*Romanian Journal of Regional Science*
*Vol. 18, No. 1, Summer 2024*

DOI: https://doi.org/10.61225/rjrs.2024.01

# FROM SECTORAL INDUSTRIAL COMPOSITION TO EMPLOYMENT AND REVERSE. THE ITALIAN CASE

**Massimo Giannini[a,*], Barbara Martini[b], Cristiana Fiorelli[c]**

[a] University of Rome Tor Vergata, Dep. of Enterprise Engineering, Italy

[b] University of Rome Tor Vergata, Dep. of Enterprise Engineering, Italy

[c] Sapienza University of Rome, Dep. of Economics and Law, Italy

[*] Corresponding author

Address: University of Rome Tor Vergata, Dep. of Enterprise Engineering, Via del Politenico 1, 00173 Rome, Italy

E-mail: massimo.giannini@uniroma2.it

**Biographical Notes**

**Massimo GIANNINI**, PhD, is Full Professor of Economics at the University of Roma Tor Vergata. His research topics include: Mathematics and Statistics applied to Economics, Spatial Econometrics, Regional Economics and Endogenous Growth Theories. ORCID ID: https://orcid.org/0000-0003-1461-0005

**Barbara MARTINI**, PhD, is Aggregate Professor of Political Economy at the University of Rome Tor Vergata. Her research topics include: Regional Economics, Spatial Econometrics, Gender studies. ORCID ID: https://orcid.org/0000-0002-2686-4259

**Cristiana FIORELLI**, PhD, is Aggregate Professor of Applied Economics at the Sapienza University of Rome. Her research topics include: Regional Economics, Spatial Econometrics, Banking systems, Macroeconomics. ORCID ID: https://orcid.org/0000-0001-9829-9466

**Abstract**
The diffusion of knowledge among firms and workers, which, in turn, depends on the nature of the knowledge itself and the relatedness or unrelatedness of the industrial composition, is an essential driver for the growth and employment of territories. The proximity between related industries allows for faster diffusion of specific knowledge, as happens in the Marshall-Arrow-Romer (MAR) theory of externalities (agglomeration) and Jacobs's theory of knowledge transmission. The literature (notably the Evolutionary Economic Geography one, EEG) has empirically investigated the relationship between employment (or GDP) and some measures of industrial composition (e.g., the related and unrelated variety). In this paper, we contribute to the literature in two ways. First, we propose a more coherent use of entropy indexes to measure the relatedness or unrelatedness of industries in an economy. Second, we empirically investigate the nexus between industrial composition and employment level by exploiting these measures. Unlike the existing literature, we argue that such a nexus could be circular. For such a reason, an empirical investigation using a spatial vector autoregressive (SpVAR) model will be performed for the Italian economy at the provincial level. The results show that such a circular nexus exists, shedding new light on the debate.

**Keywords:** Entropy, Variety, Sectoral composition, Knowledge diffusion, Spatial VAR.
**JEL Classification**: B52, J21, L16, O33

## 1. Introduction

The idea that knowledge spills through the economy in a pervasive way has always been present in economic thought, especially when firms and individuals live geographically closed. Marshall (1890) described cities as having "ideas in the air". The idea was later developed by Arrow (1962) and, more recently, by Romer (1986), where the growth of an economy is intimately related to the increasing returns stemming from knowledge spillovers. The acronym Marshall-Arrow-Romer (MAR) synthesizes this channel of knowledge transmission based on the proximity of firms belonging to common industries; the closer industries are, geographically and for the type of goods, the faster firms share and diffuse ideas, organizations, products, and technologies.

However, in the endogenous growth theory (notably the neo-Schumpeterian approach), knowledge (and human capital, to cite Lucas, 1988) spreads in the economy independently of proximity. It is an intimate externality and a public good available to everyone. In other words, it is not related to industrial tissue, even when the latter is more diversified. Nevertheless, following Jacobs (1969), it is reasonable to assume that ideas flow among different but related industries, complementing and sustaining each other, thanks to a cross-fertilization effect that triggers increasing returns. Furthermore, diversification avoids the "lock-in" effect, which strikes a specialization in MAR theory. Thus, diversification in local industries is seen as a measure of "richness" and dynamism of the local economy; however, there must be a certain degree of "relatedness" between industries, as ideas flow easily when technologies, skills, markets, and organizations are close.

The literature, particularly in evolutionary economic geography (EEG), debates the benefits and drawbacks of relatedness versus unrelatedness. Although unrelatedness provides barriers to fast knowledge diffusion, it protects industrial tissue from asymmetric shocks. As in portfolio theory, a highly diversified “basket” is less vulnerable to a specific shock that strikes a limited number of “assets”. The unrelatedness of assets prevents the diffusion of the shock throughout the portfolio. In contrast, relatedness allows knowledge to diffuse faster among similar technologies. For example, labor mobility between firms is crucial in spreading knowledge and skills (Boschma, Eriksson and Lindgren, 2014; Duranton and Puga, 2004; Gertler, 2003). However, most job flows are local and between related industries; social and cultural barriers, labor regulation, and market structure affect this knowledge exchange. Strong ICT development could relax such barriers, as it favors distant technological interactions and knowledge flow among workers and peripheral regions (Baldwin, Martin and Ottaviano, 2001).

If relatedness or unrelatedness matters for knowledge diffusion and hence for the development process, proxied by the employment or GDP, that is a question of empirical investigation. The literature does not provide a final answer. It is out of the scope of this paper to review this large literature, but only to cite more recent contributions. For instance, Pinheiro et al. (2018) find that countries developing activities in unrelated sectors show a small but significant increase in economic growth. Xiao, Boschma and Andersson (2018) showed that European regions with a higher innovation capacity are more inclined to enter less related industries. Similar results in Boschma and Capone (2016), where Western European economies tend to diversify more into unrelated industries than Eastern European economies. The role of external firms also plays a relevant role. Neffke et al. (2018) show that firms outside the region are more likely to introduce unrelated diversification and shift regions' specializations. But institutions also matter. Boschma and Capone (2015) show that countries with more liberal and less coordinated forms of capitalism are more likely to diversify into more unrelated activities. The endowment is another ingredient. Montresor and Quatraro (2017) found that regions with a strong presence of key enabling technologies tended to diversify into more unrelated technologies. Finally, the initial condition matters (path dependency); Petralia, Balland and Morrison (2017) showed that high-income countries have a higher tendency to diversify into more unrelated and sophisticated technologies.

Contrastingly, several contributions support the importance of relatedness. More recently, Boschma et al. (2023) used new measures of (un)relatedness built on patents as a source of breakthrough innovation and generally support the relatedness assumption: “Our study shows that relatedness matters for the occurrence of breakthrough inventions in regions, not unrelatedness: breakthroughs are more likely to occur in a region, the more related their technology is to the local stock of technological knowledge" (page 132). However, they also concluded that both relatedness and unrelatedness could interact: “Hence, both relatedness and unrelatedness could enhance the development of breakthrough inventions in regions. In other words, the two factors do not need to be mutually exclusive" (page 122).

Moreover, Van Dam and Frenken (2022) investigate whether the presence of implemented technologies fosters further technological innovation through the recombination of existing technologies. They argue that new products will be similar or related to existing ones because they largely consist of already present capabilities. Their findings indicate that local GDP per worker is a significant driver of product variety, but this relationship follows a non-linear, U-shaped curve. This suggests that countries initially diversify their product offerings as they develop but later specialize again. Their study raises the question of how the growth process impacts product variety. If GDP per capita signals economic complexity, which in turn affects GDP growth, a circular relationship emerges. Dynamically, this reverse causation can help explain different countries' different performances. Developing countries can easily acquire new varieties by mimicking technologies from developed countries, while the latter must innovate at the technological frontier. However, due to diminishing returns, the pace of new recombinations and innovations tends to slow down in more mature economies. This slowdown in growth in high-income countries may be attributed to the reduced rate of acquiring and recombining new capabilities. Overall, these recent contributions show that relatedness and unrelatedness cannot be seen as a dichotomic process (one against the other) but as a more interactive phenomenon.

Given this extensive debate, this study aims to investigate the potential bidirectional relationship between local growth and sectoral composition, drawing on the contributions of Van Dam and Frenken (2022). To explore how the relatedness or unrelatedness of industrial sectors impacts knowledge diffusion and economic development, we employ a novel decomposition of total entropy.

First of all, we have to properly measure the relatedness and unrelatedness of sectoral composition. Earlier empirical literature tended to measure the two effects by indexes of Related Variety (RV) and Unrelated Variety (UV), as in Frenken, Van Oort and Verburg (2007). The authors proposed a particular decomposition of the Shannon entropy index, which they call "variety", to capture the effects of two types of industrial diversity. The decomposition brings to RV, capturing "local" spillovers among related industries, and to UV, measuring "global" spillovers among distant and unrelated industries (not only geographically, but for technological, managerial, capabilities, market powers, and human capital). The latter resembles the idea that technology spreads freely within the economy as a sort of externality and public goods, while the former lies in Jacobs' contribution. We do not use this decomposition for two reasons. First, the interpretation of RV and UV as dichotomic and independent measures of relatedness and unrelatedness needs to be more mathematically correct. Second, they are applied to a single local entity, detached from the rest of the economy. We propose decomposing the total entropy, or variety, into components "within" (alpha-entropy) and "between" (beta-entropy). This type of decomposition is widely used in natural sciences to investigate the diversity within a habitat and between different habitats. In our context, this decomposition is essential to measure how the local economy diverges from the national one; this allows us to capture regional peculiarities concerning the national average. Moreover, by adopting a spatial econometric approach, we consider how local economies affect each other; knowledge spills over time and space. The standard measures of RV and UV do not own this property, as each local entity is analyzed per se. But local economies are not islands and cannot be analyzed "in vitro." The sectoral composition of a geographic entity could affect, or be affected, by what happens in their neighborhood and the whole economy (meta-community). Moreover, this type of decomposition can take into account the possible interplay between local (related) and macro (unrelated) components, as suggested by Boschma et al. (2023).

Once we have measured within and between varieties, we turn back to investigate the circular nexus between local employment and sectoral composition. As said, if sectoral composition affects employment via knowledge spillovers, we also wonder whether the reverse is true. The current literature considers only a unidirectional link, but, as van Dam and Frenken (2022) suggested, the two variables could be recursively linked. The growth process, induced by an increase in employment, can modify industries' relatedness through several channels: segmented labor demand, sector-specific technical progress induced by asymmetric ITC or R&D investments, market segmentation due to globalization, and international competitiveness are some examples. In other words, if the industrial sectoral composition could affect the local employment process, it seems clear that the reverse should work as well. Regional development and regional disparities are the visible effects of such an interplay. For this reason, in this paper, we examine this bidirectional relationship estimating a vector autoregressive (VAR) model for Italian provinces with two endogenous variables, one measuring the development process via employment level, and the second one the within component of local industrial composition, a novelty in the literature. We extend a standard VAR model to account for spatial components, including spatial correlation among the dependent variables, bringing to exploit an innovative econometric technique the spatial VAR (SpVAR - Beenstock and Felsenstein, 2007; Baltagi, Fingleton and Pirotte, 2019; Giannini, Fiorelli and Martini, 2022). Spatial spillovers should account for knowledge transmission among close productive units. Moreover, beta-entropy is used as a covariate for controlling how far the local economy is relative to the national average (between entropy). It is worth noting that a SpVAR does not allow many covariates; they must be parsimonious to be interpretable.

The paper is organized as follows. Section 2 introduces the different entropy measures used in our analysis. Section 3 reports a description of the sectoral composition in Italy. Section 4 describes the data, while Section 5 provides an econometric investigation of the possible circular nexus between employment and sectoral composition. In Section 7, we discuss the main results of Section 6. The conclusions follow.

## 2. A measure of industrial composition: gamma, alpha and beta entropy

As shown in Giannini, Fiorelli and Martini (2022), a more coherent approach should start from the total entropy ($\gamma$ entropy) calculated on the entire economy (meta-community), which is then decomposed in a within component (alpha) entropy and the divergence between the local and the meta-community (beta-entropy). This approach differs from the usual decomposition of total entropy in RV and UV, often used in EEG. In particular, the latter measures the divergence between the distribution at the lowest slice (usually five-digit) of a Standard International Classification tree (SIC) and the one at the upper level (two-digit). It is a measure of relative entropy, hence referred to as a given level (the two-digit in our case); a low value cannot be interpreted *"per se"* as a sign of concentration or vice versa, as usually interpreted in the current literature (see Appendix I for details). Furthermore, a measure of the interaction between the local and national (average) structure is missing in the EEG approach. Beta-entropy delivers important information as it underlines sectoral specificity in local communities compared to national productive tissue. Technically it measures the statistical similarity between two distributions. It allows us to assess how regional peculiarities are relevant and how a single community behaves differently from the national economy. A high value of beta-entropy shows that the single community is more dispersed with respect to the national one. In the last twenty years, spatial econometrics has stressed the importance of spatial externalities among close geographical entities. Focusing on a single geographical entity could misestimate the proper channels of knowledge diffusion.

Hence, we calculate entropy at the meta-community or national level ($\gamma$ entropy). A meta-community (Italy) is divided into several local communities (provinces - indexed by $i = 1,2, \dots, I$). $n_i$ individuals (workers) are sampled in community $i$. Let $s = 1,2, \dots, S$ denote the species (five-digit industries) that compose the meta-community, $n_{s,i}$ the number of individuals (workers) of species $s$ (five-digit industry) sampled in the local community (province) $i, n_s = \sum_i n_{s,i}$ the total number of individuals of species $s$ (i.e. the total number in Italy of workers in a given five-digit industry), $n = \sum_s \sum_i n_{s,i}$ the total number of sampled individuals (total workers in Italy). Within each local community $i$, the probability $p_{s,i}$ for an individual to belong to species $s$ is estimated by $\hat{p}_{s,i} = n_{s,i}/n_i$. The same probability for the meta-community is $p_s = n_s/n$. Communities have a weight $w_i$, satisfying $p_s = \sum_i w_i p_{s,i}$. The commonly used $w_i = n_i/n$ is a possible weight; in this case $\hat{p}_{s,i} = n_{s,i}/n$.

This entropy can be decomposed in within (inside a local community - $\alpha$ entropy) and between (among local communities - $\beta$ entropy) components. Marcon et al. (2014) derived the decomposition of the $\gamma$ entropy[1]:

$$\begin{aligned} H_\gamma &= \sum_s p_s \, ln(1/p_s) \\ H_\alpha^i &= \sum_s p_{s,i} \, ln(1/p_{s,i}) \\ H_\beta^i &= \sum_s p_{s,i} \, ln(p_{s,i}/p_s) \end{aligned} \tag{1}$$

From the formula for $H_\beta^i$ is clear that beta-entropy is a measure of divergence between the local community and the meta-community in each sectors $s$. If the local community shares mimic exactly the national ones, $p_{s,i} = p_s \; \forall s \in S$, then $H_\beta^i = 0$. As far as alpha is concerned, it is zero when all employment is concentrated in a single sector and assumes *ln(S)* when the distribution of workers in the *S* sectors is uniform. In general, a low value of alpha means that few sectors employ large shares of the population. The maximum theoretical alpha-entropy is 6.71, as *S* equals 820 in our data. As we shall see in the next section, in 2018, the average alpha in Italy was 5.25. From this point of view, the average is not so far from the maximum, bringing to the conclusion that employment is rather widespread inside the sectors. But this would be a wrong conclusion, as we have to bear in mind that the entropy formula is highly non-linear. So, 5.25 is high or low? To manage the question more correctly, it is useful use the uniform equivalent transformation. The latter responds to the following question: Given an entropy level (5.25 in our case), how many sectors are needed to have a uniform distribution? The answer is easy; it involves the inverse natural logarithm. In our case, we need only 190 sectors out of the 820 original ones. Figure 1A in Appendix III shows the relationship between alpha-entropy and the equivalent transformation; as is clear, the relationship is strictly convex. The two bars represent the maximum and minimum alpha-entropy values in 2018 in our data. This leads to the conclusion that Italian provinces generally have a low level of equivalent sectors, that is, the alpha values are in the middle-low range.

The alpha-entropy corresponds to the variety of the EEG. By trying to unify the two approaches, the gamma entropy of the meta-community can be calculated in the following way:

[1] In general, the entropy measures use a generalized logarithmic form, i.e. the q-log. When q=1 the measure brings to the Shannon index. This what we are going to assume.

$$H_\gamma = E(H_\alpha) + E(H_\beta) = E(RV) + E(UV) + E(H_\beta) \quad (2)$$

where expectation E runs through the *i*-local communities. Consequently, RV and UV do not provide a final picture of the total entropy. The beta-entropy delivers important information: how the local level diverges from the average (national) one. In the gamma-entropy decomposition, the two components can be treated as two independent measures: the alpha measures the within dispersion and the beta between one. By doing so, we obtain a more coherent and reliable measure of industrial composition and its relatedness among subunits.

## 3. Descriptive analysis of the Italian sectoral composition

As is well known, the Italian production system shows significant differences between regions. However, it is possible to identify the economic sectors that are the most driving forces for each area. In this regard, we consider the level of employment. *Ceteris paribus*, we assume that the dynamics of local development is positively correlated with the development of employment. Therefore, the sectors that employ the most people also contribute the most to the formation of the local GDP. The Italian National Institute of Statistics (ISTAT) provides the number of workers at the NUTS3 level for each economic sector (five-digit), according to the Italian Industry Classification (ATECO). Local employment (*Empl*) is defined as the total number of workers by province in the private sector, excluding agriculture. In 2018, we counted a total of 17,287,891 employees in 820 sectors (five-digit) in 107 Italian provinces (Figure 4A in Appendix III shows the geographical distribution of the provinces).

### 3.1 Distribution of employees by sector

What is relevant to our aim is the distribution of employees among sectors. A concise tool for this purpose is the heatmap. It shows how many people are employed in each sector and province. The range goes from gray (less people employed per sector) to red (more people employed per sector). The map is shown in Figure 2A in Appendix III, together with a descriptive table (Table 1A). At first glance, employment is not equally distributed across sectors and provinces, as most sectors employ few workers (in absolute value). In Figure 2A, we report the most relevant sectors and provinces according to the degree of employment. The majority of workers are employed in sectors related to services, specifically in food and beverage activities, temporary employment agency activities, real estate, transport, personal care services, and professional, scientific, and technical activities. The construction sector ranks sixth, while the roles of sectors related to industrial activities are less prominent, except for the “Custom software development” and “Machining” sectors.

Focusing on the top 10 provinces, many workers are located in large cities, especially in the North, except Rome and Naples. Milan has the highest number of workers, followed by Rome, Turin, Naples, Brescia, and Bologna. The two major provinces, Rome and Milan, have a high number of employees in almost all 30 sectors (more than 20,000 employees). The food and beverage sector remains the main industry in all provinces.

To account for the employment share, we repeat the same exercise considering the sectoral distribution of employees over the total number of employees within the province. As Figure 3A in Appendix III shows, food and beverage services remain the sector with the highest concentration of employment in all provinces. The heatmap is almost entirely gray (employees per sector account for less than 5% of total employment in each province). This implies that workers are not evenly distributed across the 820 sectors but are concentrated in very few sectors of activity. The exceptions are Prato, with a percentage of workers around 16% in the outerwear manufacturing sector, and Bolzano and Rimini in the hotels and similar accommodation, which employ a percentage of workers between 9 and 11.

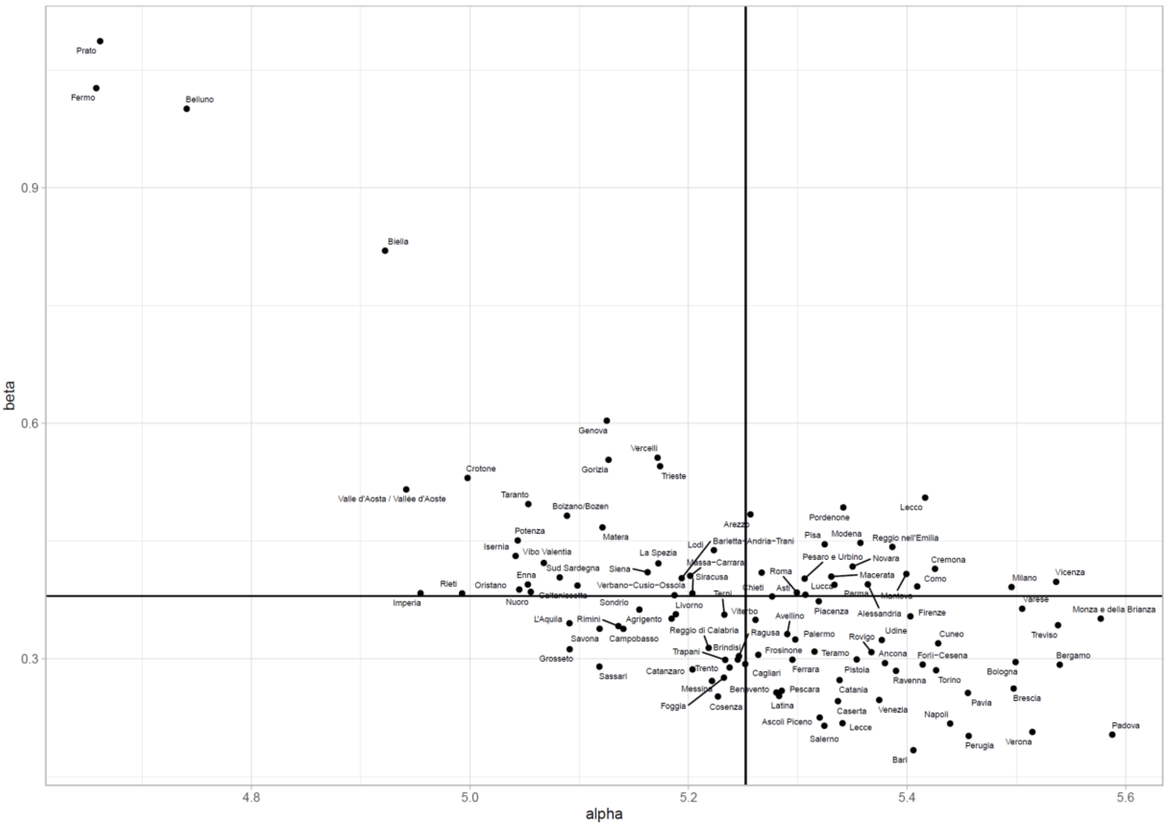
Therefore, the 2018 data show a relatively high concentration of workers in a few sectors characterized by low human capital and technology. Figure 3A in Appendix III is rather clear, the sector coded 56101, “Restaurants and mobile food service activities” is the one employing the largest share of workers in all of the 107 provinces in a broadly equal share (from 5% to 7%). The second largest sector is the “Temporary employment agency activities” but with greater dispersion between geographical units. Finally, the third is 56300 “Beverage serving activities” which, once again, is largely represented in all provinces, as expected. As previously noted, the “hottest” spot (the red bar in Figure 3A in Appendix III) is represented by the Prato province, which is mainly specialized in the 14131 sector (Manufacture of other outerwear) as the largest textile district in Europe. Finally, we remark that Milan and Rome show a higher share of employment in business support and activity and software development relative to the rest of the provinces. In other words, sectors related to ICT and higher education are most concentrated in the two Italian business centres, as expected. Furthermore, the provinces belonging to the Emilia-Romagna region, such as Bologna, Parma, Modena and Reggio Emilia, present a relatively higher share of workers in sectors related to automotive, agriculture and mechanics. Modena in particular hosts the most important 'supercar' and 'superbike' companies (Ferrari, Lamborghini, Ducati, just to cite some) and firms working in accessories and mechanical components.

The investigation of employment sectoral distribution shows that Italy is still stuck in traditional low-technology sectors, despite data referring to 2018, where we would have expected the emergence of high-skilled, high-tech sectors. Furthermore, except for restaurants and beverages, the provinces are rather dispersed among five-digit sectors, with each province more specialized in a few sectors than another.

### 3.2 Calculation of Alpha and Beta entropies

The variety in sectoral composition can be well represented by decomposition into alpha and beta entropies. The decomposition provides two numbers for each province. By taking the average values of both measures, we obtain Figure 1; it shows the distribution of alpha and beta entropies with respect to the average values. We recall that higher values for alpha point to a larger variety of the single territory, while a higher beta value means that the territory behaves differently from the national distribution (meta-community).

**Figure 1**: Alpha and Beta entropies.




**Source:** Own elaboration.

The average alpha-entropy is 5.25, and the beta one is 0.38; so the gamma entropy of Italy is 5.63. Consequently, most of the total entropy is mainly due to the within component. However, some distinct characteristics can be identified, consistent with the heatmap in Appendix III. The two largest metropolitan areas, Rome and Milan, have higher than average values for both alpha and beta; they are the most diversified and more dissimilar than the average. To this group also belong other performing provinces, such as Parma, Modena, and Reggio Emilia, as previously remarked. The group with higher alpha but lower beta is more geographically dispersed values. It is worth noting that neither southern province is in this group, being dominated mainly by the northern ones (except Rome) and basically the same results for the converse situation (lower alpha and higher beta). Finally, provinces characterized by lower alpha and beta values, compared to averages, are mainly clustered in the southern regions.

As we noted in the heatmap, there are some clear outliers, namely Prato, Fermo, Belluno, and Biella. The former two are provinces with the highest specialization (lowest alpha value) in manufacturing activities (shoes and textiles). Belluno is an important district for glasses (Luxottica and Safilo are among the main companies), while Biella is another important textile district, particularly for wool manufacturing.

A brief focus on Italian districts helps to clarify the data, as they represent areas of high productive specialization. ISTAT identified 141 industrial districts in 2011 (they were 181 in 2001) based on Local market areas (LMAs, 'labor market areas – SLL' in Italy) and the analysis of their economic specialization by using the data on economic units surveyed in the 9$^{th}$ Industry and Services Census (the last available being collected on a decennial base). They represent 64.1% of mainly manufacturing LMAs which employ 65.8% of workers in the manufacturing sector. Figure 5A in Appendix III shows the geographical distribution. A comparison with 2001 underlines that, not only the number of districts reduced by 40, but also their employment level (from 70.9% in 2001 to 65.8% in 2011). Moreover, they are mainly located in northern provinces and, in particular, in the North-East part. Districts belonging to southern regions represent only 14.9% of the total. However, when we focus on the type of specialization, districts mainly specialize in agriculture, food, textile, and house goods (see Figure 5A in Appendix III). Moreover, firms are in general of small size with a low propensity to innovate. A more recent analysis has been performed by the Italian bank Intesa San Paolo with the annual report on industrial districts.[2] Although the report has a different sample of districts (they identify 159 districts), the picture is not far from the 2011 one. There is still a large prevalence of agriculture, textile, fashion, and agricultural machine. According to Figure 32 in the report, the best performer is the district of Padova and Vicenza, followed by Reggio Emilia and Modena for the agricultural machine, as we stressed previously. However, the report stresses the low dimensionality of firms operating in such districts (small and micro firms account for about 82% of the sample) and their low propensity to innovate: the percentage of firms adopting investment in Industry 4.0 is low and concentrated in large firms, which in their turn, are a strict minority (4% of total firms in the sample). Interesting enough, the report underlines also how foreign firms are increasing their participation in the districts, in particular in micro firms.

---

[2] "Economy and finance of industrial districts", annual report – n. 14, March 2022.

Concluding, this brief explorative data analysis draws an industrial composition mainly located in traditional sectors and characterized by low dimensionality of firms with a low propensity to innovate. The proposed measures of sectorial composition (alpha and beta entropies) can correctly capture such features. They summarise the information we can filter by the heatmap and by the district analysis; the latter provides a finer level of investigation, but they are heavier, being a matrix with 820 sectors and 107 provinces each year. The proposed entropy measures are an easier way to capture differences in the sectorial composition and can be useful to exploit in an empirical investigation.

## 4. Data

The dataset consists of six annual data, from 2013 to 2018, at NUTS3 level in Italy, 107 provinces, for a total of 642 observations.[3] Some descriptive statistics for the entire panel dataset are reported in Table 1. The vector of endogenous variables contains the employment level (*Empl*) in each province and a measure of local industrial diversification. The latter is measured by the entropy index (*Alpha*) to determine the degree of dispersion into the local economic tissue. We calculate the alpha starting from the units of workers within the last level of the industrial sector classification (five-digit) (see Section 2 for details).

**Table 1:** Descriptive statistics.

| | Mean | St. Dev. | Median | Min | Max |
|---|---|---|---|---|---|
| *Empl* | 155,447.18 | 202,943.92 | 99,490.40 | 16,820.89 | 1,545,669 |
| *Alpha* | 5.26 | 0.18 | 5.28 | 4.51 | 5.62 |
| *Beta* | 0.38 | 0.15 | 0.35 | 0.18 | 1.14 |
| *wage* | 25,206.51 | 2,950.88 | 25,226.91 | 13,842.78 | 37,568.07 |
| *dens* | 5.19 | 0.80 | 5.17 | 3.66 | 7.88 |

**Source:** Own elaboration.

We control for several components. First, we expand the classical EEG approach for measuring industrial variety and consider alpha-entropy and beta-entropy (*Beta*). Beta emphasizes the sectoral characteristics of local communities with respect to national productive tissue. It is calculated as a divergence between the local community and the meta-community. As shown in Table 1, the level of beta-entropy varies in a range between 0.18 and 1.14, showing high heterogeneity in the distribution when local communities are compared with meta-community.

We also account for the wage component (*wage*). ISTAT provides data on the compensation of employees on a NUTS2 basis. Hence, the regional wages are allocated by the province using the

[3] We cut our sample to remove the outliers caused by Covid-19 and the Russia-Ukraine war from the analysis.

share of NUTS3 GDP on the regional GDP. We consider per capita and real wages divided by the GDP deflator. Finally, population density (*dens*) has been included to capture the impact of agglomeration economies. The ratio between the total population and the regional land-use area measures it. Moreover, we control for the agglomeration effect in separate clusters, as local economies can be characterized by membership. We build a series of dummy variables where 1 means the province belongs to the *i-th* cluster. Clusters were obtained by simple network analysis (igraph and icluster packages in R). More in detail, an adjacency matrix is built on each cross-section by using a cosine similarity measure. Later, a fast greedy algorithm for community detection is applied to the matrix. The number of identified clusters ranges from nine in 2013 to eleven in 2017, with a low level of churning, confirming the polarization of provinces into "clubs". It is worth noting that the eleven clusters in 2017 broadly encompass the other years, preserving the hierarchy among clusters; it represents only a finer classification and not a reshuffle. In other words, it is possible to use the 2017 eleven clusters as a proxy for the entire time horizon. This allows us to build a matrix of 107 provinces for 11 dummy variables each year.

## 5. Empirical method: spatial VAR model

Once the sectoral composition has been properly measured by means of the alpha and beta entropies, we ask whether the latter affect the change in the employment level. However, at the same time, we are interested in assessing whether reverse causation works as well. To cope with possible spatial dependence, we employ a spatial extension of the standard VAR (SpVAR) model by assuming that spatial spillovers could exist among variables. In this fashion, we analyze not only the time dependency of the model but also the spatial one and capture possible contagious effects among neighboring provinces. The assumption leads us to consider a system of two dynamic equations as follows:

$$Y_t = \rho W Y_t + B Y_{t-1} + \Gamma X_t + \varepsilon_t \tag{3}$$

where $Y$ is a 2x1 (employment[4] and alpha-entropy) vector of the endogenous variables (each endogenous is *Nx1* where *N* is the number of units), $X$ is a matrix of exogenous variables, and $\varepsilon$ is a vector of error terms. $\rho$ is the vector of spatial autoregressive coefficients and $W$ the spatial matrix, with 1 for neighbouring provinces and zero otherwise. Moreover, to capture possible latent endogeneity among regressors, we assume that the residuals follow a Spatial Autoregressive model as well, with a random individual effect (Baltagi, Fingleton and Pirotte, 2014 - BFP):

[4] To check the robustness of our findings, we estimate the model using a measure of local productivity provided by ISTAT. The results, available upon request, are qualitatively the same.

$$\varepsilon_{it} = \lambda \sum_{j=1}^{N} w_{ij}\varepsilon_{it} + u_{it}$$
$$u_{it} = \mu_i + v_{it} \tag{4}$$

where $\lambda$ is the spatial autoregressive coefficient, $\mu_i$ is an individual-specific time-invariant component which is assumed to be $i.i.d.(0,\sigma_\mu^2)$. $v_{it}$ is the error term $i.i.d.(0,\sigma_v^2)$, with $E(\mu_i, v_{it}) = 0 \quad \forall t$.

Since the number of parameters exceeds the available info, some assumptions must be made for SpVAR identification. In particular, we assume that spatial spillovers involve only one endogenous variable per equation. Such assumption brings us to rewrite the SpVAR as two dynamic spatial panel models:

$$\begin{aligned} Empl_{it} &= \rho_1 WEmpl_{it} + \beta_1 Empl_{it-1} + \beta_2 Alpha_{it-1} + \gamma X_{it} + \varepsilon_{1it} \\ Alpha_{it} &= \rho_2 WAlpha_{it} + \beta_3 Empl_{it-1} + \beta_4 Alpha_{it-1} + \gamma X_{it} + \varepsilon_{2it} \end{aligned} \tag{5}$$

where $Empl$ is the employment level and $Alpha$ is the local sectoral diversity for each province $i$ ( $i = 1...N$ ) and time $t$ ; $\gamma$ is the matrix of coefficients and $X$ is the matrix of covariates. $\rho_1$ and $\rho_2$ are the spatial autoregressive coefficients.

Estimation of the SpVAR opens several questions about the inconsistency of the estimator. Generally, the literature suggests estimating panel by panel (or block by block) using estimators related to the dynamic panel model, such as Arellano and Bond (1991). That is, by using a GMM approach with differences and levels of the lagged variables as instruments (for instance Jacobs, Ligthart and Vrijburg, 2009; Bouayad-Agha and Védrine, 2010).[5]

In this paper, we follow the approach of Baltagi, Fingleton and Pirotte (2014), with a slight difference in the assumptions about exogenous variables (see Appendix II). These authors propose a spatial GMM estimator in the spirit of Arellano and Bond (1991) under the assumptions that the model includes temporal and spatial lags on the endogenous variable together with SAR-RE disturbances.[6]

[5] Different approaches may be considered. Yu, De Jong and Lee (2008) propose a Quasi Maximum Likelihood Estimator (QMLE) for spatial dynamic panel data with fixed effects when both T and N are large. Beenstock and Feldstein (2007) suggest, being GMM rather consuming in terms of moment conditions, using an LSDV estimator by correcting the bias due to the fixed effect using Hsiao correction.
[6] Technical details are in Appendix II.

We rearrange the dynamical system described by the equations in the following way, where a first difference operator is applied in order to erase the fixed components, as usual:

$$\begin{aligned}
&\Delta Empl_t = (I - \rho_1 W)^{-1}[\beta_1 \Delta Empl_{t-1} + \beta_2 \Delta Alpha_{t-1} + \gamma \Delta X_t + \Delta\varepsilon_{1t}];\\
&\Delta\varepsilon_{1t} = (I - \lambda_1 W)^{-1} v_{1t}\\
&\Delta Alpha_t = (I - \rho_2 W)^{-1}[\beta_3 \Delta Empl_{t-1} + \beta_4 \Delta Alpha_{t-1} + \gamma \Delta X_t + \Delta\varepsilon_{2t}];\\
&\Delta\varepsilon_{2t} = (I - \lambda_2 W)^{-1} v_{2t}
\end{aligned} \tag{6}$$

In matrix form:

$$\Delta Y_t = C \Delta Y_{t-1} + \Gamma \Delta X_t + \Delta \Psi_t \tag{7}$$

The $\Delta Y_t$ vector is composed by 2 vectors ( $\Delta Empl, \Delta Alpha$) of 107x1 provinces, $X_t$ is the 2x1 vector of predetermined and exogenous variables, $C$ is a 2x2 block-matrix and $\Gamma$ is a 3x2 block matrix. Finally, $\Delta\Psi_t$ is the vector of error terms corrected by spatial component.

$$C = \begin{bmatrix} \dfrac{\beta_1}{(I - \rho_1 W)} & \dfrac{\beta_2}{(I - \rho_1 W)} \\ \dfrac{\beta_3}{(I - \rho_2 W)} & \dfrac{\beta_4}{(I - \rho_2 W)} \end{bmatrix} \tag{8}$$

Unlike the single equation models, VAR interpretation does not lie in the coefficient estimate, as the latter has little interest. A VAR describes the law of a dynamical system, its temporal evolution, its convergence properties, and how it behaves when perturbated. In the 70s, econometricians solved the question by exploiting what is called an “Impulse Response Function” and its orthogonalized version of the latter (OIRF), usually obtained by a Cholesky decomposition. As is standard in VAR models, IRFs represent the change in the dependent variables induced by a one-standard deviation shock to each variable, under the assumption that no other shocks occur at the same time. In other words, IRFs trace the temporal effects of an innovation shock on one variable on the response of all variables in the system (partial derivative). This is clearer by rewriting the model in MA form (in the levels) by a simple recursive substitution:

$$Y_t = C^t Y_0 + \left( \sum_{i=0}^{t-1} C^i X_{t-i} \right) \Gamma + \sum_{i=0}^{t-1} C^i \Psi_{t-i} \tag{9}$$

where the power expansion $\sum_{i=0}^{t-i} C^i$ accounts for impulse response functions to a unitary shock coming from the residual's component at time *t-i*, and $\sum_{i=1}^{t-i} C^i \Gamma$ are the dynamic multipliers. If the system is covariance stationary, i.e., the roots of $C$ are inside the unitary circle, the term $C^t Y_0$ can be neglected. To obtain the OIRF, a Cholesky factorization of the *C* matrix is needed.

In our case, a shock can be propagated in the system not only over the time (the *i-periods*) but in the space also, as provinces are spatially correlated. So, the OIRF must be corrected to take into consideration the spatial dynamics as well. As seen before, the *C* matrix is formed by sub-matrices containing, other than the coefficients estimate, also the spatial weights *W*. Unlike non spatial VAR, the OIRF are represented by a power expansion of a block matrix, which summarizes the shock propagation both in time and in space. The expansion is rather heavy and difficult to interpret in general. For making the OIRF more interpretable, we follow the suggestion by LeSage and Pace (2009[7]), interpreting OIRFs' matrix trace as the average direct impact (ADI), and the sum over rows and columns as the average total impact (ATI). The difference between ATI and ADI provides the average indirect impacts (AII). This decomposition allows us to identify how local spillovers work and quantify the effect of a shock on the innovation component $\varepsilon_t$ on the endogenous variable.

By calculating ADI and ATI at each lag *i*, with *i*=12, it is possible to plot the temporal evolution of the OIRFs.

## 6. Empirical results

In this paper we ask, from a macroeconomic point of view, whether sectoral composition matters for the development of a local economy via technological diffusion, and vice versa. To answer this question, we develop an empirical framework that allows us to model this interplay in a dynamic setting. Following sections discuss the results of the spatial econometric analysis.

### 6.1 Estimates of Dynamic Spatial Panel

Table 2 shows the two-step GMM estimation results. There are some highlights. First, the speed of adjustment (the coefficients of lagged endogenous) is not so fast. This means that the transitional paths of variety slowly adjust to long-run equilibrium. Second, the spatial variables coefficients ($\hat{\rho}_1$ vector) are also not high. However, spatial components cannot be omitted from the analysis, as confirmed by the cross-sectional dependence test (Pesaran, 2007) in Table 3. Both endogenous show a significant magnitude of spatial autoregressive coefficients ($\hat{\lambda}$ vector) of the residuals, performed in the second step of the BFP approach.

[7] The methodology refers to a Spatial Autoregressive model composed by a single equation but can easily adapted to a multi-equation model.

**Table 2:** Estimates of Dynamic Spatial Panel (BFP approach).

| | (I) $Empl_t$ | (II) $Alpha_t$ |
|---|---|---|
| $\hat{\gamma}$ | 0.293 | 0.366 |
| | (0.014) | (0.001) |
| $\hat{\rho}_1$ | 0.235 | 0.367 |
| | (0.009) | (0.005) |
| $\Delta Empl_{t\text{-}1}$ | | -0.129 |
| | | (0.009) |
| $\Delta Alpha_{t\text{-}1}$ | -0.248 | |
| | (0.001) | |
| $\Delta Beta_t$ | -0.291 | 0.004 |
| | (0.003) | (0.002) |
| $\Delta wage_t$ | -0.636 | 0.002 |
| | (0.005) | (0.004) |
| $\Delta dens_t$ | -0.925 | -0.159 |
| | (0.001) | (0.001) |
| Observations | 428 | 428 |
| Number of units | 107 | 107 |
| $\hat{\lambda}$ | 0.326 | 0.364 |
| $\hat{\sigma}_v^2$ | 0.001 | 0.001 |

**Note:** Standard errors are in parentheses.
**Source:** Own elaboration.

**Table 3:** CD Test for cross-sectional dependence (Pesaran, 2007).

| | $\varepsilon_e$ | $\varepsilon_{alpha}$ |
|---|---|---|
| z | 27.31 | 7.77 |
| p-value | 2.2e-16 | 7.579e-15 |

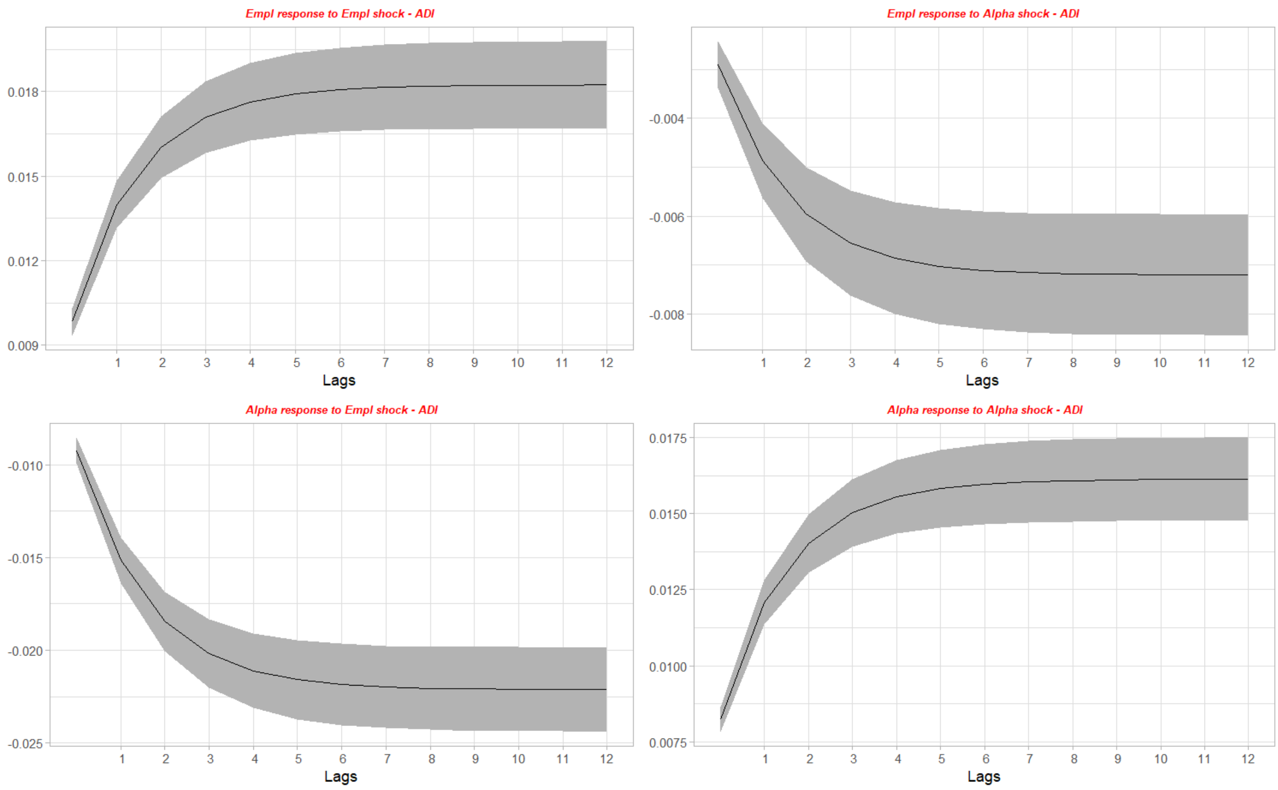
**Note:** The test of cross-sectional dependence in residuals. $H_O$: no cross-sectional dependence. The test is performed on the residuals of the first step of the BFP approach.

**Source:** Own elaboration.

## 6.2 Impulse Response analysis

As far as the causative nexus among variables, in the spatial panels the interpretation of the point estimates does not follow the one obtained in a non-spatial panel (Le Sage and Pace 2009), as they do not take into consideration global and local spatial effects embodied in the spatial model. To assess the effect of a change in endogenous variables, estimated coefficients must be corrected, bringing to the ATI, which provides the "right" magnitude of a change in an explanatory variable on the endogenous once we control for global and feedback spatial spillovers. ATI is then decomposed into ADI - the direct effect inside a territory, and, by difference, the AII – the feedback effects. Each decomposition must be repeated over the estimated horizon (twelve periods). We produce the cumulative impacts, calculated equation by equation, clearly showing the transitional paths to long-run equilibrium after a shock. The usual bootstrapping procedure calculates the standard errors.

**Figure 2:** OIRFs for the Average Direct Impact of the employment shock (left) and the entropy (right) shock. Error bands calculated using the bootstrapped approach.



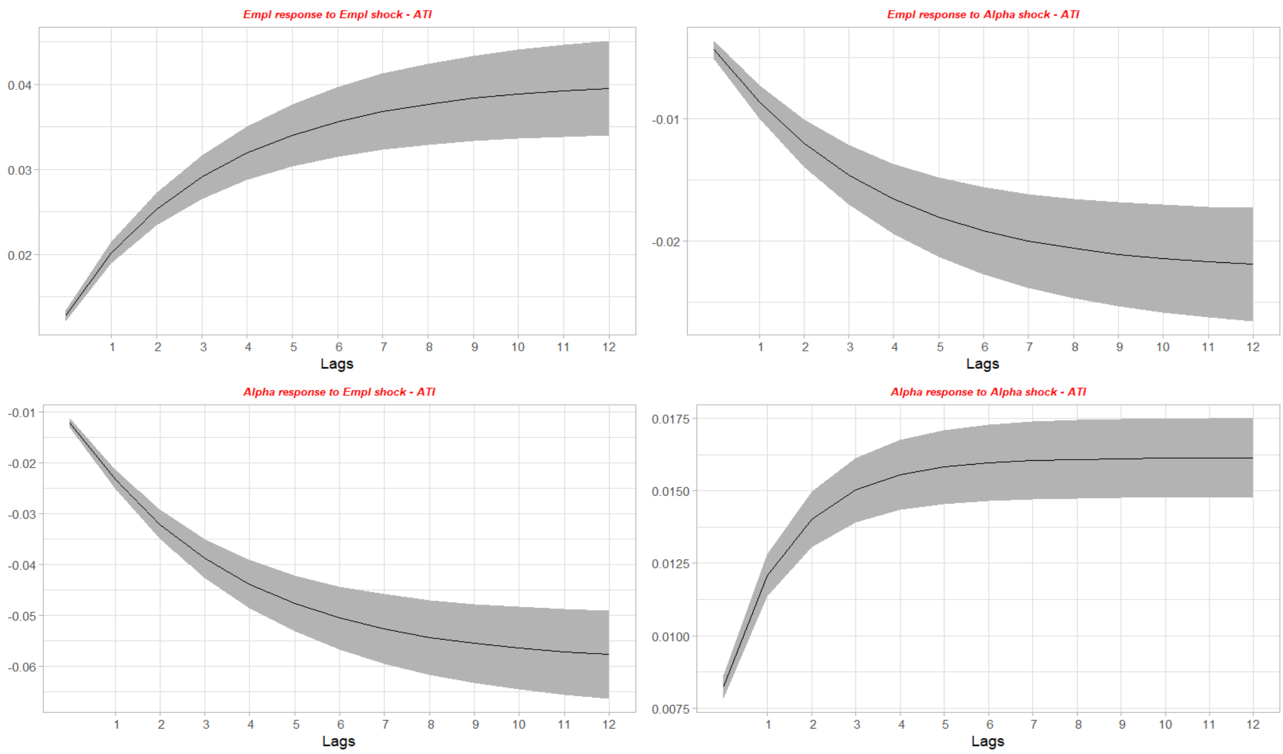

**Source:** Own elaboration.

On the left side, Figure 2 displays the cumulated OIRFs of ADI of the employment shock on the endogenous, while on the right shows the ADI of the entropy shock on the endogenous. The direct effects on the two endogenous are positive: a positive shock in employment has a significant and positive impact on the employment itself, and, likewise, a positive shock on alpha induces an increase in alpha. Both responses tend to converge to the steady-state after five periods, which was expected given the sign and magnitude of the speed adjustment coefficient in Table 2. Concerning the cross-impact of exogenous shocks, the analysis highlights that the effect is negative and significant. A positive employment shock leads to an alpha reduction. Therefore, if the number of employees increases, industrial diversification is reduced. Similarly, an increase in alpha drives a reduction in employment. The patterns return to the steady-state after five periods. Figure 3 shows the ATI, that is, how a change in exogenous variables spreads across provinces; it encompasses local spillovers. The dynamical paths are quite similar to the ADI ones.

In general, the empirical findings suggest a circular nexus between local employment and industrial variety. The relationship is negative: an increase in alpha causes a decrease in employment levels, and an increase in employment level induces a reduction in alpha. In other words, employment is always negatively related to the variety of industrial composition (alpha).

The results are consistent with the data analysis in Section 3; the actual industrial structure in Italy favors specialization. Only some sectors employ the most workers in each province. A change in the sectoral composition may reduce the number of employees. Moreover, if the industrial structure of provinces is more dispersed with respect to the national one, the impact on employment is negative.

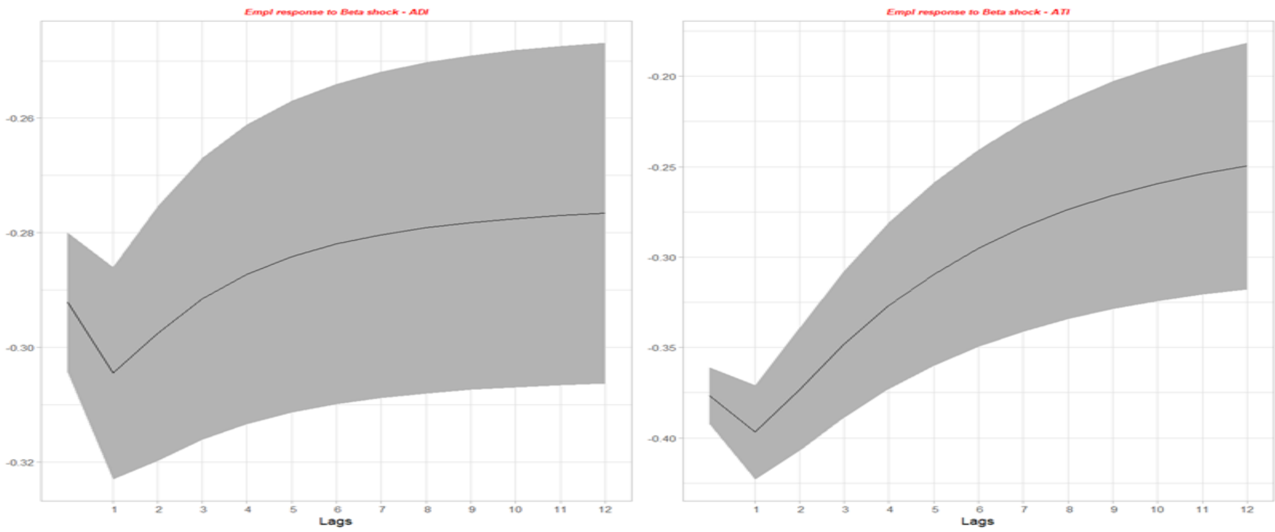
**Figure 3:** OIRFs for the Average Total Impact of the employment shock (left) and the entropy (right) shock. Error bands calculated using the bootstrapped approach.




**Source:** Own elaboration.

Finally, it is worth underlining the role that beta-entropy plays. Figure 4 shows the dynamic multiplier of a change in the beta-entropy on employment level. It is negative, meaning that when territories become more heterogeneous, this hampers local and national employment.

**Figure 4:** Dynamic multiplier for the Beta-entropy shock.




**Source:** Own elaboration.

## 7. Discussion and policy implications

This paper investigates the relationship between sectoral composition, as responsible for knowledge diffusion among firms, and the employment level. We did this by building an empirical bridge between the macro and the geographic approach: we exploit the VAR approach, augmented by spatial spillovers, and the impulse response functions, typical of macro econometrics, with entropy measures related to sectorial composition, typical of economic geography. Such an empirical framework should answer a twofold question: Does sectoral composition matter for employment level, hence for economic performance? Furthermore, at the same time, is the reverse true? The empirical findings suggest that both questions have a positive answer, at least for the Italian data. Section 3 reviews some stylized facts related to the sectorial composition as recorded by the ATECO classification, stressing that Italian workers are mainly distributed in a few sectors. The magnitude of alpha-entropy (variety) is medium-low, confirming the result.

Regarding geographical distribution, the average and median variety values are broadly equal. This means the provinces are split almost equally above or below the average. However, of the 57 provinces with alpha higher than the average, only nine belong to southern regions; conversely, 21 provinces, out of 50, with alpha lower than the average value, are in the north-center areas. In other words, the variety tends to be larger in provinces located in the north center and vice versa. However, this picture does not mean, per se, that the southern regions are disadvantaged because they are more specialized (alpha lower than the average); there are dynamic local territories in the South and less-performing provinces in the North. However, as pointed out in Sections 2 and 3, the alpha magnitude is generally more biased towards a sign of concentration than diversification. There is a certain degree of variability among provinces, but always in the middle-low values of alpha. In the spirit of MAR externalities, a brief analysis of the Italian districts provides further clues of an industrial specialization in low-tech, low-skilled sectors characterized by small and micro firms mainly located in northern provinces.

So, the question needs, once again, an empirical investigation. Our results are relatively straightforward; the impulse response functions conclude that the relationship between employment and variety is negative and circular: an increase in sectoral dispersion causes a decrease in employment performance, and an increase in employment requires a lower variety. The result is not surprising, considering what we found in Section 3. Only some sectors employ the most workers in each province. The empirical findings confirm that this broad specialization is beneficial for the Italian industrial tissue hence for the employment level, both at a local and a national level. From this point of view, the policy implications are straightforward: they should aim to foster specialization rather than introducing diversification in the industrial structure.

Moreover, the specialization model must be equalized among territories, reducing local disparities (the beta-entropy effect). In other words, the development model of the Italian economy still supports a broad sectoral specialization in traditional sectors characterized by low-tech, low-skill, low-size firms, such as wholesale and retail manufacturing, tourism, hospitality, food service, and agriculture (the “Made in Italy” in brief), which are well founded in a local context; see Giannini, Fiorelli and Martini (2022). In some sense, it resembles the Ricardian comparative advantage theory; rather than diversifying Italian production, it could be better to concentrate on the traditional sectors. At the same time, the dynamic multiplier on beta-entropy shows that such specialization should be balanced among the local economies, bringing a more common industrial tissue.

This opens an exciting policy question for the near future: Should Italy insist on the traditional productive tissue, as our results suggest, or should it engage a “structural break” more “future-oriented”? The answer is two-fold. If we look at the past (our sample lasted in 2018), our results suggest the traditional sectors. They were undoubtedly essential for Italian development in the last 50 years, and the relatedness of the sectors presumably fostered knowledge diffusion. However, concentration makes an economy more vulnerable to idiosyncratic shocks. Since 2020, due to pandemic crises, traditional sectors have been particularly hit by sanitary restrictions. In addition, traditional sectors are mainly populated by small, familiar firms, highly dependent on energy and supply costs, and the dramatic war in Eastern Europe is hitting hard. In other words, the traditional composition of low-skilled, low-tech small firms seems no longer suitable for a near future. The adverse effects of the pandemic and war crisis on the districts are well analyzed in the cited Intesa San Paolo report, where the weakness of this model in managing external shocks is evident.

So, on the one hand, we have an industrial model that was able to sustain the Italian economy, but on the other, it is fragile to specific shocks. How can we bring together such a contrasting view? Firms in these sectors are generally small, low-skilled, and obsolete from a technological point of view. So, investing in efficient technologies, human capital, new organizational models, alternative energy sources, and stimulating firm networks to share the risk is a way to make them more resilient. In other words, digitalization, green transition, and human capital investments are the keywords of the Next-Generation EU plan, implemented in Italy by the National Plane of Resilience and Recovery (NPRR).

Furthermore, this is true for the districts and the entire industrial structure. Too few firms exploited the Industry 4.0 plan, with some exceptions in the larger ones. Italian policymakers are advised to take advantage of this exceptional aid for sustaining SMEs to invest in more efficient productive technologies, training on the job, more collaborative organizational models, digital sales, and logistics. The NPRR implementation is not proceeding fast, mainly because of too complex bureaucratic and administrative procedures and a need for more professional skills. Moreover, our results show that actions must point to reducing the territorial divide instead of going towards a more territorial diversified productive structure. A strong form of local federalism, as the one on the Italian political agenda, goes in the opposite direction. Finally, the banking system must support entrepreneurs, and policies aiming to relax the credit constraint still need to be included.

Italy still needs a valid industrial policy; it disappeared from the agenda over the last 20 years, and data show the adverse effects. The NPRR represents a critical opportunity not only for recovery but also for laying the foundations for a sustainable and resilient economic future. To ensure long-term sustainability, it is essential to develop strategies that embrace innovation, inclusive growth, and environmental sustainability. This means investing in research and development to foster innovation and create an environment where new industries can thrive alongside traditional ones. Policies must also prioritize economic inclusivity, ensuring that the benefits of growth reach all regions and communities, especially those that have been left behind. This includes targeted investment in education and digital infrastructure to reduce the digital divide and improve the skills of the workforce. It is also necessary to promote green technologies and sustainable practices across all industries.

## 8. Conclusions

In this paper, we start from the literature focusing on knowledge spillovers as a driver for the development of an economy. The idea was old back in time since Marshall and had a vast echo both in the growth model of the 1960s and, more recently, in the endogenous growth literature during the 1990s. The question has been managed under several empirical and theoretical perspectives: macro, micro, and geographic economics. The latter, in particular, has adopted a more detailed approach to look inside the "black box" of industrial composition by building suitable measures of knowledge and technological diffusion, as recalled in the introduction. In particular, the related and unrelated variety measures were largely used in economic geography as a measure of industrial composition. We argue that such measures should be more correctly decomposed into alpha and beta components that are easier to interpret and more coherent with the mathematical foundations. In particular, the beta-entropy is a measure of the local geographical entity divergence. Local economies are not islands: they share knowledge, skills, and markets with the rest. In the last twenty years, spatial econometrics has enlarged our horizons, accounting for spatial spillovers among local economies. The decomposition of total entropy into gamma and beta components could be beneficial for integrating such an approach into a spatial econometric model.

Nonetheless, entropy measures are statistical tools for measuring industrial tissue concentration/dispersion, and more is needed to address the challenging question of how knowledge spills between sectors and, consequently, affects growth. For such a reason, we provide an empirical investigation. We improve the current empirical literature on the subject in a two-fold way. Firstly, we wonder whether a reverse nexus exists from employment levels to industrial composition. Second, by introducing spatial components in this bidirectional causation. Our results suggest a negative, circular relationship between employment and sectoral diversity: greater sectoral dispersion leads to poorer employment performance, while higher employment benefits from less diversity. This confirms that specialization within certain sectors is beneficial for local employment levels. Policy should therefore encourage specialization rather than diversification in industrial structures. In addition, it is important to balance specialization across regions to reduce disparities; a fair distribution of this industrial model is crucial to promote even economic development across the country. Italy is still a nation based mainly on the concentration of SMEs in traditional sectors. Rather than diversifying, policies must make this productive tissue more resilient by fostering its investment in human capital, technology, and digitalization. That is the core of the NPRR.

In conclusion, this empirical exercise confirms the significant role of sectoral composition in regional economic performance, and it also brings new insights into the bidirectional relationship between employment levels and industrial diversity. These insights not only validate but also extend theories of economic agglomeration and diversification.

Although our empirical results are clear, we also acknowledge inherent limitations. One is the aggregation of data by province. A logical next step would be to use microdata from individual firms to more accurately assess firm performance across sectors. In addition, extending this research to different economic environments could help to verify the strength of our conclusions.

**References**


Anderson, T. W. and Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. *Journal of econometrics*, 18(1), pp. 47-82. https://doi.org/10.1016/0304-4076(82)90095-1

Arellano, M. and Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, 58(2), pp. 277-297. https://doi.org/10.2307/2297968

Arrow, K. J., 1962. The economic implications of learning by doing. *The review of economic studies*, 29(3), pp. 155-173. https://doi.org/10.2307/2295952

Baldwin, R. E., Martin, P. and Ottaviano, G. I., 2001. Global income divergence, trade, and industrialization: The geography of growth take-offs. *Journal of Economic Growth*, 6(1), pp. 5-37. https://doi.org/10.1023/A:1009876310544

Baltagi, B. H., Fingleton, B. and Pirotte, A., 2014. Estimating and forecasting with a dynamic spatial panel data model. *Oxford Bulletin of Economics and Statistics*, 76(1), pp. 112-138. https://doi.org/10.1111/obes.12011

Baltagi, B. H., Fingleton, B. and Pirotte, A., 2019. A Time-Space Dynamic Panel Data Model with Spatial Moving Average Errors. *Regional Science and Urban Economics*, 76, pp. 13-31. https://doi.org/10.1016/j.regsciurbeco.2018.04.013

Beenstock M. and Felsenstein, D., 2007. Spatial Vector Autoregressions, *Spatial Economic Analysis*, 2(2), pp. 167-196. https://doi.org/10.1080/17421770701346689

Boschma, R., Eriksson, R. H. and Lindgren, U., 2014. Labour market externalities and regional growth in Sweden: The importance of labour mobility between skill-related industries. *Regional Studies*, 48(10), pp. 1669-1690. https://doi.org/10.1080/00343404.2013.867429

Boschma, R. and Capone, G., 2015. Institutions and diversification: Related versus unrelated diversification in a varieties of capitalism framework. *Research Policy*, 44(10), pp. 1902-1914. https://doi.org/10.1016/j.respol.2015.06.013

Boschma, R. and Capone, G., 2016. Relatedness and diversification in the European Union (EU-27) and European Neighbourhood Policy countries. *Environment and Planning C: Government and Policy,* 34(4), pp. 617-637. https://doi.org/10.1177/0263774X15614729

Boschma, R., Miguelez, E., Moreno, R. and Ocampo-Corrales, D. B., 2023. The Role of Relatedness and Unrelatedness for the Geography of Technological Breakthroughs in Europe. *Economic Geography*, 99(2), pp. 117-139. https://doi.org/10.1080/00130095.2022.2134005

Bouayad-Agha, S. and Vedrine, L. 2010. Estimation strategies for a spatial dynamic panel using GMM. A new approach to the convergence issue of European regions. *Spatial Economic Analysis*, 5(2), pp. 205-227. https://doi.org/10.1080/17421771003730711

Chavarro, D., Tang, P. and Rafols, I., 2014. Interdisciplinarity and research on local issues: evidence from a developing country. *Research Evaluation,* 23(3), pp. 195-209. https://doi.org/10.2139/ssrn.2736855

Duranton, G. and Puga, D., 2004. Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, 4, pp. 2063-2117. https://doi.org/10.3386/w9931

Eagle, N., Macy, M. and Claxton, R., 2010. Network diversity and economic development. *Science*, 328(5981), pp. 1029-1031. https://doi.org/10.1126/science.1186605

Frenken, K., Van Oort, F. and Verburg, T., 2007. Related variety, unrelated variety and regional economic growth. *Regional studies,* 41(5), pp. 685-697. https://doi.org/10.1080/00343400601120296

Gertler, M. S., 2003. Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of economic geography*, 3(1), pp. 75-99. https://doi.org/10.1093/jeg/3.1.75

Giannini, M., Fiorelli, C. and Martini, B., 2022. Ageing in the labour market: a spatial VAR approach. *Spatial Economic Analysis,* 17(4), pp. 538-556. https://doi.org/10.1080/17421772.2022.2036361

Giannini, M., Fiorelli, C. and Martini, B., 2022. Are Related and Unrelated Variety Suitable Measures at Analyzing Industrial Sectoral Composition? A Critical Review. *International Journal of Business and Social Sciences*, 13(1). https://doi.org/10.13140/RG.2.2.28462.84804

Guevara, M. R., Hartmann, D. and Mendoza, M., 2016. diverse: an R Package to Analyze Diversity in Complex Systems. *R J.*, 8(2), 60. https://doi.org/10.32614/rj-2016-033

Hidalgo, C. A., Klinger, B., Barabási, A. L. and Hausmann, R. 2007. The product space conditions the development of nations. *Science,* 317(5837), pp. 482-487. https://doi.org/10.1126/science.1144581

Jacobs, J., 1969. *The Economy of Cities*. Vintage, New York. https://doi.org/10.1080/14036096.2013.843125

Jacobs, J. P. A. M., Ligthart, J. E. and Vrijburg, H., 2009. Dynamic Panel Data Models Featuring Endogenous Interaction and Spatially Correlated Errors. *CentER Discussion Paper*, 2009. https://doi.org/10.2139/ssrn.1517257

Kapoor, M., Kelejian, H. H. and Prucha, I. R., 2007. Panel data models with spatially correlated error components. *Journal of econometrics*, 140(1), pp. 97-130. https://doi.org/10.1016/j.jeconom.2006.09.004

LeSage, J. and Pace, R., 2009. *Introduction to Spatial Econometrics*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781420064254

Lucas Jr, R. E., 1988. On the mechanics of economic development. *Journal of monetary economics*, 22(1), pp. 3-42. https://doi.org/10.1016/0304-3932(88)90168-7

Marcon, E., Scotti, I., Hérault, B., Rossi, V. and Lang, G., 2014. Generalization of the partitioning of Shannon diversity. *PloS ONE,* 9(3), e90289. https://doi.org/10.1371/journal.pone.0090289

Montresor, S. and Quatraro, F., 2017. Regional branching and key enabling technologies: Evidence from European patent data. *Economic Geography*, 93(4), pp. 367-396. https://doi.org/10.1080/00130095.2017.1326810

Neffke, F., Hartog, M., Boschma, R. and Henning, M., 2018. Agents of structural change: The role of firms and entrepreneurs in regional diversification. *Economic Geography*, 94(1), pp. 23-48. https://doi.org/10.1080/00130095.2017.1391691

Pinheiro, F. L., Alshamsi, A., Hartmann, D., Boschma, R. and Hidalgo, C., 2018. Shooting low or high: do countries benefit from entering unrelated activities? *Papers in Evolutionary Economic Geography*, 18(07). https://api.semanticscholar.org/CorpusID:169581834 .

Pesaran, M. H., 2007. A simple panel unit root test in the presence of cross-section dependence. *Journal of applied econometrics,* 22(2), pp. 265-312. https://doi.org/10.2139/ssrn.457280

Petralia, S., Balland, P. A. and Morrison, A., 2017. Climbing the ladder of technological development. *Research Policy*, 46(5), pp. 956-969. https://doi.org/10.1016/j.respol.2017.03.012

Rafols, I., Porter, A. L. and Leydesdorff, L., 2010. Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for information Science and Technology,* 61(9), pp. 1871-1887. https://doi.org/10.1002/asi.21368

Romer, P. M., 1986. Increasing returns and long-run growth. *Journal of political economy*, 94(5), pp. 1002-1037. https://doi.org/10.1086/261420

Van Dam, A. and Frenken, K., 2022. Variety, complexity and economic development. *Research Policy*, 51(8), 103949. https://doi.org/10.1016/j.respol.2020.103949

Yu, J., De Jong, R. and Lee, L. F., 2008. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics*, 146(1), pp. 118-134. https://doi.org/10.1016/j.jeconom.2008.08.002

Xiao, J., Boschma, R. and Andersson, M., 2018. Industrial diversification in Europe: The differentiated role of relatedness. *Economic Geography*, 94(5), pp. 514-549. https://doi.org/10.1080/00130095.2018.1444989

# Appendix I

## Properties of entropy measures

The entropy index ($H$) can be used to measure the dispersion degree in a distribution. It has many mathematical properties, particularly the additive one, allowing the index decomposition into subdimensions and subcategories. In a discrete random variable, if $p_i$ is the probability that an event occurs, which is usually estimated by its frequency in a sample (as examples: the share of income of a given group, the share of workers in a given industry and so on), the entropy measure $H(p_1, p_2, \ldots p_n)$ of the system $P$ must have these properties:

1) Positivity: $H(P) \geq 0$
2) Expansibility: Expansion of $P$ by a new component equal to 0 does not change $H(P)$
3) Symmetry: $H(P)$ is invariant under permutations of $p_1, \ldots, p_n$
4) Continuity: $H(P)$ is a continuous function of $P$ (for fixed $n$)
5) Additivity: $H(P \times Q) = H(P) + H(Q)$
6) Subadditivity: $H(X, Y) \leq H(X) + H(Y)$
7) Strong additivity: $H(X, Y) = H(X) + H(Y|X)$
8) Recursivity: $H(p_1, \ldots, p_n) = H(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$
9) Sum property: $H(P) = \sum_{i=1}^{n} g(p_i)$, for some continuous function $g$

   The logarithmic law owns this property. The Shannon entropy index:

$$H(P) = -\sum_{i=1}^{n} p_i \, ln(p_i) \tag{10}$$

is an example of an entropy measure, as it obeys the basic axioms. This entropy has a maximum when events are equiprobable (or uniformly distributed) $p_1 = p_2 = \ldots = p_n = 1/n$ so that $H(P) = ln(n)$. If the value of the index is near zero, the degree of dispersion is lower.

## Related and Unrelated Variety

The concept of variety, and its decomposition between related and unrelated variety, is known since the contribution by Frenken, Van Oort and Verburg (2007).[8] They apply the Shannon entropy index $H(p_i)$ (called "variety") to the three- or five-digit level of the SIC tree; then the variety is decomposed into UV (the entropy at the two-digit level) and RV, obtained by averaging over the entropies at five-

[8] The additive property of entropy index allows the decomposition in related and unrelated variety.

digit industries belonging to the same two-digit sector. Other contributions on the topic in Economics are Hidalgo et al. (2007), Rafols, Porter and Leydesdorff (2010), Chavarro, Tang and Rafols (2014), Guevara, Hartmann and Mendoza (2016), Eagle, Macy and Claxton (2010).

The total entropy (variety) at the finest tree's slice (five-digit as an example) is hence decomposed by the following formula:

$$\begin{aligned} H(p_i) &= \sum_{s=1}^{n} p_s \, ln(1/p_s) = \sum_{s=1}^{n} p_{si} \frac{p_g}{p_g} ln\left(\frac{1}{p_s}\frac{p_g}{p_g}\right) = RV + UV \\ RV &= \sum_{g=1}^{G} p_g \left( \sum_{s \in S_g} \frac{p_s}{p_g} ln\left(\frac{p_g}{p_s}\right) \right) = -\sum_{g=1}^{G} \left[ \sum_{s \in S_g} p_s \, ln\left(\frac{p_s}{p_g}\right) \right] \\ UV &= \sum_{g=1}^{G} p_g \, ln(1/p_g) \end{aligned} \tag{11}$$

For a given local community or entity (the province in our case), $n$ is the number of workers. The index compares two-digit ($p_g = n_g/n$) and five-digit ($p_s = n_s/n$) employment shares ; $s \in S_g$ are the species (five-digit industries) belonging to the same $g$ sector (two-digit) where $g = 1, \dots, G$. Hence $p_g = \sum_{i \in S_g} p_i$ ; it is the share of employment of a two-digit level. The UV component is simply the entropy calculated at the coarse level of the tree (two-digit, or slice-branch 1). The UV measures the degree to which employment shares are evenly distributed across sectors, in this case two-digit sectors. The UV values can vary from 0 (when all employment is concentrated in one two-digit sector) up to $ln(G)$ when all sectors employ an equal number of employees (uniform distribution).

In the following, we refer to two-digit classification as "sector" and five-digit as "subgroups" or "industries" belonging to the same sector. A given sector (two-digit) is hence composed of industries (five-digit) close to each other for type of production, or "related", according to the Frenken's definition. Following Giannini, Fiorelli and Martini (2022), RV is a measure of divergence between two- and five-digit distributions; technically speaking, it is the Jensen-Shannon divergence. More precisely, from equation (RV), it is the sum of the $G$ Kullback-Leibler (KL) divergences (the term inside the square brackets); it is always a positive term, as $p_i < p_g = \sum_{i \in S_g} p_i$. Therefore, RV is obtained as the difference between H and UV and measures the relative entropy between the two distributions. When both two and five digits are uniformly distributed $H = 2\ln(G)$, it implies $RV = UV = \ln(G)$. So, the maximum values for RV and UV can be the same. The lower RV, the more similar the two distributions are. Therefore, when $RV < UV$ the five-digit distribution is more concentrated than the two-digit one and vice versa. RV must be interpreted relative to UV, not by itself. Each territory has its own value for UV, depending on how many G sectors are covered (two-digit); this makes comparisons among territories difficult because they are dependent on the metrics. According to the EEG literature, a high value for RV should foster knowledge spillovers among widespread related industries via Jacobs's externalities, while a low-value points to very close and specialized five-digit subsectors, which more resembles the idea of MAR externalities. UV undoubtedly measures concentration/dispersion in the two-digit structure, and RV measures the same with respect to the five-digit structure once compared with UV. However, the interpretations given by the EEG literature do not seem coherent with what UV and RV really measure. In particular, the RV component cannot be used to assess the impact of Jacobs’s externalities as this measure does not account for the variety within sectors but assesses only how much the general variety H is different from the variety calculated between sectors. Furthermore, the EEG approach considers each entity as a single community, neglecting the role of the entire economic system.

## Appendix II - Estimation of the dynamic spatial panel

Starting from a general setting of dynamic spatial model:

$$y_{it} = \gamma y_{it_{-1}} + \rho_1 \sum_{j=1}^{N} w_{ij} y_{it} + x_{it}\beta + \varepsilon_{it} \qquad (12)$$

$i = 1..N, t = 1..T,$ where $y_{it}$ is the dependent variable for each elementary unit $i$ at time $t$, $x_{it}$ is a $Kx1$ vector of exogenous variables, $\gamma$ and $\beta$ are the corresponding coefficients. $w_{ij}$ is the $(i,j)$ element of the $NxN$ spatial matrix $W$. $\rho_1$ is the spatial lag coefficient. $\rho_1$ must lie in a province that makes $(I - \rho_1 W)$ non -singular. Moreover, stationarity requires $|\gamma| < 1.$ To capture a possible residual latent endogeneity among regressors, the residuals follow a spatial autoregressive (SAR) model with a random individual effect (RE):

$$\begin{aligned} \varepsilon_{it} &= \lambda \sum_{j=1}^{N} w_{ij} \varepsilon_{it} + u_{it} \\ u_{it} &= \mu_i + v_{it} \end{aligned} \qquad (13)$$

where $\mu_i$ is an individual-specific time-invariant component, which is assumed to be $i.i.d.(0, \sigma_\mu^2)$. $v_{it}$ is the error term $i.i.d.(0, \sigma_v^2)$, with $E(\mu_i, v_{it}) = 0 \quad \forall t.$ Moreover, the variance and covariance matrix of $\varepsilon_t$ is:

$$E(\varepsilon_t \varepsilon_t^{'}) = \sigma_\varepsilon^2 (B_N^{'} B_N)^{-1} \qquad (14)$$

with $\sigma_\varepsilon^2 = \sigma_\mu^2 + \sigma_v^2$ and $B_N = (I_N - \lambda W).$

As usual, the model is diff-transformed in order to remove all the time-invariant components, including the random effect $\mu_i$ and its correlation with time and space lagged variables:

$$\begin{aligned} \Delta y_{it} &= \gamma \Delta y_{it_{-1}} + \rho_1 \sum_{j=1}^{N} w_{ij} \Delta y_{it} + \Delta x_{it}\beta + \Delta\varepsilon_{it} \\ \Delta\varepsilon_{it} &= \lambda \sum_{j=1}^{N} w_{ij} \Delta\varepsilon_{it} + \Delta v_{it} \end{aligned} \qquad (15)$$

For a given cross-section $t$ this brings to:

$$\Delta y_t = \gamma \Delta y_{t_{-1}} + \rho_1 W \Delta y_t + \Delta x_t \beta + (I - \lambda W)^{-1} \Delta v_t \qquad (16)$$

The equation (delta model) is consistently estimated by the GMM or IV approach, to obtain a preliminary estimate of the coefficients $\hat{\gamma}$, $\hat{\rho}_1$ and $\hat{\beta}$. BFP suggest using the Anderson and Hsiao (1982) IV estimator by increasing the instruments list by $Wy_{t-2}$. The coefficient estimates allow calculating the residuals $\Delta\hat{\varepsilon}_t$ by Equation (delta model). The second step involves the estimation of $\hat{\lambda}$, and $\hat{\sigma}_v^2$ on the estimated variable $\Delta\hat{\varepsilon}_t$. As in BFP, we follow the GMM approach based on the Kapoor, Kelejian and Prucha (2007 - KKP) contribution. In particular, we use the first three moments condition of Equation 13 of KKP, as the differenced model erases the random effect component $\mu$, leading to two unknown coefficients ( $\lambda$ and $\sigma^2$ ) in three moments equations. The third step is actually the one-step GMM estimation of $\hat{\gamma}$, $\hat{\rho}_1$ and $\hat{\beta}$, given $\hat{\lambda}$ and $\hat{\sigma}_v^2$. The procedure is described in BFP, Equations 30 and 31. Basically, it relies on the Arellanno Bond (1991) moments conditions augmented by the spatial lags of $y$ and $x$. The fourth and last step follows Arellano and Bond (1991) again, by updating the weight matrix of the one-step GMM with the $\Delta v_t$ residuals obtained from the one-step GMM estimation, as described in Equation 32. The two-step GMM estimates of $\hat{\gamma}$, $\hat{\rho}_1$ and $\hat{\beta}$ is obtained. The procedure was implemented in R by the authors.

For both Equation of the model (spdpm1) and (spdpm3) only the variables $dens$ and clusters are treated as strictly exogenous in the vector of controls ( $X$ ). The other "exogenous" variables in the equations, i.e. alpha and beta entropies and compensations, have been treated as predetermined, changing their moment conditions with respect to the BFP contribution. For instance, in the first equation, the employment growth depends on variety at time $t-1$, but the current level of $Empl$ affects the sectoral diversity in the following period, since it enters in the alpha equation.

Hence, for each equation to be estimated, the moment conditions in our GMM model are the following.

$$\begin{aligned}
&E(y_{il}\Delta v_{it}) = 0\ \forall i, \quad l = 1,2...T-2;\ \ t = 3,4...T \text{ for the endogenous} \\
&E(x_{kil}\Delta v_{it}) = 0\ \forall i,k\ \ l = 1,2...T-1;\ \ t = 3,4...T \text{ for the predetermined} \\
&E(x_{k,il}\Delta v_{it}) = 0\ \forall i,k,l = 1,2...T;\ \ t = 3,4...T \text{ for the exogenous}
\end{aligned} \tag{17}$$

The moment conditions are gathered in a block-diagonal matrix $Z$ of instruments. In our case, we have $N = 107, T = 6, K_{pred.} = 3, K_{eso.} = 6, t = 3..6)$. Hence, for each equation to be estimated, $Z$ has $N \cdot (T-2) = 428$ rows and $196$ columns; $0.5 \cdot (T-2)(T-1) = 10$ for the endogenous variable, $[T \cdot (T-2)] = 24$ for each strictly exogenous variable and finally $[0.5 \cdot (T-2)(T+1)] = 14$ for each predetermined variable. The same procedure is then applied for the spatially lagged variables, bringing to the $Z_S$ matrix. Finally, the $Z$ and $Z_S$ are horizontally stacked to obtain the total instruments matrix $Z^*$.

Once the instruments matrix is obtained, we apply the standard GMM procedure; in the first stage GMM (the third step in BFP) we minimize, for each equation, the loss function:

$$L^2(\hat{\Gamma}) = \Delta\hat{v}'(\hat{\Gamma})Z^*A_N Z^{*'}\Delta\hat{v}(\hat{\Gamma}) \tag{18}$$

where $\hat{\Gamma}$ is the $11 \times 1$ vector of unknown coefficients to be estimated by $L^2$ minimization and $\hat{v}$ are the estimated residuals of the first step BFP procedure. $A_N$ is the GMM weight matrix:

$$A_N = Z^{*'}E(\Delta\varepsilon\Delta\varepsilon')Z^* \tag{19}$$

where:

$$E(\Delta\varepsilon\Delta\varepsilon') = (I_{T-2} \otimes H_N)(G \otimes I_N)\left(I_{T-2} \otimes H_N'\right) \tag{20}$$

and $G$ is the tri-diagonal matrix suggested by Arellano and Bond (1991), with 2 on the main diagonal and -1 up and below the diagonal. Finally, $H_N = (I_N - \hat{\lambda}W_N)^{-1}$. The minimization of the value function (step 1) by a Quasi-Newton algorithm provides the first step GMM estimator. The second step (fourth step in BFP) performs the same minimization but updating the weight matrix $A_N$ with the residuals of the model obtained by applying the first step estimates:

$$V_N = Z^{*'}(I_{T-2} \otimes H_N)(\Delta\hat{v}\Delta\hat{v}')\left(I_{T-2} \otimes H_N'\right)Z^* \tag{21}$$

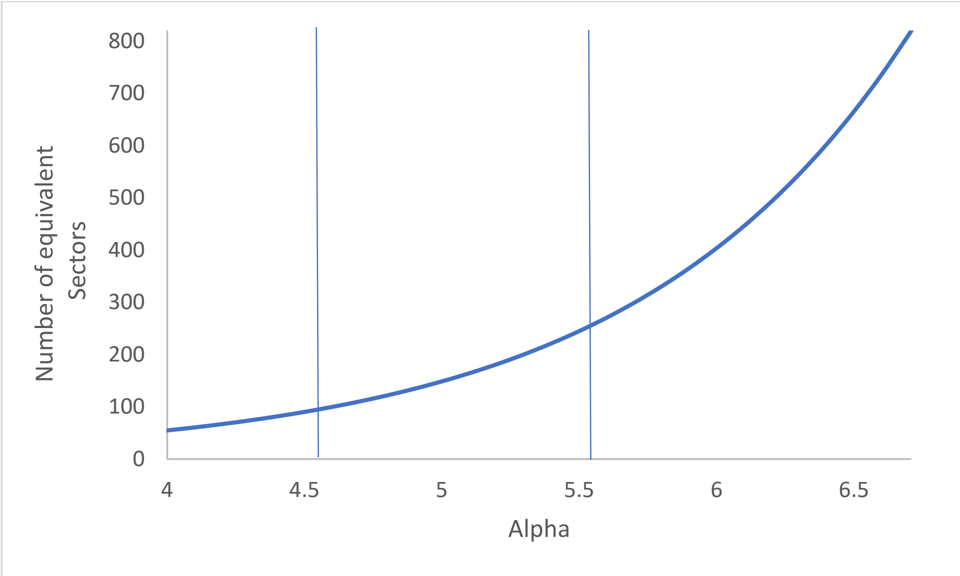
The variance-covariance matrix of the estimates is given by:

$$(\Delta X' Z^* V_N Z^* \Delta X)^{-1} \tag{22}$$

where $\Delta X$ is the matrix of the variables of the right-hand side of the estimated equation.

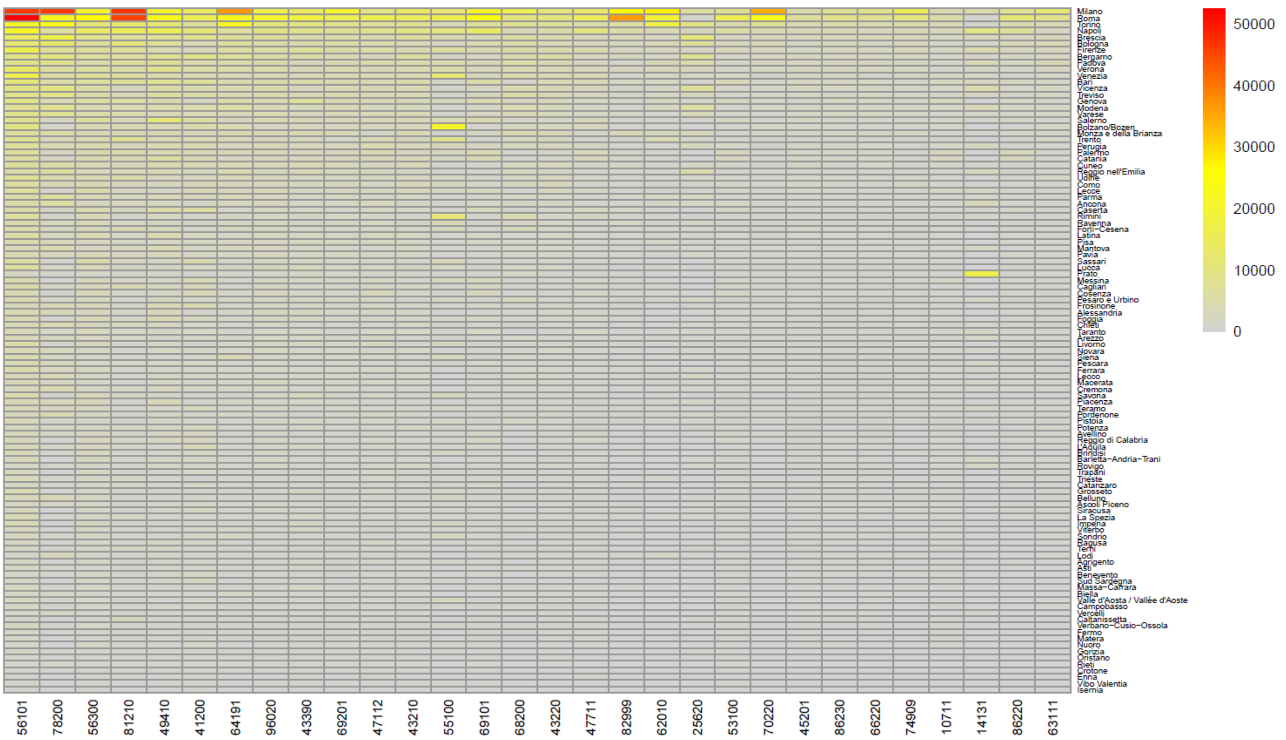
# Appendix III

**Figure 1A:** Alpha-entropy and equivalent sectors.



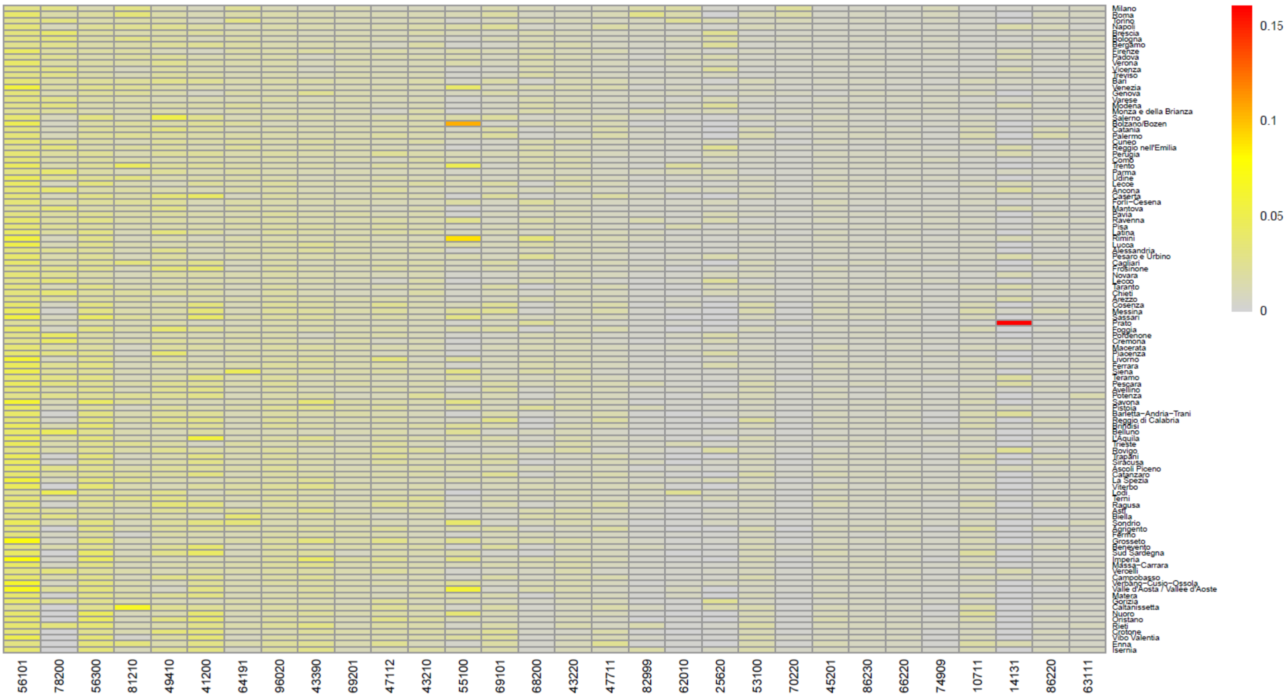

**Source:** Own elaboration.

**Figure 2A:** Heatmap of employment in the top 30 five-digit ATECO sectors.



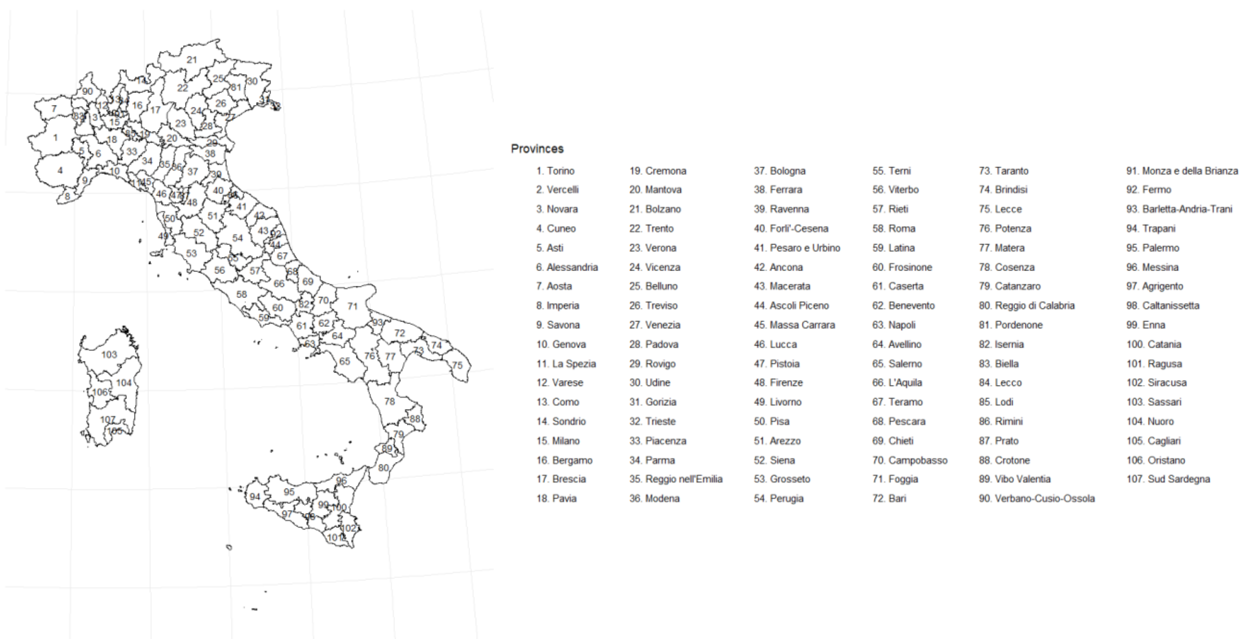

**Source:** Own elaboration.

**Figure 3A:** Heatmap for employment in the top 30 five-digit ATECO sectors (percentage on provincial employment).



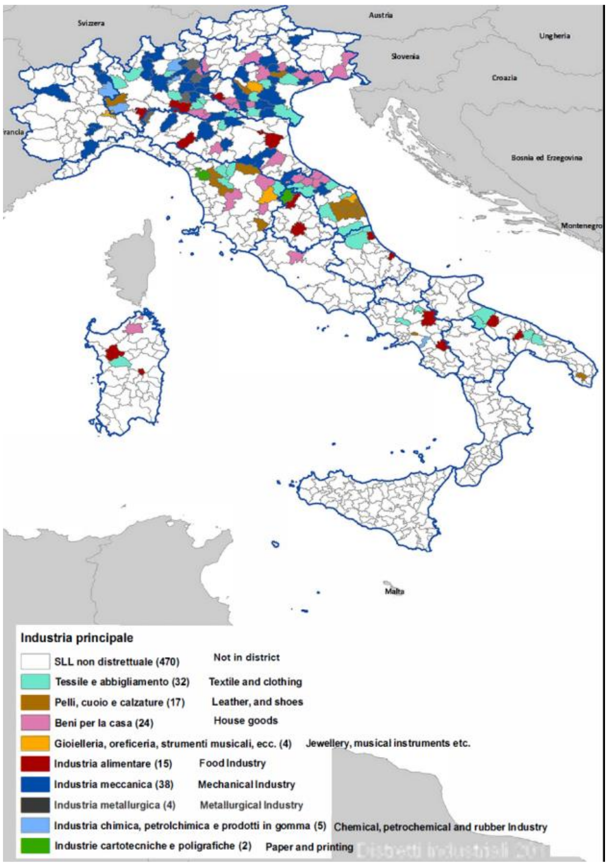

**Source:** Own elaboration.

**Figure 4A:** Map of Italian Provinces.




**Source:** Own elaboration.

**Figure 5A**: Italian districts map – 2011.




**Source:** ISTAT.

**Table 1A:** Descriptive statistics.

| ATECO code | Sector Description |
|---|---|
| 56101 | Restaurants and mobile food service activities |
| 78200 | Temporary employment agency activities |
| 56300 | Beverage serving activities |
| 81210 | General cleaning of buildings |
| 49410 | Freight transport by road |
| 41200 | Construction of residential and non-residential buildings |
| 64191 | Other monetary intermediation |
| 96020 | Hairdressing and other beauty treatment |
| 43390 | Other building completion and finishing |
| 69201 | Accounting, bookkeeping, and auditing activities; tax consultancy |
| 47112 | Supermarkets and similar stores |
| 43210 | Electrical installation |
| 55100 | Hotels and similar accommodation |
| 69101 | Legal activities |
| 68200 | Renting and operating of own or leased real estate |
| 43220 | Plumbing, heat and air-conditioning installation |
| 47711 | Retail sale of clothing in specialized stores |
| 82999 | Other business support service activities n.e.c. |
| 62010 | Custom software development |
| 25620 | Machining |
| 53100 | Postal activities under universal service obligation |
| 70220 | Business and other management consultancy activities |
| 45201 | Maintenance and repair of motor vehicles |
| 86230 | Dental practice activities |
| 66220 | Activities of insurance agents and brokers |
| 74909 | Other professional, scientific and technical activities n.e.c. |
| 10711 | Manufacture of bread; manufacture of fresh pastry goods and cakes |
| 14131 | Manufacture of other outerwear |
| 86220 | Specialist medical practice activities |
| 63111 | Data processing, hosting and related activities |

**Source:** Italian Industry Classification (ATECO).