. First, empirical studies have investigated the role of relatedness in economic development. New products will be related to existing products in that new products are produced using both existing and newly acquired capabilities. Following this reasoning, studies have analysed the extent to which national and regional economies diversify over time from existing products into related products (Hidalgo et al., 2007; Neffke et al., 2011).

Second, there have been several attempts to measure the average complexity of products produced by a country. The proposed measures build on methods that infer the complexity of economies by iteratively weighing the variety of products produced in a country and the ubiquity of these products in other countries. Such indirect measures of complexity have been used to explain income differences across countries and their growth rates over time (Hidalgo and Hausmann, 2009; Tacchella et al., 2012; Cristelli et al., 2015).

Notwithstanding the explanatory power of aforementioned studies, the economic complexity framework so far neglects a salient and fundamental feature of economic development. While new products enter a country's portfolio as it develops, already existing products may also exit (Cadot et al., 2011). One reason that countries lose products from their portfolio holds that wages, over time, become so high that a country cannot remain competitive in certain products (Sutton and Trefler, 2016). Products exiting the portfolio may thus be the products with low profit margins domestically, which can be imported at lower prices from low-wage countries. In addition, some products may become obsolete once their functionality is substituted by new products. Either way, understanding economic development will logically have to take into account both products entering and products exiting at any moment in time.

Empirically, it has been shown that the variety of products that an economy produces, is positively related to the income per capita of its workers (Hesse, 2008; Herzer and Nowak-Lehnmann, 2006; Al-Marhubi, 2000). This relationship, however, only holds up to a certain level of income per capita, as countries with the highest income per capita display lower variety. This inverted-U pattern between income per capita and variety is known as 'the hump' (Imbs and Wacziarg, 2003; Cadot et al., 2011). In a dynamic sense, then, the hump suggests that in the course of development, countries first diversify and then specialize again. This empirical pattern is inconsistent with the basic model of economic complexity, which would predict an ever-increasing variety as more capabilities are acquired over time.

We argue that products exiting a country's portfolio are likely to be the least complex ones. Such simple products can be imported at lower prices from low-income countries or substituted by new products entering a country's portfolio. Below, we extend the elementary combinatorial model underlying the framework of economic complexity by imposing a constraint on the range of the complexity of products that an economy can engage in. As a result, at a certain stage of development, countries will start losing their least complex products. The introduction of this constraint results in a theoretical model that i. is consistent with the principle of related diversification, ii. recovers the stylized fact of 'the hump', and iii. predicts that the growth in economic complexity of an economy accelerates as a function of newly acquired capabilities. From the model, we further derive a number of research questions, in particular, regarding the nature of products exiting countries' portfolios and the variations across countries in terms of the timing of the hump. Finally, we will argue that, as our model suggests that complexity continues to increase while variety starts decreasing, empirical measures of complexity that rely on the measurement of variety are theoretically unsupported.

6.2 A basic combinatorial model

Following Inoua (2016), we start with a simple model in which every product is represented as a string of capabilities. The *product length* is given by the number of capabilities required to produce it and indicates a product's sophistication or complexity.

The capabilities present in an economy determine the set of products that an economy can produce. For simplicity, we will assume that every possible combination of capabilities leads to a viable product (this assumption will be relaxed below where we introduce a 'recipe book'). A country that has n capabilities can make $\binom{n}{s}$ different combinations of lengths s. The most complex, sophistcated product it can produce is the one product that recombines all n capabilities. The total number of products that a country can make is given by the total number of strings one can make out of n capabilities

$$d(n) = \sum_{s=0}^{n} \binom{n}{s} = 2^n.$$

The average complexity of products is given by the total length of all products divided by the total number of products

$$\bar{s}(n) = \frac{\sum_{s=0}^{n} s\binom{n}{s}}{2^n} = \frac{n}{2}.$$

This leads to the following basic properties of the model:

1. The product variety is given by $d(n) = 2^n$, so that $\log(d(n)) \propto n$.
2. The average product length in a country is given by $\bar{s}(n) = \frac{n}{2}$, so that $\bar{s}(n) \propto n$.

Combining both properties, it follows that the logarithm of product variety is linearly proportional to the average product length:

$$\log(d) \propto \bar{s}(n).$$

Further note that in this basic model both the logarithm of the product variety and the average product length in an economy could provide a measure of economic complexity, as they are both proportional to the number of capabilities present (Inoua, 2016). Furthermore, the exponential relation between product variety and the number of capabilities reflects that a country with many capabilities can increase its variety more by acquiring a new capability compared to a country with only few capabilities, since the former has more capabilities with which the new capability can be recombined than the latter (Hausmann and Hidalgo, 2011).

6.3 Product exit

We now extend the model to incorporate the possibility of an economy losing products. We pose that as the average complexity of products in a country keeps on rising as part of its economic development - and wages rise accordingly - a country cannot remain competitive in the simplest products. As a result, a country will see its simplest products exit from its portfolio. This is modeled by imposing a product range r, which determines the range of product lengths a country produces. A large r indicates that a country makes both long and short products, essentially allowing for a large heterogeneity of product lengths. A small r means that there is little room for variation in product lengths, and all products produced will be of approximately the same length. It follows that countries produce products in the range of length $n - r$ to n, as n is the maximum product length. The product variety given r is thus given by

$$d(n, r) = \sum_{s=n-r}^{n} \binom{n}{s},$$

where $\binom{n}{s}$ is the number of products of length s that can be made out of n capabilities.

The average product length given r is given by (see Appendix A.1)

$$\bar{s}(n, r) = n \frac{d(n-1, r)}{d(n, r)}.$$

As long as $r \geq n$, the product range forms no constraint on the product lengths and no products are lost. In particular, since $d(n, r) = 2^n$ for $r > n$, we retrieve $\frac{d(n-1,r)}{d(n,r)} = \frac{1}{2}$ so that $\bar{s}(n) = \frac{n}{2}$ as before. When $r < n$, we find that (see Appendix A.2)

$$\frac{1}{2} < \frac{d(n-1, r)}{d(n, r)} < 1.$$

Assuming that an economy acquires new capabilities one-by-one, the dynamics of the model can then be represented as in Figure 6.1. Once products start exiting a country's portfolio, the rate at which the average product length increases in n goes up as the number of capabilities increases, but never exceeds 1. At the same time, the pace of diversification levels off as more products exit, but a country never loses more products than it gains.

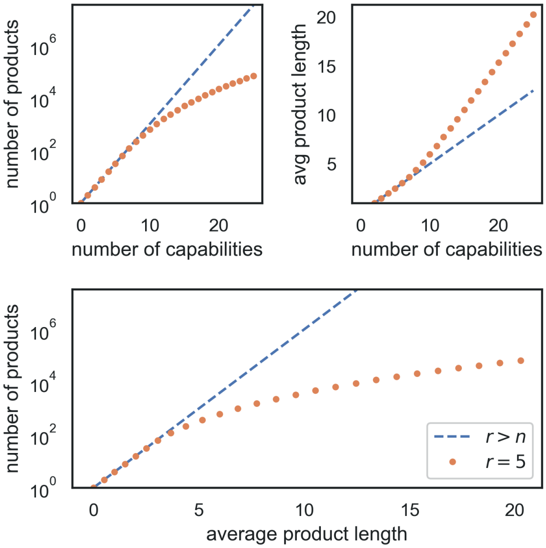


Figure 6.1: The basic dynamics of the model. The blue dashed line shows the case where no products are lost ($r > n$). The orange dots represent the case where $r = 5$, so a country only makes products with a length in the range of $n - 5$ and n. The top left panel shows the relation between product variety $d(n, r)$ and the number of capabilities. For the unconstrained case, there is an exponential relation between the two, showing a linear relationship on a logarithmic scale. Imposing a constraint of $r = 5$, the increase in variety slows down as n increases beyond 5. The top right panel shows average product length $\bar{s}(n)$ as a function of the number of capabilities. In the unconstrained case there is a linear relation, whereas the constrained case shows an acceleration in the increase of average product length once $n > 5$ and short products are lost. The bottom panel shows the resulting relation between the product variety and average product length.

6.4 Full model

The assumption that any combination of capabilities leads to a viable product is arguably too strong. More realistically, one may assume that only a fraction of combinations of capabilities result into meaningful products. This can be thought of as imposing a ’recipe book’, which describes the combinations of capabilities that are complementary in that they lead to viable products (Hausmann and Hidalgo, 2011; Inoua, 2016; Fink et al., 2017).

The basic model can be generalized by assuming that every capability is part of a viable product with a given probability ρ (Inoua, 2016). Parameter ρ can thus be thought of as reflecting the difficulty to innovation in the sense that not all combinations of capabilities, or 'recipes', lead to viable products. The lower the value of ρ, the harder it is to find useful recipes. A combination of s capabilities has probability ρ^s of representing a viable product of length s. Hence, it becomes increasingly unlikely that a combination of capabilities leads to a viable product as more capabilities are added, since ρ^s is decreasing in s when $\rho < 1$. For $\rho = 1$, we recover the basic model described before.

Since there are $\binom{n}{s}$ possible combinations of s components one can make from the total of n components, and each combination of length s has probability ρ^s of being viable, the expected number of products of length s a country with n components can make is given by $d(n,s) = \binom{n}{s}\rho^s$. Summing this quantity over all product lengths s gives the expected product variety for a given number of components n

$$d(n) = \sum_{s=0}^{n} \binom{n}{s} \rho^s = (1+\rho)^n.$$

Since the share of products of length s in a country is given by $\frac{\binom{n}{s}\rho^s}{d(n)}$, the expected product length given n components can be computed as (see Appendix A.3) (Inoua, 2016)

$$\bar{s}(n) = \sum_{s=0}^{n} s \frac{\binom{n}{s}\rho^s}{d(n)} = \frac{\rho}{1+\rho} n. \tag{6.1}$$

Note that, as in the basic model before, the expected product length increases linearly with n, where the exact rate at which the average product length increases is solely determined by the difficulty parameter ρ.

Incorporating the product range using parameter r gives

$$d(n,r) = \sum_{s=n-r}^{n} \binom{n}{s} \rho^s,$$

and the average product length given r is given by (see Appendix A.4)

$$\bar{s}(n,r) = \rho n \frac{d(n-1,r)}{d(n,r)}.$$

In Appendix A.5 it is shown that for $r < n$, the average product length is bounded from below by

$$\frac{\rho}{1+\rho} n < \bar{s}(n,r).$$

Thus once a country starts losing products, the increase of average product length with the number of capabilities starts accelerating. Furthermore, as long as variety is increasing, the increase in average product length is bounded as $\bar{s}(n,r) < \rho n$.

The model with $\rho < 1$ shows an important qualitative difference with the basic model with $\rho = 1$, in that for $\rho < 1$ a decrease in variety will occur, which happens when more products exit than enter. The condition for a decline in product variety, i.e. for 'the hump' to occur, is given by (see Appendix A.6)

$$d(n,r) < \binom{n}{r} \rho^{n-r-1}.$$

Once this condition is met, i.e. when a country starts losing more products than it gains, the average product length grows with a rate larger than ρ

$$\bar{s}(n,r) = \rho n \frac{d(n-1,r)}{d(n,r)} > \rho n$$

The model thus predicts that a decrease in variety is accompanied by further acceleration of the increase of the average product length with the number of capabilities as the shortest products are dropped.

Finally note that the increasing rate of the average product length is bounded from above by n (see Appendix A.5):

$$\bar{s}(n,r) < n,$$

which describes the limiting case in which only the single longest product of length n is produced.

In summary, the model exhibits the two 'stages of diversification' as identified empirically by Imbs and Wacziarg (2003), along with a transitory phase in between, known as 'the hump'. An overview of the three stages and their conditions is given in Table 6.1. Figure 6.2 further shows the dynamics of the model for the example of $\rho = 0.5$. In the model, the first stage of diversification is characterized by an exponential increase in product variety. During this stage no products are lost since the allowed range of product complexities exceeds the total number of capabilities. The average product length increases linearly in n with a rate that is determined by parameter ρ. In the transition stage, the simplest products are not produced anymore but the economy is still diversifying, although the rate of diversification slows down. The increases in average product length on the other hand accelerates as the shortest products exit a country's portfolio. In the final stage of diversification, then, more products are lost than gained, so variety decreases as more capabilities are acquired. During this stage, the rate of the average product length further increases and approaches the limiting rate of 1.

stage	condition	variety	avg. product length
developing	$r > n$	exponentially increasing	$\bar{s}(n) = \frac{\rho}{1+\rho} n$
transitioning	$r < n$, $d(n,r) > \binom{n}{r}\rho^{n-r-1}$	increasing with a decreasing rate	$\frac{\rho}{1+\rho} n \leq \bar{s}(n) \leq \rho n.$
developed	$r < n$, $d(n,r) < \binom{n}{r}\rho^{n-r-1}$	decreasing	$\rho n \leq \bar{s}(n) \leq n.$

Table 6.1: The conditions for the three stages in the model, with the corresponding values of product variety and average product length.

A final feature of the model hold that the product range r determines at what number of capabilities a country enters a new stage of diversification. A country with a large product range r will start losing products at a higher number of capabilities than a country with a small product range. Thus, a large r causes a country to go through the hump later than a country with lower r. And, countries with relatively low product range will experience the hump already at a low number of capabilities. The effect of r on the onset of the hump is shown in Figure 6.3.

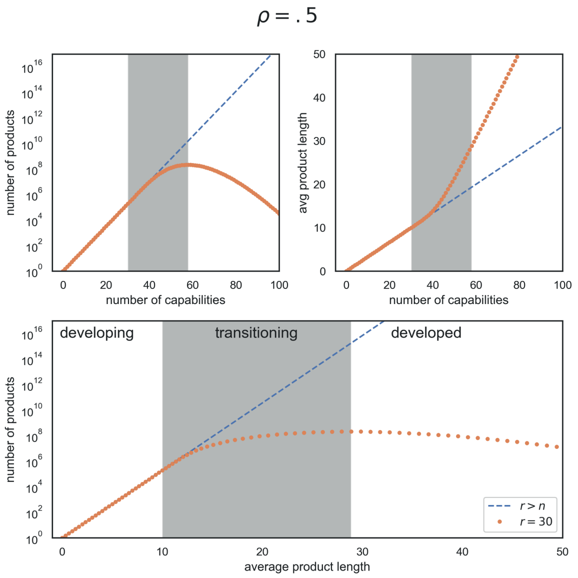


Figure 6.2: The dynamics for the model including a 'recipe book', where every capability is used in a product with probability $\rho = 0.5$. The dashed blue line represent the case where there is no constraint by the product range ($r > n$), and the orange dots show the case where $r = 30$. The grey panel's left border indicates the point where $r = n$, and products start to be lost due to the restricted product range. The right edge of the grey panel indicates the location of the 'hump', i.e. when more products exit than products enter. The top left panel shows how the expected product variety increases and then decreases in n for the constrained case. The top right panel shows the expected average product length, which shows an increase in the rate during the transitioning period. The bottom panel shows the relation between product variety and average product length for the three stages of development.

6.5 Conclusions

Elaborating on the combinatorial framework of economic development proposed by Hausmann and Hidalgo (2011) and Inoua (2016), we have modelled an economy as developing over time by acquiring new capabilities one-by-one. Every new capability is recombined with existing capabilities to allow for the production of an ever greater variety of products with increasing complexity. As long as a country produces every product it can produce given its capabilities, variety increases exponentially with the

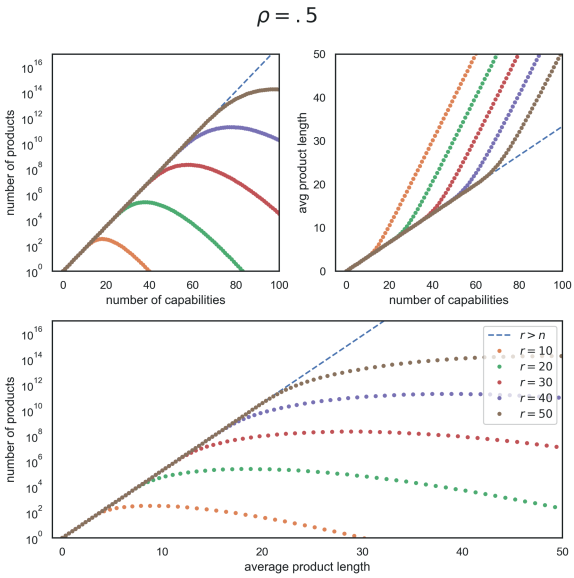


Figure 6.3: The dynamics of the model for different values of product range r. The wider the product range (the larger r), the later the hump occurs.

number of capabilities present, while the average product complexity as measured by the required number of capabilities, only increases linearly with the number of capabilities.

Assuming that there is a maximum range of product complexities that can be made in a country, one is able to recover 'the hump' in variety, which refers to the stylized fact that economies first increase and then decrease their product variety as they develop (Imbs and Wacziarg, 2003; Cadot et al., 2011). The larger the range of product complexities that an economy tolerates to be produced, the longer it takes for the hump to occur. As anticipated by Imbs and Wacziarg (2003), the empirical question that follows holds what country characteristics affect this range, so as to be able to explain why some countries experience the hump earlier in their development than others. For example, the size of a country may be of importance as larger countries may keep on producing low-complexity products for much longer in their low-wage regions compared to small countries where such low-wage regions may be absent. Furthermore, institutional factors including the absence of a minimum wage (prolonging the production of simple products) and trade barriers (preventing the import of simple products from low-wage countries) may further explain a delayed occurrence of the hump.

One objection to the model presented here may regard the way the recipe book is modelled. It was assumed that every capability has an equal chance to be part of any product. This implies that capabilities can be recombined at random to produce viable products. Alternatively, and more realistically, one may choose to impose structural features to the recipe book. For example, particular subsets of capabilities may recombine more easily than other capabilities ('modularity'), and some capabilities may be used more often than other capabilities ('prevalence') (Fink et al., 2017). For the present purposes of our model, a more refined notion of the recipe book is however of minor relevance; for the hump to occur, it only matters whether a recipe book shows a single-peaked distribution of product complexities because the lower bound of the range of tolerated product complexities will then inevitably pass this peak, as the number of capabilities increases.

6.6 Implications

Turning to the burgeoning literature on economic complexity in recent times, our model bears an important implication. We have argued that the relationship between product variety (the number of products) and economic complexity (the average number of capabilities used in products) is highly dependent on the stage of development. During an economy's first stage of development, product variety and economic complexity evolve in tandem with variety increasing exponentially and complexity increasing linearly with the number of capabilities. Hence, one could derive an economy's unobservable economic complexity from the logarithm of the observable product variety (Inoua, 2016). This relationship, however, changes in a transition stage during which the increase in economic complexity accelerates while the increase in product variety slows down, to eventually reach the final stage of diversification during which product variety even starts declining after going through the hump. As the relationship between variety and complexity depends on the stage of an economy's development, empirical attempts to derive economic complexity from product variety are not grounded theoretically by this model. More precisely, following our model, such attempts may only be meaningful for developing countries being in the first stage of diversification. Note that our theoretical argument to fundamentally distinguish between variety and complexity adds to a recent methodological contribution by Mealy et al. (2019) who disentangle the alleged association between variety and complexity in empirical studies measuring economic complexity.

One way ahead in empirical research, then, is to collect direct measurements of complexity from observable characteristics of products. For example, and in line with the notion that more complex products are those that require more capabilities to be produced, one could measure a product's complexity from the number of professions involved in its production. In the context of the model we just presented, a direct measure of complexity is important for two reasons. First, using such a measure, one would be able to verify the assertion that economies drop the simplest products from their portfolio, next to other exit determinants as already investigated (Neffke et al., 2011; Essletzbichler, 2015). While there is some indirect evidence that countries do so (Cadot et al., 2011), we should attempt to verify this empirically. Second, a direct measure of complexity would also be required to further scrutinize the phenomenon of the hump. While our model can replicate the hump as a stylized fact of economic development over time, one is in need of a complexity measure to estimate the exact shape of the hump as a relation between complexity and variety, as predicted by the model.

The question that remains is how we should understand the relationship between economic development (as understood in terms of stages of diversification) and economic growth (as understood in terms of GDP per capita). The notion of capabilities as complementary inputs to produce an output, still needs to be integrated in models of economic growth. Some models that already built from the assumption that output increases with a larger variety of inputs (Romer, 1987; Kremer, 1993), can serve as a starting point for future modelling and empirical validation. In our model, the main focus has been on showing that as countries lose the simplest products from their portfolio, they will experience faster increases in the average complexity of the products that remain. One may be tempted, then, to associate the economic complexity of an economy with GDP per capita if one assumes that the average complexity of products in the economy is reflected in the average wage paid to labour. Following this reasoning, our model would predict that the average wage has accelerated over the past decades, while the opposite has been observed. Our model, however, does not assume that the average wage increases with the average complexity of products. Instead, we assume that with every new capability acquired, a country starts producing more complex products and stops producing the simplest products. The logic of the simplest products existing from a country's portfolio, then, reflects that the minimum wage (not the average wage) in a country keeps on increasing as new capabilities are acquired. Further note that equating the average wage with average complexity of products confuses the variety of products in an economy (extensive margin) with their relative shares (intensive margin). To consider average wage as a proxy for average product complexity would assume that all products have an equal share in the economy as well as in the workforce.

More fundamentally, while economic development can be understood as stemming from the acquisition of new capabilities, there is no reason to believe that new capabilities arrive at a constant rate. For developing countries, the challenge to acquire capabilities may be largely sought in the adoption of capabilities that already exist in the world through channels like imitation, immigration, cooperation and learning. By contrast, countries at the frontier of technological development have to rely on the invention of new capabilities and finding the new combinations with the capabilities they already have (Klinger and Lederman, 2004). Theoretically, then, it is conceivable that the slowdown of growth in high-income countries over the past decades is solely the result of a slowdown in the rate at which new capabilities are acquired and recombined with existing capabilities. This links to observed changes in the return to R&D, which arguably underlie to an important extent the acquisition of new capabilities. Indeed, evidence is mounting that the return on R&D has been declining over the past decades, precisely because of the difficulty to recombine an ever larger number of knowledge domains into new inventions (Jones, 2009; Gordon, 2016). Yet, to fully appreciate this recent finding in the light of the framework of economic complexity, we are in need of direct measures of capabilities to understand the evolution of economic complexity across space and time.

Chapter 7

Vertical vs. horizontal policy in a capabilities model of economic development*

Abstract

Against the background of renewed interest in vertical support policies targeting specific industries or technologies, we investigate the effects of vertical vs. horizontal policies in a combinatorial model of economic development. In the framework we propose, an economy develops by acquiring new capabilities allowing for the production of an ever greater variety of products with an increasing complexity. Innovation policy can aim to expand the number of capabilities (vertical policy) or the ability to combine capabilities (horizontal policy). The model shows that for low-income countries, the two policies are complementary. For high-income countries that are specialised in the most complex products, focusing on horizontal policy only yields the highest returns. We reflect on the model results in the light of the contemporary debate on vertical policy.

*This chapter is available online as a working paper as: van Dam, A. and Frenken, K. Vertical vs. Horizontal Policy in a Capabilities Model of Economic Development. *arxiv.org/abs/2006.04624*, 2020b.

7.1 Introduction

Vertical policy is back on policy agendas globally (Rodrik, 2004; Cimoli et al., 2009; Chang and Andreoni, 2020). Regarding low-income countries, vertical policies comprise a variety of instruments to allow a country to catch up with the global technology frontier. One popular strategy historically has been to temporarily protect 'infant industries' from global competition as to build up knowledge capabilities and institutions required for developing particular technologies or industries (Freeman, 1987; Chang, 2002). A more recent approach has become known as modern industrial policy, and attempts to back entrepreneurs who discover new export industries with complementary public investments (Rodrik, 2004). Evaluation studies confirmed the positive role that vertical policies policies can play in fostering economic development, though success depends on the exact policy design and contextual conditions (Lane, 2020). These insights, combined with the spectacular success of China fully embracing vertical policy, has put vertical policy high on the policy agenda as a strategy for economic development for low-income countries.

Vertical policy is also experiencing popularity in high-income countries (Aiginger, 2007; Aghion et al., 2011). In the light of disappointing growth rates, the effectiveness of horizontal policies is increasingly questioned (Mazzucato, 2011; Mazzucato et al., 2015). In high-income contexts, vertical policy is called for to push the technological frontier itself rather than to catch-up with technologies already developed in other countries. Vertical policies come in different versions and with different labels, including industrial policy, policies for key enabling technologies, smart specialisation, transformative innovation policy, and mission-oriented innovation policy. Though the rationales and instruments tend to differ for each of these policies, but they share a vertical orientation towards supporting only specific industries or technologies (Foray, 2019; Mazzucato, 2018; Bailey et al., 2019).

The renewed interest in vertical innovation policy and the proliferation of new policy concepts has not been matched with new theoretical frameworks. The lack of theorising is in itself not surprising given that innovation and development are complex and elusive phenomena. What is more, economic growth models have long neglected the role of the exact industries or technologies in an economy, and the process of diversification leading to new industries and technologies. However, with the recent advent of a new capability theory of economic growth as developed by Hausmann and others (Hausmann et al., 2007; Hidalgo and Hausmann, 2009; Hausmann and Hidalgo, 2011; Inoua, 2016; Sutton and Trefler, 2016; van Dam and Frenken, 2020a), a new framework has become available to theorise about policy and its effects on economic development. This paper sets out to develop a policy framework based on the capability theory of economic growth as to assess and compare the returns of vertical and horizontal policies.

The capability theory starts from an explicit representation of specific outputs and the inputs required to produce each output. Outputs are generally considered (export) products and inputs as 'capabilities', which include assets, knowledge and skills, but also products-specific regulations and institutions (Lall, 2000; Hidalgo and Hausmann, 2009; Hausmann and Hidalgo, 2011). Economic development stems from diversification into new products made possible by the acquisition of new capabilities. Once acquired, firms start recombining the new capability with existing ones, thus increasing both the variety of products (number of products) and complexity of products (number of capabilities used in each product) in the economy. A vertical policy can be thought of as any policy that targets the acquisition of a particular capability. For example, an industrial policy focusing on aircraft production, would lead to the acquisition of one or more new capabilities, which - combined with already existing capabilities - enable a country to start producing aircraft. Once a new capability is acquired, it can also be used in other recombinations of inputs allowing further diversification into new products. The practical challenge for any vertical policy, then, is to target a capability that can be effectively recombined with existing capabilities as to increase the variety and complexity of an economy (e.g., industrial policy. targeted R&D investment, new teaching programs, selective Foreign Direct Investment, selective migration policy, etc.).

The capability theory of economic development is, in its current form, still a limited framework as it stands on two strong assumptions. First, it assumes that that countries produce every product that their capabilities base would enable them to produce. This assumption is at odds with the common observation that high-income countries lose industries to countries with lower wages over the product lifecycle (Vernon, 1966). If one instead assumes that countries stop producing low-complexity products as the average complexity of their products continues to increase with the acquisition of new capabilities, it can be shown that, over time, the trend of increasing variety changes into a trend of decreasing variety, consistent with the empirical phenomenon of the hump (Cadot et al., 2011; Sutton and Trefler, 2016; van Dam and Frenken, 2020a).

The second strong assumption in capability models is that countries would not face any limitation in being able to recombine capabilities. Put differently, it views countries as having unlimited abilities to effectively coordinate any number of capabilities required for a product. It follows from this assumption that the only objective for a policy maker would be to acquire new capabilities. If so, the policy question boils down to selecting which capabilities should be acquired and in what manner. Once one would relax this assumption and would view countries as facing constraints in the complexity of products that their firms are able to make, a more fundamental policy question arises: how much effort should a country put on acquiring a specific new capability vs. how much effort should it put on learning how to make more complex products from the capabilities already present. It is the latter policy that we will consider as a horizontal policy, which aims to increase the ability of a country to produce more complex products. Here, horizontal policy refers all policies that improve the coordination and integration of capabilities required for the production of products (e.g., basic research, public research organizations, standardization institutes, public consultation schemes, collaboration subsidies, generic social and managerial skills, laws, and regulations), similar to what has been referred to as a country's 'national innovation system' (Freeman, 1987; Lundvall, 1992).

It follows that policy for economic development can be understood as a combination of two policies: a vertical policy focusing on acquiring a new capability providing a country with opportunities to produce a larger variety of products, and a horizontal policy focusing on improving a country's ability to recombine capabilities in ever more complex products. Given these two types of policies, the question then becomes how to allocate their efforts on one or the other policy. Intuitively, one may expect the two policies to be complementary: the combinatorial logic of products stemming from combinations of capabilities implies that the ability to recombine capabilities is most valuable for countries that already have many capabilities.

Building on previous combinatorial models of economic development (Hausmann and Hidalgo, 2011; Inoua, 2016; van Dam and Frenken, 2020a), we propose a model in which we conceive of economic development as the outcome of increases in the number of capabilities residing in a country and of improvements in the ability to recombine capabilities in a country. National government decides, at each time step, whether to increase the number of capabilities in a country or to improve the ability to recombine capabilities. This decision depends on the expected increase in the average complexity of products. This basic model set-up will explain the complementarity between vertical and horizontal policy. We then turn to our extended model by introducing a minimum wage that bounds the minimum complexity of products produced in a country. As a country enters more complex products, it increases its minimum wage and abandons its products with lowest complexity. This extension of the model leads to three further contributions. First, the resulting model reproduces the stylised fact of the hump. Second, it explains the growing importance of horizontal policies over vertical policies as economies develop over time. Third, it can localise the shift in optimal policy close to the hump, suggesting that high-income countries should focus on horizontal rather than vertical policies.

7.2 The Model

Our understanding of economic development has long been guided by the notion of a production function that specifies how inputs such as capital and labor translate into the total output of an economy. More recently, models are more explicit about the products produced in an economy. At the level of products, inputs can be considered to be strictly complementary (Kremer, 1993; Hausmann and Hidalgo, 2011; Brummitt et al., 2017). This assumption is based on the idea that the production of any product or service requires a particular combination of complementary inputs.

Inputs required to produce a product have been referred to as 'capabilities' (Hidalgo and Hausmann, 2009; Hausmann and Hidalgo, 2011). Following this reasoning, the ability of an economy to produce a product depends on the capabilities present in a country. Developing new products consists of recombining old and new inputs into configurations that have economic value (Inoua, 2016). It also follows that with the acquisition of a new capability, the variety of products that a country can produce grows in a non-linear fashion. An elementary model of this kind is that each possible combination of capabilities results in one unique product. The total number of products that a country can make is then given by summing the number of possible combinations of a given length that can be made out of n capabiltiies over all possible lengths s:

$$d(n) = \sum_{s=0}^{n} \binom{n}{s} = 2^n.$$

The average complexity of products is given by the total length of all products divided by the total number of products:

$$\bar{s}(n) = \frac{\sum_{s=0}^{n} s\binom{n}{s}}{2^n} = \frac{n}{2}.$$

The assumption that any combination of capabilities leads to a viable product is arguably too strong. Instead, one can safely assume that only some combinations of capabilities lead to meaningful products. The set of combinations of capabilities resulting in meaningful products has been referred to as a 'recipe book', which describes the combinations of capabilities that are complementary in that they lead to viable products (Hausmann and Hidalgo, 2011; Inoua, 2016; Fink et al., 2017).

The model can be generalized by assuming that every capability is part of a viable product with a given probability ρ (Inoua, 2016; van Dam and Frenken, 2020a). A combination of s capabilities then has probability ρ^s of representing a viable product of length s. Hence, it becomes increasingly unlikely that a combination of capabilities leads to a viable product as more capabilities are added, since ρ^s is decreasing in s when $\rho < 1$. For $\rho = 1$, we recover the initial simple model described above.

Since there are $\binom{n}{s}$ possible combinations of s components one can make from the total of n components, and each combination of length s has probability ρ^s of being viable, the expected number of products of length s a country with n components can make is given by $d(n, s) = \binom{n}{s}\rho^s$. Summing this quantity over all product lengths s gives the expected variety of products that can be made with n components

$$d(n) = \sum_{s=0}^{n} \binom{n}{s}\rho^s = (1 + \rho)^n.$$

Since the share of products of length s in a country is given by $\frac{\binom{n}{s}\rho^s}{d(n)}$, the expected average complexity given n components can be computed as (Inoua, 2016; van Dam and Frenken, 2020a)

$$\bar{s}(n) = \sum_{s=0}^{n} s \frac{\binom{n}{s}\rho^s}{d(n)} = \frac{\rho}{1+\rho} n. \tag{7.1}$$

Note that while variety increases exponentially with n, complexity increases only linearly with n. The rate of increase in product complexity viz. economic growth is solely determined by the difficulty parameter ρ.

7.3 Vertical vs. horizontal policy

The key assumption in the combinatorial model, albeit an implicit one, holds that a country can recombine any number of capabilities. That is, the sole challenge for a country is to acquire additional capabilities, leading to an increase in n, which automatically translates into a stable growth path in the form of a linear increase in average product complexity.

Dropping the assumption that countries can recombine any number of capabilities, we introduce the parameter l referring to the maximum length of products that a country is able to produce. The expected product variety and product complexity are then given by

$$\begin{aligned} d(n,l) &= \sum_{s=0}^{l} \binom{n}{s} \rho^s \\ \bar{s}(n,l) &= \frac{\sum_{s=0}^{l} s\binom{n}{s}\rho^s}{d(n,l)}, \end{aligned}$$

respectively. Figure 7.1 shows how a constraint on the maximum complexity of products, as expressed by l, hampers economic development as product variety (left) grows less than exponentially and average product complexity (right) reaches a ceiling converging asymptotically to l with n approaching infinity.

A policy maker now has two options to foster economic development. First, (s)he can increase the number of capabilities n, to which we refer as vertical policy. We model the decision to increase the number of capabilities as a unit increase in n. The second option is to improve the country's ability to recombine capabilities, to which we refer

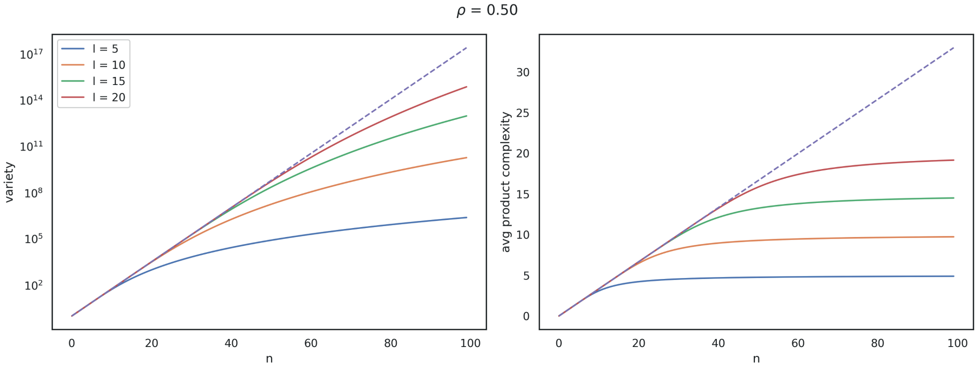


Figure 7.1: Limits to coordination

to as horizontal policy. We model a decision to increase the ability to recombine capabilities as a unit increase in l.

Note at this point that our model remains agnostic about the specific type of vertical policy that is being employed. Rather, we model vertical policy as any policy that leads to some new capability that has random complementarities with already existing capabilities. In this sense, vertical policies are targeted only in the sense that they lead to one new capability, but blind with regard to the exact complementarities that can be exploited between the new capability and the already existing ones.

Starting with the initial condition in a country with n=1 and l=1, the policy maker alternates between the two policies depending on which policy yields the highest expected increase in average complexity. For vertical policy, the expected gain in average product complexity from adding a capability is given by

$$\begin{aligned}\frac{\Delta \bar{s}}{\Delta n} &= \bar{s}(n+1,l) - \bar{s}(n,l)\\ &= \frac{\sum_{s=0}^{l} s\binom{n+1}{s}\rho^s}{d(n+1,l)} - \frac{\sum_{s=0}^{l} s\binom{n}{s}\rho^s}{d(n,l)}.\end{aligned}$$

For a horizontal policy, i.e. increasing l, the gain in average product complexity is given by

$$\begin{aligned}\frac{\Delta\bar{s}}{\Delta l} &= \bar{s}(n, l+1) - \bar{s}(n, l)\\ &= \frac{\sum_{s=0}^{l+1} s\binom{n}{s}\rho^s}{d(n, l+1)} - \frac{\sum_{s=0}^{l} s\binom{n}{s}\rho^s}{d(n, l)}.\end{aligned}$$

At any given stage in the development process (characterized by n and l), a policymaker chooses for vertical policy when $\frac{\Delta\bar{s}}{\Delta n} > \frac{\Delta\bar{s}}{\Delta l}$, and for horizontal policy when $\frac{\Delta\bar{s}}{\Delta l} > \frac{\Delta\bar{s}}{\Delta n}$.

Following this policy decision model, we simulated the evolution of product variety and average product complexity over time (upper left and middle left panel in Figure 7.2) as well as the incidence rates of vertical policy and horizontal policy (lower left panel in Figure 7.2). The two policies are clearly complementary as vertical policies (increasing n) are alternated by horizontal policies (increasing l) as to leverage the increased potential to make more complex products due to the recent rise in capabilities. We further observe that the exact incidence rates of both policies are sensitive to ρ (compare lower left panel of Figures 7.2, 7.3 and 7.4).

7.4 Full model

While our model explains the complementarity between vertical policy and horizontal policy, it falls short in reproducing the the inverted-U shape relationship between income per capita and product variety commonly known as 'the hump'. In terms of economic development, this pattern indicates that countries first diversify and then, at some level of income, start specialising again (Imbs and Wacziarg, 2003; Cadot et al., 2011).

In our combinatorial framework, the hump can be understood as resulting from low-complexity products exiting a country's portfolio as a country continues to diversify into high-complexity products (van Dam and Frenken, 2020a). Labour involved in low-complexity products arguably has lower productivity, resulting in lower wages, than labour involved in high-complexity products. Economic development leading to products with higher complexity will thus push the highest wages in a country upwards. Assuming minimum wages to increase with maximum wages, a country cannot remain competitive in low-complexity products and will lose these product to low-wage countries.

Implementing such a mechanism of product exit in our model, we assume that countries only produce products with a complexity in the range of [l - r, l], where $r > 0$. This range is based on the idea that given the minimum and maximum wage in a country, it can only be competitive in a certain range of product complexities. It follows that once $l > r$, a country starts abandoning products with the lowest complexity from its portfolio.

The second to fifth columns in Figure 7.2 show the results when we re-run the baseline model (shown in the first column), but now including parameter r = 25, r = 20, r = 10 and r = 1 respectively. Figures 7.3 and 7.4 show the same results, but now for $\rho = 0.25$ and $\rho = 0.75$.

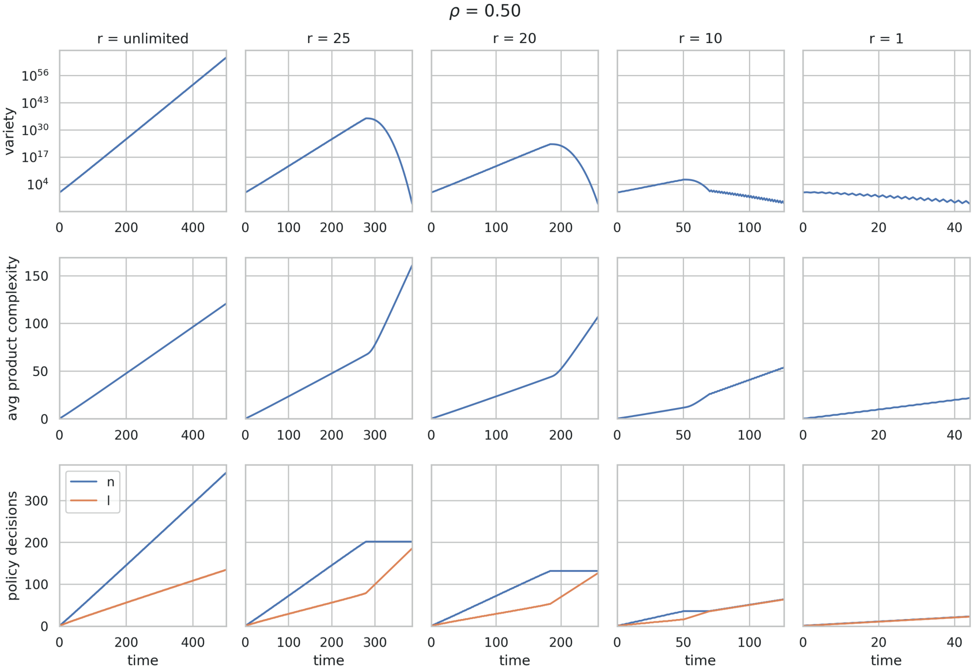


Figure 7.2: Model results for $\rho = .5$.

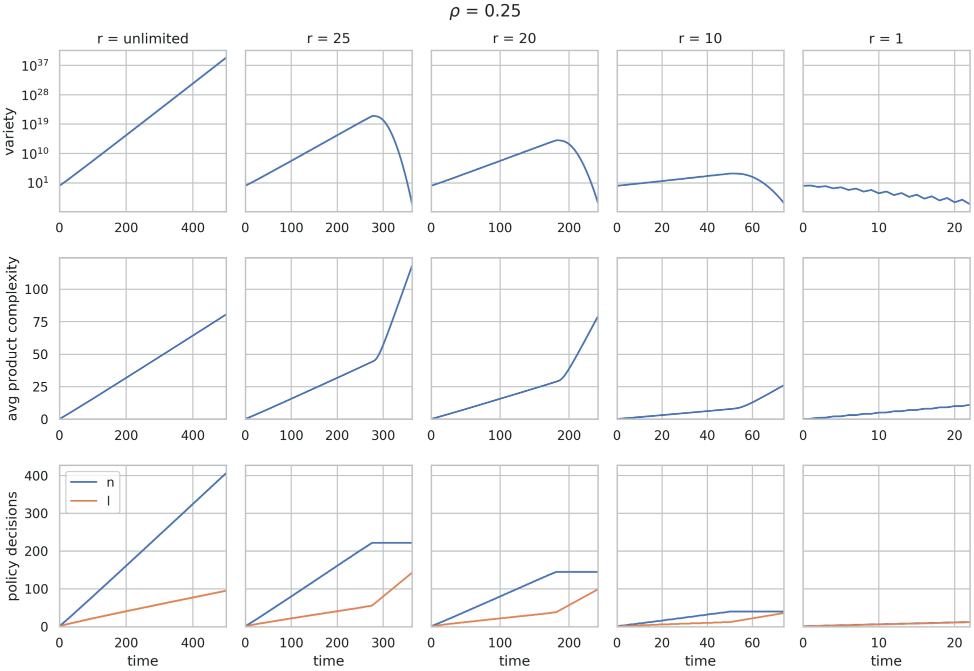


Figure 7.3: Model results for $\rho = .25$.

Three observations can be made. First, the model reproduces the hump for non-trivial values of r ($r = 25$, $r = 20$, $r = 10$), as can be seen in the first row of each figure. This is consistent with the empirical phenomenon of the 'hump' (Imbs and Wacziarg, 2003; Cadot et al., 2011), and reproduces the theoretical result of a similar capability model by van Dam and Frenken (2020a).

The second observation to be made is that product complexity starts accelerating once the variety of a country starts decreasing. Once $l > r$, a horizontal policy will then push complexity upwards in two ways: the policy increases the maximum product complexity that a country is able to produce (l) by one and it increases the minimum product complexity that is being produced in a country ($l - r$) by one.

The final observation to be made holds that during the 'hump-period', the optimal policy is to focus solely on horizontal policies, which maximizes the increase in product complexity. Such policy leverages the high number of capabilities already present

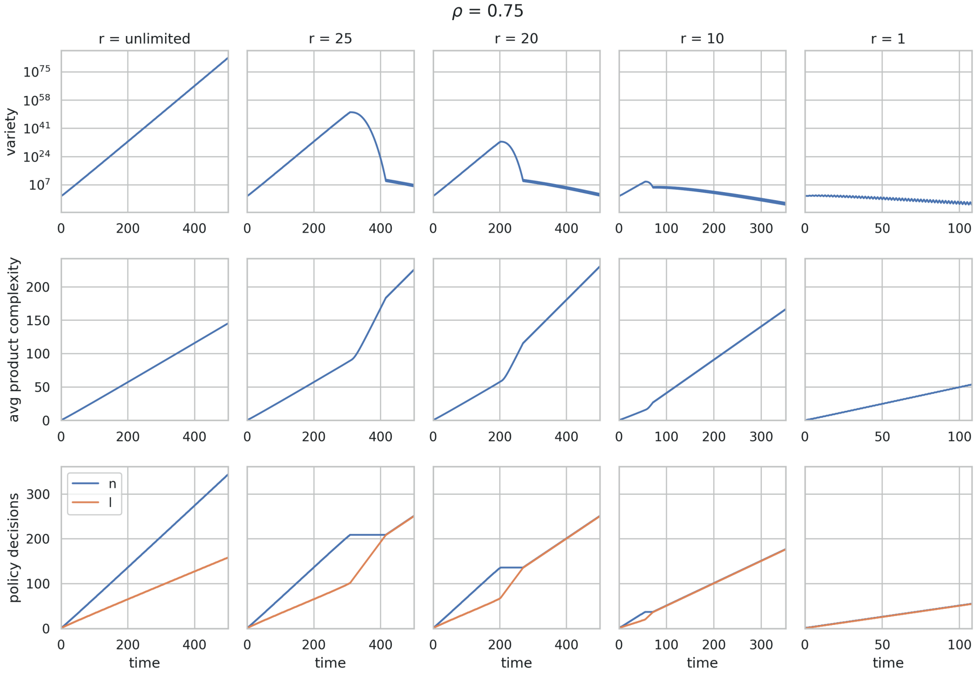


Figure 7.4: Model results for $\rho = ..75$.

by improving a country's ability to recombine its capabilities in ever more complex products. This process continues until $l = n$, reflecting a most 'advanced' economy producing solely the most complex products within the range $[n - r, n]$. It is also at this stage that vertical policy becomes relevant again next to horizontal policy, as further progress can only be reached by alternating between adding a capability and increasing maximum complexity. Importantly, the focus on horizontal policy in the hump-period is robust for different values of parameters ρ and r. And, as the hump phenomenon historically tends to occur only at a certain levels of income per capita, our model can pinpoint the countries that, on theoretical grounds, could benefit most from focusing on horizontal policies (the hump tends to occur at around 24,000 US Dollar (PPP in constant 2000) (Cadot et al., 2013)).

7.5 Discussion

Elaborating on the capabilities framework of economic development proposed by Hausmann and Hidalgo (2011), Inoua (2016) and van Dam and Frenken (2020a), we have modelled an economy as developing over time by acquiring new capabilities one-by-one. Every new capability can, with some probability, be recombined with existing capabilities to allow for the production of an ever greater product variety and product complexity. Different from previous models, however, we pose that countries are potentially constrained in the level of product complexity they can handle, due to an under-investment in basic research, managerial skills and an underdeveloped ’innovation system’.

It follows from our model that public policy can focus on two development strategies: the addition of a new capability, which we refer to as vertical policy, or an improvement of a country’s generic ability to recombine capabilities, which we refer to as horizontal policy. The key result that we draw from the model is that for low-income countries, vertical policy focused on capability acquisition is to be complemented with horizontal policy so the increasing number of capabilities can be effectively recombined in more valuable products. A second insight holds that once a country starts abandoning low-complexity products from its portfolio, horizontal policy becomes even more important. In this stage, a country loses competitiveness in relatively simple products, and needs to focus on mastering the coordination of the large number of capabilities required for the production of more complex products.

Our model is flexible in that other policies can be simulated as well. Our choice for vertical policy as the addition of one new capability and horizontal policy as the unit improvement of maximum product complexity are ideal-types of vertical and horizontal policies, respectively. In between the two policies, one can put hybrid policies. Two such policies follow naturally from our model.

First, rather than viewing vertical policy as the addition of some random capability, one could further specify a vertical policy as one that specifically targets a capability that, following our model, can be recombined with already existing capabilities in ways that would maximize the increase in average product complexity in the economy. For low-income countries with few capabilities, the targeting of such capabilities may be relatively easy to gauge as the increase in the number of new recombinations resulting from one new capability, is still rather limited. For high-income countries with many capabilities, such a targeted vertical policy may be harder to determine. Yet, the underlying idea of targeting capabilities that can be recombined in many and complex ways clearly speaks to the logic of focusing on ’general purpose technologies’ (as the term suggests) (Bresnahan and Trajtenberg, 1995).

Second, rather than viewing horizontal policy as a unit increase in the maximum product complexity that an economy can produce, one can imagine a more hybrid policy in which a government seeks to improve the maximum product complexity only in a certain broad sector (like healthcare, mobility, agriculture, etc.). In the model, sectors would correspond to a subset of products that would fall within sectoral boundaries. This would mean that horizontal policies can be made more specific to coordination challenges in certain sectoral contexts rather than across the board. Such policies remain horizontal in nature, but targeted in their scope. In particular, a policy maker would wish to target those sectors for which many relevant capabilities are already present, but which fail to leverage those capabilities in complex product due to present limits to coordination failures. This type of policy has also been discussed in the innovation policy literature under the heading of ’systemic policy’ (Smits and Kuhlmann, 2004; Wieczorek and Hekkert, 2012).

Finally, turning to the revival of industrial policy as a form of vertical policy, our model provides both support and a critique to industrial policy as a means to spur economic development. For low-income countries, there is a strong rationale for industrial policy as to increase their capability base. For such countries, focusing only on improving the ability to coordinate many capabilities makes little sense as long the number of capabilities present is still low. For high-income countries, however, the rationale for modern vertical policy is less obvious. As such countries can only compete on complex products with high value-added, the main challenge for these countries is to improve the ability to produce more complex products from the large set of capabilities that they already master. Here, horizontal policies alone could be, theoretically, sufficient to continue economic development. The exact distinction between policies for low-income countries and high-income countries could be determined empirically by looking at the inverted-U patterns between average income and product variety (with maximal variety located around 24,000 US Dollar (PPP in constant 2000) (Cadot et al., 2013)). As countries go through this ‘hump’, they should start focusing more on horizontal policies.

In this light, the plea for industrial policy in the context of high-income countries, and equally for technology missions (Mazzucato, 2011; Mazzucato et al., 2015), needs more grounding. If such missions are articulated in terms of the alleged need to master a specific new technology domain or industry, our model would cast doubt about its effects on growth. While a new technological capability could indeed be beneficial for growth, it will generate little comparative advantage if actors within the innovation system are not able to combine and integrate the new capability with the existing set of capabilities, including complementary technologies, skills and institutions.

Chapter 8

Conclusions

8.1 Overview

The starting point of this thesis was a view of economic development based on capabilities. This framework focuses on the heterogeneous and complementary nature of economic inputs, and its consequences for economic development. This lead to three central concepts: diversity, complexity and relatedness.

The thesis was built up in three parts. The first part consisted of Chapter 2, which provided an overview of the current literature on related variety, economic complexity and related diversification. It reviewed the role of diversity in each of these literatures, as well as the methodologies used to capture it. It also proposed new methodology to test the hypotheses that follow from the theory.

The second part of this thesis focused purely on methodology. Chapter 3 discussed the measurement of diversity using Hill numbers, and proposed a new measure of diversity that takes into account disparity. It measures diversity as the 'number of compositional units', given by the exponential of mutual information. It differs from current measures of diversity in that it takes into account disparity based on overlap of features of the whole set, as opposed to only the pairwise similarities. The chapter also proposed a decomposition of diversity into its separate components of variety, balance and disparity.

Chapter 4 reviewed the economic complexity indices (Hidalgo and Hausmann, 2009) as a statistical technique called correspondence analysis (CA). It showed multiple ways in which CA can be derived: as a way of ranking nodes in a bipartite network based on some latent feature (known as 'ordination' in Ecology), as a clustering algorithm applied to the network of similarities between nodes, and as a graph embedding technique, leading to a low-dimensional representation of the similarity network in a

Euclidean space. Each of these derivations leads to a different interpretation of the complexity indices, and also show how higher order eigenvectors and eigenvalues can be interpreted. As of yet, these have not been considered in the literature on economic complexity.

Chapter 5 introduced an information-theoretic framework for the measurement of relatedness between economic activities based on co-location. The framework provides a formal approach of inferring associations between economic activities, and leads to measures of (co-)location, specialization and localization. A key feature of the framework is that it is able to relate a number of widely used indices used in the literature that where hitherto considered unrelated. The framework allows for uncertainty estimates for each of the measures, and can be readily extended to multivariate analyses.

The third part of the thesis focused on capturing the capabilities framework in a simple theoretical model, building on the work of Inoua (2016). Chapter 6 extended the model to include product exit, so as to make it consistent with the stylized fact known as 'the hump'. This was done by introducing a parameter that represents the range of product complexities that a country can be competitive in, arguing that countries with high wages cannot be competitive in low-complexity products.

Chapter 7 further extended the model by introducing a 'coordination constraint', arguing that the ability of a country to combine large number of capabilities into complex products may be constrained. A policy maker can then choose to focus efforts towards adding capabilities (vertical policy), or enabling the recombination of larger numbers of capabilities in order to make more complex products (horizontal policy). While for developing countries the two policies are complementary, the model shows that developed countries may benefit more from horizontal policies than from vertical policies.

8.2 A methodological framework - connecting the dots

The methodological chapters in this thesis share a feature: the object of analysis is an incidence matrix (or contingency table) describing the frequency or intensity of one variable with respect to another. Such data can be thought of as a bipartite network, where the nodes in each layer represent possible values each variable can take. In the capabilities framework, the main object of study is the incidence matrix of economies and capabilities. However, since the capabilities are typically unobservable, they have to be studied indirectly, either by assessing the incidence matrix of economies and the activities that take place in them (for example the occupation-city matrix in Chapter 5 or the country-product matrix in Chapter 4), or the incidence matrix of economic activities and some proxy for the capabilities they use (such as the industry-occupation matrix in Chapter 3).

The study of incidence matrices is common in many fields of science: ecologists study species occurrences in sites, biologists study the expression of genes in samples, linguists study occurrences of words in documents, and psychologists study the presence of attributes in people, to name a few. Thus although each of the methodological chapters in this thesis can be read with a capabilities model in the back of the mind, the methods discussed are applicable in a much broader context.

Naturally, much can be learned from other disciplines, as exemplified in this thesis: the pointwise mutual information in Chapter 5 was already used in Linguistics, the framework of Hill numbers in Chapter 3 originated in Ecology, and the different interpretations of the complexity indices in Chapter 4 are drawn from work in statistics, spectral graph theory and computer science.

Chapter 5 takes a probabilistic approach to the challenge of extracting information from incidence matrices. This was done by assuming the data was generated by a multinomial process, and estimating the underlying probabilities. These probabilities can then be studied using tools from information theory, leading to measures of association and dependence. This approach has some attractive properties: it is consistent under aggregation, it naturally extends to multivariate analyses, and provides uncertainty estimates for each quantity. In the following, I show how the measures of diversity in Chapter 3 and correspondence analysis discussed in Chapter 4 relate to the information theoretic framework in Chapter 5, with the aim to reconcile all methodologies within a single encompassing framework. Relating these concepts methodologically may lead to a better understanding of the conceptual and empirical relations between them.

8.2.1 Diversity and entropy

Chapter 3 built on the framework of Hill numbers, which defines diversity as the inverse of the average 'commonality' of elements considered. If elements are on average rare, the diversity is high. Interestingly, there is a close relation between Hill numbers and information-theoretic measures. In particular, the Hill number (measuring the effective number of species) is given by the exponential of the Shannon entropy of the distribution of elements under consideration. Diversity and entropy are thus related by a simple transformation. Despite this close relation, this does not mean that entropy and diversity are the same. Their relation is described by Jost as:

> "Diversity is not meaningless but has been confounded with the indices used to measure it; a diversity index is not necessarily itself a "diversity". The radius of a sphere is an index of its volume but is not itself the volume, and using the radius in place of the volume in engineering equations will give dangerously misleading results ... Entropies are reasonable indices of diversity, but this is no reason to claim that entropy *is* diversity." (Jost, 2006, p. 363)

Both entropy and diversity are a function of the multinomial probabilities that can be estimated for one variable from the incidence matrix. Although they can be inferred from each other, they each describe a different property.

Chapter 3 shows that taking into account disparity (or relatedness) between the elements in terms of their overlap with some other variable (representing features) leads to a measure of diversity in terms of the 'number of compositional units', given by the exponential of the mutual information of the two variables under consideration. Note that this mutual information also appeared in 5 as a measure of overall specialization. Hence, the number of compositional units and overall specialization as conceptualized in Chapter 5 are also related by a simple transformation. The number of compositional units is exactly the exponential of the overall specialization. When considering activities and their capabiltities, the more specialized each activity is (in terms of the capabilities it uses), the higher the disparity, leading to a high number of compositional units.

Hence, the diversity measures following from Hill numbers relate directly to the proposed information-theoretic framework, and suggest a relation between specialization and diversity. It also follows that uncertainty estimates for the diversity measures, although not given explicitely in Chapter 3, can be obtained through the Bayesian approach taken in Chapter 5.

8.2.2 Correspondence analysis and associations

Chapter 4 discussed correspondence analysis (CA) and its relation to the economic complexity index (ECI). Although the empirical example in the chapter focused on the analysis of countries based on the products they export, note that it can equally well be applied to study the products based on which countries exports them. This connects the concept of the product space (the product-product similarity network) to the product complexity index (PCI): the PCI is a one-dimensional representation of (a slight variation of) the product space. Such interpretations show that the complexity indices are not a measure of complexity or 'generalized diversity' but rather of similarity. Despite this deviation from the original narrative behind the complexity indices, CA is a tool that was developed exactly for the analysis of aforementioned incidence matrices, rendering it a useful tool for studying economic data in the context of the capabilities framework.

How does CA fit into the information-theoretic framework? To see the connection, note that CA consists of a particular way of decomposing a particular type of similarity matrix. Analyzing the rows of an incidence matrix A using CA yields a weighted eigenvector decomposition of the similarity matrix

$$\tilde{S} = D_r^{-1} A D_c^{-1} A^T D_r^{-1} = V \Lambda V^T,$$

where $V^T D_r V = I$, and the columns of V contain the CA axes of the rows of A. Here, D_r and D_c are diagonal matrices with the row and column sums of A on their diagonal, respectively (see Chapter 4).[1] The entries of the matrix $\tilde{S_r}$ can be rewritten

[1] Likewise, analyzing the column of A leads to

$$\tilde{S} = D_c^{-1} A^T D_r^{-1} A D_c^{-1} = U \Lambda U^T,$$

with $U^T D_c U = I$.

using a notation in terms of probabilities (taking maximum likelihood estimates) as:

$$\tilde{S}_{ii'} = \frac{1}{\sum_j A_{ij}} \frac{1}{\sum_j A_{i'j}} \sum_j \frac{A_{ij} A_{i'j}}{\sum_i A_{ij}} = n \frac{p_{ii'}}{p_i p_{i'}},$$

where $n = \sum_{ij} A_{ij}$, $p_{ij} = \frac{A_{ij}}{n}$, and $p_i = \frac{\sum_j A_{ij}}{n}$. Hence, CA can be seen as a decomposition of the matrix containing the ratio's of joint 'co-location' probabilities $p_{ii'}$ and the marginal probabilities p_i and $p_{i'}$.

Alternatively, the matrices U and V can be obtained directly from A through a weighted singular value decomposition of the matrix (Greenacre, 1984)

$$\tilde{A} = D_r^{-1} A D_c^{-1} = V \Sigma U^T,$$

where $\Sigma^2 = \Lambda$ and $V^T D_r V = U^T D_c U = I$. Writing this in terms of probabilities yields

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sum_j A_{ij}} \sum_i A_{ij} = n \frac{p_{ij}}{p_i p_j},$$

which holds exactly the index of revealed comparative advantage (RCA), up to a factor n, discussed in Chapter 5.

CA can thus be retrieved as a decomposition of either the matrix containing the ratios between co-location probabilities and their marginals, or directly as a decomposition of the matrix containing the RCA's. Note that these matrices are exactly the basis of the association measures in Chapter 5: normalizing by n and taking the logarithm yields

$$\log(\tilde{S}_r/n) = PMI(p_{ii'})$$
$$\log(\tilde{A}/n) = PMI(p_{ij}),$$

A direct connection to the information-theoretic framework can thus be made by adapting the CA framework to include a logarithmic transformation of these matrices.

Performing CA with log-transformed matrices has been explored in (Greenacre and Lewi, 2009; Greenacre, 2009) and is known as log-ratio analysis or spectral analysis. These methods provide an opportunity to connect to the proposed measures of diversity and (co-)location. An added practical advantage of the logarithmic transformation is that it may be able to deal with the skewed distributions typically found in economic data, and that it yields results that are consistent under aggregation. Adapting the CA framework to analyze data within the proposed probabilistic framework may help in exploring and visualizing the structure underlying the associations obtained through application of PMI.

Interestingly, in the definition of ECI and PCI, the incidence matrix A is taken to be the matrix that is binarized based on the RCA index, which is justified theoretically through the concept of comparative advantage (Hidalgo and Hausmann, 2009). Applying CA to this matrix takes the RCA of these binarized values yet again, so that the complexity indices essentially analyze a matrix who's RCA is taken twice. It is unclear whether this double normalization has a deeper underlying theoretical meaning. Other ways of pre-processing the data (e.g. by taking logarithms) may lead to different results, and be better justified from a methodological point of view. Thorough understanding of which similarities are being decomposed exactly may help in exploring the empirical relation between economic performance and ECI that has motivated much of the empirical literature. Up to now, only a very specific combination of pre-processing of the data, the type of similarities, and type of decomposition have been considered in these type of analyses.

8.2.3 Multivariate dependencies

The proposed information-theoretic framework has a key advantage over currently used methods: it provides a principled way to extend analyses to multiple variables, for example moving from an incidence matrix describing two variables to a multi-way table descrbing three variables. Information theory then allows to study the multiway associations between the three variables. This allows for instance analysis of data containing the industry, occupation and educational profile of a group of individuals, and quantifying associations between each variable. Section 3.4 contains a discussion in this direction in the context of measuring diversity. The diversity measures then allow to assess the diversity of industries taking into account the disparity in terms of occupations they employ, the educational profiles they employ, or both.

A special case of these multivariate extensions emerges when dealing with different levels of aggregation, such as within hierarchical classifications or along different geographical scales. Since information theory is built in such a way that it is consistent under aggregation, it can be a useful tool either to exploit the classification structure (as in the measurement of related variety), or to find levels of aggregation (either geographically or classification based) that hold relevant information.

A second example for which multivariate extensions are useful is the analysis of co-location of multiple activities. Up to now, studies in relatedness have been limited to quantifying relatedness between pairs of economic activity. For example, Chapter 5 considered the associations between pairs of occupations based on their co-location in cities. However, one can also consider co-location of triplets or larger sets. Taking expectation of these associations leads to system-level variables that take the form of multivariate mutual information. Such measures could for example shed light on questions regarding the evolution of team sizes and the division of labor.

A final note on multivariate analysis holds that CA also readily extends to multivariate data, which is known as multiple correspondence analysis. Application of these methods is yet to be considered in the current context.

8.3 Capturing capabilities

I now turn to building blocks underlying the work discussed in this thesis: capabilties. As we have seen, in the current context the concepts of diversity, complexity and relatedness all depend crucially on distribution of capabilities underlying economic activities. Despite their importance however, their exact definition and measurement remains elusive, as they are generally not observable. There are rougly two ways to tackle this issue: either analyze data that gives a direct proxy for capabilities, or use algorithm that infer them from data.

8.3.1 Proxy for capabilites

Several attempts have been done to find suitable proxies for capabilities. An example is given by the incidence matrix of industries and occupations, where occupations represent bundles of skills or tacit knowledge that can be considered to be the necessary inputs for any given industry. This can be taken even further by using data on the skills and activities that occupations represent, and using these as a proxy for capabilities.

Such approaches are challenging however as they require both a specific dataset and a specific definition of a capability, making results hard to generalize. They further neglect one of the central ideas of the capabilities approach, namely that capabilities represent a heterogeneous set of inputs, ranging from tangible assets to tacit knowledge and the right institutional conditions.

Altneratively, one can find proxies not for capabilities but for the complexity of economic activities (i.e. the number of capabilities embodied in an economic activity). This can be done, for example, by defining the complexity of a product as the average income of countries it is produced in (Lall et al., 2006; Hausmann et al., 2007). Other scholar have used quantities like average team size or the average years of schooling of people involved in an industry as a proxy for its complexity (Balland et al., 2020). Other possible proxies for the complexity of products include skill intensity, required on-the-job-learning, intermediate inputs or non-routine content of jobs involved in their production (Schetter, 2020).

Use of such proxies allow to establish stylized facts, and may serve as a 'sanity check' for the implications of the capabilities framework and the concept of complexity itself. However, it requires strong assumptions and may give an incomplete picture of complexity. Moreover, it does not allow for studying how capabilities get recombined into products, and thus cannot be used for studying the direct connections between complexity, relatedness and diversity.

8.3.2 Inferring capabilities

An alternative to using proxies is to infer (counts of) capabilities from data, as was the objective of the complexity indices proposed in (Hidalgo and Hausmann, 2009). The capabilities can be considered a 'latent layer' that connects economic activities to the locations in which they take place. In Chapter 4 however we have seen that the economic complexity index reflect similarities rather than the diversity of capabilities, leaving us without a measure of complexity. Are there alternatives to infer the underlying capabilities from economic data? This type of inference is the domain of computer science, machine learning and network science, in which there are many methods available to perform latent variable analysis or dimensionality reduction.

A group of methods that may be particularly promising in this respect is topic modeling (Blei et al., 2003). These algorithms are used in natural language processing to assign topics to documents according to the words that occur in them. They estimate for each document a distribution over topics, and for each topic a distribution over words that it consist of. One can then think of the words in a document as being generated by the topics that that document is associated with.

Replacing document with economies, and words with economic activities, the estimated topics are readily interpreted as capabilities. The methods estimate directly the distribution of capabilities in each location, and the capability requirement of each activity. These distributions can then be analyzed using exactly the type of information-theoretic measures proposed in this thesis, enabling to quantify relatedness and diversity based directly on the inferred capabilities. These methods thus thus fit perfectly in the methodological framework proposed in this thesis.

Recent advancements further extend the original topic modeling algorithms with more suitable priors, and draws direct connections with stochastic block models and community detection in bipartite networks (Gerlach et al., 2018). The applicability of these type of methods to economic data remains to explored, but may provide a promising way forward for inferring capabilitites using a method that stays close to the capabilities framework, yielding interpretable results.

8.4 The capabilities framework - extending the model

Turning back from methodology to the capabilities model, I discuss the extensions of the model proposed in Chapter 6 and Chapter 7. In Chapter 6, we extended the model to incorporate the exit of low complexity products as countries develop. This yielded a diversification pattern consistent with the stylized fact known as 'the hump'. Chapter 7 focused on policy, and distinguished vertical from horizontal policies. Vertical policies aim at adding a specific capability, while horizontal policies aim at increasing the ability to recombine capabilities, increasing the maximum product length. Here I address some shortcomings of the model, and propose possible alternatives to extensions of the capabilities framework.

8.4.1 Modeling product exit

In Chapter 6, we incorporated product exit by arguing that as countries develop, the average income goes up so that they can no longer be competitive in low complexity products. This relaxes the assumption that countries produce all products they possibly can given their set of capabilities. It also introduces a new dimension to the model: the competition between countries and the wage structure within a country. The model investigated in this thesis is a single country model, neglecting interactions between countries. Exploring what drives the production of certain products in countries next to or on top of the capabilities structure, and how this is affected by interactions between countries is an open issue and will require taking into account interactions between countries. This leads to possible connections between more classic trade models and the economic complexity framework (Schetter, 2019, 2020).

Sticking to the simple model proposed here, a simple question one can ask is what determines the range of complexities r that a society can produce. It may depend, for instance, on the type of society in a country. Large, heterogeneous societies or countries with large inequalities for example may have a larger range, leading to a more diversified production structure.

8.4.2 Adding structure to the recipe book

The simple capabilities model explains the relation between average product complexity and variety given the number of capabilities in an economy. Due to the underlying combinatorial logic, it is also consistent with the concept of related diversification. However, it fails to capture the path-dependent nature of development since the 'recipe book' that dictates which capabilities are used by which products has a very simple structure (i.e. capabilities are used in products at random), and there is little difference between capabilities in this respect. Incorporating more complex structure into the model may provide more insight into how the dynamics of development depend on the structure of the 'recipe book'.

Work by Fink et al. (2017) explicitly studies the consequences of such structure in a model that is directly applicable to the capabilities framework. By looking at data on building blocks and outcomes (e.g. ingredients and culinary dishes), it shows how the order of acquiring capabilities matters in how the number of dishes one can make evolves as more ingredients are gained, depending on the usefulness of the ingredients. In the context of economic development, this translates directly into relevant policy questions: is there an optimal order in which capabilities should be obtained?

This also relates to the discussion of industrial policy in 7. One could model the 'coordination constraint' introduced in 7 not as a separate parameter, but as a capability that is used specifically in products requiring many other capabilities. This is arguably a more elegant way to model and stays closer to the idea of capabilities being generic inputs that can include abstract things like institutions and regulations. Capabilities that are used in many products for example (representing the equivalent of salt and pepper in the recipe book) can be considered as general purpose technologies, and may spur diversification either early or late in the development process. The question of horizontal versus vertical policies then becomes one of at which stage of development it is better to focus on general purpose capabilities, rather than capabilities that give the highest short term return. A vertical policy would consist of aiming at specific capabilities needed for a single new product, while a horizontal policy would aim at capabilities that are used in many (possibly complex) products.

Again, the biggest constraint here is that we do not have access to the recipe book mapping capabilities to products. A possible way around this is to not model explicitly the connections between capabilities and products, but to make assumptions on the distributions of product complexity and capability usefulness, as in (Fink and Reeves, 2019). The baseline model studied in this thesis implicitly assumes a binomial distribution of product complexities. However, one can create models with other distributions of product complexities, that may lead to different diversification patterns. Such extensions may provides a fruitful way forward in studying how much the development process is constrained by the structure of the underlying recipe book.

8.5 The future of economic complexity

In less than two decades, the idea of studying the structure of economic output and taking into account the heterogeneity of necessary inputs in the form of capabilities has led to a new approach to study economic development that goes beyond the traditional growth models used in economics. It has sparked a wide range of mostly empirical papers by scholars in economics, geography, physics and complexity science. Many of the underlying ideas bear close resemblance to work in Ecology. This has lead to an interdisciplinary field that is gaining momentum, and has raised interest in the policy arena.

In this thesis I have attempted to integrate the central concepts in this field by working towards an encompassing methodological framework, and extending the theoretical models underlying them. By taking a more formal approach to the methodological challenges at hand, work in economic complexity can build on state of the art methodologies from other rapidly developing fields like network science and computer science. Theoretical models may help theorizing and deriving policy implications from a rich body of empirical work.

What remains to be done is to apply the proposed methodological frameworks more extensively as to test the key hypotheses put forward by complexity economists regarding economic development. A first attempt to such an undertaking was made in Chapter 2, in which we engaged with the framework of Hill numbers (further developed in Chapter 3) and highlighted the limits of the economic complexity index (as further elaborated in Chapter 4). Regarding the information-theoretic framework to measure location and co-location as proposed in Chapter 5, some practical challenges remain, including the definition of suitable priors for the Bayesian estimation and the sensitivity of information-theoretic measures to low-frequency data. Also, investigating higher-order associations would require larger and more detailed datasets than most that are currently available. Hopefully, this thesis will inspire new empirical work in the years to come.

What this thesis did not do is to put the proposed methodology to the test empirically. A first step would be to replicate some of the key studies in the literature using the methodologies proposed in this thesis. A first attempt to such an undertaking was made in Chapter 2. Practical issues in this respect remain, such as the definition of suitable priors for the Bayesian estimation, and dealing with the sensitivity of information-theoretic measures to low-frequency data. Other challenges lie in the representation of negative and higher-order associations, as these are no longer easily represented as a network. Many of the proposed extensions in this thesis further require large, detailed datasets that span large time scales and geographical scales.

Chapter 9

Supplementary material to The concept of diversity in economic geography: related variety, economic complexity and the product space

A 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);[1] the Local Area Unemployment Statistics (LAUS) for unemployment (U);[2] and the Occupational Employment Statistics (OES) for occupations.[3]

A.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.[4]

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;[5] i.e., the same used by

[1] `https://www.bls.gov/cew/downloadable-data-files.htm`

[2] `https://download.bls.gov/pub/time.series/la/la.data.64.County`

[3] `https://www.bls.gov/oes/`

[4] 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).

[5] `https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2003/historical-delineation-files/0312cbsas-csas.xls`

BLS).[6] 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.[7] 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.[8]

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.[9]

A.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

[6] `https://www.bls.gov/cew/questions-and-answers.htm`

[7] 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.

[8] `https://www.bls.gov/sae/additional-resources/what-is-naics.htm`
`https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2002`

[9] `https://www.bls.gov/cew/overview.htm#confidentiality`

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).

B List of variables

Table 9.1: 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 9.2: 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
$\ln E_{cT} \big/ E_{ct}$	369	0.44	0.24	-0.08	0.29	0.41	0.57	1.77
$\ln w_{cT} \big/ w_{ct}$	369	0.58	0.09	0.22	0.52	0.57	0.63	1.21
$\ln U_{cT} \big/ U_{ct}$	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	1.34	0.23	0.76	1.23	1.28	1.35	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	-2.73	0.28	-3.18	-2.94	-2.75	-2.61	-1.26

Note: Wherever not necessary, the subscript t is omitted for brevity.

Table 9.3: 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
$\ln \mathrm{var}_c^i$	43315	-1.25	0.65	-4.93	-1.42	-1.06	-0.81	-0.49
$\ln \mathrm{bal}_c^i$	43315	-0.66	0.19	-2.82	-0.74	-0.65	-0.57	-0.04
$\ln \mathrm{disp}_c^i$	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 9.4: Descriptive statistics for city-industry level data, using the industry space as defined by $\tilde{\phi}_{ii'}$ of eq. (2.7).

	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	5.87	1.40	0.95	4.75	5.69	6.83	10.10
$\ln D_c^i$	43322	-1.39	0.28	-3.45	-1.55	-1.39	-1.21	-0.34
$\ln D_I(\mathbf{p}_c^i)$	43322	4.20	0.36	1.95	3.97	4.24	4.47	4.95
$\ln D_Z(\mathbf{p}_c^i)$	43322	0.72	0.03	0.65	0.71	0.72	0.72	1.19
$\ln \mathrm{var}_c^i$	43322	-0.76	0.38	-3.23	-1.04	-0.74	-0.47	-0.05
$\ln \mathrm{bal}_c^i$	43322	-0.66	0.16	-2.09	-0.72	-0.65	-0.57	-0.16
$\ln \mathrm{disp}_c^i$	43322	-3.48	0.37	-4.22	-3.76	-3.53	-3.26	-1.08
$\ln E_c$	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 9.5: Descriptive statistics for city-industry level data, using the industry space as defined by $\tilde{\psi}_{ii'}$ of eq. (2.12).

	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
$\ln \mathrm{var}_c^i$	43322	-0.85	0.43	-4.02	-1.15	-0.82	-0.54	-0.05
$\ln \mathrm{bal}_c^i$	43322	-0.70	0.19	-2.51	-0.79	-0.68	-0.58	-0.02
$\ln \mathrm{disp}_c^i$	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 9.6: Descriptive statistics for city-industry level data, using the industry space as defined by $\rho_{ii'}$ of eq. (2.13).

	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
$\ln E_{icT}/E_{ict}$	38484	0.32	0.75	-5.02	-0.07	0.29	0.68	5.84
$\ln w_{icT}/w_{ict}$	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
$\ln \mathrm{var}_c^i$	38484	-1.90	0.74	-4.93	-2.19	-1.72	-1.39	-0.60
$\ln \mathrm{bal}_c^i$	38484	-0.58	0.22	-2.59	-0.65	-0.54	-0.46	-0.00
$\ln \mathrm{disp}_c^i$	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.

Chapter 10

Supplementary material to Diversity and its decomposition into variety, balance and disparity

A Hill numbers and entropy

In the definition of diversity we rely on the concept of Hill numbers, following (Hill, 1973a) and (Jost, 2006). The Hill number of order q is given by the reciprocal of a generalized mean of the relative frequencies. The generalized weighted mean of the relative frequencies of types is given by

$$^{q}\bar{p} = \sqrt[q-1]{\sum_i p_i p_i^{q-1}}, \tag{10.1}$$

where the weights are given by the relative frequencies p_i. The parameter q determines which mean is considered. For example, $^{0}\bar{p}$ denotes the Harmonic mean, $^{1}\bar{p}$ the geometric mean and for $^{2}\bar{p}$ the arithmetic mean (Hill, 1973a). The Hill number of order q measures the diversity of types as the reciprocal of the mean

$$^{q}D(S) = \frac{1}{^{q}\bar{p}} = \left(\sum_i p_i^q\right)^{\frac{1}{1-q}}.$$

The parameter q determines how heavily the average weights common or rare species. Values of $q > 1$ weigh more heavily types with high relative frequency, and values of $q < 1$ weigh more heavily the presence of types with small relative frequency. The minimal value of $q = 0$ considers every type to contribute equally to the mean, regardless of its relative frequency. For $q = 0$ the diversity is given by

$$^0D(S) = \sum_i 1 = n$$

and gives simply a count of the number of types in S. The Hill number of order 0 is thus a measure of *variety*, which is also known as species richness in ecology.

For $q = 2$, one obtains

$$^2D(S) = \sum_i \frac{1}{p_i^2},$$

which relates directly to Simpson's index of concentration and the Gini-index (Jost, 2006).
In general, the Hill numbers are related to the Rényi entropy (Renyi and Rényi, 1961) by $^qD(S) = e^{^qH(X)}$, where

$$^qH(X) = \frac{1}{1-q} \log \left(\sum_i p_i^q \right).$$

The Shannon entropy arises as a special case when taking the limit of $q \to 1$. This corresponds to the unique Hill number that does not favor either rare or common types and is given by

$$D(S) = \lim_{q \to 1} {}^qD(S) = e^{-\sum_i p_i \log(p_i)} = e^{H(X)}.$$

The relationship between Hill numbers and entropies described above tell us how to transform measures of uncertainty, given by entropies in units of bits or nats, into measures of diversity, given in units of the 'effective number of types'. The more uncertain one is about the type of a randomly sampled element from S (i.e. the higher $^qH(X)$), the more diverse the set S in considered to be.

B Properties from Leinster & Cobbold

In their introduction of a diversity measure that takes into account disparity by including pairwise similarities between types, Leinster & Cobbold (Leinster and Cobbold,

2012) show that their measure satisfies nine properties that 'encode basic scientific intuition' that every diversity measure should satisfy. The nine properties are divided into three categories: partitioning properties, elementary properties, and similarity properties. In this section it is shown that the properties posed in (Leinster and Cobbold, 2012) also hold for the number of compositional units $D_\beta(S')$.

We follow the notation as introduced in the main text: a collection of features $i \in S$, a collection of types $j \in S'$, and their corresponding random variables X, Y and XY with probabilities $p_i = P(X = i)$, $p_j = P(Y = j)$, and $p_{ij} = P(X = i, Y = j)$ respectively.

Partitioning

Effective number: the diversity of a community of n equally abundant, totally dissimilar types is n.

Note that when all types are totally dissimilar, there is no uncertainty about the type j of an element given that one knows its feature i. That is, for every feature i we have that $p_{j|i} = 1$ for one specific type j. This implies that $H(Y|X) = -\sum_i p_i \sum_j p_{j|i} \log(p_{j|i}) = 0$, so that

$$\begin{aligned} MI(X,Y) &= H(X) + H(Y) - H(XY) \\ &= H(Y) - H(Y|X) \\ &= H(Y). \end{aligned}$$

Then

$$D_\beta(S') = e^{MI(X,Y)} = e^{H(Y)} = D(S').$$

Hence for totally dissimilar types the number of compositional units reduces to the effective number of types. In particular, for equally abundant types we have $D_\beta(S') = e^{H(Y)} = n$.

Modularity: if a collection of types consists of multiple non-overlapping sub-collections of types, for which types in different sub-collections are totally dissimilar, then the total diversity is entirely determined by the size and diversity of every sub-collection.

We can implement the sub-collections by adding a third label k to every element, which denotes the sub-collection $k \in S''$ it belongs to. Hence, we now have elements with labels i, j, k, where i denotes a feature, j denotes a type, and k denotes the sub-collection. Further introducing the corresponding random variable Z, this defines probabilities $p_{ijk} = P(X = i, Y = j, Z = k)$. Since sub-collections are non-overlapping, there is no uncertainty about the sub-collection k of an element given that one know its type j, so that $H(Z|Y) = 0$. Furthermore, since types from different sub-collections are totally dissimilar, sub-collections do not share any features, so there is no uncertainty about the sub-collection k of an element given that one knows its feature i, so $H(Z|X) = 0$. These properties imply that $H(YZ) = H(Y)$ and $H(XZ) = H(X)$. Defining

$$MI(X,Y|Z) = \sum_k p_k \sum_{ij} p_{ij|k} \log\left(\frac{p_{ij|k}}{p_{i|k}p_{j|k}}\right),$$

we can then write

$$\begin{aligned} MI(X,Y|Z) &= H(X|Z) + H(Y|Z) - H(XY|Z) \\ &= H(XZ) - H(Z) + H(YZ) - H(Z) - H(XY) + H(Z) \\ &= H(X) + H(Y) - H(XY) - H(Z) \\ &= MI(X,Y) - H(Z) \end{aligned}$$

so that

$$MI(X,Y) = MI(X,Y|Z) + H(Z).$$

Taking the exponential, this shows how the total number of compositional units of types S' relates to the number of compositional units in each sub-collection k, their relative size p_k, and the effective number of sub-collections $D(S'')$:

$$\begin{aligned} D_\beta(S') &= e^{MI(X,Y)} \\ &= e^{MI(X,Y|Z)+H(Z)} \\ &= e^{\sum_k p_k MI(X,Y|k)+H(Z)} \\ &= D(S'') \prod_k D_\beta(S'_k)^{p_k}, \end{aligned} \tag{10.2}$$

where $D(S'') = e^{H(Z)}$ denotes the effective number of sub-collections.

Replication: if m non-overlapping sub-collections are of equal size and diversity d, the diversity of the whole collection is given by md.

Using (10.2), it is easily seen that if the number of compositional units in every sub-collection is d, and there are m sub-collections with relative size $\frac{1}{m}$, we have

$$D_\beta(S') = m \prod_k d^{\frac{1}{m}} = md.$$

Elementary

Symmetry: diversity is independent of the order of the listing of types.

This property follows directly from the properties of the Shannon entropy.

Absent types: diversity is unchanged by adding a type of zero abundance.

This property follows directly from the properties of the Shannon entropy.

Identical types: for two identical types, merging the types leaves diversity unchanged.

Recall that XY is defined as the random variable with probabilities $p_{ij} = P(X = i, Y = j)$, where $i \in S$ and $j \in S'$. For two identical types j' and j'', we have that $p_{i|j'} = p_{i|j''}$ since they have an identical distribution over features.

Define a random variable $X\tilde{Y}$ in which j' and j'' are merged, i.e. $\tilde{p}_{ij'} = P(X = i, \tilde{Y} = j) = p_{ij'} + p_{ij''}$, $\tilde{p}_{ij''} = 0$ and $\tilde{p}_{ij} = p_{ij}$ for all $j \neq j', j''$. Then

$$\begin{aligned} MI(X,Y) &= \sum_{ij, j\neq j', j''} p_{ij} \log\left(\frac{p_{i|j}}{p_i}\right) + \sum_i p_{ij'} \log\left(\frac{p_{i|j'}}{p_i}\right) + \sum_i p_{ij''} \log\left(\frac{p_{i|j''}}{p_i}\right) \\ &= \sum_{ij, j\neq j', j''} p_{ij} \log\left(\frac{p_{i|j}}{p_i}\right) + \sum_i (p_{ij'} + p_{ij''}) \log\left(\frac{p_{i|j'}}{p_i}\right) \\ &= MI(X, \tilde{Y}). \end{aligned}$$

Hence, $D_\beta(S') = D_\beta(\tilde{S}')$, so merging identical types does not affect the number of compositional units.

Effect of similarity on diversity

Monotonicity: when similarity between types is increased, diversity decreases.

Although we do not have an explicit measure of pairwise similarity between types, similarity in our framework is given by the (average) overlap of features between types. This overlap may increase in two ways: either the total diversity of features $D_\gamma(S)$ decreases while the average within-type diversity $D_\alpha(S)$ remains constant, or the average within-type diversity $D_\alpha(S)$ increases while the total diversity of features $D_\gamma(S)$ remains constant. From the definition of the number of compositional units $D_\beta(S') = \frac{D_\gamma(S)}{D_\alpha(S)}$ it follows that in both cases the number of compositional units decreases.

Naive model: when similarities are ignored, diversity is greater or equal than when similarities are taken into account.

This follows directly from the definitions of $D_\beta(S')$ (which takes into account disparity) and $D(S')$ (which does not take into account disparity), and the known property that $MI(XY) \leq H(Y)$. This leads to

$$D(S') = e^{H(Y)} \geq e^{MI(X,Y)} = D_\beta(S').$$

Range: the diversity of a collection of n types is between 1 and n.

We have that $0 \leq MI(XY) \leq H(Y) \leq \log(n)$. Taking exponentials, this gives $1 \leq D_\beta(S') \leq n$.

C Multiple feature sets

This section elaborates on the results given in the main text on diversity when taking into account two feature sets, described by random variables X and Y. The feature pairs are then described by the joint distribution $p_{ij} = P(X = i, Y = j)$. Using the simple additive properties of information-theoretic quantities, we show some simple results regarding diversities. The calculations are easily verified by considering the Venn diagrams in Figure 10.1.

Here, we rewrite the diversity of types corresponding to random variable Z given the overlap among a pair of features given by random variables X and Y as

$$\begin{aligned} D_\beta^{XY}(S') &= e^{MI(XY,Z)} \\ &= e^{H(XY)-H(XY|Z)} \\ &= e^{H(X)+H(Y)-MI(X,Y)-H(X|Z)-H(Y|Z)+MI(X,Y|Z)} \\ &= e^{MI(X,Z)+MI(Y,Z)-MI(X,Y)+MI(X,Y|Z)}, \end{aligned} \tag{10.3}$$

where we used that $H(XY) = H(X)+H(Y)-MI(X,Y)$ and $H(XY|Z) = H(X|Z)+H(Y|Z)-MI(X,Y|Z)$. From this, it becomes clear that the diversity becomes lower as the features X and Y have a larger dependence, i.e. are more correlated, as indicated by a large value of $MI(X,Y)$.

In the special case that features X and Y share no information, i.e. $MI(X,Y) = 0$, we have

$$\begin{aligned} D_\beta^{XY}(S') &= e^{MI(XY,Z)} \\ &= e^{MI(X,Z)+MI(Y,Z)} \\ &= D_\beta^{X}(S')D_\beta^{Y}(S'). \end{aligned} \tag{10.4}$$

Hence, for independent feature sets the diversities are multiplicative.

D Aggregation

Here we consider the types described by random variable Z to be composed of features described by random variable Y, and the features Y themselves have features described by a random variable X (this reflects the situation described in the modularity property in Section B, where Z denotes the sub-collections, Y denotes the types, and X denotes the features). Hence the links between types and features are given by the joint probability distribution p_{jk}, and the links between features and 'sub-features' by a joint distribution p_{ij}. When the joint probabilities p_{ij} are independent of the joint probabilities p_{jk}, we have $p_{ijk} = p_{ij}p_{k|j} = p_{i|j}p_{k|j}p_j$. In other words, the random variables Z and X are conditionally independent given Y, which means that $MI(X, Z|Y) = 0$. The diversity given feature pairs XY can then be rewritten as

$$\begin{aligned} D_\beta^{XY}(S') &= MI(XY, Z) \\ &= e^{H(Z)-H(Z|XY)} \\ &= e^{H(Z)-(H(ZX|Y)-H(X|Y))} \\ &= e^{H(Z)-H(Z|Y)} = e^{MI(Z,Y)}, \end{aligned} \tag{10.5}$$

where we used that $MI(X, Z|Y) = 0$ implies that $H(XZ|Y) - H(X|Y) = H(Z|Y)$. In other words, considering X is superfluous when considering the diversity of Z in terms of features XY.

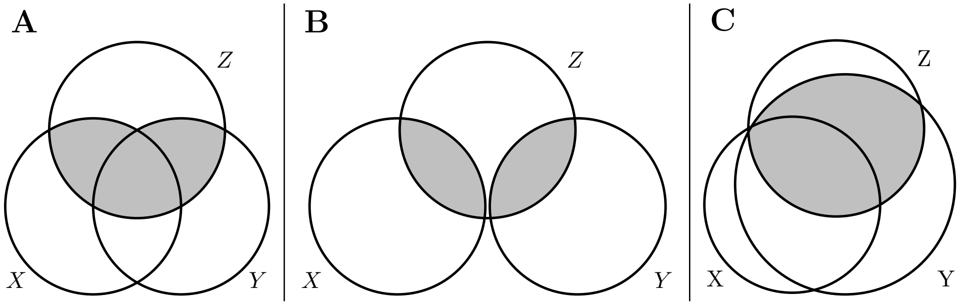


Figure 10.1: The entropies and mutual information can be represented using Venn diagrams, where each circle corresponds to the entropy $H(X)$ of the associated random variable X. The intersection of the two circles associated to X and Y represents the mutual information $MI(X, Y)$, and their union represents the joint entropy $H(XY)$. The conditional entropy $H(X|Y)$ is given by subtracting the intersection from the total uncertainty $H(X)$. **A** shows the mutual information $MI(XY, Z)$ from equation (10.3). The diversity of variable Z given the overlap in features XY is given by the exponential of the shaded area. **B** shows the special case of (10.4) in which the features X and Y are independent, i.e. $MI(X, Y) = 0$. From the figure it is clear that $MI(XY, Z) = MI(X, Z) + MI(Y, Z)$, such that associated diversity in this case is multiplicative. **C** shows the case of (10.5) in which Z and X are conditionally independent on Y, i.e. $MI(Z, X|Y) = 0$. In this case, taking into account features X becomes irrelevant in computing diversity of Z given feature pairs XY.

Chapter 11

Supplementary material to A network view of correspondence analysis: applications to ecology and economic complexity

A Constructing the country-product matrix

The data contains for every country how much of each product is has exported in the year 2016. To obtain a binary 'presence-absence' matrix, we consider whether a country exports a product with 'revealed comparative advantage' (RCA). The RCA index compares the share of a product within a countries' export portfolio to the global share of that product in world trade to evaluate whether a country exports more than expected by the global share. If q_{ij} denotes the exports of product j in country i (given in dollars), the RCA is defined as $RCA(i,j) = \frac{q_{ij}/\sum_j q_{ij}}{\sum_i q_{ij}/\sum_{i,j} q_{ij}}$. The matrix A is then defined as

$$A_{ij} = \begin{cases} 1 & \text{if } RCA(i,j) > 1 \\ 0 & \text{if } RCA(i,j) \leq 1, \end{cases}$$

B Relation between normalized cut and correlation analysis

We show that the eigenvalues and eigenvectors of Eq. (4.2) that result from the correlation analysis are directly related to those that follow from minimizing the normalized cut (Eq. (4.3)). Recall that minimizing the normalized cut results in the generalized eigenproblem

$$(D_r - S_r)\mathbf{v} = \tilde{\lambda} D_r \mathbf{v}.$$

Pre-multiplying by D_r^{-1}, this can be rewritten as

$$\begin{aligned}(I - D_r^{-1} S_r)\mathbf{v} &= \tilde{\lambda}\mathbf{v}\\ D_r^{-1} S_r \mathbf{v} &= (1 - \tilde{\lambda})\mathbf{v}\\ D_r^{-1} A D_c^{-1} A^T &= \lambda\mathbf{v},\end{aligned}$$

which shows that solutions to Eq. (4.2) are solutions to Eq. (4.3) with $\lambda = 1 - \tilde{\lambda}$.

C Country clusters

cluster	countries
2	Algeria, Angola, Brunei Darussalam, Chad, Congo, Congo (Democratic Republic of the), Equatorial Guinea, Gabon, Iraq, Kuwait, Libya, Nigeria, Papua New Guinea, Qatar, South Sudan, Turkmenistan, Venezuela
3	Afghanistan, Albania, American Samoa, Andorra, Anguilla, Antarctica, Antigua and Barbuda, Argentina, Armenia, Aruba, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bonaire, Bosnia and Herzegovina, Botswana, Bouvet Island, Brazil, British Indian Ocean Territory, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chile, China, Christmas Island, Cocos (Keeling) Islands, Colombia, Comoros, Costa Rica, Croatia, Cuba, Curaçao, Cyprus, Czech Republic, Côte d'Ivoire, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Eritrea, Estonia, Eswatini, Ethiopia, Fiji, Finland, France, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guam, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Heard and McDonald Islands, Honduras, Hong Kong, Hungary, India, Indonesia, Iran, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Lithuania, Luxembourg, Macao, Madagascar, Malawi, Malaysia, Mali, Malta, Mauritius, Mexico, Moldova, Mongolia, Montenegro, Montserrat, Morocco, Mozambique, Myanmar, Namibia, Nauru, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Niue, Norfolk Island, North Korea, North Macedonia, Northern Mariana Islands, Norway, Oman, Pakistan, Palestine, Panama, Paraguay, Peru, Philippines, Pitcairn, Poland, Portugal, Romania, Russian Federation, Rwanda, Saint Barthélemy, Saint Helena, Ascension and Tristan da Cunha, Saint Kitts and Nevis, Saint Lucia, Saint Pierre and Miquelon, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, Somalia, South Africa, South Georgia and South Sandwich Islds., South Korea, Spain, Sri Lanka, St-Martin / St Maarten, Sudan, Suriname, Sweden, Switzerland, Syrian Arab Republic, Taiwan, Tajikistan, Tanzania, Thailand, Timor-Leste, Togo, Tokelau, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turks and Caicos Islands, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States Minor Outlying Islands, United States of America, Uruguay, Uzbekistan, Vatican City, Vietnam, Wallis and Futuna, Western Sahara, Yemen, Zambia, Zimbabwe
4	Bahamas, Bermuda, Cayman Islands, Cook Islands, Falkland Islands, Faroe Islands, French Polynesia, French Southern and Antarctic Lands, Gibraltar, Greenland, Iceland, Kiribati, Maldives, Marshall Islands, Mauritania, Micronesia, New Caledonia, Palau, Saint Vincent and the Grenadines, Seychelles, Solomon Islands, Tuvalu, Vanuatu, Virgin Islands (British)

Chapter 12

Supplementary material to An information-theoretic approach to the analysis of location and co-location patterns

A Estimation of $\mathrm{Var}[p_{ij}]$

Estimates for the expectation and variance of $PMI(p_{ij})$ are obtained in a similar fashion as (5.7) and (5.8). This requires computation of $\mathrm{Var}[p_{ij}]$. We have

$$\begin{aligned}\mathrm{Var}[p_{ij}] &= \mathrm{Var}[\sum_c p_{i|c}p_{j|c}p_c] \\ &= \sum_c \mathrm{Var}[p_{i|c}p_{j|c}p_c] + \sum_{c\neq c'} \mathrm{Cov}[p_{i|c}p_{j|c}p_c, p_{i|c'}p_{j|c'}p_{c'}] \\ &= \sum_c \big(\mathrm{Var}[p_{i|c}p_{j|c}]\mathbb{E}[p_c]^2 + \mathrm{Var}[p_c]\mathbb{E}[p_{i|c}p_{j|c}]^2 + \mathrm{Var}[p_{i|c}p_{j|c}]\mathrm{Var}[p_c]\big) + \sum_{c\neq c'} \mathbb{E}[p_{i|c}p_{j|c}]\mathbb{E}[p_{i|c'}\end{aligned} \tag{12.1}$$

where in the second equality we used that $p_{i|c}p_{j|c}$ is independent of p_c, p_c' and $p_{i|c'}p_{j|c'}$. Furthermore, we used that

$$\begin{aligned}\mathrm{Cov}[p_{i|c}p_{j|c}p_c, p_{i|c'}p_{j|c'}p_{c'}] &= \mathbb{E}[p_{i|c}p_{j|c}p_c p_{i|c'}p_{j|c'}p_{c'}] - \mathbb{E}[p_{i|c}p_{j|c}p_c]\mathbb{E}[p_{i|c'}p_{j|c'}p_{c'}] \\ &= \mathbb{E}[p_{i|c}p_{j|c}]E[p_{i|c'}p_{j|c'}](\mathbb{E}[p_c p_{c'}] - \mathbb{E}[p_c]\mathbb{E}[p_{c'}]) \\ &= \mathbb{E}[p_{i|c}p_{j|c}]\mathbb{E}[p_{i|c'}p_{j|c'}]\mathrm{Cov}[p_c, p_{c'}].\end{aligned}$$

Note that the vector of p_c's follows a Dirichlet distribution, so that p_c and $p_{c'}$ are not independent.

Using the product-moment formula (Nadarajah and Kotz, 2004), we know that for $i \neq j$

$$\mathbb{E}[p_{i|c}^n p_{j|c}^n] = \frac{\Gamma(\tilde{q}_{ci}+n)\Gamma(\tilde{q}_{cj}+n)\Gamma(\tilde{q}_c)}{\Gamma(\tilde{q}_{ci})\Gamma(\tilde{q}_{cj})\Gamma(\tilde{q}_c+2n)},$$

so that

$$\begin{aligned}\mathrm{Var}[p_{i|c}p_{j|c}] &= \mathbb{E}[p_{i|c}^2 p_{j|c}^2] - \mathbb{E}[p_{i|c}p_{j|c}]^2 \\ &= \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)\tilde{q}_{cj}(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)(\tilde{q}_c+2)(\tilde{q}_c+3)} - \left(\frac{\tilde{q}_{ci}\tilde{q}_{cj}}{\tilde{q}_c(\tilde{q}_c+1)}\right)^2.\end{aligned}$$

The last term of (12.1) consists of

$$\sum_{c\neq c'} \mathbb{E}[p_{i|c}p_{j|c}]\mathbb{E}[p_{i|c'}p_{j|c'}]\mathrm{Cov}[p_c, p_{c'}] = -\sum_{c\neq c'} \frac{\tilde{q}_{ci}\tilde{q}_{cj}}{\tilde{q}_c(\tilde{q}_c+1)} \frac{\tilde{q}_{c'i}\tilde{q}_{c'j}}{\tilde{q}_c(\tilde{q}_c+1)} \frac{\tilde{q}_c\tilde{q}_{c'}}{(\tilde{q}+1)}.$$

For $i = j$ we have

$$\mathrm{Var}[p_{ii}] = \sum_c \Big(\mathrm{Var}[p_{i|c}^2]\mathbb{E}[p_c]^2 + \mathrm{Var}[p_c]\mathbb{E}[p_{i|c}^2]^2 + \mathrm{Var}[p_{i|c}^2]\mathrm{Var}[p_c]\Big) + \sum_{c\neq c'} \mathbb{E}[p_{i|c}^2]\mathbb{E}[p_{i|c'}^2]\mathrm{Cov}[p_c, p_{c'}].$$

Since $p_{i|c}$ is beta-distributed, we have

$$\mathbb{E}[p_{i|c}^2] = \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)}$$

and

$$\begin{aligned}\mathrm{Var}[p_{i|c}^2] &= \mathbb{E}[p_{i|c}^4] - \mathbb{E}[p_{i|c}^2]^2 \\ &= \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)(\tilde{q}_{ci}+2)(\tilde{q}_{ci}+3)}{\tilde{q}_c(\tilde{q}_c+1)(\tilde{q}_c+2)(\tilde{q}_c+3)} - \left(\frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right)^2\end{aligned}$$

The last term consists of

$$\sum_{c\neq c'} \mathbb{E}[p_{i|c}^2]\mathbb{E}[p_{i|c'}^2]\mathrm{Cov}[p_c, p_{c'}] = -\sum_{c\neq c'} \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)} \frac{\tilde{q}_{c'i}(\tilde{q}_{c'i}+1)}{\tilde{q}_{c'}(\tilde{q}_{c'}+1)} \frac{\tilde{q}_c\tilde{q}_{c'}}{\tilde{q}^2(\tilde{q}+1)}.$$

For larger data sets, the following approximation can be made to keep things tractable and computationally feasible. We approximate the variance of p_{ij} by assuming that there is no uncertainty about p_c, and assume $p_c = \hat{p}_c$. We then have

$$\mathrm{Var}[p_{ij}] \approx \sum_c \hat{p}_c^2 \mathrm{Var}[p_{i|c}p_{j|c}].$$

This gives for $i \neq j$:

$$\begin{aligned}\mathrm{Var}[p_{ij}] &= \sum_c \hat{p}_c^2 (\mathbb{E}[p_{i|c}^2 p_{j|c}^2] - \mathbb{E}[p_{i|c}p_{j|c}]^2) \\ &= \sum_c \frac{\tilde{q}_c^2}{\tilde{q}^2} \left(\frac{\tilde{q}_{ci}\tilde{q}_{cj}(\tilde{q}_{ci}+1)(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)(\tilde{q}_c+2)(\tilde{q}_c+3)} - \frac{\tilde{q}_{ci}^2\tilde{q}_{cj}^2}{\tilde{q}_c^2(\tilde{q}_c+1)} \right)\end{aligned}$$

and for $i = j$

$$\begin{aligned}\mathrm{Var}[p_{ii}] &= \sum_c \hat{p}_c^2 (\mathbb{E}[p_{i|c}^4] - \mathbb{E}[p_{i|c}^2]^2) \\ &= \sum_c \frac{\tilde{q}_c^2}{\tilde{q}^2} \left(\frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)(\tilde{q}_{ci}+2)(\tilde{q}_{ci}+3)}{\tilde{q}_c(\tilde{q}_c+1)(\tilde{q}_c+2)(\tilde{q}_c+3)} - \left(\frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1}{\tilde{q}_c(\tilde{q}_c+1)} \right)^2 \right)\end{aligned}$$

B MCMC simulations for $PMI(p_{ci})$ and $PMI(p_{ij})$

We implement the estimation procedure in Python using the package pymc3. Figure 12.1 compares the analytical approximations to Markov chain Monte Carlo simlulations. The results show a good fit between simulated results and the analytical pproximations, except for the standard deviation of the $PMI(p_{ij})$, where the analytical approximation slightly over-estimates the standard deviation.

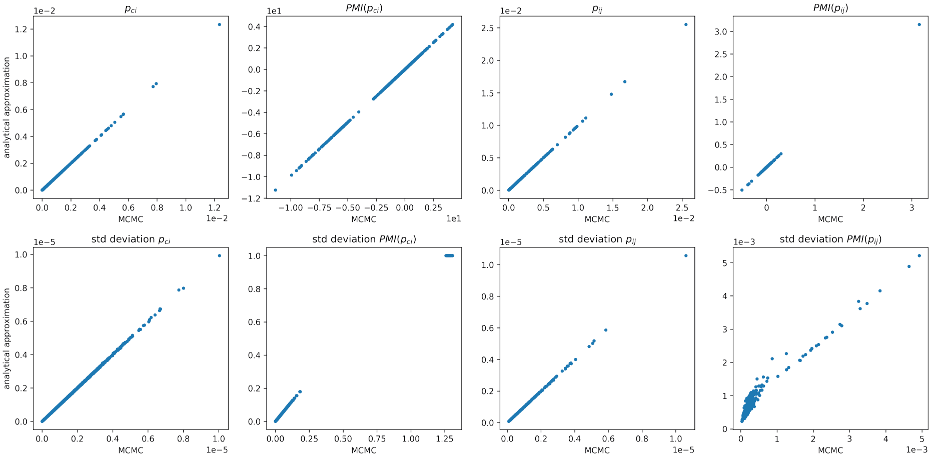


Figure 12.1: Simulated values versus analytical approximations.

C Estimation of $\mathrm{Var}[KL(p_{j|i}|p_j]$

We estimate $KL(p_{j|i}|p_j)$ in a similar way as the $PMI(p_{ij})$. By computing the Taylor expansion and taking the expectation we obtain

$$\begin{aligned}\mathbb{E}[KL(p_{j|i}|p_j)] &\approx KL(\hat{p}_{j|i}|\hat{p}_j) + \sum_j \mathrm{Var}(p_{ij}) \frac{\partial^2}{\partial p_{ij}^2} p_{j|i} \log\left(\frac{p_{ij}}{p_i p_j}\right) \\ &= KL(\hat{p}_{j|i}|\hat{p}_j) + \sum_j \mathrm{Var}(p_{ij}) \left(\frac{2p_{ij}}{p_i p_j^2} + \frac{p_{ij}}{p_i^2 p_j} + \frac{3p_{ij}}{p_j}^3 + \frac{1}{p_i p_{ij}} - \frac{4}{p_j^2} - \frac{2}{p_i p_j} \right. \\ &\left. + \left(\frac{2p_{ij}}{p_j^3} - \frac{2}{p_j^2}\right) \log\left(\frac{p_{ij}}{p_i p_j}\right)\right).\end{aligned}$$

For the variance of $KL(p_{j|i}|p_j)$ we obtain, using the delta method,

$$\begin{aligned}\mathrm{Var}(KL(p_{j|i}|p_j) &\approx \sum_j \mathrm{Var}(p_{ij}) \frac{\partial}{\partial p_{ij}} p_{j|i} \log\left(\frac{p_{ij}}{p_i p_j}\right) \\ &= \mathrm{Var}(p_{ij}) \left(\frac{1}{p_j} + \frac{p_{ij}}{p_j^2}\right) \left(\log\left(\frac{p_{ij}}{p_i p_j} - 1\right) - \frac{p_{ij}}{p_i p_j}\right).\end{aligned}$$

The standard deviations are given as the square root of this variance. Similar equations can be derived for $KL(p_{c|i}|p_c)$ and $KL(p_{i|c}|p_i)$.

Chapter 13

Supplementary material to Variety, complexity and economic development

A Derivations of full model quantities

A.1 Average product length given r

First, note that the total product length is given by

$$
\begin{aligned}
d(n,r)\bar{s}(n,r) &= \sum_{s=n-r}^{n} s\binom{n}{s} \\
&= n\sum_{s=n-r}^{n} \frac{(n-1)!}{(s-1)!(n-s)!} \\
&= n\sum_{s'=n-r-1}^{n-1} \frac{(n-1)!}{s'!(n-s'-1)!} \\
&= n\sum_{s'=n-r-1}^{n-1} \binom{n-1}{s'} \\
&= nd(n-1,r)
\end{aligned}
$$

so that the average product length is given by

$$
\bar{s}(n,r) = n\frac{d(n-1,r)}{d(n,r)}.
$$

A.2 Bounds on average product length

We show that

$$\frac{1}{2} \leq \frac{d(n-1,r)}{d(n,r)} < 1.$$

First, note that

$$\begin{aligned}
d(n+1,r) - d(n,r) &= \sum_{s=n+1-r}^{n+1} \binom{n+1}{s} - \sum_{s=n-r}^{n} \binom{n}{s} \\
&= \sum_{s=n+1-r}^{n+1} \binom{n+1}{s} - \sum_{s=n-r}^{n+1} \frac{n+1-s}{n+1} \binom{n+1}{s} \\
&= \sum_{s=n+1-r}^{n+1} \binom{n+1}{s} - \sum_{s=n+1-r}^{n+1} \frac{n+1-s}{n+1} \binom{n+1}{s} - \frac{r+1}{n+1} \binom{n+1}{n-r} \\
&= \sum_{s=n+1-r}^{n+1} \frac{s}{n+1} \binom{n+1}{s} - \binom{n}{r} \\
&= d(n,r) - \binom{n}{r},
\end{aligned}$$

so that

$$d(n+1,r) = 2d(n,r) - \binom{n}{r}. \tag{13.1}$$

Now since

$$d(n-1,r) = \sum_{s=n-1-r}^{n-1} \binom{n-1}{s} = \sum_{s=n-r}^{n-1} \binom{n-1}{s} + \binom{n-1}{n-1-r} > \binom{n-1}{n-1-r},$$

we have that

$$\begin{aligned}
d(n,r) &= 2d(n-1,r) - \binom{n-1}{r} \\
&= 2d(n-1,r) - \binom{n-1}{n-1-r} \\
&> 2d(n-1,r) - d(n-1,r) \\
&> d(n-1,r),
\end{aligned}$$

so that $\frac{d(n-1,r)}{d(n,r)} < 1$.

Furthermore, since $\binom{n-1}{r} > 0$, (13.1) gives that

$$d(n,r) < 2d(n-1,r)$$

and thus $\frac{d(n-1,r)}{d(n,r)} > \frac{1}{2}$ for $r < n$.

A.3 Average product length including ρ

$$\begin{aligned}
\bar{s}(n) = \sum_{s=0}^{n} s\frac{d(n,s)}{d(n)} &= \frac{\sum_{s=0}^{n} s\binom{n}{s}\rho^s}{(1+\rho)^n} \\
&= (1+\rho)^{-n}\sum_{s=1}^{n} s\frac{n}{s}\binom{n-1}{s-1}\rho^s \\
&= \rho(1+\rho)^{-n}n\sum_{s=1}^{n}\binom{n-1}{s-1}\rho^{s-1} \\
&= \rho(1+\rho)^{-n}n\sum_{x=0}^{n-1}\binom{n-1}{x}\rho^{x} \\
&= \rho(1+\rho)^{-n}n(1+\rho)^{n-1} \\
&= \frac{\rho}{1+\rho}n.
\end{aligned}$$

A.4 Average product length given r including ρ

The total product length is given by

$$\begin{aligned}
d(n,r)\bar{s}(n,r) &= \sum_{s=n-r}^{n} s\binom{n}{s}\rho^{s} \\
&= n\rho \sum_{s=n-r}^{n} \frac{(n-1)!}{(s-1)!(n-s)!}\rho^{s-1} \\
&= n\rho \sum_{s'=n-r-1}^{n-1} \frac{(n-1)!}{s'!(n-s'-1)!}\rho^{s'} \\
&= n\rho \sum_{s'=n-r-1}^{n-1} \binom{n-1}{s'}\rho^{s'} \\
&= n\rho d(n-1,r)
\end{aligned}$$

so that the average product length is given by

$$\bar{s}(n,r) = n\rho\frac{d(n-1,r)}{d(n,r)}.$$

A.5 Bounds on average word length

We show that

$$\frac{\rho}{1+\rho}n < \bar{s}(n,r) < n.$$

The growth in product variety is given by

$$\begin{aligned}
d(n+1,r)-d(n,r) &= \sum_{s=n+1-r}^{n+1}\binom{n+1}{s}\rho^s - \sum_{s=n-r}^{n}\binom{n}{s}\rho^s \\
&= \sum_{s=n+1-r}^{n+1}\binom{n+1}{s}\rho^s - \sum_{s=n-r}^{n+1}\frac{n+1-s}{n+1}\binom{n+1}{s}\rho^s \\
&= \sum_{s=n+1-r}^{n+1}\binom{n+1}{s}\rho^s - \sum_{s=n+1-r}^{n+1}\frac{n+1-s}{n+1}\binom{n+1}{s}\rho^s - \frac{r+1}{n+1}\binom{n+1}{n-r}\rho^{n-r} \\
&= \sum_{s=n+1-r}^{n+1}\frac{s}{n+1}\binom{n+1}{s}\rho^s - \binom{n}{r}\rho^{n-r} \\
&= \rho d(n,r) - \binom{n}{r}\rho^{n-r}.
\end{aligned}$$

This leads to the identity

$$d(n,r) = (1+\rho)d(n-1,r) - \binom{n-1}{r}\rho^{n-r-1},$$

so that

$$\begin{aligned}
\frac{d(n-1,r)}{d(n,r)} &= \frac{d(n-1,r)}{(1+\rho)d(n-1,r) - \binom{n-1}{r}\rho^{n-r-1}} \\
&> \frac{d(n-1,r)}{(1+\rho)d(n-1,r)} = \frac{1}{1+\rho}.
\end{aligned}$$

This means that a lower bound on average word length is given by

$$\bar{s}(n,r) = n\rho\frac{d(n-1,r)}{d(n,r)} > n\frac{\rho}{1+\rho}$$

since $\binom{n-1}{r}\rho^{n-r-1} > 0$.

To find an upper bound, first note that

$$\binom{n-1}{r}\rho^{n-r-1} < \sum_{s=n-r-1}^{n-1}\binom{n}{s}\rho^s = d(n-1,r).$$

We then find that

$$\begin{aligned} d(n,r) &= (1+\rho)d(n-1,r) - \binom{n-1}{r}\rho^{n-r-1} \\ &= (1+\rho)d(n-1,r) - \binom{n-1}{n-1-r}\rho^{n-r-1} \\ &> (1+\rho)d(n-1,r) - d(n-1,r) \\ &> \rho d(n-1,r), \end{aligned}$$

so that

$$\frac{d(n-1,r)}{d(n,r)} < \frac{1}{\rho},$$

and

$$\bar{s}(n,r) = n\rho\frac{d(n-1,r)}{d(n,r)} < n.$$

A.6 Diversification including r

From A.5 we have that

$$d(n+1,r) - d(n,r) = \rho d(n,r) - \binom{n}{r}\rho^{n-r}.$$

Hence variety starts decreasing for n, r when

$$d(n,r) < \binom{n}{r}\rho^{n-r-1}.$$

Bibliography

Aghion, P., Boulanger, J., and Cohen, E. Rethinking Industrial Policy. *Bruegel Policy Brief*, 04, 2011.

Aiginger, K. Industrial policy: Past, diversity, future; introduction to the special issue on the future of industrial policy. *Journal of Industry, Competition and Trade*, 7 (3-4):143–146, 2007. doi:10.1007/s10842-007-0023-9.

Al-Marhubi, F. Export diversification and growth: an empirical investigation. *Applied Economics Letters*, 7(9):559–562, 2000. doi:10.1080/13504850050059005.

Albeaik, S., Kaltenberg, M., Alsaleh, M., and Hidalgo, C. Improving the Economic Complexity Index. *arxiv.org/abs/1707.05826*, 2017a.

Albeaik, S., Kaltenberg, M., Alsaleh, M., and Hidalgo, C. 729 new measures of economic complexity (Addendum to Improving the Economic Complexity Index). *arxiv.org/abs/1708.04107*, 2017b.

Alshamsi, A., Pinheiro, F., and Hidalgo, C. Optimal diversification strategies in the networks of related products and of related research areas. *Nature Communications*, 9(1), 2018. doi:10.1038/s41467-018-03740-9.

Bailey, D., Glasmeier, A., Tomlinson, P. R., and Tyler, P. Industrial policy: New technologies and transformative innovation policies? *Cambridge Journal of Regions, Economy and Society*, 12(2):169–177, 2019. doi:10.1093/cjres/rsz006.

Balassa, B. Trade Liberalisation and ”Revealed” Comparative Advantage. *The Manchester School*, 33(2):99–123, 1965. doi:10.1111/j.1467-9957.1965.tb00050.x.

Ballance, R., Forstner, H., and Murray, T. Consistency Tests of Alternative Measures of Comparative Advantage. *The Review of Economics and Statistics*, 69(1):157, 1987. doi:10.2307/1937915.

Balland, P.-A. and Rigby, D. The Geography of Complex Knowledge. *Economic Geography*, 93(1):1–23, 2017. doi:10.1080/00130095.2016.1205947.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., and Hidalgo, C. A. Complex economic activities concentrate in large cities. *Nature Human Behaviour*, 4(3):248–254, 2020. doi:10.1038/s41562-019-0803-3.

Baudena, M., Sanchez, A., Georg, C.-P., Ruiz-Benito, P., Rodriguez, M. A., Zavala, M. A., and Rietkerk, M. Revealing patterns of local species richness along environmental gradients with a novel network tool. *Scientific Reports*, 5(1):11561, 2015. doi:10.1038/srep11561.

Beh, E. J. Simple Correspondence Analysis: A Bibliographic Review. *International Statistical Review*, 72(2):257–284, 2004. doi:10.1111/j.1751-5823.2004.tb00236.x.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi:10.1162/089976603321780317.

Benzécri, J.-P. and Coll. *L'analyse des données. Vol. 2: Analyse des Correspondances*, volume 2. Dunod, Paris, 1973.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. doi:10.1162/jmlr.2003.3.4-5.993.

Boschma, R. Relatedness as driver of regional diversification: a research agenda. *Regional Studies*, 51(3):351–364, 2017. doi:10.1080/00343404.2016.1254767.

Boschma, R., Minondo, A., and Navarro, M. The Emergence of New Industries at the Regional Level in Spain: A Proximity Approach Based on Product Relatedness. *Economic Geography*, 89(1):29–51, 2013. doi:10.1111/j.1944-8287.2012.01170.x.

Bresnahan, T. F. and Trajtenberg, M. General purpose technologies 'Engines of growth'? *Journal of Econometrics*, 65(1):83–108, 1995. doi:10.1016/0304-4076(94)01598-T.

Brummitt, C. D., Huremović, K., Pin, P., Bonds, M. H., and Vega-Redondo, F. Contagious disruptions and complexity traps in economic development. *Nature Human Behaviour*, 1(9):665–672, 2017. doi:10.1038/s41562-017-0190-6.

Brummitt, C. D., Gómez-Liévano, A., Hausmann, R., and Bonds, M. H. Machine-learned patterns suggest that diversification drives economic development. *Journal of the Royal Society Interface*, 17(162), 2020. doi:10.1098/rsif.2019.0283.

Bustos, S. and Yildirim, M. A. Production Ability and Economic Growth. *CID Working Papers*, 110, 2019.

Cadot, O., Carrère, C., and Strauss-Kahn, V. Export Diversification: What's behind the Hump? *Review of Economics and Statistics*, 93(2):590–605, 2011. doi:10.1162/REST_a_00078.

Cadot, O., Carrère, C., and Strauss-Kahn, V. Trade diversification, income, and growth: what do we know? *Journal of Economic Surveys*, 27(4):790–812, 2013. doi:10.1111/j.1467-6419.2011.00719.x.

Caldarelli, G., Cristelli, M., Gabrielli, A., Pietronero, L., Scala, A., and Tacchella, A. A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. *PLoS ONE*, 7(10):1–11, 2012. doi:10.1371/journal.pone.0047278.

Chang, H. J. *Kicking Away the Ladder: Development Strategy in Historical Perspective*. Anthem Studies in Development and Globalization. Anthem Press, 2002. ISBN 9780857287618.

Chang, H. J. and Andreoni, A. Industrial Policy in the 21st Century. *Development and Change*, 51(2):324–351, 2020. doi:10.1111/dech.12570.

Chao, A., Chiu, C.-H., and Jost, L. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):297–324, 2014. doi:10.1146/annurev-ecolsys-120213-091540.

Chávez, J. C., Mosqueda, M. T., and Gómez-Zaldívar, M. Economic complexity and regional growth performance: Evidence from the Mexican economy. *Review of Regional Studies*, 47(2):201–219, 2017.

Chiu, C. H., Jost, L., and Chao, A. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, 84(1):21–44, 2014. doi:10.1890/12-0960.1.

Chung, F. *Spectral Graph Theory*. American Mathematical Society, Providence, Rhode Island, 1997. ISBN 9780821803158.

Church, K. and Hanks, P. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, volume 16, pages 76–83, Morristown, NJ, 1989. Association for Computational Linguistics. ISBN 0891-2017. doi:10.3115/981623.981633.

Cimoli, M., Dosi, G., and Stiglitz, J. E. *Industrial Policy and Development*. Oxford University Press, Oxford; Toronto, 2009. ISBN 9780199235261. doi:10.1093/acprof:oso/9780199235261.001.0001.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi:10.1016/j.acha.2006.04.006.

Content, J. and Frenken, K. Related variety and economic development: a literature review. *European Planning Studies*, 24(12):2097–2112, 2016. doi:10.1080/09654313.2016.1246517.

Coscia, M. and Neffke, F. M. Network backboning with noisy data. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 425–436, 2017. doi:10.1109/ICDE.2017.100.

Cover, T. and Thomas, J. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2005. ISBN 9780471241959. doi:10.1002/047174882X.

Cristelli, M., Tacchella, A., and Pietronero, L. The Heterogeneous Dynamics of Economic Complexity. *PLOS ONE*, 10(2):e0117174, 2015. doi:10.1371/journal.pone.0117174.

Daly, A., Baetens, J., and De Baets, B. Ecological Diversity: Measuring the Unmeasurable. *Mathematics*, 6(7):119, 2018. doi:10.3390/math6070119.

Daru, B. H., Elliott, T. L., Park, D. S., and Davies, T. J. Understanding the Processes Underpinning Patterns of Phylogenetic Regionalization. *Trends in Ecology & Evolution*, 32(11):845–860, 2017. doi:10.1016/j.tree.2017.08.013.

Diniz-Filho, J. A. F., Rodriguez, M. A., Bini, L. M., Olalla-Tarraga, M. A., Cardillo, M., Nabout, J. C., Hortal, J., and Hawkins, B. A. Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. *Journal of Biogeography*, 36(12):2222–2236, 2009. doi:10.1111/j.1365-2699.2009.02163.x.

Diodato, D., Neffke, F., and O'Clery, N. Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26, 2018. doi:10.1016/j.jue.2018.05.002.

Dixit, A. K. and Stiglitz, J. E. Monopolistic competition and optimum product diversity. *The American Economic Review*, 67(3):297–308, 1977.

Duranton, G. and Puga, D. Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2063–2117. Elsevier, 2004.

Ellison, G. and Glaeser, E. Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy*, 105(5):889–927, 1997. doi:10.1086/262098.

Ellison, G. and Glaeser, E. The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration? *American Economic Review*, 89(2):311–316, 1999. doi:10.1257/aer.89.2.311.

Ellison, G., Glaeser, E., and Kerr, W. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review*, 100(3):1195–1213, 2010. doi:10.1257/aer.100.3.1195.

Essletzbichler, J. Relatedness, Industrial Branching and Technological Cohesion in US Metropolitan Areas. *Regional Studies*, 49(5):752–766, 2015. doi:10.1080/00343404.2013.806793.

Ester, M., Kriegel, H.-P., Sanger, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231. AAI Press, 1996.

Faggio, G., Silva, O., and Strange, W. Heterogeneous Agglomeration. *Review of Economics and Statistics*, 99(1):80–94, 2017. doi:10.1162/REST_a_00604.

Fano, R. *Transmission of Information: A statistical theory of communications.* Wiley, New York, 1961. ISBN 9780262060011.

Faurby, S., Davis, M., Pedersen, R., Schowanek, S. D., Antonelli, A., and Svenning, J. C. PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. *Ecology*, 99(11):2626, 2018. doi:10.1002/ecy.2443.

Fink, T. M. A. and Reeves, M. How much can we influence the rate of innovation? *Science Advances*, 5(1):eaat6107, 2019. doi:10.1126/sciadv.aat6107.

Fink, T. M. A., Reeves, M., Palma, R., and Farr, R. S. Serendipity and strategy in rapid innovation. *Nature Communications*, 8(1):2002, 2017. doi:10.1038/s41467-017-02042-w.

Finn, C. and Lizier, J. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. *Entropy*, 20(4):297, 2018. doi:10.3390/e20040297.

Fisher, R. A. The precision of discriminant functions. *Annals of Eugenics*, 10(1): 422–429, 1940. doi:10.1111/j.1469-1809.1940.tb02264.x.

Foley, R. A. and Mirazon Lahr, M. The evolution of the diversity of cultures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1080–1089, 2011. doi:10.1098/rstb.2010.0370.

Foray, D. On sector-non-neutral innovation policy: towards new design principles. *Journal of Evolutionary Economics*, 29(5):1379–1397, 2019. doi:10.1007/s00191-018-0599-8.

Fouss, F., Saerens, M., and Shimbo, M. *Algorithms and Models for Network Data and Link Analysis*. Cambridge University Press, Cambridge, 2016. ISBN 9781316418321. doi:10.1017/CBO9781316418321.

Freeman, C. *Technology, policy, and economic performance: lessons from Japan*. Pinter Publishers, 1987. ISBN 9780861879281.

Frenken, K., Van Oort, F., and Verburg, T. Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies*, 41(5):685–697, 2007. doi:10.1080/00343400601120296.

Gabrielli, A., Cristelli, M., Mazzilli, D., Tacchella, A., Zaccaria, A., and Pietronero, L. Why we like the ECI+ algorithm. *arxiv.org/abs/1708.01161*, 2017.

Gao, J. and Zhou, T. Quantifying China's regional economic complexity. *Physica A: Statistical Mechanics and its Applications*, 492:1591–1603, 2018. doi:10.1016/j.physa.2017.11.084.

Gao, J., Barzel, B., and Barabási, A.-L. Universal resilience patterns in complex networks. *Nature*, 530(7590):307–312, 2016. doi:10.1038/nature16948.

Gauch, H., Whittaker, R., and Wentworth, T. A Comparative Study of Reciprocal Averaging and Other Ordination Techniques. *The Journal of Ecology*, 65(1):157, 1977. doi:10.2307/2259071.

Gerlach, M., Peixoto, T. P., and Altmann, E. G. A network approach to topic models. *Science Advances*, 4(7):eaaq1360, 2018. doi:10.1126/sciadv.aaq1360.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., and Shleifer, A. Growth in Cities. *Journal of Political Economy*, 100(6):1126–1152, 1992. doi:10.1086/261856.

Gomez-Lievano, A. Methods and Concepts in Economic Complexity. *arxiv.org/abs/1809.10781*, 2018.

Gomez-Lievano, A. and Patterson-Lomba, O. Uncovering the drivers behind urban economic complexity and their connection to urban economic performance. *arxiv.org/pdf/1812.02842.pdf*, 2018.

Gomez-Lievano, A., Patterson-Lomba, O., and Hausmann, R. Explaining the prevalence, scaling and variance of urban phenomena. *Nature Human Behaviour*, 1(1): 0012, 2016. doi:10.1038/s41562-016-0012.

Gordon, R. J. *The Rise and Fall of American Growth.* Princeton University Press, Princeton, 2016. ISBN 9781400873302. doi:10.1515/9781400873302.

Greenacre, M. *Correspondence Analysis in Practice.* Taylor & Francis Group, Boca Raton, FL, 2007. ISBN 1-58488-616-1.

Greenacre, M. Power transformations in correspondence analysis. *Computational Statistics and Data Analysis*, 53(8):3107–3116, 2009. doi:10.1016/j.csda.2008.09.001.

Greenacre, M. and Lewi, P. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification*, 26(1):29–54, 2009. doi:10.1007/s00357-009-9027-y.

Greenacre, M. J. *Theory and Applications of Correspondence Analysis.* Academic Press, Orlando, FL, 1984. ISBN 0122990501.

Greenacre, M. J. Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):613–619, 2010. doi:10.1002/wics.114.

Hausmann, R. and Hidalgo, C. A. The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342, 2011. doi:10.1007/s10887-011-9071-4.

Hausmann, R. and Klinger, B. Structural transformation and patters of comparative advantage in in the product space. *CID Working Papers*, 128, 2006.

Hausmann, R., Hwang, J., and Rodrik, D. What you export matters. *Journal of Economic Growth*, 12(1):1–25, 2007. doi:10.1007/s10887-006-9009-4.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., and Yildirim, M. A. *The Atlas of Economic Complexity*. Puritan Press, New Hampshire, 2011. ISBN 9780262525428.

Herzer, D. and Nowak-Lehnmann, F. D. What does export diversification do for growth? An econometric analysis. *Applied Economics*, 38(15):1825–1838, 2006. doi:10.1080/00036840500426983.

Hesse, H. Export Diversification. *Commission on Growth and Development Working Paper*, 21, 2008.

Hidalgo, C. A. and Hausmann, R. A Network View of Economic Development. *Developing Alternatives*, 12(1):5–10, 2008.

Hidalgo, C. A. and Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009. doi:10.1073/pnas.0900943106.

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., and Zhu, S. The Principle of Relatedness. In Morales, A. J., Gershenson, C., Braha, D., Minai, A. A., and Bar-Yam, Y., editors, *Unifying Themes in Complex Systems IX*, pages 451–457. Springer International Publishing, Cham, 2018. ISBN 978-3-319-96661-8. doi:10.1007/978-3-319-96661-8_46.

Hidalgo, C. H., Klinger, B., Barabási, A.-L., and Hausmann, R. The Product Space Conditions the Development of Nations. *Science*, 317(5837):482–487, 2007. doi:10.1126/science.1144581.

Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432, 1973a. doi:10.2307/1934352.

Hill, M. O. Reciprocal Averaging: An Eigenvector Method of Ordination. *The Journal of Ecology*, 61(1):237, 1973b. doi:10.2307/2258931.

Hill, M. O. Correspondence Analysis: A Neglected Multivariate Method. *Applied Statistics*, 23(3):340, 1974. doi:10.2307/2347127.

Hirschfeld, H. O. A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935a. doi:10.1017/S0305004100013517.

Hirschfeld, H. O. A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935b. doi:10.1017/S0305004100013517.

Holt, B. G., Lessard, J.-P., Borregaard, M. K., Fritz, S. A., Araújo, M. B., Dimitrov, D., Fabre, P.-H., Graham, C. H., Graves, G. R., Jønsson, K. A., Nogués-Bravo, D., Wang, Z., Whittaker, R. J., Fjeldså, J., and Rahbek, C. An Update of Wallace's Zoogeographic Regions of the World. *Science*, 339(6115):74–78, 2013. doi:10.1126/science.1228282.

Hoover, E. The Measurement of Industrial Localization. *The Review of Economics and Statistics*, 18(4):162, 1936. doi:10.2307/1927875.

Hotelling, H. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321, 1936. doi:10.2307/2333955.

Hutter, M. and Zaffalon, M. Distribution of mutual information from complete and incomplete data. *Computational Statistics & Data Analysis*, 48(3):633–657, 2005. doi:10.1016/j.csda.2004.03.010.

Imbs, J. and Wacziarg, R. Stages of diversification. *American Economic Review*, 93 (1):63–86, 2003. doi:10.1257/000282803321455160.

Inoua, S. A Simple Measure of Economic Complexity. *arxiv.org/abs/1601.05012*, 2016.

Isard, W. *Methods of Regional Analysis: an Introduction to Regional Science.* MIT Press, Cambridge, MA, 1960.

Jacobs, J. *The economy of cities.* Random House, New York, 1969.

Jara-Figueroa, C., Jun, B., Glaeser, E. L., and Hidalgo, C. A. The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proceedings of the National Academy of Sciences*, 115(50): 12646–12653, 2018. doi:10.1073/pnas.1800475115.

Jones, B. F. The Burden of Knowledge and the “Death of the Renaissance Man”: Is Innovation Getting Harder? *Review of Economic Studies*, 76(1):283–317, 2009. doi:10.1111/j.1467-937X.2008.00531.x.

Jost, L. Entropy and diversity. *Oikos*, 113(2):363–375, 2006. doi:10.1111/j.2006.0030-1299.14714.x.

Jost, L. Partitioning diversity into independent alpha en beta components. *Ecology*, 88(10):2427–2439, 2007. doi:10.1890/06-1736.1.

Jost, L. The relation between evenness and diversity. *Diversity*, 2(2):207–232, 2010. doi:10.3390/d2020207.

Kemp-Benedict, E. An interpretation and critique of the Method of Reflections. *MPRA Paper No. 60705*, 2014.

Klinger, B. and Lederman, D. Discovery and Development: An Empirical Exploration of ”New” Products. *World Bank Policy Research Working Paper*, 3450(November 2004):1–48, 2004. doi:10.1596/1813-9450-3450.

Kogler, D. F., Rigby, D. L., and Tucker, I. Mapping Knowledge Space and Technological Relatedness in US Cities. *European Planning Studies*, 21(9):1374–1391, 2013. doi:10.1080/09654313.2012.755832.

Kremer, M. The O-Ring Theory of Economic Development. *The Quarterly Journal of Economics*, 108(3):551–575, 1993. doi:10.2307/2118400.

Krugman, P. Increasing Returns and Economic Geography. *Journal of Political Economy*, 99(3):483–499, 1991a. doi:10.1086/261763.

Krugman, P. *Geography and trade*. MIT Press, Cambridge, MA, 1991b. ISBN 9780262111591.

Kullback, S. and Leibler, R. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi:10.1214/aoms/1177729694.

Kunimoto, K. Typology of Trade Intensity Indices. *Hitotsubashi Journal of Economics*, 17(2):15–32, 1977. doi:10.15057/7981.

Lall, S. The technological structure and performance of developing country manufactured exports, 1985-98. *Oxford Development Studies*, 28(3):337–369, 2000. doi:10.1080/713688318.

Lall, S., Weiss, J., and Zhang, J. The "sophistication" of exports: A new trade measure. *World Development*, 34(2):222–237, 2006. doi:10.1016/j.worlddev.2005.09.002.

Lane, N. The New Empirics of Industrial Policy. *Journal of Industry, Competition and Trade*, 20(2):209–234, 2020. doi:10.1007/s10842-019-00323-2.

Legendre, P. and Legendre, L. Ordination in reduced space. In *Numerical Ecology*, pages 425–520. Elsevier Science B.V., Amsterdam, second edition, 1998. doi:10.1016/B978-0-444-53868-0.50009-5.

Leinster, T. and Cobbold, C. A. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012. doi:10.1890/10-2402.1.

Leontief, W. *Input-output economics.* Oxford University Press, New York, 1966.

Leydesdorff, L., Wagner, C. S., and Bornmann, L. Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics*, pages 2–3, 2019. doi:10.1016/j.joi.2019.03.016.

Lundvall, B. *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning.* Pinter Publishers, 1992. ISBN 9781855670631.

MacArthur, R. Fluctuations of Animal Populations and a Measure of Community Stability. *Ecology*, 36(3):533, 1955. doi:10.2307/1929601.

Manning, C., Raghavan, P., and Schutze, H. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, 2008. ISBN 9780511809071. doi:10.1017/CBO9780511809071.

Marshall, A. *Principles of Economics.* MacMillan, London, 8 edition, 1920.

Mazzucato, M. *The entrepreneurial state.* Demos, London, 2011. ISBN 9781906693732. doi:10.3898/136266211798411183.

Mazzucato, M. Mission-oriented innovation policies: Challenges and opportunities. *Industrial and Corporate Change*, 27(5):803–815, 2018. doi:10.1093/icc/dty034.

Mazzucato, M., Cimoli, M., Dosi, G., Stiglitz, J. E., Landesmann, M. A., Pianta, M., Walz, R., and Page, T. Which Industrial Policy Does Europe Need? *Intereconomics*, 50(3):120–155, 2015. doi:10.1007/s10272-015-0535-1.

Mealy, P. and Coyle, D. To Them That Hath: Economic Complexity and Local Industrial Strategy in the UK. *SSRN Electronic Journal*, (November):1–28, 2019. doi:10.2139/ssrn.3491153.

Mealy, P., Farmer, J. D., and Teytelboym, A. Interpreting economic complexity. *Science Advances*, 5(1):eaau1705, 2019. doi:10.1126/sciadv.aau1705.

Morales-Castilla, I., Olalla-Tárraga, M., Purvis, A., Hawkins, B., and Rodríguez, M. The imprint of Cenozoic migrations and evolutionary history on the biogeographic gradient of body size in new world mammals. *American Naturalist*, 180(2):246–256, 2012. doi:10.1086/666608.

Morales-Castilla, I., Davies, T. J., Pearse, W. D., and Peres-Neto, P. Combining phylogeny and co-occurrence to improve single species distribution models. *Global Ecology and Biogeography*, 26(6):740–752, 2017. doi:10.1111/geb.12580.

Mori, T., Nishikimi, K., and Smith, T. A divergence statistic for industrial localization. *Review of Economics and Statistics*, 87(4):635–651, 2005. doi:10.1162/003465305775098170.

Morrison, G., Buldyrev, S. V., Imbruno, M., Doria Arrieta, O. A., Rungi, A., Riccaboni, M., and Pammolli, F. On Economic Complexity and the Fitness of Nations. *Scientific Reports*, 7(1):15332, 2017. doi:10.1038/s41598-017-14603-6.

Muneepeerakul, R., Lobo, J., Shutters, S. T., Goméz-Liévano, A., and Qubbaj, M. R. Urban Economies and Occupation Space: Can They Get “There” from “Here”? *PLoS ONE*, 8(9):e73676, 2013. doi:10.1371/journal.pone.0073676.

Muthukrishna, M. and Henrich, J. Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690):20150192, 2016. doi:10.1098/rstb.2015.0192.

Nadarajah, S. and Kotz, S. Exact and Approximate Distributions for the Product of Dirichlet Components. *Kybernetika*, 40(6):735–744, 2004.

Neffke, F. and Henning, M. Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3):297–316, 2013. doi:10.1002/smj.2014.

Neffke, F., Henning, M., and Boschma, R. How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Economic Geography*, 87(3):237–265, 2011. doi:10.1111/j.1944-8287.2011.01121.x.

Neffke, F., Hartog, M., Boschma, R., and Henning, M. Agents of Structural Change: The Role of Firms and Entrepreneurs in Regional Diversification. *Economic Geography*, 94(1):23–48, 2018. doi:10.1080/00130095.2017.1391691.

Neffke, F. M., Otto, A., and Weyh, A. Inter-industry labor flows. *Journal of Economic Behavior & Organization*, 142:275–292, 2017. doi:10.1016/j.jebo.2017.07.003.

Newman, M. E. J. Spectral methods for community detection and graph partitioning. *Physical Review E*, 88(4):042822, 2013. doi:10.1103/PhysRevE.88.042822.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On Spectral Clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, pages 849 – 856, 2002.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., and van den Oord, A. Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7): 1016–1034, 2007. doi:10.1016/j.respol.2007.04.003.

Page, S. E. *Diversity and Complexity.* Princeton University Press, 2011. ISBN 9780691137674.

Pietronero, L., Cristelli, M., Gabrielli, A., Mazzilli, D., Pugliese, E., Tacchella, A., and Zaccaria, A. Economic Complexity: ”Buttarla in caciara” vs a constructive approach. *arxiv.org/abs/1709.05272*, 2017.

Purvis, A. and Hector, A. Getting the measure of biodiversity. *Nature*, 405(6783): 212–219, 2000. doi:10.1038/35012221.

Rafols, I. and Meyer, M. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010. doi:10.1007/s11192-009-0041-y.

Rao, C. R. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982. doi:10.1016/0040-5809(82)90004-1.

Renyi, A. and Rényi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 1 of *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, Berkeley, Calif., 1961. University of California Press. ISBN 0097-0433.

Rodrik, D. Industrial Policy for the 21st Century. *Unido*, (3-4):122–126, 2004.

Romer, P. M. Growth Based on Increasing Returns Due to Specialization. *The American Economic Review*, 77(2):56–62, 1987.

Roy, A. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, 3 (2):135–146, 1951.

Rueda, M., Rodríguez, M., and Hawkins, B. Identifying global zoogeographical regions: Lessons from Wallace. *Journal of Biogeography*, 40(12):2215–2225, 2013. doi:10.1111/jbi.12214.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. IPUMS USA: Version 8.0 [dataset]. *Minneapolis, MN: IPUMS*, 2018. doi:10.18128/D010.V8.0.

Sattinger, M. Assignment Models of the Distribution of Earnings. *Journal of Economic Literature*, 31(2):831–880, 1993.

Saviotti, P. P. *Technological Evolution, Variety and the Economy.* Edward Elgar Publishing, Cheltenham, 1996.

Saviotti, P. P. and Frenken, K. Export variety and the economic performance of countries. *Journal of Evolutionary Economics*, 18(2):201–218, 2008. doi:10.1007/s00191-007-0081-5.

Schetter, U. A Structural Ranking of Economic Complexity. *CID Working Papers*, 119, 2019. doi:10.2139/ssrn.3485842.

Schetter, U. Quality Differentiation and Comparative Advantage. *CID Working Papers*, 126, 2020. doi:10.2139/ssrn.3091581.

Seung-Seok, C., Sung-Hyuk, C., Tappert, C. C., and Science, C. A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics & Informatics*, 8(1):43–48, 2010. doi:10.1.1.352.6123.

Shannon, C. E. *A mathematical theory of communication*, volume 5. 2001. ISBN 0252725484. doi:10.1145/584091.584093.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. doi:10.1109/34.868688.

Smits, R. and Kuhlmann, S. The rise of systemic instruments in innovation policy. *International Journal of Foresight and Innovation Policy*, 1(1-2):4–32, 2004. doi:10.1504/IJFIP.2004.004621.

Stirling, A. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, 2007. doi:10.1098/rsif.2007.0213.

Sutton, J. and Trefler, D. Capabilities, wealth, and trade. *Journal of Political Economy*, 124(3):826–878, 2016. doi:10.1086/686034.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. A New Metrics for Countries' Fitness and Products' Complexity. *Scientific Reports*, 2(1): 723, 2012. doi:10.1038/srep00723.

Tacchella, A., Mazzilli, D., and Pietronero, L. A dynamical systems approach to gross domestic product forecasting. *Nature Physics*, 14(August):861–866, 2018. doi:10.1038/s41567-018-0204-y.

ter Braak, C. J. F. Ordination. In Jongman, R. H. G., Braak, C. J. F. T., and van Tongeren, O. F. R., editors, *Data analysis in community and landscape ecology*, chapter 5, pages 0–5. Cambridge University Press, 1995. ISBN 5212522363326.

The Growth Lab at Harvard University. International Trade Data (HS, 92), 2019.

Theil, H. *Economics and information theory.* North-Holland Publishing Company, Amsterdam, 1967.

Theil, H. *Statistical decomposition analysis: With applications in the social and administrative sciences.* North-Holland Publishing Company, Amsterdam, 1972.

Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(2):411–423, 2001. doi:10.1111/1467-9868.00293.

Tilman, D., Isbell, F., and Cowles, J. M. Biodiversity and Ecosystem Functioning. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):471–493, 2014. doi:10.1146/annurev-ecolsys-120213-091917.

Tuomisto, H. A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia*, 164(4):853–860, 2010. doi:10.1007/s00442-010-1812-0.

van Dam, A. Diversity and its decomposition into variety, balance and disparity. *Royal Society Open Science*, 6(7):190452, 2019. doi:10.1098/rsos.190452.

van Dam, A. and Frenken, K. Variety, complexity and economic development. *Research Policy*, 2020a. doi:10.1016/j.respol.2020.103949.

van Dam, A. and Frenken, K. Vertical vs. Horizontal Policy in a Capabilities Model of Economic Development. *arxiv.org/abs/2006.04624*, 2020b.

van Dam, A., Gomez-Lievano, A., Neffke, F., and Frenken, K. An information-theoretic approach to the analysis of location and co-location patterns. *arxiv.org/abs/2004.10548*, 2020.

van den Bergh, J. C. Optimal diversity: Increasing returns versus recombinant innovation. *Journal of Economic Behavior & Organization*, 68(3-4):565–580, 2008. doi:10.1016/j.jebo.2008.09.003.

van Eck, N. J. and Waltman, L. 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, 2009. doi:10.1002/asi.21075.

Vernon, R. International Investment and International Trade in the Product Cycle. *The Quarterly Journal of Economics*, 80(2):190–207, 1966. doi:10.2307/1880689.

Vollrath, T. A theoretical evaluation of alternative trade intensity measures of revealed comparative advantage. *Weltwirtschaftliches Archiv*, 127(2):265–280, 1991. doi:10.1007/BF02707986.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. doi:10.1007/s11222-007-9033-z.

Wang, J., Thijs, B., and Glänzel, W. Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE*, 10(5):1–18, 2015. doi:10.1371/journal.pone.0127298.

Weitzman, M. L. Recombinant Growth. *The Quarterly Journal of Economics*, 113 (2):331–360, 1998. doi:10.1162/003355398555595.

Whittaker, R. H. Gradient analysis of vegetation. *Biological Reviews*, 42(2):207–264, 1967. doi:10.1111/j.1469-185X.1967.tb01419.x.

Whittaker, R. H. Evolution and Measurement of Species Diversity. *Taxon*, 21(2/3): 213, 1972. doi:10.2307/1218190.

Wieczorek, A. J. and Hekkert, M. P. Systemic instruments for systemic innovation problems: A framework for policy makers and innovation scholars. *Science and Public Policy*, 39(1):74–87, 2012. doi:10.1093/scipol/scr008.

Wolpert, D. and Wolf, D. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995. doi:10.1103/PhysRevE.52.6841.

Yeats, A. On the appropriate interpretation of the revealed comparative advantage index: Implications of a methodology based on industry sector analysis. *Weltwirtschaftliches Archiv*, 121(1):61–73, 1985. doi:10.1007/BF02705840.

Yen, L., Saerens, M., and Fouss, F. A link analysis extension of correspondence analysis for mining relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 23(4):481–495, 2011. doi:10.1109/TKDE.2010.142.

Yildirim, M. A. and Coscia, M. Using Random Walks to Generate Associations between Objects. *PLoS ONE*, 9(8):e104813, 2014. doi:10.1371/journal.pone.0104813.

Yu, S. X. and Shi, J. Multiclass spectral clustering. *Proceedings of the IEEE International Conference on Computer Vision*, 1(2):313–319, 2003. doi:10.1109/iccv.2003.1238361.

Zaccaria, A., Cristelli, M., Kupers, R., Tacchella, A., and Pietronero, L. A case study for a new metrics for economic complexity: The Netherlands. *Journal of Economic Interaction and Coordination*, 2015. doi:10.1007/s11403-015-0145-9.

Zelnik-Manor, L. and Perona, P. Self-Tuning Spectral Clustering. In *Advances in Neural Information Processing Systems*, number 17, 2004.

Zha, H., He, X., Ding, C., Gu, M., and Simon, H. Bipartite graph partitioning and data clustering. *International Conference on Information and Knowledge Management, Proceedings*, pages 25–32, 2001. doi:10.1145/502590.502591.

Zhang, L., Rousseau, R., and Glänzel, W. Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67 (5):1257–1265, 2016. doi:10.1002/asi.23487.

Nederlandse samenvatting

Traditionele modellen van economische ontwikkeling gaan typisch uit van een productiefunctie waarin kapitaal en arbeid leiden tot economische waarde. Die waarde wordt vaak uitgedrukt in hoe groot een economie is, bijvoorbeeld door middel van het bruto binnenlands product. Dit proefschrift start met een alternatief model van economische ontwikkeling waarin niet de grootte van een economie centraal staat, maar haar structuur. Dat wil zeggen dat er niet gekeken wordt naar de totale productie in een economie, maar naar de specifieke producten en diensten die zij in staat is te leveren.

Dit alternatieve model gaat ervan uit dat er voor het produceren van een specifiek product of dienst altijd een aantal complementaire competenties nodig is. Deze competenties worden geïnterpreteerd in de breedste zin van het woord: voor het maken van een bepaald product zijn er bijvoorbeeld bepaalde grondstoffen nodig maar ook specifieke kennis, technologie, institutionele condities en infrastructuur. Competenties kunnen dus geïnteresseerd worden als de bouwstenen die aanwezig moeten zijn in een economie om het mogelijke te maken verschillende producten en diensten te leveren.

Economische ontwikkeling berust dan op het ontwikkelen van de juiste competenties, en die op een juiste manier met elkaar combineren. Nieuwe competenties kunnen gerecombineerd worden met de reeds aanwezige competenties, en leiden zo tot nieuwe producten. Het vergaren van meer competenties maakt het zo mogelijk om steeds meer verschillende en ook complexe producten te maken, die gebruik maken van meer competenties. Dit model leidt tot drie begrippen die centraal staan in het beschrijven van economische ontwikkeling.

Ten eerste is er de notie van diversiteit. Naarmate een economie meer competenties ontwikkelt, groeit het aantal combinaties dat gemaakt kan worden met deze competenties exponentieel. Het resultaat is dat economische ontwikkeling gepaard gaat met een grote toename in het aantal verschillende producten dat gemaakt kan worden. Ten tweede stelt een toenemend aantal competenties een economie in staat om steeds complexere producten te maken, die bestaan uit combinaties van steeds meer verschillende competenties. Tot slot leidt het model tot de notie van gerelateerdheid tussen producten, waarmee bedoeld wordt dat producten grotendeels berusten op dezelfde competenties, en dus op elkaar lijken in termen van wat er nodig is om te ze te produceren. Het ontwikkelen van extra competenties stelt economieën dus in staat om nieuwe producten te maken, maar deze zullen wel gerelateerd zijn aan de producten die ze al maakte, aangezien ze grotendeels zullen berusten op dezelfde competenties.

Dit proefschrift draagt op drie manieren bij aan het verder ontwikkelen van competenties model van economische ontwikkeling. Ten eerste levert het een methodologische bijdrage, waarin wordt gevraagd hoe variëteit, complexiteit en gerelateerdheid gemeten kunnen worden. Ten tweede levert het een theoretische bijdrage in de vorm van simpele theoretische modellen van economische ontwikkeling waarin competenties centraal staan. Zulke modellen kunnen kunnen helpen bij het redeneren over hoe de verschillende begrippen met elkaar verband houden, en bieden een raamwerk om na te denken over de beleidsimplicaties van een model van economische ontwikkeling waarin competenties centraal staan. Ten derde is er een aantal empirische bijdragen waarin de maten van diversiteit, complexiteit en gerelateerdheid zijn toegepast als ook een nieuw model waarin het empirisch fenomeen van de "hump" (eerst toe- en dan afname van diversiteit) uit de handelseconomische literatuur wordt verklaard aan de hand van een model van competenties en producten.

Hoofdstuk 2 geeft een beknopt overzicht van bestaande onderzoek in disciplines die aan de basis staan van het competenties model: economische geografie en economische complexiteit. In beide literaturen staat het begrip variëteit (van zowel producten als competenties) centraal, zij het op verschillende manieren. Ook worden de methodologieën besproken die gangbaar zijn voor het kwantificeren van diversiteit, gerelateerdheid en complexiteit. Ten slotte wordt nieuwe methodologie voorgesteld voor het meten van diversiteit, die het mogelijk maakt de verschillende hypotheses die volgen uit de literatuur te testen.

Hoofdstuk 3, 4 en 5 bestaan uit methodologische bijdrages. In hoofdstuk 3 bespreek ik het meten van diversiteit. Hierbij ligt de nadruk op hoe diversiteit gekwantificeerd kan worden met inachtneming van de gerelateerdheid tussen de elementen die in beschouwing genomen worden (zoals bijvoorbeeld economische activiteiten). Hiertoe bouw ik voort op het begrip van 'Hill numbers' uit de ecologie, dat een formeel raamwerk biedt waarin diversiteit gemeten kan worden. Ook stel ik een decompositie voor die leidt tot afzonderlijke maten van de factoren van diversiteit - te weten variëteit (het aantal verschillende elementen), balans (de verdeling over die elementen) en dispariteit (hoe (on)gerelateerd zijn de elementen).

Hoofdstuk 4 gaat over het meten van economische complexiteit. De index van economische complexiteit, voorgesteld door Hidalgo and Hausmann (2009), biedt een manier om het aantal competenties in een economie (haar 'economische complexiteit' te kwantificeren op basis van welke producten zij produceert. De economische complexiteit van landen gebaseerd op export data correleert met het bruto binnenlands product en blijkt een sterke voorspeller van economische groei. Echter blijven de precieze interpretatie en de relatie tot de theorie van deze economische complexiteits-index onduidelijk. Mealy et al. (2019) geven meer duidelijkheid over de wiskundige achtergrond van deze index. In dit hoofdstuk bespreek ik een aantal statistische methodes die wiskundig equivalent zijn aan de economische complexiteits index, die elke leiden tot alternatieve interpretaties van de index. Dit laat zien dat de index van economische complexiteit een her-uitvinding is van manieren om netwerken te clusteren, en om de dimensionaliteit van data te reduceren. Deze interpretaties gaan in tegen het idee dan de index van economische complexiteit iets zegt over het aantal competenties dat in een economie aanwezig is, en suggereren dat het eerder een weergave is van in hoeverre economieën op elkaar lijken. Deze nieuwe inzichten, gepaard met het sterke verband met economische groei, bieden nieuwe perspectieven voor empirisch onderzoek in deze richting.

Hoofdstuk 5 is het laatste methodologische hoofdstuk en gaat over het kwantificeren van gerelateerdheid op basis van co-locatie van economische activiteiten. In de praktijk wordt de gerelateerdheid tussen producten vaak gemeten door te kijken naar de co-locatie van producten, onder de aanname dat producten die vaak samen geproduceerd worden (in hetzelfde land, stad of regio) waarschijnlijk berusten op dezelfde competenties. Op deze manier kunnen netwerken worden geconstrueerd die de gerelateerdheid tussen producten weergeven, en zo ook de waarschijnlijke groeipaden van landen of regio's weergeven. In dit hoofdstuk stel ik een nieuw methodologisch raamwerk dat een meer formele manier biedt om de co-locatie tussen economische activiteiten te kwantificeren, op basis van informatie theorie en Bayesiaanse statistiek. Uit dit raamwerk volgen ook maten van lokalisatie en specialisatie, een het legt op deze manier zowel conceptuele als methodologische verbanden tussen ogenschijnlijk onafhankelijke concepten uit verschillende delen van de economische literatuur.

Hoofdstukken 6 en 7 zijn theoretisch van aard en beschrijven een simpel model van economische ontwikkeling waarin competenties gerecombineerd kunnen worden tot producten. Het model beschrijft hoe een economie die steeds meer competenties verzamelt, een toenemende variëteit van steeds complexere producten maakt, die onderling gerelateerd zijn. Het model wordt vervolgens uitgebreid door te stellen dat naarmate een economie zich ontwikkeld, de simpele producten op den duur niet meer geproduceerd kunnen worden, bijvoorbeeld omdat ze niet waardevol genoeg zijn en de lonen te hoog liggen. Dit leidt op een bepaald moment in de ontwikkeling van een land tot een afname van het aantal producten dat een land kan maken. Dit is consistent met een empirische regulariteit die bekend staat als de "hump", waarin de meest ontwikkelde landen juist een afname zien van de variëteit aan producten die ze maken.

Hoofdstuk 7 bouwt voort op het model uit hoofdstuk 6 en breidt het model uit door te stellen dat een economie naast het ontwikkelen van extra competenties ook in staat moet zijn om competenties te combineren tot complexe producten, bijvoorbeeld door middel van instituties, regelgeving en bedrijven. Dit kan gezien worden als een goed functionerend innovatiesysteem. Het verzamelen van specifieke competenties daarentegen kan bewerkstelligd worden door industrieel beleid in bepaalde domeinen. In dit hoofdstuk worden deze twee beleidskeuzes tegen elkaar afgewogen, met de conclusie dat het voor ontwikkelde economieën meer kan lonen op in te zetten op een goed functionerend innovatie systeem dan industrieel beleid te voeren.

In de conclusie reflecteer ik op de methodologische hoofdstukken in dit proefschrift en bespreek ik hoe ze met elkaar in verband staan binnen een informatie-theoretisch raamwerk, en hoe de maten gegeneraliseerd zouden kunnen worden voor gebruik in multivariate analyses. Ook bespreek ik andere technieken die gebruikt zouden kunnen worden bij het kwantificeren van competenties uit economische data. Daarnaast bespreek ik hoe het simpele model dat in hoofdstuk 6 gepresenteerd is uitgebreid zou kunnen worden om tot een meer complexere en realistische beschrijving van de economie te komen, bijvoorbeeld door de structuur op te leggen aan hoe competenties gecombineerd kunnen worden tot producten.

Acknowledgements

This thesis is full of ideas, work and support of others. I would like to thank my supervisor Koen Frenken for providing me the opportunity to do this PhD, and the freedom to explore academia as well as my personal interests in the past five years. Your mentorship has been invaluable.

I thank Andres Gomez-Lievano for his generous help and inspiration in developing research ideas, and for the numerous crash-courses on a variety of topics. It never stops to amaze me how hard it is to come up with something you haven't thought about before.

I also thank Frank Neffke for the warm welcomes at the Growth Lab and for joining this project as co-promotor and supervisor towards the end. Your contributions are very much appreciated.

Next I would like to thank all other co-authors who's work is included in this thesis: Carlo, Sultan, Matias, Mara, Mark, David, Nacho, Miguel. It's been a pleasure to work with you.

I also thank all colleagues at Copernicus, and in particular my offices mates: Marjolein, Matthijs, and Mary. I'll miss the sharing of grand theories, frustrations and business ideas.

Also a big thanks to everyone at the CCSS and all YCREW members, and especially Qingyi, for all effort mades in creating a nice place to work and meet people. It was always a pleasure to be involved in activities at the Centre.

I also thank all members and visitors at the Growth Lab for making every visit there insightful, productive and fun.

I would like to thank all other people I have met during this PhD at a variety of conferences, summer schools and visits - Shanee, Nils, Mathieu, and many others -

with whom I had the pleasure of sharing beers and good conversation, often far away from home.

I also thank my family - Anne, Liselot & Rob and Andrea, Michiel & Kasper - for always supporting me in any way possible. It's not to be underestimated.

Rob & Sjak - thanks for your never-ending enthousiasm (even for this dissertation).

Lars & Tes, thanks for joining me on the next project, which has provided the much needed distraction and a clear deadline for this PhD.

Also a big thanks to all occupants of Granatstraat 64 over the past years for keeping me sane and insane at set times.

And finally thanks again Tes for sticking around, even during periods of physical or mental absence. I hope you will for many years to come.

CV

Alje van Dam was born on the 3rd of November 1991 in Lilongwe, Malawi. He obtained an interdisciplinary Bachelor degree in Natural and Social Sciences (Beta-Gamma) with a major in Mathematics from the University of Amsterdam in 2013. During his bachelor's, he spent a semester at the University of New South Wales, Sydney. He obtained a Master degree in Mathematical Sciences (cum laude) at Utrecht University in 2016, specializing in Computational Science and Complex Systems. Within the master's program, he participated in the transdisciplinary Tesla minor at the University of Amsterdam, and was a visiting researcher at the Laboratory of Economics and Management at Scuola Superiore Sant'Anna, Pisa, and the Growth Lab at the Center for International Development at the Harvard Kennedy School, Cambridge, USA.

In 2016, Alje started as a PhD candidate at the Copernicus Institute of Sustainable Development of Utrecht University, in a project funded by the Netherlands Organization for Scientific Research (NWO) within the Vici program, led by prof. dr. Koen Frenken. His research focused mainly on developing methods to quantify diversity, complexity and relatedness in economic systems. At Utrecht University, he was also a fellow of the Centre for Complex Systems Studies, where he was involved in organizing workshops and providing advice to the board as a member of the Young Complexity Researchers Utrecht.

During his PhD, Alje was a visiting fellow at the Growth Lab at Harvard University and attended multiple international summer schools, including the Complex Systems Summer School at Santa Fe (2017) and the Oxford Summer School on Economic Networks (2018). He has presented his work at multiple academic conferences and seminars, including the Conference on Complex Systems (2017) and the Geography of Innovation Conference (2018).

After his PhD, Alje will spend a year sailing from the Netherlands to Australia on his 38ft sailing yacht *Risa*.

List of publications

van Dam, A. Diversity and its decomposition into variety, balance and disparity. *Royal Society Open Science*, 6(7):190452, 2019. doi:10.1098/rsos.190452

van Dam, A. and Frenken, K. Variety, complexity and economic development. *Research Policy*, 2020a. doi:10.1016/j.respol.2020.103949

van Dam, A., Gomez-Lievano, A., Neffke, F., and Frenken, K. An information-theoretic approach to the analysis of location and co-location patterns. *arxiv.org/abs/2004.10548*, 2020

van Dam, A. and Frenken, K. Vertical vs. Horizontal Policy in a Capabilities Model of Economic Development. *arxiv.org/abs/2006.04624*, 2020b