# THESIS

BENNOUR MOHAMED HSIN

2025-12-03

# Table of contents

# List of Figures

# List of Tables

# Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

# 1 Introduction

## 1.1 General background

The Covid 19 pandemic showed structural challenges of national economies all over the world, specifically the fragility of neoliberal policies in times of crisis and the lack of industrial and economic resilience. Six years after the fact, our societies are now confronted with inevitable novel challenges and looming shocks. We are already witnessing the consequences of AI development as it paves the way for a new technological revolution that would render most local economies obsolete and cause massive unemployment in white collar sectors. Sadly, that doesn't seem to be the end of the shocks the world has seen recently as the war in Ukraine, Gaza, Iran, and most importantly the current, chaotic trade wars, seem to foster ever increasing uncertainties. Facing all this, policy makers are confronted with a simple choice; strategise and plan for more resilient local economies. In different streams of literature, resilience is directly related to diversification/variety, whether in portfolio management in finance, trade partnerships/linkages, industrial activities, or in terms of knowledge as well. Thus we can equivalently say that resilience is the capacity

to resist and/or adapt to external shocks by relying on exiting internal capabilities that evolve in the face of such shocks. This means that for an economy to survive uncertainties, it needs to evolve, change, and innovate its way to the other end of the bleak challenges it's confronted with. However, to evolve, change and innovate, a baseline of knowledge should be leveraged since the consensus is that variety is a buffer against external shocks and shields uncertainties. This chain of thoughts, takes us back to the conceptual basics of sustainable economic development and growth; the knowledge fabric is what facilitate any long term strategy, and has been shown in the literature as clear catalyst for societal prosperity and economic resilience. For this reason the study of knowledge is detrimental for policy making, and understanding how to increase its diversity is more relevant than ever before.

Knowledge is a set of information that covers one or many topics, and its characteristics are contingent on the different forms it can take or how it was created, generally speaking, academia and businesses are the main knowledge creators in any economy through research and patents. Essentially, knowledge can be codified (accessible by anyone through any medium), or tacit (personal information based on social connections, intuition, experience, etc, that's hard to share with others). The consensus in the literature is that the main driver of competitive advantage for firms is the tacit form of knowledge, which is also widely acknowledged that it's space dependent. However, the knowledge produced by firms can be reliably seen in patents, although they capture codified information, they also reveal tacit knowledge and its geographic footprint in space. This means that its detrimental to assess the knowledge in patenting activities (we refer to this knowledge as

technologies), and more so to focus on the local aspect of these activities i.e: sub-national regions.

This framing, however, is not at all new or novel. In fact, this is the entire aim of the literature of the geography of innovation; to study how innovation is created and diffused to different actors in different geographical contexts. Specifically, relatedness and economic complexity (REC) is one of the main streams of literature that focus on the relationships between activities and geographies. The conceptual and methodological framework that REC provides is widely used and adopted in academia and among policy practitioners and was one of the main contributors to the smart specialisation policy literature. The ideas embedded in this framework, to put it simply, rely on the premise of spatial dependence of tacit knowledge in local/regional economies/geographies and focus on simplifying these relationships using network science to model the relationships between knowledge and regions. Albeit these simplifications provide valuable scope for analysis and interpretation, the cost from the loss of granular information implies that there's much more conceptual, methodological, and empirical work needed. The reason for this is because the loss of information bias the empirical interpretation in the sense that we end up with a homogeneous implication with weak regards to the regional and national contexts as well as the technological characteristics. This work is motivated by this gap and aims to simply contextualise the study of knowledge diversification using the same granular information publicly available and commonly used in the REC literature. The idea is simple, account for endogenous and exogenous contexts using granular data and understand the contexts and contingencies that drives regional diversification.

## 1.2 Problem statement

The main problem that this works aims to solve is directly embedded in the methodological and empirical framework of REC. REC models diversification through network aggregation based on co-location and co-occurrence patterns. Using these patterns, different aggregations are used to quantify relatedness (the frequency of observing a pair of activities in the same region), and relatedness density (how much of the activities frequently observed together a region has). However, these measures are often interpreted not as aggregations of frequent observations but rather as relationship models. Empirically, diversification is studied using these constructs as predictors of entry, that is a region's entry to a new specialisation in a new technology, often considered as a binary outcome. In here we briefly outline the big picture in the REC methodological and empirical framework, its conceptual issues, its empirical consequences, and highlight the research gap.

### 1.2.1 Relatedness, relatedness density, and diversification

Relatedness and relatedness density are essentially measures of proximity. In a sense they describe how close two technologies are close to each other, or how close a given technology is to a region given its portfolio of technologies. To further decompose the problem here, we will first establish the methodological constructs for proximity measures. For relatedness that's co-occurrence, and for relatedness density that's the linear aggregation of relatedness.

First, co-occurrence is essentially the frequency of observing two activities together. In

the REC literature, this frequency describes the strength of the relationship. Activities frequently observed together are more related than the pairs rarely observed together.

Second, linear aggregation of relatedness essentially measures the percentage of co-located technologies in a region that are related to a reference technology. Thus, we can think of relatedness density as the link between related technologies and co-located technologies. The idea in the REC literature states that relative to a given technology, the more related technologies a region has, the more likely that it can develop that technology.

These two constructs are used together to predict the probability that a region will enter a new technology. The REC literature shows that relatedness density is consistently associated with higher probabilities in almost all studies. These results among others were one of the major latent contributors to smart specialisation strategy (S3) policy. Thus the consensus in the literature was clear: In order to diversify into new technologies with the highest likelihood of success, regions must prioritise investment in related technologies.

### 1.2.2 Empirical consequences

The idea of resilience is not a main focus for the REC literature nor it is ours. However, falling back to this concept allows us to further assess the empirical consequences of the mainstream interpretation of relatedness and relatedness density. The idea is that in order to be resilient to external shocks and subsequent uncertainties diversification is key. But what kind of diversification is required and feasible and how to achieve it is the focus here. The REC literature tells us that the most likely successful diversification strategy

is the one that targets related capacity in the regions, often referred to as related variety. However, generalising this recommendation is not that straight forward. Aggregate regional capacities and their national and broader geographic contexts differ significantly. The initial landscape of the regional technological portfolio is detrimental here because this strategy could favour regions with already diverse portfolio but it's questionable that regions with limited portfolios would equally benefit. This is aligned with the concept of path dependency, related variety without context enforces that dependency and locks regions within their limited capacities. This brings us back to the core focus of in this work; context is key. However, context on its own here might not be enough since path dependency of related variety is a direct result of how relatedness and relatedness density is calculated and interpreted. The constructs that enable these measures (co-occurrence and linear aggregation) are the core problem that we're highlighting here. The reason behind this specific focus relied on the implicit assumptions embedded in these methodological constructs.

Co-occurrence assumes that technologies frequently observed together are likely related. Although this is within the boundaries of common sense, it's highly unlikely that's actually the case. A frequency measure is only informative when we have more observations than items—that is, more geographies than technologies. Almost always, we will have more technologies than geographies, this means that relatedness is at best noisy. Additionally, relatedness as interpreted in the literature quantifies the relationship between pairs of technologies. However, as it stands, there's no differentiation in the direction of that relationship thus, assuming that the relationship between two technologies is sym-

metrical. Albeit this assumption in itself is not problematic, it exacerbates the linearity issue when we measure relatedness density as we loose information in co-occurrence, symmetry, and linear aggregations. This takes us to the final issue we would like to highlight; relatedness density. Simply put, relatedness density measures the sum of technologies related to a reference technology present in a given region. The implicit assumption in here is that technologies are linked through linear combinations, and those combinations predict the likelihood of successful diversification. However, relatedness density is often interpreted as a value that quantifies the existing requirements a region has relative to a technology, whereas the sum of existing related technologies do not inform us on the actual requirements.

In summary, relatedness and relatedness density measures suffer from diverse methodological issues embedded in the implicit assumptions in their core constructs. Co-occurrence and linear aggregation of observed frequency are misinterpreted, accrues information loss, and poorly handles the granular data often used. This means that the empirical and methodological work ahead must account for these issues to further contextualise the study of diversification strategies.

### 1.2.3 Research problem

In the light of all the mentioned in this section, we fall back again on the core idea that we started this text with; How can we contextualise diversification strategies? The answer to this question is multi-layered and complex. In this section we started by outlining the importance of diversification strategies for regions into new technologies via patenting

activities. We explained that the REC literature provides interesting methodological and conceptual framework of analysis and showed that despite their usefulness they suffer from structural issues that limit the advantages of the used granular data, thus limit the incorporation of a broader context empirically. Essentially, the research problem we focus on here, is both methodological and empirical in nature. We highlighted the structural methodological issues as the research gap which will be the focus of our methodological and empirical contribution.

To address the methodological gaps in REC, we propose a multilayered contextual framework that distinguishes between different contextual factors. Relatedness measures alone conflate these dimensions. By decomposing context into endogenous technology-region fit and exogenous environmental conditions, we can identify which combinations enable successful diversification across different regional contexts, following this design:

Endogenous factors (within-region characteristics):

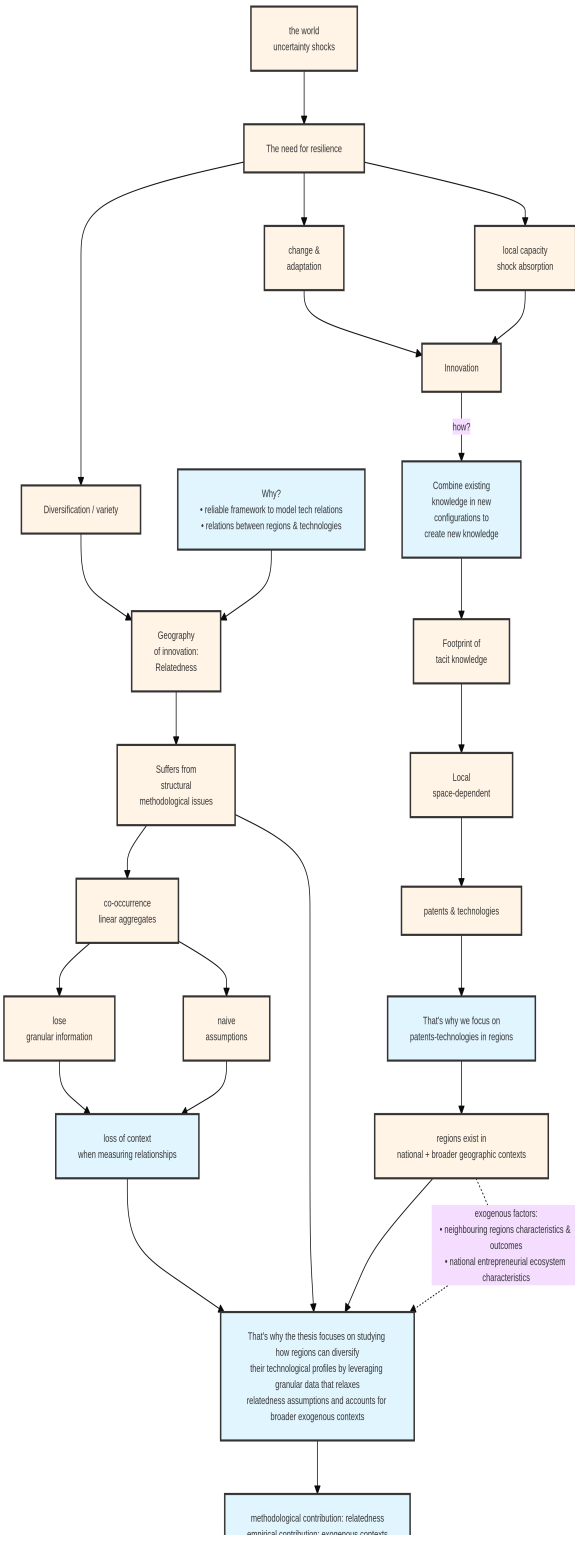
- Technology-specific attributes (network position, centrality, embeddedness)
- Regional knowledge infrastructure (coherence between technology relationships and regional specialization patterns)
- Regional technological volume (number of patents in a specific technology)

Exogenous factors (beyond-region influences):

- National entrepreneurial ecosystem characteristics (institutions, financing, R&D culture)
- Spatial spillovers from neighbouring regions (outcome diffusion, knowledge flows)

- Geographic proximity to coherent knowledge clusters

The interaction between these layers is critical: endogenous capacity determines what a region can develop, while exogenous context shapes realisation possibilities. For instance, a technology may align with regional knowledge infrastructure (endogenous coherence), but without supportive national institutions or proximity to innovative neighbours (exogenous support), entry probability remains low.


the world
uncertainty shocks
The need for resilience
change &
adaptation
local capacity
shock absorption
Innovation
how?
Diversification / variety
Why?
• reliable framework to model tech relations
• relations between regions & technologies
Combine existing
knowledge in new
configurations to
create new knowledge
Geography
of innovation:
Relatedness
Footprint of
tacit knowledge
Suffers from
structural
methodological issues
Local
space-dependent
co-occurrence
linear aggregates
patents & technologies
lose
granular information
naive
assumptions
That's why we focus on
patents-technologies in regions
loss of context
when measuring relationships
regions exist in
national + broader geographic contexts
exogenous factors:
• neighbouring regions characteristics & outcomes
• national entrepreneurial ecosystem characteristics
That's why the thesis focuses on studying
how regions can diversify
their technological profiles by leveraging
granular data that relaxes
relatedness assumptions and accounts for
broader exogenous contexts
methodological contribution: relatedness
empirical contribution: exogenous contexts

## 1.3 Objective

The objective from this work is to extend the methodological framework of relatedness starting from the structural issues it suffers from. In order to show how our methodological contribution benefit the study of regional technological diversification we rely on an empirical study that shows how broader context can be included and how such context inform granular policy insights. The broader context we aim to include empirically relied on relaxing the implicit relatedness assumptions, explicitly include the endogenous characteristics of the technologies and regions contingent on the regional knowledge infrastructure while simultaneously account for the characteristics of the national ecosystems. With such an approach we end up accounting for more contextual layers than the mainstream approach.

## 1.4 Research questions and hypothesis

**RQ1: How to further contextualise diversification strategy based on the relatedness framework?**

- H1a: Asymmetric relatedness measures better predict technology entry than symmetric measures.

**RQ2: Is successful diversification contingent on regional knowledge infrastructure?**

- H2a: Technology-specific characteristics impact on diversification is contingent on the regional knowledge coherence.
- H2b: Technology-specific characteristics impact on diversification is contingent on the regional knowledge stock.

**RQ3: Does the national ecosystem influence diversification?**

- H3a: National Entrepreneurial Ecosystem characteristics positively affects regional entry into new technologies.

**RQ4: What role does space have? Do neighbouring regions influence diversification?**

- H4a: Spatial spillovers of outcome from neighbouring regions positively affect technology entry.
- H4b: Neighbouring regions' technological characteristics influence focal region diversification outcomes.
- H4c: Geographic proximity to regions with coherent knowledge infrastructure increases entry probability.

## 1.5 Structure

We structure this work around our methodological and empirical contributions. Given the complex nature of the problems we aim to solve, we will first start with more literature context that underlines geography of innovation, and more importantly relatedness

and economic complexity literatures. We then outline our approach in the methodology chapter where we discuss in more details measures of relatedness and proximity and provide alternative conceptualisation to modeling the relationships between technologies and regions. In the same chapter we also detail other elements that will be relevant to the empirical part specifically knowledge coherence. In the empirical chapter we outline the research design and the modeling procedures that implements our methodological contribution. The results chapter outline the results of our work and its empirical consequences and interpretation, which we discuss in more detail in the discussion and conclusion of this work in which we also outline different further extensions and future research directions.

## 1.6 Relatedness(?)

# 2 Literature review

# 3 The interconnectedness between regions and technologies

## 3.1 Methodological Motivation

Traditional relatedness frameworks model diversification through symmetric co-occurrence matrices and linear aggregation of relatedness density (Hausmann and Hidalgo 2011). However, as established in our problem statement, these constructs suffer from three structural limitations: (1) symmetry assumptions that obscure directional dependencies between technologies, (2) linear aggregation that loses granular information about technology-specific requirements, and (3) noise when technologies outnumber regions.

We address these limitations by replacing traditional measures with a machine learning training strategy based on the Random Forest algorithm (RF). Specifically, we use RF to generate: (1) **Feature Importance Technology Space (FITS)** following Fessina et al. (2024), analogous to the mainstream Technology Space. FITS captures directional, hierarchical technology relationships, replacing symmetric relatedness measures; and (2)

**predicted probabilities (technological potential)** following Albora et al. (2023), analogous to relatedness density. The technological potential estimates region-specific feasibility of technology adoption given the region's own technological portfolio, replacing linear relatedness density. This approach enables the contextualization of diversification strategies by accounting simultaneously for technology-specific characteristics, regional knowledge infrastructure, and the broader non-linear interconnectedness between regions and technologies that traditional measures cannot adequately capture.

Before detailing our methodology and how we leverage the contributions in Fessina et al. (2024) and Albora et al. (2023), we will briefly situate this approach within the REC literature. Essentially, REC formalizes the empirical observation that shared input requirements (knowledge, resources, capabilities) determine diversification feasibility (Hidalgo et al. 2018; Hidalgo 2021). While this framework have proven valuable for policy (Zaccaria et al. 2018; E and A 2021), deriving granular, context-specific implications remains challenging (Hidalgo 2023; Li and Neffke 2024). Recent work on unrelated diversification (Flávio L. Pinheiro et al. 2022; Boschma et al. 2023), geographic inequalities (Flavio L. Pinheiro et al. 2025; Hartmann et al. 2017), and emerging technologies (Lee et al. 2018; Fessina et al. 2024) highlights the need for methodologies that capture contextual nuances beyond path dependency identification which we will elaborate on in subsequent sections.

Our approach is a response to the nuances expressed in the literature. But in order to highlight adequately our work we must first elaborate on the mainstream approach conceptually, methodologically and its empirical implications which we briefly outlined in the introduction. From there we expand on the mainstream approach and construct ours

conceptually, mathematically and showcase our training strategy. To this end, we also provide a brief example on how the mainstream approach and the machine learning approach to modeling the technology space compare. For the sake of clarity and consistency we will establish our notation system and the main definition in this paragraph and use them in the subsequent sections and chapter in this dissertation unless we specify otherwise.

We consider the following sets:

- $\mathcal{T} = \{t_i\}_{i=1}^{N_T}$: set of technologies
- $\mathcal{R} = \{r_i\}_{i=1}^{N_R}$: set of regions

- $\mathcal{Y} = \{y_i\}_{i=1}^{N_Y}$: set of years
- $\mathcal{C} = \{c_i\}_{i=1}^{N_C}$: set of technology categories
- $\mathcal{K} = \{\kappa_i\}_{i=1}^{N_K}$: set of countries

These sets are mirrored by the European Patent Office data (1978-2021) classified at the 4-digit IPC level, yielding 641 distinct technologies across 345 NUTS2 regions in 34 European countries. IPC classifications provide hierarchical structure: section (letter, we also refer to this as categories), class (two digits), subclass (letter). For example, F16H encompasses Section F (Mechanical engineering), Class 16 (engineering elements for mechanical power transmission), and Subclass H (gearing systems). We supplement patent data with Eurostat regional socio-economic indicators detailed in subsequent sections.

We quantify regional specialization using the Revealed Comparative Advantage (Balassa 1965). Although originally designed for trade data, this metric has been widely adopted

in innovation geography literature. Following the nomenclature in P.-A. Balland and Rigby (2017), we refer to it as Revealed Technological Advantage (RTA) in our patent data context. The RTA measures relative specialization, enabling simultaneous capture of expertise depth and portfolio diversity. Despite critiques regarding patent-based applications (P. Balland and Boschma 2019; Diodato et al. 2023), RTA aligns with our objective of capturing meaningful technology relationships through machine learning rather than raw co-occurrence. The RTA for region $r$ in technology $t$ during year $y$ is defined as:

$$\text{RTA}_{r,t,y} = \frac{\frac{X_{r,t,y}}{\sum_{t'} X_{r,t',y}}}{\frac{\sum_{r'} X_{r',t,y}}{\sum_{r',t'} X_{r',t',y}}} = \frac{X_{r,t,y} \sum_{r',t'} X_{r',t',y}}{\left(\sum_{t'} X_{r,t',y}\right)\left(\sum_{r'} X_{r',t,y}\right)}$$

Where, $X_{r,t,y}$: patent count for region $r$ in technology $t$ during year $y$

For each year $y$, we construct the RTA matrix $\mathbf{R}^{(y)}$ with entries $\text{RTA}_{r,t,y}$:

$$\mathbf{R}^{(y)} = \left[\text{RTA}_{r,t,y}\right]_{r=1,\ldots,N_R}^{t=1,\ldots,N_T}$$

These yearly matrices form the foundation for all subsequent modeling. We then binarize specialization for classification tasks:

$$z_{r,t,y} = \begin{cases} 1 & \text{if } \text{RTA}_{r,t,y} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $z_{r,t,y} = 1$ indicates region $r$ has comparative advantage (specialization) in technol-

ogy $t$ at year $y$.

To further elaborate, let's consider $c_{t,t'}$, the count of patents containing both technologies $t$ and $t'$ ($t \neq t'$) at a given time period, such that $t, t' \in \mathcal{T}$, and $c_t$, $c_{t'}$ denote the total patent counts for each class individually.

In the following we will begin by outlining how the REC literature approach relatedness, how it's calculated and discuss in detail different technical contributions and their limitations. We will then present our approach and detail our methodological and conceptual contribution in response to these limitations.

## 3.2 The mainstream approach

### 3.2.1 Relatedness Through Co-occurrence

The REC literature operationalises relatedness through the observed frequency of co-occurrences. The intuition is appealingly simple: technologies that frequently appear together—whether in the same patent, the same firm, or the same region—likely share underlying knowledge requirements. If mastering one technology facilitates developing another, we should observe them co-located more often than chance would predict. This reasoning has proven influential, providing the empirical foundation for technology space visualisations and diversification recommendations across academic and policy settings.

Yet the inference from co-occurrence to capability-based relatedness rests on assumptions that warrant scrutiny. Co-occurrence is a frequency measure; it tells us how often two

technologies appear together, not why they do so or whether that co-presence reflects shared inputs, complementary applications, or statistical artefact. The distinction matters: a measure intended to guide capability-building strategies must capture genuine knowledge relationships rather than patterns that emerge from data structure alone.

Let $c_{t,t'}$ denote the count of patents containing both technologies $t$ and $t'$ (where $t \neq t'$) within a given time period, and let $c_t$ and $c_{t'}$ denote the total patent counts for each technology individually. The simplest relatedness measure would be $c_{t,t'}$ itself, but raw co-occurrence counts are problematic: frequently patented technologies will naturally co-occur more often simply due to their prevalence. Various normalisation schemes address this issue.

**Association Strength** compares observed co-occurrence to expected co-occurrence under independence:

$$\phi_{t,t'}^{\text{assoc}} = \frac{c_{t,t'} \cdot T}{c_t \cdot c_{t'}}$$

where $T = \sum_{t \in \mathcal{T}} c_t$ is the total patent count. Values exceeding one indicate that the technologies appear together more frequently than would be expected if their occurrences were independent—they are positively associated. This measure properly corrects for size effects: a rare technology and a common technology can still exhibit strong association if they co-occur more than their marginal frequencies would predict.

**The Probability Index** refines this probabilistic approach by accounting for the fact that a patent cannot co-occur with itself:

$$\phi_{t,t'}^{\text{prob}} = \frac{c_{t,t'}}{\frac{1}{2}\left[\frac{c_t}{T} \cdot \frac{c_{t'}}{T-c_t} + \frac{c_{t'}}{T} \cdot \frac{c_t}{T-c_{t'}}\right] \cdot T}$$

This correction, while subtle, provides more accurate estimates for patent co-occurrence analysis and is implemented in standard software packages used in the field.

**Cosine Similarity** offers a set-theoretic alternative:

$$\phi_{t,t'}^{\text{cosine}} = \frac{c_{t,t'}}{\sqrt{c_t \cdot c_{t'}}}$$

This measures relative overlap between technology profiles. While intuitive, it systematically favours frequent technologies over rare ones and does not properly correct for size effects, making comparisons across technologies with different prevalences problematic.

**Minimum Conditional Probability** takes a different approach by examining regional co-location rather than patent-level co-occurrence:

$$\phi_{t,t'}^{\text{min}} = \min\left\{P(\text{RTA}_{r,t}|\text{RTA}_{r,t'}), P(\text{RTA}_{r,t'}|\text{RTA}_{r,t})\right\}$$

This computes the minimum of the conditional probabilities: given that a region specialises in one technology, how likely is it to specialise in the other? The minimum operator ensures symmetry while avoiding inflation from imbalanced conditional relationships.

A clarification is warranted here regarding two distinct but often conflated concepts. Co-occurrence typically refers to technologies appearing together within the same patent or

firm, while co-location refers to technologies appearing together within the same region. Both capture related but different phenomena: co-occurrence within patents reveals direct technical complementarities, while co-location within regions reveals shared capability requirements that may operate at greater conceptual distance. The literature sometimes uses these terms interchangeably, but the distinction matters for interpreting what relatedness measures actually capture.

### 3.2.2 Structural Limitations of Co-occurrence-Based Relatedness

Despite the variety of normalisation approaches, co-occurrence-based relatedness measures share structural limitations that affect their reliability and interpretation.

**Dimensional mismatch produces noisy estimates.** The reliability of frequency-based measures depends on having sufficient observations to estimate co-occurrence patterns accurately. In typical applications, the number of technologies substantially exceeds the number of locations—estimating a matrix of several thousand technology pairs from fewer than two hundred country observations, for instance, produces correlation structures that are largely random (Tacchella et al. 2023). The number of observations required for reliable estimation of such matrices easily reaches tens of thousands (Bun, Bouchaud, and Potters 2017), far exceeding available data. This problem is not resolved by moving to subnational levels, where harmonised data is often unavailable and the fundamental dimensional mismatch persists.

**Nested structure dominates the relatedness signal.** Location-activity networks typically exhibit strong nested structure: diversified regions produce almost everything while

less diversified regions produce only ubiquitous activities (Mariani et al. 2019). This triangular pattern means that when counting co-occurrences, one primarily detects the diversification hierarchy rather than genuine capability-based relationships. The relatedness signal becomes second-order relative to the structural pattern that generates nestedness (Tacchella et al. 2023). This explains why even randomised relatedness matrices produce predictions comparable to co-occurrence-based measures—they still capture the fact that diversified regions are likely to add activities regardless of which specific activities are involved (Bustos et al. 2012).

**Spurious associations are rarely filtered.** Studies employing co-occurrence measures often fail to test the statistical significance of the relationships they identify (Saracco et al. 2015, 2017). Without appropriate null models, spurious co-occurrences due to sampling variability are mistaken for genuine relatedness. When such tests are applied to trade data, the results are sobering: the null hypothesis of random diversification—no path dependence—is rejected for only approximately half of countries examined, with most high-income and large economies exhibiting diversification patterns that defy path dependence entirely (Coniglio et al. 2021). Whether these findings extend to patent-based technology data remains an open question, but they raise concerns about the generalisability of path-dependent diversification as a universal empirical regularity.

**Symmetry obscures directional dependencies.** All normalisation schemes above produce symmetric measures: $\phi_{t,t'} = \phi_{t',t}$. This symmetry assumes that the relationship between any two technologies operates bidirectionally with equal strength. Yet technological development often exhibits clear hierarchies (Boschma 2017). Expertise in semicon-

ductor fabrication may strongly predict future development of integrated circuit design, but the reverse need not hold—circuit designers may not subsequently develop fabrication capabilities. The capabilities required to transition from technology $i$ to technology $j$ may differ vastly from those required for the reverse transition. Symmetric measures cannot distinguish prerequisite relationships from derived relationships, enabling technologies from their applications, or identify the stepping-stone pathways that matter for strategic planning.

### 3.2.3 Relatedness Density Through Linear Aggregation

Relatedness density (Hidalgo et al. 2007) bridges the gap between pairwise technology relationships and region-specific diversification potential. Where co-occurrence measures describe technology-to-technology proximity, relatedness density aggregates these proximities relative to a region's existing portfolio, producing a single value meant to capture how feasible a new technology is for a given region. The measure has become central to diversification analysis and policy guidance, serving as the primary predictor in entry models and the basis for strategic recommendations.

The aggregation, however, imposes strong assumptions about how capabilities combine. By summing pairwise relationships, relatedness density treats the regional portfolio as a linear accumulation of independent components. Whether this adequately represents how knowledge and capabilities actually interact in enabling new activities is an empirical question that the measure's structure forecloses.

For region $r$ and technology $t$, relatedness density aggregates the relatedness between technology $t$ and all technologies in which the region currently specialises:

$$\text{relatedness density} : \Phi_{r,t} = \frac{\sum_{t' \in \mathcal{T}_r, t' \neq t} \phi_{t,t'}}{\sum_{t' \neq t} \phi_{t,t'}} \times 100$$

where $\mathcal{T}_r$ denotes the set of technologies in which region $r$ specialises (those with $\text{RTA}_{r,t} \geq 1$).

The interpretation is intuitive: the numerator sums relatedness values to technologies the region already has, while the denominator normalises by total relatedness to all technologies. High relatedness density indicates that technology $t$ is "close" to the region's existing capabilities—many of its related technologies are already present in the regional portfolio. Low relatedness density suggests that $t$ requires capabilities the region currently lacks.

This measure has proven remarkably predictive: across numerous studies using different data, time periods, and geographic contexts, higher relatedness density consistently associates with higher probability of future entry. Regions are more likely to develop specialisations in technologies that relate to what they already do well.

### 3.2.4 Structural Limitations of Linear Aggregation

The predictive success of relatedness density is not in question. The limitations concern what it cannot reveal and how its structure constrains interpretation.

**Linear aggregation obscures combinatorial information.** Relatedness density sums

pairwise relationships without regard to their interactions. This implicitly assumes that having two related technologies is simply twice as beneficial as having one—capabilities combine additively. Yet capabilities may interact in ways that linear summation cannot capture. Certain combinations may be more than the sum of their parts, creating complementarities where the joint presence of technologies $A$ and $B$ enables technology $C$ in ways that neither alone provides. Other combinations may be redundant, contributing less than their individual relatedness values would suggest. By reducing many-body relationships to independent two-body terms, linear aggregation discards precisely the combinatorial information that may distinguish feasible diversification paths from infeasible ones (Tacchella et al. 2023). Empirically, prediction models that capture many-product interactions substantially outperform density-based predictions, suggesting that linear aggregation loses significant information (Albora et al. 2023).

**Noise compounds through aggregation.** The dimensional mismatch and spurious associations affecting pairwise co-occurrence estimates do not disappear when aggregated into relatedness density—they compound. Summing noisy pairwise values produces a noisy aggregate. The normalisation by total relatedness provides some correction but cannot recover signal that was lost at the pairwise level. For regions with portfolios concentrated in rare technologies, where pairwise estimates are least reliable, relatedness density inherits and potentially amplifies these estimation problems.

**Frequency of entry conflates observation with feasibility.** The empirical regularity that higher relatedness density associates with higher entry probability is often interpreted as evidence that related diversification is more feasible and therefore should be priori-

tised. But this inference conflates observed frequency with success rate as emphasised in C. Pinheiro (2025). We observe which entries occurred, not which were attempted; the denominator—attempts—remains unobserved. Related diversification may appear more common because it is attempted more frequently, perhaps precisely because existing frameworks recommend it, rather than because it succeeds more often when attempted. Coniglio et al. (2021) addresses precisely this issue and shows that related diversification attempts succeeded 80% in Germany in contrast to 40% in the United States and France. Moreover, the observed product and technology spaces may themselves reflect past policies promoting related diversification (Andreoni and Chang 2019), creating circularity when used to justify future policy.

### 3.2.5 Implications for Contextualisation

The limitations identified above are not independent—they compound. Noisy co-occurrence estimates feed into symmetric matrices that are then linearly aggregated, producing context-free measures whose predictive success may reflect structural properties of the data (nestedness, diversification hierarchies) as much as genuine capability-based relationships. The resulting framework provides a supply-side instrument: it identifies what regions might produce given their existing portfolios, while remaining silent on demand dynamics, competitive pressures, and the institutional conditions that shape whether capability-based proximity translates into actual entry (C. Pinheiro 2025; Andreoni and Chang 2019).

This supply-side orientation has consequences. When applied to policy, relatedness-based

recommendations may systematically favour already-diversified regions whose dense portfolios generate high relatedness density across many targets, while offering limited guidance to less-diversified regions whose sparse portfolios produce low density values regardless of strategic potential. The concern that such recommendations reinforce rather than ameliorate path dependency—locking peripheral regions into constrained trajectories while core regions accumulate further advantages—has been explicitly raised in recent literature examining regional inequality and the distributional consequences of relatedness-based strategies (Flavio L. Pinheiro et al. 2025; Mealy and Coyle 2022; Hidalgo 2023).

The issue is not that relatedness is empirically wrong—the predictive regularity is robust—but that the measures as currently constructed cannot inform us about the conditions under which relatedness matters more or less, about which technologies serve as stepping stones toward others, or about how regional knowledge infrastructure, national ecosystems, and spatial factors moderate diversification possibilities. Addressing these limitations requires reconstructing the core methodological constructs in ways that relax the problematic assumptions while preserving the fundamental insight that prior capabilities shape future possibilities. This reconstruction motivates the machine learning approach developed in the following sections.

---

## 3.3 An Alternative Approach: Machine Learning for Contextualised Relatedness

The limitations outlined above—dimensional mismatch producing noisy correlations, symmetric measures obscuring directional dependencies, and linear aggregation discarding interaction effects—are not merely technical inconveniences. They represent fundamental obstacles to using relatedness for prediction and policy. If co-occurrence methods perform no better than trivial baselines, as demonstrated in Tacchella et al. (2023), then their assessment of diversification feasibility offers insufficient guidance for strategic decisions. This motivates a shift from descriptive correlation to predictive modelling.

### 3.3.1 Machine Learning as Prediction Framework

The core reframing is methodological: rather than inferring relatedness from co-occurrence patterns and hoping this correlates with future outcomes, we directly model diversification as a prediction problem. This follows recent work positioning relatedness estimation within supervised learning frameworks (Tacchella et al. 2023; Albora et al. 2023; Fessina et al. 2024). The shift offers three advantages.

First, it provides a falsifiable evaluation standard. Co-occurrence measures proliferate without systematic comparison—proximity, association strength, minimum conditional probability, and numerous variants coexist with no principled basis for selection. Out-of-sample prediction performance offers an objective criterion: methods that better forecast actual diversification outcomes are, by definition, better capturing whatever underlying

relatedness structure matters for real transitions (Albora et al. 2023). This addresses the circularity where relatedness is inferred from outcomes and then used to explain those same outcomes.

Second, prediction-based approaches can capture the many-body correlations that density-based measures discard. As Tacchella et al. (2023) emphasise, describing diversification paths as sums of binary relatedness relationships is an oversimplification. A region's probability of developing technology $i$ depends not just on pairwise similarities but on the full configuration of its existing portfolio—which combinations are present, which are absent, and how these interact. Tree-based algorithms learn precisely these complex conditional patterns.

Third, the framework naturally accommodates the asymmetric, directed relationships that symmetric co-occurrence measures cannot represent. When we train separate models for each target technology, we obtain directional importance scores: technology $t$ may strongly predict entry into technology $i$ while $i$ weakly predicts entry into $t$. This asymmetry reveals hierarchical structure—prerequisites, enabling technologies, and developmental sequences—that bidirectional similarity metrics obscure.

We apply this framework through two complementary constructs:

**Technological Potential** estimates $p_{r,t,y}$, the probability that region $r$ develops specialisation in technology $t$ by year $y$, given its existing portfolio. This replaces relatedness density. Rather than linearly summing symmetric co-occurrence frequencies, we predict region-specific feasibility using the full structure of past portfolios. The measure captures non-linear interactions and regional heterogeneity that additive approaches miss.

**Feature Importance Technology Space (FITS)** extracts directional dependencies $I_{t \to t'}$ between technologies from the trained models. This replaces symmetric technology networks. Instead of co-occurrence frequencies, we identify which technologies predict others' future adoption—and crucially, which do not predict in the reverse direction. The resulting asymmetric network reveals hierarchical technological structure invisible to symmetric measures.

### 3.3.2 Why Random Forest

Among supervised learning approaches, tree-based ensemble methods—particularly Random Forest—offer properties well-suited to relatedness estimation. We briefly outline these before presenting the technical details.

**High-dimensional sparse data.** Regional technology portfolios are high-dimensional (hundreds of technology classes) and sparse (most regions specialise in few technologies). Tree-based methods handle this naturally through recursive partitioning: each split considers only a subset of features, and the hierarchical structure means irrelevant technologies simply do not appear in decision paths. Linear methods struggle with the collinearity and sparsity typical of specialisation data; neural networks require larger samples than regional patent data typically provide.

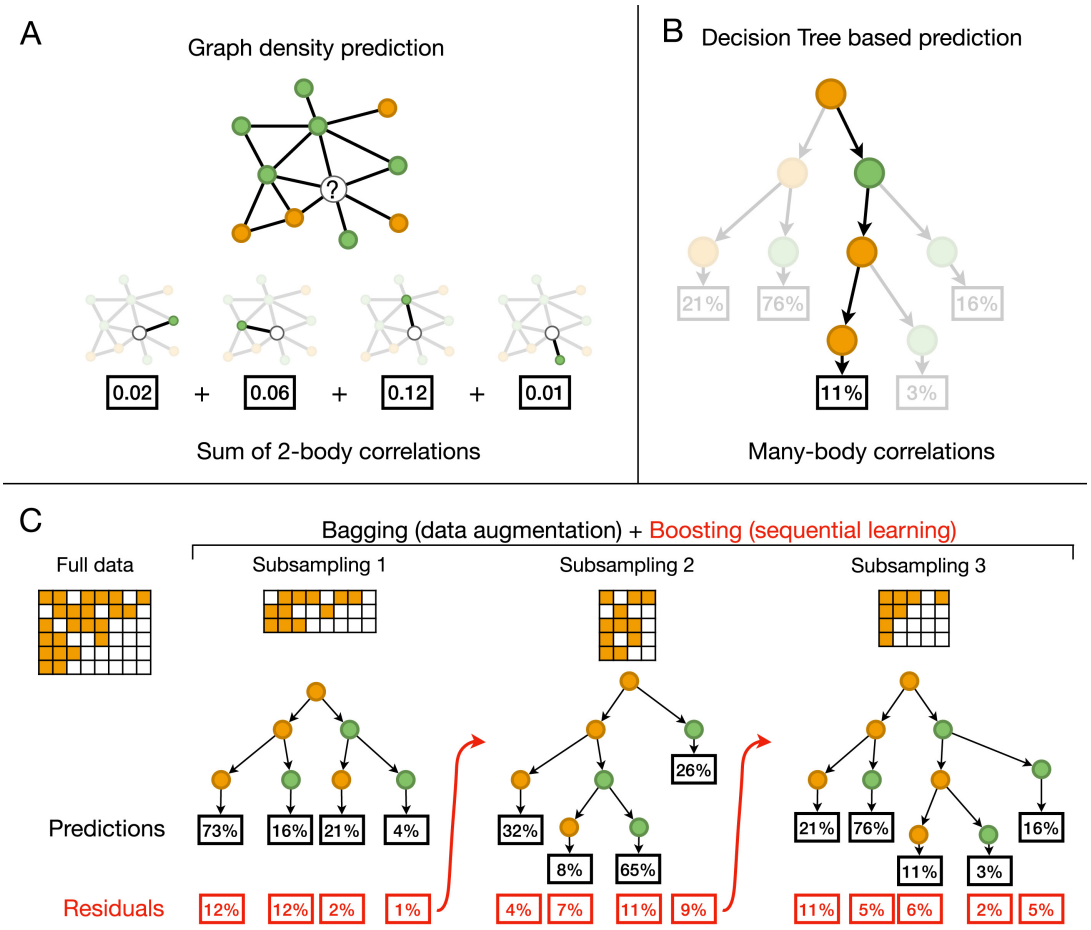
**Non-linear interactions.** The recursive splitting structure captures interactions without explicit specification. If technology $t_1$ matters for predicting $t_3$ only when $t_2$ is also present, this emerges naturally as a conditional split pattern. Density-based measures, by

construction, cannot represent such conditionality—they assume each related technology contributes independently.

**Interpretable feature importance.** Unlike black-box models, Random Forests produce feature importance scores that aggregate each predictor's contribution across all trees. These scores become the basis for FITS construction, translating prediction performance into an interpretable technology network. This maintains the explanatory value of relatedness measures while grounding them in predictive validity.

**Ensemble variance reduction.** Bootstrap aggregation across many trees reduces the variance that would plague individual decision trees, particularly given the moderate sample sizes (regions $\times$ years) available in patent data. The ensemble structure provides stability without sacrificing the flexibility needed to capture complex capability interactions.

The choice of Random Forest over alternatives thus reflects both practical constraints (sample size, interpretability requirements) and the specific structure of relatedness estimation (sparse, high-dimensional, interaction-rich). We now present the algorithm formally before describing our application.




### 3.3.3 Random Forest Algorithm

Random Forest constructs an ensemble of $B$ decision trees through bootstrap aggregating (bagging) with random feature subsampling. Each tree $T_b$ is built on a bootstrap sample $\mathcal{D}_b^*$ drawn with replacement from the original dataset.

Each tree recursively partitions the feature space through binary splits. At node $t$, we randomly select $m$ features (typically $m = \sqrt{p}$) and evaluate all possible splits within this subset. For feature $j$ and threshold $\tau$, the split creates two child nodes: $t_L = \{i : x_{ij} \leq \tau\}$ and $t_R = \{i : x_{ij} > \tau\}$.

Split quality is assessed via Gini impurity:

$$G(t) = 1 - \sum_{k=0}^{1} p_k^2(t) = 2p_0(t)p_1(t)$$

where $p_k(t) = n_k(t)/n(t)$ represents the proportion of class $k$ observations at node $t$. Gini impurity quantifies node heterogeneity: $G = 0$ indicates perfect purity (homogeneous class), while $G = 0.5$ indicates maximum impurity (equal class distribution). The optimal split maximises weighted impurity reduction:

$$\Delta G(j, \tau) = G(t) - \left[ \frac{n(t_L)}{n(t)} G(t_L) + \frac{n(t_R)}{n(t)} G(t_R) \right]$$

Weighting by relative node size prevents trivial splits that isolate single observations into pure but uninformative leaves.

Recursive splitting continues until predefined stopping criteria: node purity ($G = 0$), minimum node size threshold, or maximum tree depth. Terminal nodes are assigned the majority class of their constituent observations.

For prediction, observation $\mathbf{x}$ traverses all $B$ trees. Final classification aggregates individual tree predictions via majority voting:

$$\hat{y}(\mathbf{x}) = \text{mode}\{\hat{y}_1(\mathbf{x}), \ldots, \hat{y}_B(\mathbf{x})\}$$

Class probabilities are estimated as the proportion of trees predicting each class:

$$\hat{P}(y = 1|\mathbf{x}) = B^{-1} \sum_{b=1}^{B} \mathbb{I}[\hat{y}_b(\mathbf{x}) = 1]$$

This probability represents empirical vote share across trees. Values near 1 indicate strong consensus for class 1 (high confidence), while values near 0.5 reflect uncertainty. Unlike parametric models, these are data-driven vote proportions rather than model-based probability estimates.

The algorithm's effectiveness stems from variance reduction through decorrelated predictions. Bootstrap sampling and random feature selection reduce inter-tree correlation $\rho$, yielding ensemble variance:

$$\text{Var}(\bar{y}) = \rho\sigma^2 + \frac{1 - \rho}{B}\sigma^2$$

As $B$ increases and $\rho$ decreases, ensemble variance diminishes while maintaining the low bias of flexible tree models.

Feature importance quantifies each predictor's contribution by aggregating Gini impurity reductions:

$$I(j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in T_b : v(t) = j} \Delta G(t)$$

where the sum runs over all nodes using feature $j$ for splitting. Higher values indicate features consistently creating purer partitions. Feature importance measures predictive

association rather than causal effect, and exhibits bias toward high-cardinality features and correlated predictor sets. It ranks predictive relevance but requires caution in causal interpretation.

### 3.3.4 Training Procedure

We follow the methodology of Fessina et al. (2024) and Albora et al. (2023), originally developed for trade data. The key innovation lies in training separate models for each technology rather than constructing a single global model. For every technology $i \in \mathcal{T}$, we train a binary classification model where:

- **Outcome**: $z_{r,i,y}$ indicates whether region $r$ specialises in technology $i$ at year $y$ (RTA $\geq 1$)
- **Features**: $\{\text{RTA}_{r,t,y-\delta} : t \in \mathcal{T}, t \neq i\}$ comprises RTA values for all other technologies $\delta$ years prior

This technology-specific approach enables each model to learn unique dependency patterns. The lag $\delta = 4$ years balances predictive horizon with data availability, following conventions in the literature on capability accumulation (Andreoni and Chang 2019).

We predict specialisation for target years $y_t \in \{2008, \ldots, 2018\}$ using data from 1978 onward. For each target year $y_t$ and technology $i$:

**Training set:**

$$X_{\text{train}} = \{\text{RTA}_{r,t,y} \mid y \in [1978, y_t - 2\delta], t \neq i\}$$

$$Y_{\text{train}} = \{z_{r,i,y} \mid y \in [1978 + \delta, y_t - \delta]\}$$

**Test set:**

$$X_{\text{test}} = \{\text{RTA}_{r,t,y_t-\delta} \mid t \neq i\}$$

$$Y_{\text{test}} = \{z_{r,i,y_t}\}$$

This temporal structure ensures strict separation: models predict future specialisation using only information available at the prediction date. The expanding training window incorporates historical diversification patterns while the fixed lag prevents information leakage.

The procedure yields 7,051 models (641 technologies × 11 years). Given computational constraints, we performed cross-validation on a stratified sample of four technologies (G06G, B67B, D02J, C08J) spanning different IPC sections, applying the modal optimal parameters across all models: `mtry = 139`, `trees = 100`, `min_n = 38`. Training was conducted in R using the Ranger package (Wright and Ziegler 2017), orchestrated via the targets pipeline (Landau 2021).

## 3.4 Technological Potential

Each model produces probabilities $p_{r,i,y} = P(z_{r,i,y} = 1 \mid \text{RTA}_{r,\cdot,y-\delta})$ representing the likelihood that region $r$ develops specialisation in technology $i$ by year $y$, conditional on

its prior portfolio. These probabilities constitute **Technological Potential**—a forward-looking, region-specific measure of diversification feasibility.

The shift from correlation to prediction reframes what the measure captures. Density asks: "How similar is technology $i$ to what region $r$ already has?" Potential asks: "Given everything we know about how regions diversify, how likely is region $r$ to develop technology $i$?" The latter question—the one relevant for policy—requires a predictive approach.

High potential ($p_{r,t,y} \approx 1$) indicates a region's existing portfolio strongly predicts future specialisation—the region likely possesses necessary complementary capabilities. Low potential ($p_{r,t,y} \approx 0$) suggests capability gaps that make diversification unlikely under current conditions. Intermediate values reflect genuine uncertainty, often indicating that diversification is possible but contingent on factors beyond current portfolio composition.

Crucially, potential varies across regions for the same technology. Two regions with similar aggregate relatedness density may have quite different potential scores because their specific portfolio configurations interact differently with the target technology's requirements. This variation enables analysis of how regional context—knowledge infrastructure, institutional environment, spatial position—moderates diversification feasibility.

## 3.5 Feature Importance Technology Space (FITS)

Traditional technology networks rely on patent citations or co-occurrence patterns, each with limitations. Citations suffer from examiner additions that may not reflect actual

knowledge flows, aggregation from patent-level to technology-level that obscures directionality, and backward-looking orientation that captures historical influence rather than predictive relationships (Fessina et al. 2024). Co-occurrence networks inherit the problems outlined in our critique: symmetry, noise, and linear assumptions.

FITS addresses these by constructing an asymmetric, predictive network from the feature importance scores of our Technological Potential models. Rather than inferring relationships from co-occurrence, FITS extracts directional dependencies revealed by which technologies predict others' future adoption.

Recall that for each technology $i$, we trained a Random Forest model predicting $z_{r,i,y}$ using $\{\text{RTA}_{r,t,y-\delta} : t \neq i\}$ as features. The feature importance $I_i(t)$ quantifies how much technology $t$ contributes to predicting future specialisation in technology $i$ across all regions and time periods.

We formalise FITS as a directed, weighted network $G = (V, E, W)$ where:

- **Nodes** $V = \mathcal{T}$: the set of 641 technologies
- **Edges** $E$: directed connections $(t \to i)$ for all $t, i \in \mathcal{T}, t \neq i$
- **Weights** $W_{t \to i} = I_i(t)$: feature importance of technology $t$ in the model predicting technology $i$

We normalise weights within each target technology:

$$W_{t \to i} = \frac{I_i(t)}{\sum_{t' \neq i} I_i(t')}$$

ensuring incoming edge weights sum to 1 for each technology. This normalisation facilitates comparison across technologies with different overall predictability.

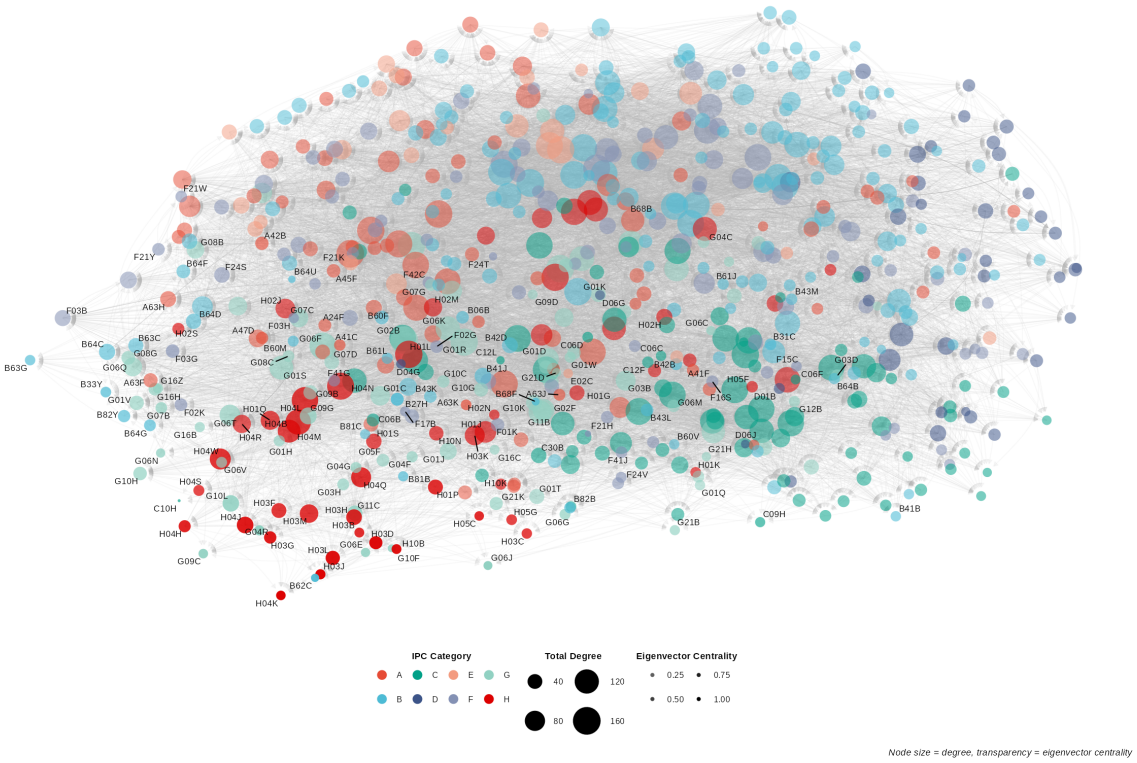



Figure 3.1: Technology Network (based on 2018 patent data)

The defining property of FITS is asymmetry: $W_{t \to i} \neq W_{i \to t}$ in general. This asymmetry captures hierarchical technological dependencies invisible to symmetric measures:

- $W_{t \to i} \gg W_{i \to t}$: Technology $t$ is a prerequisite or enabler for $i$. Expertise in $t$ predicts future development of $i$, but not vice versa. This pattern identifies foundational technologies and developmental sequences.
- $W_{t \to i} \approx W_{i \to t}$: Technologies are complementary peers with mutual predictive relationships, suggesting shared capability requirements or co-evolution.

- $W_{t \to i} \ll W_{i \to t}$: Technology $i$ enables $t$, reversing the hierarchical relationship.

This directional structure reveals technological trajectories: regions can identify which current capabilities open paths toward desired future technologies, and which apparent similarities mask asymmetric dependencies.

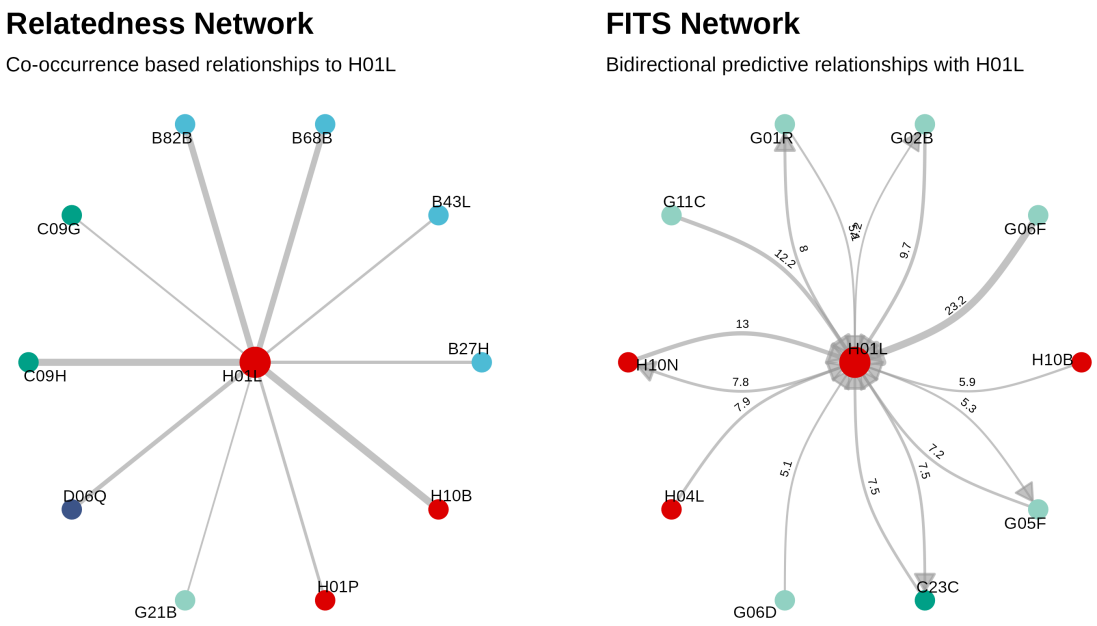
The contrast between co-occurrence relatedness and FITS is sharpest when examining specific technologies. Consider H01L (semiconductors) and D06Q (textile decoration).

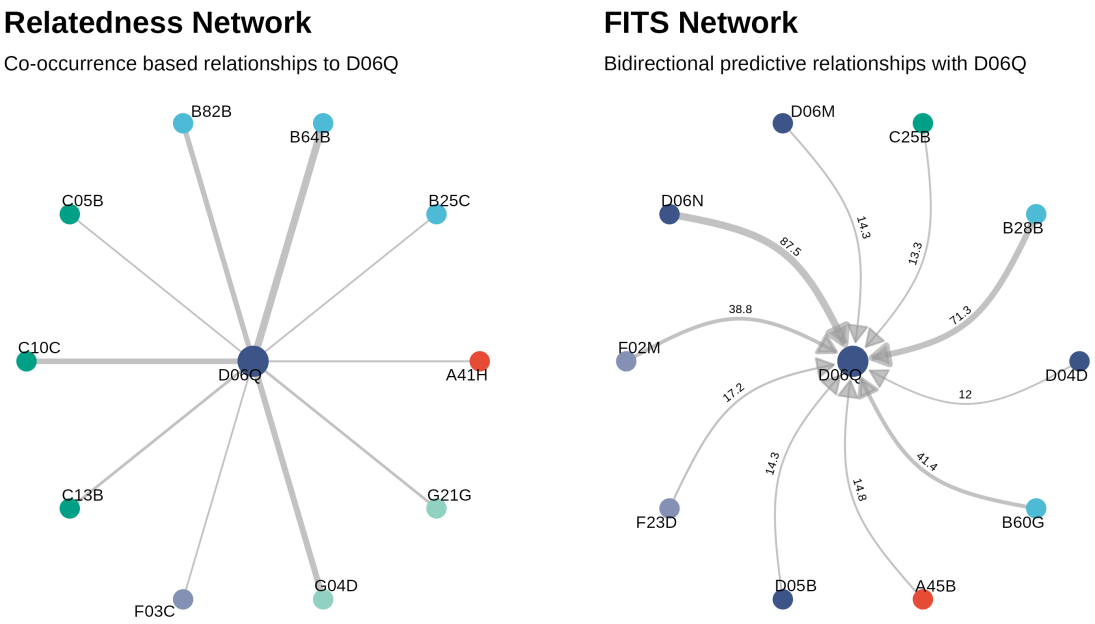
For semiconductors, co-occurrence relatedness identifies technologies that frequently appear together in regional portfolios. Some connections are sensible: semiconductor memories (H10B), packaging technologies (B68B). But the measure also surfaces implausible links—textile decoration (D06Q), woodworking (B27H), fusion reactors (G21B)—that appear not because of genuine technological affinity but because they happen to co-occur in diverse portfolios. This is the noise problem: frequency-based aggregation treats all co-presence equally, regardless of whether it reflects shared capabilities or statistical coincidence.

FITS reveals a different structure. Rather than asking which technologies appear alongside semiconductors, it asks which technologies predict future semiconductor development—and which technologies semiconductors predict. The asymmetry is striking: strong edges flow FROM application domains (digital computing G06F, telecommunications H04L, control systems G05F) TO semiconductors, indicating these fields depend on semiconductor capabilities. Weaker reverse edges show that semiconductor expertise does not particularly predict entry into computing or telecommunications. H01L emerges as an enabling technology whose importance lies in what it makes possible rather than what leads to it.

The contrast sharpens further with D06Q (textile decoration), a sparse technology where co-occurrence produces near-total noise: elevators (B64B), timing devices (G04D), chemical production (C13B) appear simply because D06Q is rare and co-occurs randomly with whatever else rare-technology regions happen to have. FITS filters this effectively, identifying genuine textile domain relationships—sewing (D05B), textile treatment (D06M), wall coverings (D06N)—that share actual functional connections.

This example illustrates how FITS captures the functional architecture of innovation networks that co-occurrence frequencies obscure. The method's ability to distinguish signal from statistical artifact, combined with its directional structure revealing technological hierarchies, provides the foundation for contextualised analysis of regional diversification patterns.






Technological Potential provides region-specific feasibility estimates grounded in predictive performance rather than descriptive correlation. FITS reveals the directional, hierarchical structure of technological dependencies that symmetric measures cannot represent. Together, these constructs enable the contextualised analysis our research questions require: examining how regional knowledge infrastructure, national innovation systems, and spatial factors moderate the relationship between technological characteristics and diversification outcomes.

## 3.6 Coherence: Bridging Technology Networks and Regional Portfolios

FITS identifies technology-to-technology relationships. Technological Potential quantifies regional diversification feasibility. **Coherence** bridges these levels by measuring the alignment between a technology's position in the FITS network and a region's existing specialization structure. In a way coherence can be viewed as sector level relatedness density.

Consider two regions, both lacking specialization in technology $i$, both having similar potential $p_{r,i,y}$. However, Region A specializes in technologies that are strong predictors of $i$ (high incoming FITS edges), while Region B specializes in technologies unrelated to $i$ in the network. Coherence captures this difference: Region A has high coherence with $i$ (its portfolio aligns with $i$'s network prerequisites), while Region B has low coherence (misalignment).

This metric operationalizes the "knowledge coherence" and "cognitive proximity" concepts from innovation literature (Neffke, Henning, and Boschma 2011; Boschma 2015) using our directional network structure. It enables testing whether diversification success depends not just on potential (predicted feasibility) but on the structural fit between regional portfolios and technology network positions.

For each region $r$, technology $i$, and IPC category $c$, we construct two embedding vectors capturing $i$'s directional network position and compare them to $r$'s average embeddings for technologies in category $c$.

**Technology embeddings** (individual technology $i$):

- Incoming: $\text{embcat\_to}_{i,c} = \frac{\sum_{t \in c} W_{t \to i}}{|\{t \in c : W_{t \to i} > 0\}|}$ (average FITS weight from category $c$ to technology $i$)
- Outgoing: $\text{embcat\_from}_{i,c} = \frac{\sum_{t' \in c} W_{i \to t'}}{|\{t' \in c : W_{i \to t'} > 0\}|}$ (average FITS weight from technology $i$ to category $c$)

**Regional average embeddings** (region $r$, category $c$):

- Incoming: $\overline{\text{embcat\_to}}_{r,c} = \frac{1}{|S_{r,c}|} \sum_{t \in S_{r,c}} \text{embcat\_to}_{t,c}$ where $S_{r,c} = \{t \in c : \text{RTA}_{r,t,y} \geq 1\}$
- Outgoing: $\overline{\text{embcat\_from}}_{r,c} = \frac{1}{|S_{r,c}|} \sum_{t \in S_{r,c}} \text{embcat\_from}_{t,c}$

**Coherence** is the cosine similarity between technology $i$'s directional embeddings and region $r$'s average directional embeddings for category $c$:

$$\text{Coherence}_{r,i,c,y} = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{||\mathbf{v}_1|| \cdot ||\mathbf{v}_2||}$$

where:

- $\mathbf{v}_1 = [\text{embcat\_to}_{i,c}, \overline{\text{embcat\_to}}_{r,c}]$
- $\mathbf{v}_2 = [\text{embcat\_from}_{i,c}, \overline{\text{embcat\_from}}_{r,c}]$

Coherence ranges from -1 to 1:

- **High coherence** ($\approx 1$): Technology $i$'s FITS network position (both incoming and outgoing connections to category $c$) closely matches the average network position of

technologies in which region $r$ specializes within category $c$. The region's existing capabilities align with the structural prerequisites and consequences of technology $i$.

- **Neutral coherence** ($\approx 0$): Misalignment between technology $i$'s relational structure and regional specialization patterns.
- **Negative coherence** ($\approx -1$): Technology $i$'s network position is opposite to the region's specialization structure (e.g., $i$ receives inputs from categories where the region sends outputs).

### 3.6.1 Empirical Application

Coherence serves two roles in our empirical analysis:

1. **Interaction with Potential**: Test whether high potential translates to actual diversification only when coherence is also high (H2a: technology-specific characteristics moderated by regional knowledge coherence)
2. **Regional Infrastructure Measure**: Aggregate coherence across a region's non-specialized technologies indicates how well the regional portfolio is "positioned" in the FITS network for future diversification (captures knowledge infrastructure quality)

By incorporating coherence, we test whether successful diversification requires not just predicted feasibility (potential) and related capabilities (traditional relatedness), but also structural alignment between regional portfolios and network prerequisites—a form of contextualization that traditional measures cannot capture.

# 4 Summary of Methodological Framework

Our approach replaces traditional relatedness constructs with machine learning-derived measures that enable contextualized diversification analysis:

- **Technological Potential** ($p_{r,t,y}$): Region-specific, non-linear, time-varying probabilities replace linear relatedness density
- **FITS Network** ($W_{t \to t'}$): Asymmetric, predictive dependencies replace symmetric co-occurrence-based relatedness

- **Coherence** ($\text{Coherence}_{r,t,c,y}$): Structural alignment between regional portfolios and technology network positions captures knowledge infrastructure quality

Together, these measures allow testing how diversification is contingent on regional knowledge infrastructure (RQ2), national ecosystem characteristics (RQ3), and spatial factors (RQ4) in ways that traditional relatedness frameworks cannot—addressing the core problem of contextualizing diversification strategies beyond path dependency identification.

# 5 Empirical Design

# 6 Results

# 7 Summary

In summary, this book has no content whatsoever.

# References


Albora, Giambattista, Luciano Pietronero, Andrea Tacchella, and Andrea Zaccaria. 2023. "Product Progression: A Machine Learning Approach to Forecasting Industrial Upgrading." *Scientific Reports* 13 (1): 1481.

Andreoni, Antonio, and Ha-Joon Chang. 2019. "The Political Economy of Industrial Policy: Structural Interdependencies, Policy Alignment and Conflict Management." *Structural Change and Economic Dynamics* 48: 136–50.

Balassa, Bela. 1965. "Trade Liberalisation and 'Revealed' Comparative Advantage 1." *The Manchester School* 33 (2): 99–123.

Balland, PA, and R Boschma. 2019. "Smart Specialization: Beyond Patents." *European Commission: Brussels, Belgium*.

Balland, Pierre-Alexandre, and David Rigby. 2017. "The Geography of Complex Knowledge." *Economic Geography* 93 (1): 1–23.

Boschma, Ron. 2015. "Towards an Evolutionary Perspective on Regional Resilience." *Regional Studies* 49 (5): 733–51.

———. 2017. "Relatedness as Driver of Regional Diversification: A Research Agenda."

*Regional Studies* 51 (3): 351–64.

Boschma, Ron, Ernest Miguelez, Rosina Moreno, and Diego B Ocampo-Corrales. 2023. "The Role of Relatedness and Unrelatedness for the Geography of Technological Breakthroughs in Europe." *Economic Geography* 99 (2): 117–39.

Bun, Joël, Jean-Philippe Bouchaud, and Marc Potters. 2017. "Cleaning Large Correlation Matrices: Tools from Random Matrix Theory." *Physics Reports* 666: 1–109.

Bustos, Sebastián, Charles Gomez, Ricardo Hausmann, and César A Hidalgo. 2012. "The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems." *PloS One* 7 (11): e49393.

Coniglio, Nicola D, Davide Vurchio, Nicola Cantore, and Michele Clara. 2021. "On the Evolution of Comparative Advantage: Path-Dependent Versus Path-Defying Changes." *Journal of International Economics* 133: 103522.

Diodato, Dario, Lorenzo Napolitano, Emanuele Pugliese, and Andrea Tacchella. 2023. "Economic Complexity for Regional Industrial Strategies." Joint Research Centre.

E, Pugliese, and Tacchella A. 2021. "Economic Complexity Analytics: Country Factsheets." https://doi.org/10.2760/368138.

Fessina, Massimiliano, Giambattista Albora, Andrea Tacchella, and Andrea Zaccaria. 2024. "Identifying Key Products to Trigger New Exports: An Explainable Machine Learning Approach." *Journal of Physics: Complexity* 5 (2): 025003.

Hartmann, Dominik, Miguel R Guevara, Cristian Jara-Figueroa, Manuel Aristarán, and César A Hidalgo. 2017. "Linking Economic Complexity, Institutions, and Income Inequality." *World Development* 93: 75–93.

Hausmann, Ricardo, and César A Hidalgo. 2011. "The Network Structure of Economic

Output." *Journal of Economic Growth* 16: 309–42.

Hidalgo, César A. 2021. "Economic Complexity Theory and Applications." *Nature Reviews Physics* 3 (2): 92–113.

———. 2023. "The Policy Implications of Economic Complexity." *Research Policy* 52 (9): 104863.

Hidalgo, César A, Pierre-Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, et al. 2018. "The Principle of Relatedness." In *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems 9*, 451–57. Springer.

Hidalgo, César A, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. 2007. "The Product Space Conditions the Development of Nations." *Science* 317 (5837): 482–87.

Landau, William Michael. 2021. "The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing." *Journal of Open Source Software* 6 (57): 2959.

Lee, Changyong, Ohjin Kwon, Myeongjung Kim, and Daeil Kwon. 2018. "Early Identification of Emerging Technologies: A Machine Learning Approach Using Multiple Patent Indicators." *Technological Forecasting and Social Change* 127: 291–303.

Li, Yang, and Frank MH Neffke. 2024. "Evaluating the Principle of Relatedness: Estimation, Drivers and Implications for Policy." *Research Policy* 53 (3): 104952.

Mariani, Manuel Sebastian, Zhuo-Ming Ren, Jordi Bascompte, and Claudio Juan Tessone. 2019. "Nestedness in Complex Networks: Observation, Emergence, and Implications." *Physics Reports* 813: 1–90.

Mealy, Penny, and Diane Coyle. 2022. "To Them That Hath: Economic Complexity and

Local Industrial Strategy in the UK." *International Tax and Public Finance* 29 (2): 358–77.

Neffke, Frank, Martin Henning, and Ron Boschma. 2011. "How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions." *Economic Geography* 87 (3): 237–65.

Pinheiro, Cristina. 2025. "Relatedness and Economic Complexity as Tools for Industrial Policy: Insights and Limitations." *Structural Change and Economic Dynamics* 72: 1–10.

Pinheiro, Flavio L, Pierre-Alexandre Balland, Ron Boschma, and Dominik Hartmann. 2025. "The Dark Side of the Geography of Innovation: Relatedness, Complexity and Regional Inequality in Europe." *Regional Studies* 59 (1): 2106362.

Pinheiro, Flávio L, Dominik Hartmann, Ron Boschma, and César A Hidalgo. 2022. "The Time and Frequency of Unrelated Diversification." *Research Policy* 51 (8): 104323.

Saracco, Fabio, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. 2015. "Randomizing Bipartite Networks: The Case of the World Trade Web." *Scientific Reports* 5 (1): 10595.

Saracco, Fabio, Mika J Straka, Riccardo Di Clemente, Andrea Gabrielli, Guido Caldarelli, and Tiziano Squartini. 2017. "Inferring Monopartite Projections of Bipartite Networks: An Entropy-Based Approach." *New Journal of Physics* 19 (5): 053022.

Tacchella, Andrea, Andrea Zaccaria, Marco Miccheli, and Luciano Pietronero. 2023. "Relatedness in the Era of Machine Learning." *Chaos, Solitons & Fractals* 176: 114071.

Wright, Marvin N, and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." *Journal of Statistical*

*Software* 77: 1–17.

Zaccaria, Andrea, Saurabh Mishra, Masud Z Cader, and Luciano Pietronero. 2018. "Integrating Services in the Economic Fitness Approach." *World Bank Policy Research Working Paper*, no. 8485.