Politecnico di Torino

DISAT

Corso di Laurea Magistrale in Physics of Complex Systems

Economic Complexity and Matrix Factorization

Inferring Hidden Capabilities in Municipal Production Networks

Relatore:
Prof. Luca Dall' Asta

Candidato:
Stefano Pietro Amedeo Massel
Matricola: 317623

Correlatori:
Prof. Andrea Tacchella
Dr. Matteo Straccamore
Dr. Alessandro Bellina

Anno Accademico 2024-2025

Contents

Abstract

In the framework of Economic Complexity, the concept of capabilities represents a fundamental theoretical construct for modeling economic systems. These capabilities, while extensively theorized as the hidden endowments driving countries' productive structures, have remained empirically elusive, never subjected to direct inference attempts. The present work addresses this critical gap by developing an approach to infer and extract information about this theoretical capabilities layer directly from empirical data. Our methodology employs Italian municipal-level economic data, specifically utilizing ATECO codes (the Italian classification system for economic activities harmonized with the European NACE nomenclature) organized in a binary matrix format, representing the presence or absence of economic activities in municipalities. We approach the problem through the lens of matrix factorization, treating the reconstruction of the municipality-activity matrix as an optimization problem where latent factors correspond to underlying economic capabilities. Employing a mask-and-predict methodology that systematically hide portions of the matrix to evaluate reconstruction accuracy, we determined that five capabilities constitute an excellent candidate in order to balance reconstruction precision against model parsimony. This dimensionality provides reconstruction accuracies already exceeding eighty percent for both the zeros and ones in the original binary matrix, demonstrating the method's robustness in capturing the underlying economic structure. The analysis reveals that the five identified degrees of freedom exhibit strong correlations with empirically observable economic-geographic properties of Italian municipalities, suggesting that our latent factors capture meaningful economic dimensions. Incorporating the fitness-complexity metrics into our model served dual purposes: consolidating the reliability of the Fitness-Complexity classification, which proved consistent with our analysis and revealing that the structural skeleton, introduced by the Fitness-Complexity algorithm, remains visible even when additional degrees of freedom are introduced. Specifically, reconstructions with more than one capability maintain a fundamental structure reminiscent of the Fitness-Complexity framework, suggesting that the approach captures essential features of economic organization that persist across different levels of dimensional reduction. Through appropriate null models, we demonstrated that the municipality-activity matrix possesses a sub-structure beyond what is captured by the fitness-complexity algorithm alone. This finding presents new challenges and opportunities for developing algorithms capable of revealing these deeper structural patterns. This thesis contributes

to the economic complexity literature by providing the first systematic attempt at capabilities inference from empirical data, trying to bridge the gap between theoretical frameworks and observable economic patterns.

Chapter 1

The Economic Complexity Framework

1.1 What is Economic Complexity

Economic Complexity constitutes a theoretical framework that applies the methods and mathematical tools of complex systems physics to the analysis of economic phenomena. This approach emerged as a fundamental departure from classical economic modeling, emphasizing that economic behavior emerges from the interactions between multiple heterogeneous agents, such as countries [19, 42], firms [27], products, and/or technologies [11], rather than from aggregated representative agents operating in equilibrium conditions. The framework fundamentally recognizes that these interactions generate non-linear dynamics, path dependencies, and emergent properties that cannot be understood through reductionist approaches.

The distinction from classical economics manifests in several critical dimensions. Traditional economic theory, rooted in neoclassical assumptions, relies heavily on equilibrium analysis, rational expectations, and representative agent models that assume homogeneity across economic actors. These models typically employ linear relationships and analytical solutions that, while mathematically tractable, fail to capture the heterogeneity and adaptive behavior observed in real economic systems. Classical approaches, exemplified by Solow growth models [37] and their derivatives, attribute economic development primarily to factor accumulation (i.e. physical capital, human capital, exogenous technological progress etc...) treating productivity as a residual unexplained by inputs. This framework struggles to explain persistent income disparities between nations with similar factor endowments and fails to account for the sudden economic transformations observed in countries like South Korea and Singapore.

Economic Complexity addresses these limitations by reframing economic systems as evolving networks where capabilities, seen as non-tradable productive knowledge embedded in organizations, institutions, and social structures, determine the feasible production space. This perspective shifts focus from aggregate quantities to the structural properties of economic networks, recognizing that, for example, the same amount of capital or labor can produce vastly different outcomes depending on how these factors combine with existing capabilities. The mathematical formalism draws from statistical physics, network theory, and non-linear dynamics, employing tools such as bipartite network analysis, fixed-point algorithms, and entropy maximization that reveal patterns invisible to traditional economic analysis.

1.1.1 Evolution of Theoretical Framework and Algorithms (ECI, EFC)

The foundational problem addressed by Economic Complexity concerned the quantification of national competitiveness through the country-product bipartite network, where edges represent revealed comparative advantages in international trade. Hidalgo and Hausmann's 2009 seminal contribution [19] introduced the Economic Complexity Index (ECI) through their Method of Reflections, an iterative linear algorithm designed to extract information about the productive capabilities of countries from their export baskets. The motivation arose from recognizing that product sophistication and country development exhibit mutual dependencies: developed countries produce complex products, while complex products are typically produced only by developed countries.

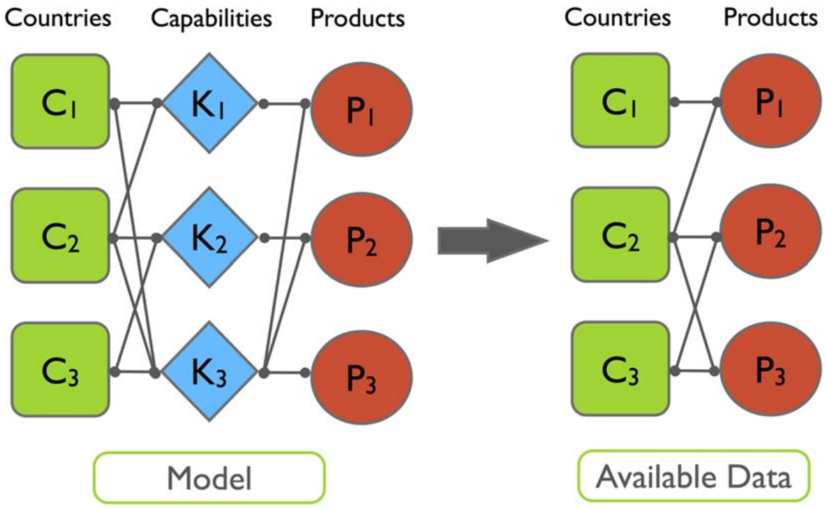


Figure 1.1: A schematic representation of the hidden capabilities layer. The real observable data is the contraction of the tripartite network Countries-Capabilities-Products: each country is connected to all and only those products for which owns all the necessary capabilities.[12]

The ECI algorithm operates on the binary country-product matrix M_{cp} (1.1), where $M_{cp} = 1$ indicates that country c has a revealed comparative advantage [3] (RCA > 1) in product p. The method initializes with two quantities: diversification $k_{c,0} = \sum_p M_{cp}$ (the number of products a country exports competitively) and ubiquity $k_{p,0} = \sum_c M_{cp}$ (the number of countries exporting a product). Through iterative refinement:

$$k_{c,N+1} = \frac{1}{k_{c,0}} \sum_p M_{cp} k_{p,N} \tag{1.1}$$

$$k_{p,N+1} = \frac{1}{k_{p,0}} \sum_c M_{cp} k_{c,N} \tag{1.2}$$

The algorithm appears conceptually straightforward: a country's complexity increases with the complexity of its exports, while a product's complexity reflects the average sophistication of its producers. However, this linear formulation harbors some mathematical and economic pathologies that limit its effectiveness.

The mathematical structure forces convergence to the eigenvectors of the matrix $\tilde{\mathbf{H}} = \mathbf{CMPM}^T$, where $\mathbf{C}$ and $\mathbf{P}$ are diagonal normalization matrices. By the Perron-Frobenius theorem, this Markov transition matrix drives the iteration exponentially toward a trivial fixed point where all countries converge to identical complexity values. Hidalgo and Hausmann circumvented this degeneracy by extracting rankings from the second eigenvector after mean subtraction, but this mathematical manipulation lacks economic justification.

Also the economic interpretation present some inconsistency. The arithmetic averaging inherent in the linear algorithm treats a country producing both sophisticated electronics and basic commodities identically to one specializing exclusively in mid-complexity manufactures. This equivalence erases crucial qualitative distinctions, indeed the diversified economy demonstrably possesses advanced capabilities absent in the specialized mid-complexity producer. The algorithm's linear structure cannot distinguish between a country's most sophisticated achievements and its average production, leading to systematic underestimation of diversified economies like China and the United States while inflating rankings of narrowly specialized nations.

Recognizing these limitations, Tacchella et al. in 2012 [42] introduced the Economic Fitness and Complexity (EFC) algorithm, implementing a non-linear approach instead of the previous linear one. The key conceptual innovation is that a product's complexity should be bounded by the least capable producer, not averaged across all producers. If both Japan and a developing nation export basic electronics, the product's accessibility to low-capability producers reveals its limited complexity requirements.

The EFC algorithm employs coupled non-linear maps with harmonic mean aggregation:

$$\tilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)} \tag{1.3}$$

$$\tilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp}/F_c^{(n-1)}} \tag{1.4}$$

followed by normalization at each iteration:

$$F_c^{(n)} = \frac{\tilde{F}_c^{(n)}}{\langle \tilde{F}_c^{(n)} \rangle_c} \tag{1.5}$$

$$Q_p^{(n)} = \frac{\tilde{Q}_p^{(n)}}{\langle \tilde{Q}_p^{(n)} \rangle_p} \tag{1.6}$$

The Fitness F_c aggregates a country's productive capabilities through linear summation weighted by product complexities, explicitly rewarding diversification. Conversely, the complexity Q_p employs harmonic averaging, ensuring that any low-fitness producer constrains the product's measured complexity. This asymmetry captures the fundamental economic insight that production requires all necessary capabilities, so their presence enables production while any absence prevents it.

The non-linear structure of the EFC generates Pareto-distributed fitness and complexity values, matching empirical observations of heavy-tailed economic distributions. In this case few countries possess disproportionate capabilities and few products require rare capability combinations. This mathematical property arises naturally from the multiplicative dynamics of capability accumulation, in which existing capabilities facilitate the acquisition of new ones through preferential attachment mechanisms. Indeed the addition of a new capability has a greater impact on export diversification when the country already possesses a larger pool of capabilities.

The predictive power of EFC has been extensively validated, particularly by Cristelli et al. [13, 14] who demonstrated its superiority in forecasting GDP growth compared to traditional economic indicators. Countries with high fitness compare to GDP per capita consistently experience subsequent growth as they monetize latent productive potential. The framework reveals regime-dependent predictability: high-fitness countries with moderate income display deterministic *laminar* growth trajectories driven by their diverse capability portfolios, while low-fitness countries experience *chaotic* dynamics dominated by exogenous shocks rather than endogenous development.

The applications of EFC have expanded across multiple domains and scales. At the scientific production level, Cimini et al. [11] applied the framework to mea-

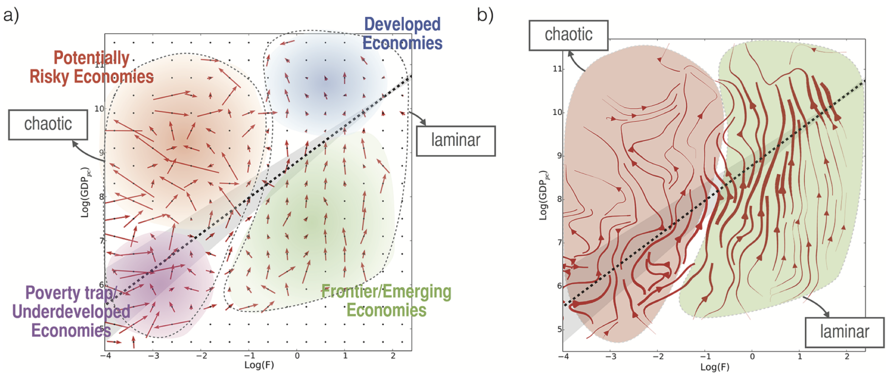


Figure 1.2: a) A finer coarse graining of the dynamics highlights two regimes for the dynamics of the evolution of countries in the fitness-income plane b) continuous interpolation of the coarse grained dynamics to illustrate the two regimes of predictability [13]

sure national research capabilities, revealing that scientific domains accessible to countries with weak research infrastructure signal modest knowledge requirements, while domains monopolized by scientifically advanced nations require extensive prerequisite capabilities. The framework has been applied to patent data for innovation analysis [4], urban economic dynamics [38, 40], and regional development strategies. In particular, recent work by Straccamore et al. [39] extended the analysis to metropolitan scales, discovering scale-dependent relationships between specialization and diversification: cities benefit from technological coherence through knowledge spillovers, while nations require broad diversification for resilience and combinatorial innovation.

Particularly relevant for the present thesis, the EFC framework has been successfully applied at increasingly fine geographical resolutions. The Enrico Fermi Research Center has developed integrated databases enabling analysis at regional and municipal levels, utilizing detailed economic activity classifications to uncover local productive structures. The application to Italian municipalities using ATECO codes, as undertaken in this thesis, represents a natural extension of this multiscale approach, seeking to identify the latent capabilities underlying local economic development patterns.

Recent methodological developments have further refined the theoretical foundations and practical applications of the Fitness-Complexity framework. Servedio et al.[36] addressed numerical stability issues in the original algorithm by introducing a modified formulation with inhomogeneous terms. Their approach ensures convergence even for countries with extremely low export volumes, which previously tended toward zero fitness in the original metric. The new formulation also provides an approximate analytical solution, enhancing interpretability and enabling parametric sensitivity analysis. Moreover the robustness of the framework to data quality issues has been systematically analyzed by Battiston et al. [5], who demonstrated that the Fitness-Complexity algorithm exhibits remarkable robustness to random noise, maintaining high correlation with noise-free rankings even when substantial proportions of matrix entries are randomly flipped. This robustness stems from the algorithm's exploitation of global network structure rather than relying on individual matrix elements.

The integration of multiple data sources represents another significant advance in the field. Recent efforts have focused on constructing integrated databases combining trade, patent, scientific publication, and firm-level data. The integrated database developed at the Enrico Fermi Research Center provides a unified framework for Economic Complexity analysis across multiple types of economic activities [29]. This enables researchers to study co-evolution of capabilities across domains. For example, how scientific capabilities in a country relate to its technological innovations and subsequent product exports (Pugliese et al.[31]). This systemic view highlights the interconnectedness of different aspects of the knowledge economy.

The proliferation of Economic Complexity methods has prompted comparative studies assessing their relative merits. Tacchella et al.[44] examined relatedness measurement in the era of machine learning, comparing traditional co-occurrence-based proximity measures with model-based approaches using neural embeddings and collaborative filtering. Machine learning techniques offer the potential to uncover latent structure in high-dimensional capability spaces that may not be apparent through simple proximity calculations. However, these methods often sacrifice interpretability for predictive performance, making it difficult to extract actionable insights for policy. The trade-off between interpretability and predictive power remains an active area of methodological research, with recent work exploring hybrid approaches that combine the strengths of both paradigms[41, 2, 17].

1.1.2 EFC: Mathematical Foundation Analysis

The mathematical foundations of the Fitness-Complexity algorithm (EFC) have undergone rigorous analysis, revealing deep connections to established mathematical frameworks and resolving questions about convergence and stability. Pugliese et al.[32] provided the first comprehensive analysis of convergence properties, demonstrating that the adjacency matrix structure determines which countries and products converge to non-zero fitness and complexity values. Their analysis revealed that convergence requires the reordered matrix's diagonal to remain within the occupied region showing that matrices violating this condition produce degenerate solutions where most entities collapse to zero values.

A critical advancement came from Servedio et al.[36], who introduced an inhomogeneous non-linear metric addressing the original algorithm's stability limitations. Their modification adds small positive parameters to the iterative equations:

$$\tilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)} + \varphi_c \tag{1.7}$$

$$\tilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp}/F_c^{(n-1)} + \pi_p} \tag{1.8}$$

After appropriate rescaling, these parameters can be set to zero, yielding a parameter-free metric that maintains the original algorithm's economic interpretation while guaranteeing convergence and stability even for challenging matrix structures. This formulation enables an approximate analytical solution at first order:

$$F_c \approx D_c - \sum_{c'} \frac{K_{cc'}}{D_{c'}} \tag{1.9}$$

where $D_c = \sum_p M_{cp}$ represents diversification and $K_{c_1 c_2} = \sum_p M_{c_1 p} M_{c_2 p}$ is the co-production matrix. This analytical form reveals that fitness primarily derives from diversification, with a correction term (inefficiency) that penalizes countries for producing products also made by low-diversification economies. The concept of net-efficiency, measuring deviation from the average inefficiency-diversification relationship, quantifies how effectively countries target high-complexity products rather than indiscriminately expanding export baskets.

However, the most profound theoretical breakthrough emerged from Mazzilli et al. (2024)[26], who demonstrated the mathematical equivalence between the Fitness-Complexity algorithm and the Sinkhorn-Knopp matrix scaling algorithm. This connection transforms EFC from an empirically motivated metric to a principled optimization problem minimizing a logarithmic barrier function:

$$g(x, y) = \sum_{ij} x_i A_{ij} y_j - \sum_{i=1}^{n} r_i \ln x_i - \sum_{j=1}^{m} c_j \ln y_j \tag{1.10}$$

The fitness and complexity emerge as logarithmic potentials in this optimization framework, with their ratio Q_p/F_c that can be seen as the energy cost for country c to produce product p. This interpretation explains the triangular nested structure observed in reordered country-product matrices: high-complexity products become energetically unfeasible for low-fitness countries, creating a natural barrier in the capability space.

The Sinkhorn-Knopp connection links EFC to optimal transport theory, where similar algorithms solve entropy-regularized Wasserstein distance problems. This mathematical unification reveals that EFC implicitly solves a resource allocation problem, finding the most efficient distribution of productive capabilities across countries and products subject to observed trade constraints. The scale invariance inherent in the Sinkhorn formulation explains why logarithmic transformations are required for forecasting applications and provides theoretical justification for comparing datasets across different years or aggregation levels.

In the end, recent extensions by Servedio et al. (2025)[35] have generalized the fitness concept beyond bipartite networks to arbitrary graph structures, introducing fitness centrality as a novel network measure. This generalization reveals that high fitness centrality identifies nodes crucial for network connectivity (those whose removal would isolate many other nodes). Applied to economic networks, this can be seen as a way to identify systemically important countries whose productive capabilities enable global value chains.

1.1.3 Capabilities: Underlying Structure and Relatedness

The capabilities framework provides the theoretical foundation for interpreting the mathematical structures revealed by complexity algorithms. This framework theorizes that economic production emerges from combinations of non-tradable endowments (capabilities) representing embedded knowledge within productive ecosystems (see Fig. 1.1). These capabilities can extend beyond traditional production factors to include tacit organizational knowledge, institutional quality that facilitates complex transactions, physical and digital infrastructure, regulatory frameworks, and social capital that facilitates collaboration [19].

Classically, mathematical formalization represents this through binary indicators where S_{ck} denotes whether the country c possesses the capability k, and T_{kp} indicates whether the capability k is required for the product p. The observed country-product matrix emerges through:

$$M_{cp} = \prod_{k} \left[1 - T_{kp}(1 - S_{ck})\right] \tag{1.11}$$

This multiplicative structure embodies a fundamental principle: production requires all necessary capabilities to be present simultaneously. The absence of a single critical capability prevents production regardless of other capability abundances, generating the nested triangular patterns observed empirically in country-product matrices where countries and products naturally order by capability endowments and requirements respectively.

Capabilities are expected to exhibit strong path dependencies and contextual embeddedness that distinguish them from transferable production factors. The pharmaceutical industry exemplifies this concept, as innovation tends to concentrate in specific locations not because of capital abundance, but due to the co-location of complementary capabilities such as research infrastructure, regulatory expertise, intellectual property systems, academic linkages, and specialized venture capital ecosystems.

While the Fitness–Complexity framework quantifies countries' capability stocks [42], understanding development dynamics requires examining the geometry of the capability space. Hidalgo et al. [20] formalized this through the *Product Space*, where proximity between products reflects the similarity of their underlying capability requirements. The proximity metric $\phi_{pp'}$ is defined as:

$$\phi_{pp'} = \min\left(P(\mathrm{RCA}_p > 1 \mid \mathrm{RCA}_{p'} > 1),\, P(\mathrm{RCA}_{p'} > 1 \mid \mathrm{RCA}_p > 1)\right) \tag{1.12}$$

Operationally, it is computed as:

$$\phi_{pp'} = \frac{\sum_c M_{cp} M_{cp'}}{\max(k_{p,0},\, k_{p',0})} \tag{1.13}$$

where $k_{p,0} = \sum_c M_{cp}$ denotes the *ubiquity* of product p.

The Product Space exhibits a core–periphery topology: sophisticated products form a densely interconnected core, while primary goods occupy isolated peripheral positions, thereby constraining countries' diversification paths. *Relatedness* captures the empirical regularity that developing a new comparative advantage is more likely when it is related to existing productive activities, and is quantified through the *Relatedness density*:

$$\omega_{cp} = \frac{\sum_{p'} M_{cp'} \phi_{pp'}}{\sum_{p'} \phi_{pp'}} \tag{1.14}$$

Empirical validation demonstrates entry probability into products at proximity 0.8 reaches approximately 15 percent versus near-zero at proximity 0.1 [20]. This principle extends across scales: regional industrial dynamics [28], urban technological development [22], and firm-level diversification [27]. Alternative proximity measures include the Taxonomy Network [46] and various skill-relatedness metrics.

The Fitness-Complexity and relatedness frameworks provide complementary perspectives: fitness quantifies capability accumulation while relatedness reveals acquisition pathways. High-fitness countries occupy dense Product Space regions enabling continuous upgrading. Low-fitness countries face sparse peripheries where sophisticated products require prohibitive capability distances. Recent advances combine both frameworks to identify optimal diversification targets maximizing feasibility through high relatedness and future opportunities through complexity and centrality [1]. The present thesis extends this framework through matrix factorization techniques applied to Italian municipal ATECO data, providing the first systematic attempt at capabilities inference from empirical data, trying to bridge the gap between theoretical frameworks and observable economic patterns.

1.2 Policy Implications and Applications

The Economic Complexity framework has gained traction among policymaking institutions as a practical tool for informing industrial policy, development strategy, and growth forecasting. This adoption reflects a shift in how governments and international organizations approach economic development planning, moving from traditional schemes to data-driven analysis of productive capabilities and their evolution.

The framework's appeal to policymakers stems from several methodological advantages that address challenges in development economics. The metrics are computed directly from observable data on trade flows, patent filings, or employment patterns, eliminating the need for subjective assessments of technological sophistication or strategic importance. This empirical grounding provides a degree of objectivity that has sometimes been lacking in traditional approaches to industrial policy, where sector selection might reflect political considerations or simplified narratives about technology rather than systematic analysis of productive capabilities.

Moreover the framework is inherently forward-looking because fitness captures latent productive potential that can become an income growth with a temporal lag. This predictive power enables proactive policy interventions designed to guide structural transformation rather than reactive responses to economic crises. Research has demonstrated that productive structure serves as a robust predictor of medium and long term growth trajectories, with prediction performance particularly strong for countries exhibiting what has been termed "laminar" growth patterns characterized by stable capability accumulation.

The granular nature of complexity analysis provides actionable guidance that generic development prescriptions cannot match. By operating at the product level for trade analysis or the technology class level for patent analysis, the framework enables identification of specific diversification opportunities tailored to each country's existing capability base. Rather than recommending broad sectoral priorities like "develop manufacturing" or "invest in technology," complexity-based analysis can identify quite precisely which products or technologies are both feasible given current capabilities and valuable for opening pathways to further sophisticated diversification. The framework measures capability breadth rather than focusing on individual champion sectors, naturally encouraging policies that build systemic competitive advantage across the productive structure. This systems perspective aligns with modern understanding of economic development as an integrated process of capability accumulation rather than isolated sectoral advances.

The distinction between predictable and unpredictable growth patterns has important policy implications. The Selective Predictability Scheme, developed through World Bank research, classifies countries into those exhibiting laminar growth whose future trajectories can be reliably forecast from productive structure, and those showing more turbulent patterns where prediction becomes difficult. This classification enables policymakers to tailor interventions based on the stability of their development pathway, with countries in laminar regimes benefiting most from incremental capability-building strategies while those in turbulent regimes may require more fundamental structural reforms.

For countries trapped in middle-income stagnation, particularly those whose resource wealth creates disincentives for diversification, the framework offers specific strategic guidance. Rather than pursuing income growth solely through intensified resource extraction, complexity analysis recommends prioritizing export diversification toward products of moderate complexity that lie within reach of existing capabilities. This strategy of "lateral escape" enables countries to build productive capacities that lower barriers to subsequent industrialization, creating pathways out of resource dependence even before reaching high income levels. The emphasis on gradual capability accumulation through related diversification contrasts with strategies that attempt to jump directly to highly sophisticated production without intermediate steps.

Numerous institutions started to implement Economic Complexity in their policies, an example is the World Bank which has broadly integrated Economic Complexity metrics into its Country Economic Memoranda and competitiveness assessments, using them to evaluate productive structures and design development interventions. Also the European Commission has adopted relatedness and complexity analysis as analytical tools for evaluating regional innovation strategies within its Smart Specialization framework [4]. Since the early 2010s, EU regional policy has required member states to develop place-based innovation strategies that identify priority areas for investment based on existing regional strengths. The empirical application of complexity methods to European patent data has revealed substantial heterogeneity in technological capabilities across regions and demonstrated that successful diversification occurs predominantly into technologies related to existing regional specializations. Research examining 285 NUTS-2 regions using data on 36 technology classes has shown that relatedness density significantly predicts technological entry, with a 10 percent increase in relatedness density raising entry probability by approximately 23-26 percent [4]. However, evaluations of implemented Smart Specialization strategies reveal a gap between theory and practice, with many regions selecting priorities that neither build on related capabilities nor target appropriately complex activities, suggesting room for improved analytical guidance. United Nations agencies have embraced Product Space analysis as a framework for identifying feasible diversification paths. UNCTAD has developed comprehensive catalogues (UNCTAD "Catalogue of Diversification Opportunities 2022") presenting potential new products for 233 economies based on analysis of their economic complexity and position in the product space. These catalogues aggregate information on over 45,000 product lines differentiated by price ranges, providing developing countries with data-driven guidance for export diversification strategies. UNCTAD's methodology emphasizes products that are both technologically accessible given existing productive capabilities and subject to favorable global demand conditions. For instance, complexity analysis identified mechanical appliances, pharmaceutical products, and plastics as priority diversification opportunities for Angola, with plastics being particularly appropriate given the country's position as Africa's second-largest oil producer.

Chapter 2

Data and Methods

2.1 Data

The empirical analysis is based on firm-level information drawn from the 2021 *Statistical Register of Active Enterprises* (ASIA), maintained by the Italian National Institute of Statistics (ISTAT). The ASIA archive provides comprehensive coverage of the productive system and, for the reference year, includes records for 4,929,379 local units belonging to non-agricultural firms. Each local unit corresponds to a geographically identifiable component of an enterprise, such as a workshop, factory, warehouse, office, mine, or depot and contains detailed attributes including its address, employment size, and activity sector. Sectors are classified according to the five-digit ATECO system [21], which is fully aligned with the first four digits of the European NACE Rev. 2 taxonomy [16].

Before being used in our analysis, the ASIA records underwent geocoding and harmonization carried out by the Study Center of the Italian Chambers of Commerce "Guglielmo Tagliacarne". This procedure ensured spatial consistency across municipalities but resulted in a data reduction of approximately 3%, primarily due to incomplete or non-standardized addresses.

Construction of the Municipality–ATECO Matrix

The central component of the empirical framework is the construction of the binary bipartite matrix M_{ma}, which encodes the presence or absence of statistically significant economic activities a within each municipality m. Starting from the ASIA microdata, we aggregated all local units by municipality and ATECO code, obtaining raw counts of activity occurrences.

Because municipalities differ in population, economic scale, and firm density, raw frequencies alone do not provide a meaningful indicator of specialization. To statistically validate each municipality–activity pair, we compared the empirical counts with an ensemble of 1,000 randomized bipartite networks generated using the Bipartite Configuration Model (BiCM) [34, 33]. The BiCM preserves the degree sequences of both municipalities and activities, making it an appropriate null model for evaluating the over- or under-representation of specific ATECO codes.

This procedure produced a matrix of p-values,

$$\mathbf{P} = \{\mathrm{P}_{ma}\},$$

computed using the `NEMtropy` Python package (`github.com/nicoloval/NEMtropy`) [45]. Each entry P_{ma} quantifies the probability that the observed count for municipality m and activity a could arise under the null hypothesis defined by the BiCM.

The binary matrix M_{ma} was then obtained by applying a significance threshold of 0.05:

$$M_{ma} = \begin{cases} 1, & \text{if } \mathrm{P}_{ma} \leq 0.05, \\ 0, & \text{if } \mathrm{P}_{ma} > 0.05. \end{cases}$$

The resulting bipartite matrix connects 7,841 Italian municipalities to 814 ATECO codes, and constitutes the foundational dataset for all subsequent analyses of productive capabilities and structural patterns.

2.2 Methods

Matrix Factorization Approach to Capability Inference

Within the capability framework [12], the observed binary country-product matrix can be modeled as the product of a latent country-capabilities and a capabilities-products binary matrices, followed by a nonlinear binarization:

$$M_{cp} = \prod_k \left[\, 1 - T_{kp}\,(1 - S_{ck})\,\right], \tag{2.1}$$

where $S_{ck} \in \{0, 1\}$ indicates whether country (or municipality) c possesses capability k, and $T_{kp} \in \{0, 1\}$ indicates whether capability k is required to produce product p. A product can therefore be exported only if all requirements (required capabilities) are present, giving rise to a strongly nonlinear mapping between latent capabilities and observed diversification.

Our aim is to invert this process, starting from the observed binary Municipality–ATECO matrix M (Figure 2.1) and inferring the two latent matrices S and T. Because the logical binarization is non-invertible since it provides no information about which capabilities are missing when an entry is zero. Direct inversion is impossible.

To proceed, we relax the binary assumption and recast the problem within a *matrix factorization* framework. Specifically, we seek two nonnegative matrices,

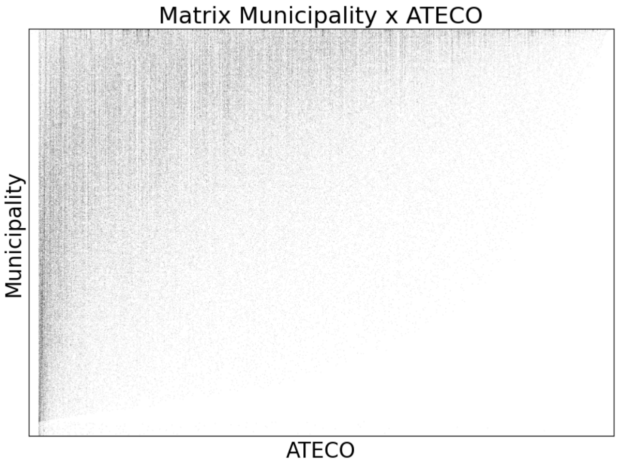


Figure 2.1: Municipality-ATECO incidence matrix M. Rows represent municipalities ordered by Fitness, columns correspond to ATECO production classes ordered by Complexity

$S \in \mathbb{R}_{+}^{N_c \times k}$ and $T \in \mathbb{R}_{+}^{k \times N_p}$, such that

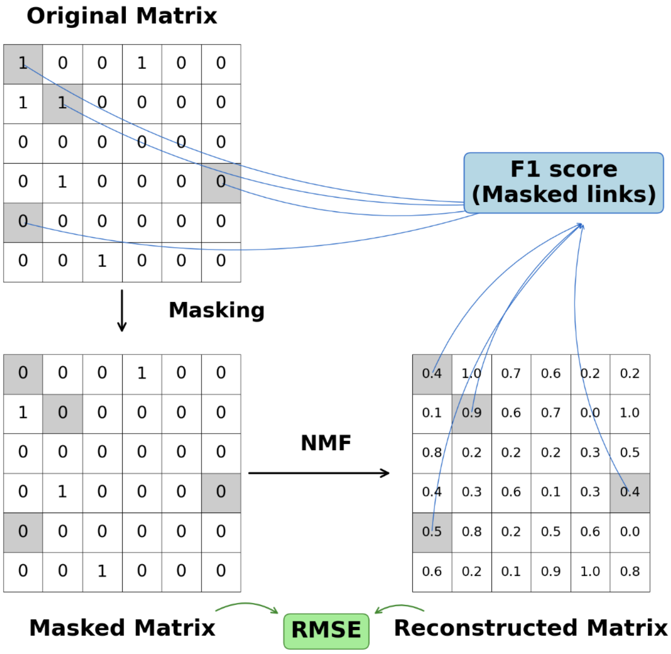
$$M \approx \theta(S \times T), \tag{2.2}$$

where $\theta(\cdot)$ is a thresholding function applied for evaluation purposes, and k represents the number of latent capabilities. The central hyperparameter of the model.

From matrix factorization theory [24, 18], exact reconstruction of a matrix requires that the latent dimensionality equals its rank. In our case, the Municipality–ATECO matrix is nearly full rank, implying that about $k = 805$ components would be needed for exact reconstruction of 814 products. Such a solution is both theoretically implausible and practically meaningless in a capability framework, where products should represent combinations of a limited number of fundamental building blocks.

Maximizing in-sample reconstruction accuracy alone does not solve this issue: the reconstruction precision increases monotonically with k because model expressivity grows with dimensionality. Without additional validation, such an approach would always favor the largest possible k, resulting in an overfitted, non-interpretable decomposition.

To identify a meaningful latent dimensionality, we adopted a *mask-and-predict* strategy inspired by validation methods in matrix completion and recommender systems [23, 7]. In our implementation, the masking procedure was performed manually: a random subset of the existing links (entries equal to one) in M was set to zero, producing a masked matrix M^{mask}. The matrix factorization was then performed on M^{mask} for a range of k values, while the quality of reconstruction was evaluated exclusively on the masked entries by comparing the binarized approxima-



1	0	0	1	0	0
1	1	0	0	0	0
0	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	1	0	0	0

0	0	0	1	0	0
1	0	0	0	0	0
0	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	1	0	0	0

0.4	1.0	0.7	0.6	0.2	0.2
0.1	0.9	0.6	0.7	0.0	1.0
0.8	0.2	0.2	0.2	0.3	0.5
0.4	0.3	0.6	0.1	0.3	0.4
0.5	0.8	0.2	0.5	0.6	0.0
0.6	0.2	0.1	0.9	1.0	0.8

Figure 2.2: Visual representation of the Mask-and-Predict process

tion $\theta(S_k T_k)$ with the true matrix M.

For the precision evaluation we used the F1-score, which is defined as the harmonic mean of precision and recall, offering a balanced indicator of classification performance when both false positives and false negatives matter. It is especially informative in settings with class imbalance, where accuracy becomes unreliable [30].

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{2.3}$$

The resulting precision displays a clear maximum as a function of k, reflecting the usual bias–variance trade-off: for small k the model underfits, while for large k it overfits to spurious patterns induced by masking. The optimal k thus corresponds to the point of maximum predictive generalization.

We evaluated several matrix factorization methods, including truncated Singular Value Decomposition (SVD), Boolean Matrix Factorization (BMF), and Non-Negative Matrix Factorization (NMF) [15]. BMF proved unsuitable for large, sparse matrices, and SVD produces components with mixed signs, which are not interpretable within a capability framework that assumes non-negativity. NMF, by contrast, respect both constraints, yielding additive and interpretable components consistent with the notion of capabilities as cumulative, non-subtractable resources. Consequently, we selected NMF as our core algorithm.

We also experimented with hybrid methods, such as combining NMF with XGBoost [9] on the residual matrix, but these more complex approaches offered no improvement over the direct application of NMF.

Non-Negative Matrix Factorization Implementation

The matrix factorization step was performed using the `NMF` class from the `scikit-learn` Python library (version 1.5). In this implementation, we approximate the masked data matrix M^{mask} by a low-rank product of two nonnegative matrices,

$$M^{\text{mask}} \approx S\,T,$$

with $S \in \mathbb{R}_{+}^{N\times k}$ and $T \in \mathbb{R}_{+}^{k\times M}$. The parameters S and T are estimated by minimizing the classical Frobenius-norm objective introduced in [24]:

$$\min_{S\geq 0,\, T\geq 0} \left\| M^{\text{mask}} - S\,T \right\|_{F}^{2}. \tag{2.4}$$

Here, M^{mask} denotes the manually masked matrix used for training, i.e. the entries reserved for validation are removed from the loss and treated as missing data.

In our experiments we rely on the coordinate-descent solver implemented in `scikit-learn` (`solver='cd'`), which is based on a Fast Hierarchical Alternating Least Squares (Fast HALS) scheme.[1] Rather than taking explicit gradient steps, Fast HALS performs an alternating minimization of (2.4) over one coordinate (or one component) at a time under nonnegativity constraints.

Concretely, the algorithm repeatedly cycles over the latent components $r = 1, \ldots, k$. For a fixed component r, it updates:

- the r-th column of S (denoted $S_{:,r}$) by solving a one-dimensional nonnegative least-squares subproblem for each row, holding all other columns and the matrix T fixed;
- the r-th row of T (denoted $T_{r,:}$) by an analogous nonnegative least-squares update.

Each such update admits a closed-form expression that can be interpreted as taking an exact minimization step along a single coordinate direction, followed by projection onto $\mathbb{R}_{+}$. In this sense, coordinate descent can be seen as a gradient-based descent method where the step sizes along each coordinate are chosen so as to minimize

[1] See the `scikit-learn` documentation for `non_negative_factorization` and `NMF`, where the `'cd'` solver is described as a coordinate-descent method using Fast HALS.

the objective exactly in that direction, rather than by a generic line search. Fast HALS algorithms of this kind are known to exhibit strong convergence properties and favorable computational efficiency for medium- to large-scale NMF problems [25, 10].

In our implementation, we use the default settings for the Frobenius-loss objective (`beta_loss='frobenius'`) in combination with the coordinate-descent solver. The optimization iterates until one of two stopping criteria is satisfied: (i) the relative decrease in the objective between successive iterations falls below the tolerance `tol=1e-4`, or (ii) the maximum number of iterations `max_iter=1000` is reached.

Because the objective (2.4) is jointly nonconvex in (S, T), the optimization is sensitive to initialization. For all experiments, we initialized the factor matrices with random nonnegative entries (`init='random'`). While deterministic schemes such as NNDSVD [6] can accelerate convergence by providing a good starting point close to a local minimum, random initialization allows the algorithm to explore a broader region of the nonconvex landscape. In the context of our application, we found that this exploration was beneficial for identifying stable, low-dimensional structures that are robust across different random seeds.

The masking procedure was applied manually prior to factorization: entries selected for evaluation were removed from the loss and treated as missing, yielding the training matrix M^{mask}. The NMF optimization thus minimizes (2.4) on the observed (unmasked) entries only. After convergence, the reconstructed matrix

$$\widehat{M} = S\,T$$

is used to impute the masked entries, and performance is evaluated exclusively on this held-out subset.

For each value of the latent dimensionality k, we select a binarization threshold τ_k on $\widehat{M}$ by maximizing the F1-score on the test mask. The optimal dimensionality $k^\star$ and its associated threshold $\tau_{k^\star}$ are then chosen as the configuration achieving the highest mask-and-predict F1-score. This protocol exploits the non-monotonic dependence of the validation F1 on k, in contrast to standard reconstruction metrics such as the unmasked F1-score or RMSE, which typically increase/decrease monotonically with increasing rank and are therefore not informative for model selection.

To verify that the mask-and-predict protocol reliably identifies the intrinsic latent dimensionality, we also applied it to synthetic “toy” matrices. These were generated by first sampling two random nonnegative continuous matrices $S_{\text{toy}} \in \mathbb{R}_+^{N\times k}$ and $T_{\text{toy}} \in \mathbb{R}_+^{k\times M}$, computing their product $X = S_{\text{toy}}T_{\text{toy}}$, and then binarizing X elementwise. The same masking and NMF procedure was then applied to these synthetic matrices.

Across a range of values of k, the maximum of the validation F1 curve consistently coincided with the true generative rank used to construct the data. This confirms that the combination of (i) gradient-based iterative descent (via coordinate descent / Fast HALS) on the Frobenius objective and (ii) the mask-and-predict selection criterion is capable of recovering the underlying latent dimensionality even after a nonlinear binarization step.

Finally, the nonnegativity constraints on both S and T imply that the reconstruction ST is a purely additive superposition of latent components. Each column of S can thus be interpreted as the intensity with which a municipality expresses a given latent "capability", while each row of T describes the association of that capability with specific ATECO sectors. The positivity constraint ensures that capabilities only contribute positively to the reconstruction and never cancel one another, which is fully consistent with the conceptual framework of capabilities in economic complexity. This parts-based, additive representation is precisely the type of interpretable structure that NMF was originally designed to capture [24].

Chapter 3

Inference of the Capabilities Layer

Optimal Dimensionality and Statistical Relevance

To infer the underlying structure of the Municipalities-ATECO bipartite system, we begin by determining its intrinsic latent dimensionality. We adopt the *mask-and-predict* strategy described in the previous chapter, whereby a subset of links is randomly removed from the empirical matrix and subsequently reconstructed through Non-negative Matrix Factorization (NMF).

Figure 3.1 illustrates the model's reconstruction performance as a function of the latent dimensionality k. The F1-score computed on masked entries exhibits a clear, non-monotonic maximum at $k = 5$. At this point, the model successfully reconstructs approximately 80% of the masked positive entries, indicating that five latent factors capture the majority of the informational content embedded in the empirical matrix.

This behaviour suggests the existence of five meaningful, interpretable capability dimensions governing the productive landscape of Italian municipalities. Increasing k beyond this point leads only to marginal increases in reconstruction capacity on the unmasked data as shown by the pannels (c) and (d) in Fig.[3.1.

The comparison with the three null models (BICM [34, 33], Curveball [8], and a purely random matrix preserving only overall density) corroborates this finding. For every value of k, the reconstruction accuracy achieved on the empirical data is consistently and substantially higher than that obtained under any of the null baselines. This demonstrates that the Italian Municipalities–ATECO matrix exhibits non-trivial, structured regularities that are effectively captured by a latent factor representation. This point is further underscored by the behaviour of the null models in panel (b). None of their curves displays a meaningful maximum: BICM and Curveball, which preserve only degree sequences, retain essentially a one-dimensional latent structure (Fitness and Complexity), and therefore reach their highest performance at $k = 1$. The fully random benchmark, by construction, contains no latent structure whatsoever, and accordingly shows no peak across values of k. Traditional performance indicators such as the unmasked F1-score or the RMSE, vary monotonically with k. As a consequence, they do not provide a reliable criterion for selecting an appropriate dimensionality of the latent space.

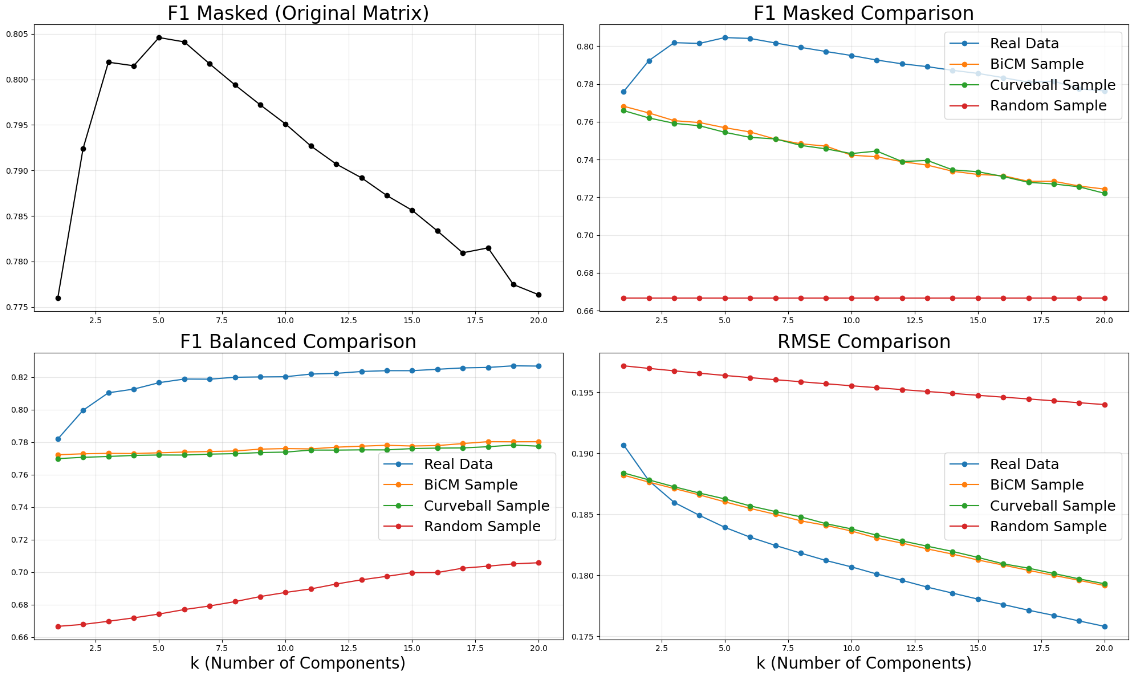


Figure 3.1: Matrix reconstruction performance as a function of latent dimensionality k. Top-left panel shows the F1-score achieved on the masked entries from the original matrix, demonstrating optimal performance at $k = 5$. Top-right panel compares the reconstruction against null models (BICM, Curveball, Random), showing that the original matrix present a more complex structure of the null models' ones. Bottom-left panel presents a balanced validation where equal numbers of ones and zeros are taken to perform the F1 evaluation (we do that to counter the matrix sparsity). Bottom-right panel displays the RMSE for the reconstruction

Analysis of the Optimal Dimensionality and Interpretation of the Latent Capabilities

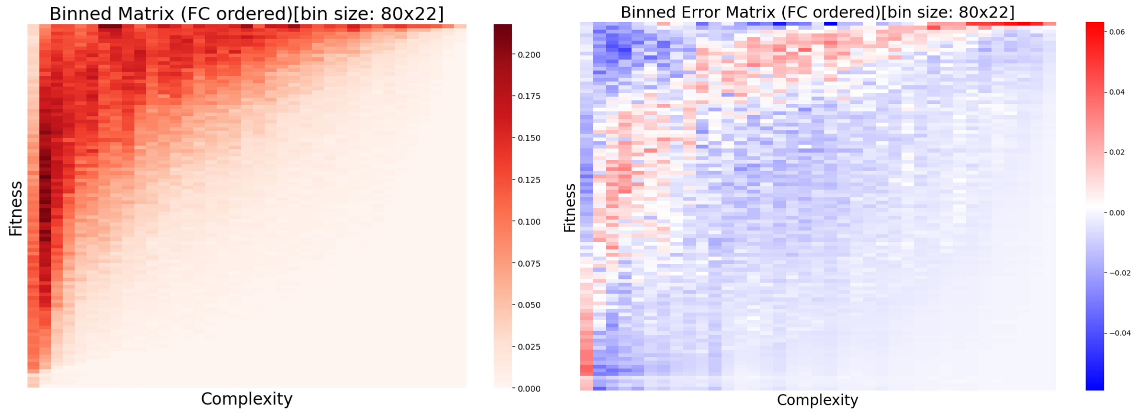
Having established $k = 5$ as the optimal dimensionality, we conduct most of the analyses within this five-dimensional latent space. The NMF model achieves an average F1-score of

$$0.8169 \pm 0.0005,$$

as estimated through bootstrapping, meaning that it correctly reconstructs more than 80% of links and zeros in the empirical matrix.

The reconstruction error is not evenly distributed across the matrix. Figure 3.2 compares the original empirical matrix, ordered by municipal Fitness and sectoral

Complexity, with the residual matrix $R = M - M_{k=5}$. The resulting error map reveals systematic and spatially localized regions of underestimation and overestimation.



(a) Binned original matrix (FC ordered)

(b) Binned error matrix (FC ordered)

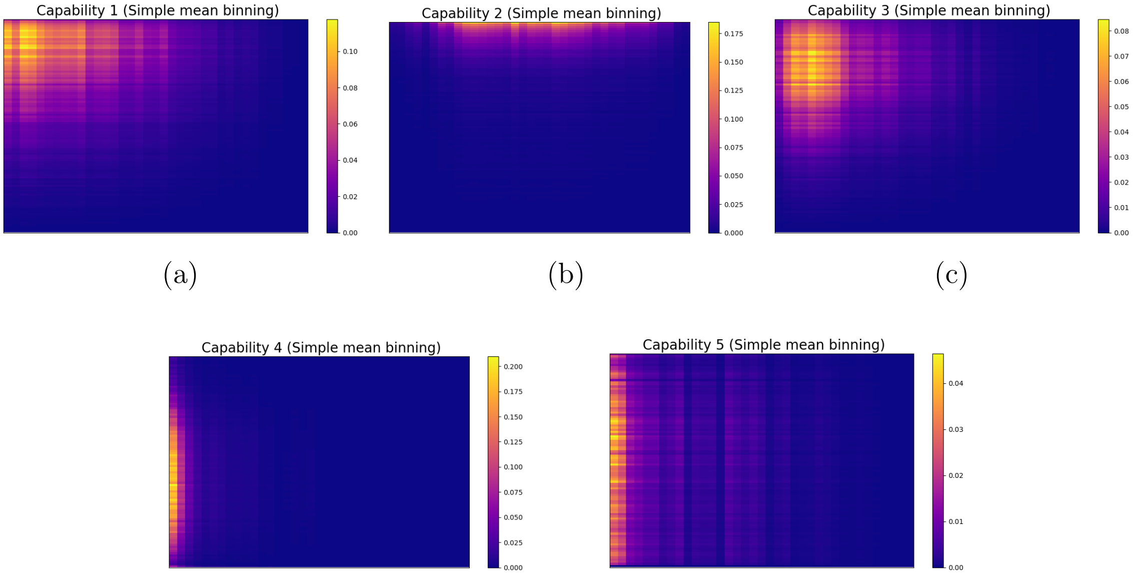
Figure 3.2: Comparison between original and error matrices after binning (bin number: 80×22). The left panel shows the nested structure of the original matrix ordered by fitness (vertical axis) and complexity (horizontal axis). The right panel displays the reconstruction error ($R = M - M_{k=5}$), with red regions indicating underestimation and blue regions indicating overestimation.

These coherent error clusters indicate that the empirical matrix displays nontrivial mesoscale structures. Certain groups of municipalities share patterns of specialization that the factorization captures only partially.

Alignment with the Fitness–Complexity Structure A remarkable feature emerges when the actions of the five latent capabilities are visualized on the matrix reordered by Economic Fitness and Complexity (Figure 3.3). Despite NMF having no access to the ordering used for visualization, each latent component activates primarily within a compact and contiguous region of the reordered matrix.

This spontaneous alignment with the Fitness–Complexity (FC) structure is conceptually striking. It indicates that the latent geometry extracted by the NMF decomposition mirrors the same productive hierarchy that the FC algorithm captures through a completely different, non-linear iterative process. While FC is rooted in mutually reinforcing notions of economic diversification and product ubiquity, NMF reconstructs the matrix by identifying additive, non-negative building blocks.

This observation is particularly remarkable because the NMF reconstruction is, by design, independent of any prior ordering of rows or columns. The algorithm has no knowledge of the EFC-based sorting applied for visualization purposes. The spontaneous emergence of such localized patterns therefore implies that the optimal



(a) (b) (c)

(d) (e)

Figure 3.3: Actions of the five latent capabilities (Simple mean binning).

decomposition identified by NMF naturally aligns with an underlying structure of the same type as that revealed by the FC framework. In other words, the EFC ordering captures a structural backbone of the matrix, one that is also independently recognized by an unsupervised factorization algorithm driven purely by reconstruction efficiency. This correspondence strengthens the interpretation of the EFC approach as a meaningful representation of the productive landscape, grounded not only in network theory but also in latent-space geometry. This drove us to try to see if the five latent components were effectively linked to relevant economic features.

Capability "City"

The first latent capability, ordered according to its magnitude in the municipality of Rome, displays a spatial distribution strongly concentrated in Italy's major urban centers (Figure 3.4). Municipalities with the highest values (Figure 3.5) include Rome, Milan, Turin, Bologna, Florence, and several regional capitals. These are high-population hubs characterized by dense economic, institutional, and infrastructural environments.

The ATECO sectors most strongly associated with this capability (Figure 3.6) provide further insight. They include highly specialized services such as financial consulting, legal and administrative advisory services, and interurban passenger transport. These sectors typically require substantial economies of scale, advanced infrastructure, and access to large pools of specialized human capital.

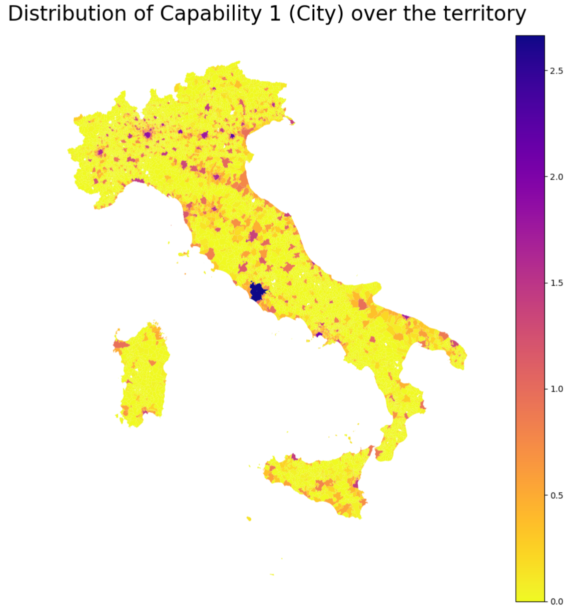
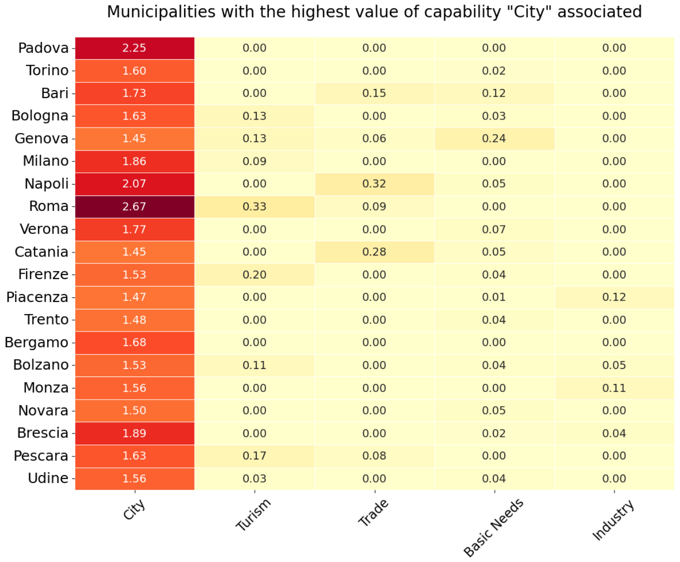

Distribution of Capability 1 (City) over the territory

Figure 3.4


Municipalities with the highest value of capability "City" associated

	City	Turism	Trade	Basic Needs	Industry
Padova	2.25	0.00	0.00	0.00	0.00
Torino	1.60	0.00	0.00	0.02	0.00
Bari	1.73	0.00	0.15	0.12	0.00
Bologna	1.63	0.13	0.00	0.03	0.00
Genova	1.45	0.13	0.06	0.24	0.00
Milano	1.86	0.09	0.00	0.00	0.00
Napoli	2.07	0.00	0.32	0.05	0.00
Roma	2.67	0.33	0.09	0.00	0.00
Verona	1.77	0.00	0.00	0.07	0.00
Catania	1.45	0.00	0.28	0.05	0.00
Firenze	1.53	0.20	0.00	0.04	0.00
Piacenza	1.47	0.00	0.00	0.01	0.12
Trento	1.48	0.00	0.00	0.04	0.00
Bergamo	1.68	0.00	0.00	0.00	0.00
Bolzano	1.53	0.11	0.00	0.04	0.05
Monza	1.56	0.00	0.00	0.00	0.11
Novara	1.50	0.00	0.00	0.05	0.00
Brescia	1.89	0.00	0.00	0.02	0.04
Pescara	1.63	0.17	0.08	0.00	0.00
Udine	1.56	0.03	0.00	0.04	0.00

Figure 3.5: Top twenty Municipality with the highest value of capability "City" associated

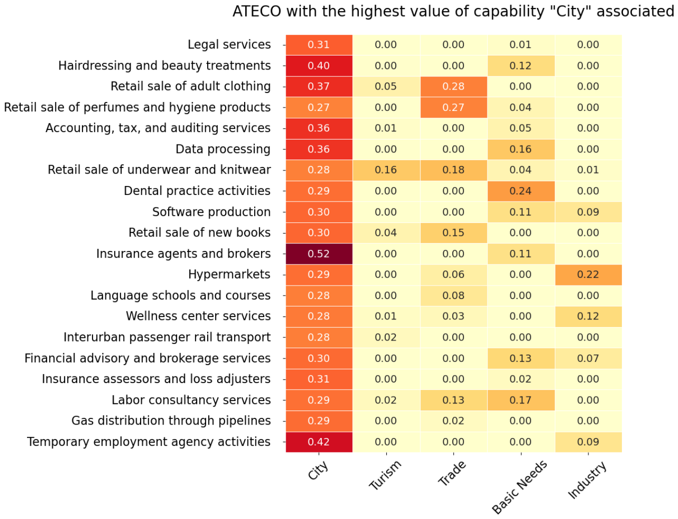


Figure 3.6: Top twenty ATECO with the highest value of capability "City" associated

This capability therefore seems to reflect the economic functions characteristic of a developed urban center. It encapsulates the concentration of high-level services, professional activities, and advanced administrative functions that distinguish cities from smaller municipalities. The relatively strong Spearman correlation with municipal Fitness (0.62, $p < 0.01$) supports this interpretation: both quantities capture the presence of diversified, high-complexity capabilities typical of economically advanced locations.

Interestingly, this capability stands apart from the others: while the remaining four (Industry, Trade, Tourism, Basic Needs) tend to co-occur cumulatively across municipalities, the "City" capability has predominantly negative correlations with them. This suggests that once a municipality has achieved a certain threshold of complexity and size, the presence of the "City" capability becomes a defining feature by itself, with other capabilities contributing only marginally to differentiation among large urban centers.

Capability "Industry"

The second capability exhibits a spatial pattern characteristic of Italy's major manufacturing districts, particularly within the Po Valley and the industrial areas around Turin (Figure 3.7). Unlike the “City” capability, this dimension tends to peak not in the metropolitan cores themselves, but rather in the productive municipalities surrounding them.

This pattern is consistent with Italy's historical industrial geography, where manufacturing activities as mechanical engineering, automotive supply chains, pre-

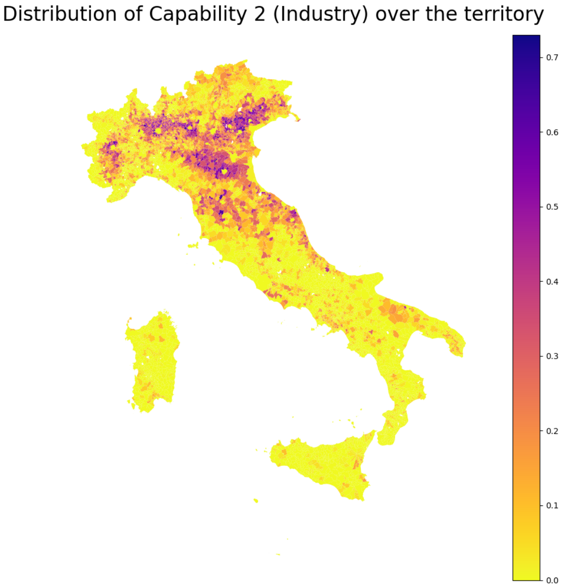


Figure 3.7

cision machinery, and various forms of specialized manufacturing, are concentrated in medium-sized municipalities and industrial clusters rather than in large urban centers. More often in the ring around them.

The municipalities with the highest values (Figure 3.8) correspond to well-established industrial districts, and the top associated ATECO sectors (Figure 3.9) include manufacturing, metalworking, industrial subcontracting, and related activities.

This capability seems to captures the specialization in manufacturing and industrial production. It reflects the presence of supply-chain dense ecosystems, a strong technical workforce, and productive infrastructures typical of industrial districts. Unlike the “City” capability, which is driven by advanced services, the “Industry” capability seems to emerges from a fundamentally different economic logic: production and manufacturing-oriented skills.

Capability "Trade"

The third capability is spatially more diffuse, with notable intensification in southern Italy and coastal regions (Figure 3.10). Unlike the previous two capabilities, its interpretation cannot be derived directly from geographic concentration alone.

A clearer picture emerges from examining the associated ATECO sectors (Figure 3.12). These include wholesale and retail trade mainly but also large commercial distributors, supermarkets, discount stores, and related logistical services. The municipalities with the highest scores (Figure 3.11) similarly tend to host significant

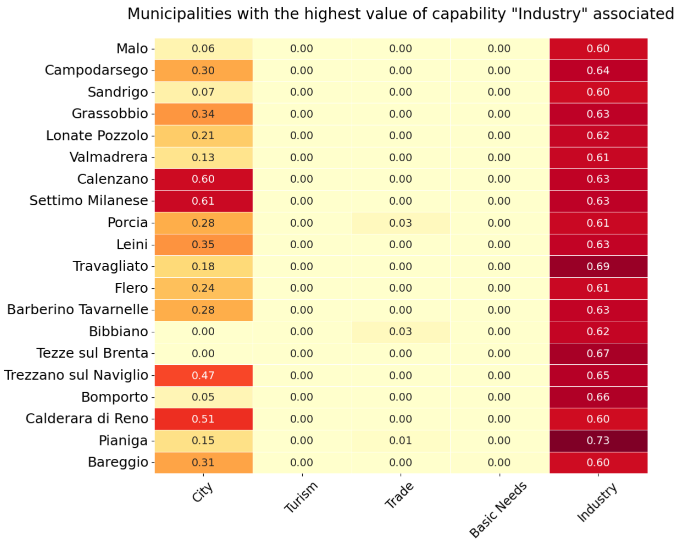


Figure 3.8: Top twenty Municipality with the highest value of capability "Industry" associated

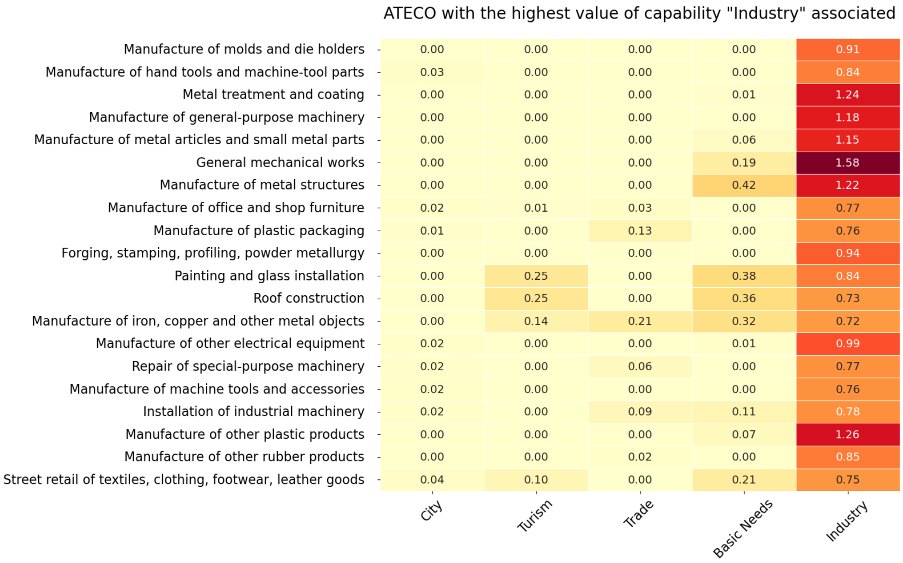


Figure 3.9: Top twenty ATECO with the highest value of capability "Industry" associated

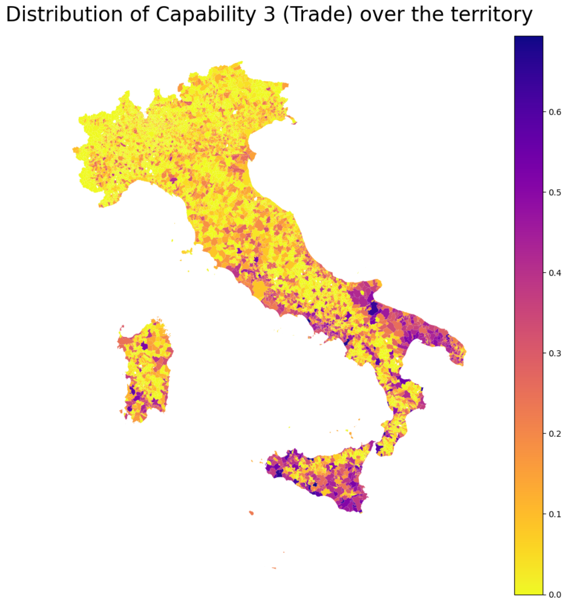

Distribution of Capability 3 (Trade) over the territory

Figure 3.10

retail infrastructures.

This capability appears to capture the commercial and distribution-oriented dimension of local economies. It does not require the dense specialization typical of industrial districts nor the scale effects found in large cities; rather, it reflects the presence of market-oriented activities that serve both local populations and surrounding territories.

Its concentration along coastal areas and in the South may reflect two factors. A reliance on commerce as a foundational economic activity in regions with fewer high-complexity industries, and the role of touristic flows in sustaining large-scale retail and wholesale infrastructures.

Capability "Turism"

The fourth capability presents one of the clearest spatial signatures among all components (Figure 3.13). High values are concentrated in some of Italy's most renowned tourist destinations: the Dolomites, northern Sardinia, the Tuscan coast, and the Adriatic Riviera. The municipalities with the highest values (Figure 3.14) include iconic locations such as Livigno, Jesolo, Campo nell'Elba, and several mountain and seaside resorts.

The associated ATECO sectors (Figure 3.15) consist almost entirely of hospitality-related services: hotels, bed and breakfasts, mountain huts, tourism-related rentals, and similar activities.

This capability therefore represents the touristic intensity of municipalities.

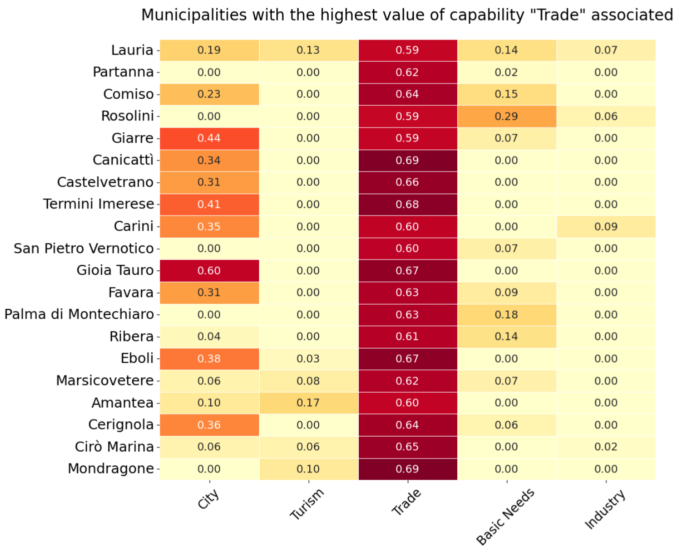


Figure 3.11: Top twenty Municipality with the highest value of capability "Trade" associated

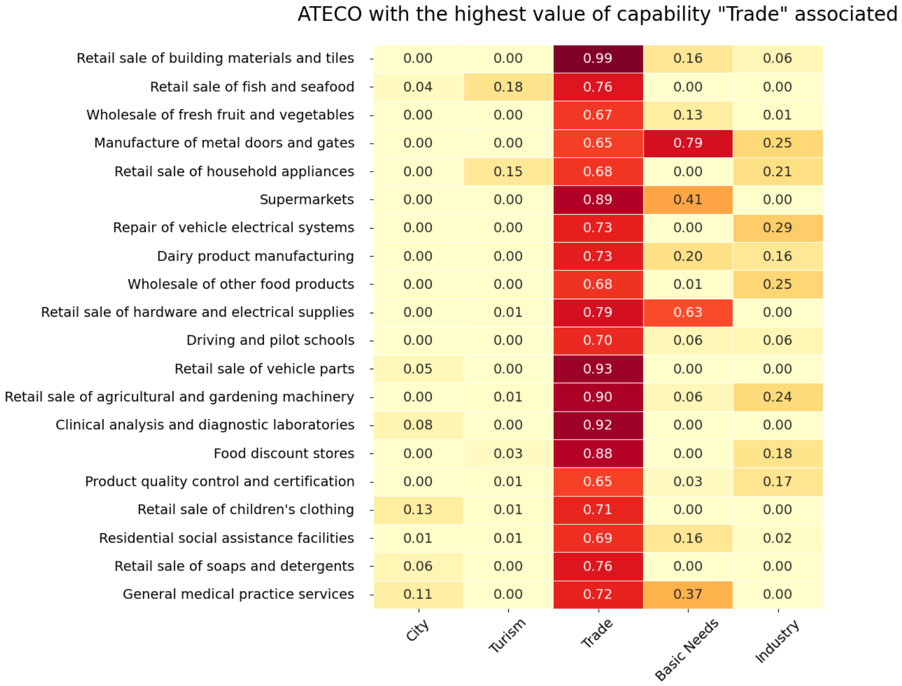


Figure 3.12: Top twenty ATECO with the highest value of capability "Trade" associated

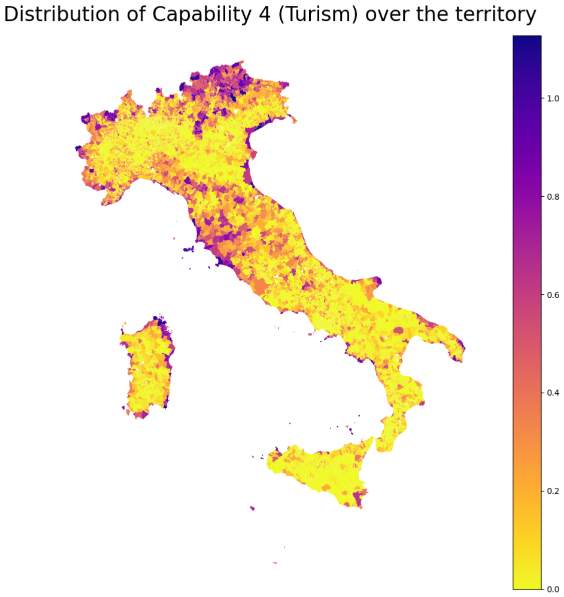


Figure 3.13

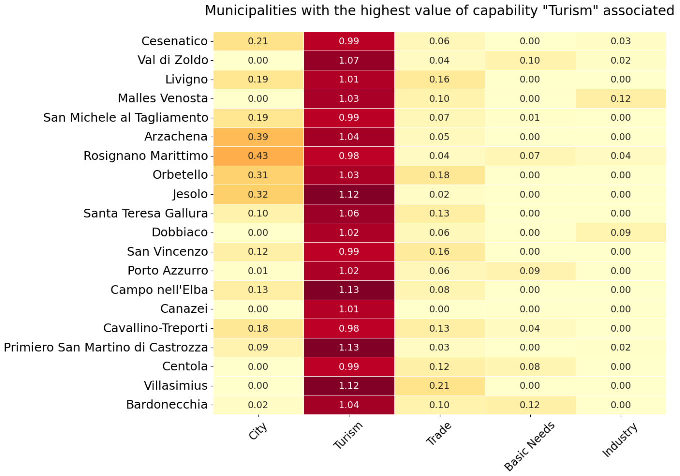


Figure 3.14: Top twenty Municipality with the highest value of capability "Turism" associated

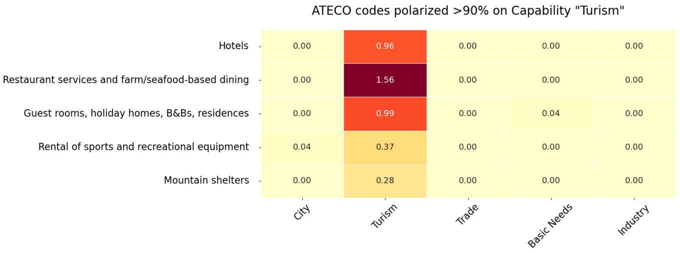


Figure 3.15: ATECO codes polarized >90% on Capability "Turism"

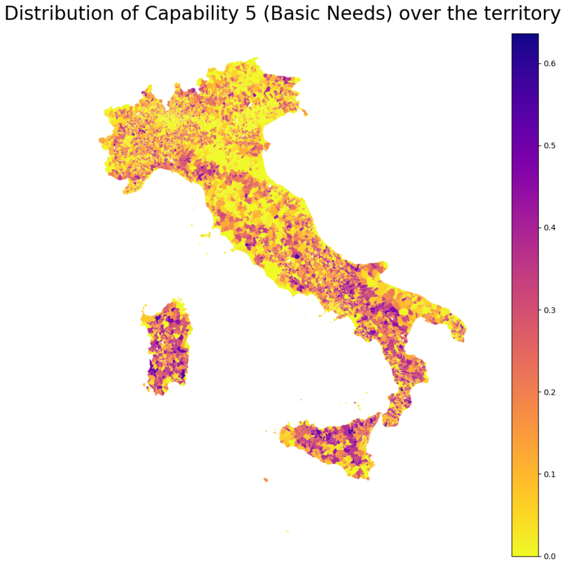

Distribution of Capability 5 (Basic Needs) over the territory
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Figure 3.16

Unlike the “City” or “Industry” capabilities, which capture structural economic transformations, the “Tourism” capability is more sensitive to geographic attractiveness and seasonality. It reflects economic systems where hospitality, accommodation, and leisure constitute a substantial component of local production.

Its spatial pattern also highlights the dual nature of Italy’s tourism economy: alpine tourism in the North and seaside tourism across coastal regions. These two forms of tourism differ in seasonality, visitor demographics, and economic impacts, but both produce a similar latent signature in the data.

Capability "Basic Need"

The fifth and final capability displays a spatial distribution that is almost uniform across the Italian territory (Figure 3.16). Unlike the other components, it does not cluster around specific geographic or economic regions.

The top municipalities (Figure 3.17) do not share any obvious similarity in terms of location, or specialization.

The key to interpreting this capability lies in the associated ATECO sectors (Figure 3.18). These include universal services such as postal offices, basic logistics, and essential urban functions that are present even in the smallest or most rural municipalities. Such activities represent the foundational layer upon which all other capabilities can build.

This capability thus reflects the baseline functional infrastructure required for municipal operation. It is not a marker of specialization, but rather a measure of the degree to which municipalities host the minimal set of services enabling social

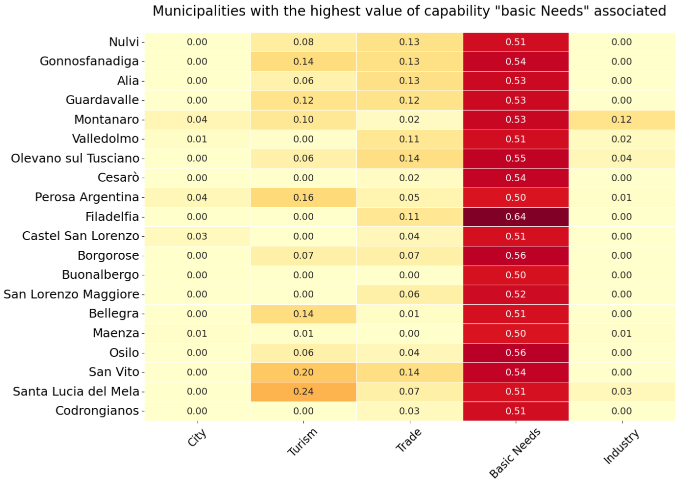


Figure 3.17: Top twenty Municipality with the highest value of capability "Basic Needs" associated

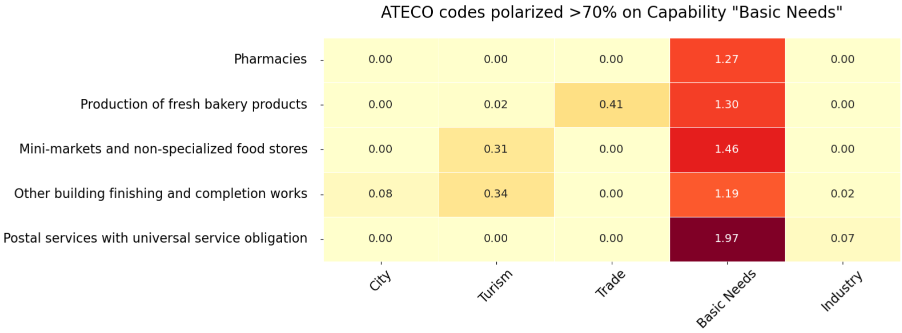


Figure 3.18: ATECO codes polarized >70 % on Capability "Basic Needs"

and economic life.

Its negative correlation with the “City” capability suggests that large urban centers rely on more advanced or differentiated forms of basic services, while smaller municipalities rely more heavily on the standard set of universal functions captured by this capability. Conversely, its positive correlations with the “Industry”, “Trade”, and “Tourism” capabilities sustain the cumulative nature of capabilities, giving "Basic Needs" as ground state.

3.1 Metrics for k different from five

In the previous sections, we have focused primarily on the case $k = 5$, as this choice emerged as a natural compromise between parsimony and predictive accuracy in the mask-and-predict experiments. Nonetheless, in order to better understand how the Non-Negative Matrix Factorization (NMF) behaves when we vary the dimensionality of the latent capability space, it is instructive to study in a systematic way how the reconstruction quality evolves as a function of k.

To this end, we considered a set of NMF decompositions of the Municipality-ATECO matrix for a wide range of values of k. For each value of k we evaluated three complementary metrics: *Total F1-score*, *Balanced F1-score* and *Root Mean Squared Error*.

The first two metrics are sensitive to the classification performance on the binary structure of the matrix, with the balanced version explicitly correcting for class imbalance, while the RMSE captures the overall goodness of fit of the continuous reconstruction.

Figure 3.19 reports the behavior of these three metrics as functions of k, together with their increments with respect to $k - 1$. The plots reveal the presence of distinct regimes. For small values of k, and in particular up to $k \approx 5$, the increase in performance is substantial: both the Total F1 and, even more clearly, the Balanced F1 grow rapidly as we add degrees of freedom in the hidden layer. In this region, the gains for the empirical Municipality-ATECO matrix are much larger than those obtained for the corresponding null/random models, which are also depicted in the figure for comparison.

The Balanced F1-score is especially informative in this context. Due to the high sparsity of the matrix, the Total F1 is inevitably more sensitive to the overwhelming prevalence of zeros and therefore to noise in the reconstruction of non-links. The Balanced F1, by construction, compares the performance on links with that on a matched sample of zeros and thus provides a clearer picture of how the

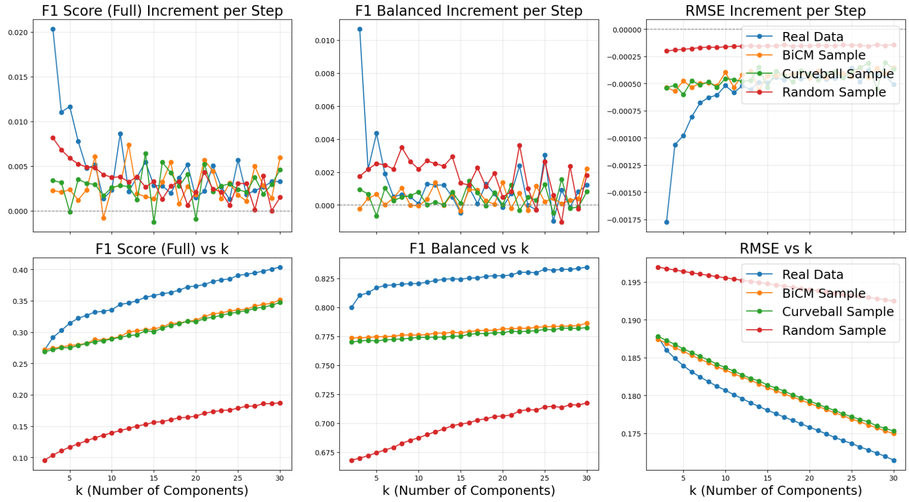


Figure 3.19: **Metrics evaluation and corresponding increments**: *Total F1-score* computed on all entries of the matrix, treating the factorization-based reconstruction as a binary classifier on links and non-links, *Balanced F1-score* computed on all observed links (entries equal to one) and on an equal number of randomly sampled zeros, in order to compensate the strong sparsity of the matrix, *Root Mean Squared Error* (RMSE), computed on all entries of the matrix

algorithm is learning the non-trivial backbone of the bipartite structure.

Up to $k = 5$, the increments of the Balanced F1 are noticeably larger than those of the null models, signaling that each additional capability dimension is used to capture genuinely informative patterns in the data, rather than simply overfitting specific entries.

Beyond $k = 5$, the situation changes qualitatively. As shown by the incremental curves in Figure 3.19, the additional improvement in all metrics becomes much smaller and, in the case of the Balanced F1, comparable in magnitude to the improvements observed in the null/random models. In other words, for $k > 5$ the factorization still leads to a mild increase in predictive accuracy, but this increase is similar to what would be obtained by adding degrees of freedom in matrices that do not contain the specific structural information of the Municipality–ATECO network. This seems to suggests that, beyond $k = 5$, the extra dimensions are mainly used to fine-tune the reconstruction and to “fix” noisy or marginal aspects of the matrix, rather than to extract new, robust information about its latent capability structure.

This interpretation is reinforced by comparison with the distribution of metrics obtained from the null models. Figure 3.20 displays, for selected values of k, the empirical values of the F1 and RMSE alongside the distributions generated by randomizations of the bipartite matrix. The Municipality-ATECO matrix lies well outside the bulk of the null-model distributions, especially in the region around

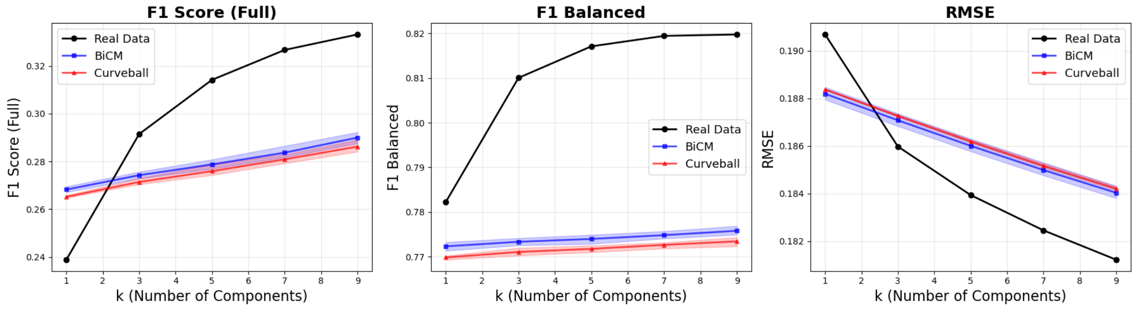


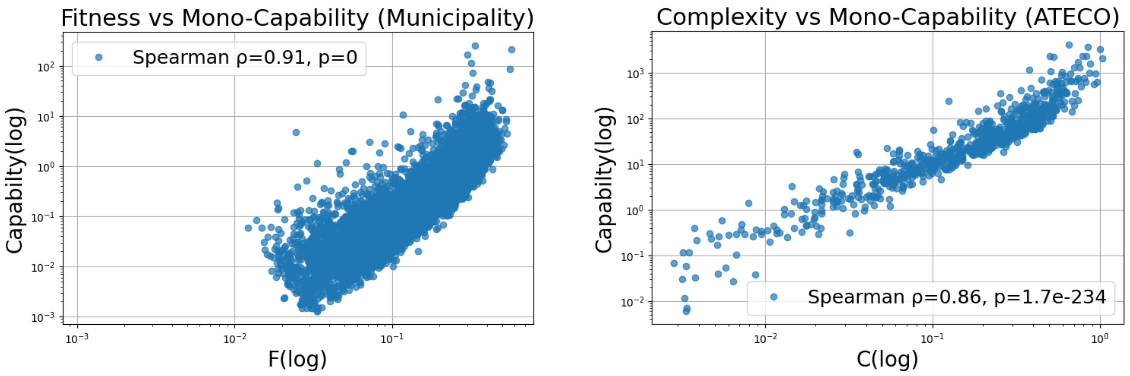
Figure 3.20: Comparison of various metrics (F1 score of the full matrix, F1 score balanced with equal number of ones and zeros, RMSE) between the Municipality-ATECO reconstructed matrix and the Null models ensambles (BiCM and Curveball) at different values of k. The blue and red intervals represent the 95% of the Null models' metrics values distributions.

$k = 5$, confirming that the observed performance is not a trivial by-product of the marginal properties of the matrix (e.g. degree sequences) but reflects statistically significant structure. At larger k, although the empirical metrics remain better than those of the null models, the relative advantage becomes less pronounced, again consistent with the picture of diminishing returns in terms of new structural information. Overall, these results support the idea that the main nested backbone of the Municipality–ATECO network can be effectively captured by a relatively low number of latent factors. In our case, $k = 5$ emerges as a particularly meaningful choice: it marks a clear change in regime in the performance curves, it is the value suggested by the mask-and-predict analysis, and it sits in a region where the empirical improvement is still clearly above that of the null models. Larger values of k yield only incremental refinements, compatible with a noise-correction role of the additional dimensions.

Connection between NMF ($k = 1$) and EFC

An additional insight into the meaning of the latent factors obtained via NMF comes from comparing the case $k = 1$ with the Economic Fitness and Complexity (EFC) algorithm introduced in the economic complexity literature [42, 43, 36]. In the previous section we have already discussed the strong Spearman correlation observed between one of the inferred capabilities (the "City" capability) and the Fitness metric computed on the same Municipality-ATECO matrix. Here we push this observation one step further by aligning the degrees of freedom of the two methods.

When we apply NMF with $k = 1$, the factorization reduces to the search for a single non-negative vector for municipalities and a single non-negative vector for ATECO codes which, when multiplied together, best approximate the original matrix in the least-squares sense. This yields a rank-one reconstruction, and the municipality and ATECO vectors can be interpreted as one-dimensional summaries of their position in the bipartite network.



(a) (b)

Figure 3.21: Dot plots showing the strong link between Fitness-Complexity ranking and the ranking coming from the single-capability decomposition made by the NMF algorithm.

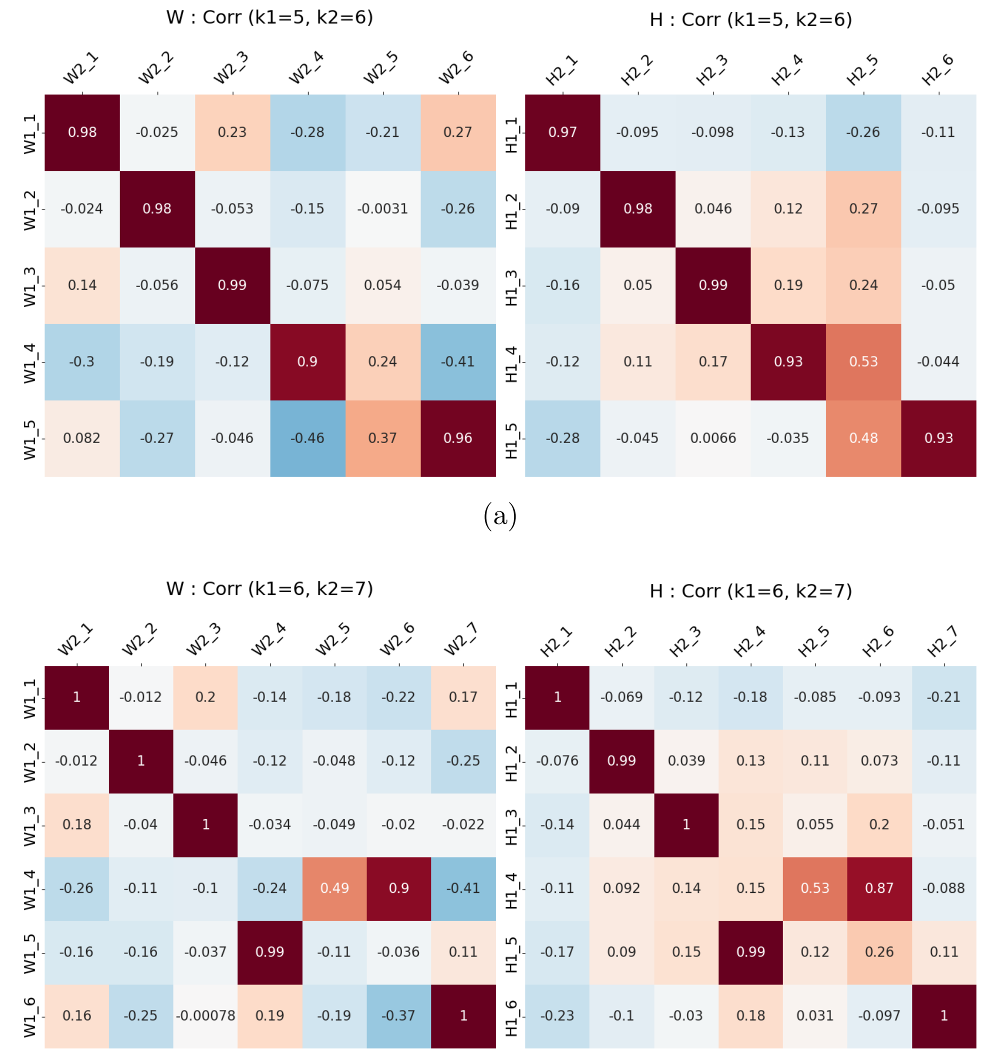
Figure 3.21 compares these NMF-derived vectors with the Fitness and Complexity, respectively. Panel (a) shows a scatter plot of the municipality scores, where each point represents a municipality with coordinates given by its NMF score and its Fitness. The two rankings are extremely similar: the Spearman correlation between the two vectors is approximately 0.91. Panel (b) displays the analogous comparison for ATECO codes, where the NMF product scores are plotted against the Complexity metric; again, the Spearman correlation is very high, around 0.86.

Such strong correlations indicate that, in the one-dimensional case, the NMF solution is essentially aligned with the EFC solution: the optimization problem solved by rank-one NMF identifies a direction in the municipality and product spaces that coincides, to a very good approximation, with the ranking induced by Fitness and Complexity. This observation is in line with recent work [26] showing deep connections between EFC-type iterative algorithms and matrix-scaling or optimization procedures. From our perspective, this result provides an additional validation of the EFC approach on this dataset and suggests that the NMF capabilities could be understood as higher-dimensional generalizations of the Fitness-Complexity ranking.

The persistence of a strong correlation between at least one of the $k = 5$ capabilities and the Fitness vector, discussed in the previous section, further supports this interpretation. It indicates that, when we increase k, the model does not discard the global Fitness-like pattern; rather, it refines it by adding orthogonal components that capture more specific dimensions of municipal and sectoral specialization.

Evolution of capabilities when varying k

To better understand how the latent capabilities evolve as we change the number of factors k, we analyzed the correlations between the columns of the municipality factor matrices obtained for consecutive values of k. More precisely, for each pair $(k, k+1)$ we computed the correlation matrix between the k capabilities at level k and the $k+1$ capabilities at level $k+1$, after aligning them in a consistent way. The resulting correlation patterns are summarized in Figure 3.22.



(a)

(b)

Figure 3.22: Correlation tables of the reconstructed matrices Municipality-Capabilities (W) and Capabilities-ATECO (H) at different values of k (number of degree of freedom in the capability layer)

Two main behaviors emerge from this analysis. In some cases, when we move

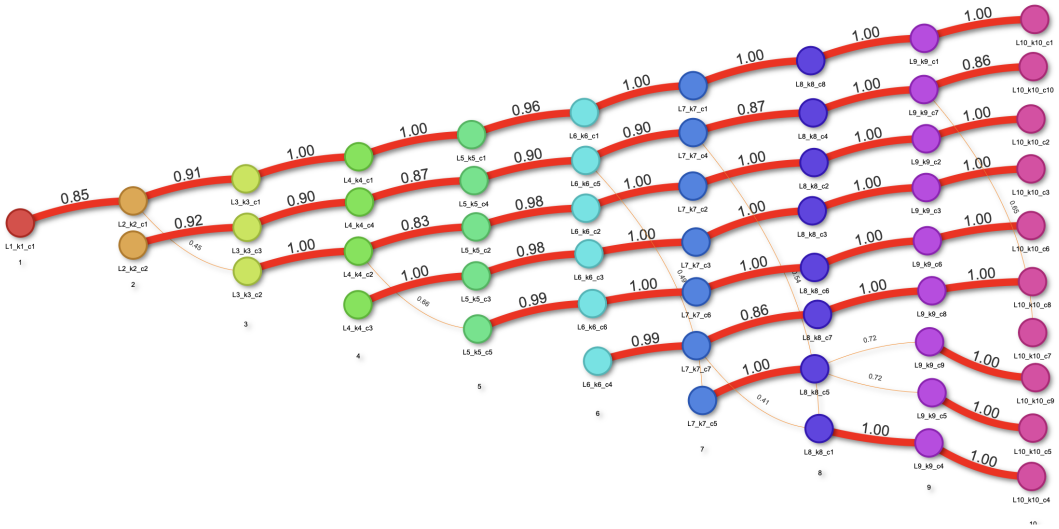


Figure 3.23: Correlation between degrees of freedom at varying number of k in the application of NMF to Municipality-ATECO matrix (Orange links: from 0.4 to 0.75, Red links: >0.75)

from k to $k+1$, one of the existing capabilities “splits” into two more specialized components. A clear example of this phenomenon occurs in the transition from $k = 6$ to $k = 7$ in the Fig[3.22b]: one of the vectors at $k = 6$ shows high correlation with two distinct vectors at $k = 7$, indicating that the algorithm is partitioning a previously broad capability into two sub-capabilities with more differentiated sectoral and spatial profiles. This is consistent with the idea that increasing k allows the model to resolve finer-grained patterns that were previously aggregated.

In other cases, the additional dimension corresponds to the emergence of a new capability as a combination of some of the previous ones. This happens, for instance, in the transition from $k = 5$ to $k = 6$ in 3.22a, where the new capability displays notable correlation with more capabilities at $k = 6$. In such situations, the added factor captures a qualitatively new direction in the space of municipal economic profiles, possibly associated with a relatively small but coherent cluster of municipalities and ATECO codes that was previously only partially represented.

Importantly, with the exception of at most one capability at each step, the majority of capabilities remain strongly correlated with their counterparts at the preceding value of k. This means that the core structure of the factorization is remarkably stable when we increase k. The new degrees of freedom typically refine or complement the existing ones, rather than radically reorganizing the entire capability space as shown in Fig.[3.23].

This stability explains why a Fitness-like capability persists even when we move from $k = 1$ to $k = 5$: the main global pattern is preserved, while additional dimensions capture deviations and specializations around it.

Taken together, the analysis of the performance metrics as functions of k, the comparison with null models, the strong alignment between NMF and the Fitness–Complexity algorithm for $k = 1$, and the stability of capabilities across consecutive values of k convey a coherent message: the Municipality-ATECO matrix appears to be governed by a relatively low-dimensional latent structure, whose essential features are already well captured by $k \approx 5$ capabilities. Beyond this point, additional factors play a secondary role, mainly associated with noise correction and finer segmentation of existing patterns, rather than with fundamentally new dimensions of economic heterogeneity.

Conclusions

This thesis set out to reconstruct the latent capability architecture underlying the productive activities of Italian municipalities, drawing on the conceptual framework of Economic Complexity and employing Non-Negative Matrix Factorization as the central analytical tool. Through a combination of rigorous validation procedures, comparisons with null models, and interpretation of the resulting latent factors, the research demonstrates that the Municipality–ATECO system possesses a remarkably strong and interpretable low-dimensional structure. This structure can be effectively summarized by five latent capabilities, which emerge as the optimal compromise between predictive accuracy, interpretability, and robustness. A central achievement of this work is to show that these capabilities are not arbitrary statistical artifacts, but correspond to recognizably distinct economic domains. Their spatial and sectoral expressions reveal coherent patterns consistent with Italy's economic geography: the concentration of advanced services in major urban centers; the industrial specialization of medium-sized northern municipalities; the pervasive commercial functions distributed across the territory; the dual model of alpine and coastal tourism; and the foundational layer of essential services that supports even the smallest localities. These findings confirm that matrix factorization techniques can indeed extract meaningful latent economic dimensions from empirical data, lending empirical substance to the long-theorized but rarely measured concept of capabilities. Beyond revealing the internal structure of the Italian productive system, the thesis uncovers a deep and unexpected connection between NMF-based factorization and the Fitness–Complexity approach. At the one-dimensional level, NMF reproduces, with striking fidelity, the municipality and sector rankings obtained through the nonlinear EFC algorithm. As the dimensionality increases, the alignment persists: several latent components activate preferentially along the same nested contours revealed by Fitness and Complexity, despite NMF having no access to the ordering used for visualization. This convergence suggests that both methods are detecting the same underlying geometric organization, one that governs the distribution of activities across municipalities and shapes the hierarchy of production possibilities. The behavior of the model when k increases above five reinforces the interpretation of $k = 5$ as the intrinsic dimensionality of the system. While additional factors continue to provide incremental improvements in reconstruction accuracy, these improvements are comparable to those obtained in randomized matrices. This seems to indicate that higher-dimensional factorizations mainly correct noise or capture minor local variations rather than unveiling new structural properties. Correlation analyses across consecutive values of k show that the core latent dimensions are stable and persistent, while new dimensions generally arise either by splitting existing capabilities into more specialized subcomponents or by capturing small-scale structural deviations. The essential structure of the capability space is therefore robust and efficiently represented with five components. While the reconstruction of latent capabilities already provides a compact and interpretable representation of the Italian productive system, its potential extends well beyond the present analysis. The methodological framework developed in this thesis can be naturally integrated (and tested) with additional layers of empirical information, enabling a richer exploration of how capabilities interact with geographical constraints and production viability. In particular, the inferred latent capabilities can be combined with viability measures, such as those capturing the minimum set of enabling conditions required for municipalities to sustain specific economic activities. Integrating viability with capability inference would make it possible to distinguish between activities that are absent because the underlying capabilities are lacking and those that are structurally unviable due to environmental, demographic, or infrastructural constraints. This distinction would provide a more nuanced understanding of how opportunity spaces vary across the territory and how municipalities differ in their potential for productive upgrading. Moreover, the methodology opens the door to a systematic analysis of geographical correlations between capabilities. By embedding municipalities in physical or functional space, through adjacency matrices and mobility networks, one could examine whether capabilities exhibit spatial autocorrelation, whether they diffuse along infrastructural corridors, or whether they cluster according to historical or institutional boundaries. Understanding these spatial interdependencies would shed light on the mechanisms through which capabilities propagate across regions, possibly revealing patterns of spatial spillovers, complementarities, or lock-ins. This line of inquiry would transform the latent capability space into a genuinely spatially grounded object, connecting structural economic heterogeneity with territorial dynamics. In this sense, the presented work represents a foundational step. It demonstrates that a low-dimensional capability space can be empirically reconstructed and meaningful links with geo-economic features can found. Future research may now build on this foundation to explore how capabilities interact with viability constraints, trade structures, and geographical proximities, ultimately contributing to a deeper understanding of the spatial organization of productive knowledge.

Bibliography

[1] Aamena Alshamsi, Flávio L Pinheiro, and Cesar A Hidalgo. Optimal diversification strategies in the networks of related products and of related research areas. *Nature communications*, 9(1):1328, 2018.

[2] Lorenzo Arsini, Matteo Straccamore, and Andrea Zaccaria. Prediction and visualization of mergers and acquisitions using economic complexity. *Plos one*, 18(4):e0283217, 2023.

[3] Bela Balassa. Trade liberalisation and “revealed” comparative advantage 1. *The manchester school*, 33(2):99–123, 1965.

[4] Pierre-Alexandre Balland, Ron Boschma, Joan Crespo, and David L Rigby. Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional studies*, 53(9):1252–1268, 2019.

[5] Federico Battiston, Matthieu Cristelli, Andrea Tacchella, Luciano Pietronero, et al. How metrics for economic complexity are affected by noise. *Complexity Economics*, 2014.

[6] Christos Boutsidis and Efstratios Gallopoulos. Svd-based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[7] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[8] Corrie Jacobien Carstens, Annabell Berger, and Giovanni Strona. A unifying framework for fast randomization of ecological networks with fixed (node) degrees. *MethodsX*, 5:773–780, 2018.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[11] Giulio Cimini, Andrea Gabrielli, and Francesco Sylos Labini. The scientific competitiveness of nations. *PloS one*, 9(12):e113470, 2014.

[12] Matthieu Cristelli, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one*, 8(8):e70726, 2013.

[13] Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. The heterogeneous dynamics of economic complexity. *PloS one*, 10(2):e0117174, 2015.

[14] Matthieu Claudio Ascagne Cristelli, Andrea Tacchella, Masud Z Cader, Kirstin Ingrid Roster, and Luciano Pietronero. On the predictability of growth. *World Bank Policy Research Working Paper*, (8117), 2017.

[15] Derek DeSantis, Erik Skau, Duc P Truong, and Boian Alexandrov. Factorization of binary matrices: Rank relations, uniqueness and model selection of boolean decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–24, 2022.

[16] Eurostat. *NACE Rev. 2: Statistical Classification of Economic Activities in the European Community*. Publications Office of the European Union, Luxembourg, 2008. ISBN 978-92-79-04741-1.

[17] Massimiliano Fessina, Giambattista Albora, Andrea Tacchella, and Andrea Zaccaria. Identifying key products to trigger new exports: an explainable machine learning approach. *Journal of Physics: Complexity*, 5(2):025003, 2024.

[18] Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12:257–291, 2014.

[19] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[20] César A Hidalgo, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[21] ISTAT – Istituto Nazionale di Statistica. *Classificazione delle Attività Economiche ATECO 2007*. ISTAT, Rome, Italy, 2009. ISBN 978-88-458-1521-3.

[22] Dieter F Kogler, David L Rigby, and Isaac Tucker. Mapping knowledge space and technological relatedness in us cities. *European planning studies*, 21(9):1374–1391, 2013.

[23] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[24] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[25] Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, 2007.

[26] Dario Mazzilli, Manuel Sebastian Mariani, Flaviano Morone, and Aurelio Patelli. Equivalence between the fitness-complexity and the sinkhorn-knopp algorithms. *Journal of Physics: Complexity*, 5(1):015010, 2024.

[27] Frank Neffke and Martin Henning. Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3):297–316, 2013.

[28] Frank Neffke, Martin Henning, and Ron Boschma. How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic geography*, 87(3):237–265, 2011.

[29] Aurelio Patelli, Luciano Pietronero, and Andrea Zaccaria. Integrated database for economic complexity. *Scientific Data*, 9(1):628, 2022.

[30] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[31] Emanuele Pugliese, Giulio Cimini, Aurelio Patelli, Andrea Zaccaria, Luciano Pietronero, and Andrea Gabrielli. Unfolding the innovation system for the development of countries: coevolution of science, technology and production. *Scientific reports*, 9(1):16440, 2019.

[32] Emanuele Pugliese, Andrea Zaccaria, and Luciano Pietronero. On the convergence of the fitness-complexity algorithm. *The European Physical Journal Special Topics*, 225(10):1893–1911, 2016.

[33] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific reports*, 5(1):10595, 2015.

[34] Fabio Saracco, Mika J Straka, Riccardo Di Clemente, Andrea Gabrielli, Guido Caldarelli, and Tiziano Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, 19(5):053022, 2017.

[35] Vito DP Servedio, Alessandro Bellina, Emanuele Calò, and Giordano De Marzo. Fitness centrality: a non-linear centrality measure for complex networks. *Journal of Physics: Complexity*, 6(1):015002, 2025.

[36] Vito DP Servedio, Paolo Buttà, Dario Mazzilli, Andrea Tacchella, and Luciano Pietronero. A new and stable estimation method of country economic fitness and product complexity. *Entropy*, 20(10):783, 2018.

[37] Robert M Solow. Growth theory and after. *The American economic review*, 78(3):307–317, 1988.

[38] Matteo Straccamore, Matteo Bruno, Bernardo Monechi, and Vittorio Loreto. Urban economic fitness and complexity from patent data. *Scientific Reports*, 13(1):3655, 2023.

[39] Matteo Straccamore, Matteo Bruno, and Andrea Tacchella. Comparative analysis of technological fitness and coherence at different geographical scales. *PLoS One*, 20(8):e0329746, 2025.

[40] Matteo Straccamore, Vittorio Loreto, and Pietro Gravino. The geography of technological innovation dynamics. *Scientific Reports*, 13(1):21043, 2023.

[41] Matteo Straccamore, Luciano Pietronero, and Andrea Zaccaria. Which will be your firm's next technology? comparison between machine learning and network-based algorithms. *Journal of Physics: Complexity*, 3(3):035002, 2022.

[42] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2(1):723, 2012.

[43] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. Economic complexity: conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control*, 37(8):1683–1691, 2013.

[44] Andrea Tacchella, Andrea Zaccaria, Marco Miccheli, and Luciano Pietronero. Relatedness in the era of machine learning. *Chaos, Solitons & Fractals*, 176:114071, 2023.

[45] Nicolò Vallarano, Matteo Bruno, Emiliano Marchese, Giuseppe Trapani, Fabio Saracco, Giulio Cimini, Mario Zanon, and Tiziano Squartini. Fast and scalable likelihood maximization for exponential random graph models with local constraints. *Scientific Reports*, 11(1):15227, 2021.

[46] Andrea Zaccaria, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. How the taxonomy of products drives the economic development of countries. *PloS one*, 9(12):e113770, 2014.