Universitas Mercatorum

Università telematica delle Camere di Commercio Italiane

PhD Programme in

Big Data and Artificial Intelligence

Curriculum "Big data management for the digital transition"

38th Cycle

PhD Dissertation in

Measuring and Modelling the Spatial Patterns of Firms: Integrating Spatial Statistics and Machine Learning for Firm-Level Analysis

Dr. Alessio Bumbea

DT00100003

Programme Coordinator
Prof. Barbara Martini

Co-Supervisor
Dr. Alessandro Rinaldi

Supervisor
Prof. Andrea Mazzitelli

External Supervisor
Dr. Emanuele Pugliese

Academic Year 2024 / 2025

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani Piano Nazionale di Ripresa e Resilienza | Universitas Mercatorum Università telematica delle Camere di Commercio Italiane

To my family,
for supporting this long and
expensive hobby.

Abstract

This thesis is developed as a collection of three independent chapters/papers and studies how productivity differences emerge and persist across firms and places, with a focus on industrial clustering and agglomeration. The first paper develops an empirical framework based on bipartite network representations of firms to characterize local productive structures of innovative startups in Lombardy. The second paper develops a deep clustering pipeline to perform bootstrap analysis of high-tech firms in Lombardy. The third paper links micro-level firm information to meso- and macro-level patterns of specialisation, the analysis identifies regularities in diversification and analyses their impact on the labour productivity.

Keywords: economic statistics, spatial statistics, GeoAI, spatial bootstrapping, spatial machine learning, economic complexity, bipartite networks, firm dynamics, deep clustering, firm productivity.

Contents

List of Figures

List of Tables

Nomenclature

β_k	Regression coefficient
$\gamma_t^{(r)}$	Year fixed effect (hexagon regressions at resolution r)
$\gamma_{t,s}$	Year-by-sector fixed effect
$\mathbb{1}_{hf}^{(r)}$	Indicator: firm f belongs to hexagon h at resolution r
$\mathbb{1}_{hp}^{(r)}$	Indicator: at least one firm in hexagon h exports product p
$\tilde{E}_{hg}^{(r)}$	Employees in education group g aggregated in hexagon h
$\tilde{L}_h^{(r)}$	Aggregated employment in hexagon h at resolution r
$\tilde{X}_h^{(r)}$	Hexagon-level aggregate of X at resolution r
$\tilde{Y}_h^{(r)}$	Aggregated added value in hexagon h at resolution r
$\varepsilon_{f,t}$	Error term (firm regression)
$\varepsilon_{h,t}^{(r)}$	Error term (hexagon regression)
$\widetilde{LP}_h^{(r)}$	Labour productivity of hexagon h at resolution r
$A_{i,j}$	Bipartite adjacency matrix (membership of point i in cluster j)
B	Number of bootstrap replicas
C	Cluster ensemble (set of base partitions)
C^r	r-th clustering solution in the ensemble
D_{KL}	Kullback–Leibler divergence
E_{gf}	Employees of firm f in education group g
F	Feature-importance score used by XGBoost
f	Firm index
f_c	Converged country Fitness

$F_c^{(n)}$ Country Fitness at iteration n

$F_h^{(r)}$ Exogenous Fitness of hexagon h at resolution r

$f_{hg}^{(r)}$ Share of education group g in hexagon h at resolution r

g Education group / field index

h Hexagon index

$H_h^{(r)}$ Entropy of education-group distribution in hexagon h at resolution r

I Moran's I (global spatial autocorrelation statistic)

K Number of clusters

K_r Number of clusters in C^r

K_{tot} Total number of clusters across all base solutions

M_{cp} Binary country–product matrix after RCA thresholding

N Number of observations

n Number of firms/observations in the sample

$N(r)$ Number of hexagons at resolution r

N_f Number of firms

Q_p Converged product Complexity

$Q_p^{(n)}$ Product Complexity at iteration n

q_{cp} Exports of country c in product p (monetary flows)

R Number of base clustering solutions

r Correlation coefficient

r H3 resolution level

RCA_{cp} Revealed Comparative Advantage of country c in product p

$S_h^{(r)}$ Scolarisation index (share above secondary education) in hexagon h

X Dataset

X_f Generic firm-level variable

X_i i-th observation in X

Acronyms

AIDA Italian company information and business intelligence database

ASIA Statistical Archive of Active Businesses

ATECO Italian classification aligned with NACE

biLouvain Bipartite Louvain community detection

BIRCH balanced iterative reducing and clustering using hierarchies

BR Business Register

COVID-19 Coronavirus Disease 2019

CREF Enrico Fermi Research Center

CSH Complexity Science Hub Vienna

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DEC Deep Embedded Clustering

ECFG Ensemble Clustering using Factor Graph

ECI Economic Complexity Index

EM Expectation Maximization

EU European Union

Extra Trees Extremely Randomized Trees

FCS Fully Conditional Specification

FE Fixed Effects

GAN Generative Adversarial Network

GDP Gross Domestic Product

GeoAI Geospatial Artificial Intelligence

GIS Geographic Information System

GTB Gradient Tree Boosting

H3 Hexagonal Hierarchical Geospatial Indexing System

HBGF Hybrid Bipartite Graph Formulation

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

ICT Information and Communication Technology

INET Institute for New Economic Thinking

IP Intellectual Property

Istat Italian National Institute of Statistics

IT Information Technology

JM Joint Modeling

KL Kullback–Leibler

KNN K-Nearest Neighbors

KPCA Kernel Principal Component Analysis

LISA Local Indicators of Spatial Association

LP Labour Productivity

MAR Missing At Random

MAUP Modifiable Areal Unit Problem

MCAR Missing Completely At Random

MICE Multiple Imputation by Chained Equations

ML Machine Learning

MNAR Missing Not At Random

NACE Nomenclature of Economic Activities

NUTS Nomenclature of Territorial Units for Statistics

OPTICS Ordering points to identify the clustering structure

PCA Principal Component Analysis

PCI Product Complexity Index

PhD Philosophiae Doctor (Doctor of Philosophy)

PNRR Piano Nazionale di Ripresa e Resilienza

R&D Research and Development

RCA Revealed Comparative Advantage

SBS Structural Business Statistics

SFI Santa Fe Institute

SHAP SHapley Additive exPlanations

Sistan National Statistical System

SMEs Small and Medium-sized Enterprises

TECFRAME Trade by Enterprise Characteristics (Istat framework)

TLC Telecommunications

VAT Value Added Tax

XGBoost eXtreme Gradient Boosting

I use maps to find out where explorers have already been. Then I go the other way.

— Expedition Map

1 Introduction

1.1 Motivation and research problem

Economic systems are complex, spatially embedded arrangements of heterogeneous agents, organisations, and technologies[20,36,206]. Firms combine workers, capital, and knowledge to produce goods and services; regions and countries aggregate these micro level activities into emergent patterns of specialization, diversification, and growth[50,138,217]. The structure of these systems is shaped by geography, institutions, and history, and it evolves through processes of innovation, imitation, entry and exit, migration, and structural change[2,109,273]. Understanding how such systems operate and transform over time is a central concern of economic statistics, economic geography, regional science and spatial econometrics[140,285].

Historically, the study of economic development and structural change has oscillated between relatively aggregate representations of the economy focused on sectors, factors, or representative agents and more granular views that emphasise heterogeneity, networks, and spatial detail[110,207]. Classical and neoclassical theories of trade and growth conceptualised countries as combinations of factors and technologies, with comparative advantage driven by differences in endowments or productivity[167,311,340]. Dual economy models and structuralist approaches highlighted the coexistence of traditional and modern sectors, often with distinct spatial footprints[172,232]. New growth theory brought ideas and human capital to the fore-

front[239,314], while new economic geography formalised the role of increasing returns, transport costs, and market size in shaping the spatial concentration of activity[140,217].

In parallel, a large body of work in urban and regional economics, spatial statistics, and economic geography has examined how agglomeration economies, knowledge spillovers, and local institutions shape the distribution of firms and workers across space[107,152,249]. A longstanding empirical tradition within this literature measures *regional specialisation* and *industrial concentration* by comparing sectoral shares across territorial units, using indicators such as location quotients and related concentration/diversification indices (e.g., Gini, Theil, Herfindahl–Hirschman, Krugman, Ellison–Glaeser). Recent systematisations stress, however, that these indicators conflate distinct concepts: sectoral concentration, geographical concentration, within-region agglomeration, and specialisation as "uniqueness"; and that an integrated view requires considering both sectoral and spatial dimensions, ideally combining cluster-based measures with approaches that exploit geocoded microdata[211].

At the same time, improvements in data availability and computational power have made it feasible to analyse large micro-level datasets on firms, workers, and trade flows, often with precise geographic information[113,361]. These developments have opened new possibilities, but they also pose new methodological challenges. Traditional econometric models struggle with the scale, dimensionality, and complexity of modern microdata[42], while Machine Learning (ML) methods that can cope with such data often lack the tools for rigorous inference and uncertainty quantification in spatially dependent settings[24,263]. Moreover, even when the object of interest is "specialisation", inference is complicated by the fact that standard indices are typically computed on discrete spatial aggregations, making results sensitive to zoning choices and hiding within-area clustering patterns, precisely the spatial dimension that theories of agglomeration emphasise.

A particularly influential development in the last fifteen years has been the rise of the economic complexity literature, which proposes network-based indicators of the knowledge intensity of economies based on the structure of their exports and production[170,171]. Importantly, these measures from statistical physics can be read as a refinement of the older specialisation indices coming from economics literature in the statistical framework. They start from the same core object: an economy-by-activity matrix that is first normalised into relative advantage or intensity measures (akin to location quotients and Balassa-type indices) and then transformed through nonlinear, iterative mappings to extract higher-order information about diversification, ubiquity, and capability composition[349]. In this view, countries and regions are not simply bundles of sectors, but ensembles of capabilities, and development is

understood as a process of accumulating and recombining these capabilities[226,349]. Economic complexity metrics have been shown to correlate with growth, diversification patterns, and a range of socio-economic outcomes[85,165], and they have inspired new approaches to industrial and regional policy[31]. However, most of this work remains at relatively aggregate scales, and the micro-foundations of capability accumulation in terms of firms, workers, and local labour markets are still being explored[30].

In parallel, recent years have seen the emergence of spatial data science and Geospatial Artificial Intelligence (GeoAI), which bring together ideas from Geographic Information System (GIS), spatial statistics, and modern ML[197]. These approaches exploit the richness of geocoded data to fit flexible models of spatial phenomena[150]. Yet, when applied to economic microdata, they raise delicate issues. Spatial autocorrelation can lead to overly optimistic validation if ignored[312], spatial heterogeneity and non-stationarity challenge simple parametric assumptions[56], and high dimensional firm-level covariates call for methods that can learn useful representations without losing interpretability[100]. Most importantly for the present work, there is limited guidance on how to perform bootstrap-based inference and uncertainty quantification when data are both high-dimensional and spatially dependent[160,224].

This thesis contributes to these debates by developing and applying ML-based methods for the analysis of spatially embedded productive structures. Empirically, it focuses on the high-tech and innovative business ecosystem of the Lombardy region in Italy. More concretely, the thesis develops a methodological "recipe" that links the ingredients introduced above: it represents productive structures by jointly considering *what* firms do (activities, capabilities, and networks) and *where* they do it (multi-scale spatial contexts, from neighbourhoods to regions); it uses modern ML to learn interpretable representations of high-dimensional firm-level data while preserving these discrete and spatial relations; it provides uncertainty quantification strategies tailored to spatially dependent and high-dimensional settings, so that empirical patterns (clusters, embeddings, and network structures) can be assessed rather than taken at face value. Across the chapters, this logic is instantiated through spatially aware clustering, bipartite graph encodings of multiple partition solutions, and spatial bootstrap schemes. The overarching goal is to produce reproducible, policy-relevant maps of regional productive structures while contributing tools that can be reused whenever rich geocoded economic microdata are available.

1.1.1 Structure of the thesis

The thesis is organised as follows.

- **Chapter 1** This introductory chapter presents an overview of the state of the art and themes that will be presented in this thesis. The rest of the thesis will be developed as a collection of papers that will be presented one after the other in three chapters, that constitute the core of this thesis.
- **Chapter 2** develops a graph-based ensemble clustering framework for innovative startups in Lombardy. Spatial bootstrap resampling is used to generate multiple clustering solutions, which are encoded in a bipartite graph between firms and cluster labels. Bipartite Louvain community detection (biLouvain) is applied to obtain stable, consensus clusters. The chapter analyses the spatial and sectoral structure of these clusters and relates them to measures of firm performance and innovation.
- **Chapter 3** moves from micro to macro scales by linking workers, firms, and hexagonal regions in Italy. Using linked employer–employee data, export-based complexity metrics, and a hexagonal grid representation of territory, it examines how worker diversity, firm productivity, and regional complexity co-evolve across spatial scales. The chapter quantifies the specialisation–diversification trade-off as one moves from firms to local clusters and larger regions and discusses the implications for economic complexity and smart specialization policies.
- **Chapter 4** introduces a Spatial ML pipeline for firm-level productivity analysis in the Lombardy high-tech ecosystem. It presents the data, describes the construction of entity embeddings and deep clustering-based strata, and details the stratified spatial bootstrap procedure. The chapter compares uncertainty estimates and variable importance obtained under different resampling schemes and discusses the implications for spatial ML practice.
- **Conclusion** summarises the main findings of the thesis, reflects on its methodological and empirical contributions, and outlines avenues for future research, including extensions to other regions, integration with causal inference frameworks, and the application of graph neural networks and other advanced ML methods to spatial economic complexity.

The following sections will introduce the elements shared by the papers presented in the following chapters: the importance of the spacial component in statistical and

economical analysis, the context that allowed the development of this thesis and the sources of the data used, an overview of the methodologies implemented and an introduction to Economic Complexity.

1.2 Economic complexity, complexity science, and multi-scale capabilities

1.2.1 Historical perspectives on development and structural change

The modern concept of economic complexity emerges against a rich historical backdrop. Classical political economists such as Smith, Ricardo, and Mill already recognised that the wealth of nations is tied not only to factor endowments but also to the division of labour, the accumulation of skills, and the adoption of new techniques[257,311,339]. Ricardo's theory of comparative advantage focused on relative productivity differences across sectors and countries[311], while Smith emphasised specialisation and the extent of the market as engines of productivity growth[339]. Subsequent structuralist and developmentalist thinkers, including Lewis, Prebisch, and Hirschman, stressed the role of structural transformation, forward and backward linkages, and the uneven diffusion of industrial capabilities[172,232,304].

Post-war development economics focused on the shift from agriculture to manufacturing, on the dynamics of dual economies, and on the role of trade and industrial policy[232,304]. Models of balanced and unbalanced growth captured different views on how investment and policy could trigger virtuous cycles of industrialisation[172,278,315]. At the same time, data limitations constrained empirical analyses to relatively coarse sectoral classifications and aggregate indicators such as Gross Domestic Product (GDP) per capita, capital-output ratios, or simple measures of export structure.

With the advent of new growth theory and endogenous growth models, attention shifted to ideas, human capital, and innovation. Romer-style models emphasised knowledge accumulation and variety expansion[314]; Lucas highlighted human capital and externalities from skill accumulation[239]; Aghion and Howitt formalised Schumpeterian creative destruction[4]. While these models brought technology and knowledge to the centre of growth analysis, their empirical implementation often relied on aggregate proxies such as Research and Development (R&D) expenditure, patent counts, or schooling rates. They also retained relatively simple representations of technology and production, typically involving a small number of sectors or a continuum of symmetric varieties[239,314].

In this context, the economic complexity literature can be seen as part of a broader move towards measuring the *structure* of economies in a more granular, combinatorial way. Rather than focusing only on the volume of production or trade, it asks *what* is produced and exported, and how these activities relate to each other in a high-dimensional space of capabilities[170,171]. This approach resonates with earlier ideas about structural change and industrialisation, but it brings new data and network-based tools to the table. Parallel to these developments, a growing strand of research in international trade and macroeconomics has emphasised how granular production structures, input–output linkages, and sectoral interdependencies shape aggregate outcomes such as productivity, volatility, and development[3,122,178]. Although not always framed explicitly in terms of economic complexity metrics, this literature shares a core concern with the internal structure of economies and the network mechanisms through which shocks and opportunities propagate.

At the same time, the Harvard Growth Lab has played a central role in systematising and popularising the economic complexity approach to development. Building on the product space framework and a family of economic complexity measures (discussed in detail in Subsection 1.2.4), researchers associated with the Growth Lab have combined high-dimensional trade data, network methods, and development diagnostics to study structural transformation, diversification paths, and the constraints faced by developing economies[164,165,171]. This work has been particularly influential in linking economic complexity to policy-oriented analyses of industrial upgrading and long-run development prospects.

1.2.2 The rise of complexity science and complexity economics

The economic complexity literature is closely linked to the broader field of *complexity science*, which studies systems composed of many interacting elements whose collective behaviour exhibits emergent properties that cannot be reduced to the sum of the parts[8,173,259]. Early conceptual foundations for complexity were laid in physics and biology. Ideas such as emergence, broken symmetry, self-organisation, and hierarchical structure challenged strictly reductionist approaches and highlighted the importance of interactions, feedbacks, and adaptation[8,173].

From the 1980s onwards, dedicated research centres for complex systems science played a central role in consolidating this perspective. The Santa Fe Institute (SFI) in New Mexico became a flagship institution for interdisciplinary complexity research[322], bringing together physicists, biologists, computer scientists, and social scientists to study phenomena such as adaptive computation, self-organising systems, and scaling laws in biology and cities. Complexity science at SFI is charac-

terised by a combination of mathematical modelling, computational simulation, and data analysis, with an emphasis on universal principles that cut across domains[259].

In economics, complexity ideas have given rise to what is now often called *complexity economics*. Rather than modelling the economy as a system in equilibrium populated by representative agents, complexity economics views the economy as a complex adaptive system composed of heterogeneous agents who interact locally and learn over time[21,352]. Agent-based models, network models, and non-linear dynamical systems are used to study phenomena such as financial crises, technological change, inequality, and macroeconomic fluctuations[352].

This perspective overlaps historically with evolutionary economics and the tradition of innovation studies, which emphasise routines, learning, and selection as drivers of industrial dynamics and long-run development[273,274]. In Europe, these ideas have been developed and empirically operationalised within communities such as SPRU at the University of Sussex (Chris Freeman, Martin Bell)[41,136], the Sant'Anna School in Pisa (Giovanni Dosi and collaborators)[101], and MERIT at Maastricht University (Luc Soete and colleagues)[137], helping to connect complex-systems thinking with evidence on technological change, innovation systems, and structural transformation. The Institute for New Economic Thinking (INET) at the Oxford Martin School, and in particular its Complexity Economics programme, is one of the main hubs for this line of work[179]. It applies tools from complex systems science—network analysis, agent-based modelling and ML[264] to economic questions ranging from climate policy to supply shocks[94] and firm-level production networks[27].

Within this broader landscape, economic complexity occupies a distinctive position at the interface between complexity science and applied development economics. While early contributions emerged from interdisciplinary environments such as SFI, subsequent work has increasingly been embedded in leading economics departments and policy institutions. In Europe, the Toulouse School of Economics represents an important node in this transition, where tools from network analysis and granular data are integrated into empirically grounded models of trade, production networks, and macroeconomic fluctuations[3,33].

In the United States, the Harvard Growth Lab has served as a key institutional hub for economic complexity research, bridging academic work and policy practice. Its research agenda combines ideas from complexity science, such as emergence, path dependence, and non-linear diversification dynamics, with development diagnostics and country-specific policy analysis. This positioning has contributed to the diffusion of economic complexity concepts beyond academia, influencing international organisations and development agencies[165,166].

In Europe, the Complexity Science Hub Vienna (CSH) provides another important node in the global complex systems network[82]. CSH focuses on data-driven complexity science, using massive administrative and digital datasets to study topics such as health, mobility, finance, and economic resilience[82]. In particular, it has contributed to the literature on transforming economies, supply chain complexity[294], multilayer networks[90] and the resilience of economic systems to shocks[83]. Together, centres such as SFI, INET Oxford, and CSH Vienna have helped legitimise complexity approaches across disciplines and have created an institutional ecosystem within which economic complexity, network science, and data-rich empirical applications can flourish[82,179,322].

Economic complexity sits at the intersection of these developments. It borrows from network science the tools to represent economies as bipartite graphs linking places and activities; from complexity science it borrows the idea that macroscopic outcomes such as development and resilience emerge from the configuration of microscopic capabilities and interactions; and from complexity economics it borrows an interest in heterogeneity, path dependence, and non-linear dynamics[21,171].

1.2.3 The Enrico Fermi Research Center and the statistical physics tradition in economic complexity

A distinctive intellectual lineage within the economic complexity literature originates from the research programme developed at the Enrico Fermi Research Center (CREF) in Rome. This contribution is best understood as part of a broader historical movement that applies concepts and methods from statistical physics to economic and social systems, rather than as a purely technical extension of trade-based complexity metrics.

The roots of this approach can be traced to earlier work by Luciano Pietronero and collaborators on scaling laws, universality, and collective phenomena in complex systems. Drawing inspiration from statistical mechanics, this line of research emphasised that many macroscopic regularities observed in social and economic data, such as power laws, hierarchical organisation, and strong heterogeneity, can emerge from the interaction of many simple units without central coordination[295,296]. This perspective challenged equilibrium-based and representative-agent models by highlighting non-linearity, irreversibility, and emergent structure as defining features of socioeconomic dynamics.

When these ideas were brought to bear on international trade and production, they led to a reinterpretation of development as an emergent property of complex productive systems. In this view, countries are not characterised by isolated

sectoral efficiencies but by their position within a high-dimensional network of complementary capabilities. This conceptual shift motivated the introduction of the Fitness–Complexity framework by Andrea Tacchella and collaborators[349]. Unlike earlier linear measures of economic complexity, the fitness approach was explicitly designed to reflect the strongly non-linear and asymmetric nature of production systems, in which the absence of a single critical capability can constrain otherwise advanced economies.

A recurring theme in this body of work is the interpretation of economic complexity indicators as relational and systemic quantities rather than as intrinsic attributes of countries or products. Fitness and complexity are understood as co-evolving outcomes of network structure, historical trajectories, and capability accumulation processes[325,332,350,371].

From a history-of-thought perspective, this program represents a clear departure from both neoclassical trade theory and early endogenous growth models. Rather than assuming smooth substitution between factors or symmetric technological varieties, it foregrounds complementarity, irreversibility, and path dependence. Development is conceived not as a movement along a stable production function but as a process of structural exploration constrained by existing capability sets. This interpretation resonates with earlier structuralist and evolutionary traditions, while grounding them in a formal, data-driven framework inspired by complexity science.

An additional contribution of the Enrico Fermi group lies in its emphasis on dynamics and historical sequencing. Several studies showed that changes in fitness anticipate changes in income, suggesting that productive structure evolves prior to observable improvements in aggregate economic performance[297,371]. This temporal ordering reinforces the idea that development is a cumulative and historically contingent process, shaped by the gradual accretion and recombination of capabilities.

Over time, the research agenda expanded beyond international trade to include regional economies, environmental constraints, and broader notions of sustainability[89,290]. These extensions reflect an ongoing effort to generalise the complexity framework to multiple spatial and thematic scales, while remaining faithful to its original conceptual core: the study of economic systems as evolving, non-equilibrium configurations of interdependent capabilities.

Taken together, the contribution of the Enrico Fermi Research Center occupies a distinctive place in the intellectual history of economic complexity. By embedding development and production within the broader tradition of statistical physics and complexity science, it has provided a conceptual bridge between network-based empirical work and a deeper theoretical understanding of economic structure, emergence, and historical change.

1.2.4 Capabilities, networks, and complexity metrics

Economic complexity starts from the intuition that production requires capabilities: broadly defined as the knowledge, skills, and organisational arrangements needed to perform an activity[171]. Individual capabilities are not directly observable, but their combination leaves a trace in the pattern of activities performed by an economy. If a country can export a product competitively, it is inferred to possess the capabilities required for that product; if it can export many sophisticated products, it is likely to have a rich capability base.

Operationally, this intuition is implemented through a bipartite network between places (usually countries) and activities (usually products in which the country has a Revealed Comparative Advantage (RCA))[28]. The structure of this network is summarised by measures of diversity (the number of products a country exports) and ubiquity (the number of countries exporting a given product). Early work showed that richer countries tend to be more diversified and to export less ubiquitous products, while poorer countries export few, common products. This observation motivated the definition of Economic Complexity Index (ECI) and the Product Complexity Index (PCI), which use iterative procedures to assign complexity scores to countries and products[171].

An alternative but related approach is the Fitness–Complexity algorithm, which defines country fitness and product complexity through a set of coupled, non-linear equations[349]. Highly fit countries are those that export many complex products; complex products are those exported by few, highly fit countries. Both ECI/PCI and Fitness–Complexity exploit the combinatorial structure of the country–product network to infer hidden capability structures[171,349]. Empirically, these indices have been shown to correlate with future growth, economic diversification paths, and various indicators of social and environmental performance[125,165].

Subsequent work has extended and critiqued these measures. Alternative formulations of complexity indices have been proposed, along with robustness checks and comparisons across different product classifications and RCA thresholds[125]. Some studies have highlighted that different complexity metrics can yield different rankings and have suggested principled ways to choose among them; others have explored how complexity relates to income inequality, environmental indicators, or regional development. In parallel, the methodology has been applied beyond trade data, including to patent portfolios, scientific production, occupational structures, and regional industry structures[272]. These extensions reveal that the basic idea of inferring hidden capabilities from observed place–activity bipartite networks is widely applicable.

From the perspective of this thesis, the key contribution of economic complexity is to provide a language and a set of tools for thinking about development as the accumulation and recombination of discrete capabilities, observable through the pattern of activities in which an economy is engaged[171]. This perspective dovetails with micro-level analyses of workers' skills and firms' product mixes, and it is naturally expressed in terms of bipartite networks and clustering, connecting it to the methods used in later chapters.

1.2.5 Related variety, diversification, and smart specialization

A closely related strand of research, arising in evolutionary economic geography, is the literature on related variety and regional diversification. While economic complexity focuses on the knowledge intensity of activities and the structure of place–activity networks, related variety emphasises the cognitive and technological proximity between industries and its implications for regional growth and diversification[51,138].

The starting point is the observation that regions can benefit from both specialisation and diversification, but in different ways. Specialisation in a particular industry can generate localisation economies, such as specialised labour markets and supplier networks, while diversification across unrelated industries can provide portfolio insurance against sector-specific shocks. Related variety seeks to reconcile these forces by distinguishing between diversification into related versus unrelated activities[138]. Empirical studies have shown that regions tend to diversify into industries that are related to their existing portfolio and that such related diversification is associated with higher growth and innovation outcomes[138,268].

These ideas have important policy implications. The European Union (EU)'s smart specialization strategy, for example, calls on regions to identify areas of competitive advantage based on their existing capabilities and to support diversification into related activities[117,131,132]. Rather than promoting the same set of high-tech sectors everywhere, smart specialization emphasises place-specific pathways of structural change grounded in the local knowledge base[131]. Economic complexity metrics and relatedness measures have been widely used to inform this agenda by mapping the "product space" or "industry space" of nearby opportunities for each region or country[163,165,170].

The thesis builds on this literature by examining how diversification and specialisation play out across multiple spatial scales, from firms to hexagonal grid cells to larger regions. It uses measures inspired by economic complexity and related variety but applies them to micro-level data on workers and firms. In doing so, it connects

macro-level patterns of diversification and complexity to the internal organisation of firms and to the composition of local labour markets.

1.2.6 From macro indicators to firm-level microdata

Most empirical applications of economic complexity and related variety are conducted at the level of countries, regions, or large cities [165,171]. This choice is partly dictated by data availability (trade data, regional accounts, patent statistics) and partly by the focus on macroeconomic outcomes such as GDP growth or regional employment. However, the capabilities that underpin complexity and relatedness are ultimately embodied in workers, organisations, and institutions. Firms combine workers with different skills and educational backgrounds; they adopt technologies and organisational routines; they engage in collaborations and networks. Similarly, local labour markets mediate the match between workers and firms, shaping the evolution of regional capability portfolios.

Recent work has started to bridge the gap between macro-level complexity metrics and micro-level data. Studies using linked employer–employee datasets, detailed occupational and educational classifications, and geocoded firm locations have shown that the composition of the local workforce and the structure of firms' skill requirements are strongly related to measures of regional complexity [272]. Other contributions have analysed how firm-level diversification in products or technologies relates to corporate performance and resilience, using firm-level export data, patent portfolios, or production networks.

Chapter 4 of this thesis contributes to this emerging literature by building a multi-scale empirical framework that links workers, firms, and hexagonal regions in Italy. Using a combination of firm-level productivity data, information on workers' educational backgrounds, and export-based measures of product complexity, the chapter investigates how patterns of diversification and complexity evolve as one moves from individual firms to local clusters and larger regions. The use of a hexagonal spatial grid allows for a flexible representation of economic space that is independent of administrative boundaries, facilitating comparisons across scales and mitigating issues related to the Modifiable Areal Unit Problem (MAUP).

1.3 The spatial dimension: regions, clusters, and firms

1.3.1 Historical roots of spatial analysis in economics

Economists and geographers have long been interested in the geographical aspect of economic activity. Von Thünen, Christaller, and Kösch's early contributions offered stylized models of land use and central place systems, emphasizing how market access and transportation costs influence the locations of urban centers and agricultural production[74,238,363]. Marshall emphasised the role of industrial districts and agglomeration economies, identifying three classic sources of local externalities: a specialised labour pool, supplier networks, and knowledge spillovers[249]. Weber developed a theory of industrial location based on transport and labour costs[365].

In the late twentieth century, new economic geography formalised these ideas using tools from industrial organisation and general equilibrium theory. Models by Krugman, Fujita, Venables and others showed how increasing returns, monopolistic competition, and trade costs can generate agglomeration and core–periphery structures[140,217].

A substantial body of empirical research on agglomeration, regional inequality, and trade integration was spurred by these models, which offered a micro-founded explanation for the spatial concentration of manufacturing and high-tech industries[218,362].

At the same time, improvements in GIS and spatial econometrics helped empirical spatial analysis. Econometricians were able to account for geographically correlated shocks and spillovers using spatial lag, spatial error models, spatial Durbin models, and related specifications[9,231]. To identify and characterize spatial autocorrelation in cross-sectional data, spatial weights matrices, Moran's I, and Local Indicators of Spatial Association (LISA) were developed[10,262]. GIS technologies facilitated the integration of disparate spatial datasets and the visualisation of spatial patterns[149].

Despite these developments, a lot of empirical research still used coarse grids or administrative regions, and it frequently viewed space as a discrete set of units rather than a continuous surface. Concerns about the selection of geographical scale and aggregation, as summed up by the MAUP, persisted[283]; especially when interpreting estimated impacts or comparing findings between research.

1.3.2 Agglomeration, innovation, and spatial concentration

The spatial concentration of innovation and high-tech activity has been documented in many countries and contexts. Silicon Valley and similar high-tech clusters are

emblematic examples[324], but more dispersed patterns of innovative activity can also be observed in regions that specialise in particular industries or technologies[26]. The literature has identified several mechanisms that contribute to such clustering: localisation economies associated with specialised supplier and customer networks; urbanisation economies linked to diversified urban environments; and various forms of knowledge spillovers that are facilitated by geographical proximity[107,144].

Empirical studies have used a wide range of techniques to characterise these patterns, from location quotients and Gini-type indices to distance-based measures of spatial concentration and point pattern analysis[108,115]. Marked point processes and spatial interaction models have been used to relate firm attributes to their spatial distribution[97]. Recent work has emphasised the importance of distinguishing between different types of agglomeration (e.g., industrial versus urban) and of accounting for underlying spatial trends when assessing the significance of observed clusters[13,245].

In the context of this thesis, the key point is that the productive and innovative fabric of Lombardy, and of Italy more generally, is far from uniformly distributed in space. Historical, institutional, and infrastructural variables influence the concentration of high-tech companies and creative startups in particular corridors, cities, and local systems[177]. Any attempt to model firm-level outcomes such as productivity, or to identify clusters of startups, must therefore take the spatial structure of the data seriously, both in the design of the analysis and in the interpretation of the results.

spatial error models, introduced spatially lagged dependent variables or error terms to capture spillovers and correlated shocks[9]. The specification of spatial

1.3.3 The Italian and Lombardy productive system

Italy offers a particularly interesting setting for the study of Spatial ML. The country is characterised by pronounced regional disparities, with a well documented North–South divide in income, employment, and industrial structure[124]. At the same time, Italy hosts a dense fabric of Small and Medium-sized Enterprises (SMEs), many of which are embedded in local production systems and industrial districts[37]. These districts specialise in a range of manufacturing activities and have historically played a central role in Italian exports[38].

Within Italy, Lombardy stands out as one of the most productive and innovative regions. It hosts a large share of Italian manufacturing and business services, a significant concentration of high-tech and knowledge-intensive firms, and major universities and research centres[280]. Large cities like Milan and smaller industrial towns coexist in the region, creating a complex mosaic of local systems with various specializations and capacity profiles[58]. Additionally, it is crucial to the success of

Italian exports and the development of new business ventures, such as cutting-edge startups in advanced manufacturing, digital technology, and life sciences[81].

The empirical focus of Chapters 2 and 3, which develop and apply new techniques in bootstrap inference and cluster analysis using firm-level and startup-level data from Lombardy, is motivated by this institutional and spatial setting.

1.3.4 Spatial scales, MAUP, and continuous space representations

A central methodological challenge in spatial analysis is the choice of spatial scale and partition. Many datasets are organised according to administrative units (municipalities, provinces, regions), which are convenient but often arbitrary from an economic viewpoint. MAUP refers to the fact that statistical relationships can depend on the choice of spatial aggregation and zoning[282,283]. For example, measures of concentration or spatial autocorrelation may differ if data are analysed at the municipal level versus the provincial level, or if regions are redefined[14].

One way to mitigate these issues is to use geometric partitions of space, such as regular grids, which are independent of administrative boundaries[134]. Hexagonal grids, in particular, have attractive properties: they provide compact, uniform coverage, each cell has the same number of neighbours, and distances between cell centroids are relatively homogeneous[46]. Hexagonal indexing systems such as H3 allow for multi-resolution representations, where each cell can be refined into smaller cells in a hierarchical fashion[357].

This thesis adopts a hexagonal grid representation of Italy to construct spatial units that can be scaled up or down as needed. Firm locations are mapped onto grid cells, enabling the construction of local labour market measures, firm densities, and complexity indicators at multiple spatial resolutions. This approach facilitates comparisons across scales and provides a flexible foundation for the multi-scale analysis conducted in Chapter 4.

1.4 Data and institutional context

1.4.1 Official business statistics and the ASIA registers

A central data source underlying the empirical analyses in this thesis is the set of Business Register (BR)s maintained by the Italian National Institute of Statistics (Istat), in particular Statistical Archive of Active Businesses (ASIA). ASIA is the official statistical register of active enterprises and local units in Italy[188]. It is

updated annually through the integration of multiple administrative and statistical sources, including tax records, social security data, and specialised surveys, in line with European standards for statistical BRs[120]. The register covers enterprises in industry and services and provides identifying information (such as location and legal form) as well as structural variables (such as employment and industry classification).

ASIA was established in the mid-1990s in response to European regulations on BRs for statistical purposes and their subsequent harmonisation at the EU level[119]. Since then, it has become the cornerstone of Italian business statistics. It serves as the sampling frame for a wide range of structural business surveys, including surveys on turnover in services, innovation, and research and development[184]. It is also used to construct official statistics on the structure of the enterprise population, such as the distribution of firms by size, sector, and territory. Annual releases provide tabulations of enterprise counts and employment by region, province, and industry, enabling the monitoring of structural changes in the productive system[191].

For the purposes of this thesis, ASIA plays multiple roles. First, it provides the basic population frame from which firm-level microdata are drawn, ensuring that analyses are grounded in a comprehensive and coherent view of the enterprise universe[120]. Second, the structural variables in ASIA, such as employment size and industry code, are used to define strata, to construct controls, and to benchmark the representativeness of the microdata used in the spatial ML and clustering exercises. Third, the geographical coordinates or location variables associated with enterprises enable the mapping of firms onto hexagonal grid cells and the construction of local indicators of firm density and industrial structure.

1.4.2 Statistical analysis by Istat

Beyond maintaining the ASIA registers, Istat produces an extensive body of statistical analysis on enterprises and territories. Structural business statistics yield regular publications on enterprise demographics (births, deaths, survival), productivity, and value added by sector and region[187,189]. Thematic analyses focus on topics such as innovation, digitalisation, internationalisation, and the integration of firms in global value chains, often combining ASIA with survey data or with external sources such as customs data[183,185]. These analyses provide a rich and consistent description of the Italian productive fabric.

Istat also plays a key role in the National Statistical System (Sistan), coordinating with regional and sectoral statistical offices in the production of official statistics[193]. Territorial statistics provide indicators of socio-economic conditions

at various spatial scales, from municipalities to regions, covering variables such as income, employment, business density, and infrastructure[190]. For this thesis, these official statistics serve both as background and as external validation. For example, aggregate indicators of enterprise structure and value added by province and region are used to contextualise the micro-level findings on Lombardy and to ensure that the patterns observed in the microdata are consistent with the broader territorial picture.

1.4.3 Firm and territory level indicators from the Guglielmo Tagliacarne Institute

Complementing Istat's official statistics, the *Centro Studi delle Camere di Commercio Guglielmo Tagliacarne* plays an important role in providing territorial socio-economic analysis in Italy[65]. The Tagliacarne Institute is the research and statistical arm of the Italian Chambers of Commerce system. Its mission is to promote economic culture and to provide data, indicators, and analyses to support the competitiveness of firms and territories. To this end, it produces studies on local production systems, SMEs dynamics, and the impact of public policies, and it maintains a variety of territorial databases and dashboards[66].

Of particular relevance are the Institute's *Statistiche territoriali*, which provide indicators such as value added by province and sector, employment, entrepreneurial density, and sectoral specialisation[67]. These indicators are often produced in collaboration with Istat and within the framework of Sistan, ensuring consistency with official statistics while adding a territorial and business-oriented perspective[193]. Tagliacarne's analyses of local production systems, industrial districts, and regional competitiveness are widely used by regional governments, Chambers of Commerce, and other institutions[64].

1.4.4 Integration of microdata and official statistics

The empirical work in this thesis relies on a careful integration of different data sources. Firm-level and startup-level microdata, including information on employment, balance sheet variables, sector, location, and, where available, innovation-related attributes, are linked to the ASIA register for consistency in identifiers, structural variables, and industry classification[188]. Workers' data, including information on educational background and occupation, are linked to firms and to hexagonal grid cells, enabling the construction of measures of worker diversity and local labour market composition[186].

These microdata are then embedded in the broader statistical infrastructure provided by Istat and the Tagliacarne Institute. Official indicators of value added, employment, and enterprise structure by province and region provide context and benchmarks[67,190]. Territorial analyses of local production systems and industrial districts help interpret the spatial clusters of startups and high-tech firms[64]. In some cases, aggregate statistics are used as explanatory variables or as controls in Spatial ML models; in others, they serve as external validation for patterns uncovered by clustering and complexity metrics.

This multi-layered data architecture, combining microdata, BRs, and territorial statistics, is a key enabler of the thesis. It allows the analyses in Chapters 2 and 3 to be grounded in a comprehensive and consistent view of the Lombardy productive system, and it supports the multi-scale perspective adopted in Chapter 4, where workers, firms, and regions are linked within a unified framework.

1.5 Spatial data science and spatial machine learning

1.5.1 From spatial statistics to spatial data science

Spatial statistics emerged in the late twentieth century to address the presence of spatial dependence in cross-sectional and panel data. Classic models, such as the spatial lag and weights matrices, typically based on contiguity or distance, became a central modelling choice[78]. Estimation techniques ranged from maximum likelihood to instrumental variables and Bayesian methods[231].

While spatial econometrics provided a rigorous statistical framework, it often assumed relatively simple functional forms and low-dimensional covariate sets. In many applications, the focus was on testing for the presence of spatial effects and estimating their magnitude, rather than on prediction or high-dimensional feature learning[114]. In parallel, GIS and exploratory spatial data analysis tools enabled the visualisation and descriptive analysis of spatial patterns[11].

In the last decade, the landscape has changed with the advent of spatial data science and GeoAI. These fields are less tightly tied to specific econometric models and more open to ML, computer vision, and data mining methods[234,335]. Spatial data science leverages large geocoded datasets, high-resolution remote sensing, and volunteered geographic information, while GeoAI applies convolutional neural networks, random forests, gradient boosting, and other algorithms to spatial prediction tasks[375]. Examples include land use classification from satellite imagery, fine-grained population mapping, spatial interpolation of environmental variables, and urban feature extraction.

Spatial data science also emphasises reproducible workflows, open-source tools, and the integration of heterogeneous data sources[337]. This ecosystem of tools and practices provides the technical environment within which the Spatial ML pipeline developed in Chapter 3 is implemented.

1.5.2 Emergence of spatial machine learning and GeoAI

Spatial ML (or GeoAI) can be defined as the application of ML methods to spatially located data, with explicit attention to spatial structure and dependence[234,335]. In the environmental sciences and ecology, spatial ML has been used to improve species distribution models, climate downscaling, and ecosystem service mapping[168]. In transportation and urban planning, it has informed demand forecasting, route optimisation, and the analysis of mobility patterns[203]. In economics and regional science, spatial ML is increasingly used to model house prices, firm performance, poverty, and other socio-economic outcomes[263].

Compared to traditional spatial econometrics, spatial ML typically places less emphasis on explicit parametric modelling of spatial dependence and more emphasis on predictive performance and flexible interactions between features[24]. Tree-based ensembles, such as random forests and gradient boosting machines, can capture non-linearities and variable interactions without specifying a functional form a priori[52,139]. Neural networks and representation learning methods can extract features from complex inputs such as images or text, and spatially explicit architectures (e.g., convolutional networks, graph neural networks) can directly model spatial adjacency or network structure[55].

However, the application of ML to spatial data raises specific challenges. Spatial autocorrelation violates the independence assumptions underlying many validation procedures; spatial non-stationarity and heterogeneity may require models that vary across space; and the interpretation of variable importance or partial dependence plots can be complicated when variables are spatially structured[312]. These challenges have motivated a growing literature on spatial feature engineering and hybrid models that combine ML with spatial statistical components.

1.5.3 Spatial bootstrap and uncertainty

In geostatistics, spatial bootstrap methods have been developed to quantify uncertainty in variograms, kriging predictions, and simulated realisations[72]. These approaches often rely on model-based conditional simulation of spatial fields and on resampling strategies that respect spatial dependence[224]. However, they are

typically designed for continuous spatial fields and relatively low-dimensional covariates, rather than for discrete firms with rich attribute sets.

In econometrics, bootstrapping with dependent data has been studied primarily in time series and panel settings, with block bootstrap and cluster bootstrap methods[59,221]. Analogous ideas have been proposed for spatial data, including spatial block bootstrap and resampling of spatial clusters[224]. Yet, in high-dimensional firm-level ML applications, straightforward block resampling may be impractical or inefficient.

Recent advances in spatial econometrics and Spatial ML have emphasized the importance of explicitly accounting for spatial heterogeneity, multiscale dependence, and algorithmic uncertainty when working with micro-level economic data. In particular Kopczewska[212] and Kopczewska[213] argue that classical resampling schemes are insufficient in spatial ML contexts, as they may fail to preserve both spatial dependence structures and similarity patterns in high-dimensional feature spaces. Their work highlights the need for bootstrap designs that are aligned with the data-generating process implied by spatial clustering, network representations, or learned latent embeddings.

This thesis contributes to this literature by proposing two complementary strategies. Chapter 2 aggregates data using bipartite graph partitioning. Chapter 3 introduces a *stratified spatial bootstrap* in which firms are grouped into strata defined jointly by spatial proximity and similarity in latent attribute representations learned via deep clustering, extending ideas of spatially aware resampling discussed by Kopczewska[212]. In both cases, the bootstrap becomes a central tool for assessing the robustness and uncertainty of patterns uncovered by spatial ML and clustering methods, rather than merely a variance estimation device.

1.6 Clustering, networks, and deep representation learning

1.6.1 Classical clustering and validation

Clustering methods aim to partition a set of observations into groups that are internally homogeneous and externally heterogeneous according to some notion of similarity. Classical algorithms such as k-means, k-medoids, and hierarchical clustering differ in how they define similarity, how they search the space of partitions, and how they represent cluster structure[200,204,241]. Density-based methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Hier-

archical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identify clusters as regions of high point density separated by regions of low density, allowing for non-spherical clusters and noise points[60,116].

In many applications, clustering is used as an exploratory tool, and the choice of algorithm and parameters (e.g., the number of clusters, density thresholds) is guided by domain knowledge and heuristic diagnostics. Cluster validation indices such as the silhouette coefficient, the Dunn index, or the Davies–Bouldin index provide quantitative measures of cluster quality based on within-cluster cohesion and between-cluster separation[87,106,316]. However, these indices can be sensitive to the scale and shape of clusters and may not align with domain-relevant notions of meaningful grouping[159].

The challenges are amplified in high-dimensional, sparse, or mixed-type data. Distance measures such as Euclidean distance can become uninformative (the "curse of dimensionality")[45], and categorical variables require specialised treatment. Firm-level microdata often combine continuous financial variables, count variables, and high-cardinality categorical variables such as industry codes or occupational categories, making off-the-shelf clustering methods difficult to apply in a principled way.

1.6.2 Ensemble and consensus clustering

An additional difficulty is that clustering results can be unstable. Small changes in the data (e.g., due to sampling) or in the algorithm's initialisation can lead to different partitions, especially when clusters are weakly separated[196]. Ensemble and consensus clustering methods address this by aggregating multiple clustering solutions into a single, more robust partition[135,346]. The basic idea is to generate a collection of clusterings (e.g. by varying initialisations, subsampling the data, or modifying hyperparameters) and then use a consensus function to combine them.

One popular approach constructs a co-association matrix that records, for each pair of observations, the proportion of clusterings in which they co-occur in the same cluster[135]. This matrix can be interpreted as a similarity measure and clustering can be applied to it to obtain a final partition. Alternatively, one can represent the ensemble as a graph where nodes are observations and edges are weighted by co-association, or as a bipartite graph between observations and cluster labels[346]. Consensus functions include majority voting, graph partitioning, and more sophisticated optimisation criteria[354].

1.6.3 Networks, bipartite graphs, and community detection

Network representations provide a unifying language for many of the concepts discussed so far. Economic complexity measures are built on bipartite networks between places and activities[171]; related variety can be represented as a network of related industries or technologies[50]; and cluster ensembles can be encoded as graphs linking observations and cluster labels[346]. Community detection algorithms seek to identify groups of nodes that are more densely connected internally than with the rest of the network[133].

A widely used class of community detection methods is based on modularity maximisation. Modularity is a quality function that compares the observed density of edges within candidate communities to the density expected under a null model with no community structure[277]. The Louvain algorithm and its successors provide fast, greedy heuristics for maximising modularity in large networks[47]. The Leiden algorithm improves on Louvain by guaranteeing well-connected communities and better convergence properties[355].

In bipartite networks, where there are two distinct types of nodes and edges occur only between types, standard modularity definitions are not directly applicable. Specialised formulations of bipartite modularity have been proposed that account for the two-mode structure[35]. These formulations have been used in the analysis of country–product networks, affiliation networks, and other two-mode systems[171]. Algorithms such as biLouvain extend modularity maximisation heuristics to the bipartite setting[265].

Chapter 2 of this thesis leverages this framework by representing the output of multiple clusterings, as a bipartite graph between firms (innovative startups) and cluster labels. By maximising bipartite modularity on this graph, the algorithm recovers stable consensus clusters.

1.6.4 Deep representation learning and deep clustering

Deep learning has introduced powerful tools for learning representations of complex, high-dimensional data[151]. Representation learning aims to discover latent features that capture salient structure in the data and that can be used for downstream tasks such as classification, regression, or clustering. In the context of tabular data with categorical variables, neural networks can learn dense *embeddings* of categories[157]. This idea has been widely used in natural language processing[256] and recommender systems[216], and it has increasingly been applied to economic and administrative data[44].

Deep clustering methods integrate representation learning and clustering into a single framework. Deep Embedded Clustering (DEC), for example, starts from an autoencoder and jointly refines latent representations and cluster centroids by minimising a clustering-oriented loss function[369]. Variants of DEC and related methods have been developed for different data types, loss functions, and architectures[258].

For firm-level microdata, deep representation learning is particularly useful because many important attributes are categorical and high-cardinality. Entity embeddings can capture similarity relations between such categories based on their co-occurrence patterns in the data[157]. When combined with continuous variables and spatial coordinates, these embeddings provide a rich, compact representation of firms in a latent space where clustering and stratification can be more effective.

Chapter 3 of this thesis uses these ideas to construct strata for a stratified spatial bootstrap. Firms are embedded in a latent space that reflects both their attributes and their spatial context; deep clustering is then applied to this space to define strata that are internally homogeneous and spatially coherent.

1.7 Bootstrap and uncertainty in high-dimensional spatial machine learning

1.7.1 Bootstrap for independent and dependent data

Bootstrap methods provide a flexible approach to approximating the sampling distribution of an estimator by resampling from the observed data[111]. In the simplest case, observations are assumed to be independent draws from a common distribution, and bootstrap samples are generated by sampling with replacement from the dataset. The estimator is recomputed on each bootstrap sample, and its sampling distribution is approximated by the empirical distribution of bootstrap estimates[88,112].

When data exhibit dependence structures, such as time series or spatial data, naive resampling at the level of individual observations can break the dependence and lead to biased or inconsistent approximations. Various extensions have been proposed to address this, including block bootstrap methods for time series (where contiguous blocks of observations are resampled)[221,301], cluster bootstraps for grouped data[59], and heteroskedastic regressions[243,367].

For spatial data, analogous ideas involve resampling spatial blocks or clusters of nearby observations, or using model-based simulation to generate pseudo-replicates of the spatial field[72,224]. In practice, however, the design of spatial bootstrap schemes is often problem-specific, and there is relatively little guidance for high-dimensional

ML applications with irregularly located units such as firms[223].

1.7.2 Design-based and model-based perspectives

Bootstrap methods can be motivated from different perspectives. In a model-based view, the observed data are treated as a sample from an underlying probability model, and the bootstrap aims to approximate the sampling distribution of an estimator under that model[88]. In a design-based view, particularly common in survey sampling, the focus is on the sampling design (e.g., stratified, cluster, multistage sampling), and bootstrap schemes are constructed to mirror the design[308,338]. In spatial applications, both perspectives are relevant: the spatial arrangement of units and the sampling scheme may not be under the analyst's control, but explicit models of spatial dependence can be used to guide resampling strategies[84,224].

In firm-level applications, the design-based perspective suggests constructing bootstrap schemes that respect the structure by which firms are distributed in space and by which data are collected (for example, oversampling certain sectors or regions)[308]. The model-based perspective suggests using knowledge about spatial autocorrelation and covariate structure to define resampling units that preserve dependence[224]. This thesis adopts a hybrid approach: it uses data-driven clustering to define strata that reflect both spatial proximity and attribute similarity, and then applies a stratified bootstrap that resamples within these strata.

1.7.3 Challenges in economic applications

Applying bootstrap methods to economic microdata with spatial structure and high-dimensional covariates presents several challenges. First, the number of potential resampling schemes is large, and naive choices can either break dependence (leading to underestimation of uncertainty) or be overly conservative (leading to inefficient estimates)[224]. Second, spatial dependence may interact with covariates in complex ways, making it difficult to define resampling units that capture the relevant structure[84]. Third, the computational cost of repeated model fitting on bootstrap samples can be substantial, especially when models are complex ML algorithms[112].

Chapters 2 and 3 address these challenges by integrating bootstrap design with representation learning and clustering. In Chapter 3, deep clustering on entity embeddings and spatial features defines strata that are used to guide resampling, thereby embedding knowledge about spatial and attribute structure into the bootstrap.

“All models are wrong, but some are useful.”

George E. P. Box

2

Bipartite graph partitioning and spatial bootstrapping: a case study of innovative startups

We present an ensemble clustering approach that builds a bipartite graph from multiple base partitions and applies biLouvain for consensus. We then recover interpretability via eXtreme Gradient Boosting (XGBoost)-based post hoc explanations and finally use the resulting clusters as strata in a spatially-aware bootstrap. The case study covers innovative startups active in Lombardy (2019–2021), matched to ASIA-Istat microdata and the BR special section.

2.1 Motivation and Introduction

Innovative startups are the source of innovation and technological development; therefore, understanding their behaviour can help better recognize the business organization’s direction. This chapter introduces a new method for clustering innovative startups using bipartite graph partitioning combined with spatial bootstrapping, improving clusters’ accuracy and interpretability. Recent advancements in clustering techniques have introduced ensemble or consensus clustering methods, which aim

to merge multiple clustering results into a superior outcome. A key challenge in this field is effectively integrating diverse clusters, and one promising solution involves utilizing graph formalism and partitioning strategies. By leveraging advanced graph partitioning techniques, we transform the task of partitioning the ensemble graph into a community detection problem. Our methodological approach improves the traditional method of bipartite graphs used in cluster ensembles by implementing the state of the art biLouvain algorithm. We also focused on techniques that could be used to increase the interpretability of the clusters themselves and how they can be used to obtain insightful information from the data. The proposed methodology was applied to a dataset of technologically advanced new businesses, located in the Lombardy region and recorded as innovative startups in the special section of the Italian Chambers of Commerce's BR.

As Schumpeter asserts, new-borns, particularly innovative startups, are the primary drivers of change and a critical strategic asset for the economic development of the country. This underscores the crucial role of startups in driving economic growth. In a world propelled by innovation, startups have emerged as the transformation pioneers, redefining norms and revolutionizing entire sectors[251,328]. Innovative startups have the potential to create and mold new industries, generating substantial economic and social impacts. Consequently, various policy initiatives aim to support innovative startups' establishment, growth, and influence. Over the past decades, innovation has received increasing attention from policymakers, managers, and entrepreneurs, so much so that it has become one of the central themes in the current economic and political debate. The purpose of this chapter is not to investigate the innovation processes of firms and innovative startups, in particular, managerial implications about the importance of innovation and its impact on business performance, but to identify, through the construction of an ad hoc algorithm, the clusters of startups and the main features that foster the development of innovative new firms. We propose a new methodology, namely an ensemble clustering algorithm for bipartite graphs based on the biLouvain algorithm and the concept of consensus clustering. The development of cluster ensemble or consensus clustering methods is a recent and intriguing advancement in clustering techniques[126,346].

The aim in cluster ensemble research is combining multiple clusters to produce a superior final clustering outcome. By leveraging advanced graph partitioning techniques, we address this issue by transforming it into a community detection problem[133]. We introduce a reduction method that constructs a bipartite graph from a given set of clusters, where the resulting graph simultaneously represents both the instances and clusters of the ensemble as vertices[127].

Our approach enhances the traditional bipartite graph method by incorporating the

biLouvain algorithm[293], which is a version of the Louvain algorithm[47] specifically made for bipartite graphs. Louvain is an unsupervised algorithm, namely a hierarchical clustering algorithm, divided into two steps: modularity optimization and community aggregation. It recursively merges communities into a single node and executes the modularity clustering on the condensed graphs. In addition, the biLouvain method allows a positive modularity gain; precisely, it displays better quality measured by bipartite modularity than existing methods.

A common challenge with clustering algorithms is that after the final clusters are formed and identified by the communities of the biLouvain algorithm, it can be challenging to determine the shared characteristics of the data points within each cluster and to identify which features were most influential in the clustering process. This problem, common to all ML algorithms, is called the loss of model interpretability. In this work, we will implement strategies to recover the explainability of the model[244] in the sense that the results of the clustering algorithm will be interpreted *post hoc* to translate the behavior of the ML model into understandable terms for humans[162,374]. Problem-solving in ML often involves classification techniques, based on Gradient Tree Boosting (GTB) algorithm[71]. XGBoost is an advanced version of the GTB algorithm. The key improvement in XGBoost over traditional GTB is the inclusion of regularization in the objective function, which helps prevent overfitting. However, one of the problems in investigating the economic growth of firms, like innovative startups, or their survival rate, is found in the lack of economic and financial data of the companies and in the lack of timeliness and punctuality in the collection of such data found in databases such as the BR[68] or Italian company information and business intelligence database (AIDA)[7]. This leads to bias in the economic analysis conducted regarding business performances.

Missing data is a common problem, even in well-designed and managed research studies. Multiple imputation[317,318] is the preferred method for addressing complex issues with incomplete data. When data is missing across multiple variables, it presents a unique challenge. Two primary approaches have been developed for imputing multivariate data: Joint Modeling (JM) and Fully Conditional Specification (FCS), also referred to as multivariate imputation by chained equations (Multiple Imputation by Chained Equations (MICE))[358].

Then, only after preprocessing the data, we applied a clustering algorithm, describe above. Once a satisfactory and explainable clustering strategy has been established, it can be leveraged to gain insights from the data. Specifically, the clusters have been utilized as strata for a stratified bootstrap algorithm. By "stratification", we refer to organizing geographical objects into subsets, known as strata, based on the similarity of their attributes or spatial relationships[364].

The proposed methodology is applied to a dataset of innovative startups [181] localized in the Lombardy region, active during 2019/2020/2021, and signed up to the innovative startups section of the Chambers of Commerce register from 2017 onwards. The goal is to show how the proposed algorithm captures both spatial and socioeconomic characteristics of clusters of enterprises. The choice of which type of firms to analyze fell on innovative startups, as these high-tech firms have strong growth potential and represent one of the key points of Italian industrial policy. On the other hand, the constructed database rich in information allows us to highlight additional characteristics of innovative startups, often overlooked in the literature, such as whether or not they belong to a industrial group or holding or differences in characteristics between being located in suburban areas or near urban areas. The chapter is organized as follows: Section 2.2 describes the current literature, Section 2.3 illustrates in detail the methodology proposed, Section 2.4 shows the results of applying the methodology to the case study of innovative startups, Section 2.5 presents the conclusions and lays the groundwork for future research.

2.2 Background

Laying the groundwork for the proposed methodology, this section is divided into three parts, each addressing the primary features of the critical building blocks of the proposal:

1. Section 2.2.1 focuses on Imputation data, briefly covering the essential concepts and relevant literature;
2. Section 2.2.2 introduces the topic of cluster ensemble and the Louvain algorithm;
3. Section 2.2.3 briefly describes the XGBoost algorithm and the problem of the spatial bootstrap that occurs when, in the geographical phenomena, attributes within strata are more similar than the between-strata.

2.2.1 Imputation of missing data

A complete and representative dataset is essential for accurate estimates in data analysis and ML. to avoid suboptimal decisions and unreliable predictions [252]. According to Rubin [317], missing data are usually classified into three distinct models [341]: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR):

- MCAR**:** missing data are independent of any other value in the dataset. The probability of a value being missing is not related to observed and unobserved data;
- MAR**:** the missingness is related to the observed data but not to the missing data itself. In other words, the probability that a value is missing depends on other available information;
- MNAR**:** this type occurs when missing values depend on the unobserved data themselves.

A simple but commonly used approach to handle missing data is the list deletion method, which involves deleting samples containing incomplete data. However, although this method is simple, it can result in significant loss of valuable statistical information and reduce precision, especially in complex multivariate analyses[225]. To overcome this problem, several advanced imputation methods have been developed to address this problem more effectively[333]:

- Regression methods that predict missing values based on their relationship to other variables in the data set;
- K-Nearest Neighbors (KNN), which identifies "k" neighbors (data points) for a given instance with missing values and imputes the missing data based on the mean (or other aggregate measure) of the neighboring instances. The KNN method is beneficial when the missingness model is MAR, where dependencies between features are present;
- Deep learning approaches, that use multilayer perceptrons based models such as autoencoders, Generative Adversarial Network (GAN) and recurrent neural networks to learn complex patterns within the data and accurately predict missing values. These methods are very effective when working with high-dimensional datasets, capturing nonlinear relationships[236] and various missing data patterns[73];
- Sophisticated statistical methods, that combine ML models with advanced statistical techniques, such as Bayesian inference, Expectation Maximization (EM) and multiple concatenated equation imputation (MICE), to provide more accurate and reliable imputations[254]. This approach help account for uncertainty and variability in the imputation process, providing robust solutions[299,307].

2.2.2 Cluster ensemble

The primary objective of clustering techniques is to optimize a global objective function defined by specific criteria, such as the similarity or distance between data points[196,334]. This process involves partitioning a dataset into clusters, where the quality of the clustering result is evaluated using the chosen objective function. While several approaches to clustering, including heuristic methods, seek acceptable solutions, most clustering algorithms tend to converge only to a local optimum rather than the global optimum. However, in some cases, such as when clusters are well-separated and of equal size, clustering algorithms like k-means can converge to a global optimum[196]. Selecting a single clustering algorithm that performs effectively across all datasets is challenging due to the variability in data structures. Consequently, numerous clustering algorithms have been proposed[233,336].
To address the limitations of relying on a single algorithm, researchers introduced the concept of consensus clustering, or cluster ensembles, which tries to overcome the limitations of any individual algorithm but employing several of them together at the same time and then combining the results to offer a superior solution[346].
The operation of putting together the results of the other clustering algorithms and using them for the creation of a final partition of the data, is performed by the consensus function.
This consensus function in practice takes the initial dataset and divides it into k clusters, where hard partitioning must satisfy two conditions: (1) each cluster must contain at least one point, and (2) each point must belong to exactly one cluster. Consensus clustering allows for integrating various clustering strategies without requiring access to the specific algorithms or features used in the initial clustering processes[161]. Ensemble clustering aims to merge multiple base clustering into a single, more accurate, and robust clustering outcome. A significant number of ensemble clustering approaches is based on graph partitioning methods[174], the idea is to transform the hard clustering task into a more manageable community detection problem on a network. Community detection algorithms look for groups of nodes on networks that are "related to each other" in some sense and identifies them as a community. More detailed information about how one famous community detection algorithm, the Louvain algorithm, works will be presented in Section 2.2.2. In order to exploit these powerfull community detection algorithms, it is necessary to create a graph by taking the initial data points and the clusters produced by the clustering algorithms and interpret them as nodes of a bipartite graph. Edges between these nodes are then subsequently created based on information derived from the base clustering algorithms[127]. This graph-based approach takes advantage

of the structural properties of the network to aggregate and refine the clustering algorithms, often leading to more coherent and meaningful clusters.

Louvain

One recently developed method for constructing graph partitioning is based on Louvain's algorithm[47].
It optimizes a quantity called modularity, a measure that quantifies the quality of the division of a network into communities.
The main steps are the following:

1. **Initial Community Assignment:** Each node in the network is initially assigned to its own community. Therefore, at the beginning, the number of communities equals the number of nodes;

2. **Modularity Optimization:** For each node in the network, the algorithm evaluates the modularity gain resulting from moving the node from its current community to the community of one of its neighbors. The modularity is calculated to determine if the movement would improve the overall modularity of the network. If modularity gain is positive (indicating an increase in modularity), the node is moved to the neighboring community with the highest modularity gain. Differently, if the modularity gain is negative, the node remains in its original community;

3. **Community Aggregation:** After all nodes have been evaluated and reassigned, the algorithm aggregates the nodes within the same community into a single "super-node", creating a new, smaller network where each node represents an entire community;

4. **Repeat:** The algorithm then repeats the process on this new network, merging and optimizing communities until no further modularity gains can be achieved.

2.2.3 Ensemble learning algorithms and spatial heterogeneity

Ensemble learning algorithms, like GTB, have gained widespread popularity in ML because they effectively improve predictive accuracy by aggregating the outputs of multiple base learners. One of the most notable algorithms in this category is XGBoost. XGBoost is a performant, robust and flexible algorithm for handling various data types, including high-dimensional and sparse datasets[71]. It fuses the principles of gradient boosting with regression trees to create a highly efficient and scalable predictive model. Additionally, XGBoost provides valuable insights into feature

importance by constructing a feature importance score F related to the number of times a particular feature is used to split the trees, enabling researchers to identify the most significant variables influencing the prediction outcomes. The core mechanism of XGBoost lies in its iterative approach to model building. The algorithm builds its predictive model by adding weak learners one after the other(usually decision trees)where each new model focuses on correcting the mistakes of the previous ones. This iterative process is guided by a gradient descent optimization method to minimize a chosen loss function. At each iteration, the residuals are calculated and used to fit the next model in the sequence. After the sequential application of all trees, XGBoost combines the results, producing a final prediction. Hence, this method ensures a gradual improvement of the model's accuracy and reduces the overall error in the final prediction, further reinforcing its reliability[139]. The algorithm's effectiveness is further enhanced by techniques such as regularization, which helps prevent overfitting, and tree pruning, which optimizes the complexity of the model[70].

The XGBoost algorithm has also been applied to handle spatially heterogeneous data. For example, XGBoost has been implemented in ecological modeling, where the algorithm outperformed traditional methods in predicting species distribution across heterogeneous landscapes[235]. To solve the problem of spatial heterogeneity, the economic and geographic literature refers to the concept of spatial contiguity, which requires that objects within a cluster or region be spatially connected or tightly grouped[364]. This notion is essential in distinguishing regionalization techniques from conventional clustering or classification algorithms, which may not consider spatial relationships[158]. Spatial contiguity can be categorized into two main types. The first is hard spatial contiguity, which implies that the objects within a class (or cluster) are geographically connected. This contiguity ensures that the resulting regions are contiguous and can be easily recognized geographically. Soft spatial contiguity allows for some flexibility in the spatial connectivity of units within a cluster. Soft contiguity can be measured by the similarity between geographical coordinates, allowing for some separation between units within the same cluster[69,212].

2.3 Methodology

In this section, all the techniques and methodologies that have been implemented will be presented. First, the problem of imputation of missing data in the dataset will be given. Then, it will show how several clustering algorithms have been trained

and how they can be combined in an ensemble to produce a single better partition of the dataset. Afterward, a simple Explainable ML strategy will be displayed to improve the explainability of the clusters obtained from the ensemble partition, making the meaning of the clusters more transparent and what characteristics define each cluster. Lastly, these clusters will be used as strata in a stratified bootstrap algorithm to compute the distribution of statistical quantities of interest.

2.3.1 Data imputation

The algorithm used for data imputation in this work is called Iterative Imputer, which is a Python implementation of MICE method originally developed in R. The idea behind MICE is to treat the missing data problem as a sequence of regression tasks, where missing values are iteratively estimated based on the observed data. Specifically, MICE at the beginning initializes the missing data with a placeholder like the mean. For each feature with missing values, a regression model is subsequently trained using the other observed features as predictors. The missing values are then updated with predictions from this model. This process is repeated for each feature with missing values in a sequential manner, cycling through the dataset multiple times until the imputations stabilize and further updates to the missing values no longer significantly change.
The Extremely Randomized Trees (Extra Trees) regressor was chosen as the base estimator for this iterative imputation process due to its ability to handle nonlinear relationships and interactions within the data. Extra Trees is an ensemble learning method that builds multiple decision trees, but unlike standard random forests, it introduces additional randomness by selecting split points at random rather than optimizing split criteria. This tends to make the overall model lighter and also reduced variance, allowing the model to generalize well and avoid overfitting. All these steps are extremely computationally intensive and therefore should be used carefully in large datasets, but they are perfectly fine for use cases like the one presented in this work where a relatively small dataset is available, and it is important to extract the maximum amount of information from the available data.

2.3.2 Clustering algorithms

Several clustering algorithms have been used thanks to their their Scikit-Learn[291] implementation:
AffinityPropagation, AgglomerativeClustering, balanced iterative reducing and clustering using hierarchies (BIRCH), DBSCAN, HDBSCAN, KMeans, BisectingK-

Means, MiniBatchKMeans, MeanShift, Ordering points to identify the clustering structure (OPTICS), SpectralClustering and GaussianMixture.
This wide range of algorithms is based on different clustering strategies, therefore ensuring the heterogeneity of the clusters that will be used in the creation of the *Cluster Ensemble*.
Some metrics have been evaluated to assess the performance of each clustering algorithm: the Silhouette score, the Davies Bouldin score, the Calinski Harabasz score with their Scikit-Learn[291] implementations and also the S_Dbw score[159]. The best combinations of the Hyperparameters for each model have been selected by doing a grid search (i.e., trying each hyperparameter combination for each algorithm and taking for each algorithm just the combination with the highest score) to maximize the Calinski Harabasz score. The Hyperparameter configurations that have been tried are shown in Table 2.1.

The procedure is time-consuming but ensures that the partition chosen is one of the best possible for each clustering algorithm. It does not ensure that every algorithm is adapted for this particular dataset. Therefore, it is essential to look at the combination of all the evaluation metrics to assess the actual quality of each one of the clustering solutions and eventually discard the algorithms that produce unsatisfactory results. Once the algorithms have been trained and the different alternative clustering solutions have been created, it is possible to use an ensemble strategy to find a synthesis and unify the various clustering algorithms in a single superior clustering strategy.

2.3.3 Consensus via Hybrid Bipartite Graph Formulation (HBGF) + biLouvain

An ensemble strategy called HBGF[127] has been implemented to exploit the clustering algorithms implemented. The algorithm consists of creating a bipartite graph starting from the results of the clustering algorithms and then applying a Community detection algorithm like biLouvain on the graph's nodes as a consensus function to obtain a single better clustering solution.
Let's assume we are given a data set $X = \{X_1, ..., X_N\}$ of N rows. A *cluster ensemble* is a set of cluster solutions $C = C^1, ..., C^R$ where R is the number of clustering algorithms. Each clustering solution C^r is a partition of X, $C^r = \{C^r_1, ..., C^r_{K_r}\}$ such that $\cup_k C^r_k = X$ where K_r is the number of clusters identified by the clustering solution C^r and therefore $K_{tot} = \sum_r K_r$ is the total number of clusters created by all the solutions. Starting from this *cluster ensemble* the objective is to find an improved partition of X into K disjoint clusters, beginning with the clusters in C. It can be

Algorithm	**Hyperparameters**
MiniBatchKMeans	n_clusters: [7, 8, ..., 29, 30]
Kmeans	n_clusters: [7, 8, ..., 24, 25]
AffinityPropagation	damping: [0.5, 0.6, 0.7, 0.8, 0.9] preference: [-50, -45, ..., 40, 45]
AgglomerativeClustering	n_clusters: [7, 8, ..., 39, 40]
GaussianMixture	n_components: [10, 11, ..., 39, 40] covariance_type: ['full', 'tied', 'diag', 'spherical']
BisectingKMeans	n_clusters: [7, 8, ..., 24, 25]
Birch	n_clusters: [17, 18, ..., 35, 36] threshold: [0.08, 0.081, ..., 0.099, 0.100] branching_factor: [2, 3, ..., 13, 14]
DBSCAN	eps: [0.3, 0.4, ..., 0.7] min_samples: [2, 3, ..., 18, 19] metric: ['euclidean', 'manhattan', 'chebyshev', 'minkowski'] algorithm: ['auto', 'ball_tree', 'kd_tree', 'brute'] leaf_size: [10, 20, 30, 40, 50] p: [3, 4, 5]
SpectralClustering	n_clusters: [17, 18, ..., 35, 36]
HDBSCAN	cluster_selection_method: ['eom', 'leaf'] min_cluster_size: [2, 3, 4] min_samples: [2, 3, 4]
MeanShift	cluster_all: [True, False]
OPTICS	min_samples: [5, 6, ..., 19, 20]

Table 2.1: List of the algorithms that have been tried and the some of the hyperparameters that can be fine-tuned. A grid search optimization strategy can been used to find the best hyperparameters configuration for each model with respect to one of the scores.

done using the aforementioned HBGF.

To define a graph, we create the associated adjacency matrix $A_{i,j}$ where $i = 1, ..., N$ corresponds to the i_{th} data point in the dataset and $j = 1, ..., K_{tot}$ corresponds to one cluster in the set of clusters C.

The matrix element $A_{i,j}$ is 1 if the data point belongs to the j_{th} cluster. Given that each clustering solution C^r puts each data point in precisely one cluster and has R clustering solutions, it follows that $\sum_{j=1}^{K_{tot}} A_{i,j} = R$. Once a bipartite graph has been created it is possible to use the biLouvain Community detection algorithm[293] to identify K Communities of data points, which are the new clusters. The algorithm also allows to identify communities among the clusters produced by the various clustering algorithms, effectively allowing to study analogies between the various clustering strategies.

The main differences between the biLouvain algorithm and the original Louvain algorithm presented in the section 2.2.2 consist of:

- the algorithm Murata+[293] used to calculate the modularity, which tells the strength of the division of the graph into communities, and is different from the one used in the Louvain algorithm because it must take into account the constraint that connections are possible only between points in the dataset and clusters in the cluster set C. Therefore, connection between data points and clusters are forbidden;

- the way movement between nodes is performed, and the candidate communities is proposed because in the Louvain algorithm, in principle, any node can be moved to any community. However, given the bipartite nature of the graphs, only certain moves are permitted. For instance, a data point can be moved to a target community only if there is at least one node in the target community that has a connection to a cluster connected to the data point that has to be moved (community of neighbors of my neighbor);

- finally, given that the modularity function has been changed, the way the modularity gain is computed must also be changed; it is important to stress that only moves that increase the modularity are accepted.

Some of the advantages that the biLouvain algorithm inherited from the Louvain algorithm are its scalability to large graphs and its automatic determination of the number of communities. The communities of data points can then be used as a final clustering partition of the data that considers the results of the clustering algorithms used to create the graphs.

2.3.4 Cluster explainability with XGBoost

One common issue of clustering algorithms is that once the final clusters have been identified, it is challenging to understand what kind of properties the data points within each cluster share and what features played a crucial role in determining the clusters. This is a common issue with all ML algorithms and is called model Interpretability loss (i.e., it is impossible to infer a causal relation between the input and the output of the model). Fortunately, some techniques have been developed in recent years to produce explainable models, allowing for determining which input features are more important for the output computation. Most algorithms designed to enhance the Explainability of models are specific for supervised tasks like regression and classification. A frequent practice to extend these algorithms to unsupervised

tasks like clustering is to use one *ex post* to understand what the clustering strategy is doing.
It is done in practice by interpreting the features in the data set X as independent variables and the cluster labels created by a clustering algorithm as dependant variables y. Hence, it becomes a multi-class classification problem with K classes, one for each cluster. This approach allows using one ML algorithm as an effective way to learn the relationship between the X and y and, therefore, discover the logic that the clustering algorithm used.
This classification problem can then be studied using state-of-the-art algorithms like XGBoost. It is an ensemble tree algorithm that has built-in explainability algorithms. For instance, this model automatically computes a feature importance score F by looking at how often a particular feature appears in decision trees while training the model. This feature importance score allows us to understand which features were used the most by XGBoost to maximize the accuracy of the predictions and, therefore, which features were more impactful in creating the clusters themselves.
The necessity to introduce feature importance in decision trees, particularly in the context of clustering interpretability, comes from its ability to extract meaningful insights about the decision boundaries implicitly learned by the clustering algorithm. Since the classification model attempts to predict the cluster assignments based on the original features, the importance scores derived from decision trees reveal which features contribute most to the separability of clusters. This approach has been widely adopted in ML interpretability studies[260], where tree-based models are used as post hoc explanatory tools.

2.3.5 Spatially-stratified bootstrap

When a satisfactory and interpretable clustering strategy is established, it can be effectively utilized to extract meaningful insights from the data. Clusters have been used as strata for a stratified bootstrap algorithm. A bootstrap algorithm consists of extracting repeated sub-samples of the dataset and computing the statistical quantity of interest, like an average or a correlation coefficient, on these sub-samples. The procedure is then iterated a fixed number of times, and all the quantities computed are then put together to obtain a distribution.
The stratified variant of the bootstrap[212] is particularly adapt for spatial data, such as firms located in a region, and identified by the geographical coordinates latitude and longitude, because it allows for the problem of spatial dependence between statistical units, such as spatial autocorrelation, to be taken into account.
Extraction with repetition is not performed from the complete dataset but from

predetermined sub-samples called strata that are comparatively uniform within and distinct from each other. The K clusters produced by HBGF are taken as the strata for the stratified bootstrap.
The connection between geographic and attribute spaces is a critical element of our algorithm. When the geographic space, representing the spatial relationships, changes, the attribute space, which encapsulates the characteristics of these entities, also shifts. Hence, grasping the interplay between space and features of the enterprises, such as economic activity, productivity, and employees, is essential to understanding the algorithm's functionality and effectiveness[214].

2.4 Case study: innovative startups in Lombardy

In this paragraph the algorithm previously presented will be applied to a dataset of innovative startups from the Italian Lombardy region.

2.4.1 Dataset

The innovative startups list was extracted from an Italian Chambers of Commerce BR special section[181]. According to the Startup Act of 2012, a national law designed to foster the development of a startup ecosystem, innovative startups are defined as companies that meet specific criteria, including having fewer than 249 employees, annual revenues below 5 million euros, and not distributing dividends. Additionally, they must satisfy at least one of the following conditions: i) they allocate at least 15% of the greater value between operating costs and revenues to research and development annually; ii) at least one-third of their employees are doctoral or graduate students, or at least two-thirds of the workforce holds a master's degree; iii) they possess or lease a patent, trademark, or registered software. The BR is an administrative database maintained by the Chambers of Commerce, where company registration is mandatory under Italian law. However, the special section for innovative startups relies on self-declarations by firms, which may introduce potential biases, such as the over-reporting of eligibility criteria to gain tax benefits and public support. Additionally, the register does not track startups that fail to register or those that do not meet the formal requirements but still engage in innovative activities. The database covers several demographic information about the companies, such as the year of establishment, year of registration, sector of activity, class of employees, turnover classes, and the profile of the shareholders type of startups. To enrich the database information, we matched the innovative startups with the firms of the ASIA-Istat[182] database by tax code.

In Figure 2.1 the number of new startups recorded in the Innovative startup section of the BR by year of registration is shown for some selected regions. The Lombardy region has the highest number of new innovative startups in the country each year.

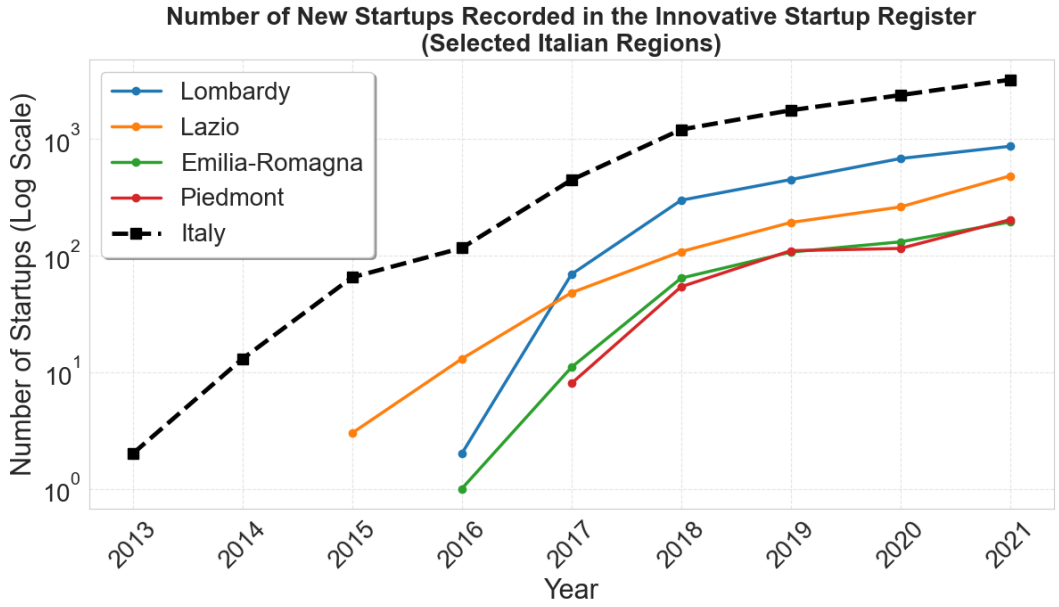


Figure 2.1: New innovative startups recorded each year in the innovative startups section of the BR by year of registration. Note that each region started having firms recorded in the register in different years, here only the four regions with the highest number of new innovative startups in the year 2021 are shown. The Lombardy region is the driving force of italian innovation.

Data preprocessing

We focused our analysis on the last three available years, 2019, 2020, and 2021, to see if there are significant differences between these years. We merged two datasets: the ASIA databases and the special section of the BR of Innovative Startups. We filtered the ASIA databases only for businesses in the Lombardy region that were active in the three years considered consecutively. Subsequently, given that each business can be an innovative startup for a maximum of five years, we discarded all the signed companies in the startup registry earlier in 2017. We then merged the two databases, looking at the tax code number as an univocal identifier of the business. It led to the creation of a dataset of 517 firms. The categorical variables in the dataset have been one-hot encoded to transform them into numerical variables. Many variables in the dataset were redundant and highly correlated; therefore, all the variables with a Pearson correlation higher than 0.8 and lower than -0.8 were removed.

This dataset had some missing values, so the imputation technique presented in2.3.1

has been implemented. Subsequently, the data were scaled in the [0, 1] range, and a dimensionality reduction technique called Kernel Principal Component Analysis (KPCA)[327] was applied.

2.4.2 Clustering of the startups dataset

Using a grid search approach, all the clustering algorithms presented in 2.3.2 have been trained on the dataset, and the hyperparameter configurations that maximized the Calinski Harabasz score have been saved, as shown in Table 2.2. The algorithms that performed significantly worse than the others have been discarded to avoid possible negative impacts on the *cluster ensemble*. From the remaining nine clustering algorithms, the HBGF strategy can be implemented. The bipartite graph can be created by identifying the associated adjacency matrix $A_{i,j}$ associated with the bipartite graph is created as follows: each row represents one of the 517 businesses, and each column is a cluster of one of the 9 cluster algorithms for a total of 93 clusters (7+7+5+7+10+7+18+2+30). The adjacency matrix is shown in Figure 2.2.

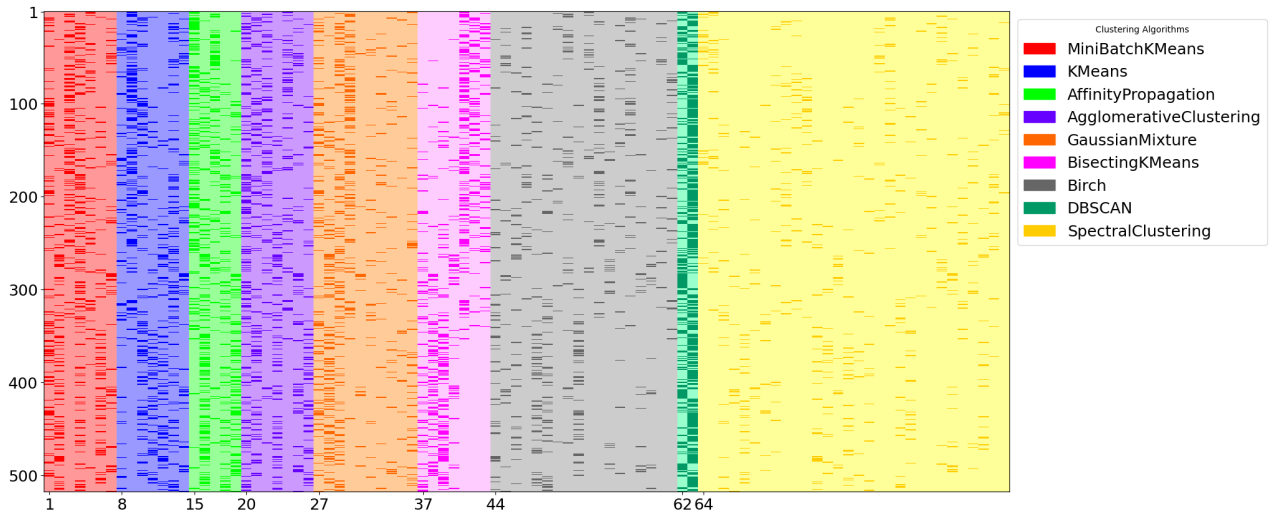


Figure 2.2: Adjacency matrix associated to the clustering ensemble, each row is one data point and each column is one cluster. Each column is colored according to its clustering algorithm. Each data point belongs to one cluster for each cluster algorithm (hard assignment), therefore in each row there must be 9 dark lines.

After the bipartite graph is built, the biLouvain algorithm is applied to identify 5 communities for the rows and 5 for the columns. This process achieved a Murata+ Modularity score of 0.7065, reflecting the effectiveness of the partitioning[293].
The row communities represent the new partition of the dataset that is the output of the *clusters ensemble*, and therefore $K = 5$, indicates that the startups can effectively be partitioned into five groups. The column communities indicate that even though

the nine clustering algorithms followed different clustering strategies, their outputs are similar and that many clusters are composed of the same startups[1]. The bipartite graph is presented in Figure 2.3.

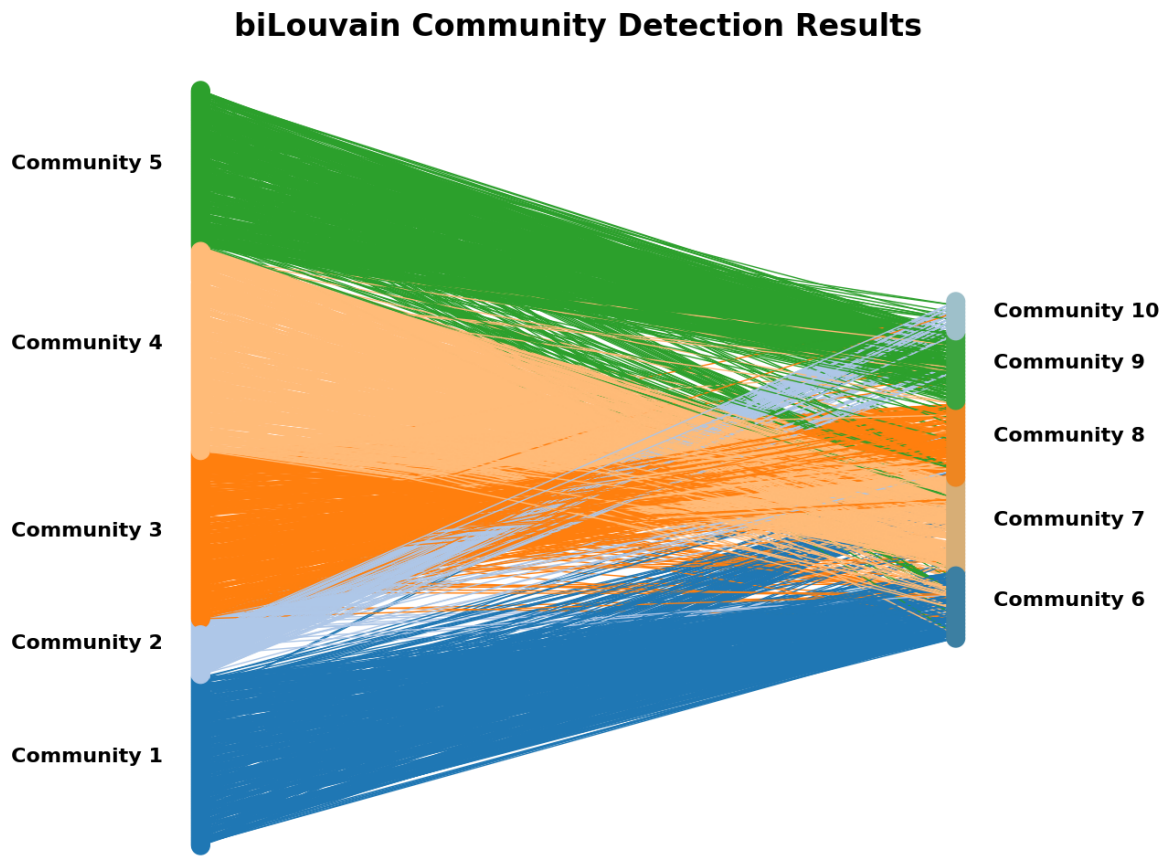


Figure 2.3: Bipartite graph where the data points corresponding to the nodes on the left and the clusters corresponding to the nodes on the right, are sorted and colored according to their community. The communities on the right in the bipartite graph have been colored starting from the colors of the communities on the left: the color on the right is the weighted average of the colors on the left, where the weights are given by the number of edges that connect each left community to the right community. For instance, this means that the community 6 is blue because the majority of edges that are connected to the community 6 start from the community 1. This highlights the fact that there's almost a 1 to 1 correspondence between the communities on the left and the communities on the right. Given the high number of edges, the edges have been randomly subsampled at the end to make the graph more readable.

[1]To verify the robustness of this approach, another algorithm not based on biLouvain has been tried. This other approach is presented in A and produces results analogous to those of biLouvain.

Algorithm	Silhouette	S_Dbw	Davies Bouldin	Calinski Harabasz	Best parameters	Number of clusters
MiniBatchKMeans	0.1777	0.7354	1.7655	67.6914	7	7
KMeans	0.1705	0.7660	1.8619	64.5171	7	7
AffinityPropagation	0.1446	0.8069	1.9936	63.2718	0.8, -10	5
AgglomerativeClustering	0.1452	0.7392	1.9295	60.2850	7	7
GaussianMixture	0.1557	0.7091	1.9578	52.5263	'tied', 10	10
BisectingKMeans	0.1303	0.7909	2.2347	49.2502	7	7
Birch	0.2105	0.6155	1.5526	48.5648	9, 18, 0.08	18
DBSCAN	0.1147	0.9950	3.1690	48.2844	'auto', 0.3, 10, 'minkowski', 17, 3	2
SpectralClustering	0.2651	0.4699	1.2412	45.0880	30	30
HDBSCAN	0.2774	0.2986	1.3187	21.5558	'eom', 4, 2	64
MeanShift	0.1613	1.0326	2.3974	21.3962	False	2
OPTICS	-0.0372	0.3239	1.4044	19.4423	16	7

Table 2.2: Clustering Algorithm, associated Evaluation Metrics, Best hyperparameters, and number of clusters each algorithm produced.

2.4.3 Explainable Machine Learning

Once the K = 5 clusters are created from the dataset, it becomes challenging to interpret the specific characteristics and attributes each cluster represents concerning innovative startups. This is a problem because, in practice, clustering algorithms behave like black boxes, where it is unclear how the clustering algorithm has used each feature in the dataset and what features have had the most impact on the clusters. The complexity of this problem is heightened when the number of features is relatively high, rendering an individual examination of each feature impractical. A multi-class classifier XGBoost algorithm has been employed to unravel the logic behind the clustering algorithm. This powerful tool has been trained on the dataset using the community labels as the target independent variable. The 10 variables out of the 125 used that had the highest importance score are shown in Figure 2.4 and are described in detail in Table 2.3 providing a clear insight into the clustering logic.

Given these key features, they can be compared with the clusters themselves to gain a deeper understanding of the types of startups each cluster comprises. In Figure 2.5 the distribution of the R&D requirement of firms is shown for the different clusters, highlighting how the clustering technique implemented gave high relevance to R&D when partitioning the startups. Figure 2.6 shown the distribution of clusters among the different activity sectors.

Within each cluster, the startups were further split into two groups: those possessing a certain characteristic and those lacking it. For numerical features, this split was based on whether the feature value was above or below the median. The count of startups in each group was then divided by the total number of startups in the cluster. This process was repeated for each feature across all clusters. Finally, the results were normalized across clusters so that the sum of each row was equal to

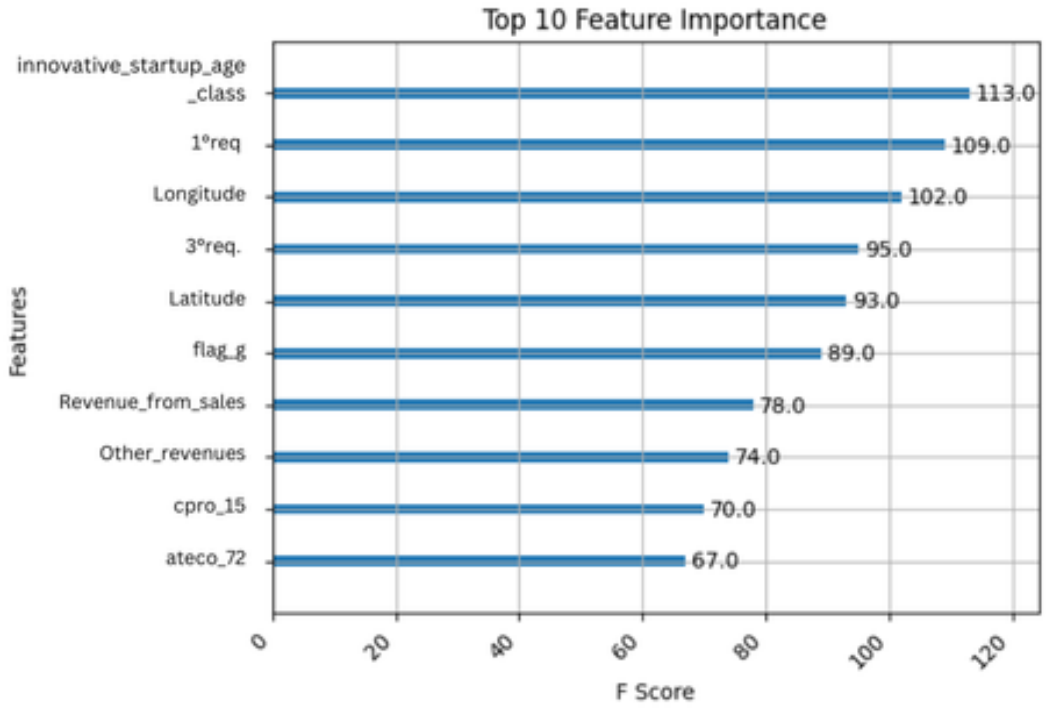


Figure 2.4: The features with the highest Feature Importance score are the business age class, R&D expenditures, longitude, employ qualification, latitude, adherence to a companies group, revenues from sales, revenues from sources other than sales, the startups is in the regional capital Milan and the startup's main activity is scientific research and development. All the other features have lower F score and the complete list of features can be found in B.

1, facilitating a clearer comparison of feature distributions across the clusters. The results are presented in Figure 2.7, where the clusters exhibit distinct characteristics that help define them.

Using the results presented in Figure 2.7 as a guide, it is possible to give an interpretation of the clusters:

1. **Established R&D Collaborators**
   These are slightly older startups (Age class 2) that invest heavily in R&D but do not patent. They do not carry out "scientific research" as their primary activity despite high R&D spending, suggesting R&D spending might be more applied or development-focused. They belong to some form of group/consortium, have higher sales revenues, and operate within Milan in non-manufacturing sectors.

2. **Young Manufacturing Patent-Holders Outside Milan**
   Younger firms (Age class 1) that have low R&D but do own patents—they may be acquiring Intellectual Property (IP) in a more defensive/strategic way, rather than by large R&D spend. They are manufacturing startups, located

Feature name	Feature description
Innovative_startup _age_class	1 if the innovative startup has less than 3 years, 2 if the startup has 3-5 years
1°req	Flag that is 1 if the firm has R&D expenditures equal to at least 15% of the higher of cost and total production value, 0 if not (some variables are called requirements because a firm must have at least one of them to be officially considered an innovative startup, for instance if a business has low R&D investments and no patents, it must demonstrate to have a highly qualified personnel)
Longitude	Longitude of the innovative startup (geographical coordinates)
3°req.	Flag that is 1 if the firm is the owner, depositary or licensee of at least one patent or holder of registered software, 0 if not
Latitude	Latitude of the innovative startup (geographical coordinates)
flag_g	Flag that is 1 if the innovative startup belongs to a business group as defined by Istat (a collection of enterprises connected through legal and/or financial ties, with overall control exercised by a group head. Such groups can have multiple decision-making centers, especially for key decisions related to production, sales, and profits, and may centralize functions like financial management and taxation. They operate as a single economic entity, capable of making strategic choices that impact all its member units[118]), 0 if not
Revenue_from_sales	Current revenues from sales (in Euros) excluding Value Added Tax (VAT), gross of indirect taxes
Other_revenues	Revenues (in Euros) from sources other than sales like royalties from patents and rental of goods owned by the startup
cpro_15	Flag that is 1 if the administrative headquarters of the startup are located within the province of Milan, 0 if not
ateco_72	Flag that is 1 if the main activity of the startup is scientific research and development, corresponding to European Nomenclature of Economic Activities (NACE) code 72

Table 2.3: Description of features that have the greatest impact in the analysis of startups, a description of all the other variables used can be found in B.

outside Milan, and tend to be standalone (not belonging to a group). They show relatively low sales but higher "other" revenues (potentially licensing, non-operating income, or external funding).

3. **Young Research-Driven Service Startups (Low R&D Spending)**
   Young startups (age class 1), that do scientific research as a main activity but spend relatively little on R&D, perhaps they conduct early-stage, mostly intellectual/academic research. They do not own patents, are not in manufacturing, and do not belong to a formal group. They have borderline high sales vs. low other revenues, indicating they might be generating sales from specialized

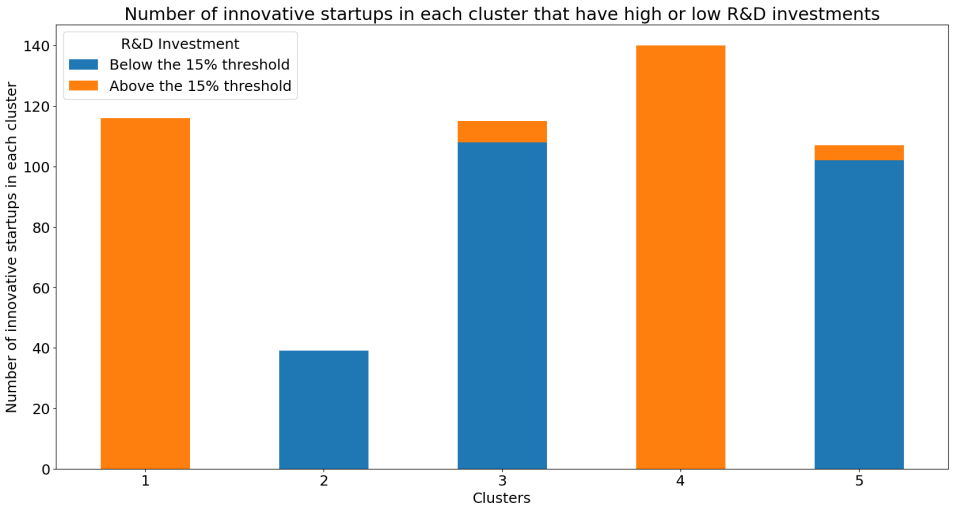


Figure 2.5: Distribution of the R&D first requirement among the different clusters. Clearly the first and forth cluster of firms are characterised only by firms that have an R&D investment above the 15% threshold. On the other hand, firms in the second, third and fifth cluster are almost exclusively composed of firms under the threshold.

research services or products without large overhead R&D investment

4. **High-R&D Research Startups (Low Sales, High External Funding)**
   Young research-focused firms that spend a lot on R&D but do not hold patents. Possibly they rely heavily on external funding or grants (hence "high other revenues" vs. "low sales"). They are not in manufacturing and are standalone (do not belong to a group).

5. **Young Patent-Owning Collaborators in Milan (Diverse Revenues)**
   Young patent-owning startups in Milan, apparently research oriented (scientific research as main activity) but with low R&D investment levels, suggesting they may have developed or licensed IP early. They belong to some group (possibly an accelerator or consortium). They also have both high sales and high other revenues, suggesting multiple revenue streams.

The spatial aspect is paramount in our analysis, so Figure 2.8 shows the spatial distribution of the startups. According to Figure 2.7 and Figure 2.8 innovative startups belonging to Cluster 2 are located outside the province of Milan. Cluster 2 mainly contains manufacturing companies. In contrast, the companies that belong to Cluster 5 (which includes startups with high patent ownership and are part of a group) are located within the province of Milan and are mostly advanced service enterprises.

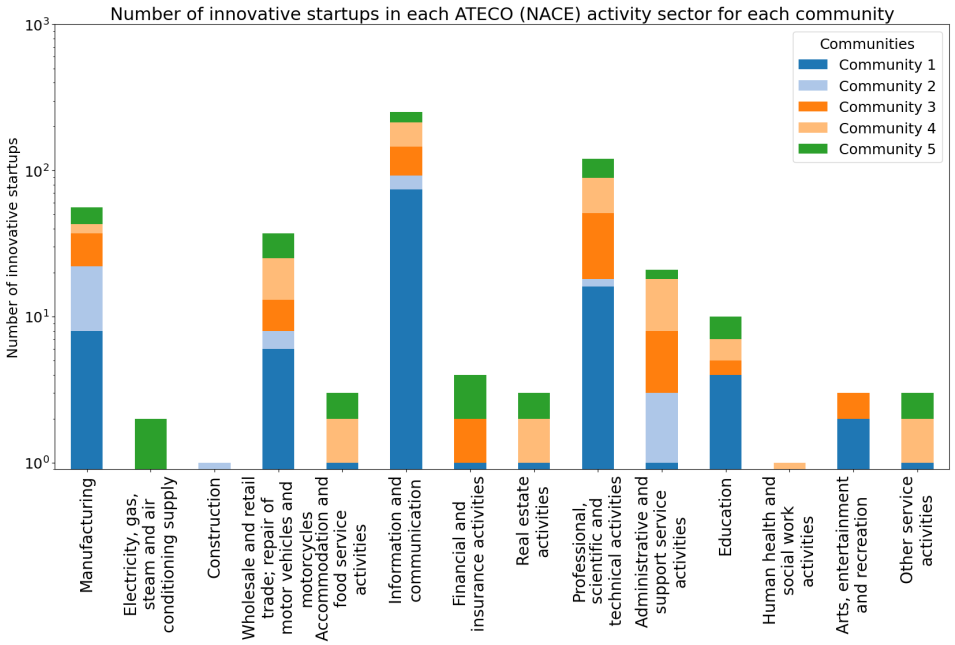


Figure 2.6: Distribution in logarithmic scale of the number of firms in the different communities for each activity sector. The activity sectors have been computed looking at the 1-digit Italian classification aligned with NACE (ATECO) code of the firms. It is possible to see that Community 1 is prevailing in the sectors where most startups are located, on the other hand other Communities are prevailing in smaller sectors. Like the "Electricity, gas, steam and air conditioning supply" startups which are all in Community 5.

2.4.4 Spatial Bootstrap

The clusters produced will be used as layers of a bootstrap stratified algorithm to calculate the distribution of average firm LP in the years 2019, 2020, and 2021.
For each firm, LP was calculated as the ratio of value added and the number of employees. In Figure 2.9 the distribution of the LP in each Community is shown. The different distributions are approximately normally distributed but they have different variances, in particular Community 4 has larger variances due to the presence of outliers.

Once a bootstrap sample is obtained by sampling from the strata, the average LP of the startups in the sample is calculated and added to the distribution. This procedure is then iterated to obtain the entire distribution. The results are shown in Figure 2.10.
The averages calculated using the bootstrap strategy for 2019 and 2021 are slightly higher than those calculated over the entire sample, while the bootstrap average for 2020 is closer to the average calculated over the total sample. The bootstrap average, thanks to resampling, is much more resilient to the presence of outliers that the cluster distributions. Generally, the averages tend to increase yearly, indicating that startups

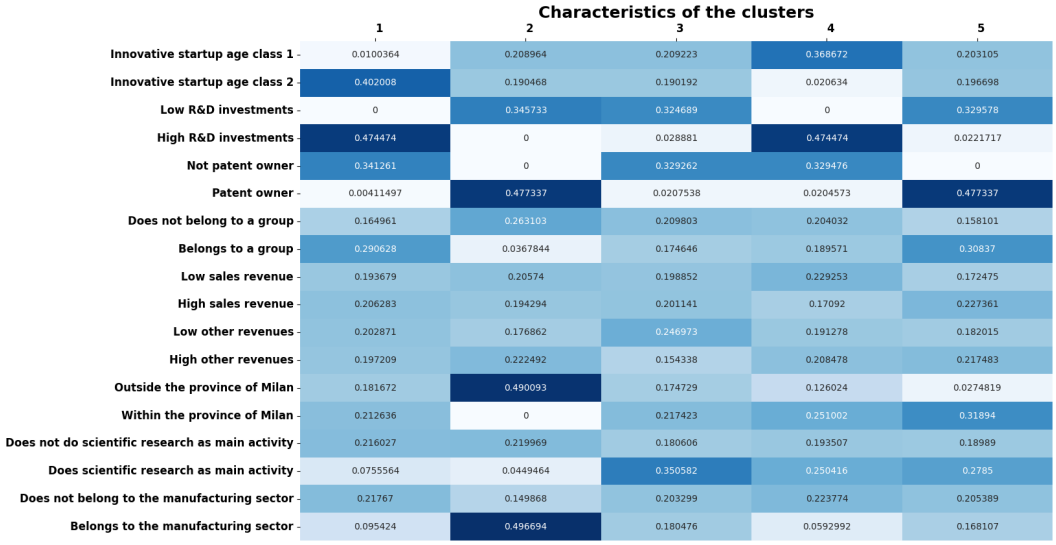


Figure 2.7: Heatmap of the characteristics highlighted by the Feature importance score vs the clusters. The sum of each row equals 1; therefore, in the case of perfect equipartition, each value should be 0.2. The values that are much larger than 0.2 should be considered as defining characteristics of the cluster.

grow even under severe economic constraints, such as in 2020 when the Coronavirus Disease 2019 (COVID-19) pandemic caused companies to lock down production activities. The impact of this restriction must be considered as it affected the shape of the distribution of the average LP. Although average LP in 2020 increased compared to 2019, the distribution is much broader, with much heavier tails on both the left and right, indicating more significant variability in startups performance. Some companies had to close temporarily or could not adapt their business model to the situation, and their revenues declined sharply. This is one possible explanation for the heavy left tail. On the other hand, interpreting the heavier right tail is not trivial. Some startups have quickly adapted their business models to the constraints imposed by their investment in digital and technology-driven companies and have seen their revenues increase. It has contributed to increased LP in 2020 for these types of companies. It could explain the shift even further to the distribution's right tail. However, most firms on the right side of the distribution did not suffer recessionary effects caused by the lockdown and continued to generate revenues. Therefore, the average LP distribution in 2020 shifts slightly to the right. After the pandemic, when most restrictions were removed in 2021, average LP increased not because of a further push toward the right side of the distribution by the best-performing startups but rather because of a decrease in the left tail. This means that the startups that survived the pandemic could return to their regular activities when the restrictions were removed and recovered the output gap compared to the innovative startups that

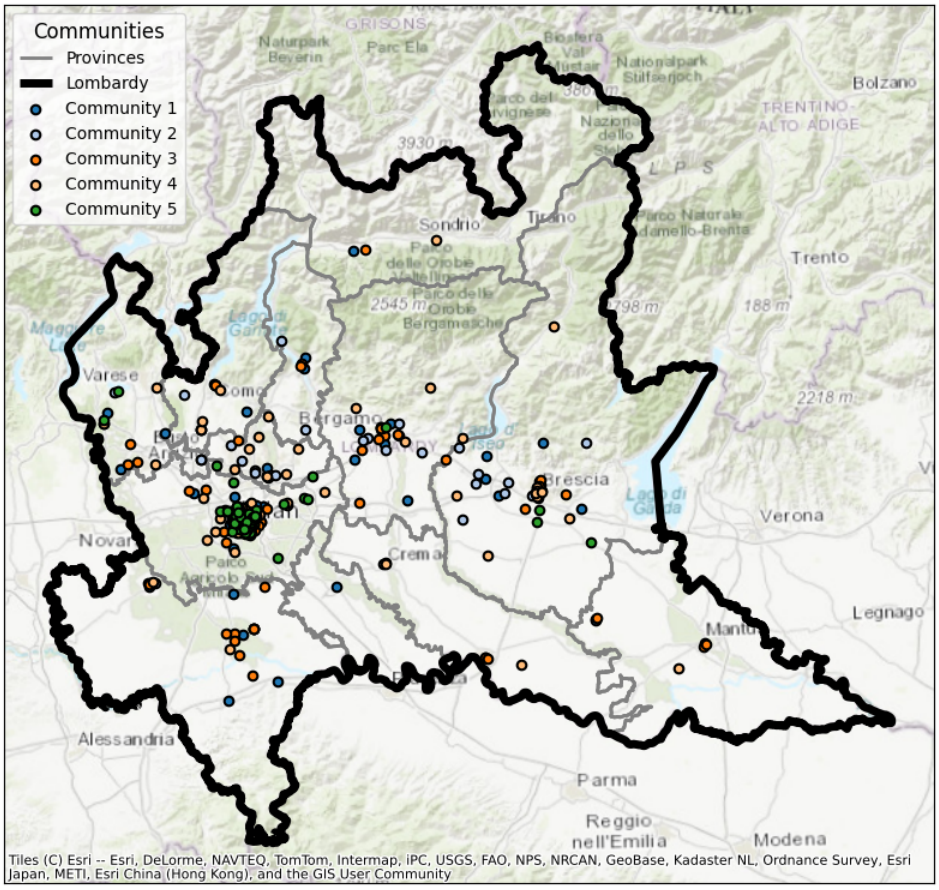


Figure 2.8: Spatial distribution of the startups in the Lombardy region: each startup is colored according to its Cluster. It is possible to see the spatial aggregation of startups around the regional capital, Milan, in the west. A clustering strategy based solely on geographical aspects would not be able to disentangle this agglomeration and distinguish effectively within Milan.

suffered less from the recessionary effects of the pandemic.

2.5 Conclusions

The new generation of innovative businesses is not just a possibility but a promising reality. They are poised to achieve rapid growth, international recognition, a culture of open innovation, and cross-sector collaboration[86]. This approach, which promotes creative entrepreneurship, technology transfer, and market competition, is designed to accelerate the industrial revolution and boost production levels, competitiveness, and efficiency in the entire economy in the long run[286]. Entrepreneurship catalyses innovation, productivity enhancement, and significant job creation. Despite their potential, innovative startups face numerous obstacles to their growth, including regulatory hurdles, administrative challenges, and a need for more financ-

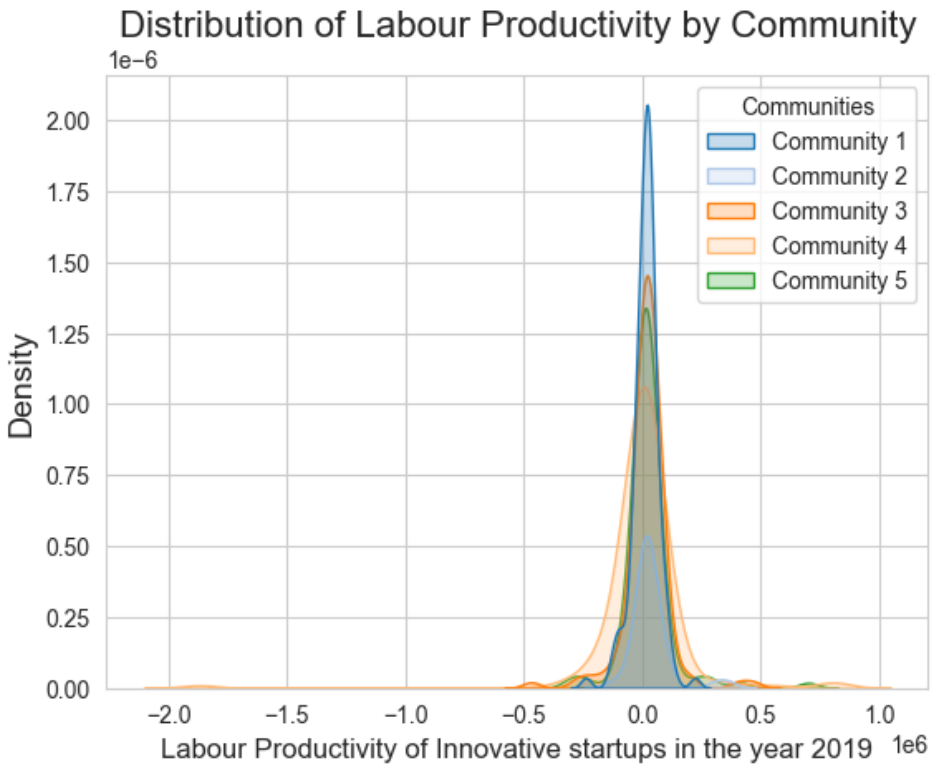


Figure 2.9: Distribution of the LP within each cluster for each Community. The distributions are approximately normal with similar mean but different variance. Community 1 and 2 have the lowest variances while Community 4 has the highest variance. Note that the negative values of productivity are not surprising, innovative startups are newborn firms that have to face upfront costs with likely no or little source of revenue. This means that their added value, which is the difference between revenues and costs, can be negative and therefore LP, computed as the ratio between the added value and number of employees can be negative, too.

ing and skills. The presence of regulatory protection for established firms, complex regulatory procedures, and an inefficient bankruptcy framework, which may involve lengthy and costly legal proceedings, can significantly impede market entry, experimentation, and necessary exits. In the early stages of a business, information asymmetries, lack of collateral, and the absence of a track record typically restrict new entrepreneurs' access to external funding. Market failures and institutional barriers underscore a pressing need for supportive policies to foster entrepreneurship. These policies are beneficial and essential for the growth and sustainability of innovative young firms [143]. In this view, the power of the algorithm succeeds in distinguishing individual spatial entities and identifying the main peculiarities of each, even in the presence of a multitude of information and characteristics of innovative enterprises. The added value of the clustering algorithm, combined with the wealth of information in the purpose-built database, allows the policymaker to have greater

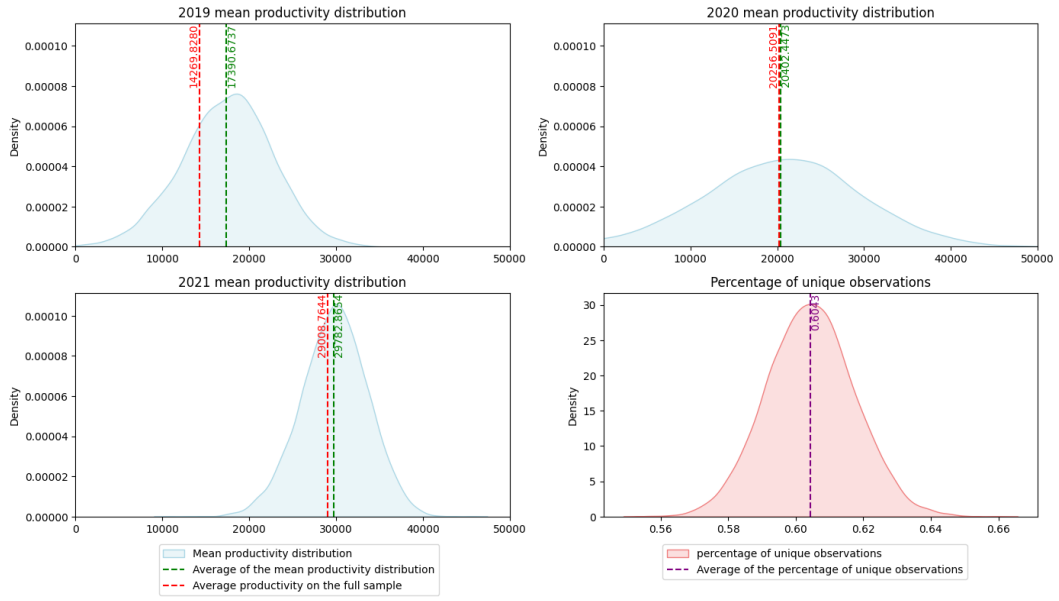


Figure 2.10: Bootstrap distribution for mean LP of the innovative startups. 10000 bootstrap replicas have been computed using the clusters as strata; for each bootstrap replica, the mean LP for 2019, 2020, and 2021 has been calculated and added to the distributions. For each distribution the average has been computed and is presented in green. The average calculated of the entire sample is also presented in red for comparison. The pink plot shows the distribution of the number of unique observations (i.e., how many different startups) contributed to each bootstrap replica. Approximately 60% of all the observations have been included in each bootstrap replica.

completeness and robustness of the results obtained, which support industrial policies to be adopted to develop innovative startups. The clustering algorithm, starting from the requirements that innovative startups must possess, allows highlighting the differences in characteristics between peripheral innovative startups and those located within urban areas, characterized mainly by high technological specialization. This makes it possible to identify five different communities of companies, each with its peculiarities but always meeting the requirements of an innovative startup.Hence in this chapter, we performed analyses of innovative startups to better assess their key features and characteristics and how these features combine to create meaningful clusters of businesses that effectively represent the innovation landscape of the Lombardy region. We used fine-tuned state-of-the-art clustering algorithms to make these clusters, united their results in a bipartite graph, and used modern community detection algorithms to obtain a superior dataset partition. Once the groups have been created, modern explainability techniques have been used to make the clusters more intelligible by humans and shed light on the logic used by the ML algorithm. These clusters were finally used in the bootstrap analysis to evaluate the distribution of the average LP of innovative startups over the years 2019-2021 and assess the

economic consequences of the lockdown caused by the COVID-19 pandemic. The work will be further enhanced by studying the strategic behaviours of innovative startups and alliances with firms belonging to or not belonging to the same group using spatial network complexity techniques[96,267]. In addition, in future works, we plan to implement this strategy to groups of businesses larger than innovative startups to scale the algorithm efficiently and also to implement even more robust explainability algorithms like SHapley Additive exPlanations (SHAP)[240] to produce even more detailed descriptions of the clusters and delve deeper into cluster ensembles, implementing other algorithms based on graphs like Ensemble Clustering using Factor Graph (ECFG)[174,310].

"The purpose of computing is insight, not numbers."

Richard Hamming

3

Spatial bootstrapping using deep clustering methods: spatial machine learning applied to Lombardy high-tech businesses

Bootstrap and clustering techniques are foundational tools across scientific disciplines, playing a particularly important role in spatial analysis. However, traditional bootstrap methods often fall short in preserving spatial dependencies and complex attribute relationships during resampling. In this chapter, we introduce a novel framework in the Spatial ML domain that leverages deep learning techniques to enhance stratified bootstrap procedures for spatial data. Deep learning has already revolutionized prediction and classification tasks in data with temporal and spatial dependencies. In this chapter we want to extend the scope of application to bootstrap analysis by using tools like entity embeddings and autoencoders. By encoding high-cardinality categorical variables into continuous representations, entity embeddings facilitate the discovery of meaningful spatial and attribute-based cluster. These embeddings are then passed to a DEC algorithm that can use them to create clusters. This algorithm is able to handle high-dimensional big data using an autoencoder-based architecture that

performs dimensionality reduction and clustering simultaneously to avoid loss of information. These clusters can be finally used as strata that guide a stratified bootstrap approach which preserves spatial autocorrelation and heterogeneity. We demonstrate the utility of our framework by performing a bootstrap analysis of high-tech firm productivity in the Lombardy region. Our approach is able to efficiently analyze large amounts of high-dimensional data with complex attributes.

3.1 Introduction

The productive structure of high-tech sectors is rich and complex, characterized by a high degree of heterogeneity in productive specializations, varying relevance to economic activity, and diverse firm sizes. Innovation is widely regarded as a key driver of economic development and firm competitiveness[313,329]. This view informs many policy strategies aimed at promoting creative entrepreneurship, technology transfer, and market competition, with the long-term goal of accelerating industrial transformation and boosting productivity and efficiency across the economy[15,143].

The spatial distribution of high-tech firms is shaped by multiple factors, including market size, access to scientific infrastructure, the availability of skilled labor, and agglomeration effects stemming from the proximity of other firms and public knowledge centers[281]. These agglomeration dynamics, facilitated by information spillovers and localized learning, are widely recognized as key drivers of innovation and regional economic growth[115,144,219,324]. According to Kichko et al.[205], high-tech clusters emerge in contexts where localized knowledge spillovers, abundant human capital, and low commuting costs intersect, underscoring the importance of spatial externalities in the formation and expansion of innovation hubs.

Through processes of localization and urbanization, agglomeration leads to both the concentration and the diversification of productive activities across geographic space[108,302]. Understanding how this spatial partitioning relates to technological specialization and tacit knowledge diffusion[300] has become a central concern in the empirical analysis of innovation systems. In this context, a variety of methods have been developed to detect and analyze spatial clusters of firms. Among these, distance-based approaches are widely used in economic geography to examine the spatial concentration of firms relative to their technological profiles[16,212,246]. Marked point processes and related tools such as Ripley's K function have become standard for capturing spatial autocorrelation and understanding how firm-level attributes interact with geographic proximity.

However, while spatial clustering offers valuable descriptive insights, its role in statistical inference, particularly in resampling methods like the bootstrap, remains underdeveloped. Traditional bootstrap techniques often neglect spatial autocorrelation and attribute heterogeneity, potentially leading to misleading inference in the presence of spatial structure. This chapter addresses that gap by proposing a novel spatial bootstrapping framework in which clustering plays a methodological role: not as an end in itself, but as a means to define strata that preserve spatial and attribute-based dependencies during resampling.

Our contribution is primarily methodological. We introduce a Spatial ML[213] pipeline that leverages recent advances in deep learning, specifically entity embeddings and DEC, to produce robust strata for stratified bootstrap procedures. Entity embeddings are used to convert high-cardinality categorical variables into continuous latent representations that capture meaningful relationships between firm-level features[6]. These embeddings are then passed to a DEC model, which performs clustering and dimensionality reduction simultaneously. Unlike traditional clustering algorithms such as k-means, which perform poorly in high-dimensional spaces due to the curse of dimensionality, DEC optimizes feature representation and cluster assignment jointly, improving the quality of strata in complex datasets.

By using these deep learning tools to construct spatially and semantically meaningful strata, our approach allows for more reliable bootstrap inference in high-dimensional, spatially dependent datasets. We demonstrate the value of our method through an empirical analysis of high-tech firm productivity in the Lombardy region. This application highlights how our framework can be used to improve the validity and interpretability of spatial resampling methods in real-world economic data.

The remainder of the chapter is structured as follows. Section 2 provides a review of the key literature underlying our approach. Section 3 describes the geodatabase constructed for high-tech firms in Lombardy. Section 4 presents the proposed methodology in detail. Section 5 discusses the empirical results. Section 6 concludes with a summary and suggestions for future research directions.

3.2 Literature review

Over the last two decades, significant advancements have been made in the field of neural networks thanks to faster computers, more data, and improved methods. Neural networks have begun to replace or are already replacing long-standing methods in various fields where the data are unstructured or it is hard to find structured data. However, neural networks are less crucial for ML problems involving well structured

data[320]. Different from unstructured data found in nature, structured data with categorical features usually lack continuity, and when they do have some continuity, it may not be immediately apparent. The continuous nature of neural networks limits their effectiveness with categorical variables. Therefore, directly applying neural networks to structured data with integer representations for categorical variables could be more effective[373]. One common workaround is to use a one-hot encoding. Still, this approach has two significant drawbacks. First, with many high-cardinality features, one-hot encoding demands unrealistic computational resources. Second, it treats different values of categorical variables as entirely independent from one another, often ignoring the informational relationships between them[157,229]. In Guo and Berkhahn[157], entity embedding is introduced to learn the representation of categorical features in multidimensional spaces automatically. The goal of entity embedding is to approximate similar values of a categorical variable in the embedding space. By using real numbers to define the similarity of values, entity embedding is closely related to the problem of embedding finite metric spaces in topology.

Entity embedding enhances the neural network's generalization ability, especially when data are sparse and statistical properties are unknown. This makes it particularly useful for datasets with many high-cardinality features, where other methods tend to overfit. Furthermore, embeddings obtained from a trained neural network significantly boost the performance of all tested ML methods when used as input features. Since entity embeddings define a distance measure for categorical variables, they can be used for visualizing and clustering categorical data.

Entity embeddings are vector representations of categorical variables or entities in a dataset. These embeddings are learned by training a neural network to capture the relationships between different entities in a high-dimensional space. They convert categorical data into continuous numerical data, allowing the use of ML algorithms that require numerical inputs. Additionally, they can be employed for tasks such as clustering, visualization, and similarity searches.

3.2.1 Clustering

Clustering is a fundamental ML problem whose performance heavily depends on the nature of the data and the quality of data representation. Therefore, feature transformations have been extensively employed to enhance data representation for clustering. This becomes particularly relevant in the field of Spatial ML where the spatial characteristics of the data are extremely correlated to the other attributes of the system.

The use and development of appropriate methodologies for clustering with spatial data is an essential topic because it can have huge implications in many different field. For example, Farzammehr and Moradi [123] employed spatial analysis and clustering methodologies to examine smuggling operations in Iranian districts from 2009 to 2017. Similarly, Gómez [146], in his analysis of spatial patterns in Bolivia, demonstrated the utility of clustering techniques in identifying key regional clusters of economic activities, highlighting how these spatial distributions can inform policy decisions and resource allocation.

Similarly, Ryu et al. [321] present a hedonic pricing model that improves real estate value predictions by using geographic data, such as latitude and longitude. Instead of treating location as just another category, the model uses it as a continuous variable, allowing it to better capture non-linear relationships within the data. It works conceptually similarly to entity embeddings in neural networks, which reveal hidden patterns in categorical data. This approach makes the model more sensitive to the complexities of the real estate market, ultimately leading to more accurate forecasts.

For this reasons recent methodological works have focused on developing techniques that are able to highlight and exploit the spatial characteristics of data for analysis. One of the most relevant algorithms in this field is ClustGeo [69]. This model is extremely useful and is based on the creation of two matrices: one with the spatial distances and one for attribute dissimilarities, sometimes computed using the Gower distance [154] and uses a convex combination of the two to create the clusters. The model is extremely powerful and adapt for many scenarios where aggregate level data are available. One of the limitations of the algorithm is that the size of the matrices scale quadratically with the size of the dataset and therefore it becomes rapidly too computationally demanding for firm level analysis scenarios like the one presented in this chapter, where tens of thousands of firms have to be considered at the same time.

Other fundamental algorithms are DBSCAN [116] which is a non-parametric density based clustering algorithm and k-prototype [175] which is an extension of the k-means algorithm that works for big data with categorical variables. These are powerful algorithms that are appropriate in most use case scenarios, but they are still based on the idea of computing the distance between data points which becomes impractical in use cases where data are high dimensional and there are collinearity issues between the features.

We decided to employ neural networks because their complexity and memory demands scale linearly with the size of the dataset and the number of features. Neural networks in our use case do not compute directly distances between all the pairs of points but rather look at the Kullback–Leibler (KL) [220] divergence between distribu-

tions. This makes the computational effort for firm level analysis more manageable on regular computers and motivated the employment of this methodology. Deep neural networks are particularly apt to learn nonlinear mappings[372] that allow transforming data into a representation that simplifies the clustering task without requiring manual feature selection and engineering. Many different deep clustering algorithms exist in the literature that the interested reader can find in literature review works like Ren et al.[309]. In this chapter, we will employ the DEC algorithm. DEC, which simultaneously learns feature representations and cluster assignments, is a promising technique for managing extensive variables and nonlinear relationships[369]. DEC learns a mapping from the embedded data space to a lower-dimensional feature space and performs a clustering task.

The algorithm employs an autoencoder, a neural network capable of summarizing a large number of variables into a reduced set of latent features. DEC utilizes these latent features to identify clusters. For instance, it uses K-means to determine initial cluster centroids, then optimizes the encoder of the autoencoder and updates the cluster centroids to form the clusters[62,91].

In the specialized literature the researchers use the bootstrapping technique to assess the stability of cluster analysis results and make statistical inferences.

Bootstrap methods are precious when dealing with non-normal data, skewed distributions, outliers, small sample sizes, or complex statistics. However, complex data structures and dependencies challenge bootstrap methods because they violate the assumption of independence and identical data distribution. For instance, with clustered data, bootstrap samples may not accurately reflect the variation within and between clusters. Similarly, with longitudinal data, bootstrap samples may fail to preserve the temporal order and correlation of the observations. Additionally, in the presence of correlated or causal variables, bootstrap samples might not correctly capture the joint distribution and direction of effects. Various approaches can address these complexities and dependencies when using bootstrap methods, depending on the nature and degree of the complexity and dependency. One such approach is the stratified bootstrap, which divides the data into strata or homogeneous subgroups based on relevant variables, factors, or clusters. This method ensures that the bootstrap samples better represent the variability within each subgroup and maintain the structure of the original data.

3.3 Data and variables

To investigate how firms are spatially clustered, we introduce a Spatial ML algorithm[213] that utilises the geographical coordinates of firms but also considers the other attributes and characteristics of the firms.
The source of information that we used is the ASIA-Istat database by tax code[1].

The analysis is conducted using micro-data including different characteristics of firms as the sector of activity of each enterprises using a 5-digit code called ATECO, the local version of the European NACE code. The available data are updated to 2020 but we restricted the dataset to firms that have been active in the years 2017, 2018, and 2019 to avoid distorting effects with the outbreak of the pandemic from COVID-19. Finally, the cleaned dataset had a total of 109 features for 24976 businesses.

3.3.1 Description of the dataset

Istat databases, specifically the ASIA databases, is the main source of the data used in this project.
Specifically, we used four ASIA databases[2]:

1. ASIA businesses with the business' demographics.
2. ASIA local units with info about the workforce of each local unit that constitutes each business.
3. ASIA Trade by Enterprise Characteristics (Istat framework) (TECFRAME)-Structural Business Statistics (SBS) with information about the export and import activity of the business, if applicable.
4. ASIA economic results with the economical performances of each local unit for each business with revenues and costs.

Together with the Tagliacarne Study Center[3], we combined the databases provided by Istat using the enterprise code as unique identifier and georeferenced the data by calculating the latitude and longitude of each enterprise; about 2% of the

[1]The ASIA-Istat contains information about the firms' structure, financial situation, and whether and what they export. Further information is available at ASIA-SBS [22]

[2]See C for more details on the structure of ASIA's databases.

[3]The Guglielmo Tagliacarne Italian Chambers of Commerce Study Center, actively creates, gathers, and evaluates data about Italian companies and the Italian economy

enterprises could not be successfully georeferenced[4].
Local units' information has been extracted, cumulated and attributed to the whole business. The economic activity code, ATECO code, was attributed to the whole enterprise. Firms outside the Lombardy region and outside the high-tech sector were filtered out. The definition of high-tech adopted is from[121], where a firm is considered high-tech if its ATECO code falls into one of the categories shown in Table 3.1.
The ATECO code[23] is the Italian version of the NACE code[266]; the two classifications are identical up to the fourth digit. The ATECO code adds additional digits to further subdivide the subcategories.
For our analysis the data needed further preprocessing:

1. Duplicated columns or almost identical columns have been discarded[5].

2. Each column about import and export for businesses not belonging to the TECFRAME-SBS dataset has been attributed to zero.

3. The 2% of businesses for which Latitude and Longitude could not be computed have been discarded.

4. All the companies that did not belong to the 2017, 2018, and 2019 versions of the ASIA databases simultaneously have also been discarded.

5. The productivity of each business for the years 2017, 2018, and 2019 has been computed as the quotient of the total revenues of the business and the total number of employees. The productivity between two years is the variable for which the correlation with the stratified bootstrap method will be computed.

6. All available data about the businesses for the most recent year, 2019, have been chosen as the variables to use to create the clusters.

The initial ~150 columns have been replaced by 109 columns in the final dataset, which is composed of 24976 businesses out of the 30704 that were active in 2019.

[4]We acknowledge the fact that the georeferencing process could introduce biases if the excluded firms have distinct characteristics. For instance, firms in remote or less urbanized areas might have been disproportionately affected, potentially impacting the representativeness of the clusters. This goes beyond the scope of the present work and we plan to address it in future works

[5]Other highly correlated features have been kept because neural networks can use them to learn complex interactions between features[151] and use them proficiently, especially in combination with regularization techniques like Dropout[342] and Batch normalization[180].

Sectors	**ATECO Codes**
Pharmaceuticals	21XXX
ICT Manufacturing: semiconductor, computer, and TLC hardware	261XX, 262XX, 263XX, 264XX, 265XX, 267XX 268XX
Aerospace: aircraft manufacturing, spacecraft and related devices, satellite communications	303XX, 3316X, 613XX
Biomedical: electromedical and medical devices	266XX, 325XX (without 32505)
ICT Trade: wholesale and retail trade of ICT equipment	465XX, 474XX
IT Services: software publishing and production, IT consulting, database management	582XX, 62XXX, 631XX
Telecommunications: fixed and mobile telecommunications	61XXX (without 613XX)

Table 3.1: High-tech categories and their corresponding ATECO codes. Each ATECO code is a five-digit code where the first two digits are the category, and the last ones are the nested subcategories. The Xs indicate that all the subcategories within that category have been taken.

3.3.2 High-tech industry and knowledge-intensive services

There are many types of enterprises developed based on various criteria, such as size, ownership structure, scope of business, or the industry in which they operate. A further classification is based on the recognition of enterprises and their level of technological advancement. Advanced technology, as a new and innovative field, dramatically impacts the economic shifts within nations. High-tech sectors play a vital role in the reorganization of industries and the transformation of economies. Their growth is essential for escaping the middle-income trap and creating a modern and robust nation. Nevertheless, the classification and conceptualization of high-tech enterprises require a multidimensional approach. The high-tech sector is so broad that it can include enterprises at various levels of technological advancement. There is no uniform taxonomy of high-tech enterprises. The product-based approach is often used to classify high-tech industries and enterprises. The definition of high-tech industry and goods mainly refers to all enterprises operating in the

high-tech sector, producing high-tech goods or providing technologically advanced services [145,242,344]. A limitation of this approach may lie in the size of firms included in the sector. As a result, the share of SMEs may be significantly underestimated [29]. A comprehensive definition of the high-tech sector is found in Świdurska [348], who states that a high-tech sector employs a high level of scientific and technical personnel, collaborates intensively with scientific and research institutions, is characterized by a rapid aging process of developed products and technologies, dynamically exchanges resources in terms of technical infrastructure and patents. Following the definitions of high-tech firms, the number of companies within each category and for each year can be computed. The results are shown in Figure 3.1. The overall number of high-tech companies in the Lombardy region is 30704 and has slightly increased in the five years data was considered. The IT services category, with approximately two-thirds of the overall companies in this sector, is the predominant one. Interestingly, the perceptual change shown in Figure 3.2 shows that the smallest industry, Aerospace, is growing faster.

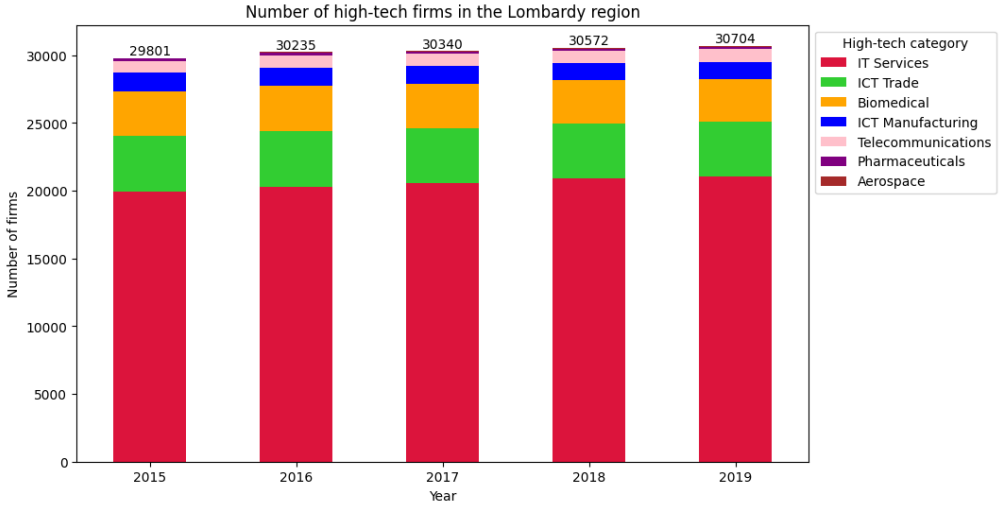


Figure 3.1: The number of companies in the High-tech sector between 2015 and 2019 has slightly increased over the years, and most of the sector is made up of IT services companies.

3.4 Methodology

3.4.1 Entity embedding

Different techniques exist in literature to treat high cardinality categorical variables without adding more variables like target encoding [255]. The one algorithm implemented in this case is entity embedding [157], as it produces a meaningful representation of data without the risk of data leakage between the features and the target.

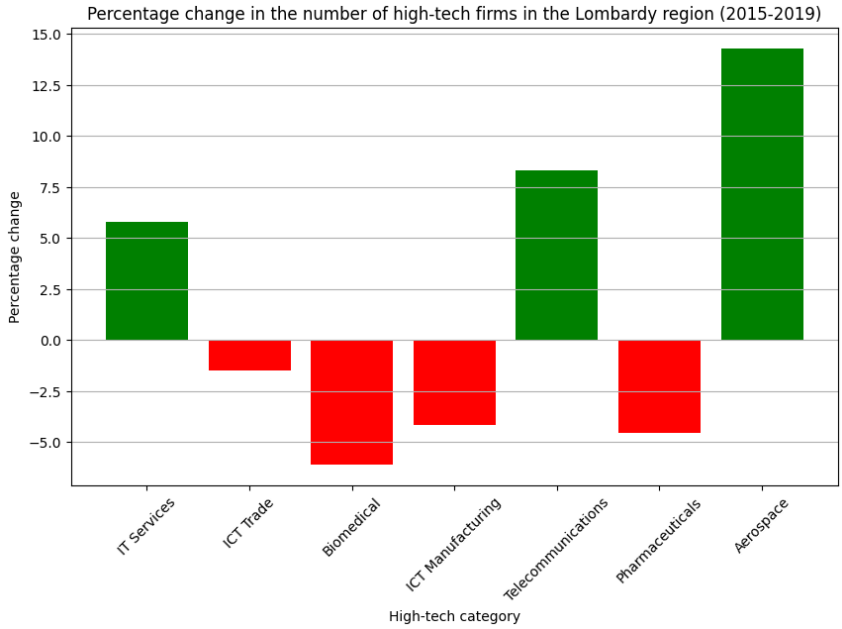


Figure 3.2: Relative change between the years 2015 and 2019 for each category. The Aerospace category, which, with 64 companies in Lombardy in 2019, is the smallest sector in absolute terms, is the fastest-growing one in relative terms.

This algorithm trains a deep neural network for a specific regression task. It uses the numerical outputs of one hidden neural network layer to embed the data.
This numerical output of one of the hidden layers of the neural network can be considered a valid proxy for the original data if the neural network can learn a meaningful representation of the data and can generalize it to unseen data.
The embedding is completely numerical and can be used for all subsequent computations as an effective proxy for the original dataset. Implementing this procedure has become necessary because several variables in the original dataset were not numerical but rather high-cardinality categorical variables.
We want to stress that this step is fundamental for the subsequent analysis as it allows to transform an initial dataset composed of highly collinear features with complex categorical variables in a simple completely numerical representation. This is possible in the bootstrap (and similar) scenarios because the relevant target variable that will be used in the bootstrap is known a priori and therefore it is possible to create regression task with the only purpose of creating an embedding preserving the relevant information for the bootstrap purpose.

3.4.2 Deep Embedded Clustering

The embedded data generated can be used for unsupervised tasks like clustering. The traditional approach to solving the problem of high dimensionality is either

the use of an algorithm like CURE[156] or the use of some dimensionality reduction technique like Principal Component Analysis (PCA)[201], factor analysis[34], or nonlinear techniques like KPCA[327], before implementing the clustering algorithm. The main issue with this approach is that dimensionality reduction techniques act independently from the subsequent clustering technique and could suppress dimensions with helpful information for the clustering algorithm. To address this and other issues, algorithms leveraging neural networks have been developed to perform clustering with high-dimensional data in the Deep Clustering field[147,309].

One of the first and most influential algorithms developed in this field is DEC[369]. DEC jointly optimizes feature representation and cluster assignments in an iterative fashion. This procedure is repeated such that a slightly different feature space with more refined clusters is generated each time, gradually increasing the cluster purity. This cluster purity is quantified using the KL divergence, a measure of how one probability distribution diverges from a second, expected probability distribution[220]. In the context of DEC, KL divergence is used as the loss function to compare the soft cluster assignments (the model's output) with a target distribution that emphasizes more confident assignments. By minimizing the KL divergence, the network improves its clustering performance with each iteration[151,369]. The interested reader can find the full architectural details in the original paper[369].

A crucial requirement of DEC is that it must be initialized with a "naive" clustering guess to improve upon. To this end, the embedded data has been trained on a MiniBatchKMeans model[331], a lightweight and scalable version of KMeans suited to large datasets.

DEC also relies on training an autoencoder, a type of neural network that constructs a low-dimensional representation of the input data through a process of encoding and decoding. The encoder maps high-dimensional input into a compressed representation, while the decoder attempts to reconstruct the original input from this compressed form. This compressed representation is often referred to as the latent space: an abstract, typically lower-dimensional space where meaningful features and structures in the data are preserved[44,151]. The key idea is that this latent space captures essential properties of the data, enabling more effective clustering. If the decoder is linear and the mean squared error is used as the loss function, the autoencoder is mathematically equivalent to PCA[151]. When nonlinear components are used, the autoencoder becomes a nonlinear generalization of PCA, allowing for more expressive representations.

The encoding part of the autoencoder and the initial clustering guess are combined into the DEC model. This model then iteratively adjusts the latent space by focusing on data points that are confidently assigned to clusters, while updating the clusters

based on the new representation. This iterative process enhances clustering quality over time.
Recent studies[62,91] show that DEC models are particularly effective at uncovering nonlinear relationships in spatial economic data. In our case, DEC enables us to capture the joint structure of spatial coordinates and categorical economic variables, which traditional clustering algorithms (e.g., k-means) cannot accommodate without preprocessing or dimensionality reduction steps.

3.4.3 A novel stratified bootstrap: geographical data with attribute space

The use of bootstrap techniques to calculate statistical quantities of interest has been widely adopted over the years[212]. The idea is to extract by repetition subsamples of the original data set and compute the statistical quantity of interest, such as the mean or, in our case, a correlation coefficient, on these subsamples.
This procedure is iterated for a fixed number of times, and all the calculated correlation coefficients are put together to obtain a correlation distribution.
A variant of the Bootstrap, called the Stratified Bootstrap[212], is particularly suitable for spatial data, as it allows for spatial features of the data, such as spatial autocorrelation, to be taken into account. For a full discussion of the topic of spatial autocorrelation and its solutions, see the work of Jiang[199] and Kopczewska[215]. The variant of Bootstrap still involves extraction with repetition, but not from the entire dataset, but from subsamples, which are relatively homogeneous within but different across subsamples. One possible way to compute these layers is to apply a clustering algorithm such as the well-known k-means to spatial coordinates.
This robust approach is then limited by the characteristics of the clustering algorithm used to compute the layers and the type of data it can use.
For example, suppose this procedure considered only latitude and longitude with a k-means algorithm and used clusters as layers for Bootstrap in the application for business use cases. In that case, it might overlook some characteristics of businesses, namely industrial and economic indexes. On the other hand, a more significant number of numerical variables fed by the k-means algorithm would lead to problems related to increased computational costs and possibly the curse of dimensionality[356].
The algorithm improvement we propose simultaneously considers the geographical location of firms and space attributes. As a result, the algorithm can distinguish between a software development activity and an aerospace activity because discrete information of this type is stored in ATECO codes[23], which are categorical variables[6] with high cardinality. These variables pose a challenge because traditional

techniques, such as using dummy variables [105], can result in an excessive increase in the total number of characteristics.
However, it is worth addressing because these variables, such as the ATECO above code for the business sector, provide essential information about the firm. Attributes inform the clustering algorithm that although two firms may be geographically distant, they may share some similarities if they operate in the same industry. Conversely, two companies may be in neighboring geographic areas but be subject to different local taxes and policies, further differentiating them.

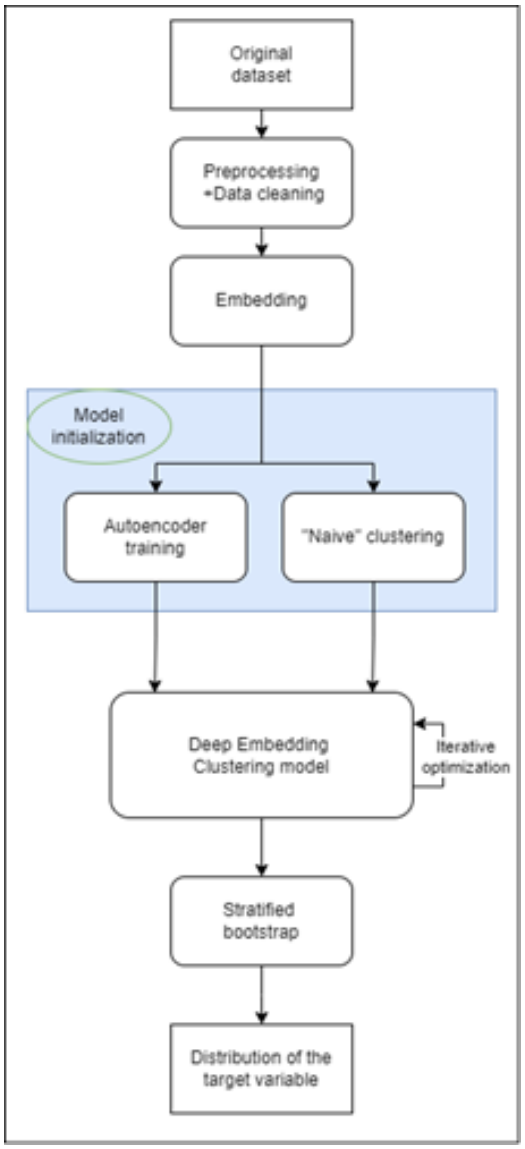


Figure 3.3: This is the proposed method for extracting insightful statistics using stratified bootstrap. After the initial data preprocessing and cleaning, the data are mapped to an embedding space; in this case, we used the Entity embedding algorithm. Subsequently, the data are fed to a deep clustering algorithm to perform both dimensionality reduction and clustering. These clusters are then used as strata in a stratified bootstrap algorithm to obtain insightful information.

The relationship between the geographic and attribute spaces is a fundamental aspect of our algorithm. When the geographic space of relations changes, the attribute space also changes. For instance, in creating a partition of the territory of enterprises, the geographic space is the areas containing the firms.

The corresponding attribute space describes enterprises' qualitative and quantitative characteristics, such as ATECO code, productivity, turnover, number of employees, and more. Understanding the relationship between geographic and attribute space is key to understanding the functionality and effectiveness of our algorithm. In our application, the specialization of high-tech firms, that plays an important role in the mainstream of economic and geographical studies, is the spatial feature for which spatial clustering is developed[213,279]. The defined industrial and socio-economic characteristics of the areas and the constraints on their definition specified by the (co)location of the activities of the firms constitute the attributes for which a functional partition of the territory is developed.

Several criteria, such as quality, complexity, and comprehensibility, combined with the chosen Spatial ML algorithm's complexity, interpretability, and computational performance, can assess the adequacy of the attributes[210].

This novel multistep estimation procedure is a mixture of traditional statistical techniques, bootstrapping and ML[215].
The following Figure 3.3 summarizes the steps of the algorithm implemented in our work.

3.5 Empirical evidence

The detailed technical description of the embedding procedure and the implementation of DEC is presented in C.1. In this section we focus on the description of the clusters that the algorithm produces and how they can be used as strata for spatial analysis.

3.5.1 Clusters as strata

When we talk about stratification, we mean the organization of geographic objects into subsets (called strata) based on the similarity of their attributes or spatial relationships[158]. In spatial analysis, stratification is typically conducted on a geographic basis, such as by dividing the study area into subregions using specific variables. This approach is beneficial for purposes such as sampling[80] or statistical inference[364]. Within each stratum, spatial objects should be as similar as possible regarding attributes relevant to the analysis. Similarity can be measured by variability within strata, distance matrices, or probability models.
Lastly, the correlation coefficient is computed for each bootstrap sample, and therefore, a distribution for this variable is obtained.
The DEC algorithm created 11 clusters, with the larger clusters comprising 11656 businesses and the smaller ones comprising 167 firms.
Their geographical division is shown in Figure 3.4, and the division according to the activity sector is shown in Table 3.2 in absolute values and in the heatmap in Figure 3.5 for the relative values.
We want to highlight the fact that the presence of a central region with extremely high concentration of firms did not imply automatically that they all belong to the same cluster, an issue observed in some density based clustering algorithms, and therefore the algorithm has been able to discern between spatially agglomerated firms.

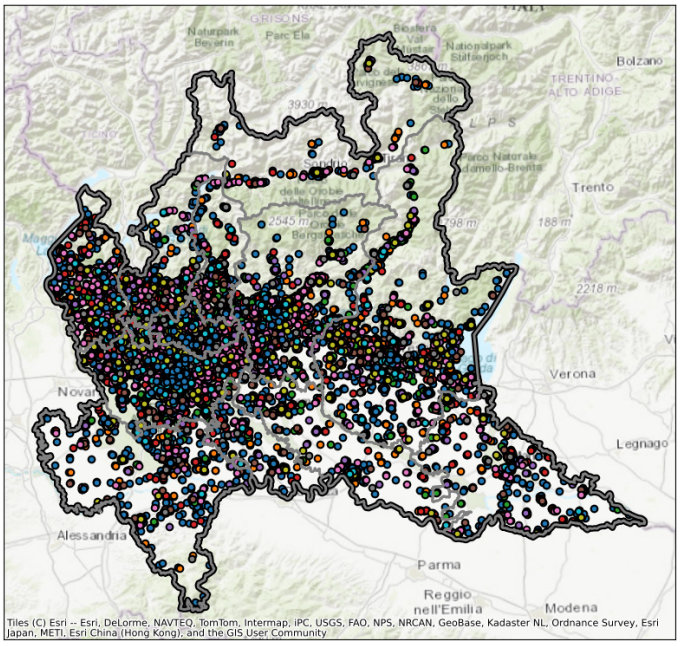


Figure 3.4: Image of the Lombardy region where each business is coloured according to its cluster. Note that even though the latitude and longitude are two considered variables, the clusters are mixed. The map has been created using Contextily[18] and Geopandas[202].

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Aerospace	17	3	11	2	1	2	15	0	1	1	0	53
Biomedical	1734	366	110	254	70	62	184	39	22	10	17	2868
ICT Manufacturing	248	74	151	71	63	73	415	23	10	11	8	1147
ICT Trade	866	150	1107	567	101	284	250	134	27	22	6	3514
IT Services	8485	2075	1069	1224	1511	1141	215	255	292	145	134	16546
Pharmaceuticals	5	3	29	6	1	5	96	1	0	0	0	146
Telecommunications	301	56	170	55	31	54	21	5	3	4	2	702
Total	11656	2727	2647	2179	1778	1621	1196	457	355	193	167	24976

Table 3.2: Cluster distribution in different high-tech industries

Figure 3.5 presents a map of the Lombardy region with the high-tech firms colored according to their cluster. The algorithm assigns neighboring firms to different clusters.

In a nutshell, spatial embedding is a feature learning approach utilized in spatial analysis, where geographic data types such as points, lines, polygons, or other spatial entities are transformed into vectors of real numbers. In essence, this process involves mapping a high-dimensional space for each geographic object into a continuous, lower-dimensional vector space. Differently, embeddings are vector representations

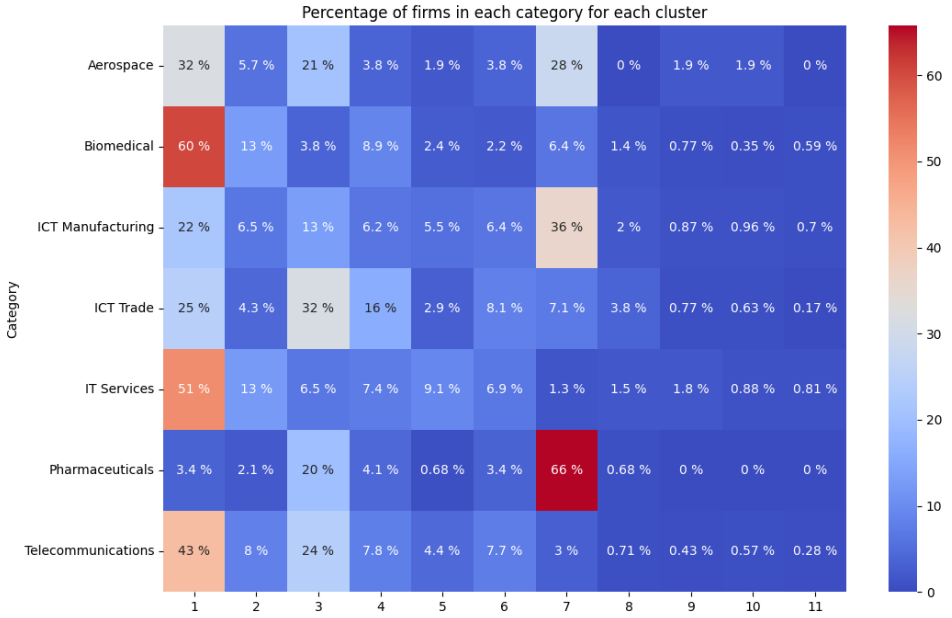


Figure 3.5: Heatmap created using Matplotlib[176], that relates each firm category to the clusters. The first clusters is the largest one and contains most of the Biomedical, IT Services and Telecommunications firms. Most of the Pharmaceuticals firms belong to the seventh cluster.

of data that capture complex relationships in a low-dimensional space. This approach enables the use of complex spatial data in neural networks and have been proven to enhance performance in spatial analysis tasks. Real-world complex data are inherently heterogeneous; they have different types of attributes, and relationships. In the context of spatial heterogeneity embedding methods like graph embeddings or geographic feature embeddings (e.g., socio-economic indicators) can represent spatial regions in a way that encodes similarities and differences between them. Thus, embeddings compress these features into a manageable size while retaining meaningful spatial relationships.

3.5.2 Spatial economic analysis

We compute, also, the correlation between business productivity between 2018 and 2019 and between 2017 and 2018. In Figure 3.6, in the left panel, the bootstrap correlation coefficient shows the distribution of productivity in the year 2018 and productivity in the year 2019. The distribution average (0.9317) is much higher than the correlation computed using the full sample(0.7150). The coefficients values are similar for the correlation between firms' productivity between 2017 and 2018 (Figure 3.7). This indicates that the direct correlation computation between two subsequent years underestimates the productivity. On the right panel is the coverage of

the bootstrap replicas, i.e., the fraction of the total sample present in each bootstrap replica. The relative sizes of the clusters can significantly affect the coverage value. What the identified clusters share is that they use the same or similar production technology nonetheless; «if [the production of] two goods...require similar institutions, infrastructure, physical factors, technology, or some combination thereof, they will tend to be produced [in the same location]»[170]. Moreover, the spatial co-production of products indirectly catches their production technology similarity[267]. Therefore, the clusters considered some characteristics of the firms, such as economic activity, technology, turnover, number of employees, geographic localization, etc. It implies that two geographically close firms may belong to distinct clusters because they differ in these attributes. For instance, adopting a specific technology in a particular geographical location is likely to be restricted not by a lack of workforce or industrial facilities but by a lack of the essential expertise required, namely know-how. Furthermore, knowledge is difficult to create or move outside the localized area where it was produced and where it has become increasingly important[30,141,248].

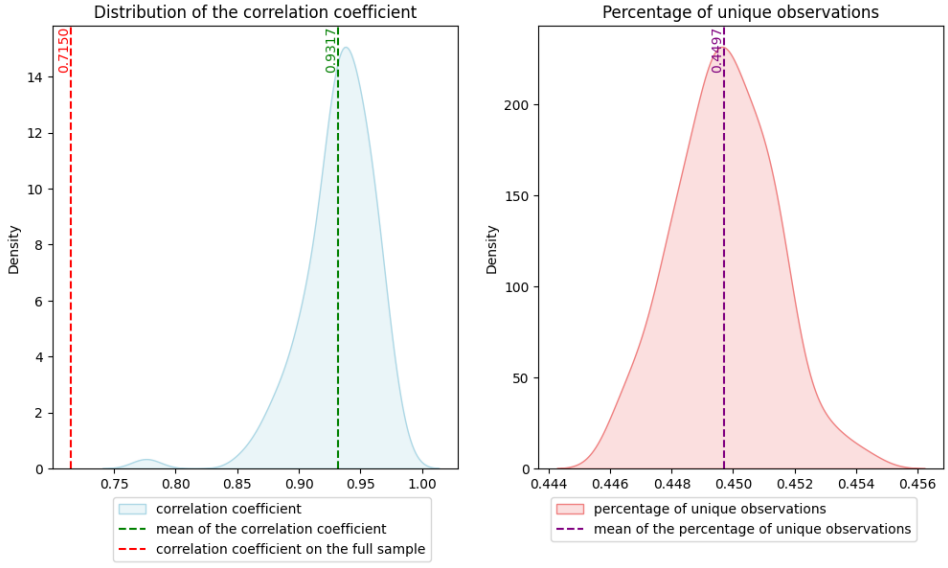


Figure 3.6: The correlation between the productivity between the years 2018 and 2019. The average coverage of each bootstrap replica (in the right figure) is low because the clusters have different sizes

In addition, recent studies in the literature indicate a decline in industrial concentration and specialization indices while employment concentration and specialization indices have risen. It suggests that regions gain a comparative advantage through their efficiency in offering specialized services such as legal support, finance, advertising, and engineering across various industries. Differences in regional productivity related to these specific functions influence the location choices of industries

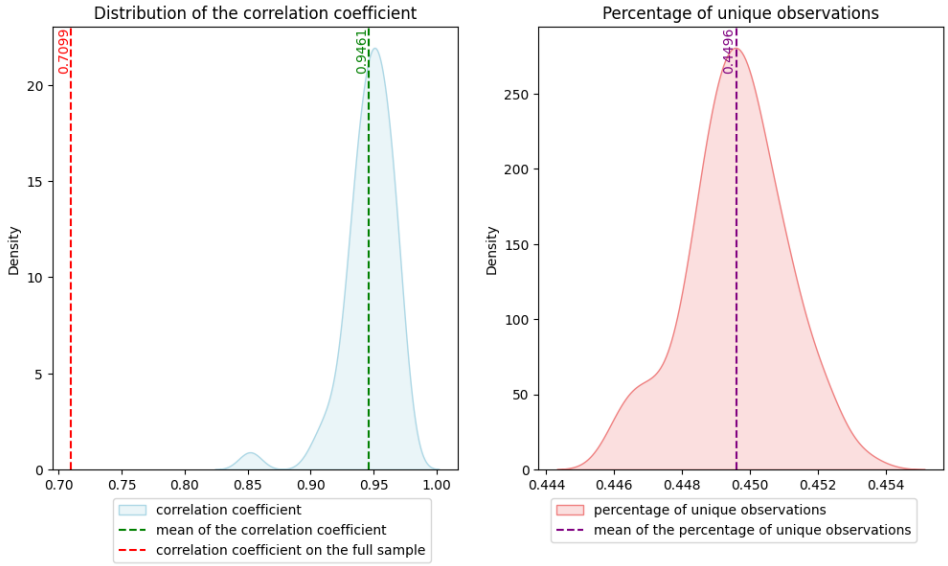


Figure 3.7: The correlation between the productivity between the years 2017 and 2018. The average coverage of each bootstrap replica (in the right figure) is low because the clusters have different sizes

reliant on multiple services, thereby shaping regional specialization patterns in functional roles and industrial activities [142].

These differences in the attribute's space result in large distances in latent space that the clustering algorithm can capture.

Nevertheless, by delving into the potential of analyzing multiple attribute spaces simultaneously, we can uncover the underlying relationship between different forms of representation of technological specialization and the dynamic interaction between attribute spaces.

The complexity of modeling spatial interactions among agents has led the theoretical literature on economic geography [140] to traditionally emphasize simplified contexts, such as a few symmetrical locations, which must be more easily aligned with real-world data. However, more recent research has developed quantitative models of the spatial distribution of economic activity. These data-intensive models contain detailed information about firm characteristics, such as the number of location units unevenly distributed across a geographic area and productivity.

Regional specialization emerges from the unique characteristics of a region, which are influenced by the benefits derived from the clustering of industrial entities and the concentration of production. This perspective emphasizes the importance of understanding the interplay between spatial and sectoral factors, as their complementary relationship significantly contributes to shaping a region's distinct specialization [279]. Thus, the clustering of economic activity within a region is closely tied to the geographical distribution of industrial entities.

Modern spatial economics plays a crucial role in studying new forms of connectivity among economic agents in a given geographic area. It analyzes economic forces' dynamic and complex interaction, particularly the equilibrium balance between these centripetal and centrifugal forces.
Given the complexity of modeling spatial interactions among economic agents, the application of complexity theory and the use of models capable of mapping complex and interconnected spatial connections become not just beneficial, but imperative. These tools are essential in navigating the challenges posed by the intricate nature of spatial interactions, and they form the backbone of our research.
However, research economic complexity is still under development, and its success depends significantly on large datasets and advanced ML methods[32].

3.6 Discussions and future prospectives

The composition of clusters across different high-tech industries (Table 3.2) reveals significant structural asymmetries. Cluster 1, which comprises nearly half the dataset, is dominated by Biomedical, IT Services, and Telecommunications firms. This suggests that it captures a significant segment of the regional high-tech ecosystem, likely centered in more significant urban hubs such as Milan. However, despite this density, the DEC algorithm distinguished additional clusters within this area, avoiding the common pitfall of conflating proximity with homogeneity.
This spatial discernment is evident in Figure 3.4, where clusters are visibly intermixed despite some firms being geographically close. This outcome reinforces the methodological strength of combining geographic data with attribute embeddings: firms are grouped by location and more profound economic, categorical, and structural similarities. For instance, two firms operating in the same city may belong to different clusters due to differences in size, sectoral classification, or export activity.
The heatmap in Figure 3.5 provides a more granular look at the relationship between clusters and industry sectors. Cluster 7 shows an unusually high relative concentration of Pharmaceutical firms, indicating that it might capture a subgroup defined by specific operational characteristics (e.g., R&D focus or international trade profile). Meanwhile, the ICT Trade Category is spread across multiple clusters, signaling internal heterogeneity, possibly related to the firm's role within supply chains, market scale, or digital maturity. Turning to Figures 3.6 and 3.7, which examine the correlation of productivity between consecutive years, the empirical results are particularly illuminating. The bootstrap-based correlation distributions yield significantly higher average values (≈ 0.93) than those obtained from a naïve correlation

on the full dataset (≈ 0.71)[6]. This suggests that when spatial and attribute-based heterogeneity is respected, through clustering and stratified bootstrap, the internal consistency of firm productivity trends becomes more evident.

Interpreting the coverage distribution (right panels of Figures 3.6 and 3.7) is also essential. The observed low average coverage reflects the variation in cluster sizes: large clusters dominate the resampling, while smaller, potentially more specialized clusters appear less frequently. However, this variability is not a limitation but a feature of the stratified approach. It ensures that heterogeneous substructures, such as niche industries or geographically peripheral firms, are not masked by the more dominant sectors.

In broader terms, these results provide evidence that the proposed pipeline, combining deep clustering and stratified bootstrapping, not only enhances methodological robustness but also amplifies the interpretability of economic phenomena. The refined correlation estimates suggest stronger year-to-year stability within semantically coherent firm groups—highlighting the importance of structural similarity over simple geographic co-location in explaining productivity dynamics.

In the present chapter, we applied a spatial bootstrapping algorithm to the high-tech businesses in the Italian Lombardy region. The spatial bootstrapping algorithm proposed is the stratified clustering algorithm. This algorithm assumes that the data are separated into strata of similar data points from which sample with repetition. These strata have been created using a deep clustering algorithm called DEC, based on an Autoencoder Neural Network. The algorithm was able to cluster the data successfully by using the characteristics of the businesses and executing Bootstrap to compute the distribution of the correlation coefficient. The next step will be implementing algorithms to create more interpretable clusters[347]. At the same time, we wish to improve the algorithm analyzing the spatial stratified heterogeneity that refers to the variation in spatial data that arises due to differences within distinct strata or regions. This concept is crucial in many scientific fields, as it acknowledges that spatial data is not uniform, and that variability often exists between different geographic or ecological zones. Recognizing and accounting for this heterogeneity is essential for accurate analysis and modeling. Accounting for spatial stratified heterogeneity[158] leads to more accurate and reliable models, as it acknowledges and incorporates the inherent variability within the data. By recognizing and analyzing the differences between strata, this approach reduces bias that can arise from treat-

[6]The entire clustering and bootstrap procedure implemented in this chapter has been repeated, for comparison purposes, with the traditional k-prototype algorithm and is presented in C.2. The traditional method results in a much lower correlation (≈ 0.31), highlighting the importance of implementing more advanced architectures.

ing the data as homogeneous. Understanding spatial heterogeneity enables more informed decision-making in policy planning, resource allocation, and intervention strategies tailored to the specific needs of different regions. Hence, addressing the complexity and variability in spatial data, embeddings and autoencoders provide powerful tools for tackling spatial heterogeneity in economics.

Moreover, determining the appropriate criteria for stratification can be challenging and requires a deep understanding of the underlying spatial processes. The clustering, embedding, and resampling processes can be computationally intensive, especially with large datasets. Therefore, efficient algorithms and computing resources are essential.

Nevertheless, the accuracy of spatial stratified heterogeneity analysis depends on the quality and granularity of the data. Incomplete or coarse data can limit the effectiveness of this approach. Research in spatial stratified heterogeneity continues to evolve, focusing on developing more sophisticated clustering algorithms, improving embedding techniques, and integrating these methods with advanced spatial analysis tools. Future directions also include: 1) extending these techniques to handle temporal-spatial data; 2) enhancing real-time data analysis capabilities and 3) implementing new geospatial specific cross-validation techniques[370] to better assess the performances of the models during training.

One fundamental issue that should be addressed in future research is the creation of an explicit balancing of the contribution of the spatial and non-spatial attributes in the creation of the clusters while still being able to exploit the information contained in high cardinality categorical variables. Recent relevant works in this direction like Lee and Lauw[230] propose the use of Graph Neural Networks to explicitly encode the spatial aspect to create spatially-aware embeddings of numerical features.

By utilizing clustering algorithms and embedding techniques, researchers and practitioners can achieve a deeper understanding of the intricate patterns and relationships in the physical world. This leads to the development of more accurate models and more informed decision-making.

"A complex system that works is invariably found to have evolved from a simple system that worked."

John Gall

4

How Scale Shapes Productivity: Skills, Capabilities and Complexity from Macro to Micro

This chapter investigates how the tension between specialization and diversification unfolds across scales, from individual firms to region-sized aggregates. Using a uniquely detailed dataset covering around 4.4 million Italian firms per year from 2015 to 2019, we link workers, firms and territories within a unified empirical framework. We measure the diversification of human capital through the entropy of academic qualifications inside each firm and hexagon, construct a multiscale spatial partition of Italy based on the Hexagonal Hierarchical Geospatial Indexing System (H3) hexagonal grid, and proxy export capabilities using an exogenous Fitness index derived from world-level product complexities and aggregated at different resolutions.

At the firm level, Fixed Effects (FE) regressions show that both the level and the diversification of academic qualifications are positively associated with Labour Productivity (LP), even after controlling for size, age, exporting, workforce composition and sector-year effects. The productivity premium of diversified workforces strengthens with firm size, suggesting that larger organizations ben-

efit disproportionately from combining heterogeneous skills. When we move from firms to H3 hexagons of increasing area, the sign and magnitude of key coefficients change systematically: workforce diversification tends to depress productivity in large, region-like hexagons, but becomes strongly beneficial at finer resolutions, while export Fitness matters more for aggregated territories and less for individual firms, where exporting status dominates. These results reconcile the apparent paradox of diversified regions and specialised firms by showing that the specialization–diversification trade-off is intrinsically scale-dependent. Our findings highlight the importance of multi-scale, geometry-based approaches for understanding how capabilities, skills and productivity co-evolve in complex economic systems.

4.1 Introduction

Since Smith[339], economics relates the efficiency of economic systems to the synergies between tasks, workers, functions, firms, and sectors composing them. Specialization, externalities and returns to scope are crucial aspects to understand the persistent heterogeneity among economic actors' performance. At the same time, the growth process of complex systems is a process of diversification[195]: economic systems grow organically like plants, adding functions and new structures and institutions that are required to sustain their size along the growth process. Economic actors grow by expanding their set of activities to fully utilize their resources and capabilities[292]. This tension between specialization and diversification is crucial to understand how complexity arises, aiming at diversifying the system's output while specializing its parts. The concept of related variety[49,138] was proposed by economic geography to ease the tension, by diversifying without wandering far from the core regional capabilities. At the policy level, this same intuition is at the basis of smart specialization strategies[5,130]. Preference for related diversification is observed also for firms[103] and countries[170].

However, different economic actors, firms, industrial clusters, regions, countries, do not solve the tension between diversification and specialization in the same way, because they have different resources and constraints. Using a very common biological metaphor, if economies are close to ecosystems, firms are individual living entities[40,273]. Firms can be modeled as adaptive organisms, evolving algorithmic routines: they decide strategically their inputs and can adjust their boundaries, which drives them toward narrower specialization to fit their niche, their diversification patterns restricted by the knowledge and capabilities embedded in their routines[104]. On

the contrary, geographical entities inherit resources that are mostly persistent, pushing stakeholders toward diversification to mobilize whatever asset and capabilities the territory offers[138]. This difference between firms and regions rises an apparent paradox that is at the core of this work: when we speak of the diversification process of regions and countries, we are actually referring to a firm dynamics, we are saying that a firm based in that region is diversifying, or that a new firm is born. Complexity arises from the aggregation of different firms in a geographical space, through regional capabilities[77,250].

In this work we will look at the combination and aggregation of capabilities, from workers to firms and from firms to regions. In doing so, we will see how the diversification-specialization tradeoff is affected by the size of the economic system, looking at two different ways in which an economic system can grow. First, for firms of different size, from micro-firms to multinational corporations, economic agents growing organically and strategically adding functions and new structures that are required to sustain their size[79]. We will look at the relation between firms function and complexity and the diversity of the education paths of their work force, for the universe of almost 4.5 milion Italian firms. The number of different education backgrounds of employees here proxy the need for different functions: when a firm will require a legal department, a marketing department, a chemistry research department, they will hire lawyers, comunication experts, chemists. Complex and multifunctional firms will need a diversity of education backgrounds, simpler firms can use similar employees.

Second, we will observe how this relates to regional dynamics when changing the scale of observation, from an area as wide as a single firm to a region. One of the unique features of our analysis is a smooth transition from single firms to regions, moving through larger and larger industrial clusters first and clusters of clusters later. We will see how larger areas will prefer diversified work force to have diversified sectoral compositions[98] to cater to different and multifunctional firms. At the regional level we will look at these two dynamics, the coevolution with the regional scale of the complexity of the product basket of regions and the diversity of their education backgrounds. We will achieve this smooth and homogeneous transition by ignoring administrative regions and looking instead at simple geometric divisions of the territory.

Looking at both dynamics will help clarify the coevolution between firms and regional economic growth. We will do so not just qualitatively, comparing firms and regions in general, but quantitatively, measuring in squared kilometers and in number of people the economic scale at which the trade-offs tend to weigh in one direction or the other for different economic actors.

4.2 Literature review

4.2.1 Firm-level specialization, related diversification and capabilities

Earlier management works refer to diversification as a matter of physical resources and administrative complexity, without mentioning capabilities. In Penrose's view, diversification proceeds through incremental branching that exploits underutilized internal resources rather than leaps into unrelated areas[292]. Similarly, Rumelt's early work on corporate strategy showed that firms following "coherent" or related diversification outperform both highly specialized and unrelatedly diversified firms[319]. Montgomery's survey of corporate diversification further systematizes this evidence, highlighting that performance premia are associated with diversification within a shared resource or capability base rather than across unrelated businesses[261]. Markides and Williamson interpret these results through a resource-based lens, arguing that related diversification creates value only when it allows the firm to access or leverage strategic assets that are valuable, scarce and difficult to imitate[247].

Starting from the '80s and '90s, evolutionary economics started referring to unobservable resources of the firm, often embedded in routines and therefore untradable, as capabilities. Firm behavior is then understood as constrained by the coherence of their internal knowledge bases. Evolutionary theories of the firm emphasize that routines and organizational capabilities are persistent, tacit, and difficult to redeploy, so that diversification possibilities remain bounded by existing knowledge[104,273]. This explains why most firms occupy relatively narrow niches, and why related diversification is more common than unrelated expansion. Early empirical work showed that multi-product firms tend to expand into technologically or organizationally related domains, preserving a stable "coherence" across their activities[103,351]. Jacobides and Winter analyse the co-evolution of capabilities and transaction costs and show that vertical specialization along a value chain requires heterogeneous capabilities; reductions in transaction costs translate into greater specialization only when capabilities are unevenly distributed[194]. Argyres and Zenger integrate capabilities and transaction-cost approaches, arguing that differences in comparative capabilities are central in determining firm boundaries and hence the degree of specialization[17]. Together, these contributions support a view of firms as adaptive, but coherently bounded, organizations: diversification is typically related and limited, while specialization along dimensions aligned with their capabilities is the default outcome.

The notion of *core competences* pushes this argument inside the firm and shift the focus to knowledge. Prahalad and Hamel conceptualize core competences as bundles of skills and technologies that underpin multiple product lines, but that are themselves relatively narrow and difficult to build[303]. In their view, the firm's ability to grow depends on its capacity to redeploy these competences across related domains; attempts to diversify far from the core typically destroy rather than create value. This is consistent with evolutionary perspectives in which firms are seen as carriers of specific, path-dependent routines and knowledge bases, so that the scope for diversification is strongly constrained by internal coherence requirements. More related works show that multi-technology firms systematically "know more than they make": their knowledge base is broader than their actual production scope, in order to coordinate loosely coupled networks of suppliers and cope with uneven technological change[57,102]. The product scope remains limited relative to the underlying knowledge base, reinforcing the idea that firms occupy relatively specialized niches.

Related to this literature, different works used the same argument to see how firms' focuses their research in few related technological domains. Early analyses of patent portfolios showed that large firms accumulate competences along persistent, path-dependent trajectories and rarely branch into unrelated fields[289]. Historical studies confirm that diversification tends to deepen existing technological strengths rather than open entirely new ones[61]. Granstrand's work characterizes technology-based firms as balancing the need for some knowledge breadth with the pressure to maintain internal coherence[155]. Patent-based evidence reinforces this view: firms tend to enter technological classes closely related to their prior competences[53], and innovative performance depends more on the coherence of the knowledge base than on sheer variety[275]. Recent work shows that scientific capabilities also condition which technological domains firms can enter, again emphasizing the importance of internal coherence in shaping diversification paths[63].

Relevant to the rest of this chapter, recent empirical approaches operationalize these ideas through measures of skill- and knowledge-relatedness.[270] shows that inter-industry worker flows reveal a latent structure of "skill relatedness," predicting which industries firms are likely to diversify into. Diversification patterns align with these skill proximity: firms overwhelmingly expand into activities that reuse the competences of their existing workforce[269]. Relatedly,[237] analyse the skill and knowledge content underlying production structures. A complementary strand examines firm-level functional complexity. Coad, Mathew and Pugliese[79] quantify firms' hierarchical capability profiles using a nestedness framework and show that the accumulation of more complex functions predicts growth, survival, and the

ability to enter new related domains. These approaches situate firms as coherent, adaptive entities whose specialization–diversification strategies are shaped by the tacit, path-dependent nature of their routines.

4.2.2 Regional and national diversification, related variety and smart specialization

A large body of work in evolutionary economic geography highlights that regions diversify by branching into activities related to their existing technological and industrial structures. The tone and arguments are similar to the firms related literature described in the previous section, diversification happens coherently to the previous industrial structure of the region. However, while at the firm level the scholars try to find a reason why firms would be interested in diversification[292], at the regional and national level the idea that diversification is an important driver of economic growth is undisputed[195]. Also macro-level studies have demonstrated the broad empirical regularity that economies diversify as they grow. Saviotti and Frenken argue that economic development involves a systematic increase in the variety of goods, functions and technologies[323]. Their theory of variety-led growth offers a formal link between the accumulation of knowledge, the expansion of economic activities and long-run productivity gains. The mystery to solve in this literature is why this regional diversification happens to be related, not why the diversification happens to begin with.

Indeed, this "related diversification" mechanism has been observed across countries, sectors and time periods. Klepper's industry life-cycle framework shows that regional clusters grow through entry of firms whose knowledge accumulates locally, generating spin-offs that remain technologically close to their parents[208]. Boschma and Wenting provide a historical demonstration of this mechanism in the evolution of the British automobile industry, where new firms emerged from related pre-existing capabilities in bicycle and engine manufacturing[48]. Such studies frame regional diversification as the outcome of a cumulative, path-dependent process in which new activities draw on the capabilities embodied in incumbent industries.

A complementary strand links regional diversification to variety in knowledge and technologies. Kogler et al.[209] examine US patent data and show that technological variety and relatedness strongly predict regional inventive output as well as future diversification paths. Balland and Rigby[30] use patent co-classification networks to map the geography of technological knowledge, showing that regional diversification depends on recombining locally embedded technological capabilities and on the structure of related technologies present in the region. These approaches

converge on the idea that regional development is driven by the recombination of existing, localized knowledge structures.

At the national scale, classic structural-change theories already stressed that development entails the progressive expansion of the productive structure. Pasinetti's multi-sectoral model formalizes how increasing variety in consumption and production leads to continuous structural change, with sectors emerging as per capita income grows[288]. Cimoli et al.[75] emphasize that catching-up requires broadening the technological base, especially in middle-income economies, where the ability to diversify into more complex sectors is a key determinant of long-run performance. Structuralist and evolutionary approaches thus converge: successful development depends on accumulating a portfolio of capabilities that supports diversification into related but progressively more complex activities.

Recent work in international trade further emphasizes the role of relatedness in structural transformation. Atkin, Khandelwal and Osman show that export upgrading in Bangladesh's garment sector depended on related product capabilities already present in local firms, and that relatedness constraints shape which varieties countries can enter[25]. Javorcik and co-authors document how foreign direct investment stimulates related diversification in host economies, especially when multinationals introduce technologies close to domestic capabilities[198]. These findings reinforce the notion that relatedness governs which sectors can be developed in a given place.

4.2.3 Measuring relatedness and diversification: Economic Complexity

Economic complexity approaches characterise development through the structure of countries' productive capabilities, building quantitative indices of capabilities based on diversification. ECI introduced by Hidalgo and Hausmann[171] aims to nterpret diversification and product ubiquity as reflections of an underlying capability set, showing that more advanced economies tend to be both more diversified and specialised in rarer products. However, Mealy et al.[253] demonstrate that, mathematically, ECI is orthogonal to simple measures of diversity, implying that complexity and diversification capture distinct dimensions of economic structure. This critique strengthens the case for non-linear alternatives such as the Fitness–Complexity algorithm of Tacchella et al.[349], where country Fitness is explicitly extensive in diversification and product Complexity depends on the least-fit exporters, aligning more directly with the idea that diversification itself is a driver of competitiveness. The economic complexity framework constitutes a "new paradigm" linking complexity to industrial policy[32,306];

At the regional level, complexity approaches have been increasingly applied to understand subnational productive structures and their diversification dynamics. Early work by Balland et al.[31] demonstrated that the geography of complex knowledge is highly uneven across regions, and that regional technological portfolios shape their ability to enter new knowledge domains, and Pinheiro et al.[298] showed how diversity in the knowledge basket affects between regions inequality, while previous results highlighted the impact on within regions inequality[325]. More recently, applications of the Fitness–Complexity framework have shown how regional productive capabilities can be inferred from detailed export or occupational structures. For Italian regions, Sbardella et al.[326] document persistent territorial divides in regional fitness, suggesting that regional growth trajectories are constrained by path-dependent capability accumulation. Extensions of these tools for policy evaluation have been proposed by Diodato et al.[99], who develop regional fitness indicators to support smart specialisation strategies in Europe. Complementary evidence from Turkey shows that regional fitness correlates strongly with sectoral LP and structural upgrading[76]. Together, these contributions reveal how economic complexity metrics can be meaningfully applied to subnational territories, providing a micro-founded account of regional diversification and structural change. Particularly relevant to this work, Operti et al.[284] apply the Fitness framework to Brazilian states, introducing an exogenous estimation of regional Fitness computed through the interaction of regional and national level.

Few works applied economic complexity inspired techniques to analyze differences in the measures at different scales, and they are therefore extremely relevant for this work. In particular Pugliese et al.[305] look at the interaction of geographical and technological scales, noticing how emerging properties of the technological diversification process are persistent at different geographical scales if observed with a finer technological lens at the same time. Straccamore et al.[345] look at the relationship between diversification and coherence at different scales. De Stefano et al.[92], using the same data used in this work, look instead at diversification in export destinations as a proxy of complexity.

4.2.4 Skills, workers and multi-scale capabilities: from individuals to firms to regions

Knowledge diffusion across industries has been modelled through labour-flow networks, where worker mobility reveals how capabilities spread between related activities[272]. Building on this approach, the Skill Space developed by Neffke[271] provides a micro-founded map of how occupations are connected through shared knowledge

requirements, offering a quantitative basis for analysing capability diffusion. Further research using job postings links regional skill coherence to economic performance, emphasizing that the composition of occupational skills in a territory constrains its diversification potential[169]. Such findings are consistent with the complexity literature at the urban scale, where occupational diversity and functional specialization jointly account for differences in city-level complexity and growth. The aggregation of worker skills into firm capabilities, and of firm capabilities into regional capability portfolios, leads to a multi-level view of development. Diodato et al.[98] make this explicit at the extensive industry margin: countries expand into new industries when their occupational structures already embed many of the required skills. This nested, multi-layered approach provides a bridge between micro-level constraints on firm diversification and macro-level patterns of regional structural change.

4.2.5 Spatial scale, aggregation and geometric approaches

Most empirical studies of regional diversification rely on administrative regions Nomenclature of Territorial Units for Statistics (NUTS)2/NUTS3, provinces or metropolitan areas as the fundamental spatial unit. While convenient, such boundaries may not capture the true geography of labor markets, supply chains or knowledge spillovers. Evolutionary economic geography typically acknowledges this limitation but seldom departs from administrative units in practice[49,138].

A smaller methodological literature explores more flexible spatial representations. van Dam et al.[359] propose an information-theoretic framework that unifies measures of localization, specialization and coagglomeration, and can in principle operate on arbitrary spatial partitions. Yet empirical applications that systematically vary the spatial scale, moving smoothly from the footprint of individual firms to clusters and then to regions, are rare. Likewise, studies employing geometric divisions of the territory, rather than institutional borders, remain limited despite their potential to reveal how specialization–diversification trade-offs change across scales.

This gap is especially relevant when connecting firm-level dynamics to regional patterns. If firms specialize due to internal coherence, while regions diversify because they must mobilize a broader capability base, then the spatial scale at which these dynamics are observed becomes an empirical question. Recent work has begun to highlight this multi-scale tension, but a fully integrated analysis remains lacking. By constructing a continuous spatial transition and examining how diversification–specialization patterns evolve across scales, the present work contributes directly to this underdeveloped area.

Beyond the substantive question of "what is the right region?", empirical results can change when the same underlying phenomenon is represented on different spatial supports. This sensitivity is commonly discussed under the *modifiable areal unit problem* (MAUP), which captures both a *scale effect* (changing the level of aggregation) and a *zoning effect* (changing the partition at a given scale)[283]. From a statistical perspective, aggregation/disaggregation can be viewed as a transformation of a latent spatial process through a measurement operator defined by the chosen territorial partition; consequently, inference may conflate the scale of the process with the scale of observation. In this spirit, Arbia and Espa[12] frame territorial databases as transformed images of an underlying "original" geographical process and organize a taxonomy of transformations across point, areal and continuous representations, highlighting how analytic choices can introduce distortions that resemble (or compound) MAUP-type effects.

Methodologically, changing spatial support often requires *cross-areal conversion* (or *areal interpolation*), where values observed on a *source* partition S must be transferred to a *target* partition T with different boundaries covering the same territory[12,287]. A key distinction concerns whether the conversion should respect a conservation constraint ("mass preservation" or *pycnophylactic* property) and whether the variable is extensive (additive over area) or intensive (e.g., rates), which typically requires converting to a compatible extensive form prior to interpolation[148,227]. These issues are pervasive in official statistics and GIS integration because administrative boundaries, sampling frames, and analytic units often do not coincide[287].

Classic areal interpolation methods range from *simple* to *model-based* approaches. Simple methods include areal weighting/map overlay, which redistributes source totals proportionally to intersection areas[148]. Volume-preserving methods enforce mass conservation while producing smoother internal surfaces; the canonical example is Tobler's pycnophylactic interpolation[353], discussed in broader reviews of spatial interpolation and areal methods[227]. More "intelligent" approaches exploit auxiliary information (e.g., land use, covariates) and statistical models, including EM-based formulations that treat target-zone values as missing data[95,129]. In this line, Bee and Espa[39] propose an EM algorithm for continuous variables that explicitly leverages auxiliary variables to improve disaggregation and reconcile incompatible zonal systems.

A further step is to represent the change-of-support problem within a fully probabilistic framework that encodes spatial dependence. Panzera[287] centers on the Bayesian Interpolation Method (BIM) introduced by Benedetti and Palma[43], which combines an aggregation matrix with a spatial prior (e.g., CAR-type dependence) to disaggregate areal totals while accounting for autocorrelation. Related families in-

clude geostatistical area-to-point approaches[222] and methods for linking misaligned aggregated datasets through geostatistical modeling[153]. More broadly, hierarchical Bayesian models provide a principled route to combining information across scales by defining a latent process and explicit measurement models at each resolution[366].

These issues are directly relevant for H3-based multi-scale analysis. A hierarchical hexagonal grid offers a controlled family of nested supports, which is useful for diagnosing scale effects and running systematic sensitivity analyses; however, MAUP does not disappear, and comparisons with administrative geographies still require careful change-of-support treatment[12,283,287]. In practice, H3 aggregation can be complemented with cross-areal interpolation when translating estimates to non-nested target partitions, choosing among mass-preserving overlays, EM/regression-based conversion, or Bayesian approaches depending on whether conservation constraints, auxiliary variables, and spatial dependence are central to the application[39,287].

4.3 Dataset

The data that we used have been produced by Istat, by looking at data collected for tax purposes and official registries for the firms and at customs for the export data. The data gathered have been used to create the following ASIA databases[1] for each year:

1. ASIA businesses which contains the information about the main structural characteristics of each business including the region, province, postal code and address of the firms;

2. ASIA economical results with which contains information about the economical performances of the firms like their revenues, the costs and the added value;

3. TECFRAME-SBS which contains information about exporting firms like what they export and where they export;

4. ASIA occupation which contains information about each one of the employees of the firms including the academic qualifications, in particular if he/she has a bachelor or master degree and what kind of degree he/she has, and what

[1]Notably, these datasets do not contain information about the agricultural enterprises that are contained in a different statistical registry and that would be extremely hard to properly georeference given that they are mainly composed by individual farmers whose address can be extremely far away from the fields where the actual production happens

percentage of the year the employee has worked in the firm as a fraction, namely working for the whole year counts as 1 and working just for six months counts as 0.5.

These different statistical registries have then been unified as a single dataset of by the Italian chambers of commerce study center Guglielmo Tagliacarne which also geo-referenced all the firms using the address of the headquarters of the firms. This procedure has been repeated for several years, in particular here the data for the years from 2015 to 2019 will be used, creating five datasets of around 4.4 millions firms each [2].

The data used in this work can be used to create usefull descriptive statistics like the one presented in Figure 4.1 with the scolarisation level of the workforce for the last available year.

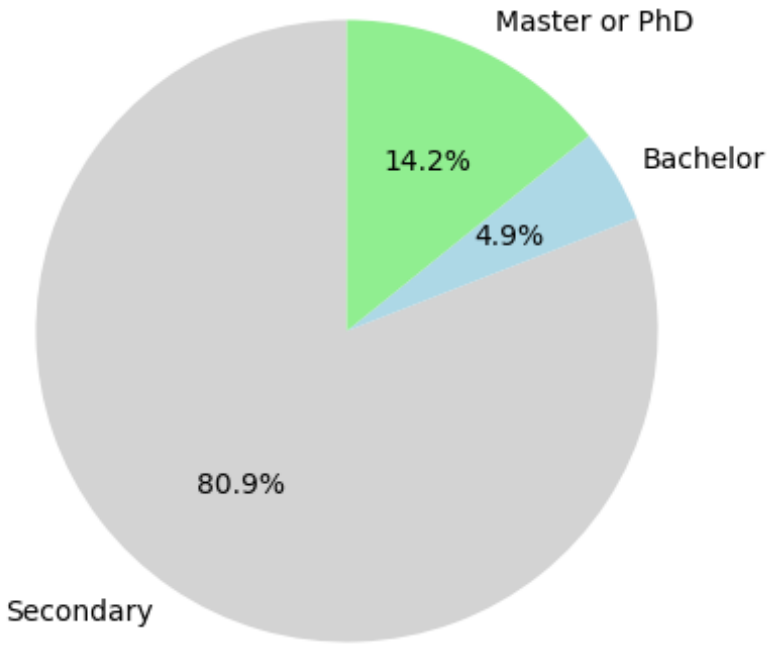


Figure 4.1: Levels of scolarisation for the year 2019

[2]The geo-referencing process was impossible for about 2 – 3% of firms whose precise latitude and longitude could not be computed and they have been discarded

4.4 Methodology

In this section we present the methods that we implemented, first we will present the H3 spatial index used to create a functional partition of the territory, then we will introduce the problem of quantifying the diversification of academic qualifications and how we approached this problem using Entropy, lastly we will introduce Exogenous Fitness and how we used it to create a capability measure that can be scaled from the single firm to the region size hexagons.

4.4.1 H3 spatial index

In this section we present the H3 spatial index, the hexagonal grid and how the firms have been grouped in the hexagons.

The H3 spatial index by Uber[54] is a geospatial indexing system designed to efficiently handle spatial data across various scales.
It uses a hexagonal grid system that offers some advantages over traditional square grids, for example the hexagons naturally tile the earth with fewer distortions than squares, especially near the poles, which makes H3 ideal for applications that require consistent area representations across latitudes.
This hexagonal grid enables the division of the earth's surface into finer resolution levels, allowing flexible scaling from large regions down to very fine details.
There are 16 resolution levels in total ranging from 0, the largest one, up to 15. Table 4.1 sums up the characteristics of the hexagons.

For each resolution level, the hexagons cover the entire surface of earth and each hexagon has its own unique identifier (index) that represents its location and level of detail, making it effective for spatial analysis, data visualization. It is also possible to move through the resolution levels because each hexagon is uniquely associated to its "parent" hexagon at a lower resolution level[3]. Given that the firms have been geo-referenced it is possible to associate to each firm's latitude and longitude, for any given level of resolution, the corresponding hexagon and cumulate the variable associated to the same hexagon. For instance, the number of employees of an hexagon is computed as the sum of the number of employees of each firm within the hexagon.

[3]This in theory could be used to pass from one scale to the other given that all the quantities of interest presented in this work are additive in the sense that the value of one hexagon can be computed by adding up the corresponding value of the "children". In practice, to avoid any possible error and to decrease the computational effort, all the computations have been performed iteratively for each resolution level and than they have been put together.

Resolution	Average Area in km^2	Average Edge Length in km	Number of Hexagons	Average number of employees per hexagon	Median number of employees per hexagon	Average number of firms per hexagon	Median number of firms per hexagon
0	4357449.42	1107.71	15	5388157.90	357134.29	1286508.07	125908.00
1	609788.44	418.67	20	4035451.31	2929726.20	962936.00	650917.00
2	86801.78	158.24	60	1345238.64	355113.79	320964.45	116065.00
3	12393.43	59.81	255	316926.84	112537.70	75668.04	42430.00
4	1770.35	22.61	1113	72579.88	25835.46	17323.68	7910.00
5	252.90	8.54	6405	12615.43	3267.68	3011.55	1071.00
6	36.13	3.23	34536	2339.35	440.41	558.30	153.00
7	5.16	1.22	164948	489.84	59.68	116.86	22.00
8	0.74	0.46	637254	126.82	12.82	30.27	5.00
9	0.11	0.17	1711404	47.22	7.36	11.26	3.00
10	0.02	0.07	4084633	19.78	4.33	4.71	2.00
11	0.002	0.02	7582633	10.64	2.93	2.52	1.00
12	0.0003	0.009	10226124	7.89	2.00	1.86	1.00
13	0.00004	0.0036	11659432	6.91	2.00	1.63	1.00
14	0.000006	0.0013	12056302	6.68	2.00	1.57	1.00
15	0.000001	0.0005	12133332	6.64	2.00	1.56	1.00

Table 4.1: The different resolution levels that can be created using H3 spatial indexing. Very low levels of resolution create hexagons not well centered on the country, leading to noisy results and had to be discarded. Very high levels of resolution create hexagons with just one firm per hexagon and lead to redundant results and therefore will be omitted in the analysis. Note that given the non perfectly spherical shape of earth some hexagons covering the world can have different size, this is not relevant when considering adjacent hexagons over a single country like in our case but to be precise we still called it the "average" area of the Hexagon.

The variables that have been computed this way for each hexagon are the added value, the overall number of employees, the number of employees for each kind of academic qualification, and the exogenous fitness as defined in 4.4.3.

The effect of the H3 spatial indexing can be simply modeled by introducing an indicator function $\mathbb{1}_{hf}^{(r)}$ where $h = 1, ..., N(r)$, $f = 1, ..., N_f$ and the total number of Hexagons $N(r)$ is a function of the resolution level r and the total number of firms is N_f. The indicator function is simply defined as follows:

$$\mathbb{1}_{hf}^{(r)} := \begin{cases} 1, & \text{if firm } f \text{ falls within hexagon } h \text{ at resolution level } r \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

With this definition, starting from the firm level value of any variable X_f and a certain resolution level r, it is possible to obtain the hexagonal aggregate $\tilde{X}_h^{(r)}$ value simply as

$$\tilde{X}_h^{(r)} = \sum_f \mathbb{1}_{hf}^{(r)} X_f \tag{4.2}$$

This aggregation procedure can be computed for all firm level variables and used to create composite variables, for instance the aggregated added value for the hexagon h at resolution r, $\tilde{Y}_h^{(r)}$ can be used in combination with the aggregate size of workforce $\tilde{L}_h^{(r)}$ to compute the LP

$$\widetilde{LP}_h^{(r)} = \frac{\tilde{Y}_h^{(r)}}{\tilde{L}_h^{(r)}} \tag{4.3}$$

This kind of analysis can be repeated for each resolution level and each year[4]. The productivity measure is shown in Figure 4.2 for the year 2019 and resolution level 9.

4.4.2 Entropy of academic qualification

In this section, we look at metrics of diversification of academic qualifications of the workforce (These data include information on employees, independent employees, temporary employees, and external employees).
In order to quantify the degree of diversification of the workforce we start by counting the number of employees within each firm for each education level and scientific-disciplinary group (these groups are available only for people with bachelor and master level of education), we present in Figure 4.3 the aggregate for the entire workforce and all the groups for one year. The number of employees for each education group and each firm E_{gf} can be grouped similarly to Eq. 4.2

$$\tilde{E}_{hg}^{(r)} = \sum_f \mathbb{1}_{hf}^{(r)} E_{gf} \tag{4.4}$$

[4]We avoided introducing another index just for the year to avoid making the notation too heavy, we will introduce it later when the data from different years are put together in a panel

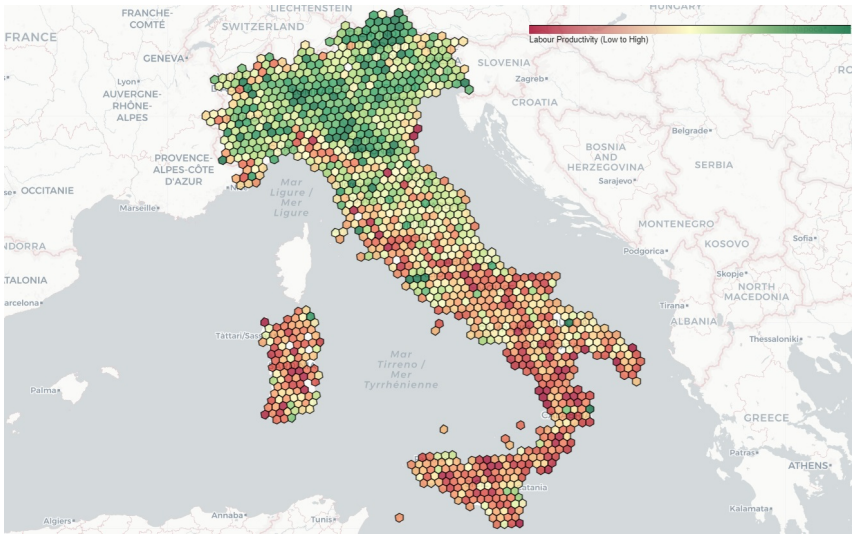


Figure 4.2: Levels of LP for the year 2019 and resolution level 9. It is possible to clearly identify the North-South divide in the country, with the highly productive areas in the North, which benefit from positive spillover effects, and the less productive South, with the exception of urban areas.

and the fraction f_g of the workforce of with a certain kind of education at resolution r can be simply computed as

$$f_{hg}^{(r)} = \frac{\tilde{E}_{hg}^{(r)}}{\sum_{g'} \tilde{E}_{hg'}^{(r)}} \tag{4.5}$$

We can simply use these frequencies to compute the Entropy of the academic qualification as

$$H_h^{(r)} = -\sum_g f_{hg}^{(r)} \log f_{hg}^{(r)} \tag{4.6}$$

One important aspect that we addressed was the fact that it is possible that the scolarisation $S_h^{(r)}$ level itself could play a key role; for instance, it could be possible that a highly educated workforce could influence positively the economic performance, irrespective of what kind of education it received. To address this issue, another "scolarisation" variable has been created that gives the fraction of the workforce of the hexagon that has more than a secondary school level of education. This variable can be simply computed starting from the fraction of the workforce that has a secondary education or lower, namely

$$S_h^{(r)} = 1 - f_{sec,h}^{(r)} \tag{4.7}$$

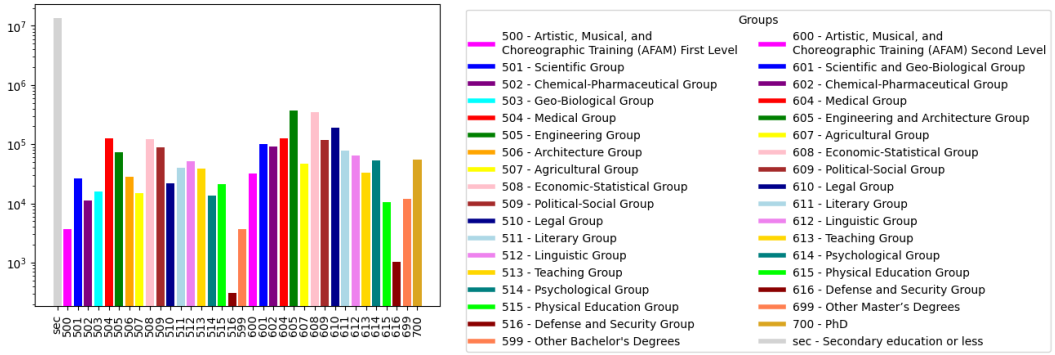


Figure 4.3: Distribution of academic qualifications for the year 2019, all the group codes starting with 5 refer to bachelor level education, groups starting with 6 refer to master level education, all secondary level education and below have been grouped together in a *sec* group.

4.4.3 Exogenous fitness

In this section we look at how we computed the economic complexity score of hexagons using the Exogenous Fitness approach. For each year the Economic Fitness and Complexity index at world level can be computed using the algorithm implemented in Tacchella et al.[349] and can be used to compute the exogenous Fitness as presented in Operti et al.[284]. The idea is that, starting from a q_{cp} matrix containing the the monetary export flows for country c for product p, it is possible to obtain the so-called RCA as follows:

$$RCA_{cp} = \frac{\frac{q_{cp}}{\sum_{c'} q_{c'p}}}{\frac{\sum_{p'} q_{cpp}}{\sum_{c',p'} q_{c'p'}}} \tag{4.8}$$

This RCA_{cp} can be transformed into an unweighted bipartite graph that imposes a threshold:

$$M_{cp} = \begin{cases} 1, & \text{if } RCA_{cp} \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

From this matrix it is possible to compute the Fitness F_c and the Complexity Q_p scores using the following iterative algoritm:

$$\begin{cases} \tilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)} \\ \tilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp} \frac{1}{F_c^{(n-1)}}} \end{cases} \Rightarrow \begin{cases} F_c^{(n)} = \frac{\tilde{F}_c^{(n)}}{\left\langle \tilde{F}_c^{(n)} \right\rangle_c} \\ Q_p^{(n)} = \frac{\tilde{Q}_p^{(n)}}{\left\langle \tilde{Q}_p^{(n)} \right\rangle_p} \end{cases} \tag{4.10}$$

where:

- $F_c^{(n-1)}$ and $Q_p^{(n-1)}$ denotes the fitness and complexity at iteration $n-1$;
- $\langle \tilde{F}_c^{(n)} \rangle_c$ represents the average value of $\tilde{F}_c^{(n)}$ computed across all countries;
- $\langle \tilde{Q}_p^{(n)} \rangle_p$ represents the average value of $\tilde{Q}_p^{(n)}$ computed across all products.

Assuming that the algorithm reaches convergence, the fitness (complexity) of each country(product) can be computed as

$$\begin{aligned} f_c &= \lim_{n\to\infty} F_c^{(n)} \\ Q_p &= \lim_{n\to\infty} Q_p^{(n)} \end{aligned} \tag{4.11}$$

These product complexities computed at world level give a good reference level for the complexity of the products at lower level and therefore can be used to compute the fitness of the hexagons. We say that the fitness of a hexagon is simply the sum of the complexities of the products exported from it, namely:

$$F_h^{(r)} := \sum_p \mathbb{1}_{hp}^{(r)} Q_p \tag{4.12}$$

where the indicator $\mathbb{1}_{hp}^{(r)}$ is defined as

$$\mathbb{1}_{hp}^{(r)} = \begin{cases} 1, & \text{if } \textit{at least one} \text{ firm from hexagon } h \text{ exports product } p \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

Once that fitness value has been computed for each hexagon and for each resolution level, it has been used to rank the hexagons like presented in[284] in percentile form[5].

To address the possibility that what matters is not *what* is exported but rather *if* anything is exported, a dummy variable has been introduced which is one if in the hexagon there is *at least one* exporting firm. This becomes extremely relevant for high resolution levels (i.e. small hexagons) where there can be one single firm per hexagon.

[5]We used the python cnde from the Pandas library Pandas.dataframe.rank(method='average', pct=True)

4.5 Results

4.5.1 Firm level analysis

Panel of firms

In this subsection we present the analysis performed by looking at the firm-level data. We performed a FE regression analysis on the unbalanced panel of firms, the regression looks like this:

$$\log(\mathrm{LP})_{f,t} = \beta_1 \log(\mathrm{size})_{f,t} + \beta_2 \,\mathrm{entropy}_{f,t} + \beta_3 \,\mathrm{scolarisation}_{f,t} + controls + \gamma_{t,s} + \varepsilon_{f,t} \tag{4.14}$$

The results are shown in Table 4.2. The dependent variable is the logarithm of the LP of the firms and the independent variables included the Entropy of the academic qualifications and a "scolarisation" variable with the fraction of workforce of the firm with a tertiary education, completely analogous to the one introduced in 4.7. The time fixed effect is based on the year t of the observation, and the sectorial fixed effect is based on the two-digit NACE[266] activity sector s of the firm. The control variables included in the regression are

1. the logarithm of the size of the firm measured as the number of employees of the firm;
2. a flag stating if the firm exports its products abroad[6];
3. the age (in years) of the firm;
4. the fraction of females employees over the total number of employees in the firm;
5. the fraction of employees under 30 years of age in the firm;
6. the fraction of employees over 50 years of age in the firm;
7. the fraction of employees born outside of Italy but within the EU over in the firm;
8. the fraction of employees born outside the EU in the firms;
9. the fraction of apprentices over the total number of employees;

[6] we didn't include the Fitness of the firms at this stage to avoid confusion, it will be included in the next stage when regressions with hexagons will be included

10. the fraction of managers over the total number of employees;

11. the fraction of executives over the total number of employees.

The errors of the regression have been computed as clustered errors, where the clusters are the years and the individual firms labels.

The regression results show that the diversification of academic qualifications contributes significantly and positively to the productivity of the firm, even when controlling for the fraction of workforce with a degree in the firm, which gives an overall estimate of how skilled is the workforce of the firm. This is particularly relevant because it shows that *ceteris paribus*, i.e.: with the same organizational structure of the firm and with the proportion of people with degrees in the firm, having a more diversified workforce leads to better productivity and performances. This is coherent with the idea that having a diversified workforce helps to develop a portfolio of competences that can help sustain the firm in its growth process.
The regression includes all the firms available in the panel with 16 or more employees. In the following subsection we will show why this threshold is far from arbitrary and that this skill diversification process has an even stronger effect as firm size grows.

Analysis for different sizes

We repeated the panel analysis presented in the previous section but dividing the employees in 10% quantiles of size from 16 employees onward. We had to cut at 16 because smaller firms have a very noisy and hetereogeneous behaviour and this threshold proved to be particularly relevant and far from random. There is a famous Italian law called The Workers' statute[343] that protects and and makes harder to fire employees when firms have more than 15 workers. This forces firms that want to grow above this threshold to be more organized and coherently managed. Several subsequent laws followed similar structures and are based on the same threshold. This creates an (apparently arbitrary) change in behaviour for firms above or under this threshold.
In Figure 4.4 we present the regression coefficient of the Entropy variable, where the regression has been repeated each time for firms with different size interval.

4.5.2 Changing scale analysis

In this section we show the results of applying the H3 index to see how the regression results scale from the region-wide level to the firm level scale. This allows

Table 4.2: FE regressions on the panel of firms

	Log-LP		
	(1)	(2)	(3)
Log-size	0.089***	0.001	0.014***
	(0.004)	(0.004)	(0.003)
Entropy	-0.075**	0.064**	0.129***
	(0.029)	(0.018)	(0.014)
Scolarisation	0.350***	0.138***	0.166***
	(0.023)	(0.022)	(0.014)
Flag exporting		0.154***	0.102***
		(0.003)	(0.002)
Firm age		0.003***	0.003***
		(0.000)	(0.000)
Female fraction		-0.338***	-0.199***
		(0.014)	(0.006)
Under 30 fraction		-0.233***	-0.138***
		(0.018)	(0.021)
50 or more fraction		-0.080***	-0.052***
		(0.019)	(0.006)
EU extra Italy fraction		-0.003	-0.003
		(0.013)	(0.013)
Extra EU fraction		-0.188***	-0.115***
		(0.012)	(0.005)
Apprentices fraction		0.237***	0.115***
		(0.028)	(0.022)
Managers fraction		0.925***	0.741***
		(0.018)	(0.018)
Executives fraction		1.529***	1.364***
		(0.053)	(0.045)
Year	-	-	x
Activity sector	-	-	x
Observations	744527	744527	744527
S.E. clustered	by: Year+Firm label	by: Year+Firm label	by: Year+Firm label
R^2	0.034	0.331	0.422

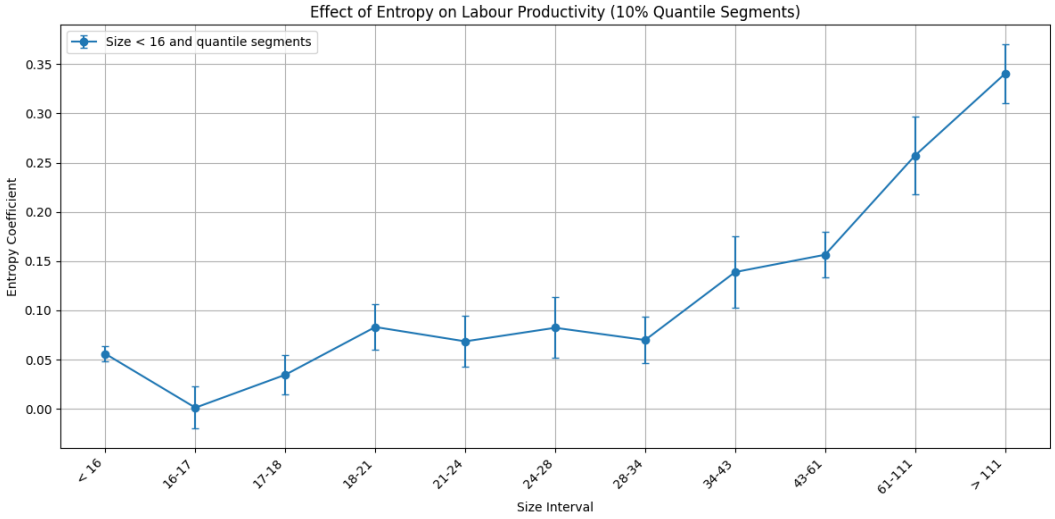
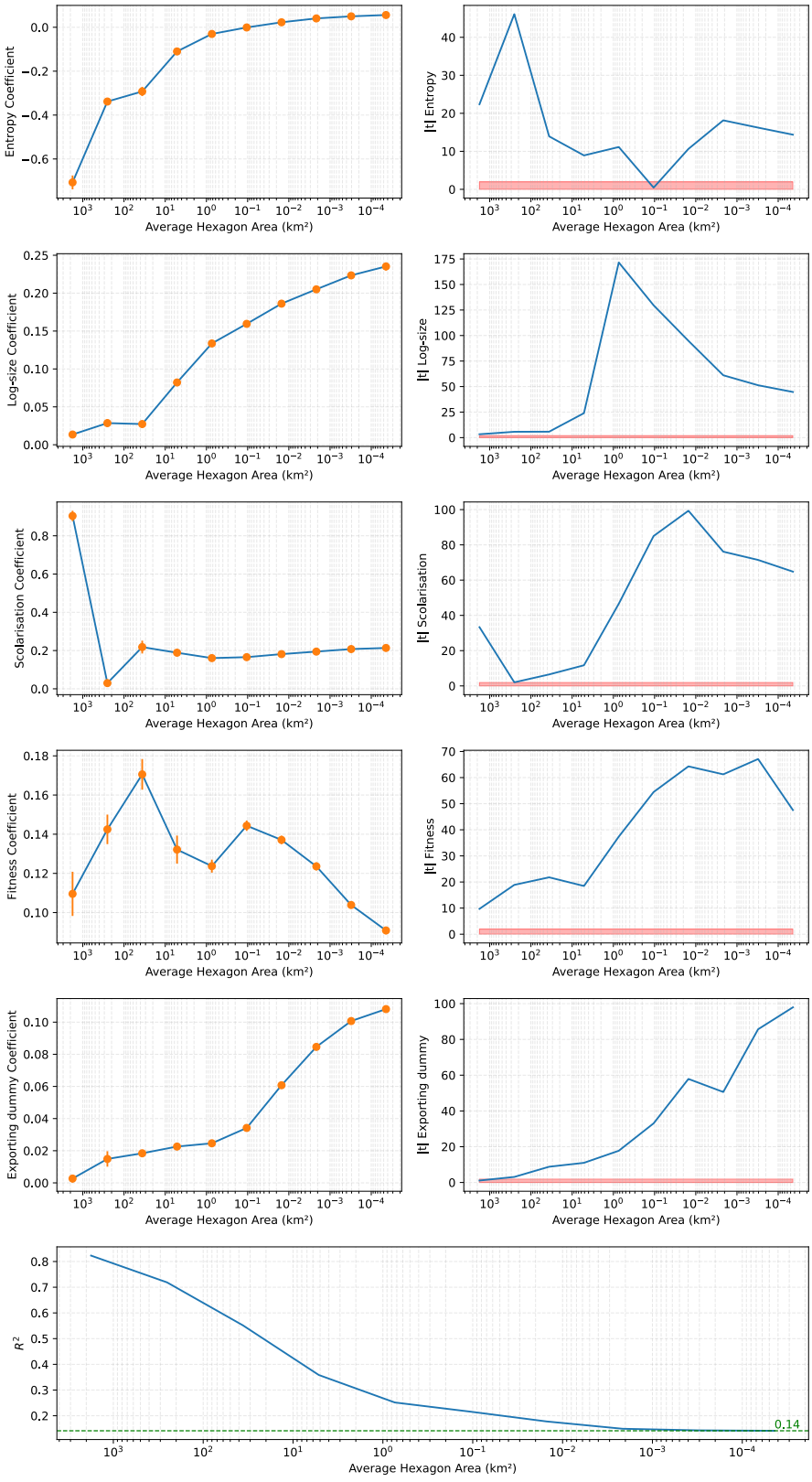


Figure 4.4: Regression β coefficients for Entropy with the associated error. Each regression in is made using firms in a certain size range, note that the number of employees is a fractional number because it takes into account part time employees, newly hired people, etc. The effect becomes larger and more significant as the size increases, testifying the increasing importance of diversification as firms scale up.

identification of how the impact of the variables of interest scales when looking at the phenomena at different levels of agglomeration. This is a different approach compared to the one presented in the previous sections because we do not look at how "tall" firms grow but rather how "dense" the ecosystem becomes. The regression is completely analogous to the one presented in the previous section but rather than having a panel of firms we have a panel of hexagons with fixed resolution r.
For instance, the LP has been computed as the quotient between the total added value of the firms within the hexagon and the total number of employees as described in Equation 4.3. Similarly all the control variables have been computed taking into account the fact that there are possibly many firms within an hexagon; for instance the age of the firms is computed by looking at the average of the age of all the firms within the hexagon. Similarly, the fraction of workers under 30 has been computed by summing the number of employees under 30 years of age within the hexagon and dividing it by the total number of employees within the hexagon.
The only differences between these regressions and the ones presented in the previous section are: the Exogenous Fitness computed as in 4.12; a "Counter" variable that counts how many firms are present within each hexagon and the absence of the sector fixed effect because there is more that one firm in each hexagon.
Each one of the regressions, for each different values of the resolution r, presented in Equation 4.15, takes the form:

Figure 4.5: Regression results at different scales vs the area of the hexagons. On the left the β coefficients, on the right the associated $|t|$ values and on the bottom the R^2.



$$\begin{aligned}\log(\widetilde{\mathrm{LP}})_{h,t}^{(r)} = \beta_1\ \log \mathrm{size}_{h,t}^{(r)} + \beta_2\, \mathrm{entropy}_{h,t}^{(r)} + \beta_3\, \mathrm{scolarisation}_{h,t}^{(r)} + \beta_4\, \mathrm{counter}_{h,t}^{(r)} \\ + \beta_5\, \mathrm{exog\_Fitness}_{h,t}^{(r)} + \beta_6\, \mathrm{exporting}_{h,t}^{(r)} + \beta_7\, \mathrm{average\_age}_{h,t}^{(r)} \\ + \beta_8\, \mathrm{Female\_fraction}_{h,t}^{(r)} + \beta_9\, \mathrm{Under\_30\_fraction}_{h,t}^{(r)} \\ + \beta_{10}\, \mathrm{50\_or\_more\_fraction}_{h,t}^{(r)} + \beta_{11}\, \mathrm{EU\_extra\_italia\_fraction}_{h,t}^{(r)} \\ + \beta_{12}\, \mathrm{extra\_EU\_fraction}_{h,t}^{(r)} + \beta_{13}\, \mathrm{Apprentice\_fraction}_{h,t}^{(r)} \\ + \beta_{14}\, \mathrm{Manager\_fraction}_{h,t}^{(r)} + \beta_{15}\, \mathrm{Executive\_fraction}_{h,t}^{(r)} \\ + \gamma_t^{(r)} + \varepsilon_{h,t}^{(r)}\end{aligned} \tag{4.15}$$

The results of the regressions are presented in Figure 4.5. Each point of the plots represent one regression with a different resolution r.
The graphs on the left side identify the β coefficients for some relevant variables with the associated error bars vs the average area of the hexagons. On the right there are the associated t values in absolute values to understand the statistical significance of the coefficients with the area of statistical non-significance at 5% highlighted in red. The errors have been computed with clustered errors on the year of the observations and therefore the significance levels are not affected by the sample size in the regression. On the bottom the R^2 of the regressions that decreases with the detail of the level of resolution. This indicates that larger hexagons average out a lot of the firm level heterogeneity. The first row with the Entropy graph shows that workforce diversification has a negative effect on productivity at low resolution(large hexagons) but when increasing resolution(smaller hexagons up to the individual firms) the impact becomes positive, this is coherent with the idea that on one hand territories have an advantage of having a coherent workforce specialized in a certain domain that can share skills, and on the other hand, on the contrary, firms have an advantage at diversifying the workforce and including different departments with different skills and purposes within their organization. The graph on the right in the first row shows that for a certain level of aggregation there is a transition from negative to positive, and the two effects balance out leading to an overall null effect. The second row presents the size effect on productivity, which corresponds in practice to the size of the overall workforce for a territory and the number of employees in a firm. The low resolution part shows that having a large workforce has a positive but relatively small and barely significant effect on the productivity of the territory. The high resolution shows the well known result that the number on employees has a huge impact on productivity, with larger firms more productive that smaller firms.

The third and forth row refer to the importance of exporting on productivity: the third row presenting the Exogenous Fitness of the hexagons which gives an idea of the importance of producing complex products on productivity; and the forth row presenting a dummy variable if at least one firm is exporting within the hexagon. The Fitness shows that what is exported matters for larger hexagons as it is a good proxy for skills and capabilities embedded within a territory, and it becomes less impactful for individual firms, this is coherent with the idea that product complexity is important for export but what is really important is to acquire the skills to export and acquire resources from the global market. This is confirmed by the fourth row, where having at least one firm exporting in the hexagon is trivially true (and non-significant) for territories but becomes fundamental for individual firms because it tells if the firm is capable of exporting its products and is competitive in the global market.
The last row presents the R^2 of the regressions; it shows that more aggregate data averages out the heterogeneity of individual firms, and therefore, just with the information available, it is possible to explain up to about 80% of the variance at the aggregate level. On the contrary looking at the individual firm the noise and heterogeneity become dominant, and the information available is able to explain only 14% of the total variance[7].

4.6 Conclusions

This chapter has investigated how the tension between specialization and diversification plays out differently for firms and for territories, moving "from macro to micro" through a continuous change of spatial scale. Building on a uniquely rich firm level dataset for Italy, we connected individual workers, firms and regions within a unified framework. We did so by combining detailed information on academic qualifications, a geometric partition of space based on the H3 spatial index, and a capability measure derived from exogenous Fitness. This allowed us to characterise how functional diversification in the workforce and export based capabilities correlate with LP at different scales, from single firms to region-sized hexagons.

At the firm level, our panel regressions show that both the level and the diversification of academic qualifications are robustly associated with higher LP, even after controlling for firm size, age, exporting status, workforce composition and sector-year FE. Importantly, the impact of the entropy of academic qualifications becomes

[7]Note that this firm level R^2 is lower that the one for the regressions in 4.2 because here we did not impose any cut-off on the size of the firms and therefore the numerous micro-firms which are extremely heterogeneous lead to a smaller R^2.

stronger as firm size increases. In other words, conditional on having a sufficiently structured organization (above the 15-employee threshold), firms appear to benefit from combining diverse types of human capital within their boundaries. This is consistent with the view of firms as carriers of coherent yet multi-functional capabilities: a diversified internal knowledge base supports more complex organizational structures and more productive use of resources.

The multi scale analysis using H3 hexagons reveals that these relationships are not scale invariant. When we aggregate firms into larger and larger hexagons, the sign and magnitude of key coefficients change in a systematic way. Workforce diversification, as captured by the entropy of academic qualifications, tends to have a negative or weak effect on productivity at coarse resolutions (large hexagons approximating regions), while turning positive at fine resolutions approaching individual firms. This suggests that, for territories, having a more specialised pool of skills can be beneficial, in line with arguments about localized specialization and coherent industrial structures. For firms, by contrast, internal diversification of competences is advantageous. The apparent paradox between diversified regions and specialised firms is thus resolved by explicitly accounting for scale: what is diversification at one level may correspond to specialization at another.

A similar pattern emerges for export-related variables. The exogenous Fitness index, built from world level product complexities and aggregated at the hexagon level, is strongly associated with productivity at larger spatial scales, where it proxies for the capability portfolio of the local economy. Its role becomes less important as we move to very fine resolutions. At the firm level, what matters is less *which* complex products are exported in the surrounding area and more *whether* the firm itself is able to export. Accordingly, the simple exporting dummy is weakly informative at coarse scales (where almost all productive territories contain some exporters) but becomes more relevant at the firm scale. Together with the scale-dependent behaviour of the R^2, these results indicate that aggregation smooths out a substantial share of micro-level heterogeneity and makes macro patterns easier to explain with a small set of variables.

Overall, the chapter contributes to the literature in three main ways. First, it provides firm-level evidence that the diversification of workers' academic qualifications is associated with higher productivity, above and beyond the average education level, and that this effect grows with firm size. Second, it offers a systematic, geometric multi-scale analysis of specialization–diversification trade-offs, showing how the sign and significance of key relationships change when moving from firms to regions. Third, it operationalizes exogenous Fitness as a scalable capability measure that can be consistently applied from the export basket of individual firms to that

of region-sized hexagons, thereby linking firm capabilities, regional complexity and productivity within a single empirical framework.

Our approach also has limitations. The entropy measure captures variety in formal academic qualifications, but does not fully reflect task-based skills, on-the-job learning or informal competences. The exogenous Fitness index is derived from export data and therefore focuses on tradable activities, potentially underrepresenting local services and non-tradable sectors that are nonetheless crucial for regional development. Moreover, while we use rich controls and FE, our regressions remain correlational and cannot definitively establish causal effects of diversification or capabilities on productivity. Finally, the use of H3 hexagons deliberately abstracts from administrative borders, which is a strength for scale analysis but makes direct comparison with policy-relevant units (regions, provinces) less immediate.

Despite these caveats, the findings provide a coherent picture of how complexity, capabilities and human capital interact across scales. Firms appear to grow by layering additional functions and competences, benefitting from diversified internal skill portfolios, while territories gain from hosting specialized yet complementary activity clusters that share compatible skill bases. The H3-based geometric approach shows that these patterns emerge gradually as we move across spatial resolutions, rather than being confined to a single "natural" unit of analysis.

Future research

Our work opens several paths for future research. A first natural extension is to move from static productivity levels to dynamic outcomes such as firm growth, entry and exit, and regional upgrading trajectories. This would allow us to test whether diversified workforces and higher exogenous Fitness predict not only higher contemporaneous productivity but also faster growth, greater resilience to shocks and more intense diversification into new activities over time. Relatedly, quasi-experimental strategies or instrumental-variable approaches could help move from correlation to causation, especially around institutional thresholds (such as the 15-employee rule) or policy changes affecting labour regulation and export conditions.

Second, the multi-scale framework could be applied to other countries and institutional contexts to assess the generality of our findings. Comparing economies with different labour-market institutions, education systems and export specialisations would clarify whether the scale-dependent trade-offs we document are specific to the Italian productive structure or represent a broader empirical regularity. Cross-country analyses using a harmonised H3 grid and comparable microdata could also shed light on how national policies shape the interplay between firm-level capabili-

ties and regional diversification.

Third, future work could refine the way skills and capabilities are measured. On the labour side, combining academic qualifications with task-based skill taxonomies, occupation-level information or data from job postings would offer a more granular view of the worker capabilities that matter for firm and regional complexity. On the production side, integrating product-based Fitness with technology or patent-based complexity measures would help disentangle the role of export complexity from that of underlying technological knowledge. Non-linear specifications and interaction terms could also be explored to capture threshold effects and complementarity between specialization and diversification.

Finally, the geometric perspective on space can be pushed further. Here we have treated H3 hexagons as independent units, but the underlying grid naturally defines adjacency and neighbourhood relations. Future research could study how capabilities and skills diffuse across neighbouring hexagons, or how commuting patterns and supply chains interact with the geometric partition of space. This would bring the analysis closer to network-based representations of regional systems while retaining the clarity of a scale-explicit approach. On the policy side, translating our multi-scale findings into guidance for smart specialization strategies, for instance, by identifying the scales at which diversification or specialization policies are most effective, represents an important avenue for collaboration between researchers and policymakers.

"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

5 Conclusions

This thesis has investigated how productive and innovative activity is organised across space and across levels of aggregation, and how such structure can be measured and analysed with modern data and methods. Starting from the view of economic systems as complex, spatially embedded configurations of heterogeneous agents, capabilities, and technologies, the thesis has pursued a unifying goal: to move beyond purely aggregate representations of regional development by linking macro-level indicators of complexity and diversification to micro-level data on firms and workers, while taking spatial dependence and uncertainty seriously.

A first, general lesson that emerges across the three empirical chapters is that the spatial and technological fabric of regional economies is strongly heterogeneous, and that this heterogeneity is only partially visible through administrative aggregates or standard complexity indicators. Macro measures remain informative as descriptive summaries and as comparative benchmarks, but they can conceal distinct local configurations of capabilities. The multi-scale perspective adopted in this thesis therefore supports a shift from treating regions as homogeneous units to modelling them as layered systems made of firms, labour markets, and localised networks of related activities.

Chapter 2 contributes to this agenda by developing a graph-based ensemble clustering framework for innovative startups in Lombardy. The combination of spatial bootstrap resampling with consensus clustering, encoded as a bipartite graph

between firms and cluster labels and solved through bipartite modularity maximisation, provides a transparent way to separate stable cluster structure from algorithmic and sampling-induced variability. Substantively, the chapter shows how innovation ecosystems can be characterised as spatially and sectorally differentiated communities, and how the resulting partitions can be related to observable dimensions of firm performance and innovation. Methodologically, it demonstrates how ideas from network science can be used to obtain clustering results that are both interpretable and robust in spatial microdata settings.

Chapter 3 focuses on the methodological challenges raised by Spatial ML on firm-level microdata, particularly the need for uncertainty quantification when observations are spatially dependent and covariates are high-dimensional. The proposed pipeline combines entity embeddings and deep clustering to construct strata that are jointly meaningful in attribute space and coherent in geographic space, and it uses these strata within a stratified spatial bootstrap procedure. By comparing uncertainty estimates and variable-importance patterns, the chapter illustrates why naive validation and resampling schemes can be misleading in Spatial ML applications, and how bootstrap design can be integrated into the modelling workflow rather than treated as an afterthought.

Chapter 4 moves from a purely firm-centred view to an explicitly multi-scale framework linking workers, firms, and territory. By combining linked employer–employee information with export-based complexity metrics and a hexagonal grid representation of Italy, the chapter examines how worker diversity, firm characteristics, and regional productive structure co-vary as one moves from firms to local clusters and broader areas. Using a continuous-space partition reduces the need for administrative boundaries and makes it easier to compare things at different spatial resolutions. The chapter clarifies how the specialisation–diversification tension manifests differently across scales, and it provides a concrete bridge between capabilities as inferred from macro networks and capabilities as embodied in workers and organisations.

Taken together, the thesis makes three main contributions. First, it contributes methodologically by showing how resampling-based inference can be adapted to the realities of spatial microdata. In both the ensemble clustering setting (Chapter 2) and the deep clustering setting (Chapter 3), bootstrap procedures are used not only for variance estimation, but as tools to assess the stability of learned structures and the robustness of model-based conclusions under spatial dependence.

Second, it contributes a set of practical, interpretable frameworks that connect ML flexibility with economic structure: bipartite graph representations and modularity-based consensus for clustering; hexagonal grids for multi-resolution spa-

tial measurement; and representation learning as a way to handle high-cardinality firm attributes without abandoning transparency about the units and relationships being modelled.

Third, it contributes empirically to the documentation and interpretation of Italy's high-tech and innovative business ecosystem, with particular emphasis on Lombardy. By integrating business registers and linked microdata with network-based complexity measures and spatial partitions, the thesis provides an evidence base that complements standard territorial statistics with micro-founded descriptions of local capability structures.

These contributions have direct implications for innovation and regional development policy, especially in the context of smart specialization. The results support policies that are sensitive to within-region heterogeneity and that build on place-specific capability combinations, rather than relying exclusively on coarse sectoral targets or aggregate rankings. They also motivate a more explicit treatment of uncertainty in empirical diagnostics used for policy design: when the stability of clusters, importance rankings, or inferred specialisation profiles depends on spatial dependence and sampling variability, robustness checks are not ancillary but central to credible evidence.

Several limitations also delineate avenues for future research. The analyses largely rely on observed and codified attributes of firms and workers, which only imperfectly proxy for informal knowledge flows, organisational practices, and institutional complementarities. Dynamic feedback mechanisms over longer horizons are only partially captured, suggesting extensions using richer longitudinal designs and explicit modelling of path dependence. Methodologically, promising directions include integrating the proposed spatial bootstrap logic with causal inference frameworks for policy evaluation, incorporating inter-firm collaboration and mobility networks to model capability diffusion more directly, and extending the network perspective with graph neural networks and other architectures that can represent multi-layer relations between firms, workers, technologies, and places.

In conclusion, this thesis advances the analysis of spatially embedded productive structures by linking complexity and relatedness concepts to microdata, and by embedding modern ML within workflows that remain attentive to spatial dependence, interpretability, and uncertainty. By combining multi-scale measurement with robust clustering and Spatial ML inference, it offers both a set of methodological tools and a coherent empirical perspective on how local capabilities are organised and how they shape regional economic evolution.

Bibliography

[1] M. Abadi. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. software available from tensorflow.org.

[2] D. Acemoglu, S. Johnson, and J. Robinson. Institutions as a fundamental cause of long-run growth. *Handbook of Economic Growth*, 1:385–472, 2005.

[3] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012.

[4] P. Aghion and P. Howitt. A model of growth through creative destruction. *Econometrica*, 60(2):323–351, 1992.

[5] P. Aghion, P. A. David, and D. Foray. Science, technology and innovation for economic growth: linking policy research and practice in 'stig systems'. *Research policy*, 38(4):681–693, 2009.

[6] A. Agresti. *Categorical data analysis*. Wiley, New York, 1990.

[7] AIDA. Database of the italian firms owned by bureau van dijk, 2024. URL `https://login.bvdinfo.com/R0/AidaNeo?SetLanguage=it`. (accessed on 23-08-2024).

[8] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[9] L. Anselin. *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht, 1988.

[10] L. Anselin. Local indicators of spatial association—lisa. *Geographical Analysis*, 27 (2):93–115, 1995.

[11] L. Anselin. Interactive techniques and exploratory spatial data analysis. *Geographical Information Systems: Principles, Techniques, Applications, and Management. New York Chichester: John Wiley*, 1999.

[12] G. Arbia and G. Espa. Metodologie statistiche per la disaggregazione di dati socio-economici a connotazione territoriale. Technical report, Università degli Studi di Trento and Università degli Studi "G. d'Annunzio" di Pescara, 1998. Working paper / manuscript (provided by the authors).

[13] G. Arbia. The role of spatial effects in the empirical analysis of regional concentration. *Geographical Systems*, 3:271–281, 2001.

[14] G. Arbia. *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Springer, Berlin, 2006.

[15] G. Arbia, G. Espa, D. Giuliani, and A. Mazzitelli. Detecting the existence of space-time clusters of firms. *Regional Science and Urban Economics, 40, 311–323*, 2010. URL https://doi.org/10.1016/j.regsciurbeco.2009.10.004.

[16] G. Arbia, G. Espa, D. Giuliani, and A. Mazzitelli. Clusters of firms in an inhomogeneous space: the high-tech industries in milan. *Economic Modelling, 29(1), 3–11*, 2012.

[17] N. S. Argyres and T. R. Zenger. Capabilities, transaction costs, and firm boundaries. *Organization Science*, 23(6):1643–1657, 2012.

[18] D. Arribas-Bel. Contextily: context geo tiles in python, 2023. URL https://contextily.readthedocs.io.

[19] L. Arsini, M. Straccamore, and A. Zaccaria. Prediction and visualization of mergers and acquisitions using economic complexity. *PLoS ONE, 18(4), e0283217*, 2023. URL https://doi.org/10.1371/journal.pone.0283217.

[20] W. B. Arthur. Complexity and the economy. *Science*, 284(5411):107–109, 1999.

[21] W. B. Arthur. *Complexity and the Economy*. Oxford University Press, Oxford, 2013.

[22] ASIA-SBS. Informations about these data can be found on the official istat website (accessed 19 may 2024), 2019. URL https://www.istat.it/it/archivio/165883.

[23] ATECO. Further information about ateco codes is available at (accessed 31 may 2024), 2015. URL https://www.istat.it/en/archivio/17959.

[24] S. Athey and G. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

[25] D. Atkin, A. K. Khandelwal, and A. Osman. Exporting and firm performance: Evidence from a randomized experiment. *Quarterly Journal of Economics*, 132(2): 551–615, 2017.

[26] D. B. Audretsch and M. P. Feldman. R&d spillovers and the geography of innovation and production. *American Economic Review*, 86(3):630–640, 1996.

[27] A. Bacilieri, A. Borsos, P. Astudillo-Estevez, and F. Lafond. Firm-level production networks: what do we (really) know. *INET Oxford Working Paper*, 2023, 2023.

[28] B. Balassa. Trade liberalisation and "revealed" comparative advantage. *The Manchester School*, 33(2):99–123, 1965.

[29] J. Baldwin and G. Gellatly. Are there high-tech industries or only high-tech firms? evidence from new technology-based firms. *Analytical Studies Branch, Micro-Economic Analysis Division, Statistics Canada (Research report style), 24–32*, 1998.

[30] P.-A. Balland and D. Rigby. The geography of complex knowledge. *Economic geography*, 93(1):1–23, 2017.

[31] P.-A. Balland, R. Boschma, J. Crespo, and D. L. Rigby. Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9):1252–1268, 2019. doi: 10.1080/00343404.2018.1437900. URL `https://doi.org/10.1080/00343404.2018.1437900`.

[32] P.-A. Balland, T. Broekel, D. Diodato, E. Giuliani, R. Hausmann, N. O'Clery, and D. Rigby. The new paradigm of economic complexity. *Research Policy*, 51:104450, 2022.

[33] D. R. Baqaee and E. Farhi. The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem. *Econometrica*, 87(4):1155–1203, 2019.

[34] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[35] M. J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.

[36] M. Batty. *Cities and Complexity*. MIT Press, Cambridge, MA, 2005.

[37] G. Becattini. The marshallian industrial district as a socio-economic notion. *Industrial Districts and Inter-Firm Co-operation in Italy*, pages 37–51, 1990.

[38] G. Becattini, M. Bellandi, and L. De Propris. *A Handbook of Industrial Districts*. Edward Elgar, Cheltenham, 2009.

[39] M. Bee and G. Espa. Metodi statistici per l'interpolazione areale: l'algoritmo EM per dati continui. *Statistica Applicata*, 11(3):466–492, 1999.

[40] E. D. Beinhocker. *The Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics*. Harvard Business School Press, Boston, 2006.

[41] M. Bell and K. Pavitt. Technological accumulation and industrial growth: Contrasts between developed and developing countries. *Industrial and Corporate Change*, 2 (2):157–210, 1993. doi: 10.1093/icc/2.2.157.

[42] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2): 608–650, 2014.

[43] R. Benedetti and D. Palma. Desegregation, interpolation and integration of spatial series. *Statistica*, 54(1):87–111, 1994.

[44] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[45] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? *ICDT*, pages 217–235, 1999.

[46] C. P. D. Birch, S. P. Oom, and J. A. Beecham. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological modelling*, 206 (3-4):347–359, 2007.

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[48] R. Boschma and R. Wenting. The spatial evolution of the british automobile industry: Does location matter? *Industrial and Corporate Change*, 16(2):213–238, 2007.

[49] R. Boschma and S. Iammarino. Related variety, trade linkages, and regional growth in italy. *Economic geography*, 85(3):289–311, 2009.

[50] R. Boschma. Towards an evolutionary perspective on regional resilience. *Regional Studies*, 49(5):733–751, 2015.

[51] R. Boschma. Relatedness as driver of regional diversification: A research agenda. *Regional Studies*, 51(3):351–364, 2017.

[52] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

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

[54] I. Brodsky. H3: Uber's hexagonal hierarchical spatial index, 2018. URL `https://eng.uber.com/h3`. Available from Uber Engineering website [22 June 2019].

[55] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017.

[56] C. Brunsdon, A. S. Fotheringham, and M. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28 (4):281–298, 1996.

[57] S. Brusoni, A. Prencipe, and K. Pavitt. Knowledge specialization, organizational coupling, and the boundaries of the firm: why do firms know more than they make? *Administrative science quarterly*, 46(4):597–621, 2001.

[58] G. Cainelli and N. De Liso. Innovation in industrial districts: Evidence from italy. *Industry and Innovation*, 12(3):383–398, 2005.

[59] A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

[60] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. *PAKDD*, pages 160–172, 2013.

[61] J. Cantwell and G. Vertova. Historical evolution of technological diversification. *Research Policy*, 33(3):511–529, 2004. doi: 10.1016/j.respol.2003.10.003.

[62] J. Castela Forte, G. Yeshmagambetova, M. L. van der Grinten, B. Hiemstra, T. Kaufmann, R. J. Eck, F. Keus, A. H. Epema, M. A. Wiering, and I. C. C. van der Horst. Identifying and characterizing high-risk clusters in a heterogeneous icu population with deep embedded clustering. *Scientific Reports, 11(1), 12109*, 2021. URL `https://doi.org/10.1038/s41598-021-91297-x`.

[63] P. Catalán, C. Navarrete, and F. Figueroa. The scientific and technological cross-space: Is technological diversification driven by scientific endogenous capacity? *Research Policy*, 51(8):104016, 2022. doi: 10.1016/j.respol.2020.104016.

[64] Centro Studi Guglielmo Tagliacarne. Local labour systems and industrial districts of italy. `https://www.tagliacarne.it`, 2019.

[65] Centro Studi Guglielmo Tagliacarne. Institutional mission and research activities. `https://www.tagliacarne.it`, 2020.

[66] Centro Studi Guglielmo Tagliacarne. Territorial economy statistics. `https://www.tagliacarne.it`, 2021. Territorial socio-economic indicators.

[67] Centro Studi Guglielmo Tagliacarne. Value added of italian provinces. `https://www.tagliacarne.it`, 2021.

[68] Chambers of commerce of Italy. Italian registry of businesses by the italian chambers of commerce, 2024. URL `https://www.registroimprese.it/`. (accessed on 23-08-2024).

[69] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Clustgeo: An r package for hierarchical clustering with spatial constraints. *Computational Statistics 33(4): 1799-1822*, 2018.

[70] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 2014.

[71] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[72] J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, Hoboken, 2012.

[73] S. Choudhury and N. Pal. Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, 182, 07 2019.

[74] W. Christaller. *Die zentralen Orte in Süddeutschland*. Gustav Fischer, Jena, 1933.

[75] M. Cimoli, G. Dosi, and J. E. Stiglitz. *The Political Economy of Capabilities Accumulation*. Oxford University Press, Oxford, 2010.

[76] I. T. Çınar, I. Korkmaz, and T. Baycan. Regions' economic fitness and sectoral labor productivity: Evidence from turkey. *Regional Science Policy & Practice*, 14(3): 575–599, 2022.

[77] X. Cirera and W. F. Maloney. *The innovation paradox: Developing-country capabilities and the unrealized promise of technological catch-up*. The World Bank, 2017.

[78] A. D. Cliff and J. K. Ord. *Spatial Processes: Models and Applications*. Pion, London, 1981.

[79] A. Coad, N. Mathew, and E. Pugliese. Positioning firms along the capabilities ladder. Working Paper 2021-031, UNU-MERIT, 2021. UNU-MERIT Working Paper Series.

[80] W. G. Cochran. Sampling Techniques, 3rd ed. *John Wiley & Sons*, 1977.

[81] M. G. Colombo and M. Delmastro. How effective are technology incubators? evidence from italy. *Research Policy*, 31:1103–1122, 2002.

[82] Complexity Science Hub Vienna. Complexity science hub vienna. `https://csh.ac.at/`, 2025. Accessed 18-12-2025.

[83] Complexity Science Hub Vienna. Supply chain science. `https://csh.ac.at/research/research-topic/supply-chain-science/`, 2025. Accessed 18-12-2025.

[84] N. Cressie. *Statistics for Spatial Data*. Wiley, New York, revised edition, 1993.

[85] M. Cristelli, A. Tacchella, M. Z. Cader, and L. Pietronero. On the predictability of growth. *EPJ Data Science*, 6(1):1–20, 2017.

[86] D. Czarnitzki and J. Delanote. R&d policies for young smes: input and output effects. *Small Business Economics*, 2015.

[87] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[88] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997.

[89] F. De Cunzo, A. Sbardella, and L. Pietronero. Economic complexity and sustainability: A network-based perspective. *Sustainability*, 12(4):1604, 2020.

[90] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature communications*, 6(1):6868, 2015.

[91] J.W.T.M. de Kok, F. van Rosmalen, J. Koeze, F. Keus, S.M.J. van Kuijk, J. Castela Forte, R.M. Schnabel, R.G.H. Driessen, T.T.W. van Herpt, J.-W.E.M. Sels, et al. Deep embedded clustering generalisability and adaptation for integrating mixed datatypes: two critical care cohorts. *Scientific Reports*, 14(1):1045, 2024.

[92] V. De Stefano, M. Mula, M. S. Mariani, and A. Zaccaria. From macro to micro: Economic complexity indicators for firm growth, 2025.

[93] N. J. De Vos. Kmodes categorical clustering library, 2015. URL `https://github.com/nicodv/kmodes`.

[94] R. M. del Rio-Chanona, P. Mealy, A. Pichler, F. Lafond, and J. D. Farmer. Supply and demand shocks in the covid-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy*, 36(Supplement_1):S94–S137, 2020.

[95] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

[96] M. M. Dickson, G. Espa, R. Gabriele, and A. Mazzitelli. Small businesses and the effects on the growth of formal collaboration agreements: additional insights and policy implications. *Applied Economics*, 2021.

[97] P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, 3 edition, 2013.

[98] D. Diodato, R. Hausmann, and U. Schetter. A simple theory of economic development at the extensive industry margin. HKS Working Paper RWP22-016, 2022.

[99] D. Diodato, L. Napolitano, E. Pugliese, and A. Tacchella. Economic complexity for regional industrial strategies. Working paper, European Commission, Joint Research Centre, 2023. URL `https://publications.jrc.ec.europa.eu/repository/handle/JRC136443`.

[100] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[101] G. Dosi. Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy*, 11 (3):147–162, 1982. doi: 10.1016/0048-7333(82)90016-6.

[102] G. Dosi, M. Grazzi, and D. Moschella. What do firms know? what do they produce? a new look at the relationship between patenting profiles and patterns of product diversification. *Small Business Economics*, 48(2):413–429, 2017.

[103] G. Dosi, N. Mathew, and E. Pugliese. What a firm produces matters: Processes of diversification, coherence and performances of indian manufacturing firms. *Research Policy*, 51(8):104152, 2022. ISSN 0048-7333. doi: https://doi.org/10.1016/j.respol.2020.104152. URL `https://www.sciencedirect.com/science/article/pii/S0048733320302274`. Special Issue on Economic Complexity.

[104] G. Dosi. *The Foundations of Complex Evolving Economies. Part One*. Oxford University Press, Oxford, 2023. doi: 10.1093/oso/9780192865922.001.0001.

[105] N. R. Draper and H. Smith. Applied Regression Analysis. *John Wiley & Sons*, 1998.

[106] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.

[107] G. Duranton and D. Puga. Micro-foundations of urban agglomeration economies. *Handbook of Regional and Urban Economics*, 4:2063–2117, 2004.

[108] G. Duranton and H. G. Overman. Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4):1077–1106, 2005.

[109] G. Duranton and D. Puga. Urban land use. *Handbook of Regional and Urban Economics*, 5:467–560, 2015.

[110] S. N. Durlauf. Complexity and empirical economics. *Economic Journal*, 115(504): F225–F243, 2005.

[111] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[112] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993.

[113] L. Einav and J. Levin. Economics in the age of big data. *Science*, 346(6210), 2014.

[114] J. P. Elhorst. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer, Heidelberg, 2014.

[115] G. Ellison and E. L. Glaeser. Geographic concentration in u.s. manufacturing industries. *Journal of Political Economy*, 105(5):889–927, 1997.

[116] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press*, 1996.

[117] European Commission. Guide to research and innovation strategies for smart specialisation (ris3). `https://ec.europa.eu/regional_policy/en/information/publications/guides/2012/guide-to-research-and-innovation-strategies-for-smart-specialisation`, 2012. Published 2012; accessed 18-12-2025.

[118] European Council. Council regulation (eec) no 696/93 of 15 march 1993 on the statistical units for the observation and analysis of the production system in the community. `https://eur-lex.europa.eu/eli/reg/1993/696/oj`, 1993. (Accessed: 07-09-2024).

[119] European Parliament and Council of the European Union. Regulation (ec) no 177/2008 establishing a common framework for business registers for statistical purposes. Official Journal of the European Union, L 61, 2008. Repealed and replaced earlier business register regulations.

[120] Eurostat. Business registers: Recommendations manual, 2010.

[121] Eurostat. High-tech industry and knowledge-intensive services (htec), 2024. URL `https://ec.europa.eu/eurostat/cache/metadata/en/htec_esms.htm`.

[122] B. Faber. Trade integration, market size, and industrialization: Evidence from china's national trunk highway system. *Review of Economic Studies*, 81(3):1046–1070, 2014.

[123] M. A. Farzammehr and S. Moradi. Classical and spatial cluster analysis of smuggling in iranian provinces. *Regional Science Policy & Practice, 16(2), 12609*, 2024. URL `https://doi.org/10.1111/rsp3.12609`.

[124] E. Felice. *Regional Value Added in Italy, 1891–2001*. Routledge, London, 2011.

[125] J. Felipe, U. Kumar, A. Abdon, and M. Bacate. Product complexity and economic development. *World Development*, 40(1):36–68, 2012.

[126] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003.

[127] X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. *Proceedings of the twenty-first international conference on Machine learning*, 2004.

[128] M. Fessina, A. Giambattista, T. Andrea, and Z. Andrea. Identifying key products to trigger new exports: an explainable machine learning approach. *Journal of Physics: Complexity, 5(2), 025,003*, 2024.

[129] R. Flowerdew and M. Green. Developments in areal interpolation methods and GIS. *The Annals of Regional Science*, 26(1):67–78, 1992. doi: 10.1007/BF01581481.

[130] D. Foray, P. A. David, and B. H. Hall. Smart specialisation: from academic idea to political instrument, the surprising destiny of a concept and the difficulties involved in its implementation, 2011.

[131] D. Foray. *Smart Specialisation: Opportunities and Challenges for Regional Innovation Policy*. Routledge, Abingdon, 2015.

[132] D. Foray, K. Morgan, and S. Radosevic. The role of smart specialisation in the eu research and innovation policy landscape. `https://ec.europa.eu/regional_policy/sources/brochure/smart/role_smartspecialisation_ri.pdf`, 2018. European Commission (DG REGIO); accessed 2025-12-18.

[133] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 2010.

[134] A. S. Fotheringham and D. W. S. Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.

[135] A. L. N. Fred and A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (6):835–850, 2005.

[136] C. Freeman. The 'National System of Innovation' in historical perspective. *Cambridge Journal of Economics*, 19(1):5–24, 1995. doi: 10.1093/oxfordjournals.cje.a035309.

[137] C. Freeman and L. Soete. *The Economics of Industrial Innovation: Third Edition*. MIT Press, 3 edition, 1997.

[138] K. Frenken, F. Van Oort, and T. Verburg. Related variety, unrelated variety and regional economic growth. *Regional Studies*, 41(5):685–697, 2007.

[139] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[140] M. Fujita, P. Krugman, and A. J. Venables. *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge, MA, 1999.

[141] M. Gertler. Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of Economic Geography, 3, 75–99*, 2003.

[142] A. Gervais, J. R. Markusen, and A. J. Venables. Regional specialization: From the geography of industries to the geography of jobs. *Canadian Journal of Economics/Revue canadienne d'économie, 57, 1236–1264*, 2024. URL `https://doi.org/10.1111/caje.12747`.

[143] D. Giuliani, D. Toffoli, M. M. Dickson, A. Mazzitelli, and G. Espa. Assessing the role of spatial externalities in the survival of italian innovative startups. *Regional Science Policy & Practice*, 16(1):12653, 2024.

[144] E. L. Glaeser. Learning in cities. *Journal of Urban Economics*, 46(2):254–277, 1999.

[145] A. Głodowska. The concept of high-tech firms and their role in the contemporary economy. *The Internationalization of High-Tech Firms: Patterns, Innovation and Research and Development, 6–35, Cambridge Scholars Publishing*, 2019.

[146] R. K. Gómez. Mapping obesity in women and chronic malnutrition in children across the municipalities of Bolivia: Spatial clusters and regionalization. *Research in Social Problems and Public Policy, 20*, 2024. URL `https://doi.org/10.1016/j.rspp.2024.100129`.

[147] F. F. Gonzales. State of the art on: Deep clustering, 2022.

[148] M. F. Goodchild and N. S.-N. Lam. Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing*, 1(3):297–312, 1980.

[149] M. F. Goodchild. Geographical information science. *International Journal of Geographical Information Systems*, 6(1):31–45, 1992.

[150] M. F. Goodchild. Reimagining the history of gis. *Annals of GIS*, 24(1):1–8, 2018.

[151] I. Goodfellow. *Deep Learning*. MIT Press, 2016. URL `http://www.deeplearningbook.org`.

[152] I. R. Gordon and P. McCann. Innovation, agglomeration, and regional development. *Journal of Economic Geography*, 5(5):523–543, 2005.

[153] C. A. Gotway and L. J. Young. A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, 16(1):115–135, 2007. doi: 10.1198/106186007X179257.

[154] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics, 27(4): 857–871*, 1971.

[155] O. Granstrand. Towards a theory of the technology-based firm. *Research Policy*, 27 (5):465–489, 1998. doi: 10.1016/S0048-7333(98)00067-5.

[156] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. *ACM SIGMOD Record, 27(2), 73–84*, 1998.

[157] C. Guo and F. Berkhahn. Entity embeddings of categorical variables. *arXiv:1604.06737*, 2016.

[158] J. Guo, J. Wang, C. Xu, and Y. Song. Modeling of spatial stratified heterogeneity. *GIScience & Remote Sensing, 59(1), 1660–1677.*, 2022. doi: 10.1080/15481603.2022.2126375. URL `https://doi.org/10.1080/15481603.2022.2126375`.

[159] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[160] P. Hall and B.-Y. Jing. On sample reuse methods for dependent data. *Journal of the Royal Statistical Society, Series B*, 58(4):727–737, 1995.

[161] S. S. Hamidi, E. Akbari, and H. Motameni. Consensus clustering algorithm based on the automatic partitioning similarity graph. *Data & Knowledge Engineering*, 2019.

[162] R. H. Hariri, E. M. Fredericks, and K. M. Bowers. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big data*, 2019.

[163] R. Hausmann and B. Klinger. Structural transformation and patterns of comparative advantage in the product space. Technical report, Center for International Development, Harvard University, 2006.

[164] R. Hausmann and B. Klinger. The structure of the product space and the evolution of comparative advantage. *CID Working Paper*, 2007.

[165] R. Hausmann, C. A. Hidalgo, S. Bustos, M. Coscia, S. Chung, J. Jimenez, A. Simoes, and M. A. Yıldırım. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. MIT Press, Cambridge, MA, 2014.

[166] R. Hausmann. The growth lab's approach to economic development. Harvard Growth Lab Working Paper, 2020.

[167] E. F. Heckscher. The effect of foreign trade on the distribution of income. *Ekonomisk Tidskrift*, 21:497–512, 1919.

[168] T. Hengl et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

[169] M. Henning, R. Eriksson, P. Garefelt, H. Martin, and Z. Elekes. Job relatedness, local skill coherence and economic performance: a job postings approach. *Regional Studies, Regional Science*, 12(1):95–122, 2025.

[170] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

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

[172] A. O. Hirschman. *The Strategy of Economic Development*. Yale University Press, New Haven, 1958.

[173] J. H. Holland. *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA, 2 edition, 1992.

[174] D. Huang, J. Lai, and C.-D. Wang. Ensemble clustering using factor graph. *Pattern Recognition*, pages 131–142, 2016.

[175] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2, 283–304*, 1998.

[176] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering, 9(3), 90–95*, 2007.

[177] S. Iammarino and P. McCann. *Multinationals and Economic Geography*. Edward Elgar, Cheltenham, 2013.

[178] J. Imbs and R. Wacziarg. Stages of diversification. *American Economic Review*, 93 (1):63–86, 2003.

[179] INET Oxford (Oxford Martin School). Complexity economics. https://www.inet.ox.ac.uk/research/programmes/complexity-economics, 2025. Accessed 18-12-2025.

[180] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In: International Conference on Machine Learning, PMLR, 448–456*, 2015.

[181] Italian Chambers of Commerce. Italian registry of innovative start-ups by the italian chambers of commerce, 2024. URL https://www.registroimprese.it/start-up-innovative. (accessed on 23-08-2024).

[182] Italian National Institute of Statistics (ISTAT). Statistical registry of active businesses by the italian national institute of statistics (istat), 2024. URL https://www.istat.it/scheda-qualita/registro-statistico-delle-imprese-attive-asia-imprese-2/. (accessed on 23-08-2024).

[183] Italian National Statistical Institute (ISTAT). Italian enterprises in global value chains. https://www.istat.it, 2018.

[184] Italian National Statistical Institute (ISTAT). Structural business statistics: Methods and definitions. https://www.istat.it, 2018. Documentation on Italian structural business surveys.

[185] Italian National Statistical Institute (ISTAT). Innovation in italian enterprises. https://www.istat.it, 2019. Based on the Community Innovation Survey.

[186] Italian National Statistical Institute (ISTAT). Labour force and employment statistics: Concepts and definitions. https://www.istat.it, 2019.

[187] Italian National Statistical Institute (ISTAT). Productivity measures in the italian economy. https://www.istat.it, 2019.

[188] Italian National Statistical Institute (ISTAT). Statistical registry of active businesses (asia): Methodology anv content. https://www.istat.it, 2020. Official documentation of the Italian business register.

[189] Italian National Statistical Institute (ISTAT). Business demography: Births, deaths and survival of enterprises. https://www.istat.it, 2020.

[190] Italian National Statistical Institute (ISTAT). Territorial indicators for development. https://www.istat.it, 2020.

[191] Italian National Statistical Institute (ISTAT). Structure and competitiveness of the italian enterprise system, annual statistical publication. `https://www.istat.it`, 2021.

[192] Italian National Statistical Institute (ISTAT). Extended registry of economical variables at firm level from the italian national institute of statistics istat, 2024. URL `https://www.istat.it/scheda-qualita/sistema-informativo-frame-territoriale`. (accessed on 03-09-2024).

[193] Italian Republic. Legislative decree no. 322 of 6 september 1989: National statistical system (sistan). Gazzetta Ufficiale della Repubblica Italiana, 1989.

[194] M. G. Jacobides and S. G. Winter. The co-evolution of capabilities and transaction costs: Explaining the institutional structure of production. *Strategic Management Journal*, 26(5):395–413, 2005.

[195] J. Jacobs. *The Economy of Cities*. Random House, New York, 1969.

[196] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 2010.

[197] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri. Geoai: Spatially explicit artificial intelligence techniques. *International Journal of Geographical Information Science*, 34(4):625–636, 2020.

[198] B. S. Javorcik. Does foreign direct investment increase the productivity of domestic firms? in search of spillovers through backward linkages. *American Economic Review*, 94(3):605–627, 2004.

[199] Z. Jiang. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering, 31(9), 1645–1664*, 2018.

[200] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[201] I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150, 202*, 2016.

[202] K. Jordahl, J. Van den Bossche, and F. M. others. geopandas/geopandas: v0, 2020. URL `https://zenodo.org/doi/10.5281/zenodo.2585848`.

[203] J. Kang et al. Geospatial big data and urban analytics. *Computers, Environment and Urban Systems*, 75:13–26, 2019.

[204] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[205] S. Kichko, W. J. Liang, J. F. Thisse, and P. Wang. The Rise (and Fall) of Tech Clusters. *Papers in Regional Science*, 2024.

[206] A. Kirman. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6(2):117–136, 1992.

[207] A. Kirman. Complex economics: Individual and collective rationality. *The Economic Journal*, 120(544), 2010.

[208] S. Klepper. Disagreements, spinoffs, and the evolution of detroit as the capital of the u.s. automobile industry. *Management Science*, 53(4):616–631, 2007.

[209] D. F. Kogler, D. L. Rigby, and I. Tucker. Mapping knowledge space and technological relatedness in us cities. *European Planning Studies*, 21(9):1374–1391, 2013.

[210] I. Kononenko and M. Kukar. *Constructive Induction in Machine Learning and Data Mining*. 213–226, Horwood Publishing, Chichester, UK, 2007.

[211] K. Kopczewska, P. Churski, A. Ochojski, and A. Polko. *Measuring Regional Specialisation: A New Approach*. Palgrave Macmillan, Cham, Switzerland, 2017. ISBN 978-3-319-51504-5. doi: 10.1007/978-3-319-51505-2. eBook ISBN: 978-3-319-51505-2.

[212] K. Kopczewska. *Applied Spatial Statistics and Econometrics: Data Analysis in R*. Routledge, London, 2020.

[213] K. Kopczewska. Spatial machine learning: New challenges for regional science. *Regional Science Policy & Practice*, 14(1):1–19, 2022.

[214] K. Kopczewska. Spatial machine learning: new opportunities for regional science. *The Annals of Regional Science*, 2022.

[215] K. Kopczewska. Spatial bootstrapped microeconometrics: Forecasting for out-of-sample geo-locations in big data. *Scandinavian Journal of Statistics, 50(3), 1391–1419*, 2023. URL `https://doi.org/10.1111/sjos.12636`.

[216] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[217] P. Krugman. Increasing returns and economic geography. *Journal of Political Economy*, 99(3):483–499, 1991.

[218] P. Krugman. *Development, Geography, and Economic Theory*. MIT Press, Cambridge, MA, 1995.

[219] P. R. Krugman. *Geography and Trade*. Cambridge: MIT Press, 1991.

[220] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics, 22(1), pp.79-86*, 1951.

[221] H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

[222] P. C. Kyriakidis. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289, 2004. doi: 10.1111/j.1538-4632.2004.tb01135.x.

[223] S. N. Lahiri. Theoretical comparisons of block bootstrap methods for spatial data. *The Annals of Statistics*, 27(1):386–404, 1999.

[224] S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003.

[225] R. Lall and T. Robinson. The midas touch: Accurate and scalable missing-data imputation with deep learning. *Political Analysis*, 30(2):179–196, 2022.

[226] S. Lall. Technological capabilities and industrialization. *World Development*, 20(2): 165–186, 1992.

[227] N. S.-N. Lam. Spatial interpolation methods: A review. *The American Cartographer*, 10(2):129–150, 1983. doi: 10.1559/152304083783914958.

[228] Y. A. LeCun. *Efficient BackProp*. In: Springer Berlin Heidelberg, Berlin, Heidelberg, 9–48, 2012. URL `https://doi.org/10.1007/978-3-642-35289-8_3`.

[229] C. Lee. How can we use neural network with entity embedding for product valuations? a case study for the car industry. *International Journal of Information Management Data Insights, 3(2023), 100187*, 2023.

[230] W. J. Lee and H. W. Lauw. Latent representation learning for geospatial entities. *ACM Transactions on Spatial Algorithms and Systems, 10(4), 1-31*, 2024.

[231] J. P. LeSage and R. K. Pace. *Introduction to Spatial Econometrics*. CRC Press, Boca Raton, 2009.

[232] W. A. Lewis. Economic development with unlimited supplies of labour. *The Manchester School*, 22(2):139–191, 1954.

[233] H. Li, X. Liu, T. Li, and R. Gan. A novel density-based clustering algorithm using nearest neighbor graph. *Pattern Recognition*, 2020.

[234] W. Li. Geoai: Where machine learning and big data converge in giscience. *Journal of Spatial Information Science*, pages 71–77, 2020.

[235] Y. Li, D. Wu, T. Zhao, X. Wu, and B. Li. Species distribution modeling based on xgboost: Application to prehistoric human migrations. *International Journal of Geographical Information Science*, pages 2009–2034, 2019.

[236] W. C. Lin, C. F. Tsai, and J. R. Zhong. Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowledge-Based Systems*, 239, March 2022. ISSN 0950-7051. Publisher Copyright: © 2021 Elsevier B.V.

[237] A. Lo Turco and D. Maggioni. The knowledge and skill content of production complexity. *Research Policy*, 51(8):104059, 2022. ISSN 0048-7333. doi: https://doi.org/10.1016/j.respol.2020.104059. URL `https://www.sciencedirect.com/science/article/pii/S0048733320301372`. Special Issue on Economic Complexity.

[238] A. Lösch. *Die räumliche Ordnung der Wirtschaft*. Gustav Fischer, Jena, 1940.

[239] R. E. Lucas. On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42, 1988.

[240] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

[241] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[242] E. J. Malecki. Industrial location and corporate organization in high technology industries. *Economic Geography, 61(4), 345*, 1985. URL `https://doi.org/10.2307/144054`.

[243] E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285, 1993.

[244] R. Marcinkevičs and J. E. Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint 2012.01805*, 2023.

[245] E. Marcon and F. Puech. Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, 10(5):745–762, 2010.

[246] E. Marcon and F. Puech. Mapping distributions in non-homogeneous space with distance-based methods. *Journal of Spatial Econometrics, 4-13*, 2023.

[247] C. C. Markides and P. J. Williamson. Related diversification, core competences and corporate performance. *Strategic Management Journal*, 15(S2):149–165, 1994.

[248] A. Markusen. Sticky places in slippery space: A typology of industrial districts. *Economic Geography*, 72(3):293–313, 1996.

[249] A. Marshall. *Principles of Economics*. Macmillan, London, 1890.

[250] N. Mathew and E. Pugliese. The less-than-one percent: How a few firms shape regional capabilities in india, forthcoming.

[251] D. Matricano. The effect of r&d investments, highly skilled employees, and patents on the performance of italian innovative startups. *Technology Analysis & Strategic Management*, pages 1195–1208, 2020.

[252] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo. *Missing Data: A Gentle Introduction*. Guilford Publications, 2007. ISBN 9781606238202.

[253] P. Mealy, J. D. Farmer, and A. Teytelboym. Interpreting economic complexity. *Science advances*, 5(1):eaau1705, 2019.

[254] M. Mera-Gaona, U. Neumann, R. Vargas, and D. López. Evaluating the impact of multivariate imputation by mice in feature selection. *PLOS ONE*, 16:e0254720, 07 2021.

[255] D. Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter, 3(1), 27–32*, 2001.

[256] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013.

[257] J. S. Mill. *Principles of Political Economy*. John W. Parker, London, 1848.

[258] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long. A survey of clustering with deep learning. *arXiv preprint arXiv:1801.07648*, 2018.

[259] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, Oxford, 2009.

[260] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2022. Online book, available at `https://christophm.github.io/interpretable-ml-book/`.

[261] C. A. Montgomery. Corporate diversification. *Journal of Economic Perspectives*, 8 (3):163–178, 1994.

[262] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1–2): 17–23, 1950.

[263] S. Mullainathan and J. Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[264] L. Mungo, F. Lafond, P. Astudillo-Estévez, and J. D. Farmer. Reconstructing production networks using machine learning. *Journal of Economic Dynamics and Control*, 148:104607, 2023.

[265] T. Murata. Detecting communities from bipartite networks based on bipartite modularity. *Proceedings of the International Conference on Computational Intelligence*, 2010.

[266] NACE. European classification system used to group businesses by their economic activities, 2010. URL `https://ec.europa.eu/competition/mergers/cases/index/nace_all.html`.

[267] Z. P. Neal, R. Domagalski, and B. Sagan. Analysis of spatial networks from bipartite projections using the r backbone package. *Geographical Analysis, 54(3), 623–647*, 2022.

[268] F. Neffke, M. Henning, and R. 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.

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

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

[271] F. Neffke. The skill space. CID Research Paper 99, Center for International Development, Harvard University, 2019.

[272] F. M. H. Neffke, A. Otto, and A. Weyh. Inter-industry labor flows. *Journal of Economic Behavior & Organization*, 142:275–292, 2017.

[273] R. R. Nelson and S. G. Winter. *An Evolutionary Theory of Economic Change*. The Belknap Press of Harvard University Press: Cambridge, MA, 1982.

[274] R. R. Nelson and S. G. Winter. Evolutionary theorizing in economics. *Journal of Economic Perspectives*, 16(2):23–46, 2002. doi: 10.1257/0895330027247.

[275] L. Nesta and P. P. Saviotti. Coherence of the knowledge base and the firm's innovative performance: evidence from the us pharmaceutical industry. *The Journal of Industrial Economics*, 53(1):123–142, 2005.

[276] Netron. Visualizer for neural network, deep learning and machine learning is available at (accessed 28 december), 2024. URL `https://github.com/lutzroeder/netron`.

[277] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E.*, page 026113, 2004.

[278] R. Nurkse. *Problems of Capital Formation in Underdeveloped Countries*. Oxford University Press, Oxford, 1953.

[279] A. Ochojski, A. Polko, and P. Churski. *Theoretical Foundations of Specialisation, Agglomeration and Concentration, in: Kopczewska, K. Measuring Regional Specialisation*. Palgrave Macmillan. Springer Nature, 2017.

[280] OECD. *OECD Territorial Reviews: Milan, Italy*. OECD Publishing, Paris, 2006.

[281] OECD. Attractiveness for innovation: location factors for international investment. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp*, 2011. URL `https://doi.org/10.1787/9789264104815-en`.

[282] S. Openshaw and P. J. Taylor. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, pages 127–144, 1979.

[283] S. Openshaw. *The Modifiable Areal Unit Problem*. Geo Books, Norwich, 1984.

[284] F. G. Operti, E. Pugliese, J. S. Andrade Jr, L. Pietronero, and A. Gabrielli. Dynamics in the fitness-income plane: Brazilian states vs world countries. *PloS one*, 13(6): e0197616, 2018.

[285] J. Paelinck and L. Klaassen. *Spatial Econometrics*. Saxon House, 1979.

[286] F. Pan and B. Yang. Financial development and the geographies of startup cities: evidence from china. *Small Business Economics*, 2019.

[287] D. Panzera. Areal interpolation methods: The Bayesian interpolation method. In Paolo Postiglione, Roberto Benedetti, and Federica Piersimoni, editors, *Spatial Econometric Methods in Agricultural Economics Using R*, pages 189–202. CRC Press, Boca Raton, FL, 2022.

[288] L. L. Pasinetti. *Structural Change and Economic Growth*. Cambridge University Press, Cambridge, 1981.

[289] P. Patel and K. Pavitt. The wide (and increasing) spread of technological competencies in the world's largest firms: A challange to conventional wisdom. In A. D. Chandler and O. S. Peter Hagström, editors, *The Dynamic Firm: The Role of Technology, Strategy, Organization, and Regions*. Oxford University Press, Oxford, 1998.

[290] A. Patelli, G. Caldarelli, and L. Pietronero. Complexity metrics for regional and environmental economic analysis. *Entropy*, 24(3):381, 2022.

[291] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 2011.

[292] E. Penrose. *The Theory of the Growth of the Firm*. Oxford University Press, 3rd edition, 1959.

[293] P. Pesántez-Cabrera and A. Kalyanaraman. Efficient detection of communities in biological bipartite networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 258–271, 2017.

[294] A. Pichler, C. Diem, A. Brintrup, F. Lafond, G. Magerman, G. Buiten, T. Y. Choi, V. M. Carvalho, J. D. Farmer, and S. Thurner. Building an alliance to map global supply networks. *Science*, 382(6668):270–272, 2023.

[295] L. Pietronero. Complexity ideas from statistical physics and their application to socioeconomic systems. *Physica A*, 299(1–2):1–6, 2001.

[296] L. Pietronero and A. Zaccaria. Towards a statistical physics of economic complexity. *European Physical Journal B*, 64(3–4):567–575, 2008.

[297] L. Pietronero, M. Cristelli, and A. Tacchella. Economic complexity and development: The role of productive structures. *Entropy*, 19(10):559, 2017.

[298] F. L. Pinheiro, P.-A. Balland, R. Boschma, and D. Hartmann. The dark side of the geography of innovation: relatedness, complexity and regional inequality in europe. *Regional Studies*, 59(1):2106362, 2025. doi: 10.1080/00343404.2022.2106362. URL `https://doi.org/10.1080/00343404.2022.2106362`.

[299] J. Podani, T. Kalapos, B. Barta, and D. Schmera. Principal component analysis of incomplete data – a simple solution to an old problem. *Ecological Informatics*, 61: 101235, 2021. ISSN 1574-9541.

[300] M. Polanyi. *The Tacit Dimension*. Anchor Books, New York, 1966.

[301] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, New York, 1999.

[302] M. E. Porter. *The Comparative Advantage of Nations*. Free Press, New York., 1990.

[303] C. K. Prahalad and G. Hamel. The core competence of the corporation. *Harvard Business Review*, 68(3):79–91, 1990.

[304] R. Prebisch. The economic development of latin america and its principal problems. Technical report, United Nations Economic Commission for Latin America (ECLA), 1950.

[305] E. Pugliese, L. Napolitano, M. Chinazzi, and G. Chiarotti. The emergence of innovation complexity at different geographical and technological scales, 2019.

[306] E. Pugliese and A. Tacchella. Economic complexity analytics: Country factsheets. *Publications Office of the European Union, Luxembourg. doi*, 10:368138, 2021.

[307] M. D. Raihan-Al-Masud and M. R. H. Mondal. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLOS ONE*, 15(2):1–21, 2020.

[308] J. N. K. Rao and C. F. J. Wu. Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241, 1988.

[309] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. Yu, and L. He. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems, PP*, 2024. URL `https://doi.org/10.48550/arXiv.2210.04142`.

[310] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[311] D. Ricardo. *On the Principles of Political Economy and Taxation*. John Murray, London, 1817.

[312] D. R. Roberts et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[313] P. M. Romer. Increasing returns and long-run growth. *The Journal of Political Economy, 94(5), 1002–1037*, 1986.

[314] P. M. Romer. Endogenous technological change. *Journal of Political Economy*, 98 (5):S71–S102, 1990.

[315] P. N. Rosenstein-Rodan. Problems of industrialisation of eastern and south-eastern europe. *The Economic Journal*, 53(210/211):202–211, 1943.

[316] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[317] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 1987. Reprinted/republished edition: 2004.

[318] D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[319] R. P. Rumelt. *Strategy, Structure, and Economic Performance*. Division of Research, Harvard Business School, Boston, MA, 1974.

[320] M. Ryan. *Deep learning with structured data*. Manning Publications, 2020.

[321] D. Ryu, J. Hong, and H. Jo. Capturing locational effects: application of the K-means clustering algorithm. *The Annals of Regional Science, 73(1), 265-289.*, 2024.

[322] Santa Fe Institute. About: Overview. `https://www.santafe.edu/about/overview`, 2025. Accessed 18-12-2025.

[323] P. P. Saviotti and K. Frenken. Export variety and the economic performance of countries. *Journal of Evolutionary Economics*, 18(2):201–218, 2008.

[324] A. Saxenian. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA, 1994.

[325] A. Sbardella, E. Pugliese, and L. Pietronero. Economic development and wage inequality: A complex system analysis. *PloS one*, 12(9):e0182774, 2017.

[326] A. Sbardella, A. Zaccaria, L. Pietronero, and P. Scaramozzino. Behind the italian regional divide: An economic fitness and complexity perspective. LEM Working Paper 2021/30, Scuola Superiore Sant'Anna, LEM, 2021. URL `https://hdl.handle.net/10419/247299`.

[327] B. Schölkopf. Kernel principal component analysis. *In: International Conference on Artificial Neural Networks, Springer, pp. 583–588*, 1997.

[328] J. Schumpeter and U. Backhaus. *The Theory of Economic Development*, chapter 2-4, pages 61–116. Springer US, Boston, MA, 2003. ISBN 978-0-306-48082-9.

[329] J. A. Schumpeter. *The Theory of Economic Development*. Harvard University Press, Cambridge, MA, 1934.

[330] scikit-learn. Official scikit-learn documentation on one-hot encoding of categorical variables, 2024. URL `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html`. (accessed on 04-09-2024).

[331] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[332] V. D. P. Servedio, A. Bellina, E. Calò, and G. De Marzo. Economic complexity in mono-partite networks. *arXiv preprint arXiv:2405.04158*, 2024.

[333] K. Seu, M.-S. Kang, and H. Lee. An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization*, 6:278, 05 2022.

[334] K. K. Sharma and A. Seal. Clustering analysis using an adaptive fused distance. *Engineering Applications of Artificial Intelligence*, 2020.

[335] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[336] K. P. Sinaga and M.-S. Yang. Unsupervised k-means clustering algorithm. *IEEE access*, 2020.

[337] A. Singleton and D. Arribas-Bel. Geographic data science. *Geographical Analysis*, 53(1):61–75, 2021. doi: 10.1111/gean.12194.

[338] R. R. Sitter. A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419):755–765, 1992.

[339] A. Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

[340] R. M. Solow. A contribution to the theory of economic growth. *Quarterly Journal of Economics*, 70(1):65–94, 1956.

[341] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse, and C. Biernacki. Model-based clustering with missing not at random data. *arXiv preprint 2112.10425*, 2023.

[342] N. Srivastava. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15(1), 1929–1958*, 2014.

[343] Statuto dei Lavoratori. Norms on the protection of workers' freedom and dignity, trade union freedom and activity in the workplace., 1970. URL `https://www.gazzettaufficiale.it/eli/id/1970/05/27/070U0300/sg`.

[344] H. J. Steenhuis and E. J. de Bruijn. High technology revisited: definition and position. *In: ICMIT 2006 Proceedings, IEEE International Conference on Management of Innovation and Technology, 21-23 June, Singapore*, 2006.

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

[346] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[347] J. Svirsky and O. Lindenbaum. Interpretable deep clustering. *arXiv:2306.04785*, 2023.

[348] A. Świdurska. Efficiency of the development of high-tech industry in poland, 2009. URL `https://prace-kgp.uken.krakow.pl/article/view/480`.

[349] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific Reports*, 2:723, 2012.

[350] A. Tacchella, G. Caldarelli, A. Gabrielli, and L. Pietronero. Economic complexity: conceptual grounding of a new metrics for global competitiveness. *European Physical Journal Special Topics*, 223(10):1893–1913, 2013.

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

[352] L. Tesfatsion and K. L. Judd, editors. *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*. Elsevier, Amsterdam, 2006.

[353] W. R. Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979. doi: 10.1080/01621459.1979.10481647.

[354] A. Topchy, A. K. Jain, and W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

[355] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

[356] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 306–307*, 1979.

[357] Uber Technologies, Inc. H3: A hexagonal hierarchical spatial index. `https://h3geo.org`, 2018.

[358] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

[359] A. van Dam, A. Gomez-Lievano, F. Neffke, and K. Frenken. An information-theoretic approach to the analysis of location and co-location patterns. *Journal of Regional Science*, 63(1):173–213, 2023.

[360] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research, 9(11)*, 2008.

[361] H. R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

[362] A. J. Venables. Equilibrium locations of vertically linked industries. *International Economic Review*, 37(2):341–359, 1996.

[363] J. H. von Thünen. *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Perthes, Hamburg, 1826.

[364] J. F. Wang, B. B. Gao, and A. Stein. The spatial statistic trinity: A generic framework for spatial sampling and inference, 2020. URL `https://doi.org/10.1016/j.envsoft.2020.104835`. Environmental Modelling & Software, 134.

[365] A. Weber. *Über den Standort der Industrien*. Mohr, Tübingen, 1909.

[366] C. K. Wikle and L. M. Berliner. Combining information across spatial scales. *Technometrics*, 47(1):80–91, 2005. doi: 10.1198/004017004000000572.

[367] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

[368] C. Xiao, S. Hong, and W. Huang. Optimizing graph layout by t-sne perplexity estimation. *International Journal of Data Science and Analytics*, 15(2):159–171, 2023.

[369] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. *In: International Conference on Machine Learning, PMLR, 478–487*, 2016.

[370] W. Yanwen. Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation, 121*, 2023. URL `https://doi.org/10.1016/j.jag.2023.103364`.

[371] A. Zaccaria, M. Cristelli, A. Tacchella, and L. Pietronero. How the taxonomy of products drives the economic development of countries. *PLOS ONE*, 9(12):e113770, 2014.

[372] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *Computer Vision ECCV 2014*, 2014.

[373] L. Zhang, H. Pan, Q. Fan, C. A. I. Ying, and Y. Jing. 1gbdt, lr & deep learning for turn-based strategy game ai. *IEEE Conference on Games (CoG), 1-8, London.*, 2019. URL https://doi.org/10.1109/CIG.2019.8848103.

[374] Z.-H. Zhou. *Machine learning*. Springer nature, 2021.

[375] X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

A

Monopartite graph

In order to verify the robustness of the biLouvain clusters and the number of clusters produced, another approach has been tried.
Starting from the adjacency matrix of the bipartite graph $A_{i,j}$, where the i index runs over the startups of the dataset and the j index runs over the various clusters of the clustering algorithm in the *clusters ensemble*, it is possible to create a monopartite graph by combining the startups and the clusters and interpret them just as nodes of a monopartite graph.
This procedure requires the imposition of the constraint that two nodes corresponding to two startups cannot be connected by an edge and, analogously, that two nodes describing two clusters cannot be connected either.
In practice, this is possible by defining a square adjacency matrix as in Equation A.1.

$$M = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{A.1}$$

where A is the adjacency matrix of the bipartite graph, A^T is its transpose.
The size of the M matrix is $K_{tot} \cdot |X| \times K_{tot} \cdot |X|$, where K_{tot} is the total number of clusters in the clusters ensemble and $|X|$ is the number of data points in the dataset.
This graph is presented in Figure A.1 and shows that even with this other matrix and using other community detection algorithms, the number of clusters produced is still 5.

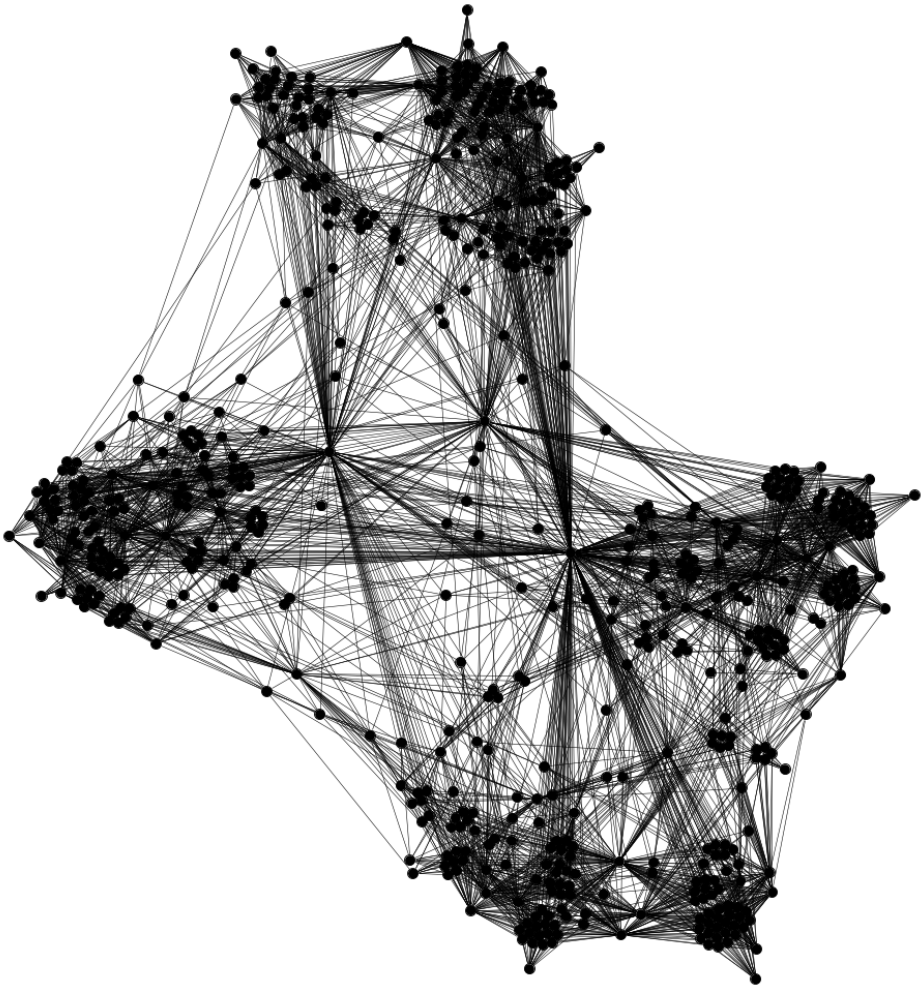


Figure A.1: Graph associated to the M adjacency matrix, implementing the strategy presented in[47], it is possible to see that the Modularity of this graph is 0.601 and that there are 5 communities which are composed of both data points and clusters. This allows to see that the 5 rows communities and the 5 column communities identified by the biLouvain algorithm have a 1 to 1 correspondence because in the monopartite graph the pairs of corresponding communities fuse together into just 5 hybrid communities.

B

List of features used

In this appendix, the list of all the features that have been used to run the algorithms will be presented.

Note that several other variables were available but they have been removed because they were highly correlated ($|corr|) > 0.8$) to the other ones and therefore would give redundant information. In particular when the categorical variables, like the province or the ATECO codes (which is the Italian localized version of the European NACE codes), have been encoded using dummy variables, it is common practice[330] to remove one of the dummy variables because it would just be a linear combination of all the other variables introducing multicollinearity.

The variables that have a number in parenthesis in their names have been obtained by the official registry of innovative startups and are some of the requisites that a business should have to be considered an innovative startup. Most of them have been kept, but some of them, the ones referring to the activity sector of the startup, have been removed because they were highly correlated with the ATECO columns that convey the same kind of information. The three features that have a (6) in the name are important because each business must have at least one of them to be considered an innovative startup.

The variables that start with the acronym *SBS* are obtained from the *Frame SBS*[192] which is the extended registry of economic variables at firm level from ISTAT within the ASIA database.

The variables with *ul* in the name have been obtained by summing over the corresponding value for each one of the local units, for instance the overall number of employees of a firm is the sum of the number of employees of each one of its local units.
The extended list of the variables is:

- **start_date**: Start Date of Activity (format = YYYYMMDD) (from this variable the day and the month have been discarded and only the year has been kept as a numerical variable);
- **independent_workers**: Number of independent workers who carries out their activity without formal subordination constraints and whose remuneration has the nature of mixed income;
- **artisan_flag**: Artisan Flag which tells if the firm respects all the requisites to be considered an artisan enterprise;
- **revenue_class**: There are four revenue class brackets: less than 2 million Euros, between 2 million and 10 million, between 10 million and 50 million and more than 50 million, that gives an approximate idea of the overall economical size of the firm;
- **business_age_class**: Business Age Class which is 1 if the business has 0-2 years, 2 if the business has 3-5 years, 3 if the business has 6-10 years, 4 if the business has 11-15 years and 5 if the business has 16 or more years;
- **sbs_n_countries_imp**: Number of countries from which imports are made by the firm;
- **sbs_n_countries_exp**: Number of countries to which exports are made by the firm;
- **sbs_valimp**: Total imports (in Euros);
- **sbs_valesp**: Total exports (in Euros);
- **sbs_imp_migs_20**: Intermediate goods imports (in Euros);
- **sbs_exp_migs_20**: Intermediate goods exports (in Euros);
- **sbs_exp_migs_41**: Durable consumer goods exports (in Euros);
- **sbs_exp_migs_42**: Non-durable consumer goods exports (in Euros);

- **sbs_exp_germany**: Exports to Germany (in Euros);
- **sbs_exp_russia**: Exports to Russia (in Euros);
- **sbs_imp_usa**: Imports from the USA (in Euros);
- **sbs_exp_usa**: Exports to the USA (in Euros);
- **sbs_exp_china**: Exports to China (in Euros);
- **sbs_exp_japan**: Exports to Japan (in Euros);
- **sbs_exp_brasil**: Exports to Brazil (in Euros);
- **sbs_imp_area_15**: Imports from non-EU European countries (in Euros);
- **sbs_exp_area_15**: Exports to non-EU European countries (in Euros);
- **sbs_exp_area_17**: Exports to North Africa (in Euros);
- **sbs_exp_area_18**: Exports to other African countries (in Euros);
- **sbs_imp_area_23**: Imports from the Middle East (in Euros);
- **sbs_exp_area_23**: Exports to the Middle East (in Euros);
- **sbs_exp_area_24**: Exports to Central Asia (in Euros);
- **sbs_exp_area_25**: Exports to East Asia (in Euros);
- **sbs_exp_area_99**: Exports to Oceania, other territories, and other destinations (in Euros);
- **sbs_revenue**: Total revenue from the sale of goods and services (in Euros);
- **sbs_labour_costs**: Costs that a firm has to pay (in Euros) associated with labour like salaries, etc.;
- **sbs_value_added**: Value added (in Euros) computed as the total sources of revenue of the firm minus the sum of all the costs;
- **n_ul** number of local units that compose the firm;
- **emp_ul_qual_worker** number of employees with worker qualification;
- **emp_ul_qual_executive** number of employees with executive qualification;

- **emp_ul_qual_apprentice** number of employees with apprentice qualification;
- **emp_ul_qual_manager** number of employees with managerial qualification;
- **emp_ul_qual_other_kind** number of employees with other qualification;
- **emp_ul_age_50_over** number of employees in the age group 50 or over;
- **emp_ul_gender_m** number of employees with male gender;
- **emp_ul_naz_eu** number of employees born within the EU;
- **emp_ul_naz_extraeu** number of employees born outside the EU;
- **flag_g**: Flag that tells if the business belongs to business groups;
- **flag_exp**: Flag that tells if the business exports;
- **revenue_sales_perform_ul**: Current revenues from sales (in Euros) excluding VAT, gross of indirect taxes;
- **other_revenues_ul**: Other revenues (in Euros) and income of the firm;
- **services_ul**: Costs (in Euros) for services of the firm;
- **enjoy_asset_ul**: Costs (in Euros) due to enjoyment of third-party assets (like rent, etc.) of the firm;
- **sal_ul**: Total expenses (in Euros) due to salaries of the firm;
- **other_costs_ul**: Miscellaneous management costs (in Euros) of the firm;
- **Date of registration in the startup section** Year of inscription in the innovative startups registry;
- **Date of registration in the Company Register** Year of inscription in the national registry of firms;
- **Final year production class (1)** Production class of the innovative startup as defined in the innovative startups registry and conveying the approximate information of the total revenues of the firm for the last year;
- **class of employees last year (2)** Employees class of the innovative startpup as defined in the innovative startups registry and conveying the approximate information of the total number of employees;

- **Capital class (3)** Capital class of the innovative startp-up as defined in the innovative startups registry and conveying the approximate information of the share capital of the firm;
- **1° req. (6)** Flag that tells if the firm has R&D expenditures equal to at least 15% of the higher of cost and total production value;
- **2° req. (6)** Flag that tells if the firm employs highly qualified personnel (at least 1/3 PhDs, PhD students or researchers, or at least 2/3 with a master's degree);
- **3° req. (6)** Flag that tells if the firm is the owner, depositary or licensee of at least one patent or holder of registered software;
- **female prevalence (8)** The prevalence of females in the firm: 0-> no prevalence (under 50% of the average between the percentage of the share capital of the firm and the percentage of administrative positions is held by women), 1->majoritarian prevalence (over 50%), 2->strong prevalence (over 66%), 3->exclusive prevalence (100%);
- **youth prevalence (8)** The prevalence of young employees in the firm: 0-> no prevalence (under 50% of the average between the percentage of the share capital of the firm and the percentage of administrative positions is held by people under 35 years of age), 1->majoritarian prevalence (over 50%), 2->strong prevalence (over 66%), 3->exclusive prevalence (100%);
- **foreign prevalence (8)** The prevalence of people non born in Italy in the firm: 0-> no prevalence (under 50% of the average between the percentage of the share capital of the firm and the percentage of administrative positions is held by people not born in Italy), 1->majoritarian prevalence (over 50%), 2->strong prevalence (over 66%), 3->exclusive prevalence (100%);
- **long** Longitude of the main registered office of the firm;
- **lat** Latitude of the main registered office of the firm;
- **sector_SERVICES** Flag that tells if the firm operates in the services activity sector;
- **legal_form_1320** Flag that tells if the firm is a Ltd (limited liability company);
- **legal_form_1330** Flag that tells if the firm is a Ltd with a single shareholder;

- **legal_form_1520** Flag that tells if the firm is a consortium company;
- **cpro_13** The firm is in the Como province;
- **cpro_14** The firm is in the Sondrio province;
- **cpro_15** The firm is in the Milano province;
- **cpro_16** The firm is in the Bergamo province;
- **cpro_17** The firm is in the Brescia province;
- **cpro_18** The firm is in the Pavia province;
- **cpro_19** The firm is in the Cremona province;
- **cpro_20** The firm is in the Mantova province;
- **cpro_97** The firm is in the Lecco province;
- **cpro_98** The firm is in the Lodi province;
- **cpro_108** The firm is in the Monza-Brianza province;
- **ateco2_13** The firms corresponding European NACE code is Manufacture of textiles;
- **ateco2_14** The firms corresponding European NACE code is Manufacture of wearing apparel;
- **ateco2_15** The firms corresponding European NACE code is Manufacture of leather and related products;
- **ateco2_18** The firms corresponding European NACE code is Printing and reproduction of recorded media;
- **ateco2_20** The firms corresponding European NACE code is Manufacture of chemicals and chemical products;
- **ateco2_21** The firms corresponding European NACE code is Manufacture of basic pharmaceutical products and pharmaceutical preparations;
- **ateco2_22** The firms corresponding European NACE code is Manufacture of rubber and plastic products;
- **ateco2_23** The firms corresponding European NACE code is Manufacture of other non-metallic mineral products;

- **ateco2_24** The firms corresponding European NACE code is Manufacture of basic metals;
- **ateco2_25** The firms corresponding European NACE code is Manufacture of fabricated metal products, except machinery and equipment;
- **ateco2_26** The firms corresponding European NACE code is Manufacture of computer, electronic and optical products;
- **ateco2_27** The firms corresponding European NACE code is Manufacture of electrical equipment;
- **ateco2_28** The firms corresponding European NACE code is Manufacture of machinery and equipment n.e.c.;
- **ateco2_30** The firms corresponding European NACE code is Manufacture of other transport equipment;
- **ateco2_31** The firms corresponding European NACE code is Manufacture of furniture;
- **ateco2_32** The firms corresponding European NACE code is Other manufacturing;
- **ateco2_33** The firms corresponding European NACE code is Repair and installation of machinery and equipment;
- **ateco2_35** The firms corresponding European NACE code is Electricity, gas, steam and air conditioning supply;
- **ateco2_43** The firms corresponding European NACE code is Specialised construction activities;
- **ateco2_45** The firms corresponding European NACE code is Wholesale and retail trade and repair of motor vehicles and motorcycles;
- **ateco2_46** The firms corresponding European NACE code is Wholesale trade, except of motor vehicles and motorcycles;
- **ateco2_47** The firms corresponding European NACE code is Retail trade, except of motor vehicles and motorcycles;
- **ateco2_56** The firms corresponding European NACE code is Food and beverage service activities;

- **ateco2_58** The firms corresponding European NACE code is Publishing activities;
- **ateco2_59** The firms corresponding European NACE code is Motion picture, video and television programme production, sound recording and music publishing activities;
- **ateco2_61** The firms corresponding European NACE code is Telecommunications;
- **ateco2_62** The firms corresponding European NACE code is Computer programming, consultancy and related activities;
- **ateco2_63** The firms corresponding European NACE code is Information service activities;
- **ateco2_64** The firms corresponding European NACE code is Financial service activities, except insurance and pension funding;
- **ateco2_66** The firms corresponding European NACE code is Activities auxiliary to financial services and insurance activities;
- **ateco2_68** The firms corresponding European NACE code is Real estate activities;
- **ateco2_69** The firms corresponding European NACE code is Legal and accounting activities;
- **ateco2_70** The firms corresponding European NACE code is Activities of head offices; management consultancy activities;
- **ateco2_71** The firms corresponding European NACE code is Architectural and engineering activities; technical testing and analysis;
- **ateco2_72** The firms corresponding European NACE code is Scientific research and development;
- **ateco2_73** The firms corresponding European NACE code is Advertising and market research;
- **ateco2_74** The firms corresponding European NACE code is Other professional, scientific and technical activities;
- **ateco2_75** The firms corresponding European NACE code is Veterinary activities;

- **ateco2_77** The firms corresponding European NACE code is Rental and leasing activities;
- **ateco2_78** The firms corresponding European NACE code is Employment activities;
- **ateco2_79** The firms corresponding European NACE code is Travel agency, tour operator and other reservation service and related activities;
- **ateco2_82** The firms corresponding European NACE code is Office administrative, office support and other business support activities;
- **ateco2_85** The firms corresponding European NACE code is Education;
- **ateco2_86** The firms corresponding European NACE code is Human health activities;
- **ateco2_90** The firms corresponding European NACE code is Creative, arts and entertainment activities;
- **ateco2_93** The firms corresponding European NACE code is Sports activities and amusement and recreation activities;
- **ateco2_96** The firms corresponding European NACE code is Other personal service activities.

C

Database structure of ASIA

In this appendix the list of all the variables contained in the ASIA databases is shown. There are duplicate variables between the different databases that have been removed and all the data have been carefully examinated and preprocessed before using them to train the neural networks.

Table A.1: Variables in ASIA Businesses

Variable	**Description**	**Variable**	**Description**
business code	Unique identifier for each business	tax code	Business's tax identification code
business name	Name of the business	address	Physical location (later discarded)
postal code	Postal code (redundant)	region code	Region identifier
province code	Province identifier	municipality code	Municipality identifier
start date of activity	Year business began (day/month discarded)	legal form	Legal classification

Continued on next page

Table A.1 – continued from previous page

Variable	Description	Variable	Description
average number of independent workers	Avg. independent workers	average number of employees	Avg. employees
sum of average independent and dependent workers	Combined average workers	flag for address completeness	Address completeness (discarded)
artisan flag	Artisan business indicator	revenue class	Revenue class
economic activity classification	Ateco2007 activity code	business age class	Business age class

Table A.2: Variables in ASIA Local Units

Variable	Description	Variable	Description
business code	Unique identifier for each business	local unit code	Identifier for the local unit
tax code	Tax identification code	business name	Name of the business for the unit
local unit region code	Region code of the local unit	province code	Province code
municipality code	Municipality code	address	Physical address of the local unit
postal code of local unit	Postal code	legal form	Legal classification of the business
local unit economic activity classification	Ateco2007 activity code	headquarters flag	Is it the headquarters?
number of employees in the local unit	Number of employees	number of workers	Total number of workers

Continued on next page

Table A.2 – continued from previous page

Variable	Description	Variable	Description
number of employees	Total employees	number of apprentices	Apprentices
number of middle managers	Middle managers	number of executives	Executives
number of employees in other categories	Other employee categories	number of employees aged 15 29	Employees aged 15–29
number of employees aged 30 49	Employees aged 30–49	number of employees aged 50 and above	Employees aged 50+
number of employees with unavailable age	Unknown age	number of female employees	Female employees
number of male employees	Male employees	number of employees with missing gender information	Unknown gender
number of employees born in italy	Born in Italy	number of employees born in eu countries	Born in EU countries
number of employees born in non eu countries	Born outside EU	number of employees with missing nationality information	Unknown birthplace
active for six months flag	Active at least 6 months		

Table A.3: Variables in ASIA tecframe-sbs

Variable	**Description**	**Variable**	**Description**
business code	Business Code	total employees	Total Employees
economic activity	Ateco2007 activity (5-digit)	province abbreviation	Province abbreviation
number of countries from which imports are made	Countries of origin for imports	number of countries to which exports are made	Countries of destination for exports
number of geographic areas from which imports are made	Import geographic areas	number of geographic areas to which exports are made	Export geographic areas
number of imported products	Imported products (8-digit)	number of exported products	Exported products (8-digit)
number of imported product groups	Product groups (CPA) imported	number of exported product groups	Product groups (CPA) exported
total imports	Total imports (€)	total exports	Total exports (€)
energy imports	Energy imports (€)	energy exports	Energy exports (€)
intermediate goods imports	Intermediate goods imports (€)	intermediate goods exports	Intermediate goods exports (€)
capital goods imports	Capital goods imports (€)	capital goods exports	Capital goods exports (€)
durable consumer goods imports	Durable consumer goods imports (€)	durable consumer goods exports	Durable consumer goods exports (€)
non durable consumer goods imports	Non-durable goods imports (€)	non durable consumer goods exports	Non-durable goods exports (€)

Continued on next page

Table A.3 – continued from previous page

Variable	**Description**	**Variable**	**Description**
non classifiable imports	Non-classifiable imports (€)	non classifiable exports	Non-classifiable exports (€)
imports from germany	From Germany (€)	exports to germany	To Germany (€)
imports from china	From China (€)	exports to china	To China (€)
imports from usa	From USA (€)	exports to usa	To USA (€)
imports from india	From India (€)	exports to india	To India (€)
imports from japan	From Japan (€)	exports to japan	To Japan (€)
imports from brazil	From Brazil (€)	exports to brazil	To Brazil (€)
imports from north africa	From North Africa (€)	exports to north africa	To North Africa (€)
imports from east asia	From East Asia (€)	exports to east asia	To East Asia (€)
revenue from the sale of goods and services	Revenue (€)	labor costs	Labor costs (€)
value added	Value added (€)	costs for raw materials supplies consumption services and third party goods	Input, materials, and third-party costs (€)

Table A.4: Variables in ASIA Economic Results

Variable	Description	Variable	Description
business code	ASIA business code	belonging to business groups	Belongs to a business group
exporting business	0/1 binary variable if the firm is exporting or not	foreign controlled business resident in italy	Foreign-controlled firm in Italy
italian controlled business resident abroad	Italian-controlled firm abroad	current revenues excluding vat gross of indirect taxes of the local unit	Revenues excl. VAT (gross of taxes)
other revenues and income of the local unit	Other income of local unit	costs of goods and services of the local unit	Costs of goods/services
costs for services of the local unit	Service costs	enjoyment of third party assets of the local unit	Use of third-party assets
salaries	Salaries of local unit	miscellaneous management costs	Miscellaneous management costs
value added	Value added by the business		

C.1 Embeddings and network architectures

In this appendix the detailed technical description of how the embeddings have been created and how they have been used to compute the clusters of firms will be shown. The procedure employed is inspired by Entity embedding[157]. The idea is that an ad-hoc regression problem is created, and then a deep neural network is trained to solve it. By solving it, the neural network learns, in the process, a meaningful representation of the data that can later be used in the clustering process.
In our work, the numerical representation in the embedding space has 128 dimen-

sions. This number has been chosen because it is common practice to select powers of two for the size of hidden layers, and this is the closest one to the original size of the dataset. Therefore, this limits possible information loss associated with choosing too small or too close to the output layer.
To create a regression task that the model must solve, one of the variables is chosen as the target, and the other ones are used to try to predict the target. The choice of the target variable is essential as it can affect the performance of the regressor and, therefore, the overall quality of the data embedding. The dataset, in this case, has a high degree of multicollinearity. Thus, most of the choices created a model with high performances on test sets, as the regression tasks themselves were instead "easy".
Two approaches for the choice of the target variable have been tried:

1. Given that the embedding subsequently will be used to create the strata for the stratified Bootstrap for the computation of the correlation coefficient of the productivity for the years 2017, 2018, and 2019, the feature that, on average, has the highest correlation with the productivity across the three years. This choice has been subsequently discarded because, in this way, the embedding used to compute the clusters had some information about the productivity. Still, the clusters should be completely independent of the productivity. Another approach has been adopted to avoid this subtle form of data leakage.

2. The variable used as the target was the overall number of employees because it was the most stable across the years and the most informative for assessing a company's actual health state. The a priori choice avoided any theoretical data leakage. Still, in practice, it did not affect the model's performance as the high degree of multicollinearity made the results robust under this choice.

Once the target variable has been chosen, the data are split into a train set, a test set. A neural network is trained on the train set (containing both numerical and categorical variables). The model's performances are then tested on the test set; if the performances are reasonable, the model has effectively learned a meaningful representation of the input data and has used it to compute the target variable. After the split, the numerical data were scaled in the range [0, 1] for the train, and then the same exact transformation (to avoid any kind of subtle data leakage) was applied to the test set. This scaling has been performed as it improves convergence speed and the stability of gradient descent during the training of neural networks [151,228].
A deep neural network has been created using Tensorflow [1], which can take numerical values and categorical variables as input and is shown in figure B.2. The output

of the dense layer with 128 neurons has been chosen as the embedding of the data and the meaningful representation of the initial data. The reason for this choice is that by choosing as an output layer for the embedding a layer not too close to the input layers, the model has enough layers to properly abstract from the inputs and create a meaningful representation of the data. The same is true for the output layer; by choosing a smaller layer too close to the output, the model risks being too influenced by the output of the regression, and it could incur information loss.

After fine-tuning, the best hyperparameters are:

1. Adam optimizer with a learning rate of 0.001;
2. Relu activation function for the dense layers;
3. Dropout rate of 0.4;
4. Batch normalization after each one of the dense layers;
5. A sigmoid activation function for the output layer, the output has been scaled between 0 and 1 and this choice turned out to give better performance than linear activation functions or other activation functions for the output.

With this configuration the model had a 0.000574 train error, a 0.000536 test error with an R2 = 0.943599. These are deemed good test performances indicating that the model has been able to learn and generalize on unseen data.
The embedded dataset is presented in Figure B.1 by implementing the t-SNE algorithm[360]; this algorithm creates a probability distribution in the original space by assigning a higher probability to a similar pair of data points and a lower probability to dissimilar data points, and another distribution in the two-dimensional space. Then, the KL divergence between the two distributions is minimized. This algorithm is widely used[19,128] and has become one of the best options for visualizing high-dimensional data. The correct use of this algorithm involves tuning some hyperparameters like Perplexity; research with some best practices exists in literature[368], and in this work, the Perplexity has been tuned until a stable result has been obtained.
From Figure B.1, IT services businesses are the more numerous and sparser category. Some ICT Trade businesses are concentrated but very close to IT services businesses. The largest companies are close to each other and confined in a particular region on the right; this means that when the size increases, the companies tend to have more similar characteristics, irrespective of the activity sector.
The data generated has 128 dimensions, which are too many for computationally intensive clustering algorithms and can be affected by the curse of dimensionality,

therefore we applied a DEC algorithm.
First, an autoencoder neural network is trained as part of the model. The autoencoder comprises an encoder comprising four layers, each containing 100, 100, 300, and 10 neurons, followed by a decoder with 300, 100, 100, and 128 neurons. Using a mean squared error loss during training, the model efficiently learns a 10-dimensional representation of the data in the last layer of the encoder. This representation must retain all the information in the 128-dimensional data because the decoder needs it to recover the original 128-dimensional data.
To train this architecture, data was divided into train and test sets. The model was trained on the train set, and its efficacy in reducing and recovering data dimensions was evaluated on the test set using the Mean Squared error between the original and reconstructed data.

The model's hyperparameters have been finetuned and are

1. Adam optimizer with a learning rate of 0.01.
2. Exponential Linear Unit (ELU) activation function for the dense layers.
3. Dropout rate of 0.5.
4. Batch normalization after each one of the dense layers.
5. A linear activation function for the output layer.

The fine-tuning process included trying the Stochastic Gradient descent optimizer; the activation function ReLU, ELU, Tanh; dropout rates between 0 and 0.5; Linear activation function for the output lair and all possible combinations in a Grid-search style of approach. The aforementioned hyperparameters combination turned out to be the best ones. With these hyperparameters the model has on the train set a Mean Squared error of 0.005612 an R2 of 0.992000 and a Mean squared error of 0.005416 and an R2 of 0.933725 on the test set. Once the autoencoder has been trained, the decoder is removed and substituted with a clustering layer. The figure of the neural network is presented in Figure B.3. The model works analogously as the t-SNE algorithm: it computes a Student's t-distribution of the compressed 10-dimensional data. Then,[369] explained that a target distribution from the current cluster assignments is created.
This encoder+clustering layer model is then trained to minimize the KL divergence between the two distributions. Autoencoders are particularly useful for handling spatial heterogeneity and learn latent features from spatial data, identifying underlying patterns (e.g., regional economic structures) that traditional methods may miss.

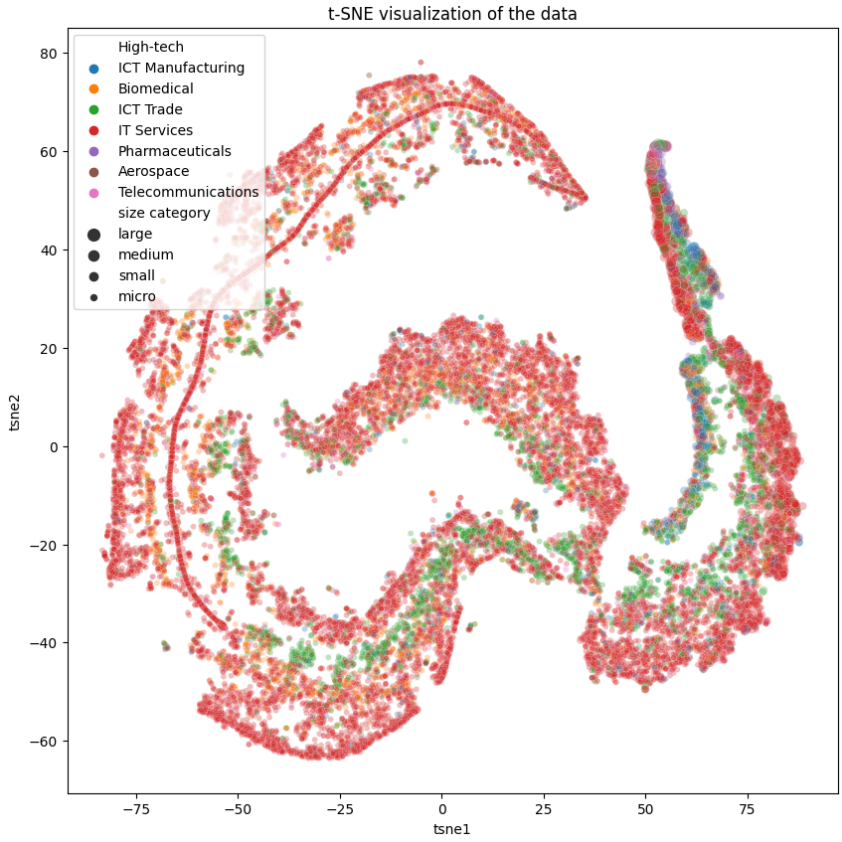


Figure B.1: The embedded data are presented and projected in two dimensions using the t-SNE algorithm. The above figure has been obtained by running the algorithm for 1000 iterations, setting the perplexity parameter to 100. Other parameter configurations yielded similar results. Each dot represents a business and has been colored according to its activity sector. Close dots mean that the represented businesses have similar characteristics. The size of the dots gives the average number of employees of the business in 2019 year. Micro businesses have up to 10 employees, small businesses up to 50, medium businesses up to 250, and large businesses have over 250 employees.

Spatial data often contains noise (e.g., measurement errors), and the autoencoders can filter out irrelevant variations, isolating the core characteristics of each region. The latent space generated by an autoencoder often reveals clusters of similar regions.

The previously described procedure changes the data's cluster assignments and low-dimensional representation because it changes the encoder's weights and, thus, how the 10-dimensional representation is shaped. It is an essential feature because it creates clustering-friendly data representation without the risk of information loss. Once the algorithm converges, the clusters can be used as strata for the Stratified bootstrap algorithm[212]

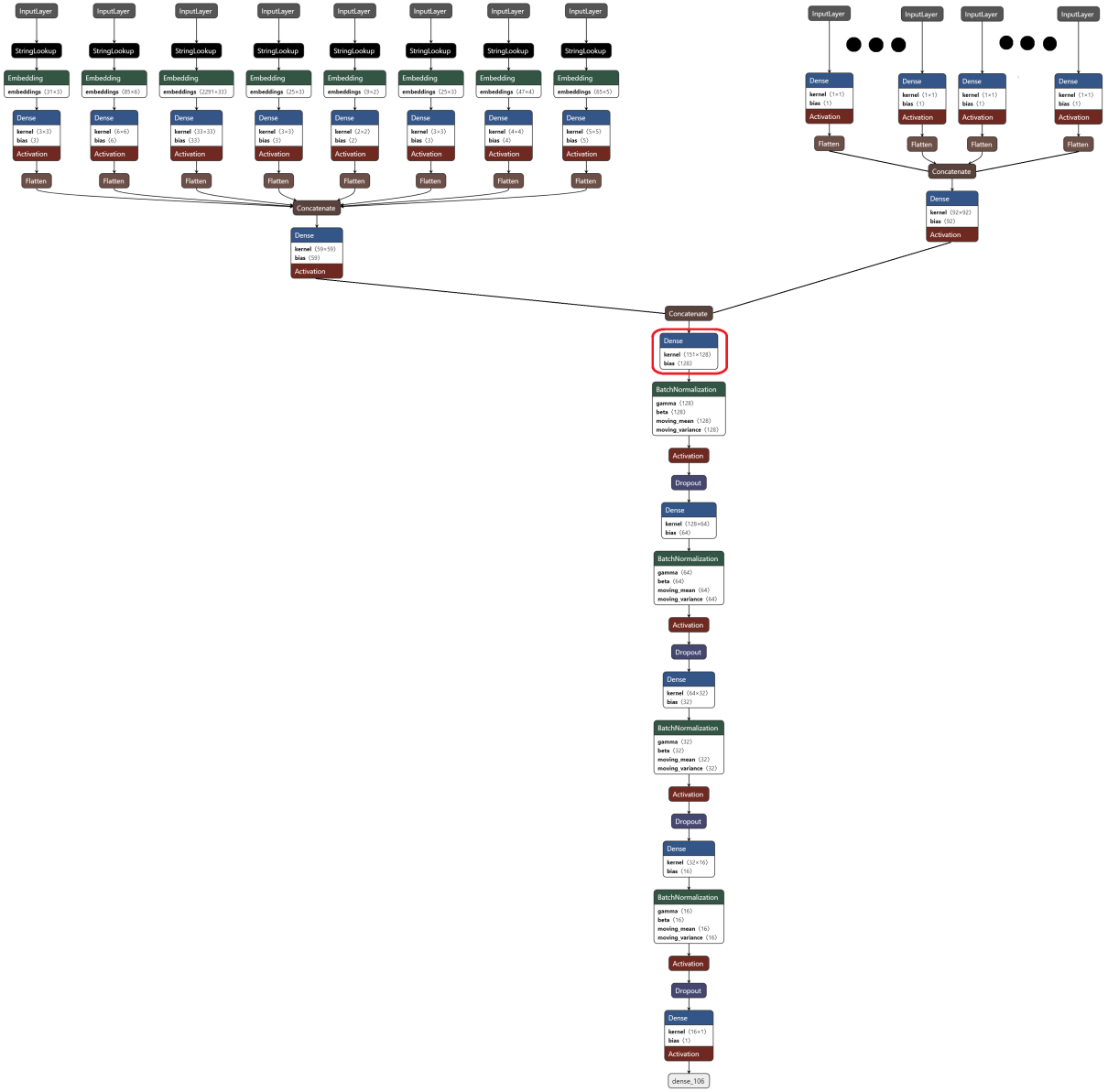


Figure B.2: The architecture of the regressor neural network. On the top left are the categorical variables that pass through a Stringlookup layer, an Embedding layer, and a dense layer before being concatenated. The numerical variables are passed in dense layers on the top right before concatenation. The numerical and categorical variables are then passed through dense layers that do not alter the number of inputs before being concatenated. After the concatenation, the data are passed through some last layers with 128 neurons, 64 neurons, 32 neurons, 16 neurons, and finally, the output layer. This graphical representation has been created using Netron [276].

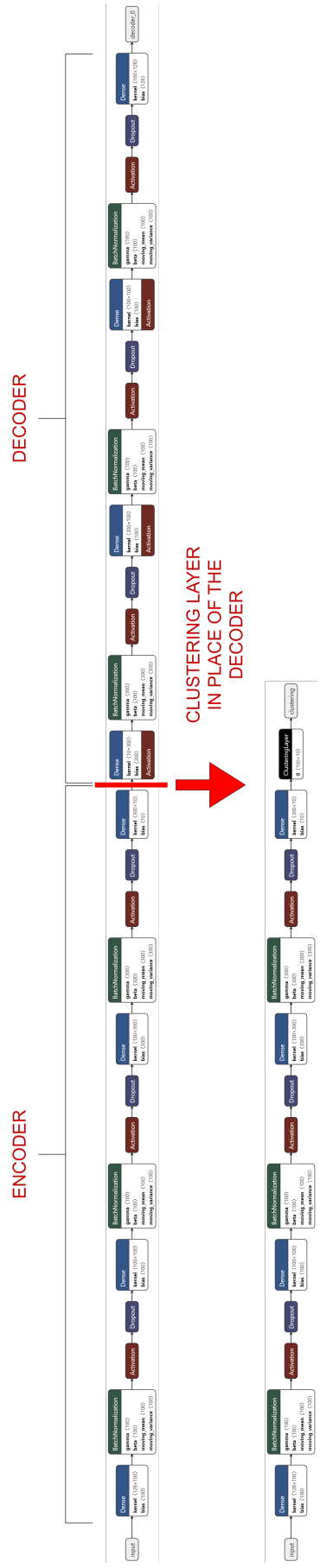


Figure B.3: The architecture of the DEC. First, an autoencoder is trained on the train set and evaluated on the test set. Then, the decoder is substituted with a clustering layer, and the encoder plus clustering layer is trained on the entire dataset by minimizing the KL divergence.

C.2 Comparison with other approach

In the appendix the comparison of the DEC-based methodology with another similar methodology aimed at clustering mixed categorical and numerical features will be presented.
In particular, we applied the k-prototype algorithm to the same data that we used for the DEC algorithm and we used the output clusters for the same stratified bootstrap strategy presented in Subsection3.5.1 and Subsection 3.5.2.
We used the implementation of k-prototype in the Python package kmodes[93] leaving all settings to default values and just changing the number of clusters to 11 to obtain comparable results with the ones produced by the DEC algorithm.
The two algorithms give radically different solutions to the same clustering problem, leading to low values for comparison metrics, as shown in Table C.1. To see if the

Metric	Score
Adjusted Rand Index (ARI)	0.0822
Normalized Mutual Information (NMI)	0.1396
Homogeneity	0.1273
Completeness	0.1546
V-Measure	0.1396

Table C.1: Clustering comparison metrics. All these metrics are scaled to have a value of 1 for perfect agreement. All the metrics agree that the clustering solutions proposed by k-prototype and DEC are different.

differences between the two clustering solutions have an impact on the bootstrap procedure, we used the k-prototype clustering solution for the bootstrap analysis of correlation coefficients, creating the exact analogue of Figure 3.6 but with this other clustering algorithm. This new bootstrap solution is presented in Figure C.1. A correlation of 0.31 between productivity of two consecutive years is unrealistically low and too far away from the value calculated on the whole sample. This underscores and justifies the importance of the use of new methodologies when clustering is used for stratified bootstrap and highlights the fact that the outcomes and results of all methodologies should be evaluated carefully.

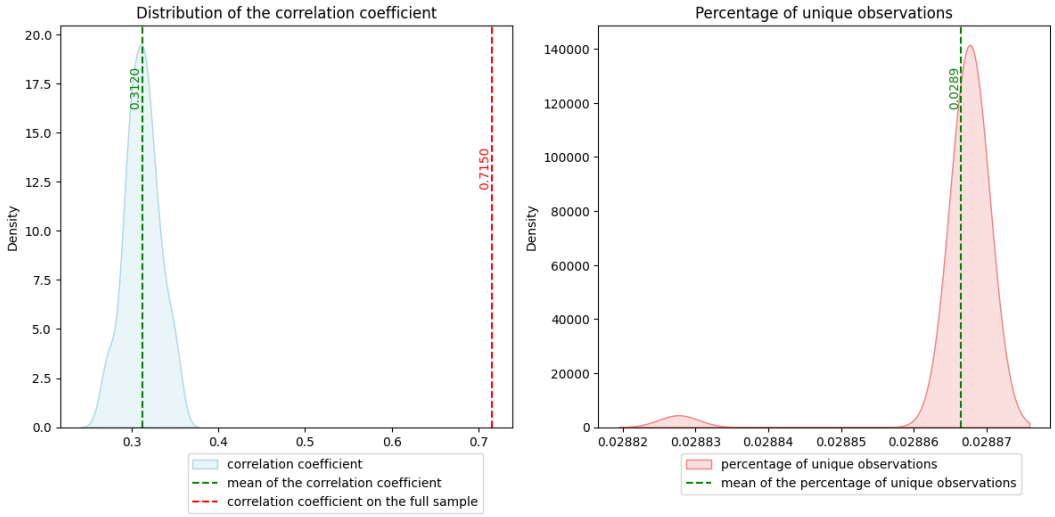


Figure C.1: The correlation between the productivity between the years 2018 and 2019 computed using the k-prototype algorithm.

Acknowledgements

This dissertation was carried out within the Philosophiae Doctor (Doctor of Philosophy) (PhD) Programme in *Big Data and Artificial Intelligence* at Universitas Mercatorum.

The author, Alessio Bumbea, gratefully acknowledges support from a three–year PhD fellowship awarded under **D.M. n. 352/2022** and funded within Italy's Piano Nazionale di Ripresa e Resilienza (PNRR), Mission 4 "*Education and Research*", Component 2 "*From Research to Enterprise*", Investment 3.3 "*Introduction of innovative PhD programmes that meet the needs of enterprises and promote the recruitment of researchers by enterprises*". The fellowship was co–funded by the Italian Chambers of Commerce Study Centre "Guglielmo Tagliacarne".

The author wishes to thank **Prof. Andrea Mazzitelli**,**Dr. Alessandro Rinaldi**,**Dr. Emanuele Pugliese** for their supervision and guidance throughout the doctoral studies. Sincere thanks are also due to **Italian Chambers of Commerce Study Centre "Guglielmo Tagliacarne"** for the collaboration, hosting, and material and emotional support in this journey. The author is grateful to the **United Nations University MERIT** for hosting the research stay abroad and for the scientific discussions that contributed to the development of the thesis.

The author wants to express sincere thanks to the members of his family for their unwavering support and encouragement.

The author also wants to express deep gratitude to his friends from Terracina, Rome and Maastricht, for their companionship, laughter, and being there throughout this journey, making him feel always welcome with them.

Unless otherwise stated, all opinions and any errors remain the author's responsibility.

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of PhD in *Big Data and Artificial Intelligence* is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

I hereby declare no financial conflicts of interest related to the concepts, technologies, or entities discussed. Institutional disclosures were made, and confidentiality was ensured by anonymizing datasets, and proprietary tools in compliance with GDPR.

During the preparation of this work, Chat-GPT 5.2, Grammarly and QuillBot were used to perform translation, grammar and spelling check. After using these tools and services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications in the Appendix "*Scientific outputs arising from the PhD*" at the end of the thesis). All figures and tables are original or used with proper permission.

Rome, April 24, 2026

_____

Alessio Bumbea

Scientific outputs arising from the PhD

Peer-reviewed papers

1. **Bumbea, A.**, Mazzitelli, A., Giuffrida, A., and Espa, G. *Spatial bootstrapping using deep clustering methods: spatial machine learning applied to Lombardy high-tech businesses*. *Regional Science Policy & Practice*, 17(12), December 2025, 100242. DOI: 10.1016/j.rspp.2025.100242

2. **Bumbea, A.**, Espa, G., Gentile, M., Giuffrida, A., Mazzitelli, A., and Pini, M. *Economic complexity as tool to assess the territorial development: a novel empirical approach inspired by network theory applied to patent data*. *Quality & Quantity* (2025). DOI: 10.1007/s11135-025-02141-7

3. **Bumbea, A.**, Mazzitelli, A., Espa, G., and Rinaldi, A. *Bipartite graph partitioning and spatial bootstrapping methods: a case study of innovative startups*. *Big Data Research* (2025), 100533. DOI: 10.1016/j.bdr.2025.100533

Conference proceedings

1. **Bumbea, A.**, Espa, G., Mazzitelli, A., and Pugliese, E. *From micro to macro: Building up regional capabilities from individual firms and employees*. In *Book of Short Papers – 3rd Italian Conference on Economic Statistics (ICES 2025) "Sustainability, Innovation and Digitalization: Statistical Measurement for Economic Analysis"*, Napoli: Enzo Albano, 2025, pp. 61–64. ISBN: 979-12-80655-52-3

2. Giuffrida, A., **Bumbea, A.**, Espa, G., and Mazzitelli, A. *Economic complexity, technological innovation, and co-location of activities*. In *Book of Short Papers – 3rd Italian Conference on Economic Statistics (ICES 2025) "Sustainability, Innovation and Digitalization: Statistical Measurement for Economic Analysis"*, Napoli: Enzo Albano, 2025. ISBN: 979-12-80655-52-3

3. **Bumbea, A.**, Straccamore, M., Bellina, A., Espa, G., Mazzitelli, A., and Tacchella, A. *Diversification in Non-Profit Services: a Statistical Analysis through Economic Fitness and Complexity approaches*. In *Methodological and Applied Statistics and Demography II: Short Papers, Solicited Sessions*. Springer Nature, 2024. ISBN: 9783031643491. DOI: 10.1007/978-3-031-64350-7_74

4. Bellina, A., **Bumbea, A.**, Espa, G., Mazzitelli, A., Straccamore, M., and Tacchella, A. *Economic Fitness of municipalities*. In *Statistical analysis of complex economic data: recent developments and applications*. 2024. ISBN: 978-88-476-2950-9

5. **Bumbea, A.**, Espa, G., Mazzitelli, A., and Rinaldi, A. *Bootstrapping methods on Innovative start-ups*. In *Statistical analysis of complex economic data: recent developments and applications*. 2024. ISBN: 978-88-476-2950-9

6. Vurro, A. E., **Bumbea, A.**, Giuffrida, A., Mazzitelli, A., and Espa, G. *Geography of business alliances and spatial network complexity: the backbone of formal collaboration agreements*. Conference Proceedings (in press), 2024.

7. Giuffrida, A., **Bumbea, A.**, Mazzitelli, A., Sbardella, A., and Zaccaria, A. *Firms' Capabilities and Economic Complexity: insights from Italian survey data*. In *Book of Abstracts – Data Science & Social Research*, 2025. ISBN: 9781326620653

Other

1. Gentile, M., **Bumbea, A.**, Giannini, D., Giuffrida, A., Macigno, L., Mariz, D., Mazzitelli, A., Pini, M., Righi, P., Rinaldi, A., and Salate Santone, F. *Looking at EU strategic technologies through the lens of patents: measuring, impact on productivity, and technological interdependencies*. Working paper, Luiss Research Center for European Analysis and Policy, 2025.

2. Vurro, A. E., **Bumbea, A.**, Giuffrida, A., Mazzitelli, A., and Espa, G. *Business Alliances and Spatial Network Backbones: A National-Scale Analysis of Formal Collaborations*. Paper under review.

3. Straccamore, M., Bellina, A., **Bumbea, A.**, Mazzitelli, A., and Tacchella, A. *Emergent description of Italian economic structure through Economic Complexity lens*. Paper under review.

4. **Bumbea A.**, Espa G., Giuffrida A. and Mazzitelli A. *Technological Relatedness and Innovation Geography: A Patent- Based Analysis Beyond Industrial Districts*. Paper under review

The PhD scholarship co-funded with resources from the European Union – NextGeneration EU National Recovery and Resilience Plan (PNRR), Mission 4, Component 2 *"From Research to Business"* – Investment 3.3 *"Introduction of innovative PhD programmes that meet the innovation needs of enterprises and promote the recruitment of researchers by companies"* – CUP D83C22001880003