LUND UNIVERSITY
School of Economics and Management

Master's Programme in Economic Development

Legacy Knowledge, Technology Cycle Time, and Regional Catch-Up: Evidence from Automotive Innovation Clusters in China and Europe

by

Wenyi Yan (we6538ya-s@student.lu.se)

Abstract: This thesis investigates how technology cycle time (TCT) and legacy knowledge accumulation (LKA) jointly shape EV innovation performance and regional catch-up dynamics across automotive clusters in China and Europe. Using PCT patent data (2000–2024) and HDBSCAN clustering to identify 22 Chinese and 57 European innovation clusters, the study employs PPML-HDFE and OLS panel estimators. The results show that both lower LKA and shorter TCT are associated with significantly higher EV innovation output, with stronger effects in Chinese latecomer clusters, where faster knowledge renewal supports more rapid convergence toward the European frontier. European incumbent clusters, by contrast, exhibit a gradual erosion of technological advantage under conditions of high legacy dependence and slower technological iteration. These findings advance the catch-up literature by integrating knowledge renewal dynamics and legacy knowledge structure into a regional framework, providing cross-regional evidence that both dimensions shape innovation performance and catch-up trajectories during paradigm shifts.

Keywords: Technological Catch-up, Path Dependence, Technology Cycle Time, Knowledge Structure, Innovation Clusters, Electric Vehicles

Course code: EKHS21
Master's Thesis (15 credits ECTS)
May 2026
Supervisor: Olof Ejermo
Examiner: Josef Taalbi
Word Count: 15883

Acknowledgements

I would first like to thank my supervisor, Olof Ejermo, for his valuable feedback, guidance, and support throughout the thesis process. I am also grateful to Tobias Axelsson for his timely assistance and helpful advice. In addition, I would like to thank all the teachers and faculty members at Lund University whose courses and discussions helped broaden my perspective throughout the programme.

I would also like to thank my classmates, friends, and corridor mates in Lund. The discussions during seminars, shared meals in the kitchen, and countless conversations outside academia made this period far more manageable and memorable.

Finally, I would like to thank my family for their constant support and encouragement.

Table of Contents

List of Tables

List of Appendix Tables

List of Figures

1 Introduction

The global automotive industry is currently undergoing a major technological transition from the long-established internal combustion engine (ICE) regime toward electrification (Song et al., 2025). This transition has generated substantial industrial restructuring across global automotive value chains and innovation systems. For decades, technological leadership in the automotive sector was concentrated within established European clusters that accumulated extensive engineering capabilities and organizational experience under the ICE paradigm (Aghion et al., 2016). However, the emergence of electric vehicle (EV) technologies has begun to reshape the competitive structure of the industry and create new opportunities for latecomer economies (Altenburg et al., 2022).

Among these latecomers, China has experienced particularly rapid growth in both EV production and technological activity. Although Chinese firms and regional innovation systems achieved substantial manufacturing scale within the traditional automotive industry, they historically remained in a catch-up position regarding core technological capabilities. However, Chinese actors have increasingly expanded their presence within emerging EV technologies (Altenburg et al., 2022). At the same time, European incumbents continue to possess substantial advantages in accumulated technological capabilities, industrial specialization, and legacy knowledge bases. This coexistence of strong incumbent capabilities and rapid latecomer catch-up presents an important empirical puzzle for the literature on technological transition and evolutionary catch-up.

Conventional evolutionary theories emphasize the cumulative nature of technological development, suggesting that historical knowledge accumulation should strengthen the long-term innovative capacity of incumbent regions. Under this logic, established automotive clusters with deep ICE-related knowledge bases would be expected to maintain their technological leadership during industrial transition. However, recent developments in the EV sector suggest that historical technological advantages may not always translate into sustained leadership under conditions of paradigm shift. In particular, rapidly evolving technological trajectories may reduce the value of existing specialization while increasing the importance of knowledge renewal and adaptive capability (Leonard-Barton, 1992; Lee, 2024).

Despite growing research on electrification and technological transition, two important limitations remain in the existing literature. First, many studies rely on national-level analysis or predefined administrative regions, which may not adequately capture the endogenous spatial organization of innovation activities. Second, while previous research has examined windows of opportunity created by electrification, fewer studies have jointly investigated how legacy knowledge structures and technological iteration dynamics influence regional catch-up trajectories during paradigm shifts.

This thesis addresses these gaps by examining automotive innovation clusters across China and Europe during the electrification transition. The primary aim of this study is to investigate the role of legacy knowledge structures and technological renewal dynamics in regional EV innovation during the automotive electrification transition. In addition, the study aims to improve the spatial identification of regional innovation systems by employing endogenously defined innovation clusters generated through a density-based clustering framework. Accordingly, this thesis addresses the following research questions:

Research Question:

How are legacy knowledge accumulation and technology cycle time associated with regional electric vehicle innovation output, and how are they associated with the catch-up trajectories of latecomer clusters and the leadership positioning of incumbent clusters during the electrification paradigm shift?

To answer the research question, the study constructs a cross-regional patent panel dataset and employs a data-driven hierarchical density-based spatial clustering algorithm with noise (HDBSCAN) to identify innovation clusters based on the spatial concentration of technological activity. The empirical analysis employs Poisson Pseudo-Maximum Likelihood estimators with high-dimensional fixed effects (PPML-HDFE) together with linear panel models to examine the relationship between historical knowledge structures, technological renewal dynamics, and regional innovation outcomes.

The findings contribute to the literature in three ways. First, the study extends research on technological catch-up by integrating legacy knowledge accumulation and technology cycle dynamics into a unified analytical framework. Second, it contributes to evolutionary economic geography by emphasizing the endogenous spatial structure of innovation clusters during periods of technological transition. Third, the study provides new empirical evidence on uneven regional adaptation patterns during the automotive electrification transition, drawing on cross-regional comparisons between Chinese latecomer and European incumbent clusters.

The remainder of the thesis is organized as follows. Chapter 2 reviews the development of the global electric vehicle industry and recent patenting trends. Chapter 3 presents the theoretical framework and develops the research hypotheses. Chapter 4 introduces the data, clustering procedures, and econometric methodology. Chapter 5 reports the empirical results and robustness checks. Chapter 6 discusses the main findings and policy implications. Finally, Chapter 7 concludes and outlines directions for future research.

2 Context

In the global electric vehicle sector, China has achieved remarkable commercial success and market penetration. This market leadership is largely driven by robust infrastructure development, proactive policy support, and a vast domestic consumer base (Liu et al., 2020; Zhao et al., 2020; Altenburg et al., 2022). However, an analysis of international PCT patent data, shown in Figure 2.1, reveals a different pattern regarding codified technological activity. Over a long observation period, European entities have maintained a substantial lead in EV patent accumulation. This long-term advantage in EV patent accumulation reflects an earlier initiation of R&D activities and a deeply established automotive engineering foundation in Europe (Aghion et al., 2016; Liu et al., 2020; Zhao et al., 2020). Despite this historical gap, the longitudinal trend indicates a rapid and accelerating technological catch-up process by internationally oriented Chinese innovation clusters. The recent trajectory shows that China's annual patent output is narrowing the gap with the European aggregate. Most notably, the Chinese output curve has recently converged with and, at the end of the observation window, marginally surpassed the annual output of Germany, which is the largest individual European automotive powerhouse. This empirical observation is consistent with recent studies suggesting that China's accelerated transition from imitative to independent innovation benefits from the technological windows of opportunity and leapfrogging strategies within the electromobility paradigm (Liu et al., 2020; Altenburg et al., 2022; Ren et al., 2025). Such dynamics may also be facilitated by the lower entry barriers associated with shorter-cycle technology domains and the proactive capability accumulation of pioneering firms (Zhu & Li, 2025; Whitfield & Wuttke, 2026). This distinct divergence between commercial market dominance and the underlying codified knowledge base highlights the complexity of the current paradigm shift. Furthermore, while national-level aggregates illustrate the overall trajectory of catch-up, they inherently mask the intense spatial heterogeneity and localized nature of knowledge creation (Asheim & Gertler, 2005; Binz & Truffer, 2017). In particular, the structural knowledge characteristics of regional innovation systems, including inherited technological specialization and the capacity for knowledge renewal, remain underexplored in explaining cross-regional differences in EV innovation dynamics at the cluster level. This, therefore, underscores the importance of moving beyond macro-level aggregates to investigate the deeper structural factors underlying the catch-up trajectories of Chinese and European innovation clusters (Li et al., 2026).

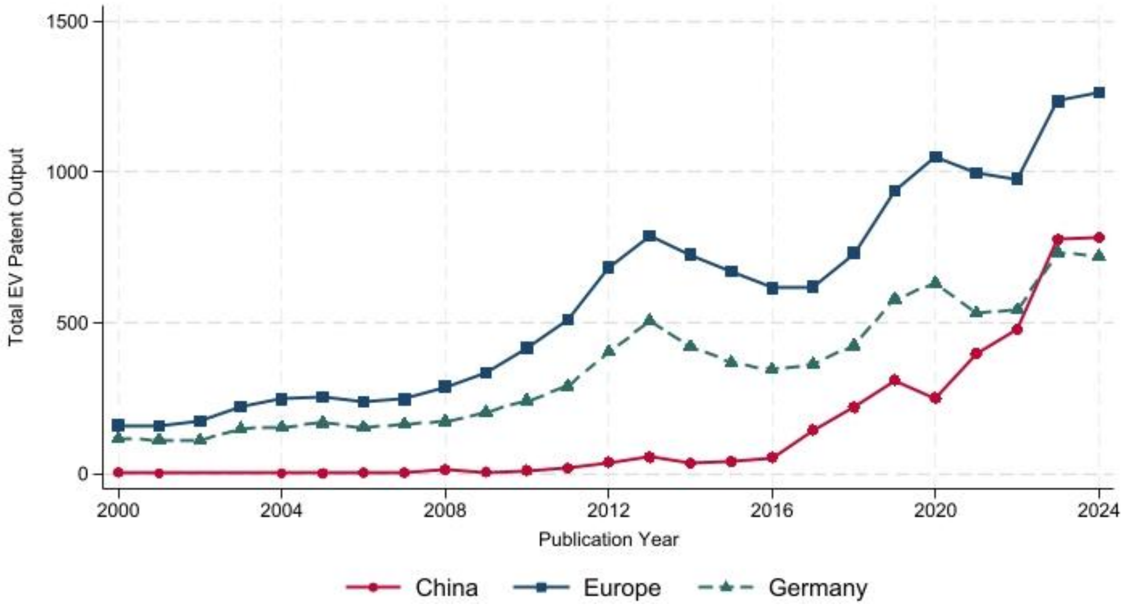

Figure 2.1: Annual EV PCT patent trends in China, Europe, and Germany
Source: Author's own elaboration based on Google Patents Public Data, retrieved from Google BigQuery.

3 Theory and Previous Research

This chapter develops the theoretical framework for analyzing regional catch-up trajectories and innovation performance during the automotive electrification transition. Section 3.1 introduces the concepts of paradigm shifts and evolutionary catch-up. Section 3.2 then turns to the spatial dimension by discussing regional innovation systems and their resilience to technological shocks. Section 3.3 focuses on two structural characteristics that may shape the adaptation process to technological paradigm shift: legacy knowledge accumulation and technology cycle time. Building on this foundation, Section 3.4 identifies the main theoretical gaps in the literature and presents the three hypotheses examined in the study."

3.1 Paradigm Shift and Technological Catch-up

The concept of paradigm shift originates from the philosophy of science. Kuhn argued that scientific development does not proceed through the linear accumulation of knowledge, but through discontinuous transformations in which an existing paradigm is replaced by a new and incompatible one during periods of crisis (Shapere, 1964; Kuhn, 2012, pp.91-111). This notion of discontinuous change was subsequently incorporated into evolutionary economics and innovation studies (von Tunzelmann et al., 2008). Within the evolutionary economics tradition, scholars distinguished between incremental improvements along established technological trajectories and radical technological changes associated with the emergence of new technological paradigms (Dosi, 1982; Cimoli & Dosi, 1995; Christensen & Rosenbloom, 1995). Existing technological regimes are often constrained by established routines, accumulated knowledge bases, institutional structures, and existing value networks, which may limit further technological progress when underlying practices lack scientific controllability (Nelson & Winter, 1982a, pp.258-260; Dosi, 1982; Christensen & Rosenbloom, 1995; Nelson, 2008). Consequently, major technological transitions frequently require the creation of new technological trajectories and organizational adaptations (Christensen & Rosenbloom, 1995; Nelson, 2008). Subsequent studies extended the paradigm shift perspective to the macroeconomic level (von Tunzelmann et al., 2008). In this view, technological revolutions involve not only the replacement of core technologies but also broader and pervasive transformations in industrial structures, institutions, and production systems (Perez, 2002, pp.8-22; Santarelli et al., 2023). This literature emphasizes that the diffusion of a new techno-economic paradigm is typically accompanied by periods of financial turbulence, institutional adjustment, and structural reorganization before a new stage of economic growth can emerge (Perez, 2002, pp.97-125; Santarelli et al., 2023).

The structural turbulence generated by a technological paradigm shift systematically generates strategic windows of opportunity for latecomer economies to catch up (Perez & Soete, 1988; Lee & Malerba, 2017). The concept of technological catch-up describes the process through which latecomers narrow the capability gap with technological leaders (Abramovitz, 1986; Malerba & Lee, 2021). Early studies emphasized that late industrialization does not simply replicate the development path of advanced economies. Instead, backwardness may provide unique opportunities for accelerated growth under appropriate institutional conditions (Gerschenkron, 1962). However, this catch-up process is not automatic. Successful convergence depends on the development of adequate social capabilities, including learning capacity, institutional support, and human capital (Abramovitz, 1986; Fakhimi & Miremadi, 2022). During periods of techno-economic transition, established technological trajectories weaken and substantially lower the entry barriers for new competitors (Perez & Soete, 1988; Lee & Malerba, 2017; Landini et al., 2017). Nevertheless, the existence of such windows alone does not guarantee success. Cohen and Levinthal (1990) emphasized that firms and regions must possess sufficient absorptive capacity to identify and assimilate external knowledge. Without sustained capability accumulation, latecomers may fail to benefit from technological change even when major disruptions occur (Bell & Pavitt, 1993; Davy & Hansen, 2023). Studies further suggest that successful latecomers often adopt strategic forms of technological development rather than directly imitating incumbents (Fagerberg & Godinho, 2005). Through linkage, leverage, and selective technological specialization, latecomers may bypass existing technological trajectories via "stage-skipping" or "path-creating" strategies and achieve technological leapfrogging in emerging sectors (Lee & Lim, 2001; Fagerberg & Godinho, 2005; Mathews, 2006; Lee, 2013, pp.30-37). This theoretical consensus highlights that catch-up dynamics are jointly shaped by exogenous technological opportunities and the endogenous capability-building and responsive processes of latecomers (Landini et al., 2017; Malerba & Lee, 2021). Such a dynamic interplay provides a crucial theoretical lens for examining how regional innovation systems restructure their capabilities during the electrification paradigm shift in the automotive industry.

This paradigm shift of the global automotive industry has created important windows of opportunity for latecomer economies within the automotive sector (Perez & Soete, 1988; Altenburg et al., 2022). China has emerged as the most significant latecomer in this transition by strategically promoting the development of new energy vehicles through industrial policy support and market expansion (Altenburg et al., 2012; Zhao et al., 2020). Existing studies suggest that these policies facilitated domestic capability accumulation and stimulated technological innovation in the electric vehicle sector (Liu et al., 2020). However, the Chinese catch-up process cannot be explained solely through state intervention. Firm-level studies emphasize that pioneering companies such as BYD and CATL played an important role in developing technological capabilities before large-scale policy support (Whitfield & Wuttke, 2026). More broadly, the literature suggests that successful catch-up in China resulted from the interaction between industrial policy, capability accumulation, and technological specialization in emerging sectors. In particular, several studies argue that short-cycle technologies such as lithium batteries created favorable conditions for latecomer entry by weakening the advantages of incumbent firms (Altenburg et al., 2022; Zhu & Li, 2025).

Despite China's rapid expansion in the electric vehicle market, the extent of technological catch-up remains debated. While some studies identify evidence of technological leapfrogging in specific component sectors, others argue that China continues to lag behind established leaders in certain dimensions of technological quality and frontier innovation (Liu et al., 2020; Zhu & Li, 2025). This ongoing debate highlights the importance of examining the deeper structural and regional drivers underlying China's innovation dynamics during the current automotive paradigm shift.

3.2 Regional Innovation Systems and Resilience

The concept of regional innovation systems (RIS) emphasizes that innovation is embedded within geographically localized networks of firms, institutions, and supporting organizations (Cooke, 2002; Iammarino, 2005). Rather than viewing innovation as an isolated firm-level activity, the RIS perspective highlights the importance of interactions among universities, firms, governments, financial institutions, and other regional actors in facilitating knowledge creation, diffusion, and application (Cooke et al., 1997; Autio, 1998; Asheim & Gertler, 2005). Within the broader framework of the learning economy, innovation is understood as a cumulative and interactive learning process shaped by institutional structures and regional capabilities (Lundvall, 1992; Lundvall, 2016). Geographical proximity plays an important role in innovation dynamics by facilitating the exchange of tacit and localized knowledge (Maskell & Malmberg, 1999; Asheim & Isaksen, 2002; Asheim & Gertler, 2005； Iammarino, 2005). Regions characterized by dense institutional networks and strong inter-firm linkages are therefore more likely to generate knowledge spillovers and sustain technological upgrading (Frenken et al., 2007). From an evolutionary perspective, regional innovation systems develop through path-dependent processes in which existing industrial structures, institutional arrangements, and historical capabilities shape future innovation trajectories (Boschma & Frenken, 2006; Martin & Sunley, 2006; Neffke et al., 2011; Tanner, 2014). As a result, regions differ substantially in their capacity to adapt to technological transitions, absorb unforeseen shocks, and exploit emerging industrial opportunities (Simmie & Martin, 2010; Pike et al., 2010; Boschma, 2015).

From an evolutionary perspective, the long-term competitiveness of regional innovation systems depends not only on the accumulation of specialized knowledge but also on the capacity to recombine existing capabilities and adapt to changing technological environments and develop new growth paths (Grabher & Stark, 1997; Asheim et al., 2011; Boschma, 2015; Balland et al., 2019). During periods of technological paradigm shift, regions face growing pressures for industrial restructuring and knowledge renewal. These uneven regional responses to technological disruption have led scholars to increasingly focus on the concept of regional resilience.

Within evolutionary economic geography, regional resilience is increasingly understood as the capacity of regions to adapt to structural change and sustain development under conditions of economic and technological disruption (Simmie & Martin, 2010; Boschma, 2015). Rather than simply returning to a previous equilibrium after external shocks, resilient regions are able to reorganize their industrial structures, renew their knowledge bases, and develop new growth paths in response to changing technological environments (Martin & Sunley, 2015). This perspective is particularly relevant during periods of technological paradigm shift, when existing industrial trajectories may become exhausted, and regions face growing pressures for industrial transformation (Christopherson et al., 2010; Pike et al., 2010). Regions with stronger innovation capabilities, diversified knowledge structures, and adaptive institutional networks are generally better positioned to support path renewal and technological upgrading (Frenken et al., 2007; Bristow & Healy, 2014; Boschma, 2015). In contrast, regions characterized by technological lock-in and limited adaptive capacity may struggle to transition toward emerging industrial paradigms (Grabher, 1993; Martin & Sunley, 2006; Hassink, 2010).

Empirical research on regional resilience and innovation dynamics commonly relies on administrative regions as the primary unit of analysis. However, the interactive learning processes and localized knowledge spillovers that underpin regional innovation systems often extend beyond formal territorial boundaries (Porter, 1998; Bathelt et al., 2004). From this perspective, innovation clusters can be understood as the spatial expression of regional innovation systems, emerging through geographic proximity, industrial concentration, and dense patterns of interaction (Audretsch & Feldman, 1996; Porter, 1998; Asheim & Isaksen, 2002).

The use of fixed administrative units may therefore fragment continuous innovation networks and obscure the actual geography of technological adaptation, as pre-defined territorial borders fail to capture complex relational activities (Binz & Truffer, 2017). This issue becomes important during periods of technological paradigm shift, when the spatial organization of innovation activities may evolve rapidly across regions and industries. Accurately examining regional adaptation consequently requires approaches capable of identifying the endogenous spatial structure of innovation clusters rather than relying solely on externally defined territorial boundaries. This consideration provides the methodological basis for applying density-based spatial approaches to capture the evolving geography of innovation and regional resilience.

3.3 Knowledge Endowment and Technological Iteration

During periods of incremental technological development, the accumulation of domain-specific knowledge tends to generate increasing returns and path-dependent dynamics (David, 1985; Arthur, 1989; Chu et al., 2026). Historical investments in a particular technological regime enhance organizational learning processes and strengthen absorptive capacity through accumulated experience (Cohen & Levinthal, 1990; Fankhauser et al., 2013). Over time, these competence-enhancing innovations become embedded within organizational routines, production systems, and institutional structures, thereby reinforcing the competitive advantages of incumbent firms and regions (Tushman & Anderson, 1986; Leonard-Barton, 1992; Klepper, 2010). Such cumulative processes may ultimately contribute to technological lock-in around an established dominant design (Unruh, 2000; Vergne & Durand, 2010; Goldstein et al., 2023). However, technological evolution is not always continuous. Periods of paradigm shift may introduce competence-destroying changes that disrupt existing technological trajectories (Tushman & Anderson, 1986). Under these conditions, previously accumulated knowledge and routines may reduce the adaptability of incumbents to emerging technologies. Existing capabilities that once supported technological leadership can gradually evolve into core rigidities, constraining the ability of firms and regions to recognize, absorb, and exploit new technological opportunities (Leonard-Barton, 1992; Unruh, 2000; Vergne & Durand, 2010). In particular, highly specialized knowledge structures may limit absorptive capacity in unfamiliar technological domains and increase the risk of technological lock-in during industrial transitions (Cohen & Levinthal, 1990; Aghion et al., 2016; Chu et al., 2026).

This mechanism is especially relevant in the transition from the internal combustion engine regime to automotive electrification (Aghion et al., 2016). Regional innovation systems characterized by substantial legacy accumulation in internal combustion technologies may face greater difficulties in adapting to the emerging electrochemical paradigm. Consequently, historical knowledge endowment may exert a negative effect on innovation performance in new energy vehicle technologies. Moreover, this effect may exhibit nonlinear characteristics, as the advantages generated by increasing returns under the old technological regime can gradually transform into structural constraints during periods of disruptive technological change (Arthur, 1989; Vergne & Durand, 2010; Goldstein et al., 2023).

Technological development is not solely a cumulative process of knowledge accumulation, but also involves the continuous renewal and replacement of existing knowledge bases. As technologies evolve, previously accumulated knowledge may gradually become obsolete, particularly in periods of rapid technological change (Bosworth, 1978; Park et al., 2006; Lee, 2024). This dynamic is especially important during technological paradigm shifts, when established technological trajectories are disrupted, and innovation systems face increasing pressures for knowledge renewal (Tushman & Anderson, 1986). However, regional innovation systems differ substantially in their ability to absorb, update, and apply new knowledge (Cohen & Levinthal, 1990; Rosiello & Maleki, 2021). Evolutionary studies emphasize that successful adaptation depends not only on the stock of existing knowledge but also on the capacity to continuously renew technological capabilities in response to changing environments (Nelson & Winter, 1982a, pp.7-18; Teece et al., 1997; Zahra & George, 2002). Regions characterized by stronger learning capabilities and more flexible innovation networks are, therefore, generally better positioned to adapt to rapidly evolving technological conditions.

This adaptive process is closely related to the concept of technology cycle time, which reflects the speed at which existing knowledge depreciates and becomes obsolete over time (Lee, 2013, pp.31-33; Park & Lee, 2015; Shi & Zhang, 2025). Empirically, it is measured using the mean backward citation lag between a focal patent and its cited predecessors (Park & Lee, 2015; Rosiello & Maleki, 2021; Lee, 2024). Short technology cycles are typically associated with faster knowledge turnover and more rapid shifts in technological frontiers (Lee, 2013, pp.32-33; Lee, 2024). Under such conditions, the advantages of incumbent regions based on accumulated experience may weaken more quickly, while the ability to access and integrate recent knowledge becomes increasingly important (Park & Lee, 2015; Shi & Zhang, 2025). For latecomer regions, short-cycle technologies may therefore create favorable conditions for technological catch-up by reducing the persistence of historical technological advantages (Park & Lee, 2015; Shi & Zhang, 2025).

During periods of technological paradigm shift, legacy knowledge accumulation and technology cycle time may influence regional catch-up through different mechanisms. Legacy knowledge accumulation refers to the historical stock of skills, capabilities, and cognitive resources embedded in preceding technological trajectories (Tushman & Anderson, 1986; Chu et al., 2026). Under radical technological change, these deeply embedded competencies may evolve into organizational rigidities that hinder adaptation to emerging technological paradigms (Leonard-Barton, 1992). Legacy knowledge accumulation primarily affects the degree of adjustment difficulty during technological transition. Regions with deeper accumulation in established technological paradigms may face greater challenges in reallocating resources, adapting organizational routines, and shifting toward new technological trajectories. Concurrently, technology cycle time influences the necessary pace of knowledge renewal and technological adaptation. Shorter technology cycles create strong imperatives for regions to rapidly access and integrate frontier knowledge. Together, these mechanisms suggest that differences in historical knowledge structures and knowledge renewal dynamics may contribute to uneven regional catch-up outcomes during periods of technological transition.

3.4 Theoretical Gaps and Hypotheses Development

A review of the existing literature reveals two important gaps in research on technological catch-up. First, most studies rely on national-level comparisons or predefined administrative regions, which may not accurately capture the spatial organization of innovation activities during periods of technological transition (Boschma, 2015; Binz & Truffer, 2017). Such approaches can overlook the endogenous structure of innovation clusters and the uneven geography of technological adaptation. Second, while previous studies have emphasized the opportunities created by electrification, fewer studies have jointly examined how legacy knowledge structures and technological iteration dynamics shape regional catch-up trajectories (Leonard-Barton, 1992; Aghion et al., 2016; Altenburg et al., 2022; Lee, 2024; Shi & Zhang, 2025).

To address these gaps, this study employs data-driven spatial clustering methods to identify innovation clusters based on actual patterns of technological concentration and interaction (Audretsch & Feldman, 1996; Binz & Truffer, 2017). Building on the theoretical arguments developed above, this study argues that regional innovation performance during the automotive electrification transition is shaped by both historical knowledge endowment and the pace of technological renewal. Legacy knowledge accumulation may constrain adaptation to emerging technological paradigms by reinforcing established routines and technological specialization (Nelson & Winter, 1982b; Leonard-Barton, 1992; Aghion et al., 2016). By contrast, shorter technology cycle times may facilitate knowledge renewal and improve adaptability under rapidly changing technological conditions (Park & Lee, 2015; Lee, 2024).

Hypothesis 1: *Legacy knowledge accumulation is negatively associated with regional electric vehicle innovation output, whereas shorter technology cycle times are associated with higher innovation output.*

Beyond innovation performance, technological catch-up also involves changes in the relative distance between latecomers and incumbents (Abramovitz, 1986; Malerba & Lee, 2021).

Latecomer regions with lower dependence on internal combustion engine knowledge may face fewer structural constraints during the electrification transition (Lee & Lim, 2001; Altenburg et al., 2022). At the same time, shorter technology cycle times may create favorable conditions for rapid knowledge renewal and technological adaptation (Lee, 2013, pp.83-84; Park & Lee, 2015).

Hypothesis 2: *For latecomer innovation clusters, lower legacy knowledge accumulation and shorter technology cycle times are associated with a reduction in technological distance from incumbent regions.*

For incumbent regions, deep specialization in established technological paradigms may reduce adaptability during periods of disruptive technological change (Tushman & Anderson, 1986; Christensen & Rosenbloom, 1995). In particular, incumbent clusters characterized by strong legacy dependence and slower technological renewal may experience greater difficulties in maintaining their technological leadership during the electrification transition (Leonard-Barton, 1992; Aghion et al., 2016).

Hypothesis 3: *For incumbent innovation clusters, higher legacy knowledge accumulation and longer technology cycle times are associated with a decline in relative technological leadership during the automotive electrification transition.*

4 Data and Methodology

This chapter describes the empirical design of the study and the construction of the patent dataset. Section 4.1 outlines the data collection process and explains how the cross-regional patent panel was assembled. Section 4.2 introduces the hierarchical density-based spatial clustering approach used to identify automotive innovation clusters and their geographic boundaries. Finally, Section 4.3 defines the main variables and presents the econometric models used to examine innovation performance and regional catch-up dynamics.

4.1 Data Source and Collection Procedure

Patents are widely used as indicators of inventive efforts, providing a consistent and measurable proxy for knowledge generation and technological advancement over time (Pavitt, 1985; Archibugi, 1992; Nagaoka et al., 2010; Lerner & Seru, 2022). This study utilizes PCT International Applications from 2000 to 2024 from mainland China and EPO countries in Europe as the primary proxy for innovation, rather than granted patents. This selection criterion is driven by three critical methodological advantages. First, PCT applications improve timeliness. Granted patents often suffer from a significant examination lag (typically 3–5 years), whereas PCT applications are published 18 months after filing. This allows the study to capture the most recent catch-up dynamics in the rapidly evolving EV sector. Second, the PCT system introduces a degree of quality selection, because international patent applications involve substantially higher filing costs and are more frequently associated with inventions targeting broader international markets. Third, it enhances comparability. By focusing on transnational filings, this study mitigates home-country bias, subjecting both Chinese and European applicants to a unified international preliminary examination standard.

The dataset is constructed by integrating two sources to ensure maximum data integrity. The first source is Google Patents Public Data. Hosted on the Google Cloud Console, this database provides extensive global bibliographic information and serves as the primary bibliographic foundation for the initial retrieval of patent records. The second source is the World

Intellectual Property Organization (WIPO) database. As the administrative body of the PCT system, the WIPO database is used as a secondary verification and imputation source. The WIPO database is specifically used to cross-reference and supplement critical bibliographic fields that are missing in the primary Google dataset. This integration process is intended to enhance the overall consistency and reliability of the sample utilized in the comparative analysis.

The dataset is constructed by aggregating PCT applications based on their publication year, rather than the filing year, covering the period from 2000 to 2024. This temporal definition helps align the analysis with the timing of publicly disclosed technological information within the observation window. The period is deliberately restricted to the end of 2024 to mitigate right truncation bias. Due to the statutory 18-month secrecy period between filing and publication, a certain portion of applications submitted in 2025 has not yet been disclosed in public databases. By setting 2024 as the cutoff based on publication dates, the study is intended to provide a more consistent representation of annual innovation outputs and to reduce the risk of an artificial downward trend in the final year of the panel.

To delineate the technological paradigm shift from the ICE vehicle technology to the emerging electrification trajectory, this study adopts a combined retrieval strategy integrating International Patent Classifications (IPC) with keyword restrictions, a methodology validated by previous studies (Borgstedt et al., 2017; Sinigaglia et al., 2022b). For the ICE regime, the identification employs a Boolean search strategy combining core subclasses (e.g., F02B, F02D) with automotive-specific keywords to reduce unrelated industrial noise, following the protocol established by Sinigaglia et al. (2022a). For electric vehicles, this study adopts the granular 'New Three Systems' standard (CNIPA, 2024). Unlike broader definitions (Schmitt et al., 2016), this framework focuses on BEV-oriented components spanning propulsion, energy storage, and vehicle control systems, such as batteries (H01M) and electric propulsion (B60L), to better capture the transition toward electrification (See Appendix B).

The structured dataset is constructed by extracting essential bibliographic fields from these selections, including filing dates, publication dates, priority dates, assignee information, and backward citation networks. To prepare the raw data for spatial analysis and enhance its geographic reliability, a systematic cleaning procedure is applied before the regional analysis. First, the WIPO database is utilized to cross-verify the records and impute any missing essential fields from the primary Google dataset. Furthermore, given the research focus on spatial innovation clusters, accurate geographic localization is important. Therefore, patents that entirely lack assignee information or are assigned exclusively to independent individuals are excluded. This step removes approximately 3.78% of the initial sample and helps mitigate potential spatial measurement errors, as individual inventors often register patents using residential addresses that may diverge from actual R&D locations. Additionally, because the initial retrieval relied on the harmonized assignee country field, the dataset inadvertently included patents originating from overseas research facilities. To maintain consistency with the research scope, patents with actual geographic coordinates outside China and Europe (EPO membership countries) are removed. This geographic filtering procedure excludes an additional 3.60% of observations and prevents the formation of clusters outside the geographic scope of the study during the subsequent clustering process. I constructed a consistent set of corporate entities by systematically standardizing the applicant names. Although the Google database provides a preliminary harmonized assignee name, it may not provide sufficient organizational consistency for the following empirical analysis. Therefore, I manually cross-referenced these initial records with the official standardized entity names provided by the WIPO database. This manual procedure reduces persistent spelling inconsistencies and improves organizational consistency when grouping subsidiary applications under their respective parent firms..

Following these procedures, the retrieval process yields a consolidated dataset of 35,202 PCT patent applications from 2000 to 2024. Geographically, this comprises 5,026 patents from Chinese applicants and 30,176 patents from European applicants. Technologically, the sample includes 15,186 patents within the ICE domain and 20,369 patents within the EV domain. Notably, 355 patents belong to both technological categories. Furthermore, the extracted backward citation network contains 87,390 cited patents for the ICE domain and 111,392 for the EV domain. This refined dataset serves as the working sample for the subsequent spatial aggregation and cluster identification.

4.2 Identification of Innovation Clusters

This section outlines the methodology used to identify and construct geographic innovation clusters from individual patent records. Section 4.2.1 describes the spatial geocoding process used to establish the coordinate locations of patent assignees. Section 4.2.2 discusses the mathematical principles of density-based clustering algorithms. Finally, Section 4.2.3 details the empirical hyperparameter calibration procedure and the criteria for establishing the final spatial sample.

4.2.1 Spatial Geocoding

The empirical foundation of this spatial analysis relies on the geographic localization of innovation actors. The name of the first assignee and the corresponding harmonized assignee country are initially extracted from the patent database, and the information serves as the primary input for the sequential spatial geocoding process. In the first phase, the Google Maps Places API is deployed to automatically query and retrieve the text addresses and spatial coordinates, including latitude and longitude for each corporate entity.

The automated retrieval process inevitably encounters missing values due to unregistered facilities or ambiguously named entities in the primary database. In this study, the data attrition is addressed through a secondary manual extraction protocol. All missing entities are cross-referenced with the official WIPO database to manually recover their registered corporate addresses. In the final phase, the Google Maps Geocoding API is utilized to process these manually supplemented text addresses and convert them into geographic coordinates. This combination of automated retrieval and manual verification helps to build a reliable geographic database to support the subsequent clustering.

4.2.2 Theoretical Principles of Spatial Clustering

Cluster analysis divides similar spatial data into meaningful groups. Identifying industrial clusters requires algorithms that can handle complex real-world geography. The development of clustering algorithms shows a continuous adaptation to different geographical features. Early methods, such as the K-Means algorithm, assign points to the nearest center. This approach assumes that all clusters are perfectly round and evenly distributed (Wu, 2012; Giordani et al., 2020). Therefore, it struggles to process isolated data points or identify the irregular industrial corridors found in modern innovation networks.

Density-based algorithms emerged to overcome these limitations. They identify highly concentrated regions separated by empty areas. The standard DBSCAN algorithm successfully finds clusters of any shape by using a fixed search radius and a minimum point requirement (Ester et al., 1996). However, economic geography naturally features both dense core cities and sparse peripheral regions. Using a single density threshold makes standard DBSCAN unsuitable for large regional datasets (Campello et al., 2013; Campello et al., 2015). Applying the same radius across different landscapes creates a major problem. A small radius correctly identifies dense urban centers but wrongly labels sparse innovation networks as meaningless noise. Conversely, a large radius wrongly combines distinct regional centers into one massive cluster.

The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm addresses these limitations by combining density-based principles with agglomerative hierarchical clustering (Campello et al., 2013; Campello et al., 2015). The algorithm highlights genuine agglomeration features by reconstructing the spatial distance matrix. The core distance is calculated for each geographic node x to quantify its local spatial density based on its proximity to the k-th nearest neighbor. Serving as an inverse proxy for the density, this metric is formally defined as:

$$core_k(x) = d(x, kNN(x)) \tag{4.1}$$

where $d(x, kNN(x))$ represents the geographic distance between node x and its k-th nearest neighbor. To reduce systematic geographic distortions across diverse latitudes, the physical distance d is calculated as the great-circle distance using the Haversine formula. The mutual reachability distance (MRD) is introduced to replace standard physical metrics. This transformed spatial metric between two geographic nodes x and y is mathematically expressed as:

$$d_{mreach}(x, y) = max\{core_k(x), core_k(y), d(x, y)\} \tag{4.2}$$

This mathematical transformation isolates sparse regional nodes by artificially expanding their topological distances while preserving the true physical proximity within dense industrial networks. The specific transformation ensures that points in sparse regions cannot easily form artificial bridges between distinct genuine clusters (Campello et al., 2013; Campello et al., 2015).

To provide a geometric intuition for the mutual reachability transformation, Figure 4.1 presents a simplified conceptual example. Node A is located within a dense spatial region and therefore exhibits a relatively small core distance, whereas Node B is situated within a sparse region and exhibits a substantially larger core distance. According to Equation (4.2), the MRD is determined by the maximum value among the direct geographic distance and the two core distances. In this example, the large core distance associated with Node B dominates the calculation and expands the final topological distance from 4.60 to 5.39. This transformation reduces the likelihood that isolated low-density observations create artificial links between dense spatial regions during the clustering process.

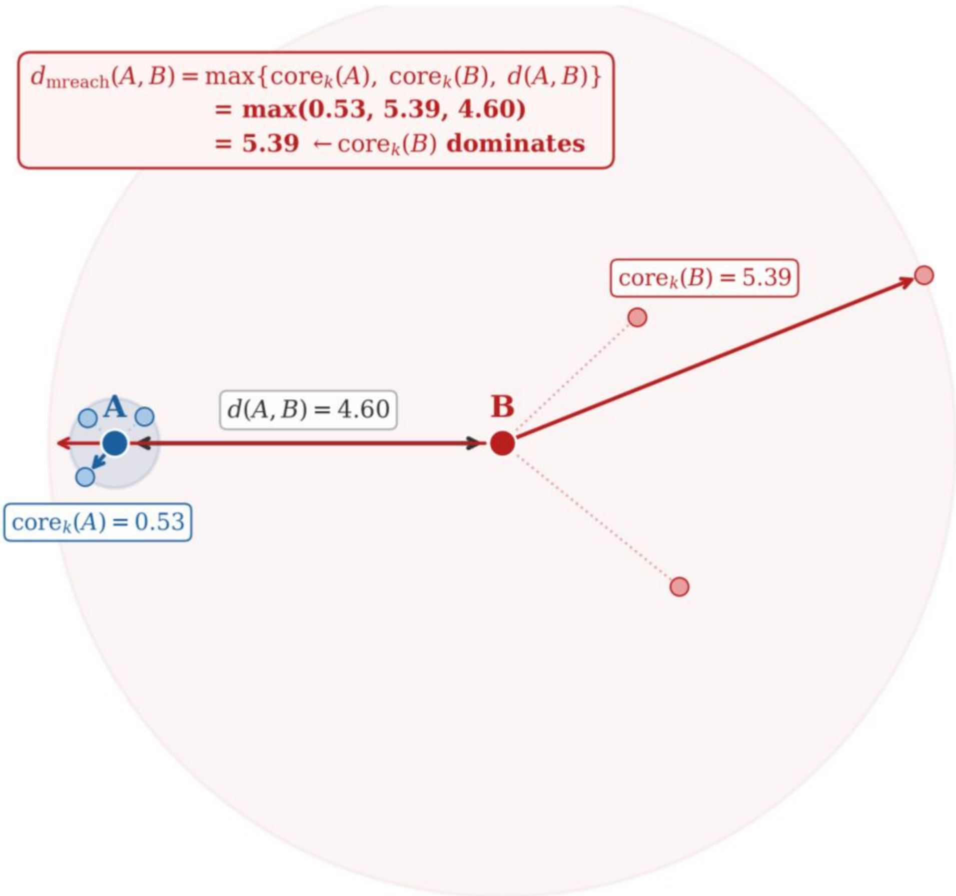

Figure 4.1: Conceptual illustration of mutual reachability distance mechanics

The algorithm utilizes this transformed distance space to construct a global minimum spanning tree and generate a clustering hierarchy. Subjective distance cutoffs are discarded in this empirical design. The algorithm calculates a persistence score for each spatial group C by integrating its structural survival across all possible density thresholds. In this framework, the density threshold λ acts as a global scanning threshold, defined as the inverse of distance. As this global threshold increases, it imposes a tightening connectivity requirement on the static MRD matrix. The stability of a cluster is quantified by summing the lifespans of all constituent nodes:

$$stability(C) = \sum_{x \in C} \lambda_{death}(x) - \lambda_{birth}(x) \tag{4.3}$$

where $\lambda_{birth}(x)$ denotes the density threshold at which the cluster C initially emerges from a broader parent agglomeration, and $\lambda_{death}(x)$ represents the threshold at which the MRD of node can no longer satisfy the density constraint, causing it to separate from the cluster as spatial noise (Campello et al., 2015; McInnes et al., 2017). The final empirical output is determined by selecting the non-overlapping cluster set that maximizes this total stability function (Campello et al., 2015; McInnes et al., 2017). Only agglomerations demonstrating structural resilience under tightening density constraints are extracted as empirically identified clusters.

To understand how this grouping process works in reality, Figure 4.2 presents a simple example based on the automotive innovation network in Bavaria. In Stage 1 of Panel A, when the density threshold λ is relatively low, spatial points remain connected even under relatively large mutual reachability distances. Because the density requirement remains relatively weak at this stage, peripheral observations like Node B can temporarily act as a structural bridge. This bridge temporarily joins the dense Munich and Ingolstadt regions together with surrounding sparse areas into one unified Bavarian parent cluster, which appears as the thick root trunk of the condensed cluster tree in Panel B. However, in Stage 2 of Panel A, as the density threshold λ increases and the connectivity requirement becomes stricter, the connection implied by the mutual reachability distance of 5.39 no longer survives under this higher density threshold. As a result, this connection disappears from the network hierarchy, and Node B falls out to be classified as spatial noise. Once this connection disappears, the original parent cluster separates into the Munich Cluster and the Ingolstadt Cluster. Finally, the algorithm determines the final partition by evaluating cluster stability, represented by the shaded areas in the tree diagram. Since the spatial points within the Munich and Ingolstadt groups remain internally connected across a wider range of density levels, these two subclusters exhibit greater structural persistence and higher stability scores than the broader parent structure. The stability optimization, therefore, favors retaining these two cities as separate final clusters instead of the broader parent structure.

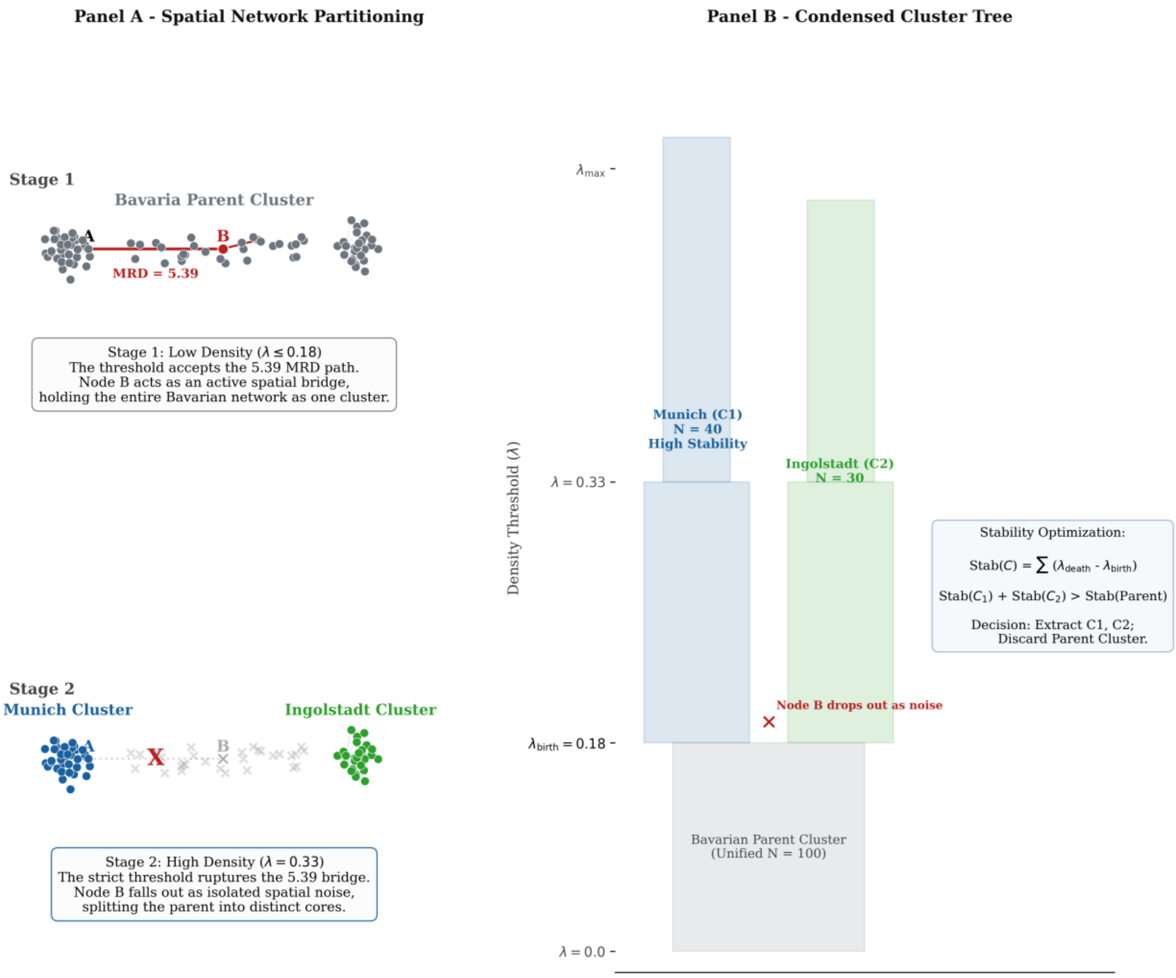

Figure 4.2: Spatial network partitioning and stability-based cluster extraction

4.2.3 Empirical Procedure of Spatial Clustering

The empirical calibration of spatial hyperparameters requires balancing economic geography considerations with machine learning principles. To maintain cross-regional comparability between the Chinese and European samples, this study rejects relative parameter scaling based on divergent sample sizes. An absolute physical baseline is established, drawing upon the regional cluster definition utilized by the United Kingdom government, which mandates a minimum of 10 independent enterprises. This policy standard serves as the theoretical lower bound for the minimum cluster size parameter (DSIT, 2025).

Building upon this theoretical foundation, a constrained grid search optimization is executed. The minimum cluster size is iteratively tested across an absolute numerical range from 10 to 60. This interval is designed to encompass the conventional machine learning heuristic of targeting cluster sizes corresponding to approximately one to two percent of the sample while preserving cross-regional consistency. The core point neighborhood parameter is simultaneously tested using a baseline of half the minimum cluster size alongside an allowable variance of plus or minus two. This grid search systematically evaluates the parameter space for both regions under an identical calibration framework.

The final parameter selection is governed by a dual-criteria optimization framework. The first criterion requires the spatial noise ratio to remain below 30 percent to avoid excessive exclusion of geographically dispersed innovation activity (Brenndoerfer, 2025). The second criterion selects the parameter configuration that maximizes the Density-Based Clustering Validation (DBCV) index within the admissible noise threshold (Moulavi et al., 2014). Following this calibration procedure, the selected parameter configuration sets the minimum cluster size at 14 and the minimum samples parameter at 5. The cluster selection epsilon is set to zero, allowing cluster boundaries to be determined solely by the stability optimization without post-hoc merging. This calibration strategy combines institutional benchmarks with internal clustering validation to improve the robustness and cross-regional consistency of spatial cluster identification.

The empirical distribution of absolute innovation capabilities across the firm sample exhibits an extremely right-skewed distribution. The bottom 99 percent of firms account for only a limited share of total patent output, while the output exhibits an extreme concentration among the top one percent. Although the calibrated parameter configuration improves the stability of the clustering structure, the HDBSCAN algorithm relies strictly on geographic coordinates and local spatial density. It identifies clusters by isolating high-density regions separated by low-density valleys. Consequently, this density-based approach could naturally exclude some high-output enterprises from the identified clusters.

This exclusion suggests that some superstar firms do not operate within locally dense innovation agglomerations but instead remain positioned within relatively sparse spatial density valleys away from other innovative entities. Given the considerable financial thresholds of PCT international applications, these firms do not necessarily require a dense local network of other PCT-filing peers to sustain their innovation output. One possible interpretation is that these firms rely more heavily on internal research capacity and broader interregional knowledge networks than on localized PCT-intensive agglomerations. Consequently, these firms are not retained within the baseline density-based cluster sample used in this study. Nevertheless, these firms remain technologically significant actors within the broader regional innovation landscape.

4.3 Econometric Specification

This section outlines the econometric approach used to test the research hypotheses. Section 4.3.1 introduces the variables used in the analysis. Section 4.3.2 explains the use of Poisson Pseudo-Maximum Likelihood models with high-dimensional fixed effects to examine innovation output. Finally, Section 4.3.3 presents the panel regression models estimated using Ordinary Least Squares to analyze regional catch-up dynamics.

4.3.1 Definition of the Variables

This subsection explains how the main variables used in the empirical analysis are constructed. For clarity, the variables are grouped into dependent, independent, and control variables.

(1) Dependent Variables

EV innovation output (EVOUT). This variable quantifies the total number of pure EV PCT patent publications generated by a cluster within a specific year. Aggregating patent counts at the geographic cluster level provides a representative measure of the cluster's collective innovation capacity (Krammer, 2009; Wu, 2011). Consequently, this variable captures the technological dynamic of the cluster and serves as an indicator of the cluster's overall technological performance and innovation output in the pure EV domain.

Distance to the technological frontier (DTF). This variable is designed for the Chinese sample to measure its capacity to close the gap with the most advanced European benchmarks. The catch-up literature generally views the reduction of technological gaps as a central component of latecomer development (Abramovitz, 1986; Fagerberg, 1994; Howitt, 2000; Wu, 2011; Pece et al., 2015). Absolute gap measures may overstate differences between clusters with very different historical accumulation (Delgado et al., 2010; Boschma, 2015). To address this issue, I first classify all Chinese and European clusters into three scale tiers based on their cumulative total number of companies, then apply the 40th and 90th percentiles of this cumulative distribution as thresholds to define small, medium, and large tiers. This stratification helps improve the comparability of clusters by restricting comparisons to competitors with broadly similar organizational scale. DTF is defined as the logarithmic difference between a focal Chinese cluster and the top-performing European cluster within the same scale tier, which aligns with established distance-to-frontier metrics widely utilized in the endogenous growth and catch-up literature (Aghion et al., 2005; Acemoglu et al., 2006; Lee & Malerba, 2017; Shi et al., 2026). This calculation is formally expressed in Equation (4.4).

$$DTF_{i,t} = \ln\left(EUFrontier_{c,t} + 1\right) - \ln(EVOUT_{i,t} + 1) \tag{4.4}$$

In this equation $EVOUT_{i,t}$ represents the absolute EV patent innovation output of Chinese cluster i in year i. $EUFrontier_{c,t}$ denotes the maximum patent output achieved by any European cluster belonging to the same scale tier c during the same year. A smaller DTF value indicates that the focal Chinese cluster is closer to the European technological frontier, which suggests a more efficient technological catch-up process.

Incumbent technological advantage (ITA). This variable is applied to the European sample to evaluate its capacity to maintain leadership against emerging Chinese competitors. Theoretically, established incumbents defend their market dominance by employing historical scale to create entry barriers and capitalizing on their accumulated knowledge base through continuous incremental innovations (Lieberman & Montgomery, 1988; Acemoglu & Cao, 2015). Building upon the scale-matching logic introduced for DTF, this metric calculates the logarithmic disparity of a European cluster over its most productive Chinese peer within the same scale tier. The calculation is formulated as Equation (4.5).

$$ITA_{j,t} = \ln\left(EVOUT_{j,t} + 1\right) - \ln(CNFrontier_{c,t} + 1) \tag{4.5}$$

Here $EVOUT_{j,t}$ represents the absolute patent output of European cluster j. $CNFrontier_{c,t}$ is the highest output among all Chinese clusters within the corresponding scale tier c. A larger ITA value indicates that the European incumbent successfully maintains a wider technological lead over its Chinese rivals, reflecting a stronger resilience against catch-up pressures. DTF and ITA are constructed as symmetric indicators. Together, the two indicators capture the catch-up relationship from both the Chinese and European perspectives, providing a comprehensive view of the bilateral catch-up dynamics.

The operationalization of both DTF and ITA uses a $\ln(x + 1)$ transformation. This choice was primarily motivated by the presence of zero-value observations in the EV patent data. During the early transition period, several clusters could record no EV patents at all. These zero observations are substantively meaningful, as they reflect regions that remained fully specialized in ICE technologies rather than missing data. Removing them would bias the sample toward clusters that had already entered the electrification trajectory.

In addition, the logarithmic specification is more consistent with the conceptual interpretation of technological distance in the literature (Aghion et al., 2005; Acemoglu et al., 2006). Catch-up dynamics are better understood in relative rather than absolute terms. Since the difference between logarithms can be interpreted as the logarithm of a ratio, the transformation captures proportional technological distance while reducing the influence of regional size differences.

A further practical consideration concerns the distribution of patent counts. Patent data are typically highly skewed and characterized by extreme outliers. Using absolute patent differences would therefore produce highly dispersed values dominated by a small number of very large clusters. The logarithmic transformation helps compress the upper tail of the distribution and improves the stability of the panel estimations.

At the same time, adding the constant term inevitably affects the distribution of observations near zero and may introduce a small degree of attenuation bias. However, given the importance of retaining zero-value observations during the technological transition process, the $\ln(x + 1)$ specification was considered a reasonable empirical adjustment for maintaining a consistent log-linear framework.

(2) Independent Variables

Technology cycle time (TCT). For each automotive PCT patent, its cited prior arts are identified, and the time difference between the focal patent's filing date and the publication date of each cited patent is calculated. This temporal difference, initially measured in days, is converted into years by dividing by 365.25 to represent the technology cycle time for a specific reference. The average technology cycle time is computed for the entire automotive patent portfolio of a specific geographic cluster within a given year. It serves as an indicator of a cluster's overall knowledge recency and technological iteration speed across its entire R&D base. A shorter average technology cycle reflects a proactive approach to incorporating recent advancements into R&D activities, thus demonstrating faster knowledge renewal and agile innovation capability (OECD, 2009; Lee, 2013, pp.83-84; Lee, 2024). Conversely, a longer average technology cycle suggests a reliance on older, established technologies, highlighting an emphasis on incremental innovations or a slower integration of cutting-edge developments.

Legacy knowledge accumulation (LKA). This indicator quantifies a cluster's structural reliance on the traditional ICE paradigm. To map the dynamic evolution of the regional knowledge structure, this study constructs the cumulative patent stock using the Perpetual Inventory Method (PIM) applied to the observation period from 2000 to 2024. Because technical knowledge gradually becomes obsolete and loses its economic relevance over time, a constant annual knowledge depreciation rate of 15 percent is applied to the patent stock (Hall et al., 2000; Noailly & Smeets, 2015). The knowledge stock for a specific technology domain is recursively calculated based on annual patent flows. Equation (4.6) defines the accumulation function.

$$S_{i,t}^{k} = F_{i,t}^{k} + S_{i,t-\tau}^{k}(1-\sigma)^{\tau} \tag{4.6}$$

In this equation, $S_{i,t}^{k}$ represents the accumulated knowledge stock of cluster i in year t for technology domain k, which represents either pure ICE patents or total automotive patents. $F_{i,t}^{k}$ denotes the absolute annual patent flow. The parameter σ is the 15 percent knowledge depreciation rate. The variable τ represents the time gap between the current observation and the previous record, which accounts for intermittent patenting activities and applies compound depreciation appropriately. To mitigate potential endogeneity and contemporaneous bias, the LKA ratio incorporates a one-year lag. It is computed as the lagged stock of ICE patents divided by the lagged total automotive patent stock and shown in Equation (4.7)

$$LKA_{i,t} = \frac{S_{i,t-1}^{ICE}}{S_{i,t-1}^{Total}} \tag{4.7}$$

In the innovation economics literature, the relative composition of accumulated patent stocks is a well-established empirical approach for capturing technological trajectories and measuring path dependence (Patel & Pavitt, 1997; Noailly & Smeets, 2015; Aghion et al., 2016). A higher LKA value indicates that a cluster's cognitive and technical base remains disproportionately anchored in the ICE paradigm. A high concentration of legacy knowledge is associated with greater structural inertia arising from path-dependent specialization (Nelson & Winter, 1982b; Leonard-Barton, 1992; Dijk et al., 2013; Cecere et al., 2014). Such a knowledge structure may increase the difficulty of reallocating innovative resources toward the emerging electric vehicle paradigm and contribute to lock-in effects within the regional innovation ecosystem (Arthur, 1989; Unruh, 2000).

(3) Control Variables

Self-citation rate (SELFCITE). This variable captures the role of a cluster's prior innovations in driving its current technological progress. It is defined as the proportion of a cluster's total annual automotive patent citations that reference prior patents originating from within the same cluster. A higher self-citation rate indicates a robust localized knowledge base and a strong capacity for continuous innovation (Hall et al., 2005). This phenomenon typically reflects a cumulative innovation process, wherein the regional ecosystem iteratively refines and expands upon its established technologies within specific domains (OECD, 2009). By contrast, a lower self-citation rate signifies a heavier reliance on external knowledge spillovers, potentially indicating an outward-looking technological strategy aimed at integrating diverse, extra-regional technologies.

Technological originality (TECHORIG). This variable measures the extent to which a regional cluster draws upon knowledge from diverse technological domains. It is constructed using the complement of the Herfindahl-Hirschman Index ($1 - HHI$) based on the distribution of patent citations across seven-digit IPC technological classes. It is first calculated at the patent level and subsequently averaged across all patents within a cluster-year observation. Higher values indicate that cited knowledge is distributed across a broader range of technological fields, reflecting a stronger capability for recombinative innovation across domain boundaries (Trajtenberg et al., 1997; OECD, 2009). Cross-domain citation diversity serves as a proxy for originality, as genuinely novel inventions tend to recombine knowledge from previously unconnected technical areas rather than refining solutions within a single domain (Trajtenberg et al., 1997). Lower values suggest a concentration of knowledge sourcing within a narrower set of technological domains, reflecting a trajectory of incremental innovation.

Average inventor team size (INVENTORS). This variable captures the micro-level scale of collaborative human capital mobilized in the knowledge creation process. It is measured as the average number of inventors listed per patent application within a geographic cluster in a given year. As technological complexity increases, innovation has progressively shifted from individual effort toward coordinated, team-based knowledge recombination (Wuchty et al., 2007; Jones, 2009). This tendency is particularly pronounced in the electric vehicle paradigm, which requires the convergence of diverse expertise across electrochemistry, power electronics, embedded software, and mechanical engineering. Econometrically, this variable is introduced to control for time-varying shifts in regional R&D labor deployment.

University citation rate (UNIVCITE). This variable captures the extent to which a cluster's innovation activities rely on academic and public research resources. It is defined as the proportion of a cluster's total annual patent citations that reference patents granted by universities or public research centers. A higher university citation rate indicates a close interaction between the regional industry and academia, reflecting the role of academic knowledge in supporting technological adaptation and innovation (Trajtenberg et al., 1997; Corrocher et al., 2007).

Controlling for university linkages is empirically important because public research institutions often function as key sources of scientific knowledge and technological spillovers during periods of technological transition. Excluding this dimension could bias the estimated effects of the core explanatory variables by omitting localized knowledge-network effects.

One limitation of this measure concerns its application to the European sample. In several European countries, academic inventions have historically been patented under firm ownership or individual inventor ownership rather than university ownership due to institutional arrangements associated with the "Professor's Privilege" system (Lissoni et al., 2008; Ejermo & Källström, 2016). As a result, the variable likely underestimates the true contribution of universities to regional innovation activity in Europe. Nevertheless, the measure still captures an observable lower bound of university-industry knowledge linkages and therefore remains informative for comparing regional differences in academic engagement.

Organizational knowledge variety (OKV). This variable quantifies the dispersion of a cluster's knowledge sources across different originating entities, such as firms, universities, or public research centers. It is constructed using the complement of the Herfindahl-Hirschman Index based on patent citation shares across different organizations, including firms, universities, and public research institutions. The index is calculated at the patent level and then averaged across all patents within a cluster-year observation. Higher values indicate that knowledge citations are distributed across a wider range of organizations, reflecting a more diversified knowledge base and broader technological search process (Corrocher et al., 2007). Conversely, a lower value suggests a higher concentration of citations on a few key dominant players.

Controlling for organizational knowledge variety is empirically important because regions with more diverse knowledge networks may possess stronger adaptive capacity during periods of technological transition. The limitation is also that the institutional differences associated with the Professor's Privilege system may introduce some measurement distortion in the European sample. This distortion may affect the observed diversity of organizational citation sources to some extent. However, because academic patents constitute a relatively small component of the broader regional citation network in the context, the resulting bias is expected to remain relatively limited at the aggregate cluster level.

4.3.2 Estimation Model for Innovation Output

In regional innovation empirical studies, the EV innovation output presents a typical non-negative integer count characteristic. This output is accompanied by an extremely right-skewed distribution and severe overdispersion. Conventionally, count data is analyzed using the Poisson regression model. Under the Poisson assumption, the dependent variable follows a Poisson distribution. The expected patent count for cluster i at year t can be expressed as Equation (4.8).

$$E(Y_{i,t}|X_{i,t}) = \mu_{i,t} = exp(X_{i,t}\beta) \tag{4.8}$$

In this equation, $Y_{i,t}$ represents the dependent variable for cluster i at year t, $X_{i,t}$ represents the independent variables, and β is the vector of coefficients. However, Poisson models impose a strict assumption that the conditional mean equals the conditional variance, which is $Var(Y_{i,t}|X_{i,t}) = \mu_{i,t}$. But in the context of regional innovation activities, patent outcomes are highly skewed and exhibit significant overdispersion, where the conditional variance substantially exceeds the conditional mean. Overdispersion violates the fundamental assumption of the Poisson model, leading to severely underestimated standard errors and biased model fit.

To address the overdispersion issue, empirical studies generally introduce an additional dispersion parameter and employ the Negative Binomial Regression (NBR) model (Hausman et al., 1984). Furthermore, regional panel data analysis requires the model to strictly control for time-invariant, unobservable cluster-specific characteristics, such as geographic location, industrial foundations, and long-term policy environments. However, the conditional fixed-effects NBR model reveals a critical flaw when applied to panel data. This model does not genuinely control for all stable covariates because its individual heterogeneity is absorbed through the dispersion parameter rather than the conditional mean (Allison & Waterman, 2002; Guimarães, 2008). Consequently, relying on the traditional conditional fixed-effects NBR model to process panel fixed effects ultimately yields severe omitted variable bias.

To overcome these theoretical flaws and obtain robust estimates, this study adopts the Poisson Pseudo-Maximum Likelihood (PPML) estimator combined with High-Dimensional Fixed Effects (HDFE) for the EV innovation output. As a pseudo-maximum likelihood approach, PPML does not impose strict distributional assumptions on the underlying data. As long as the conditional mean is correctly specified, the estimator provides consistent coefficient estimates even if the true data-generating process diverges from a Poisson distribution (Santos Silva & Tenreyro, 2006). Furthermore, traditional nonlinear panel models frequently suffer from the incidental parameter problem when incorporating high-dimensional individual and time fixed effects. The PPML-HDFE algorithm overcomes this limitation by absorbing multi-dimensional unobservable heterogeneity (Correia et al., 2020).

Building upon the aforementioned theoretical foundation, this study constructs the baseline empirical model for the EV innovation output. The specific empirical specification is presented in Equation (4.9).

$$\begin{aligned} E\left(EVOUT_{i,t} \middle| X_{i,t}, \mu_i, \delta_t\right) \\ = \exp\left(\beta_0 + \beta_1 TCT_{i,t} + \beta_2 LKA_{i,t} + \gamma Control_{i,t} + \mu_i + \delta_t\right) \end{aligned} \tag{4.9}$$

In this equation, the subscript i denotes the specific geographic cluster, and t denotes the observation year. The dependent variable $EVOUT_{i,t}$ represents the EV innovation output of cluster i in year t. $TCT_{i,t}$ and $LKA_{i,t}$ represents the technology cycle time and the legacy knowledge accumulation for cluster i in year t, respectively. $Control_{i,t}$ is the vector of control variables for a given cluster-year observation. The parameter μ_i captures cluster-specific fixed effects. The parameter δ_t captures year fixed effects. To address potential heteroskedasticity and residual overdispersion interference in the patent data, this model strictly applies cluster-robust standard errors across all PPML regressions.

4.3.3 Estimation Model for Relative Catch-Up Dynamics

The DTF and the ITA measure the relative catch-up dynamics. By definition, these two indicators are calculated as logarithmic differences. This mathematical transformation not only converts them into continuous variables but also alters their statistical distribution. Unlike the raw patent counts that exhibit severe right-skewness and overdispersion, the logarithmic transformation compresses the scale of the data and mitigates the strong right-skewness in the distribution. Because the underlying data-generating process no longer conforms to a skewed count distribution, nonlinear count models become structurally inappropriate. For these continuous and relatively symmetrically distributed variables in panel data, employing Ordinary Least Squares combined with High-Dimensional Fixed Effects (OLS-HDFE) provides a more accurate and straightforward interpretation of marginal effects. The specific linear panel model specification is presented in Equation (4.10).

$$Gap_{i,t} = \alpha_0 + \alpha_1 TCT_{i,t} + \alpha_2 LKA_{i,t} + \theta Control_{i,t} + \mu_i + \delta_t + \epsilon_{i,t} \tag{4.10}$$

In Equation (4.10), the subscript i refers to the geographic cluster, and t refers to the observation year. The dependent variable $Gap_{i,t}$ acts as a unified indicator for the relative technological gap. Depending on the specific empirical estimation, this variable represents either the DTF measured within the Chinese cluster sub-sample or the ITA measured within the European cluster sub-sample. All explanatory variables and unobservable fixed effects remain identical to the baseline model. The term $\epsilon_{i,t}$ represents the idiosyncratic error term. This linear model also employs cluster-robust standard errors to improve the reliability and robustness of the statistical inference.

5 Result and Analysis

This chapter presents the empirical results of the study. Section 5.1 reports the spatial clustering results generated by the HDBSCAN algorithm and maps the geographic distribution of automotive innovation clusters. Section 5.2 provides a descriptive overview of the Chinese and European panel datasets. Section 5.3 presents the Pearson correlation matrix together with diagnostic checks for multicollinearity. Section 5.4 discusses the baseline regression results for innovation performance and regional catch-up dynamics. Finally, Section 5.5 presents a series of robustness checks to examine the stability of the main findings.

5.1 Clustering Results

Following the application of the HDBSCAN algorithm, the spatial noise ratio is strictly constrained below the 30 percent threshold. The empirical procedure ultimately identifies 22 distinct automotive clusters in China comprising 2,988 patents. Concurrently, the European sample yields 57 clusters encompassing 25145 patents.

As shown in Figure 5.1, the spatial pattern of automotive innovation in China demonstrates significant regional heterogeneity. High-density clusters are predominantly concentrated along the eastern coastal regions, particularly within the Pearl River Delta and the Yangtze River Delta. These coastal ecosystems exhibit a highly dense spatial distribution where multiple distinct innovation units operate in close geographic proximity. Conversely, the vast inland areas are largely classified as spatial noise. The identified inland clusters exhibit weaker spatial continuity and remain geographically separated from one another as isolated regional innovation centers.

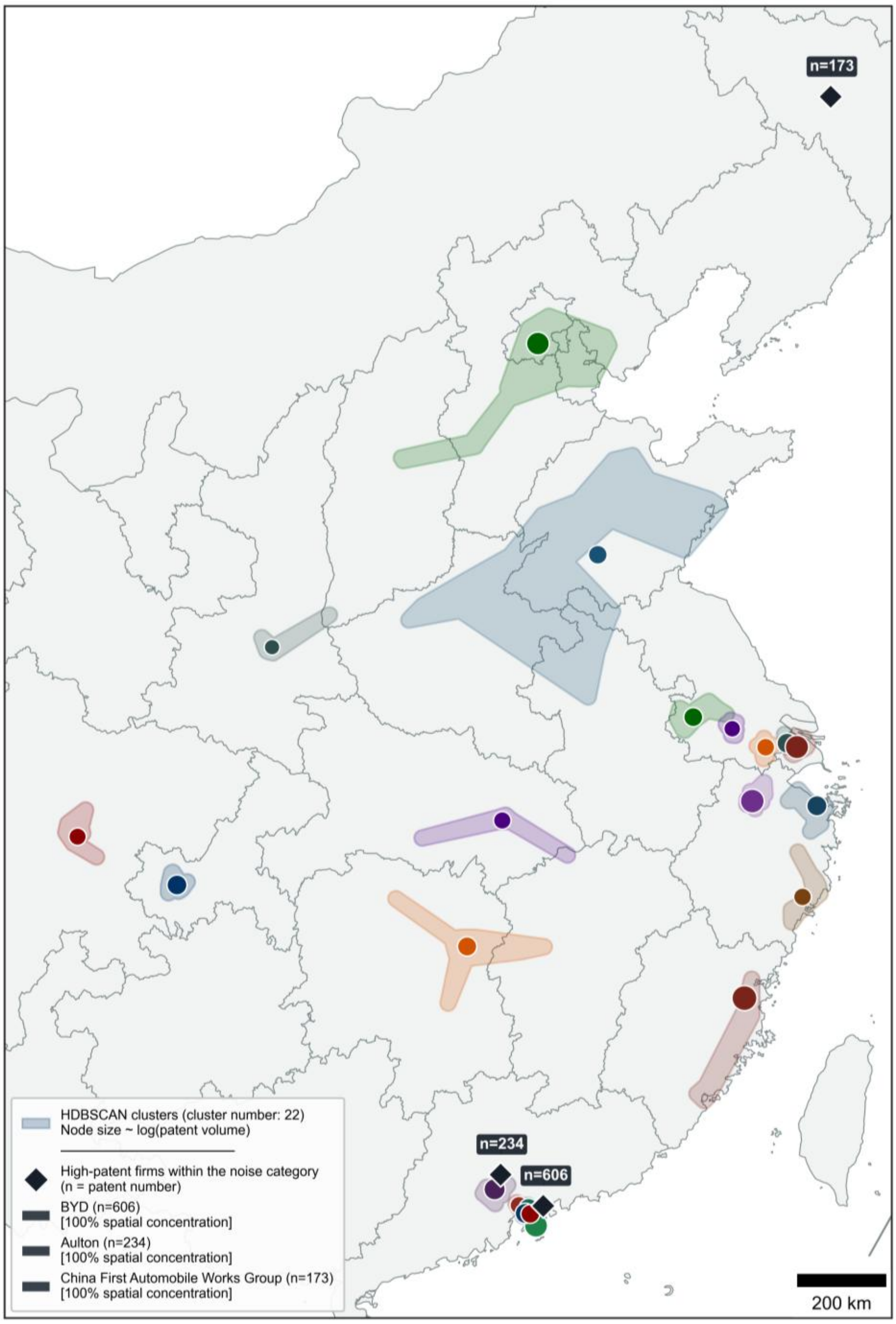

Figure 5.1: Automotive innovation cluster distribution in China

Note: Minor visual overlap between certain clusters is a graphical rendering artifact generated during polygon smoothing in the mapping process. The HDBSCAN procedure assigns each spatial coordinate exclusively to a single cluster or to the noise category. This note applies to all subsequent spatial cluster visualizations in this chapter.

As shown in Figure 5.2, the spatial distribution of automotive innovation clusters in Europe exhibits strong geographic concentration and substantial spatial continuity. High-density clusters are predominantly concentrated across Central and Western Europe. These spatial units remain in close geographic proximity and collectively form a broadly continuous agglomeration belt that frequently transcends national borders. The visual results indicate that

European automotive innovation activities display a relatively dense and geographically contiguous spatial structure rather than a highly dispersed regional pattern.

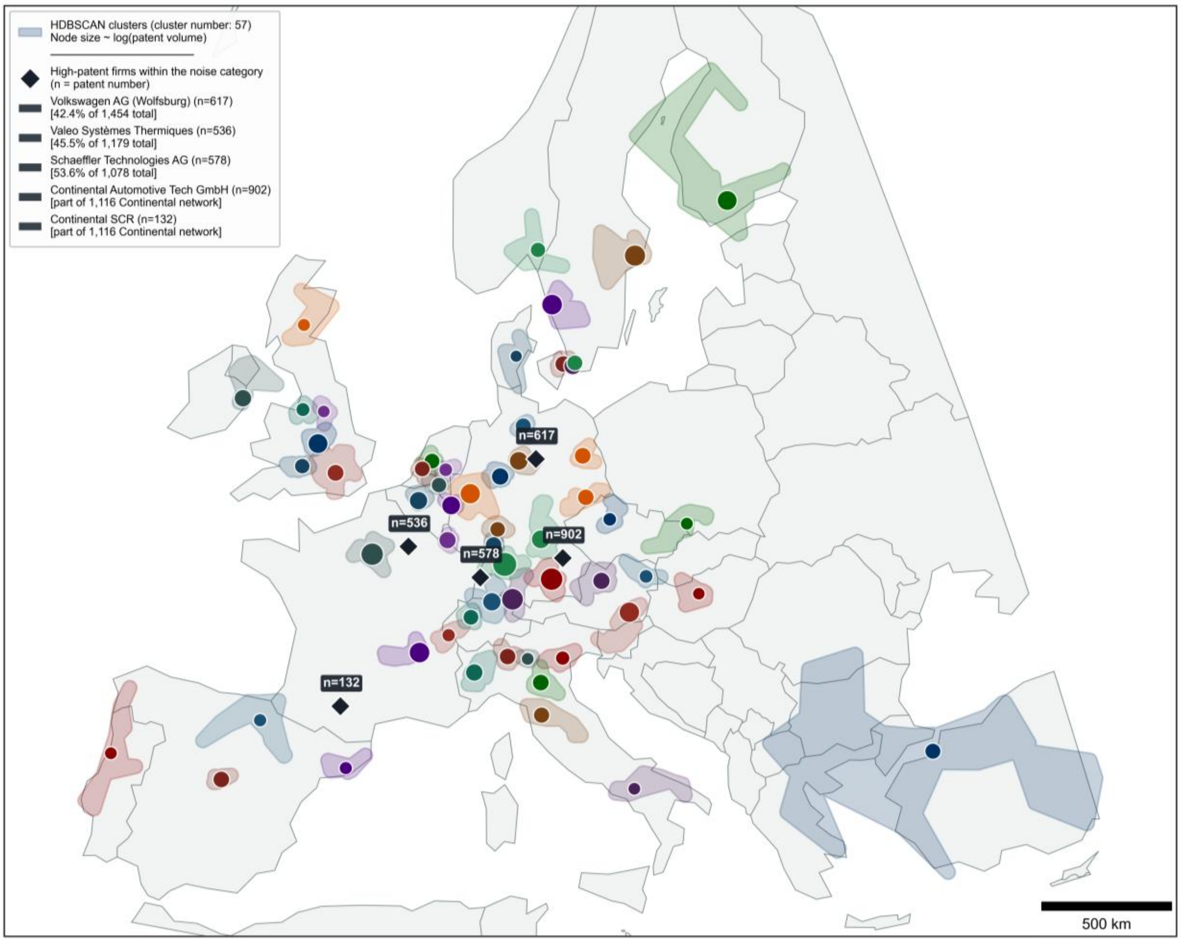

Figure 5.2: Automotive innovation cluster distribution in Europe

Building upon the physical spatial identification of these innovation networks, Figure 5.3 projects these localized spatial units into a two-dimensional technological feature space. This projection allows for an objective examination of their evolutionary characteristics. The horizontal axis measures the aggregated technology cycle time at the cluster level. The vertical axis represents the end of period legacy knowledge accumulation ratio. The area of each bubble corresponds to the EV patent output of the respective cluster. The empirical mapping reveals a pronounced structural divergence between the two regions.

The blue bubbles representing European clusters are predominantly concentrated in the upper right quadrant of the feature space. The data indicate that these highly interconnected European networks possess significant historical technological endowments and extended technology cycle periods. I identify the strongest technological cores in the European sample, including the Stuttgart-Karlsruhe cluster, the Munich-Ingolstadt cluster, and the Paris-Île-de-

France cluster. These highly productive spatial units are anchored in the core areas of Western and Central Europe. This geographic distribution illustrates that the traditional manufacturing heartland continues to function as a major center of automotive innovation activity during the transition period.

Conversely, the red bubbles representing Chinese clusters are densely aggregated in the lower left quadrant. The data show that these spatially fragmented Chinese networks exhibit minimal legacy knowledge accumulation and highly compressed technology cycle times. We extract the most prominent technological cores in China, including the Ningde-Xiamen cluster, the Beijing-Jingjinji cluster, and the Hangzhou cluster. The spatial coordinates of these top-tier innovation networks reveal a strict concentration along the eastern coastal regions. This specific geographic positioning suggests that the eastern coastal belt currently constitutes the primary geographic center of Chinese EV innovation activity.

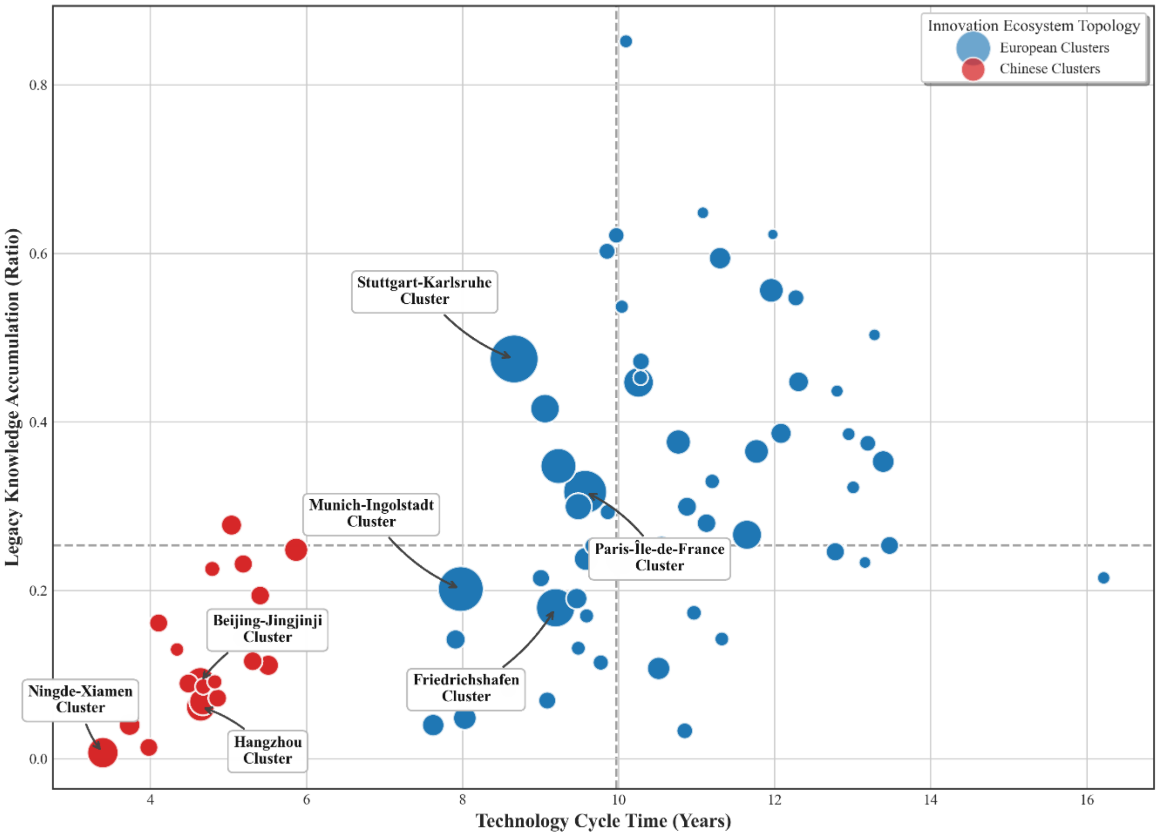

Figure 5.3: Two-dimensional technological feature space of Chinese and European automotive innovation clusters

To further dissect the micro-foundations of the identified regional clusters, Table 5.1 identifies the top 10 innovation contributors within the Chinese and European samples. A comparative analysis of these leading entities reveals four critical characteristics regarding the spatial boundaries scale, knowledge base, and organizational typology of the anchor firms driving regional innovation.

First, an analysis of the spatial noise reveals a structural divergence. This divergence is observed within the top 1% of geographic coordinates measured by patent volume in the noise pool. In the Chinese sample, three of the leading innovation contributors operate as spatial noise. Specifically, BYD possesses 606 patents, Aulton possesses 234 patents, and China First Automobile Works Group possesses 173 patents. Within the observed PCT spatial network, the patenting activities of these firms are entirely concentrated at a single geographic coordinate. By contrast, the high-output noise coordinates in Europe represent only part of broader corporate innovation networks. For example, Volkswagen AG in Wolfsburg registers 617 noise patents, accounting for only 42.4 percent of its 1454 total regional patents. Similar partial distribution patterns are repeated across other European giants. Valeo Systemes Thermiques holds 536 noise patents out of 1179 total patents for Valeo. Schaeffler Technologies AG accounts for 578 noise patents out of 1078 total patents for Schaeffler. Continental displays a clear enclave structure. Continental Automotive Tech GmbH has 902 patents, and Continental SCR has 132 patents, out of 1116 total regional patents for Continental. The evidence suggests that, within the noise pool, major Chinese firms appear substantially more spatially concentrated within the observed PCT innovation network, whereas European incumbents exhibit more geographically distributed innovation structures.

Second, the data show a disparity in the absolute scale of the representative firms within the clusters between the two regions. After excluding spatial noise, the core nodes supporting European clusters are vastly larger in patent volume than those in China. For instance, the top contributor within the European clusters, Robert Bosch, generated 5381 total patents. This scale exceeds the 441 patents generated by CATL, which represents the leading entity within the Chinese clusters. Despite the massive difference in absolute volume, both regions exhibit a high level of innovation concentration. The top 10 entities in China and Europe account for 45.42% and 53.23% of their respective total outputs. It suggests that regional innovation capabilities are not evenly distributed but are dominated by a highly concentrated group of apex organizations.

Third, the comparative data highlight a divergence in legacy knowledge accumulation among these top contributors. In this context, LKA is measured as the proportion of ICE patents relative to the total automotive patent output of the firm. The output-weighted average LKA for the European top 10 is 0.525, which is substantially higher than the 0.048 for the Chinese top 10. The higher proportion reflects a greater accumulation of historical knowledge in the traditional ICE paradigm. European clusters are likely anchored by a profound traditional technological base. Meanwhile, Chinese clusters operate with a knowledge structure that exhibits substantially weaker dependence on ICE technological trajectories.

Fourth, the composition of Table 5.1 points to a notable divergence in the organizational ecology of the regional clusters. The European innovation network is predominantly composed of established incumbents. suggests that European clusters possess deep technological heritage alongside potential institutional inertia. Such an observation aligns with traditional path-dependent innovation trajectories. In contrast, the top contributors within the Chinese clusters display a different organizational composition. This ecosystem features a diverse group of new entrants, including specialized leaders like CATL, startups like NIO, and cross-boundary technology entities like Huawei. This organizational composition is broadly consistent with the technological window of opportunity framework associated with electrification transitions. The structural absence of an old knowledge base among these entities is consistent with a dynamic of strategic agility. Such an organizational ecology plausibly facilitates the reallocation of resources toward the EV paradigm.

Table 5.1: Top innovation contributors in China and Europe

	China			Europe		
Rank	Name	Output	LKA	Name	Output	LKA
1	BYD	606	0.023	ROBERT BOSCH	5381	0.707
2	CONTEMPORARY AMPEREX TECHNOLOGY	441	0	VOLKSWAGEN	1454	0.369
3	HUAWEI	307	0.016	RENAULT	1317	0.440
4	AULTON	234	0	MERCEDES BENZ	1263	0.481
5	CHINA FIRST AUTOMOTIVE WORKS GROUP	173	0.069	BAYERISCHE MOTORENWERKE	1239	0.274
6	GEELY	146	0.151	VALEO	1179	0.303
6	NIO	146	0	CONTINENTAL	1116	0.722
8	GREAT WALL MOTOR	140	0.221	SCHAEFFLER	1078	0.350
9	GUANGZHOU AUTOMOBILE GROUP	56	0.446	STELLANTIS	1078	0.445
10	SCHAEFFLER (China)	34	0.029	SIEMENS	958	0.570
Total	Top 10 in China	2283	0.048	Top 10 in Europe	16063	0.525
	CN CR10	45.42%		EU CR10	53.23%	

5.2 Descriptive Analysis

Table 5.2 presents the descriptive statistics for all relevant variables evaluated at the cluster-year level. Both the Chinese and European clusters exhibit highly right-skewed distributions of electric vehicle innovation output (EVOUT). The standard deviations wildly exceed the corresponding means in both subsamples, and the medians remain extremely low compared to the absolute maximum values. This highly skewed distribution strongly suggests the presence of substantial overdispersion in regional innovation outputs. Such a statistical property motivates the use of the PPML estimator in the subsequent baseline regressions. Furthermore, the data reveal the structural reality of the bilateral catch-up dynamics through the continuous dependent variables. The distance to the technological frontier (DTF) for Chinese clusters reports a mean of 1.658, which indicates a persistent but quantifiable gap relative to the top European benchmarks. Conversely, the incumbent technological advantage (ITA) for European clusters shows a negative mean of -0.354 and a median of 0. This distribution suggests that, apart from a limited number of leading incumbents, many European clusters no longer maintain a substantial innovation output advantage relative to Chinese competitors within the same scale tier. Regarding the core explanatory variables, the two regional subsets reveal divergent underlying technological structures. The European clusters report a mean legacy knowledge accumulation (LKA) of 0.521 alongside an average technology cycle time (TCT) of 11.155 years. This data distribution reflects a higher concentration of traditional internal combustion engine technologies within their knowledge stock and corresponds to a longer technological iteration cycle. In contrast, the Chinese clusters demonstrate a lighter historical endowment with an average LKA of 0.281 and a shorter technology cycle time of 5.409 years. This statistical divergence reflects substantial differences in the underlying technological structures of the two regional innovation systems. Furthermore, the control variables reveal compelling organizational differences. Chinese clusters exhibit a higher observable university citation rate (UNIVCITE) than the European sample. Nevertheless, this contrast should be interpreted cautiously because institutional differences associated with the Professor's Privilege system may mechanically understate the observable contribution of universities within several European patent systems. Additionally, the average inventor team size (INVENTORS) in China stands at 3.470, which is larger than the European average of 2.246. This difference is consistent with a more collaborative and team-oriented innovation structure during the Chinese catch-up process. Finally, all remaining control variables fall within expected economic ranges and exhibit no statistical anomalies that would threaten the empirical estimations.

Table 5.2: Descriptive analysis

	N	Mean	SD	Min	Max	Median
Panel A: Chinese Clusters						
EVOUT	270	9.819	21.544	0	225	3
DTF	270	1.658	1.247	-2.572	4.754	1.749
TCT	270	5.409	2.740	0.339	21.682	4.852
LKA	270	0.281	0.275	0	1	0.208
SELFCITE	270	0.085	0.095	0	0.5	0.065
TECHORIG	270	0.628	0.119	0	0.9	0.644
INVENTORS	270	3.47	1.843	1	16	3.261
UNIVCITE	270	0.08	0.105	0	0.611	0.05
OKV	270	0.788	0.078	0	0.876	0.804
Panel B: European Clusters						
EVOUT	1155	11.366	27.263	0	220	2
ITA	1155	-0.354	1.703	-4.97	4.754	0
TCT	1155	11.155	6.254	1.25	58.142	9.809
LKA	1155	0.521	0.275	0	1	0.544
SELFCITE	1155	0.066	0.090	0	0.667	0.028
TECHORIG	1155	0.599	0.138	0	0.906	0.616
INVENTORS	1155	2.246	1.024	.5	13	2.111
UNIVCITE	1155	0.009	0.033	0	0.5	0
OKV	1155	0.749	0.090	0	0.911	0.758

5.3 Correlation Analysis

Table 5.3 presents the Pearson correlation matrices for the Chinese and European cluster subsamples, respectively. Across both matrices, the absolute values of the correlation coefficients among all explanatory variables remain significantly below the conventional threshold of 0.70. This initial observation suggests that severe multicollinearity is unlikely to threaten the empirical models. To systematically validate this assumption, a Variance Inflation Factor analysis was conducted separately for both regional datasets before the regressions. The diagnostic results confirm that the maximum VIF value is 1.16 for the Chinese sample and 1.32 for the European sample. Both values are strictly below the conservative threshold of 5. This collective evidence suggests that multicollinearity is unlikely to pose a serious concern for the subsequent high-dimensional fixed effects estimations.

Table 5.3: Pearson Correlation Matrix for Chinese and European Clusters

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Chinese Clusters									
(1)EVOUT	1								
(2)DTF	-0.471***	1							
(3)TCT	-0.154**	0.238***	1						
(4)LKA	-0.202***	0.300***	0.110*	1					
(5)SELFCITE	0.100*	-0.101*	-0.223***	-0.006	1				
(6)TECHORIG	-0.016	0.018	0.060	-0.036	-0.047	1			
(7)INVENTORS	0.068	-0.045	-0.194***	0.066	0.164***	-0.044	1		
(8)UNIVCITE	-0.014	-0.005	-0.074	-0.022	0.265***	0.105*	0.095	1	
(9)OKV	0.096	-0.101*	0.054	-0.123**	-0.055	0.316***	0.118*	0.040	1
Panel B: European Clusters									
(1)EVOUT	1								
(2)ITA	0.388***	1							
(3)TCT	-0.130***	-0.067**	1						
(4)LKA	-0.055*	0.151***	0.082***	1					
(5)SELFCITE	0.301***	0.225***	-0.128***	-0.059**	1				
(6)TECHORIG	-0.026	-0.125***	0.014	0.017	-0.055*	1			
(7)INVENTORS	0.132***	0.050*	-0.183***	-0.067**	0.166***	-0.013	1		
(8)UNIVCITE	-0.006	-0.170***	-0.081***	-0.070**	0.001	-0.007	0.041	1	
(9)OKV	-0.023	0.000	0.138***	0.038	-0.119***	0.456***	0.043	0.023	1

*Note: Pearson correlation matrix for the Chinese and European cluster subsample. Significance levels are Bonferroni-adjusted for multiple comparisons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

5.4 Regression Results

Table 5.4 presents the baseline regression results estimated using the PPML estimator with high-dimensional fixed effects. Columns (1) and (4) include only the core explanatory variables. Columns (2) and (5) introduce SELFCITE, TECHORIG, and INVENTORS as controls while excluding UNIVCITE and OKV, to assess the sensitivity of the core estimates to potential measurement error arising from Professor's Privilege arrangements in several European countries. Columns (3) and (6) present the full models incorporating all variables. Given the exponential conditional mean specification of the PPML estimator, the estimated coefficients are interpreted as semi-elasticities. All specifications include both cluster and year fixed effects.

The stability of the estimated coefficients for TCT and LKA across Columns (1) through (3) and Columns (4) through (6), respectively, confirms that the core findings are not sensitive to the inclusion or exclusion of UNIVCITE and OKV, suggesting that the potential measurement issues associated with the university citation variable are unlikely to materially affect the primary estimates

The core explanatory variable of TCT exhibits a highly significant negative effect across all baseline models. The estimates provide strong empirical evidence that the rate of knowledge renewal is significantly associated with innovation output. In Column (3), the coefficient for TCT is -0.098 ($p < 0.01$). This implies that, holding other multidimensional heterogeneous factors constant, a one-year increase in TCT is associated with an approximate 9.3% ($e^{-0.098}$-1) significant decrease in the expected EV innovation output of Chinese clusters. Comparatively, the negative output effect for a one-year increase in TCT within the full European model in Column (6) is approximately 3.3% ($e^{-0.034}$-1). These findings are consistent with the argument that rapid knowledge iteration may facilitate the exploitation of technological windows of opportunity. Rapid knowledge iteration is associated with clusters to more effectively absorb frontier external knowledge and adapt to the emerging technological trajectory.

The estimation results for LKA further quantify the penalty costs of path dependence and technological lock-in effects. Given that LKA is a proportional variable strictly bounded between 0 and 1, the study adopts more rigorous percentage-point changes to characterize its economic impact. Based on the estimates in Column (3) (β = -1.297, $p < 0.05$), a 10 percentage-point increase in the proportion of traditional ICE patents within the knowledge base of Chinese clusters is associated with a substantial 12.2% decline ($e^{-1.297 \times 0.1}$-1) in expected EV innovation output. In Column (6), European clusters facing the same 10% incremental historical burden (β = -0.615, $p < 0.1$) would experience an approximate 6.0% drop ($e^{-0.615 \times 0.1}$-1) in output. These coefficients indicate that a heavy reliance on the legacy paradigm is associated with a technological lock-in effect. This negative relationship is consistent with the presence of path-dependent adjustment costs that hinder the reallocation of R&D resources toward the new electric vehicle trajectory. The cross-regional empirical evidence suggests that legacy knowledge is significantly associated with reduced innovation productivity during the transition towards new technological paradigms. Collectively, the significant negative coefficients for both LKA and TCT across the Sino-European samples provide consistent empirical support for Hypothesis 1.

Table 5.4: Main result with EVOUT as dependent variable

	EVOUT (China)			EVOUT (Europe)		
	(1)	(2)	(3)	(4)	(5)	(6)
TCT	-0.116***	-0.113***	-0.098***	-0.036***	-0.032***	-0.034***
	(0.030)	(0.031)	(0.030)	(0.008)	(0.008)	(0.008)
LKA	-1.165**	-1.348**	-1.297**	-0.605*	-0.640*	-0.615*
	(0.511)	(0.541)	(0.516)	(0.357)	(0.337)	(0.329)
SELFCITE		-1.383**	-1.581**		0.539	0.663
		(0.624)	(0.762)		(0.429)	(0.432)
TECHORIG		-1.310*	-0.827		-0.711***	-0.981***
		(0.705)	(0.713)		(0.180)	(0.185)
INVENTORS		0.002	0.024		0.082**	0.081*
		(0.051)	(0.045)		(0.041)	(0.041)
UNIVCITE			-2.005***			-2.006**
			(0.603)			(0.921)
OKV			-1.913			0.782
			(1.265)			(0.480)
Constants	3.812***	4.793***	6.050***	4.259***	4.374***	3.969***
	(0.158)	(0.662)	(1.384)	(0.202)	(0.275)	(0.371)
Cluster FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Wald χ^2	19.845	24.079	35.018	22.339	46.635	71.296
Prob > χ^2	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.674	0.677	0.683	0.839	0.841	0.841
Observations	264	264	264	1155	1155	1155
Clusters	22	22	22	57	57	57

Standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.5 presents the baseline estimation results utilizing OLS-HDFE. The dependent variables of DTF and ITA are specified as logarithmic differences. The linear panel model captures how micro-level technological attributes are associated with the rate of technological convergence and the erosion of leadership.

Similarly, Columns (2) and (5) retain SELFCITE, TECHORIG, and INVENTORS while excluding UNIVCITE and OKV. The estimated coefficients for TCT and LKA remain stable in sign, magnitude, and significance across all three specifications in both the DTF and ITA models, confirming that the potential measurement issues associated with the university citation variable are unlikely to affect the core findings.

The DTF model for Chinese clusters characterizes the micro-foundations of the latecomer catch-up process. The results indicate that both the TCT and LKA are significantly associated with the distance to the European frontier. In Column (3), the positive coefficient for TCT (0.058, p < 0.05) highlights the critical value of innovation agility. A one-year reduction in TCT is associated with an approximate 5.8% contraction ($e^{0.058}$-1) in the distance to the

European benchmark. Furthermore, the coefficient for LKA (0.755, $p < 0.05$) quantifies the burden of historical path dependence. A 10 percentage-point decrease in LKA facilitates an approximate 7.8% narrowing ($e^{0.755\times0.1}$-1) of the technological gap. These findings highlight the simultaneous importance of technological agility and legacy knowledge structure in explaining catch-up performance. These results suggest that the ability to maintain a lightweight knowledge structure and rapid iteration is important for technological convergence. Based on the estimated coefficients, the magnitude of a 1 percentage-point increase in LKA is numerically comparable to approximately a 0.13-year increase in TCT. Clusters with lower legacy burdens may therefore face fewer sunk-cost-related adjustment constraints during the transition toward the EV paradigm. Collectively, this evidence provides empirical support for Hypothesis 2.

Table 5.5: Main result with DTF and ITA as dependent variables

	DTF (China)			ITA (Europe)		
	(1)	(2)	(3)	(4)	(5)	(6)
TCT	0.058^{**}	0.058^{**}	0.058^{**}	-0.024^{***}	-0.024^{***}	-0.025^{***}
	(0.023)	(0.023)	(0.024)	(0.004)	(0.004)	(0.004)
LKA	0.751^{**}	0.766^{**}	0.755^{**}	-0.325^{**}	-0.312^{**}	-0.301^{**}
	(0.312)	(0.317)	(0.314)	(0.146)	(0.146)	(0.148)
SELFCITE		0.582	0.406		0.531	0.611^{*}
		(0.495)	(0.519)		(0.349)	(0.340)
TECHORIG		0.443	0.411		-0.265^{**}	-0.470^{**}
		(0.530)	(0.547)		(0.128)	(0.185)
INVENTORS		0.002	0.001		0.012	0.009
		(0.019)	(0.017)		(0.025)	(0.026)
UNIVCITE			0.644^{*}			-0.865
			(0.311)			(0.605)
OKV			-0.160			0.632^{**}
			(0.577)			(0.293)
Constants	1.131^{***}	0.787^{*}	0.903	0.085	0.169	-0.162
	(0.182)	(0.437)	(0.565)	(0.097)	(0.131)	(0.200)
Cluster FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Within R^2	0.071	0.078	0.085	0.048	0.053	0.059
Adjusted R^2	0.635	0.633	0.632	0.833	0.834	0.835
Wald F	4.110	2.192	2.134	17.052	11.071	9.724
Observations	266	266	266	1155	1155	1155
Clusters	22	22	22	57	57	57

Standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The ITA model for the European sub-sample symmetrically validates the fragility of the incumbent position in the face of structural change. The results in Column (6) show that TCT and LKA exhibit significant negative impacts on this relative advantage. A one-year extension in the technology cycle is associated with an approximate 2.5% reduction ($e^{-0.025}$-1) in the European leadership margin. Conversely, an increase of 10 percentage points in the legacy knowledge share further reduces the incumbent technological advantage by approximately 3.0% ($e^{-0.301\times0.1}$-1). A similar internal trade-off exists within the European innovation ecosystem. Within the European sample, the estimated marginal effect of a 1 percentage-point increase in LKA is numerically comparable to approximately a 0.12-year increase in TCT. Taken together, the results suggest that incumbent technological advantage becomes increasingly difficult to sustain when clusters remain heavily dependent on legacy ICE knowledge and exhibit slower knowledge renewal. These findings suggest that maintaining incumbent technological advantage during periods of paradigm transition may require faster knowledge renewal and a gradual reduction in dependence on legacy ICE knowledge structures. These findings substantiate Hypothesis 3, showing that deep legacy dependence and delayed technological renewal are significantly associated with the erosion of incumbent technological advantage.

5.5 Robustness Checks

Table 5.6 presents the robustness checks for the baseline EVOUT estimations. The construction of the baseline LKA indicator relies on a standard 15 percent annual depreciation rate. To ensure the empirical findings are not merely artifacts of this specific parameter selection, Columns (1) and (4) recalculate the indicator using a 10 percent depreciation rate to simulate slower technological obsolescence. Subsequently, Columns (2) and (5) apply a 20 percent depreciation rate to model faster knowledge decay (Aghion et al., 2016). Across these variations, the estimated coefficients for LKA remain negative and statistically significant. Additionally, Columns (3) and (6) substitute the PPML estimator with a conditional fixed-effects Poisson model to assess the sensitivity of the results to alternative count-data estimation approaches. The regression results reveal no substantive changes in the sign, magnitude, or statistical significance of the core explanatory variables.

Table 5.6: Robustness check for regressions with EVOUT as dependent variable

	EVOUT (China)			EVOUT (Europe)		
	(1)	(2)	(3)	(4)	(5)	(6)
TCT	-0.099***	-0.097***	-0.098***	-0.034***	-0.034***	-0.034***
	(0.030)	(0.030)	(0.029)	(0.008)	(0.008)	(0.008)
LKA	-1.480***	-1.114**	-1.297**	-0.632*	-0.591*	-0.615*

	(0.555)	(0.472)	(0.504)	(0.340)	(0.321)	(0.326)
SELFCITE	-1.580**	-1.566**	-1.581**	0.641	0.688	0.663
	(0.763)	(0.751)	(0.744)	(0.432)	(0.432)	(0.428)
TECHORIG	-0.874	-0.773	-0.827	-0.971***	-0.992***	-0.981***
	(0.716)	(0.711)	(0.697)	(0.183)	(0.188)	(0.184)
INVENTORS	0.024	0.025	0.024	0.080*	0.082**	0.081**
	(0.045)	(0.045)	(0.044)	(0.041)	(0.041)	(0.041)
UNIVCITE	-2.022***	-1.997***	-2.005***	-1.992**	-2.021**	-2.006**
	(0.610)	(0.597)	(0.589)	(0.922)	(0.920)	(0.912)
OKV	-1.925	-1.901	-1.913	0.778	0.792	0.782*
	(1.261)	(1.270)	(1.236)	(0.479)	(0.482)	(0.475)
Constants	6.139***	5.962***	-	3.992***	3.943***	-
	(1.408)	(1.359)	-	(0.382)	(0.362)	-
Cluster FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Wald χ^2	35.571	34.408	19357.178	72.907	69.385	2680.399
Prob > χ^2	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.685	0.682	-	0.841	0.841	-
Log-likelihood	-992	-1002	-1468	-2982	-2980	-4269
Observations	264	264	270	1155	1155	1155
Clusters	22	22	22	57	57	57

Standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.7 presents the corresponding robustness checks for the DTF and the ITA models. Mirroring the previous empirical design in Table 5.6, Columns (1) and (4) apply the 10 percent depreciation rate, while Columns (2) and (5) utilize the 20 percent parameter. These estimations confirm that the continuous gap variables remain qualitatively stable under alternative knowledge depreciation assumptions. Moving beyond depreciation sensitivity, Columns (3) and (6) address the potential sensitivity of the baseline estimations to the specific threshold selection of the scale tiers. To execute this test, the dependent variables DTF and ITA are reconstructed utilizing standard statistical terciles at the 33rd and 67th percentiles instead of the baseline 40th and 90th percentiles. Under this alternative tercile-based partition, the estimated coefficients for both the TCT and the LKA maintain consistent signs and broadly similar levels of statistical significance.

Table 5.7: Robustness check for regressions with DTF and ITA as dependent variables

	DTF (China)			ITA (Europe)		
	(1)	(2)	(3)	(4)	(5)	(6)
TCT	0.058**	0.058**	0.048**	-0.025***	-0.025***	-0.023***
	(0.024)	(0.024)	(0.023)	(0.004)	(0.004)	(0.004)
LKA	0.786**	0.719**	0.600*	-0.291*	-0.306**	-0.440***
	(0.322)	(0.305)	(0.302)	(0.155)	(0.141)	(0.153)
SELFCITE	0.406	0.404	0.281	0.610*	0.612*	0.825**
	(0.521)	(0.519)	(0.476)	(0.340)	(0.340)	(0.394)
TECHORIG	0.411	0.412	0.300	-0.473**	-0.468**	-0.511***

	(0.548)	(0.547)	(0.595)	(0.185)	(0.184)	(0.183)
INVENTORS	0.001	0.001	-0.021	0.009	0.010	-0.014
	(0.017)	(0.017)	(0.022)	(0.026)	(0.026)	(0.027)
UNIVCITE	0.645*	0.643*	0.789***	-0.862	-0.866	0.263
	(0.313)	(0.310)	(0.244)	(0.604)	(0.605)	(0.628)
OKV	-0.147	-0.174	-0.229	0.638**	0.627**	0.692**
	(0.579)	(0.576)	(0.606)	(0.294)	(0.293)	(0.325)
Constants	0.880	0.928	0.754	-0.166	-0.160	-0.029
	(0.573)	(0.557)	(0.580)	(0.203)	(0.197)	(0.213)
Cluster FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Within R^2	0.087	0.083	0.068	0.059	0.060	0.056
Adjusted R^2	0.633	0.631	0.743	0.835	0.835	0.793
Wald F	2.170	2.088	2.453	9.699	9.763	7.498
Observations	266	266	266	1155	1155	1155
Clusters	22	22	22	57	57	57

Standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6 Discussion and Conclusion

This chapter discusses the main findings of the study and their implications for regional technological transition and economic development. Section 6.1 summarizes the key empirical results on regional innovation performance, regional catch-up, and changing leadership patterns in the automotive industry. Section 6.2 discusses the study's theoretical and methodological contributions to the literature on evolutionary economic geography and technological catch-up. Finally, Section 6.3 discusses the policy implications for both latecomer and established regional innovation systems.

6.1 Summary of Main Findings

This study examines how legacy knowledge accumulation and technology cycle time jointly shape regional innovation performance and catch-up trajectories during the automotive electrification transition. Drawing on a cross-regional patent panel of Chinese and European clusters identified through data-driven spatial clustering, the empirical analysis produces three main findings.

First, the baseline results indicate that both lower legacy knowledge accumulation and shorter technology cycle times are positively associated with higher electric vehicle innovation output, consistent with Hypothesis 1. Across the full sample, clusters characterized by heavier dependence on ICE knowledge and slower knowledge renewal processes tend to generate lower levels of EV patenting activity after controlling for cluster-specific and year-specific heterogeneity. The estimated effects are also more pronounced among Chinese clusters than among European clusters. For example, a one-year increase in technology cycle time is associated with an approximately 9.3% reduction in expected EV patent output in China, compared to 3.3% in Europe. Similarly, a ten-percentage-point increase in legacy knowledge share corresponds to a 12.2% negative output effect among Chinese clusters versus 6.0% among European clusters. These findings suggest that the structure and renewal dynamics of knowledge bases may play a particularly important role in explaining innovation outcomes within latecomer innovation systems (Tushman & Anderson, 1986; Aghion et al., 2016; Lee & Malerba, 2017; Altenburg et al., 2022).

Second, the results indicate that these mechanisms are also associated with regional catch-up trajectories, supporting Hypotheses 2 and 3. Among Chinese clusters, lower reliance on legacy ICE knowledge and shorter technology cycle times are both associated with a narrowing of the technological gap relative to European frontier regions (Griffith et al., 2004; Lee & Malerba, 2017). The estimated trade-off suggests that, on the margin, a 0.13-year reduction in technology cycle time may offset the catch-up penalty associated with a one-percentage-point increase in legacy knowledge share. This finding is consistent with evolutionary arguments emphasizing the importance of both reducing technological lock-in and sustaining knowledge renewal during periods of paradigm shift (Park & Lee, 2015; Lee & Malerba, 2017; Lee, 2024).

Finally, for European incumbent clusters, the results point to a gradual weakening of relative technological leadership during the electrification transition. Higher levels of legacy dependence and longer technology cycle times are both associated with larger declines in relative technological advantage. These findings suggest that historical specialization in ICE technologies may constrain adaptability under rapidly changing technological conditions (Henderson & Clark, 1990; Leonard-Barton, 1992; Aghion et al., 2016).

Taken together, the findings highlight the importance of knowledge structure and knowledge renewal dynamics in explaining uneven regional adaptation during technological transition. The results suggest that regional differences in historical technological specialization and innovation renewal capacity may influence how effectively clusters respond to paradigm shifts in the automotive industry. In particular, regions characterized by lighter legacy dependence and faster knowledge renewal processes appear to be better positioned to adapt to emerging technological trajectories during the electrification transition.

6.2 Theoretical Contributions

This study makes three contributions to the existing literature. First, it integrates legacy knowledge accumulation and technology cycle time into a unified analytical framework for examining regional catch-up during technological transitions. While prior studies have separately examined the constraining effects of path dependence on incumbent innovation systems (Aghion et al., 2016) and the role of technology cycle dynamics in shaping catch-up opportunities (Lee, 2013, pp.139-140; Lee, 2024), these two mechanisms have rarely been examined jointly at the regional cluster level. The findings suggest that legacy dependence and knowledge renewal capacity operate as independent structural dimensions of regional innovation systems, each exhibiting independent associations with innovation output and relative catch-up trajectories. This integrated perspective offers a more complete characterization of the knowledge-structural conditions underlying uneven regional adaptation during paradigm shifts.

Second, the study contributes methodologically by employing data-driven spatial clustering to identify innovation clusters as the primary unit of analysis. Existing empirical research on regional innovation systems and catch-up dynamics predominantly relies on predefined administrative boundaries, which may not accurately reflect the spatial organization of knowledge creation and technological spillovers (Binz & Truffer, 2017). By applying the HDBSCAN algorithm to the geographic distribution of patent assignees, this study allows cluster boundaries to emerge endogenously from observed patterns of spatial concentration. This approach provides a more precise delineation of the geographic units within which knowledge structures and renewal dynamics operate, and offers a complementary identification strategy for future empirical work on regional innovation systems.

Third, the study provides additional insights into the spatial organization of innovation during technological catch-up. The clustering results reveal noticeable differences in the spatial morphology of Chinese and European automotive innovation systems. In particular, the HDBSCAN algorithm identifies several high-output Chinese firms, including BYD and Aulton, as spatial noise because their innovation activities are concentrated within highly localized geographic coordinates rather than embedded within dense multi-firm agglomeration structures. By contrast, European incumbents more frequently exhibit geographically distributed cluster configurations associated with long-established industrial regions and subsidiary networks.

While these findings should be interpreted cautiously, they suggest that latecomer innovation systems may not always conform to the highly agglomerated spatial patterns commonly emphasized in the regional innovation systems literature. More broadly, the results indicate that differences in organizational structure and spatial concentration may represent an additional dimension explaining regional innovation dynamics during technological transition.

6.3 Policy Implications

The empirical findings carry several implications for innovation policy in both latecomer and incumbent regional contexts, although the correlational nature of the analysis warrants caution in drawing strong causal policy conclusions. The following discussion focuses specifically on the conditions under which regional knowledge structures may support or constrain EV innovation performance.

For latecomer regions, the findings suggest two directions for policy consideration. First, evaluation frameworks governing public R&D support and industrial incentive programs may benefit from incorporating knowledge frontier orientation alongside aggregate patent volume as an indicator of innovation quality. The stronger sensitivity of Chinese clusters to technology cycle time suggests that the speed of knowledge renewal may play a particularly important role in latecomer innovation systems. This finding suggests that the recency of knowledge inputs, rather than the scale of patent accumulation alone, may influence EV innovation performance in these regions. Second, within the specific objective of strengthening EV innovation capacity, the allocation of public research funding may benefit from explicit attention to the directional composition of regional knowledge stocks. The findings indicate that the share of legacy ICE knowledge within the accumulated patent stock is associated with meaningfully lower EV innovation output, even at relatively low baseline levels of legacy dependence. This suggests that the technological composition of regional knowledge stocks, and not only their aggregate scale, may warrant consideration in the design of innovation support programs.

For incumbent regions, the findings point to two complementary policy directions. First, sustaining competitiveness in electric vehicle technologies may require policy instruments that support the organizational and technological adjustment processes associated with industrial transition, rather than relying exclusively on direct EV research subsidies. The results indicate that higher legacy knowledge dependence is associated with erosion of both absolute EV innovation output and relative technological advantage, suggesting that the weight of accumulated ICE knowledge may constrain the reallocation of organizational resources and capabilities toward the emerging paradigm. Policies that facilitate the reallocation of research capabilities, skilled labor, and innovation investment toward emerging EV technologies may therefore complement direct R&D subsidies. Second, the findings suggest that the speed of knowledge renewal represents a dimension of competitive positioning that is analytically distinct from the scale of research investment. Innovation funding frameworks that place greater emphasis on the novelty, recency, and technological relevance of research activities, rather than focusing primarily on expenditure scale, may provide more effective support during periods of rapid technological transition.

7 Limitations and Future Research

Several limitations of this study should be acknowledged alongside directions for future research.

First, the analysis does not explicitly model spatial dependence across innovation clusters. This omission reflects both methodological and theoretical considerations. The HDBSCAN procedure identifies clusters based on endogenous spatial density rather than predefined administrative units, which complicates the application of conventional spatial econometric specifications and fixed spatial weight matrices. In addition, the primary focus of this study concerns the relationship between internal knowledge structures and cluster-level innovation performance rather than inter-cluster diffusion processes. Nevertheless, unobserved spatial spillovers may still influence regional innovation dynamics. Future research could explore alternative spatial identification strategies capable of incorporating more flexible forms of inter-cluster connectivity and knowledge diffusion.

Second, the measurement of regional innovation relies exclusively on PCT international patent applications. While this choice ensures cross-regional comparability, it excludes domestic patent filings that may capture a substantial share of localized technological activity. This limitation may also help explain the spatial structure observed in the noise pool, where several Chinese giants appear highly concentrated at single geographic coordinates, whereas European incumbents more frequently exhibit geographically distributed innovation networks. One possible mechanism is that some localized supplier and collaborative innovation activities surrounding Chinese giants are more heavily represented in domestic patent filings and therefore remain only partially visible within the PCT framework. As a result, the current analysis primarily captures internationally oriented innovation networks rather than the full domestic industrial ecosystem. Future research integrating domestic patent applications could provide a more comprehensive view of these localized innovation structures and complement the spatial analysis with targeted corporate case studies.

Third, the observation window is bounded to the 2000–2024 period. Patent activity in the most recent years may be affected by reporting and indexing delays within international patent databases. Although the analysis does not extend beyond 2024 in order to limit additional truncation bias, recent patent cohorts should still be interpreted with caution. Furthermore, because the technology cycle time and some other indicators within the study rely on backward patent citations, incomplete citation histories for recent cohorts may constrain the full observation of knowledge trajectories. Consequently, the estimated effects may understate the speed of knowledge renewal and the extent of regional catch-up dynamics during the final years of the sample. Future longitudinal extensions using updated database releases could provide a more comprehensive assessment of these recent cohorts as additional citation information becomes available.

Fourth, the $ln(x + 1)$ transformation applied to the technological gap indicators introduces a minor attenuation bias at the lower bound. This smoothing constant is empirically necessary to retain zero-value observations, although it slightly reduces estimation precision for clusters with minimal patent output. Alternative transformations such as the inverse hyperbolic sine (IHS) specification handle zero-valued right-skewed data without this distortion (Bellemare & Wichman, 2020). However, its coefficients require non-linear transformations of marginal effects for economic interpretation, making them less directly comparable to the log-linear elasticity framework. For this reason, the $ln(x + 1)$ specification is retained to preserve a more transparent interpretation of proportional catch-up dynamics. Future research may nonetheless consider IHS as a methodological extension.

Fifth, this study employs a keyword-based method to identify universities and public research institutions from the broader pool of patent assignees. This identification strategy may underestimate the contribution of European academic institutions operating under the Professor's Privilege systems. Although the intermediate regressions suggest that this measurement bias has a limited impact on the core findings, the organizational classification remains imperfect. Future studies could address this limitation by using matched inventor–employer datasets to better trace academic patents across corporate boundaries.

References

Abramovitz, M. (1986). Catching Up, Forging Ahead, and Falling Behind, *The Journal of Economic History*, vol. 46, no. 2, pp. 385-406, https://doi.org/10.1017/S0022050700046209

Acemoglu, D., & Cao, D. (2015). Innovation by Entrants and Incumbents, *Journal of Economic Theory*, vol. 157, pp. 255-294, https://doi.org/10.1016/j.jet.2015.01.001

Acemoglu, D., Aghion, P., & Zilibotti, F. (2006). Distance to Frontier, Selection, and Economic Growth, *Journal of the European Economic Association*, vol. 4, no. 1, pp. 37-74, https://doi.org/10.1162/jeea.2006.4.1.37

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and Innovation: An Inverted-U Relationship, *The Quarterly Journal of Economics*, vol. 120, no. 2, pp. 701-728, https://doi.org/10.1093/qje/120.2.701

Aghion, P., Dechezleprêtre, A., Hemous, D., Martin, R., & Van Reenen, J. (2016). Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry, *Journal of Political Economy*, vol. 124, no. 1, pp. 1-51, https://doi.org/10.1086/684581

Allison, P. D., & Waterman, R. P. (2002). Fixed-Effects Negative Binomial Regression Models, *Sociological Methodology*, vol. 32, no. 1, pp. 247-265, https://doi.org/10.1111/1467-9531.00117

Altenburg, T., Bhasin, S., & Fischer, D. (2012). Sustainability-Oriented Innovation in the Automobile Industry: Advancing Electromobility in China, France, Germany and India, *Innovation and Development*, vol. 2, no. 1, pp. 67-85, https://doi.org/10.1080/2157930X.2012.664036

Altenburg, T., Corrocher, N., & Malerba, F. (2022). China's Leapfrogging in Electromobility: A Story of Green Transformation Driving Catch Up and Competitive Advantage, *Technological Forecasting and Social Change*, vol. 183, p. 121914, https://doi.org/10.1016/j.techfore.2022.121914

Archibugi, D. (1992). Patenting as an Indicator of Technological Innovation: A Review, *Science and Public Policy*, vol. 19, no. 6, pp. 357-368, https://doi.org/10.1093/spp/19.6.357

Arthur, W. B. (1989). Competing Technologies, Increasing Returns, and Lock-in by Historical Events, *The Economic Journal*, vol. 99, no. 394, pp. 116-131, https://doi.org/10.2307/2234208

Asheim, B. T., & Gertler, M. S. (2005). The Geography of Innovation: Regional Innovation Systems, in J. Fagerberg, D. C. Mowery & R. R. Nelson (eds), *The Oxford Handbook of Innovation*, Oxford: Oxford University Press, pp. 291-317

Asheim, B. T., & Isaksen, A. (2002). Regional Innovation Systems: The Integration of Local "Sticky" and Global "Ubiquitous" Knowledge, *Journal of Technology Transfer*, vol. 27, pp. 77-86, https://doi.org/10.1023/A:1013100704794

Asheim, B. T., Smith, H. L., & Oughton, C. (2011). Regional Innovation Systems: Theory, Empirics and Policy, *Regional Studies*, vol. 45, no. 7, pp. 875-891, https://doi.org/10.1080/00343404.2011.596701

Audretsch, D. B., & Feldman, M. P. (1996). R&D Spillovers and the Geography of Innovation and Production, *The American Economic Review*, vol. 86, no. 3, pp. 630-640, https://www.jstor.org/stable/2118216

Autio, E. (1998). Evaluation of RTD in Regional Systems of Innovation, *European Planning Studies*, vol. 6, no. 2, pp. 131-140, https://doi.org/10.1080/09654319808720451

Balland, P. A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart Specialization Policy in the European Union: Relatedness, Knowledge Complexity and Regional Diversification, *Regional Studies*, vol. 53, no. 9, pp. 1252-1268, https://doi.org/10.1080/00343404.2018.1437900

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and Knowledge: Local Buzz, Global Pipelines and the Process of Knowledge Creation, *Progress in Human Geography*, vol. 28, no. 1, pp. 31-56, https://doi.org/10.1191/0309132504ph469oa

Bell, M., & Pavitt, K. (1993). Technological Accumulation and Industrial Growth: Contrasts between Developed and Developing Countries, *Industrial and Corporate Change*, vol. 2, no. 1, pp. 157-210, https://doi.org/10.1093/icc/2.2.157

Bellemare, M. F., & Wichman, C. J. (2020). Elasticities and the Inverse Hyperbolic Sine Transformation, *Oxford Bulletin of Economics and Statistics*, vol. 82, no. 1, pp. 50-61, https://doi.org/10.1111/obes.12325

Binz, C., & Truffer, B. (2017). Global Innovation Systems—A Conceptual Framework for Innovation Dynamics in Transnational Contexts, *Research Policy*, vol. 46, no. 7, pp. 1284-1298, https://doi.org/10.1016/j.respol.2017.05.012

Borgstedt, P., Neyer, B., & Schewe, G. (2017). Paving the Road to Electric Vehicles – A Patent Analysis of the Automotive Supply Industry, *Journal of Cleaner Production*, vol. 167, pp. 75-87, https://doi.org/10.1016/j.jclepro.2017.08.161

Boschma, R. (2015). Towards an Evolutionary Perspective on Regional Resilience, *Regional Studies*, vol. 49, no. 5, pp. 733-751, https://doi.org/10.1080/00343404.2014.959481

Boschma, R., & Frenken, K. (2006). Why is Economic Geography Not an Evolutionary Science? Towards an Evolutionary Economic Geography, *Journal of Economic Geography*, vol. 6, no. 3, pp. 273-302, https://doi.org/10.1093/jeg/lbi022

Bosworth, D. L. (1978). The Rate of Obsolescence of Technical Knowledge—A Note, *The Journal of Industrial Economics*, vol. 26, no. 3, pp. 273-279, https://doi.org/10.2307/2097871

Brenndoerfer, M. (2025). *Machine Learning from Scratch*, https://mbrenndoerfer.com/books/machine-learning-from-scratch

Bristow, G., & Healy, A. (2014). Regional Resilience: An Agency Perspective, *Regional Studies*, vol. 48, no. 5, pp. 923-935, https://doi.org/10.1080/00343404.2013.854879

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates, in J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (eds), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, vol. 7819, Berlin, Heidelberg: Springer, pp. 160-172

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection, *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 1, pp. 1-51, https://doi.org/10.1145/2733381

Cecere, G., Corrocher, N., Gossart, C., & Ozman, M. (2014). Lock-in and Path Dependence: An Evolutionary Approach to Eco-Innovations, *Journal of Evolutionary Economics*, vol. 24, no. 5, pp. 1037-1065, https://doi.org/10.1007/s00191-014-0381-5

China National Intellectual Property Administration (CNIPA). (2024). Patent Classification System for "New Three" Related Technologies, https://www.cnipa.gov.cn/art/2024/8/9/art_75_194145.html

Christensen, C. M., & Rosenbloom, R. S. (1995). Explaining the Attacker's Advantage: Technological Paradigms, Organizational Dynamics, and the Value Network, *Research Policy*, vol. 24, no. 2, pp. 233-257, https://doi.org/10.1016/0048-7333(93)00764-K

Christopherson, S., Michie, J., & Tyler, P. (2010). Regional Resilience: Theoretical and Empirical Perspectives, *Cambridge Journal of Regions, Economy and Society*, vol. 3, no. 1, pp. 3-10, https://doi.org/10.1093/cjres/rsq004

Chu, H., Hassink, R., Martin, R., Sunley, P., & Unruh, G. (2026). Rethinking Path Dependence and Lock-ins in Regions, Economy and Society, *Cambridge Journal of Regions, Economy and Society*, vol. 19, no. 1, pp. 1-15, https://doi.org/10.1093/cjres/rsag001

Cimoli, M., & Dosi, G. (1995). Technological Paradigms, Patterns of Learning and Development: An Introductory Roadmap, *Journal of Evolutionary Economics*, vol. 5, pp. 243-268, https://doi.org/10.1007/BF01198306

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation, *Administrative Science Quarterly*, vol. 35, no. 1, pp. 128-152, https://doi.org/10.2307/2393553

Cooke, P. (2002). Regional Innovation Systems: General Findings and Some New Evidence from Biotechnology Clusters, *The Journal of Technology Transfer*, vol. 27, pp. 133-145, https://doi.org/10.1023/A:1013160923450

Cooke, P., Gomez Uranga, M., & Etxebarria, G. (1997). Regional Innovation Systems: Institutional and Organisational Dimensions, *Research Policy*, vol. 26, no. 4-5, pp. 475-491, https://doi.org/10.1016/S0048-7333(97)00025-5

Correia, S., Guimarães, P., & Zylkin, T. (2020). Fast Poisson Estimation with High-Dimensional Fixed Effects, *The Stata Journal*, vol. 20, no. 1, pp. 95-115, https://doi.org/10.1177/1536867X20909691

Corrocher, N., Malerba, F., & Montobbio, F. (2007). Schumpeterian Patterns of Innovative Activity in the ICT Field, *Research Policy*, vol. 36, no. 3, pp. 418-432, https://doi.org/10.1016/j.respol.2007.01.002

David, P. A. (1985). Clio and the Economics of QWERTY, *The American Economic Review*, vol. 75, no. 2, pp. 332-337, https://www.jstor.org/stable/1805621

Davy, E., & Hansen, U. E. (2023). A Firm-Level Perspective on Windows of Opportunity, *Technology in Society*, vol. 75, p. 102374, https://doi.org/10.1016/j.techsoc.2023.102374

Delgado, M., Porter, M. E., & Stern, S. (2010). Clusters and Entrepreneurship, *Journal of Economic Geography*, vol. 10, no. 4, pp. 495-518, https://doi.org/10.1093/jeg/lbq010

Department for Science, Innovation and Technology (DSIT). (2025). *Innovation Clusters Map: Summary and Methods*, Independent report, London: GOV.UK, https://www.gov.uk/government/publications/innovation-clusters-map-summary-and-methods/innovation-clusters-map-summary-and-methods

Dijk, M., Kemp, R., & Valkering, P. (2013). Incorporating Social Context and Co-evolution in an Innovation Diffusion Model—With an Application to Cleaner Vehicles, *Journal of Evolutionary Economics*, vol. 23, no. 2, pp. 295-329, https://doi.org/10.1007/s00191-011-0241-5

Dosi, G. (1982). Technological Paradigms and Technological Trajectories: A Suggested Interpretation of the Determinants and Directions of Technical Change, *Research Policy*, vol. 11, no. 3, pp. 147-162, https://doi.org/10.1016/0048-7333(82)90016-6

Ejermo, O., & Källström, J. (2016). What is the Causal Effect of R&D on Patenting Activity in a "Professor's Privilege" Country? Evidence from Sweden, *Small Business Economics*, vol. 47, no. 3, pp. 677-694, https://doi.org/10.1007/s11187-016-9752-7

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR: AAAI Press, pp. 226-231

Fagerberg, J. (1994). Technology and International Differences in Growth Rates, *Journal of Economic Literature*, vol. 32, no. 3, pp. 1147-1175, https://www.jstor.org/stable/2728605

Fagerberg, J., & Godinho, M. M. (2005). Innovation and Catching-up, in J. Fagerberg, D. C. Mowery & R. R. Nelson (eds), *The Oxford Handbook of Innovation*, Oxford: Oxford University Press, pp. 514-543

Fakhimi, M., & Miremadi, I. (2022). The Impact of Technological and Social Capabilities on Innovation Performance: A Technological Catch-Up Perspective, *Technology in Society*, vol. 68, p. 101890, https://doi.org/10.1016/j.techsoc.2022.101890

Fankhauser, S., Bowen, A., Calel, R., Dechezleprêtre, A., Grover, D., Rydge, J., & Sato, M. (2013). Who Will Win the Green Race? In Search of Environmental Competitiveness and Innovation, *Global Environmental Change*, vol. 23, no. 5, pp. 902-913, https://doi.org/10.1016/j.gloenvcha.2013.05.007

Frenken, K., Van Oort, F., & Verburg, T. (2007). Related Variety, Unrelated Variety and Regional Economic Growth, *Regional Studies*, vol. 41, no. 5, pp. 685-697, https://doi.org/10.1080/00343400601120296

Gerschenkron, A. (1962). *Economic Backwardness in Historical Perspective: A Book of Essays*,

Cambridge, MA: The Belknap Press of Harvard University Press
Giordani, P., Ferraro, M. B., & Martella, F. (2020). Introduction to Clustering, in P. Giordani, M. B. Ferraro & F. Martella (eds), *An Introduction to Clustering with R*, Singapore: Springer, pp. 3-5
Goldstein, J. E., Neimark, B., Garvey, B., & Phelps, J. (2023). Unlocking "Lock-in" and Path Dependency: A Review across Disciplines and Socio-Environmental Contexts, *World Development*, vol. 161, p. 106116, https://doi.org/10.1016/j.worlddev.2022.106116
Grabher, G. (1993). The Weakness of Strong Ties: The Lock-in of Regional Development in the Ruhr Area, in G. Grabher (ed.), *The Embedded Firm: On the Socio-Economics of Industrial Networks*, London: Routledge, pp. 255-277
Grabher, G., & Stark, D. (1997). Organizing Diversity: Evolutionary Theory, Network Analysis and Postsocialism, *Regional Studies*, vol. 31, no. 5, pp. 533-544, https://doi.org/10.1080/00343409750132315
Griffith, R., Redding, S., & Reenen, J. V. (2004). Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries, *Review of Economics and Statistics*, vol. 86, no. 4, pp. 883-895, https://www.jstor.org/stable/40042976
Guimarães, P. (2008). The Fixed Effects Negative Binomial Model Revisited, *Economics Letters*, vol. 99, no. 1, pp. 63-66, https://doi.org/10.1016/j.econlet.2007.05.030
Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2000). Market Value and Patent Citations: A First Look, National Bureau of Economic Research, working paper, no. 7741, https://www.nber.org/papers/w7741
Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2005). Market Value and Patent Citations, *RAND Journal of Economics*, vol. 36, no. 1, pp. 16-38, https://www.jstor.org/stable/1593752
Hassink, R. (2010). Regional resilience: A Promising Concept to Explain Differences in Regional Economic Adaptability?, *Cambridge Journal of Regions, Economy and Society*, vol. 3, no. 1, pp. 45-58, https://doi.org/10.1093/cjres/rsp033
Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R&D Relationship, *Econometrica*, vol. 52, no. 4, pp. 909-938, https://doi.org/10.2307/1911191
Henderson, R. M., & Clark, K. B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms, *Administrative Science Quarterly*, vol. 35, no. 1, pp. 9-30, https://doi.org/10.2307/2393549
Howitt, P. (2000). Endogenous Growth and Cross-Country Income Differences, *American Economic Review*, vol. 90, no. 4, pp. 829-846, https://www.jstor.org/stable/117310
Iammarino, S. (2005). An Evolutionary Integrated View of Regional Systems of Innovation: Concepts, Measures and Historical Perspectives, *European Planning Studies*, vol. 13, no. 4, pp. 497-519, https://doi.org/10.1080/09654310500107084
Jones, B. F. (2009). The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation

Getting Harder?, *The Review of Economic Studies*, vol. 76, no. 1, pp. 283-317, https://doi.org/10.1111/j.1467-937X.2008.00531.x

Klepper, S. (2010). The Origin and Growth of Industry Clusters: The Making of Silicon Valley and Detroit, *Journal of Urban Economics*, vol. 67, no. 1, pp. 15-32, https://doi.org/10.1016/j.jue.2009.09.004

Krammer, S. M. S. (2009). Drivers of National Innovation in Transition: Evidence from a Panel of Eastern European Countries, *Research Policy*, vol. 38, no. 5, pp. 845-860, https://doi.org/10.1016/j.respol.2009.01.022

Kuhn, T. S. (2012). *The Structure of Scientific Revolutions*, 4th edn, Chicago, IL: The University of Chicago Press

Landini, F., Lee, K., & Malerba, F. (2017). A History-Friendly Model of the Successive Changes in Industrial Leadership and the Catch-Up by Latecomers, *Research Policy*, vol. 46, no. 2, pp. 431-446, https://doi.org/10.1016/j.respol.2016.09.005

Lee, K. (2013). *Schumpeterian Analysis of Economic Catch-up: Knowledge, Path Creation, and the Middle-income Trap*, Cambridge: Cambridge University Press

Lee, K. (2024). Economics of Technology Cycle Time (TCT) and Catch-up by Latecomers: Micro-, Meso-, and Macro-Analyses and Implications, *Journal of Evolutionary Economics*, vol. 34, no. 2, pp. 319-349, https://doi.org/10.1007/s00191-024-00847-9

Lee, K., & Lim, C. (2001). Technological Regimes, Catching-up and Leapfrogging: Findings from the Korean Industries, *Research Policy*, vol. 30, no. 3, pp. 459-483, https://doi.org/10.1016/S0048-7333(00)00088-3

Lee, K., & Malerba, F. (2017). Catch-up Cycles and Changes in Industrial Leadership: Windows of Opportunity and Responses of Firms and Countries in the Evolution of Sectoral Systems, *Research Policy*, vol. 46, no. 2, pp. 338-351, https://doi.org/10.1016/j.respol.2016.09.006

Leonard-Barton, D. (1992). Core Capabilities and Core Rigidities: A Paradox in Managing New Product Development, *Strategic Management Journal*, vol. 13, no. S1, pp. 111-125, https://www.jstor.org/stable/2486355

Lerner, J., & Seru, A. (2022). The Use and Misuse of Patent Data: Issues for Finance and Beyond, *Review of Financial Studies*, vol. 35, no. 6, pp. 2667-2704, https://www.jstor.org/stable/48825284

Li, Y., Xu, J., & Lee, C. C. (2026). Evolution and Driving Mechanism of Green Transportation Multinational Cooperation Innovation Network, *Technology in Society*, vol. 86, p. 103294, https://doi.org/10.1016/j.techsoc.2026.103294

Lieberman, M. B., & Montgomery, D. B. (1988). First-Mover Advantages, *Strategic Management Journal*, vol. 9, no. S1, pp. 41-58, https://www.jstor.org/stable/2486211

Lissoni, F., Llerena, P., McKelvey, M., & Sanditov, B. (2008). Academic Patenting in Europe: New Evidence from the KEINS Database, *Research Evaluation*, vol. 17, no. 2, pp. 87-102, https://doi.org/10.3152/095820208X287171

Liu, L., Zhang, T., Avrin, A. P., & Wang, X. (2020). Is China's Industrial Policy Effective? An Empirical Study of the New Energy Vehicles Industry, *Technology in Society*, vol. 63, p. 101356, https://doi.org/10.1016/j.techsoc.2020.101356

Lundvall, B.-Å. (2016). *The Learning Economy and the Economics of Hope*, London: Anthem Press

Lundvall, B.-Å. (ed.) (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*, London: Pinter Publishers

Malerba, F., & Lee, K. (2021). An Evolutionary Perspective on Economic Catch-up by Latecomers, *Industrial and Corporate Change*, vol. 30, no. 4, pp. 986-1010, https://doi.org/10.1093/icc/dtab008

Martin, R., & Sunley, P. (2006). Path Dependence and Regional Economic Evolution, *Journal of Economic Geography*, vol. 6, no. 4, pp. 395-437, https://doi.org/10.1093/jeg/lbl012

Martin, R., & Sunley, P. (2015). On the Notion of Regional Economic Resilience: Conceptualization and Explanation, *Journal of Economic Geography*, vol. 15, no. 1, pp. 1–42, https://doi.org/10.1093/jeg/lbu015

Maskell, P., & Malmberg, A. (1999). Localised Learning and Industrial Competitiveness, *Cambridge Journal of Economics*, vol. 23, no. 2, pp. 167-185, https://doi.org/10.1093/cje/23.2.167

Mathews, J. A. (2006). Catch-up Strategies and the Latecomer Effect in Industrial Development, *New Political Economy*, vol. 11, no. 3, pp. 313-335, https://doi.org/10.1080/13563460600840142

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical Density Based Clustering, *Journal of Open Source Software*, vol. 2, no. 11, https://doi.org/10.21105/joss.00205

Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation, in *Proceedings of the 2014 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Philadelphia, PA: SIAM, pp. 839-847, https://doi.org/10.1137/1.9781611973440.96

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent Statistics as an Innovation Indicator, in B. H. Hall & N. Rosenberg (eds), *Handbook of the Economics of Innovation*, vol. 2, Amsterdam: Elsevier, pp. 1083-1127

Neffke, F., Henning, M., & Boschma, R. (2011). How do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions, *Economic Geography*, vol. 87, no. 3, pp. 237-265, https://doi.org/10.1111/j.1944-8287.2011.01121.x

Nelson, R. R. (2008). Factors Affecting the Power of Technological Paradigms, *Industrial and Corporate Change*, vol. 17, no. 3, pp. 485-497, https://doi.org/10.1093/icc/dtn010

Nelson, R. R., & Winter, S. G. (1982a). *An Evolutionary Theory of Economic Change*, Cambridge, MA: The Belknap Press of Harvard University Press

Nelson, R. R., & Winter, S. G. (1982b). The Schumpeterian Tradeoff Revisited, *The American Economic Review*, vol. 72, no. 1, pp. 114-132, https://www.jstor.org/stable/1808579

Noailly, J., & Smeets, R. (2015). Directing Technical Change from Fossil-Fuel to Renewable Energy

Innovation: An Application Using Firm-Level Patent Data, *Journal of Environmental Economics and Management*, vol. 72, pp. 15-37, https://doi.org/10.1016/j.jeem.2015.03.004

OECD. (2009). *OECD Patent Statistics Manual*, Paris: OECD Publishing

Park, G., Shin, J., & Park, Y. (2006). Measurement of Depreciation Rate of Technological Knowledge: Technology Cycle Time Approach, *Journal of Scientific & Industrial Research*, vol. 65, pp. 121-127

Park, J., & Lee, K. (2015). Do Latecomer Firms Rely on 'Recent' and 'Scientific' Knowledge More than Incumbent Firms Do? Convergence or Divergence in Knowledge Sourcing, *Asian Journal of Technology Innovation*, vol. 23, no. sup1, pp. 129-145, https://doi.org/10.1080/19761597.2015.1011261

Patel, P., & Pavitt, K. (1997). The Technological Competencies of the World's Largest Firms: Complex and Path-Dependent, but Not Much Variety, *Research Policy*, vol. 26, no. 2, pp. 141-156, https://doi.org/10.1016/S0048-7333(97)00005-X

Pavitt, K. (1985). Patent Statistics as Indicators of Innovative Activities: Possibilities and Problems, *Scientometrics*, vol. 7, no. 1-2, pp. 77-99, https://doi.org/10.1007/BF02020142

Pece, A. M., Simona, O. E. O., & Salisteanu, F. (2015). Innovation and Economic Growth: An Empirical Analysis for CEE Countries, *Procedia Economics and Finance*, vol. 26, pp. 461-467, https://doi.org/10.1016/S2212-5671(15)00874-6

Perez, C. (2002). *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*, Cheltenham: Edward Elgar Publishing

Perez, C., & Soete, L. (1988). Catching up in Technology: Entry Barriers and Windows of Opportunity, in G. Dosi, C. Freeman, R. Nelson, G. Silverberg & L. Soete (eds), *Technical Change and Economic Theory*, London: Pinter Publishers, pp. 458-479

Pike, A., Dawley, S., & Tomaney, J. (2010). Resilience, Adaptation and Adaptability, *Cambridge Journal of Regions, Economy and Society*, vol. 3, no. 1, pp. 59-70, https://doi.org/10.1093/cjres/rsq001

Porter, M. E. (1998). Location, Clusters, and the "New" Microeconomics of Competition, *Business Economics*, vol. 33, no. 1, pp. 7-13, https://www.jstor.org/stable/23487685

Ren, R., Wang, C., Sun, X., Liu, G., & Wang, Y. (2025). The Evolving National Technological Innovation Modes: Evidence from New Energy Vehicle Industry, *Asian Journal of Technology Innovation*, pp. 1-34, https://doi.org/10.1080/19761597.2025.2566461

Rosiello, A., & Maleki, A. (2021). A Dynamic Multi-Sector Analysis of Technological Catch-up: The Impact of Technology Cycle Times, Knowledge Base Complexity and Variety, *Research Policy*, vol. 50, no. 3, p. 104194, https://doi.org/10.1016/j.respol.2020.104194

Santarelli, E., Staccioli, J., & Vivarelli, M. (2023). Automation and Related Technologies: A Mapping of the New Knowledge Base, *The Journal of Technology Transfer*, vol. 48, pp. 779-813, https://doi.org/10.1007/s10961-021-09914-w

Santos Silva, J. M. C., & Tenreyro, S. (2006). The Log of Gravity, *The Review of Economics and Statistics*, vol. 88, no. 4, pp. 641-658, https://www.jstor.org/stable/40043025

Schmitt, G., Scott, J., Davis, A., & Utz, T. (2016). Patents and Progress; Intellectual Property Showing the Future of Electric Vehicles, *World Electric Vehicle Journal*, vol. 8, no. 3, pp. 635-645, https://doi.org/10.3390/wevj8030635

Shapere, D. (1964). The Structure of Scientific Revolutions, *The Philosophical Review*, vol. 73, no. 3, pp. 383-394, https://doi.org/10.2307/2183664

Shi, J., & Zhang, J. (2025). Short Technology Cycle Time and Firm Innovation: Evidence from China, *Research Policy*, vol. 54, no. 8, p. 105305, https://doi.org/10.1016/j.respol.2025.105305

Shi, J., Yan, W., Jiang, K., & Sadowski, B. (2026). Disruptive Innovation and Technological Catch-up in the NLP Sector: The Moderating Role of the Technology Cycle, *Journal of Evolutionary Economics*, vol. 36, no. 1, p. 12, https://doi.org/10.1007/s00191-025-00930-9

Simmie, J., & Martin, R. (2010). The Economic Resilience of Regions: Towards an Evolutionary Approach, *Cambridge Journal of Regions, Economy and Society*, vol. 3, no. 1, pp. 27-43, https://doi.org/10.1093/cjres/rsp029

Sinigaglia, T., Martins, M. E. S., & Siluk, J. C. M. (2022a). Technological Evolution of Internal Combustion Engine Vehicle: A Patent Data Analysis, *Applied Energy*, vol. 306, p. 118003, https://doi.org/10.1016/j.apenergy.2021.118003

Sinigaglia, T., Martins, M. E. S., & Siluk, J. C. M. (2022b). Technological Forecasting for Fuel Cell Electric Vehicle: A Comparison with Electric Vehicles and Internal Combustion Engine Vehicles, *World Patent Information*, vol. 71, p. 102152, https://doi.org/10.1016/j.wpi.2022.102152

Song, Y., Li, Y., Jiang, J., & Ye, B. (2025). The Industrial Prospect of Electric Vehicles—Time Delay Stochastic Evolutionary Game Evidence from the US, China, the EU, and Japan, *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1-17, https://doi.org/10.1057/s41599-025-05342-5

Tanner, A. (2014). Regional Branching Reconsidered: Emergence of the Fuel Cell Industry in European Regions, *Economic Geography*, vol. 90, no. 4, pp. 403-427, https://doi.org/10.1111/ecge.12055

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic Capabilities and Strategic Management, *Strategic Management Journal*, vol. 18, no. 7, pp. 509-533, https://www.jstor.org/stable/3088148

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus Corporate Patents: A Window on the Basicness of Invention, *Economics of Innovation and New Technology*, vol. 5, no. 1, pp. 19-50, https://doi.org/10.1080/10438599700000006

Tushman, M. L., & Anderson, P. (1986). Technological Discontinuities and Organizational Environments, *Administrative Science Quarterly*, vol. 31, no. 3, pp. 439-465, https://doi.org/10.2307/2392832

Unruh, G. C. (2000). Understanding Carbon Lock-in, *Energy Policy*, vol. 28, no. 12, pp. 817-830, https://doi.org/10.1016/S0301-4215(00)00070-7

Vergne, J.-P., & Durand, R. (2010). The Missing Link Between the Theory and Empirics of Path Dependence: Conceptual Clarification, Testability Issue, and Methodological Implications, *Journal of Management Studies*, vol. 47, no. 4, pp. 736-759, https://doi.org/10.1111/j.1467-6486.2009.00913.x

von Tunzelmann, N., Malerba, F., Nightingale, P., & Metcalfe, S. (2008). Technological Paradigms: Past, Present and Future, *Industrial and Corporate Change*, vol. 17, no. 3, pp. 467-484, https://doi.org/10.1093/icc/dtn012

Whitfield, L., & Wuttke, T. (2026). China's Technological Catch-up and Leapfrogging in Electric Vehicles: A Firm-Level Study of BYD and CATL, *Progress in Economic Geography*, vol. 4, p. 100054, https://doi.org/10.1016/j.peg.2025.100054

Wu, J. (2012). Cluster Analysis and K-Means Clustering: An Introduction, in J. Wu (ed.), *Advances in K-Means Clustering*, Berlin, Heidelberg: Springer, pp. 1-16

Wu, Y. (2011). Innovation and Economic Growth in China: Evidence at the Provincial Level, *Journal of the Asia Pacific Economy*, vol. 16, no. 2, pp. 129-142, https://doi.org/10.1080/13547860.2011.564752

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge, *Science*, vol. 316, no. 5827, pp. 1036-1039, https://doi.org/10.1126/science.1136099

Zahra, S. A., & George, G. (2002). Absorptive Capacity: A Review, Reconceptualization, and Extension, *The Academy of Management Review*, vol. 27, no. 2, pp. 185-203, https://www.jstor.org/stable/4134351

Zhao, S., Kim, S. Y., Wu, H., Yan, J., & Xiong, J. (2020). Closing the Gap: The Chinese Electric Vehicle Industry Owns the Road, *Journal of Business Strategy*, vol. 41, no. 5, pp. 3-14, https://doi.org/10.1108/JBS-03-2019-0059

Zhu, Z., & Li, H. (2025). Technological Catch-up: A New Measure and Patent-Based Evidence from China's Manufacturing Industries, *Research Policy*, vol. 54, p. 105299, https://doi.org/10.1016/j.respol.2025.105299

Appendix A

The following statement outlines how Artificial Intelligence tools were utilized in the completion of this degree project.

1. Tools: ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google).

2. Degree of Use: The core intellectual work, including the research design, critical analysis, and interpretation of results, was conducted by the author. AI tools were used in a supplementary capacity during several stages of the project.

Preparation and Idea Generation (Chapters 1 & 3): The core research questions and theoretical frameworks were developed by the author. Claude was used as a supplementary research tool to help identify foundational literature in fields such as Evolutionary Economic Geography and related innovation studies, which supported the author's broader exploration of the literature. Claude was also used to assist in organizing ideas into clearer and more coherent chapter structures and literature review frameworks. All references were selected and reviewed by the author.

Coding and Data Processing (Chapter 4): The author provided the underlying analytical logic and all mathematical specifications for the study's variables. Gemini served as a technical coding assistant to facilitate the implementation of these specifications. Specifically, Gemini assisted in drafting initial SQL queries to extract bibliographic records from the Google Patents Public Data database, executing Stata scripts for calculating key indicators such as Legacy Knowledge Accumulation (LKA) and relative catch-up dynamics (DTF and ITA), and developing Python code for the HDBSCAN algorithm. Critically, the author maintained full intellectual control over the research design. All substantive methodological decisions, including the selection of spatial algorithms, the calibration of hyperparameters, and the determination of knowledge depreciation rates and lag structures, were made independently by the author.

Language Editing and Formatting (All Chapters): The author drafted the entirety of the thesis independently. ChatGPT, Claude, and Gemini were subsequently employed to improve academic phrasing, correct grammatical errors, and enhance the overall flow and readability of the text. AI tools did not replace the author's own analysis, interpretation, or substantive research contributions.

Appendix B

This appendix describes how the patent datasets used in the study were constructed. Section B.1 explains the retrieval strategies used to identify Internal Combustion Engine and Electric Vehicle patents based on IPC classifications and keyword filters. Section B.2 presents the SQL queries used to retrieve patent records and bibliographic information from the Google Patents Public Data database for the citation network analysis.

B.1 Patent Classification Standards

Internal combustion engine (ICE) patents were identified following the approach proposed by Sinigaglia et al. (2022a), which combines IPC classifications with keyword filters to reduce unrelated industrial patents.

Table B. 1: The retrieval strategy used in ICE vehicles

Technology Category	Sub-System	Relevant IPC code	Refinement
1. Internal combustion engine		F02B, F02D, F02F, F02M, F02N, F02P	Internal combustion engine and (vehicle* or car or automobile*)
2. Advanced combustion modes	HCCI	F02B	HCCI or homogeneous charge compression ignition
	RCCI	F02D, F02M, F02B	RCCI or Reactivity controlled compression

	Octane on-demand		Octane on-demand
	Gasoline compression ignition		Gasoline compression ignition
	Alternative Fuels, gaseous fuels, bi-fuel	F02D041/38, F02M021/02, F02M037/00, F02M043/00, F02D041/00	Alternative fuel* or gaseous fuel* or bi-fuel*
3. Efficiency Optimization & Emission Control	Friction loss, heat management, and mechanical resistance	F02F	Friction loss or heat management or mechanical resistance
	Variable compression ratio	F02D013/02, F02D 015/00	Variable compression ratio
	EGR	F02D021/08, F02M026/00, F02D041/00	EGR or Exhaust gas recirculation
4. Air management & boosting	Super- and turbocharging	F02B027/00, F02B 029/00, F02B033/ 00, F02B037/00, F02B039/00, F02D 023/00, F02D041/ 00	Turbocharging or turbocharger or supercharger or supercharging
	Downsizing		Downsizing and internal combustion engine
	Variable valve actuation	F02D013/02, F01L 001/34, F01L009/ 02,	Variable valve actuation

		F01L009/04, F01L013/02, F01L 013/00	
5. Fuel injection & mixture formation	Injection strategy	F02D041/38, F02M045/00, F02M069/04	Injection strategy
	Direct injection	F02B023/06, F02B 023/10, F02D041/ 38, F02M047/00, F02M051/06, F02M055/00, F02M057/00, F02M059/00, F02M063/02	Direct injection
	Water injection	F02M	Water injection

Note: The patent search strategy presented here was developed with reference to Sinigaglia et al. (2022a).

The identification of electric vehicle patents strictly adopts the New Three Systems standard defined by the China National Intellectual Property Administration. This framework isolates pure battery electric vehicle components and completely excludes transitional hybrid technologies.

Table B. 2: The retrieval strategy used in EVs

Technology Category	Sub-System	Primary IPC code	Refinement
1. Whole Vehicle Manufacturing	Vehicle Body	B60K1*, B60L50/30, B60L50/40, B60L50/51, B60L50/52, B60L50/53, B60L50/60*, B60L50/90, B60L8*	Motor vehicles, excluding fuel-powered vehicles.
		B60K11*, B60K17*, B60K25*, B60K26*, B60K6*(excluding B60K6/08、B60K6/24、B60K6/32), B60K7*	Electric vehicle manufacturing, excluding fuel-powered vehicles.

		B62D21*, B62D31*, B65G47*, B62D65*	Electric vehicle and motor vehicle manufacturing, excluding fuel-powered vehicles.
2. Manufacturing of electric vehicle devices and components	Electric vehicle powertrain (Manufacturing of electric motors and engines)	H02K1*, H02K5*, H02K15*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
	(Manufacturing of electric motors and engines)	B60L*	Electric vehicles and automotive electric motors, electric drives, and batteries, excluding fuel-powered vehicles.
	(Electric traction)	B60L8*, B60L9*, B60L15*, B60L50*	Energy-saving, environmentally friendly, and green electric power.
	(Installation or arrangement of power or transmission units)	B60K1*(excluding B60K1/04), B60K6*(excluding B60K6/28, B60K6/30), B60K11*, B60K16*, B60K17*	Energy-saving, environmentally friendly, and green electric power.
	(Electric vehicle control)	B60W10*, B60W20*, B60W30*, B60W40*, B60W50*, B60W60*	Electric vehicles
		G05D1*, B60L15*, B60G17/0195, G05B19/00	Electric vehicle manufacturing, excluding fuel-powered vehicles.
	Manufacturing of energy storage	B60L58*, H01M50/249, H01M10/625	Excluding elderly mobility vehicles, motorcycles, other

	systems for electric vehicles		nonmotor vehicles, and fuel-powered vehicles.
		H01M10*, H01M12*, H01M4/04, H01M4/13*, H01M4/14*, H01M4/24*, H01M50*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
	Manufacturing of electric vehicle parts and components	B60L1*, B60L15*, B60L3*, B60L5*, B60L7*, B60W20*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
		H02J7*	Motor vehicles, excluding fuel-powered vehicles.
		B60L50/60*	Charging devices.
		B62D5/04, F16H3*, F16H59*, F16H61*, F16H63*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
		B60G*, B60K20*, B60Q5*, B60R16*, B60T13*, B60T17*, B60T7*, B60T8*, B60W10*, B60W30*, B60W40*, B60W50*, B60W60*, B60L7*	Electric vehicles and automotive electric motors, electric drive units, and vehicle batteries, excluding fuel-powered vehicles.
3. Manufacture of facilities related to electric vehicles	Manufacturing of power supply equipment	B60L53*, B60L55*	
		H02J7*	Charging devices.
	Manufacture of testing and diagnostic equipment.	G01R31/34, G01R27*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.

		G01L3*, G01M13*, G01M15*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
		G01M17*	Electric vehicles and automotive electric motors, electric drive units, and vehicle batteries, excluding fuel-powered vehicles.
	Manufacture of other related supporting facilities.	C08K3/04, F04C18/02, F04C18/356	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
4. Services related to electric vehicles	Electric vehicle maintenance services	B23K37*	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.
		B60S5*	Electric vehicles and automotive electric motors, electric drive units, and vehicle batteries, excluding fuel-powered vehicles.
	Electric vehicle charging and battery swapping services	B60L53/80, B60S5/06, B60L53/10, B60L53/126, B60L53/16, B60L53/18, B60L53/22, B60L53/24, B60L53/30, B60L53/302, B60L53/31, B60L53/34, B60L53/35, B60L53/38, B60L53/50, B60L53/57, B60L53/60, B60L53/66, B60L53/67	Electric vehicles and other motor vehicle types, excluding fuel-powered vehicles.

Note: The patent search strategy is developed with reference to the CNIPA (2024) document.

B.2 SQL Data Extraction Script

B.2.1 SQL for ICE Data Extraction

The following SQL query extracts patent families related to Internal Combustion Engine (ICE) technologies from the Google Patents Public Data database. The query combines publication records with patent classification information and limits the sample to patents published between 2000 and 2024.

Listing B.1 SQL query for base data extraction of ICE patents

```
WITH JoinedData AS (
  -- === Step 1: Base data extraction and text field concatenation ===
  SELECT
    pubs.publication_number,
    pubs.family_id,
    pubs.country_code,
    pubs.kind_code,

    pubs.filing_date,
    pubs.publication_date,
    pubs.grant_date,
    pubs.priority_date,

    pubs.assignee_harmonized,
    pubs.assignee,
    pubs.inventor_harmonized,
    pubs.inventor,

    pubs.ipc,
    pubs.cpc,
    pubs.citation,

    pubs.title_localized,
    pubs.abstract_localized,

    LOWER(CONCAT(
      IFNULL(res.title, IFNULL((SELECT text FROM UNNEST(pubs.title_localized) WHERE
language = 'en' LIMIT 1), '')),
      ' ',
      IFNULL(res.abstract, IFNULL((SELECT text FROM UNNEST(pubs.abstract_localized)
WHERE language = 'en' LIMIT 1), '')),
      ' ',
      IFNULL((SELECT text FROM UNNEST(pubs.title_localized) LIMIT 1), '')
    )) AS text_search_field

  FROM
    `patents-public-data.patents.publications` AS pubs
  LEFT JOIN
    `patents-public-data.google_patents_research.publications` AS res
```

```
  ON
    pubs.publication_number = res.publication_number
  WHERE
    pubs.country_code = 'WO'
    AND pubs.publication_date BETWEEN 20000101 AND 20241231
),

FilteredData AS (
  SELECT * FROM JoinedData
  WHERE
    -- === Step 2: Assignee geographical filtering ===
    (
      ARRAY_LENGTH(assignee_harmonized) > 0
      AND EXISTS (
        SELECT 1 FROM UNNEST(assignee_harmonized) AS a
        WHERE a.country_code IN (
          'CN', 'EP', 'AL', 'AT', 'BE', 'BG', 'CH', 'CY', 'CZ', 'DE', 'DK', 'EE',
'ES', 'FI',
          'FR', 'GB', 'GR', 'HR', 'HU', 'IE', 'IS', 'IT', 'LI', 'LT', 'LU', 'LV',
'MC',
          'MK', 'MT', 'ME', 'NL', 'NO', 'PL', 'PT', 'RO', 'RS', 'SE', 'SI', 'SK',
'SM', 'TR'
        )
      )
    )
    OR (ARRAY_LENGTH(assignee_harmonized) = 0)
),

LogicLayer AS (
  SELECT
    *,
    -- Preprocess IPC array into a clean searchable string
    (SELECT STRING_AGG(REGEXP_REPLACE(i.code, r'[^A-Za-z0-9]', ''), ' ') FROM
UNNEST(ipc) AS i) AS clean_ipc_string,

    --
==================================================================================
=
    -- Technological Classification Rules (Upgraded with robust fuzzy matching
architecture)
    --
==================================================================================
=

    -- 1. General Internal Combustion Engine (ICE)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i WHERE REGEXP_CONTAINS(i.code,
r'^(F02B|F02D|F02F|F02M|F02N|F02P)'))
      AND REGEXP_CONTAINS(text_search_field, r'(combustion|engine|motor)')
      AND REGEXP_CONTAINS(text_search_field,
r'(vehicle|car|auto|truck|bus|powertrain)')
    ) AS is_01_ICE,

    -- 2. Water Injection
    -- ☑ Regex Refinement: Strict boundary constraint applied to ensure match
begins precisely with class F02M.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i WHERE REGEXP_CONTAINS(i.code, r'^F02M'))
      AND REGEXP_CONTAINS(text_search_field, r'water') AND
REGEXP_CONTAINS(text_search_field, r'injection')
```

```
    ) AS is_02_WaterInjection,

    -- 3. Homogeneous Charge Compression Ignition (HCCI)
    -- ☑ Regex Refinement: Strict boundary constraint applied for class F02B.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i WHERE REGEXP_CONTAINS(i.code, r'^F02B'))
      AND REGEXP_CONTAINS(text_search_field, r'(hcci|homogeneous charge)')
    ) AS is_03_HCCI,

    -- 4. Reactivity Controlled Compression Ignition (RCCI)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i WHERE REGEXP_CONTAINS(i.code,
r'^(F02D|F02M|F02B)'))
      AND REGEXP_CONTAINS(text_search_field, r'(rcci|reactivity controlled)')
    ) AS is_04_RCCI,

    -- 5. Friction Reduction and Thermal Management
    -- ☑ Regex Refinement: Safely isolate engine construction parameters starting
with F02F.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i WHERE REGEXP_CONTAINS(i.code, r'^F02F'))
      AND REGEXP_CONTAINS(text_search_field, r'(friction|heat management|mechanical
resistance|thermal)')
    ) AS is_05_Friction,

    -- 6. Injection Strategy
    -- ☑ Regex Refinement: Upgraded to \D*0* structure to guarantee accurate
extraction of specific subgroups (e.g., strictly capturing F02D 41/38).
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02D\D*0*41\D+0*38')
          OR REGEXP_CONTAINS(i.code, r'^F02M\D*0*45\D+0*00')
          OR REGEXP_CONTAINS(i.code, r'^F02M\D*0*69\D+0*04'))
      AND REGEXP_CONTAINS(text_search_field, r'injection') AND
REGEXP_CONTAINS(text_search_field, r'(strategy|control|method|timing)')
    ) AS is_06_InjectionStrategy,

    -- 8. Exhaust Gas Recirculation (EGR)
    -- ☑ Regex Refinement: Refined fuzzy matching algorithms to capture precise
EGR classifications without interference.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02D\D*0*21\D+0*08')
          OR REGEXP_CONTAINS(i.code, r'^F02M\D*0*26\D+0*00')
          OR REGEXP_CONTAINS(i.code, r'^F02D\D*0*41\D+0*00'))
      AND REGEXP_CONTAINS(text_search_field, r'(egr|exhaust gas recirculation)')
    ) AS is_08_EGR,

    -- 9. Supercharging and Turbocharging
    -- ☑ Regex Refinement: Target specific supercharging main groups F02B 27, 29,
33, 37, and 39 utilizing boundary markers.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02B\D*0*(27|29|33|37|39)\D+0*00')
          OR REGEXP_CONTAINS(i.code, r'^F02D\D*0*(23|41)\D+0*00'))
      AND REGEXP_CONTAINS(text_search_field, r'(turbo|supercharg)')
    ) AS is_09_Turbo,

    -- 10. Downsizing (Text-based semantic matching)
    (REGEXP_CONTAINS(text_search_field, r'downsizing') AND
REGEXP_CONTAINS(text_search_field, r'(combustion engine|ice)')
    ) AS is_10_Downsizing,

    -- 11. Variable Compression Ratio (VCR)
```

```
    -- ☑ Regex Refinement: Strictly match specified engine control subclasses F02D
13/02 and 15/00.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02D\D*0*(13\D+0*02|15\D+0*00)'))
      AND REGEXP_CONTAINS(text_search_field, r'variable') AND
REGEXP_CONTAINS(text_search_field, r'compression')
    ) AS is_11_VCR,

    -- 12. Variable Valve Actuation (VVA)
    -- ☑ Regex Refinement: Safely isolate target valve control architectures
distributed across classes F02D and F01L.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02D\D*0*13\D+0*02')
          OR REGEXP_CONTAINS(i.code,
r'^F01L\D*0*(1\D+0*34|9\D+0*02|9\D+0*04|13\D+0*02|13\D+0*00)'))
      AND REGEXP_CONTAINS(text_search_field, r'variable') AND
REGEXP_CONTAINS(text_search_field, r'valve')
    ) AS is_12_VVA,

    -- 13. Alternative Fuels
    -- ☑ Regex Refinement: Target specific alternative fuel delivery mechanisms
including subgroups 21/02, 37/00, and 43/00 within F02M.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02D\D*0*41\D+0*38')
          OR REGEXP_CONTAINS(i.code, r'^F02M\D*0*(21\D+0*02|37\D+0*00|43\D+0*00)')
          OR REGEXP_CONTAINS(i.code, r'^F02D\D*0*41\D+0*00'))
      AND REGEXP_CONTAINS(text_search_field, r'(alternative fuel|gaseous|bi-
fuel|hydrogen|methanol|ethanol)')
    ) AS is_13_AltFuel,

    -- 14. Direct Injection
    -- ☑ Regex Refinement: Accurately capture direct injection components
utilizing defined constraints across F02B and F02M.
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^F02B\D*0*23\D+0*(06|10)')
          OR REGEXP_CONTAINS(i.code, r'^F02D\D*0*41\D+0*38')
          OR REGEXP_CONTAINS(i.code,
r'^F02M\D*0*(47\D+0*00|51\D+0*06|55\D+0*00|57\D+0*00|59\D+0*00|63\D+0*02)'))
      AND REGEXP_CONTAINS(text_search_field, r'direct') AND
REGEXP_CONTAINS(text_search_field, r'injection')
    ) AS is_14_DirectInjection,

    -- 15. Octane On-Demand (Text-based semantic matching)
    (REGEXP_CONTAINS(text_search_field, r'octane') AND
REGEXP_CONTAINS(text_search_field, r'(demand|variable)')
    ) AS is_15_Octane,

    -- 16. Gasoline Compression Ignition (Text-based semantic matching)
    ((REGEXP_CONTAINS(text_search_field, r'gasoline') AND
REGEXP_CONTAINS(text_search_field, r'compression ignition')) OR
REGEXP_CONTAINS(text_search_field, r'\bgci\b')
    ) AS is_16_GCI

  FROM FilteredData
)

-- === Final Output Extraction ===
SELECT
```

```
  publication_number,
  family_id,
  filing_date,
  publication_date,
  grant_date,
  priority_date AS earliest_priority_date,

  (SELECT STRING_AGG(name, ' | ') FROM UNNEST(assignee_harmonized)) AS
assignee_harmonized_name,
  (SELECT STRING_AGG(country_code, ' | ') FROM UNNEST(assignee_harmonized)) AS
assignee_harmonized_country,
  (SELECT STRING_AGG(a, ' | ') FROM UNNEST(assignee) AS a) AS assignee_list,
  (SELECT STRING_AGG(a, ' | ') FROM UNNEST(assignee) AS a) AS
assignee_address_proxy,

  (SELECT STRING_AGG(name, ' | ') FROM UNNEST(inventor_harmonized)) AS
inventor_harmonized_name,
  (SELECT STRING_AGG(country_code, ' | ') FROM UNNEST(inventor_harmonized)) AS
inventor_harmonized_country,
  (SELECT STRING_AGG(inv, ' | ') FROM UNNEST(inventor) AS inv) AS inventor_list,

  (SELECT STRING_AGG(code, ', ') FROM UNNEST(ipc)) AS ipc_codes,
  (SELECT STRING_AGG(code, ', ') FROM UNNEST(cpc)) AS cpc_codes,

  (SELECT STRING_AGG(publication_number, ', ') FROM (SELECT publication_number FROM
UNNEST(citation) LIMIT 20)) AS cited_patents,
  ARRAY_LENGTH(citation) AS citation_count,

  (SELECT text FROM UNNEST(title_localized) WHERE language = 'en' LIMIT 1) AS
title_en,

  -- Compile array of all matched ICE category labels
  ARRAY_TO_STRING(
    [
      IF(is_01_ICE, '01. Internal Combustion Engine General', NULL),
      IF(is_02_WaterInjection, '02. Water injection', NULL),
      IF(is_03_HCCI, '03. HCCI', NULL),
      IF(is_04_RCCI, '04. RCCI', NULL),
      IF(is_05_Friction, '05. Friction loss/heat management', NULL),
      IF(is_06_InjectionStrategy, '06. Injection strategy', NULL),
      IF(is_08_EGR, '08. EGR', NULL),
      IF(is_09_Turbo, '09. Super- and turbocharging', NULL),
      IF(is_10_Downsizing, '10. Downsizing', NULL),
      IF(is_11_VCR, '11. Variable compression ratio', NULL),
      IF(is_12_VVA, '12. Variable valve actuation', NULL),
      IF(is_13_AltFuel, '13. Alternative Fuels', NULL),
      IF(is_14_DirectInjection, '14. Direct injection', NULL),
      IF(is_15_Octane, '15. Octane on-demand', NULL),
      IF(is_16_GCI, '16. Gasoline compression ignition', NULL)
    ], ' | '
  ) AS matched_technologies

FROM LogicLayer
WHERE
  (is_01_ICE OR is_02_WaterInjection OR is_03_HCCI OR is_04_RCCI OR is_05_Friction
OR
   is_06_InjectionStrategy OR is_08_EGR OR is_09_Turbo OR
   is_10_Downsizing OR is_11_VCR OR is_12_VVA OR is_13_AltFuel OR
```

```
  is_14_DirectInjection OR is_15_Octane OR is_16_GCI)

ORDER BY filing_date DESC
```

B.2.2 SQL for EV Data Extraction

This section presents the SQL query used to identify Electric Vehicle patents. The query follows the same basic data structure as the ICE extraction process but applies a different set of IPC classifications and keyword filters designed for EV-related technologies.

The classification framework follows the "New Three Systems" standard published by the China National Intellectual Property Administration, which focuses on battery electric vehicle technologies while excluding hybrid transition technologies.

Listing B.2 SQL query for base data extraction of EV patents

```
WITH JoinedData AS (
  -- ... (Step 1: Base data extraction logic remains completely unchanged) ...
  SELECT
    pubs.publication_number,
    pubs.family_id,
    pubs.country_code,
    pubs.kind_code,
    pubs.filing_date,
    pubs.publication_date,
    pubs.grant_date,
    pubs.priority_date,
    pubs.assignee_harmonized,
    pubs.assignee,
    pubs.inventor_harmonized,
    pubs.inventor,
    pubs.ipc,
    pubs.cpc,
    pubs.citation,
    pubs.title_localized,
    pubs.abstract_localized,

    LOWER(CONCAT(
      IFNULL(res.title, IFNULL((SELECT text FROM UNNEST(pubs.title_localized) WHERE
language = 'en' LIMIT 1), '')),
      ' ',
      IFNULL(res.abstract, IFNULL((SELECT text FROM UNNEST(pubs.abstract_localized)
WHERE language = 'en' LIMIT 1), '')),
      ' ',
      IFNULL((SELECT text FROM UNNEST(pubs.title_localized) LIMIT 1), '')
    )) AS text_search_field

  FROM
    `patents-public-data.patents.publications` AS pubs
  LEFT JOIN
    `patents-public-data.google_patents_research.publications` AS res
```

```
  ON
    pubs.publication_number = res.publication_number
  WHERE
    pubs.country_code = 'WO'
    AND pubs.publication_date BETWEEN 20000101 AND 20241231
),

FilteredData AS (
  SELECT * FROM JoinedData
  WHERE
    (
      ARRAY_LENGTH(assignee_harmonized) > 0
      AND EXISTS (
        SELECT 1 FROM UNNEST(assignee_harmonized) AS a
        WHERE a.country_code IN (
          'CN', 'EP', 'AL', 'AT', 'BE', 'BG', 'CH', 'CY', 'CZ', 'DE', 'DK', 'EE',
'ES', 'FI',
          'FR', 'GB', 'GR', 'HR', 'HU', 'IE', 'IS', 'IT', 'LI', 'LT', 'LU', 'LV',
'MC',
          'MK', 'MT', 'ME', 'NL', 'NO', 'PL', 'PT', 'RO', 'RS', 'SE', 'SI', 'SK',
'SM', 'TR'
        )
      )
    )
    OR (ARRAY_LENGTH(assignee_harmonized) = 0)
),

LogicLayer AS (
  SELECT
    *,
    (SELECT STRING_AGG(REGEXP_REPLACE(i.code, r'[^A-Za-z0-9]', ''), ' ') FROM
UNNEST(ipc) AS i) AS clean_ipc_string,

    --
==================================================================================
=
    -- 1. Whole Vehicle Manufacturing
    --
==================================================================================
=

    -- Cat1-Row1: Vehicle Body (Product)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
      WHERE (
        -- ☑ Regex Refinement: Applied \D*0* to accurately isolate main group
B60L 8 (preventing false matches like B60L 58) and used \D+0* for strict subgroup
separation.
        REGEXP_CONTAINS(i.code,
r'^(B60L\D*0*(8|50\D+0*30|50\D+0*40|50\D+0*51|50\D+0*52|50\D+0*53|50\D+0*60|50\D+0*
90))')
        -- ☑ Regex Refinement: Safely match main group B60K 1.
        OR REGEXP_CONTAINS(i.code, r'^B60K\D*0*1')
      ))
      AND REGEXP_CONTAINS(text_search_field,
r'(vehicle|car|automotive|motor|automobile)')
      AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat1_Row1,
```

```
    -- Cat1-Row2: Vehicle Body (Mfg)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE (
         -- ☑ Regex Refinement: Added \D*0* for precise main group isolation.
         REGEXP_CONTAINS(i.code, r'^B60K\D*0*(11|17|25|26|7)')
         -- ☑ Regex Refinement: Match B60K 6 while explicitly excluding specific
subgroups (08, 24, 32).
         OR (REGEXP_CONTAINS(i.code, r'^B60K\D*0*6') AND NOT
REGEXP_CONTAINS(i.code, r'B60K\D*0*6\D+0*(08|24|32)'))
       ))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new energy)')
       AND REGEXP_CONTAINS(text_search_field,
r'(manufactur|assembl|process|method|production)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat1_Row2,

    -- Cat1-Row3: Vehicle Body (Assembly)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE (
         -- ☑ Regex Refinement: Added \D*0* for precise main group isolation.
         REGEXP_CONTAINS(i.code, r'^(B62D\D*0*(21|31|65)|B65G\D*0*47)')
       ))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new energy)')
       AND REGEXP_CONTAINS(text_search_field,
r'(manufactur|assembl|process|method|production)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat1_Row3,

    --
==================================================================================
=
    -- 2. Manufacturing of electric vehicle devices and components
    --
==================================================================================
=

    -- Cat2-Row1: Powertrain (H02K)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Safely capture main groups H02K 1, 5, and 15.
       WHERE REGEXP_CONTAINS(i.code, r'^H02K\D*0*(1|5|15)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row1,

    -- Cat2-Row2: Powertrain (B60L*)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^B60L')) -- Broad class B60L requires no
structural modification.
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|electric
motor|electric drive|battery)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row2,
```

```
    -- Cat2-Row3: Electric Traction (B60L8...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Safely isolate B60L 8, 9, 15, and 50 using robust
boundary parameters.
       WHERE REGEXP_CONTAINS(i.code, r'^B60L\D*0*(8|9|15|50)'))
       AND REGEXP_CONTAINS(text_search_field, r'(energy saving|environmentally
friendly|green|efficiency)')
    ) AS is_Cat2_Row3,

    -- Cat2-Row4: Arrangement (B60K...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE (
         -- ☑ Regex Refinement: Match B60K 1 but explicitly exclude subgroup /04.
         (REGEXP_CONTAINS(i.code, r'^B60K\D*0*1') AND NOT REGEXP_CONTAINS(i.code,
r'B60K\D*0*1\D+0*04')) OR
         -- ☑ Regex Refinement: Match B60K 6 but explicitly exclude subgroups /28
and /30.
         (REGEXP_CONTAINS(i.code, r'^B60K\D*0*6') AND NOT REGEXP_CONTAINS(i.code,
r'B60K\D*0*6\D+0*(28|30)')) OR
         REGEXP_CONTAINS(i.code, r'^B60K\D*0*(11|16|17)')
       ))
       AND REGEXP_CONTAINS(text_search_field, r'(energy saving|environmentally
friendly|green|efficiency)')
    ) AS is_Cat2_Row4,

    -- Cat2-Row5: EV Control (B60W)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Safely capture B60W main groups 10 through 60.
       WHERE REGEXP_CONTAINS(i.code, r'^B60W\D*0*(10|20|30|40|50|60)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new energy
vehicle)')
    ) AS is_Cat2_Row5,

    -- Cat2-Row6: EV Control (Misc)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Target specific control architectures across
distinct subclasses (e.g., G05D 1).
       WHERE REGEXP_CONTAINS(i.code, r'^(G05D\D*0*1|B60L\D*0*15|G05B\D*0*19)') OR
REGEXP_CONTAINS(i.code, r'B60G\D*0*17\D+0*0195'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev)')
       AND REGEXP_CONTAINS(text_search_field, r'(manufactur|assembl|produc)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row6,

    -- Cat2-Row7: Energy Storage (Specific)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Precisely target designated battery storage
subgroups (e.g., H01M 50/249).
       WHERE REGEXP_CONTAINS(i.code,
r'^(B60L\D*0*58|H01M\D*0*50\D+0*249|H01M\D*0*10\D+0*625)'))
       AND NOT REGEXP_CONTAINS(text_search_field,
r'(elderly|motorcycle|scooter|wheelchair|non[- ]?motor|bicycle|tricycle|internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row7,

    -- Cat2-Row8: Energy Storage (General H01M)
```

```
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE (
         -- ☑ Regex Refinement: Target general electrochemical storage classes
H01M 10, 12, and 50.
         REGEXP_CONTAINS(i.code, r'^H01M\D*0*(10|12|50)')
         -- ☑ Regex Refinement: Target specialized subgroups under H01M 4 (04, 13,
14, 24).
         OR REGEXP_CONTAINS(i.code, r'^H01M\D*0*4\D+0*(04|13|14|24)')
       ))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|motor
vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row8,

    -- Cat2-Row9: Parts (B60L1... B60W20)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Safely capture B60L operational components (1, 3, 5,
7, 15).
       WHERE REGEXP_CONTAINS(i.code, r'^(B60L\D*0*(1|15|3|5|7)|B60W\D*0*20)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|motor
vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row9,

    -- Cat2-Row10: Parts (H02J7)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Accurately isolate charging circuit class H02J 7.
       WHERE REGEXP_CONTAINS(i.code, r'^H02J\D*0*7'))
       AND REGEXP_CONTAINS(text_search_field, r'(motor
vehicle|vehicle|automotive)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row10,

    -- Cat2-Row11: Parts (B60L50/60)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Explicit boundary match for subgroup B60L 50/60.
       WHERE REGEXP_CONTAINS(i.code, r'B60L\D*0*50\D+0*60'))
       AND REGEXP_CONTAINS(text_search_field, r'(charging device|charger)')
    ) AS is_Cat2_Row11,

    -- Cat2-Row12: Parts (B62D5...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       -- ☑ Regex Refinement: Added \D*0* to ensure precise main group matching
across steering and transmission classes.
       WHERE REGEXP_CONTAINS(i.code,
r'^(B62D\D*0*5\D+0*04|F16H\D*0*(3|59|61|63))'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|motor
vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row12,

    -- Cat2-Row13: Parts (B60G...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
```

```
      -- ☑ Regex Refinement: Robust parsing structure for an extensive array of
relevant automotive subcomponents.
       WHERE REGEXP_CONTAINS(i.code,
r'^(B60G|B60K\D*0*20|B60Q\D*0*5|B60R\D*0*16|B60T\D*0*(13|17|7|8)|B60W\D*0*(10|30|40
|50|60)|B60L\D*0*7)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|electric
motor|electric drive|battery)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat2_Row13,

    --
====================================================================================
=
    -- 3. Manufacture of facilities related to electric vehicles
    --
====================================================================================
=

    -- Cat3-Row1: Power Supply (B60L53...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^B60L\D*0*(53|55)'))
       AND REGEXP_CONTAINS(text_search_field, r'(charging|power supply|electric
vehicle|bev)')
    ) AS is_Cat3_Row1,

    -- Cat3-Row2: Power Supply (H02J7)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^H02J\D*0*7'))
       AND REGEXP_CONTAINS(text_search_field, r'(charging device|charger|charging
station)')
       AND REGEXP_CONTAINS(text_search_field, r'(vehicle|car|automotive)')
    ) AS is_Cat3_Row2,

    -- Cat3-Row3: Testing (G01R)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^(G01R\D*0*31\D+0*34|G01R\D*0*27)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat3_Row3,

    -- Cat3-Row4: Testing (G01L...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^(G01L\D*0*3|G01M\D*0*(13|15))'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat3_Row4,

    -- Cat3-Row5: Testing (G01M17)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^G01M\D*0*17'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|electric
motor|electric drive|battery)')
```

```
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat3_Row5,

    -- Cat3-Row6: Supporting (C08K...)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code,
r'^(C08K\D*0*3\D+0*04|F04C\D*0*18\D+0*02|F04C\D*0*18\D+0*356)'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat3_Row6,

    --
=================================================================================
=
    -- 4. Services related to electric vehicles
    --
=================================================================================
=

    -- Cat4-Row1: Maintenance (B23K37)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^B23K\D*0*37'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat4_Row1,

    -- Cat4-Row2: Maintenance (B60S5)
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE REGEXP_CONTAINS(i.code, r'^B60S\D*0*5'))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|electric
motor|electric drive|battery)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat4_Row2,

    -- Cat4-Row3: Charging Services
    (EXISTS(SELECT 1 FROM UNNEST(ipc) AS i
       WHERE (
         REGEXP_CONTAINS(i.code, r'^B60S\D*0*5\D+0*06') OR
         -- ☑ Regex Refinement: Comprehensive mapping of charging infrastructure
subgroups.
         REGEXP_CONTAINS(i.code,
r'^B60L\D*0*53\D+0*(80|10|126|16|18|22|24|30|302|31|34|35|38|50|57|60|66|67)')
       ))
       AND REGEXP_CONTAINS(text_search_field, r'(electric vehicle|bev|new
energy|motor vehicle)')
       AND NOT REGEXP_CONTAINS(text_search_field, r'(internal
combustion|gasoline|diesel|fuel engine|fuel[- ]?powered)')
    ) AS is_Cat4_Row3

  FROM FilteredData
)
```

```
-- === Final Output ===
SELECT
  publication_number,
  family_id,
  filing_date,
  publication_date,
  grant_date,
  priority_date AS earliest_priority_date,

  (SELECT STRING_AGG(name, ' | ') FROM UNNEST(assignee_harmonized)) AS
assignee_harmonized_name,
  (SELECT STRING_AGG(country_code, ' | ') FROM UNNEST(assignee_harmonized)) AS
assignee_harmonized_country,
  (SELECT STRING_AGG(a, ' | ') FROM UNNEST(assignee) AS a) AS assignee_list,
  (SELECT STRING_AGG(a, ' | ') FROM UNNEST(assignee) AS a) AS
assignee_address_proxy,

  (SELECT STRING_AGG(name, ' | ') FROM UNNEST(inventor_harmonized)) AS
inventor_harmonized_name,
  (SELECT STRING_AGG(inv, ' | ') FROM UNNEST(inventor) AS inv) AS inventor_list,

  (SELECT STRING_AGG(code, ', ') FROM UNNEST(ipc)) AS ipc_codes,
  (SELECT STRING_AGG(code, ', ') FROM UNNEST(cpc)) AS cpc_codes,

  (SELECT STRING_AGG(publication_number, ', ') FROM (SELECT publication_number FROM
UNNEST(citation) LIMIT 20)) AS cited_patents,
  ARRAY_LENGTH(citation) AS citation_count,

  (SELECT text FROM UNNEST(title_localized) WHERE language = 'en' LIMIT 1) AS
title_en,

  -- Compile array of all matched category labels
  ARRAY_TO_STRING(
    [
      IF(is_Cat1_Row1, 'Cat1-Row1 (Body-Prod)', NULL),
      IF(is_Cat1_Row2, 'Cat1-Row2 (Body-Mfg)', NULL),
      IF(is_Cat1_Row3, 'Cat1-Row3 (Body-Assy)', NULL),
      IF(is_Cat2_Row1, 'Cat2-Row1 (Motor-H02K)', NULL),
      IF(is_Cat2_Row2, 'Cat2-Row2 (Motor-B60L)', NULL),
      IF(is_Cat2_Row3, 'Cat2-Row3 (Traction)', NULL),
      IF(is_Cat2_Row4, 'Cat2-Row4 (Arrange)', NULL),
      IF(is_Cat2_Row5, 'Cat2-Row5 (Control-B60W)', NULL),
      IF(is_Cat2_Row6, 'Cat2-Row6 (Control-Misc)', NULL),
      IF(is_Cat2_Row7, 'Cat2-Row7 (Storage-Spec)', NULL),
      IF(is_Cat2_Row8, 'Cat2-Row8 (Storage-Gen)', NULL),
      IF(is_Cat2_Row9, 'Cat2-Row9 (Parts-1)', NULL),
      IF(is_Cat2_Row10, 'Cat2-Row10 (Parts-H02J)', NULL),
      IF(is_Cat2_Row11, 'Cat2-Row11 (Parts-L50/60)', NULL),
      IF(is_Cat2_Row12, 'Cat2-Row12 (Parts-2)', NULL),
      IF(is_Cat2_Row13, 'Cat2-Row13 (Parts-3)', NULL),
      IF(is_Cat3_Row1, 'Cat3-Row1 (Supply-B60L)', NULL),
      IF(is_Cat3_Row2, 'Cat3-Row2 (Supply-H02J)', NULL),
      IF(is_Cat3_Row3, 'Cat3-Row3 (Test-G01R)', NULL),
      IF(is_Cat3_Row4, 'Cat3-Row4 (Test-G01L/M)', NULL),
      IF(is_Cat3_Row5, 'Cat3-Row5 (Test-G01M17)', NULL),
      IF(is_Cat3_Row6, 'Cat3-Row6 (Supporting)', NULL),
      IF(is_Cat4_Row1, 'Cat4-Row1 (Maint-B23K)', NULL),
      IF(is_Cat4_Row2, 'Cat4-Row2 (Maint-B60S)', NULL),
```

```
      IF(is_Cat4_Row3, 'Cat4-Row3 (Service-Charge)', NULL)
    ], ' | '
  ) AS matched_rows

FROM LogicLayer
WHERE
  (is_Cat1_Row1 OR is_Cat1_Row2 OR is_Cat1_Row3 OR
   is_Cat2_Row1 OR is_Cat2_Row2 OR is_Cat2_Row3 OR is_Cat2_Row4 OR is_Cat2_Row5 OR
is_Cat2_Row6 OR is_Cat2_Row7 OR is_Cat2_Row8 OR is_Cat2_Row9 OR is_Cat2_Row10 OR
is_Cat2_Row11 OR is_Cat2_Row12 OR is_Cat2_Row13 OR
   is_Cat3_Row1 OR is_Cat3_Row2 OR is_Cat3_Row3 OR is_Cat3_Row4 OR is_Cat3_Row5 OR
is_Cat3_Row6 OR
   is_Cat4_Row1 OR is_Cat4_Row2 OR is_Cat4_Row3)

ORDER BY filing_date DESC
```

B.2.3 Searching for Patent Fields

After constructing the ICE and EV patent datasets, additional bibliographic fields were retrieved for the citation network analysis. The following query matches cited patent identifiers with the global Google Patents database and extracts key metadata, including filing dates, assignee information, inventor information, and patent classification codes.

Listing B.3 SQL query for detailed patent field extraction from citation networks

```
SELECT
  -- Step 1 Retain the original identifier from the source citation table to ensure
traceability.
  t.cited_patents AS input_id,

  -- Step 2 Retrieve the standardized Google publication number for cross-
referencing.
  pub.publication_number AS google_id,

  -- Step 3 Extract fundamental chronological metrics including filing and priority
dates.
  pub.filing_date,
  pub.publication_date,
  pub.grant_date,
  pub.priority_date,

  -- Step 4 Compile corporate assignee information.
  -- Extract specific attributes from the harmonized nested array.
  (SELECT STRING_AGG(name, '; ') FROM UNNEST(pub.assignee_harmonized)) AS
assignee_harmonized_name,
  (SELECT STRING_AGG(country_code, '; ') FROM UNNEST(pub.assignee_harmonized)) AS
assignee_harmonized_country,

  -- Aggregate raw assignee text elements directly from the flat array.
  (SELECT STRING_AGG(a, '; ') FROM UNNEST(pub.assignee) AS a) AS assignee_list,

  -- Step 5 Compile individual inventor details.
  -- Extract specific attributes from the harmonized nested array.
```

```
  (SELECT STRING_AGG(name, '; ') FROM UNNEST(pub.inventor_harmonized)) AS
inventor_harmonized_name,

  -- Aggregate raw inventor text elements directly from the flat array.
  (SELECT STRING_AGG(i, '; ') FROM UNNEST(pub.inventor) AS i) AS inventor_list,

  -- Step 6 Aggregate technological classification codes.
  -- Extract specific classification codes from the nested structural arrays.
  (SELECT STRING_AGG(code, '; ') FROM UNNEST(pub.ipc)) AS ipc_codes,
  (SELECT STRING_AGG(code, '; ') FROM UNNEST(pub.cpc)) AS cpc_codes

FROM
  -- Define the source table containing the initial set of Internal Combustion
Engine patent citations.
  `[PROJECT_ID].[DATASET].citations_table` AS t

JOIN
  `patents-public-data.patents.publications` AS pub
ON
  -- Perform string manipulation to remove formatting dashes and ensure accurate
key matching between databases.
  t.cited_patents = REPLACE(pub.publication_number, '-', '');
```

Appendix C

This appendix presents the main procedures used for the spatial clustering analysis. Section C.1 describes how patent assignee addresses were converted into geographic coordinates. Section C.2 reports the parameter search process used to calibrate the clustering model. Section C.3 applies the HDBSCAN algorithm to identify geographically concentrated automotive innovation clusters and separate geographically isolated firms from the main cluster sample.

C.1 Spatial Data Preparation and Geocoding Implementation

The following Python script details the automated pipeline utilized to convert unstructured patent assignee addresses into precise geographic coordinates. It employs heuristic text denoising and unrestricted application programming interface querying to ensure high-fidelity spatial data extraction. This rigorous geocoding process establishes the necessary foundational coordinates required for the subsequent density-based spatial clustering analysis.

Listing C.1 Python script for API geocoding

```
import pandas as pd

import requests

import time

import os

import re



API_KEY = ""

INPUT_CSV = "address_data.csv"

OUTPUT_CSV = "geocoded_final_output.csv"
```

```
CACHE_FILE = "address_geocode_cache.csv"


def clean_patent_address(raw_address):
    """
    Perform deep text denoising to optimize address strings for the Geocoding API matching logic.
    """
    if pd.isna(raw_address):
        return ""
    addr_str = str(raw_address).strip()

    # Remove trailing country codes enclosed in parentheses (e.g., (GB), (CN))
    addr_str = re.sub(r'\s*\([A-Z]{2}\)$', '', addr_str)

    # Remove redundant and potentially confusing consecutive geographic suffixes (e.g., ,KONG,CN)
    addr_str = re.sub(r',\s*[A-Z]{2,4}\s*,\s*[A-Z]{2}$', '', addr_str)

    # Deduplicate consecutive identical postal codes (e.g., 999077,999077)
    addr_str = re.sub(r'(\b\d{5,6}\b)(?:,\s*\1)+', r'\1', addr_str)

    return addr_str.strip()
```

```
def geocode_free_text_address(address_text):
    """
    Remove underlying regional constraints to rely entirely on the API's natural language parsing of the physical address.
    """
    if not address_text:
        return None, None, "EMPTY_ADDRESS"

    url = "https://maps.googleapis.com/maps/api/geocode/json"

    # Completely omit the 'components' parameter to allow unrestricted, global-scale geographic matching
    params = {
        "address": address_text,
        "key": API_KEY
    }

    try:
        response = requests.get(url, params=params, timeout=10)
        data = response.json()
```

```
        if data['status'] == 'OK' and len(data['results']) > 0:
            location = data['results'][0]['geometry']['location']
            formatted_addr = data['results'][0].get('formatted_address', 'N/A')
            return location['lat'], location['lng'], formatted_addr

    except Exception as e:
        print(f"Geocoding API connection failed: {e}")

    return None, None, "GEOCODE_FAILED"

print("Initializing unrestricted free-text spatial parsing pipeline...")

df_main = pd.read_csv(INPUT_CSV)

# Initialize caching mechanism to prevent redundant API calls and handle unexpected interruptions
cache_dict = {}
if os.path.exists(CACHE_FILE):
    df_cache = pd.read_csv(CACHE_FILE)
    for _, row in df_cache.iterrows():
        cache_dict[str(row['publication_number'])] = row.to_dict()
```

```
results = []
total_records = len(df_main)


# Execute spatial coordinate extraction iteratively
for idx, row in df_main.iterrows():
    pub_num = str(row['publication_number']).strip()
    assignee = str(row['first_assignee']).strip()
    raw_addr = row['raw_address']


    # Retrieve existing coordinates from the local cache if available
    if pub_num in cache_dict:
        results.append(cache_dict[pub_num])
        continue


    print(f"Processing unrestricted geocoding for entity [{idx + 1}/{total_records}]:
{assignee}")


    clean_addr = clean_patent_address(raw_addr)
    lat, lng, std_addr = geocode_free_text_address(clean_addr)


    if lat is not None:
        source = "GEOCODING_API_SUCCESS"
```

```
else:
    source = "FAILED"

record = {
    'publication_number': pub_num,
    'first_assignee': assignee,
    'Raw_Address': raw_addr,
    'Cleaned_Address': clean_addr,
    'Standardized_Address': std_addr,
    'Latitude': lat,
    'Longitude': lng,
    'Source': source
}

results.append(record)
cache_dict[pub_num] = record

# Update local cache dynamically
pd.DataFrame(list(cache_dict.values())).to_csv(CACHE_FILE, index=False)

# Implement rate limiting to comply with API usage policies
```

```
    time.sleep(0.5)


df_final = pd.DataFrame(results)

df_final.to_csv(OUTPUT_CSV, index=False)

print("Coordinate matrix reconstruction complete. Spatial network nodes have been fully resolved.")
```

C.2 Parameter Calibration and Robustness Diagnostics for HDBSCAN Clustering

The following script implements a systematic grid search to calibrate the hyperparameters for the spatial clustering algorithm. It evaluates a range of minimum cluster sizes and sample thresholds to identify the most stable configuration for the regional innovation systems. The diagnostic output incorporates internal validity metrics and noise ratio assessments to ensure that the final results are robust across both Chinese and European samples.

Listing C.2 Python script for hyperparameter grid search and spatial cluster diagnostics

```
import os

os.environ['JOBLIB_TEMP_FOLDER'] = 'C:/temp'


import pandas as pd

import numpy as np

import hdbscan

from sklearn.metrics.pairwise import haversine_distances


np.random.seed(42)
```

```
print("Loading master spatial panel...")

df_master = pd.read_csv("master_spatial_nodes.csv")

KMS_PER_RADIAN = 6371.0



def evaluate_endogenous_grid_with_weight_diagnostics(sub_df, region_name):

    print(f"\n============================================================")

    print(f"Initiating 0km endogenous high-dimensional weight diagnostics for
{region_name}...")



    coords_rad = np.radians(sub_df[['latitude', 'longitude']].values)

    total_unique_nodes = len(sub_df)



    # Implement a standardized cross-national calibration benchmark scanning minimum
cluster sizes from 10 to 56

    grid_mcs = list(range(10, 57))



    MAX_NODE_NOISE_RATIO = 0.30

    results = []



    print(f" >>> Scanning the unconstrained physical topological space for {region_name}
(Total Nodes {total_unique_nodes})...")
```

```
    for mcs in grid_mcs:
        # Enforce the half-size heuristic for core density estimation allowing a localized
deviation boundary
        baseline_ms = max(2, mcs // 2)
        grid_ms = list(range(max(2, baseline_ms - 2), baseline_ms + 3))


        for ms in grid_ms:
            eps_km = 0.0
            eps_rad = eps_km / KMS_PER_RADIAN


            clusterer = hdbscan.HDBSCAN(
                min_cluster_size=mcs,
                min_samples=ms,
                cluster_selection_epsilon=eps_rad,
                cluster_selection_method='eom',
                metric='haversine',
                gen_min_span_tree=True,
                core_dist_n_jobs=-1
            )


            clusterer.fit(coords_rad)
            labels = clusterer.labels_
```

```
        unique_clusters = len(set(labels)) - (1 if -1 in labels else 0)

        noise_mask = labels == -1

        noise_count_nodes = np.sum(noise_mask)

        node_noise_ratio = noise_count_nodes / total_unique_nodes



        # Diagnostic Metric 1 Extract the top five weighted nodes misclassified as spatial
noise

        if noise_count_nodes > 0:

            top_noise_weights = sub_df.loc[noise_mask, 'patent_weight'].nlargest(5).tolist()

            top_noise_weights.extend([0.0] * (5 - len(top_noise_weights)))

        else:

            top_noise_weights = [0.0] * 5



        min_cluster_dist = np.nan

        mean_persistence = np.nan

        weight_top5 = [0.0] * 5

        weight_lightest = 0.0



        if unique_clusters >= 2:

            sub_df_temp = sub_df.copy()

            sub_df_temp['temp_label'] = labels
```

```
valid_nodes = sub_df_temp[sub_df_temp['temp_label'] != -1]

mean_persistence = np.mean(clusterer.cluster_persistence_)

# Aggregate the total patent weight for valid and economically meaningful regional clusters
cluster_stats = valid_nodes.groupby('temp_label').agg(
    total_weight=('patent_weight', 'sum')
).reset_index()

cluster_stats = cluster_stats.sort_values('total_weight', ascending=False)

# Diagnostic Metric 2 Extract the top five and the lowest weighted clusters to evaluate size disparity
weights_list = cluster_stats['total_weight'].tolist()
weights_list.extend([0.0] * (5 - len(weights_list)))
weight_top5 = weights_list[:5]
weight_lightest = cluster_stats.iloc[-1]['total_weight'] if len(cluster_stats) > 0 else 0.0

# Diagnostic Metric 3 Calculate the minimum geographic distance between cluster centroids
centroids = valid_nodes.groupby('temp_label')[['latitude', 'longitude']].mean()
```

```
            centroids_rad = np.radians(centroids.values)

            dist_matrix = haversine_distances(centroids_rad, centroids_rad) *
KMS_PER_RADIAN

            np.fill_diagonal(dist_matrix, np.inf)

            min_cluster_dist = np.min(dist_matrix)



            try:

                dbcv_score = clusterer.relative_validity_

            except Exception:

                dbcv_score = np.nan

        else:

            dbcv_score = np.nan



        # Real-time console monitoring output

        print(f"     - [Size {mcs:2d}, Sensitivity {ms:2d}] | Clusters {unique_clusters:3d} |
Noise {node_noise_ratio*100:4.1f}% | Distance {min_cluster_dist if
pd.notna(min_cluster_dist) else 0.0:4.1f}km")

        print(f"       >>> Noise Extremes {top_noise_weights[0]:.0f},
{top_noise_weights[1]:.0f}, {top_noise_weights[2]:.0f}, {top_noise_weights[3]:.0f},
{top_noise_weights[4]:.0f}")

        print(f"       >>> Cluster Distribution (Top5 & Min) [{weight_top5[0]:.0f},
{weight_top5[1]:.0f}, {weight_top5[2]:.0f}, {weight_top5[3]:.0f}, {weight_top5[4]:.0f} |
Min {weight_lightest:.0f}] | Persistence {mean_persistence if pd.notna(mean_persistence)
else 0.0:.3f} | DBCV {dbcv_score if pd.notna(dbcv_score) else -1.0:.4f}")
```

```
results.append({
    'min_cluster_size': mcs,
    'min_samples': ms,
    'cluster_count': unique_clusters,
    'node_noise_ratio': node_noise_ratio,
    'noise_w1': top_noise_weights[0],
    'noise_w2': top_noise_weights[1],
    'noise_w3': top_noise_weights[2],
    'noise_w4': top_noise_weights[3],
    'noise_w5': top_noise_weights[4],
    'mean_persistence': mean_persistence,
    'cluster_w1': weight_top5[0],
    'cluster_w2': weight_top5[1],
    'cluster_w3': weight_top5[2],
    'cluster_w4': weight_top5[3],
    'cluster_w5': weight_top5[4],
    'cluster_w_min': weight_lightest,
    'min_cluster_dist_km': min_cluster_dist,
    'dbcv_score': dbcv_score
})
```

```
    df_results = pd.DataFrame(results)


    # Execute plateau validation to identify structurally stable parameter configurations
    df_results = df_results.sort_values(['min_cluster_size',
'min_samples']).reset_index(drop=True)
    diff_bwd = df_results['cluster_count'].diff().abs()
    diff_fwd = df_results['cluster_count'].diff(-1).abs()
    df_results['is_stable'] = (diff_bwd <= 2) | (diff_fwd <= 2)


    condition_noise = df_results['node_noise_ratio'] <= MAX_NODE_NOISE_RATIO
    condition_stable = df_results['is_stable'] == True
    condition_dbcv = df_results['dbcv_score'].notna()


    df_results['meets_criteria'] = np.where(condition_noise & condition_stable &
condition_dbcv, 1, 0)


    output_filename = f"evaluation_grid_0km_weight_diag_{region_name}.csv"
    df_results.to_csv(output_filename, index=False)


    valid_configs = df_results[df_results['meets_criteria'] == 1]
    print(f"\n >>> Evaluation for {region_name} completed. Captured {len(valid_configs)}
robust empirical parameter sets.")
```

```
    print(f" >>> Diagnostic matrix safely stored to {output_filename}")

    return df_results

# Execute empirical calibration
df_cn = df_master[df_master['true_in_china'] == 1].copy()
df_eu = df_master[df_master['true_in_europe'] == 1].copy()

eval_cn = evaluate_endogenous_grid_with_weight_diagnostics(df_cn, "China")
eval_eu = evaluate_endogenous_grid_with_weight_diagnostics(df_eu, "Europe")

print("\n==================================================")
print("Standardized cross-national high-dimensional weight diagnostic matrices generated successfully.")
```

C.3 Spatial Clustering and Identification of Isolated Firms

The following script applies the Python-based HDBSCAN clustering algorithm to identify geographically concentrated automotive innovation clusters across the sample. The procedure also identifies firms that are classified as spatial noise even though they have substantial patent output; I refer to these as isolated giants. These firms, labelled as “cis” entities in the extraction process, operate outside the main geographic clusters and were excluded from the primary cluster-level panel used in the regression analysis. This approach allows the analysis to focus on geographically embedded regional clusters while separately accounting for isolated high-output firms.

Listing C.3 Python script for spatial clustering and geographic network construction

```
import os

os.environ['JOBLIB_TEMP_FOLDER'] = 'C:/temp'



import pandas as pd

import numpy as np

import hdbscan



np.random.seed(42)



print("Loading Master Spatial Panel...")

df_master = pd.read_csv("master_spatial_nodes.csv")



# Standardized baseline parameters calibrated from prior diagnostics

TARGET_MCS = 14

TARGET_MS = 5



print("\n=====================================================")

print(">>> Step 1. Calculate the Global Exogenous Cut-off Threshold <<<")



# Establish a threshold based on the 99th percentile to identify potential Superstar Firms

EXOGENOUS_THRESHOLD = df_master['patent_weight'].quantile(0.99)
```

```
print(f"Total observations in master sample {len(df_master)}")

print(f"Defining the Superstar Firm threshold at the 99th percentile.")

print(f"Exogenous entry threshold for isolated giants locked at >=
{EXOGENOUS_THRESHOLD:.0f} patents")

print("==================================================\n")



def execute_dual_track_clustering(sub_df, region_offset, region_name):

    print(f"Initiating Dual-Track Identification for the {region_name} region...")



    coords_rad = np.radians(sub_df[['latitude', 'longitude']].values)



    # --------------------------------------------------

    # Step 2. Pure Physical Topological Partitioning

    # --------------------------------------------------

    clusterer = hdbscan.HDBSCAN(

        min_cluster_size=TARGET_MCS,

        min_samples=TARGET_MS,

        cluster_selection_epsilon=0.0,

        cluster_selection_method='eom',

        metric='haversine',

        gen_min_span_tree=True,
```

```
    core_dist_n_jobs=-1
)

sub_df = sub_df.copy()
sub_df['temp_raw_id'] = clusterer.fit_predict(coords_rad)

# Assign globally unique identifiers with regional prefixes to valid clusters
sub_df['cluster_id'] = sub_df['temp_raw_id'].apply(
    lambda x: x + region_offset if x != -1 else -1
)

# Initialize indicator variable (coded as 'is_cis') to flag isolated Superstar Firms
sub_df['is_cis'] = 0

ris_count = sub_df['cluster_id'].nunique() - (1 if -1 in sub_df['cluster_id'].values else 0)
print(f"  > [Physical Track] Identified {ris_count} Regional Innovation Systems")

# ---------------------------------------------------
# Step 3. Attribute-based Cross-Identification and Extraction
# ---------------------------------------------------
noise_mask = sub_df['cluster_id'] == -1
```

```
    # Condition A. Identify spatially isolated entities as Superstar Firms based on weight thresholds
    cis_mask = noise_mask & (sub_df['patent_weight'] >= EXOGENOUS_THRESHOLD)
    cis_count = cis_mask.sum()

    if cis_count > 0:
        print(f"  > [Extraction Track] Identified {cis_count} isolated giants as Superstar Firms")

        # Flag these entities (is_cis = 1) for targeted extraction and separate qualitative discussion
        sub_df.loc[cis_mask, 'is_cis'] = 1

        max_id = sub_df['cluster_id'].max() + 1

        # Allocate independent identifiers to facilitate subsequent exclusion from the spatial regression panel
        for idx in sub_df[cis_mask].index:
            sub_df.at[idx, 'cluster_id'] = max_id
            max_id += 1

    # Condition B. Confirm remaining low-weight entities as genuine spatial noise for final removal
```

```
    true_noise_mask = (sub_df['cluster_id'] == -1)

    true_noise_count = true_noise_mask.sum()

    print(f"  > [Fragment Removal] Confirmed {true_noise_count} entities as genuine spatial fragments\n")



    return sub_df



# Execute the identification pipeline separately for China and Europe

df_cn = df_master[df_master['true_in_china'] == 1].copy()

df_eu = df_master[df_master['true_in_europe'] == 1].copy()



final_cn = execute_dual_track_clustering(df_cn, region_offset=0, region_name="China")

final_eu = execute_dual_track_clustering(df_eu, region_offset=1000, region_name="Europe")



final_panel_nodes = pd.concat([final_cn, final_eu], ignore_index=True)

final_panel_nodes = final_panel_nodes.drop(columns=['temp_raw_id'])



output_filename = "master_spatial_panel_final4.10.csv"

final_panel_nodes.to_csv(output_filename, index=False)

print(f"Dual-track identification complete. Superstar firms have been isolated for separate analysis.")
```