APPLICATION OF MACHINE LEARNING FOR EVALUATING THAILAND'S ECONOMIC COMPLEXITY

PORNPINUN YEERONG

MASTER OF SCIENCE
IN
INFORMATION TECHNOLOGY

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2024

APPLICATION OF MACHINE LEARNING FOR EVALUATING THAILAND'S ECONOMIC COMPLEXITY

PORNPINUN YEERONG

THIS THESIS IS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2024

THESIS APPROVAL

MAE FAH LUANG UNIVERSITY

FOR

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

Thesis Title: Application of Machine Learning for Evaluating Thailand's Economic Complexity

Author: Pornpinun Yeerong

Examination Committee:

Assistant Professor Teerawat Kamnardsiri, Ph. D.	Chairperson
Assistant Professor Surapong Uttama, Ph. D.	Member
Assistant Professor Santichai Wicha, Ph. D.	Member
Assistant Professor Worasak Rueangsirarak, Ph. D.	Member

Advisor:

....Advisor

(Assistant Professor Surapong Uttama, Ph. D.)

Dean:

....

(Assistant Professor Nacha Chondamrongkul, Ph. D.)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Assistant Professor Dr. Surapong Uttama, for his invaluable guidance, constructive feedback, and unwavering support throughout the course of this research. His academic rigor and encouragement have been instrumental in shaping the quality and direction of this thesis.

I am also sincerely thankful to Assistant Professor Dr. Nathapornpan Uttama, who not only encouraged me to pursue my master's degree but also provided continuous support and motivation throughout my academic journey. Her belief in my potential has been a source of inspiration and strength.

My appreciation extends to the faculty and staff of the School of Information Technology, Mae Fah Luang University, for fostering a supportive academic environment. I am especially grateful to the Center of Excellence in AI and Emerging Technologies for offering technical resources essential to this research.

I also acknowledge the contributions of public institutions such as the National Statistical Office of Thailand, the Office of the National Economic and Social Development Council (NESDC), and the Bureau of the Budget, whose data formed the empirical foundation of this study.

I would also like to express my sincere gratitude to Mae Fah Luang University for providing financial support and facilitating the completion of this thesis, as well as to the Center of Excellence in AI and Emerging Technologies for additional financial support towards the publication of this research.

Finally, I am deeply thankful to my family for their unconditional love, patience, and support throughout this journey, as well as to my friends and colleagues for their encouragement and companionship during my studies.

Pornpinun Yeerong

Thesis Title	Application of Machine Learning for Evaluating Thailand's Economic Complexity
Author	Pornpinun Yeerong
Degree	Master of Science (Information Technology)
Advisor	Assistant Professor Surapong Uttama, Ph. D.

ABSTRACT

This study investigates Thailand's economic complexity at the subnational level by constructing a panel dataset of 77 provinces across 20 economic activities from 2011 to 2021. The Economic Complexity Index (ECI) is developed using employment data, applying the Location Quotient and Method of Reflection techniques. Fixed-effects panel regression, quantile regression, and generalized additive models (GAMs) are employed to explore the relationship between ECI and two key development outcomes: economic growth and income inequality. The results reveal a nonlinear and distribution-sensitive relationship. While ECI tends to promote economic growth after surpassing a complexity threshold—particularly in provinces with initially lower income levels, its inequality-reducing effect is most pronounced in high-Gini provinces. Clustering analysis is used to group provinces based on economic complexity, income, inequality, and demographic indicators. Among the algorithms tested, K-means clustering performs best, revealing distinct regional development patterns and increasing structural divergence over time. This research advances the literature on economic complexity by integrating machine learning techniques into subnational economic diagnostics and highlights the potential of ECI as a tool for promoting inclusive and region-specific development policies in middle-income countries like Thailand.

Keywords: Economic Complexity Index, Regional Economic Development, Income Inequality, Panel Data Analysis, Fixed Effects Model, Machine Learning, Clustering, Location Quotient, Thailand, Subnational Analysis

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND SYMBOLS

ECI	Economic Complexity Index
GDP	Gross Domestic Product
GINI	Gini Coefficient
FE	Fixed Effects
QREG	Quantile Regression
GAM	Generalized Additive Model
ML	Machine Learning
RGPPPC	Real Gross Provincial Product Per Capita
GAM	Generalized Additive Model
LQ	Location Quotient
VIF	Variance Inflation Factor
LFS	Labor Force Survey
NSO	National Statistical Office (Thailand)
POP	Population
POPD	Population Density

CHAPTER 1

INTRODUCTION

1.1 Background and Importance of the Research Problem

Modern economic development increasingly unfolds in a global environment characterized by volatility, uncertainty, complexity, and ambiguity (VUCA). Originally conceptualized in strategic planning and later adopted in policy discourse, the VUCA framework underscores the limitations of conventional models in responding to rapidly shifting and interdependent systems (Bennett & Lemoine, 2014). Under such conditions, development strategies that rely solely on capital accumulation, labor expansion, or aggregate output growth have proven insufficient, particularly at the sub-national level. Common indicators such as Gross Domestic Product (GDP), although widely used, offer limited spatial granularity and responsiveness due to infrequent data updates, collection costs, and aggregation biases (Gao & Zhou, 2018; NESDC, 2021). These constraints point to the need for more granular, dynamic, and structurally informative metrics.

The Economic Complexity Index (ECI), developed by Hidalgo and Hausmann (2009), provides a powerful alternative by capturing the embedded knowledge and productive capabilities of an economy. Rather than focusing on traditional inputs, the ECI evaluates the diversity and sophistication of the products an economy is able to produce competitively, serving as a proxy for its underlying know-how. More complex economies are generally better positioned for innovation, diversification, and long-term resilience. Recent studies have shown that higher economic complexity is strongly associated with faster, more inclusive growth and lower structural inequality (Hidalgo, 2021; Fritz & Manduca, 2021; Aghion, Cherif, & Hasanov, 2021).

Thailand's development trajectory exemplifies both the benefits and the limitations of structural transformation. Between 1965 and 2005, the country emerged as a "great transformer" economy—alongside Brazil, Indonesia, Türkiye, Malaysia, China, South Korea, and Singapore—driven by rapid industrialization and labor absorption (Hartmann et al., 2017). However, since graduating to Upper Middle-Income Country (UMIC) status in 2011, Thailand's growth has stagnated, while regional disparities have persisted (Jitsuchon, 2012; Kittiprapas & Wiboonchutikula, 2019). Despite this, Thailand's productive structure continues to evolve: its global ECI ranking improved from 64th in 2000 to 29th in 2023 (Hausmann, Pietrobelli, & Santos, 2021), surpassing several high-income economies and signaling untapped potential.

Yet this national-level progress conceals wide intra-national disparities. While the National Economic and Social Development Council (NESDC) employs planning instruments such as the Sustainable Development Goals (SDG) monitoring framework and the Development Potential Assessment Index (DPAI), these tools emphasize aggregate indicators like Gross Provincial Product (GPP) per capita and the Gini coefficient. Though useful, they fail to account for the structural composition of provincial economies. The exclusion of productive complexity indicators limits policymakers' ability to identify latent capabilities and design effective upgrading strategies—thereby perpetuating reliance on static or incomplete development diagnostics (NESDC, 2021; Apaitan, Ananchotikul, & Disyatat, 2017).

Recent empirical studies underscore the value of applying economic complexity metrics at the sub-national scale. Numerous studies across diverse national contexts have demonstrated that economic complexity can reveal structural inequalities and inform targeted policy interventions. For instance, complexity has been shown to correlate negatively with income inequality in Mexican states (Gómez-Zaldívar, Osorio-Caballero, & Saucedo-Acosta, 2022), while in Australia, Canada, and Italy, sub-national complexity analysis has been used to support regional industrial strategies, assess global competitiveness, and explain productivity polarization (Reynolds et al., 2018; Wang & Turkina, 2020; Basile & Cicerone, 2022). In Europe, researchers have applied economic complexity frameworks to Romanian counties, Spanish provinces, and the Kaliningrad region—illustrating their versatility in both interregional and international trade contexts (Török, Benedek, & Gómez-Zaldívar, 2022; Balsalobre et al., 2017, 2019; Roos et al., 2021).

Despite this expanding body of research, Thailand currently lacks institutional tools for systematically measuring economic complexity at the provincial level. A notable exception is the study by Apaitan, Ananchotikul, and Disyatat (2017), which employed firm-level and export registration data. However, this approach is constrained by fragmented data systems, limited public access, and low scalability. In contrast, employment-based methods—such as those proposed by Fritz and Manduca (2021) and Mealy, Farmer, and Teytelboym (2019)—offer a replicable, timely, and more inclusive alternative, particularly in contexts where sub-national trade data are unavailable.

This thesis addresses the empirical and operational gap by constructing a sub-national ECI for Thailand using employment data from the Labor Force Survey (LFS), which records employment across 20 economic sectors in all 77 provinces. Both ECI and Product Complexity Index (PCI) scores are derived using the method of reflection and location quotient techniques. These indices are then empirically examined in relation to provincial-level outcomes in economic growth and income inequality using fixed-effects and quantile regression models.

In addition, the study applies unsupervised machine learning techniques—including K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM)—to classify provinces based on complexity, socio-economic indicators, and structural characteristics. Such hybrid approaches have been endorsed in recent economic research for their ability to capture nonlinear patterns and structural heterogeneity (Mullainathan & Spiess, 2017; Athey & Imbens, 2019). By integrating complexity measurement with advanced analytical techniques, this thesis aims to produce timely, scalable, and policy-relevant insights to inform Thailand's regional development planning.

1.2 Research Objectives

This study investigates the role of economic complexity in shaping regional development outcomes in Thailand, with a particular focus on economic growth and income inequality at the provincial level. It addresses existing empirical and methodological gaps by constructing a subnational Economic Complexity Index (ECI) and applying a combination of econometric and machine learning techniques.

The three main research objectives are as follows:

1.2.1 To collect and preprocess employment data from Thailand's Labor Force Survey (LFS) at the provincial level to construct Economic Complexity Index

1.2.2 To investigate the impact of ECI on economic growth and income inequality using regression techniques

1.2.3 To grouping provinces based on socio-economic profiles using clustering algorithms

1.3 The Importance of Research

This research is significant for both academic inquiry and policy formulation in the context of regional development. From a theoretical perspective, the study advances the literature on economic complexity by extending its application to the subnational level in an emerging economy. While the Economic Complexity Index (ECI) has been widely used to explain cross-country differences in growth and inequality, its relevance within countries, particularly across diverse provincial contexts remains underexplored. By constructing a provincial ECI using labor market data, this study contributes a novel and replicable framework for capturing productive capabilities where trade data are unavailable or incomplete.

Methodologically, the study integrates econometric techniques with machine learning to offer a multidimensional analysis of the complexity–development relationship. The use of panel quantile regression allows for the investigation of how the effects of complexity vary across different levels of inequality and income, while generalized additive models (GAMs) uncover nonlinear dynamics that conventional models may miss. Clustering algorithms further enrich the analysis by identifying spatial and structural patterns among provinces, providing insight into latent development typologies.

From a policy perspective, the study addresses a critical gap in Thailand's provincial planning framework. Current development indicators overlook structural economic diversity, relying instead on aggregate measures such as GPP per capita and household income. By introducing complexity-based metrics and classifying provinces based on socio-economic profiles, this research equips policymakers with tools to design more targeted and capability-sensitive strategies. The findings support Thailand's goals under the Sustainable Development Goals (SDGs) and decentralization reforms by promoting evidence-based, regionally adaptive development planning.

1.4 Research Hypotheses

This study tests the relationship between economic complexity and regional development outcomes in Thailand and explores structural provincial typologies using clustering techniques. The research is guided by the following hypotheses:

1.4.1 Economic Complexity and Economic Growth

H1.1: Provinces with higher economic complexity, as computed from employment data, exhibit stronger economic growth.

H1.2: The effect of complexity on growth varies across the income distribution, captured through quantile-based modelling.

H1.3: A nonlinear relationship exists between complexity and growth, detectable through nonparametric modelling.

1.4.2 Economic Complexity and Income Inequality

H2.1: Economic complexity is negatively associated with income inequality across provinces.

H2.2: The inequality-reducing effect of complexity is heterogeneous and distribution-sensitive.

H2.3: Complexity and inequality are linked through a nonlinear, potentially threshold-based relationship.

1.4.3 Provincial Clustering and Development Typologies

H3.1: Unsupervised machine learning can classify provinces into development clusters based on complexity and socio-economic indicators.

H3.2: Cluster membership patterns reflect regional structural characteristics and evolve over time.

1.5 Research Questions

This study is guided by the following research questions, which aim to explore the link between economic complexity and regional development in Thailand through a data-driven, computational approach:

1.5.1 How can a provincial-level Economic Complexity Index (ECI) be constructed using high-frequency employment data in Thailand?

1.5.2 To what extent does economic complexity influence economic growth and income inequality at the provincial level, and how do these effects vary across the distribution and overtime?

1.5.3 Can clustering techniques effectively group Thai provinces into distinct development clusters based on complexity and socio-economic characteristics?

These questions are addressed through the integration of econometric modeling and machine learning techniques, providing a comprehensive understanding of structural disparities and development trajectories across provinces.

1.6 Scopes of Research

This study is situated within the context of Thailand's provincial economies and investigates the role of economic complexity in shaping two key development outcomes: economic growth and income inequality. The spatial scope of the research encompasses all 77 provinces of Thailand, treating each as an individual unit of analysis. Bueng Kan province, established in 2011, is included in the analysis from 2013 onward due to the absence of baseline data for the year of its inception.

The study focuses on five benchmark years—2011, 2013, 2015, 2017, and 2019, which correspond to the availability of income inequality data from the Household Socio-Economic Survey. For the analysis of economic growth, the dataset is extended through 2021 using interpolated values of the Economic Complexity Index (ECI), allowing for a more complete panel structure. Thematically, the research is centered on the construction of a subnational ECI and its application in explaining real gross provincial product per capita (RGPPPC) and the Gini index, which serve as proxies for growth and inequality, respectively. In addition, the study incorporates a structural classification of provinces using clustering techniques to uncover latent development typologies.

Methodologically, the research adopts a multi-method framework that integrates fixed effects panel regression, panel quantile regression, and generalized additive models (GAMs) to estimate both average and distributional effects, as well as nonlinearities in the relationship between economic complexity and development outcomes. Unsupervised machine learning techniques—specifically K-means, hierarchical agglomerative clustering (HAC), and Gaussian mixture models (GMM) are employed to classify provinces based on multi-dimensional socio-economic profiles. Together, these spatial, temporal, thematic, and methodological boundaries define the analytical scope of the study and ensure its relevance to both academic inquiry and policy application in regional development.

1.7 Preliminary Agreement

This study is conducted in partial fulfillment of the requirements for the Master of Science degree in Information Technology at Mae Fah Luang University, under the supervision of Assistant Professor Dr. Surapong Uttama. The research topic, objectives, and methodological framework have been reviewed and approved by the thesis advisory committee in accordance with the university's academic regulations. All data used in the study are obtained from publicly available and officially recognized sources, primarily the National Statistical Office (NSO), the Office of the National Economic and Social Development Council (NESDC), and the Bureau of Budget. Ethical considerations have been observed throughout the research design and data handling processes. The preliminary agreement ensures that the study complies with institutional guidelines and academic standards, and that the research topic contributes meaningfully to both the field of information technology and public policy discourse on regional development.

1.8 Research Limitations

While this study adopts a comprehensive and data-driven approach to analyzing economic complexity and regional development, several limitations should be acknowledged. First, the construction of the provincial-level Economic Complexity Index (ECI) relies on employment data from the Labor Force Survey, which, although high-frequency and nationally representative, may not fully capture informal sector dynamics or knowledge-based activities not classified under standard economic sectors.

Second, the measurement of income inequality is constrained by data availability, as the Gini coefficient is only reported biennially, limiting the temporal resolution of the inequality analysis. Third, while fixed effects and quantile regression models control unobserved heterogeneity, potential endogeneity between complexity and development outcomes may still affect causal inference. Fourth, the clustering analysis, based on unsupervised machine learning techniques, is exploratory in nature and sensitive to variable selection and algorithmic parameters.

Finally, the generalizability of the findings may be limited to contexts with similar institutional structures and data availability as Thailand. Despite these limitations, the study provides a robust and replicable framework for subnational complexity analysis and contributes novel insights to the intersection of information technology, economics, and public policy.

1.9 Terminology Definition

Economic Complexity Index (ECI): A measure of the knowledge intensity of an economy, based on the diversity and ubiquity of the products it exports. Higher ECI values indicate a more complex and sophisticated economy (Hidalgo & Hausmann, 2009).

Economic Growth: The increase in the production of goods and services in an economy over a specific period is typically measured by the change in Gross Domestic Product (GDP) or Gross Provincial Product (GPP).

Income Inequality: The unequal distribution of income among individuals or households within an economy. Common measures of income inequality include the Gini coefficient, Theil index, and Palma ratio.

Labor Force Survey (LFS): A survey was conducted by the National Statistical Office of Thailand to collect data on employment, unemployment, and other labor market characteristics.

Machine Learning: A subset of artificial intelligence that involves training computer algorithms to learn patterns and make predictions or decisions based on data, without being explicitly programmed (Mitchell, 1997).

Clustering Algorithms: Machine learning techniques are used to group similar data points or observations into clusters based on their inherent characteristics or patterns. Examples include K-means, hierarchical clustering, and DBSCAN.

Regression Models: Statistical models are used to investigate the relationship between a dependent variable and one or more independent variables. Common types of regression models include linear regression, logistic regression, and polynomial regression.

Middle-Income Trap: A situation in which a country's economic growth slows down or stagnates after reaching a middle-income level, making it difficult to transition to a high-income economy (Gill & Kharas, 2015).

Sustainable Development Goals (SDGs): A set of 17 global goals was adopted by the United Nations Member States in 2015, aiming to address social, economic, and environmental challenges and promote sustainable development by 2030 (United Nations, 2015).

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the theoretical and empirical foundations relevant to this study, which applies two major analytical approaches: regression analysis and clustering techniques. The chapter begins by exploring the core concepts underpinning the measurement of economic complexity, particularly the Economic Complexity Index (ECI), Location Quotient (LQ), and the Method of Reflection. These indicators offer a multidimensional view of economic structure, productivity potential, and regional competitiveness.

The second part of the review is organized around two methodological strands. The first strand focuses on regression-based studies that examine the relationship between economic complexity and key developments, namely, economic growth and income inequality. A wide range of techniques are considered in this context, including panel regression models, fixed and random effects models, quantile regression, and generalized additive models (GAM). These approaches allow for both linear and nonlinear modelling of the complexity-growth-inequality nexus. The second strand explores clustering-based research, particularly the use of unsupervised machine learning algorithms to group regions or provinces based on multiple socio-economic indicators. Methods such as K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) are reviewed for their suitability in regional classification tasks.

These two strands provide the conceptual and empirical grounding for this study's dual-method research design. The chapter also highlights key gaps in the literature, particularly the limited use of economic complexity and clustering techniques in sub-national policy contexts in Thailand.

2.1 Theoretical Foundations

This section presents the theoretical underpinnings that inform the study's approach to analyzing economic complexity and its relationship to regional development outcomes. The foundation begins with a discussion of the Economic Complexity Theory, which provides a framework for understanding the diversity and sophistication of productive activities across provinces. This is followed by two technical concepts; Location Quotient (LQ) and the Method of Reflection that are integral to computing sub-national measures of economic complexity using employment data.

These theoretical tools serve dual purposes. First, they underpin the construction of the Economic Complexity Index (ECI) used in regression models that evaluate the impact of complexity on economic growth and income inequality. Second, the outputs derived from these theoretical constructs, such as economic complexity scores serve as key inputs for clustering Thailand's provinces based on shared socio-economic characteristics, enabling more targeted policy design and regional classification. These foundations enable the study to assess not only how economic complexity is distributed across Thailand but also how it correlates with broader development patterns through both causal inference (regression) and descriptive categorization (clustering).

2.1.1 Theoretical Foundations

The theoretical foundations of economic complexity arise from a fundamental rethinking of what drives economic development. Traditional growth models—such as Solow's (1956) neoclassical framework—emphasize capital accumulation, labor input, and exogenous technological progress. In contrast, economic complexity theory posits that long-term growth is driven by a society's ability to accumulate and recombine productive knowledge embedded in human capital, institutions, and industrial ecosystems (Hidalgo & Hausmann, 2009; Romer, 1990; Aghion & Howitt, 1992).

At the heart of this approach is the Economic Complexity Index (ECI), which quantifies an economy's latent productive capabilities by analyzing the diversity and ubiquity of the products it exports. The ECI rests on the premise that sophisticated economies export a broad array of complex products that few others can make. High

ECI scores thus reflect a deep reservoir of productive knowledge, while low scores indicate limited specialization and capability accumulation (Hidalgo et al., 2007; Mealy et al., 2019). The Product Complexity Index (PCI) complements the ECI by measuring the knowledge-intensity required to produce individual products. Both indices are derived using the Method of Reflection, an iterative technique that captures the structure of productive know-how by assessing the diversity of an economy (the number of products it exports) and the ubiquity of each product (how many other economies export it) (Hidalgo & Hausmann, 2009).

Unlike endogenous growth models that treat knowledge as a generic, non-rival input, economic complexity theory emphasizes the non-fungibility and embeddedness of productive capabilities. Knowledge cannot be easily transferred across sectors or regions; it is embedded in networks of firms, institutions, and human skills (Acemoglu, Akcigit, & Kerr, 2016; Hidalgo, 2023). This embedded knowledge constrains or enables structural transformation, depending on how closely new economic activities relate to existing capabilities—a concept formalized through the product space (Hausmann & Klinger, 2007). Recent developments enrich the framework through network science and machine learning techniques, which are used to map patterns of diversification, upgrading, and structural change. Tools such as relatedness-complexity diagrams and product space visualizations support policymakers in identifying feasible pathways for economic transformation (Hausmann et al., 2014; Hidalgo, 2023).

Empirically, the economic complexity framework has been shown to explain variations in income levels, growth trajectories, inequality, and resilience to shocks. High-ECI regions tend to demonstrate stronger growth, more innovation, and more inclusive development outcomes (Hartmann et al., 2017; Balland & Boschma, 2021). At the sub-national level, complexity measures have been applied to states, provinces, and cities to reveal spatial heterogeneities in productive structures (Reynolds et al., 2018; Wang & Turkina, 2020; Roos et al., 2021). Nevertheless, the framework faces limitations. Its reliance on international trade data excludes much of the service sector and informal economy. These limitations are particularly acute at the sub-national level, where trade data is often unavailable. To address this, scholars increasingly use employment data or firm-level information to calculate local complexity indices (Fritz & Manduca, 2021; Török, Benedek, & Gómez-Zaldívar, 2022).

In sum, economic complexity offers a powerful complement to traditional development theory. Rather than focusing solely on increasing factor inputs or adopting external technologies, this framework emphasizes the recombination and accumulation of embedded knowledge as the engine of sustainable structural transformation.

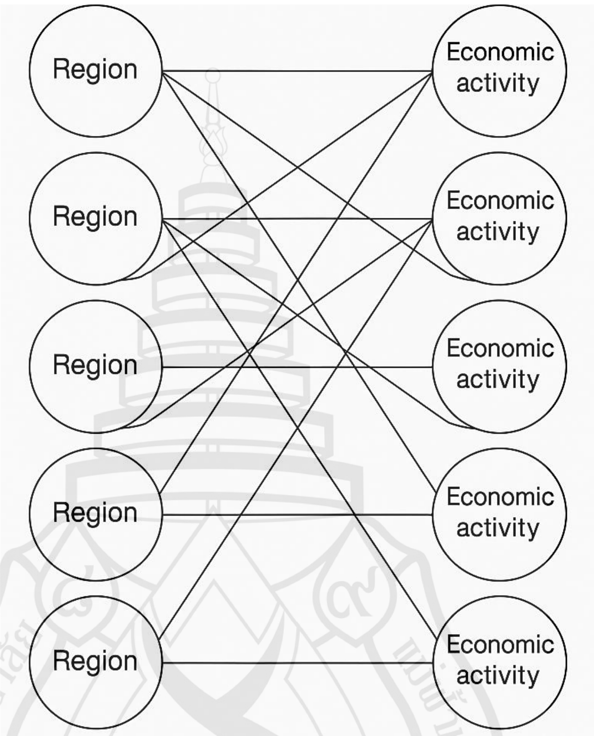
2.1.2 Location Quotient

The Location Quotient (LQ) is a widely used tool in regional economic analysis to identify areas of industrial specialization. It compares the concentration of employment in a specific industry within a province to the concentration of that industry at the national level (Isserman, 1977; Miller, Gibson, & Wright, 1991). An LQ value greater than one indicates that the province has a higher-than-average concentration in that industry, suggesting a comparative advantage or regional specialization. LQ is especially useful in sub-national studies where trade data may be unavailable or limited. It provides a proxy for understanding the relative importance of different sectors across provinces and is often used to inform economic development planning and industrial policy (Billings & Johnson, 2012). In the context of this study, LQ serves as a precursor for constructing the binary matrix required for computing economic complexity indices based on employment data. While LQ helps to identify sectors with regional employment strengths, it does not capture the depth or sophistication of capabilities. Therefore, it is best understood as a diagnostic tool—one that complements more advanced techniques such as the Method of Reflection, which is used to compute the Economic Complexity Index (ECI) and Product Complexity Index (PCI).

2.1.3 Method of Reflection

The Method of Reflection is the foundational algorithm used to calculate the Economic Complexity Index (ECI) and Product Complexity Index (PCI). Developed by Hidalgo and Hausmann (2009), it provides a structured way to infer a country's or region's latent productive capabilities by examining the presence of observed economic output namely, products or activities that a location can produce competitively. Rather than assuming capabilities are directly measurable, the method infers them indirectly from patterns of specialization across a bipartite network of locations and products.

The method begins with a binary matrix that records whether a country or region produces a particular product—or engages in a specific economic activity—with Revealed Comparative Advantage (RCA). This concept, introduced by Balassa (1965), is typically operationalized using a threshold of RCA ≥ 1, indicating that a location is specialized in each product relative to the global average. The matrix structure forms a bipartite network linking locations and products, where each link represents an instance of specialization (see Figure 2.1).

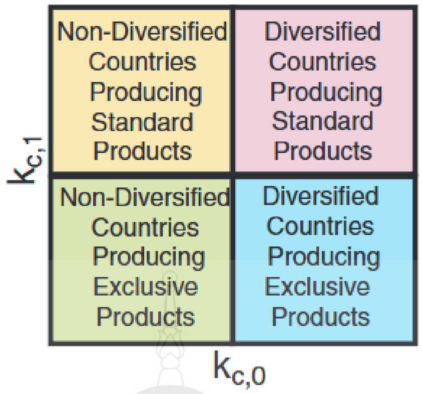


Source Adapted from Hidalgo and Hausmann (2009)

Figure 2.1 Bipartite Network Connecting Regions and Economic Activities

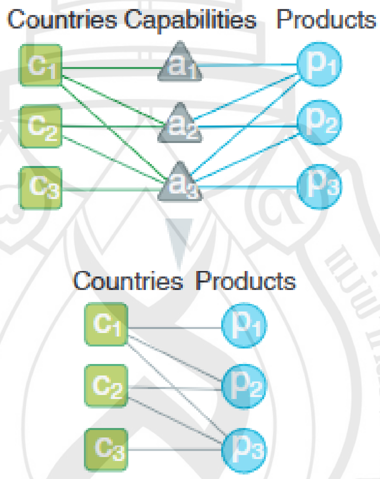
The method infers complexity by evaluating two core dimensions: (1) Diversity: the number of different products or activities a region specializes in (2) Ubiquity: the number of regions that specialize in each product or activity

The logic of this approach is based on the idea that sophisticated regions tend to produce a wide array of rare (non-ubiquitous) products, while less complex regions specialize in common outputs (Hidalgo & Hausmann, 2009; Hidalgo, 2021). This principle is illustrated in Figure 2.2, which maps regions along axes of diversity and ubiquity, identifying high-complexity economies in the bottom-right quadrant—those that are both highly diversified and produce exclusive products.



Source Hidalgo and Hausmann (2009)

Figure 2.2 Typology of countries based on product diversity and ubiquity



Source Hidalgo and Hausmann (2009)

Figure 2.3 Tripartite Network Connecting Countries, Capabilities, and Products

Through iterative averaging, the method updates diversity and ubiquity over multiple rounds. A region that produces a product shared only with other highly diversified regions receives a higher complexity score than one connected to more ubiquitous outputs. The final outcome is a pair of indices—ECI for regions and PCI for products—that quantify the relative depth and uniqueness of their embedded productive knowledge (Hidalgo & Hausmann, 2009; Hidalgo, 2021). Although originally applied to national export data, the Method of Reflection has since been adapted for sub-national contexts using employment, firm registration, or industry-level production data. These adaptations have been particularly valuable in data-constrained settings, where trade statistics are unavailable or incomplete (Mealy et al., 2019). For example, in countries like Thailand, where provincial export data are scarce, complexity can instead be estimated using employment-based matrices constructed from labor force surveys. These applications are discussed further in Section 2.2.

Conceptually, the MoR can also be viewed as a projection from a more granular tripartite network, in which countries are linked to capabilities, and capabilities are linked to products (Hidalgo, 2021). Since capabilities are not directly observable, they are abstracted out of the final country–product matrix, allowing co-occurrence patterns of product specializations to serve as proxies for latent knowledge. This abstraction highlights how the structure of observed outputs reflects deeper productive capabilities. Despite its power and flexibility, the method has limitations. It assumes that all relevant capabilities manifest in measurable outputs, potentially underrepresenting services, informal sectors, or institutional factors (Mealy et al., 2019; Kemp-Benedict, 2014). Moreover, threshold sensitivity such as the choice of RCA ≥ 1 can significantly affect complexity scores and must be carefully calibrated. Ourens (2013) questions the predictive performance of the MoR for long-term growth, while Kemp-Benedict (2014) critiques its use of linear averaging, which may compress variation and mask structural heterogeneity. Mealy et al. (2019) further highlight the need for transparency in interpreting complexity metrics across empirical applications.

Nevertheless, the Method of Reflection remains one of the most widely used and theoretically grounded approaches in complexity economics. Its transparent structure, adaptability to multiple data environments, and empirical relevance make it central to this study's assessment of Thailand's regional development trajectories.

2.1.4 Linking Growth and Inequality Theories

Economic complexity is not only a framework for describing productive structures but also a lens through which scholars and policymakers analyze long-term economic performance and distributional outcomes. Its relevance is grounded in theoretical traditions within both growth economics and the study of income inequality. From a growth perspective, economic complexity resonates most directly with endogenous growth theory, which argues that economic expansion arises from internal factors such as human capital accumulation, innovation, and technological diffusion

(Romer, 1990; Aghion & Howitt, 1992). These models reject the neoclassical assumption of diminishing returns by emphasizing increasing returns from knowledge-based activities. In this view, an economy's capacity to diversify into more sophisticated products reflects the accumulation and recombination of intangible capabilities thus directly linking complexity to sustained growth potential (Hausmann et al., 2014; Mealy et al., 2019). In contrast to earlier theories such as the Solow-Swan model (Solow, 1956), which treated technological progress as exogenous, complexity theory introduces a data-driven mechanism to empirically trace the internal evolution of productive knowledge. This aligns with new structural economics, which also emphasizes upgrading and diversification as central to development (Lin, 2012).

The connection between economic complexity and inequality is supported by both theoretical and emerging empirical insights. On one hand, complex economies often generate more varied and higher-skilled employment opportunities, increasing the absorptive capacity of labour markets and enabling social mobility (Pugliese et al., 2017; Hartmann et al., 2017). Additionally, institutions that support complexity such as education systems, innovation networks, and governance quality tend to reduce exclusion and dualism in labour markets (Andrews et al., 2023). This echoes Kuznets' (1955) classic hypothesis, which suggests that inequality initially rises and later falls as economies structurally transform. On the other hand, spatial inequality may increase when productive capabilities are concentrated in urban cores or advanced regions, leaving peripheral areas in low-complexity traps. This perspective is supported by regional development literature, which shows that capability divergence can reinforce uneven economic geographies (Balland et al., 2020; Rodríguez-Pose, 2018; Chávez et al., 2017). Therefore, economic complexity serves as both a development enabler and a diagnostic tool for understanding how regions differ in their growth trajectories and inequality outcomes.

These theoretical linkages form the conceptual basis for the present study's inquiry into how economic complexity relates to provincial growth and inequality in Thailand. The next sections will review empirical evidence from both national and sub-national studies that test these propositions.

2.2 Measuring Economic Complexity

Building upon the theoretical foundations outlined in the previous section, this part of the chapter examines the methodological approaches used to operationalize economic complexity in empirical research. Central to this endeavor are the Economic Complexity Index (ECI) and the Product Complexity Index (PCI), which serve as quantitative proxies for the latent capabilities embedded within regional or national economies. These indices are grounded in the logic of the Method of Reflection and aim to capture the degree of diversification and specialization present in each economic unit.

Originally constructed using international trade data, the ECI and PCI have since been adapted for use with alternative data sources, most notably, employment-based datasets to facilitate sub-national applications. This is particularly relevant in contexts such as Thailand, where trade data at the provincial level are limited or unavailable. The transition to employment data introduces both new opportunities and methodological challenges, necessitating careful attention to data structure, industry classification, and the criteria for determining economic specialization.

This section is organized into three subsections. The first section outlines the procedures for computing ECI and PCI using employment data. The second section reviews recent empirical studies that apply economic complexity metrics to sub-national regions, highlighting their analytical potential and policy relevance. Third section evaluates the diverse methodological adaptations proposed in the literature to improve the robustness, comparability, and interpretability of complexity measures across different spatial scales and data environments. These discussions collectively inform the measurement strategy adopted in this thesis and lay the groundwork for the empirical analysis that follows.

2.2.1 Employment-Based Economic Complexity

Given the limitations of trade data at the sub-national level, particularly in developing countries where detailed export statistics are often unavailable or unreliable, researchers have increasingly turned to employment data as a practical alternative for constructing economic complexity indices. Employment records are typically more accessible, systematically collected across regions, and capable of capturing a broader range of economic activities, including service-oriented and informal sectors, which are frequently omitted from trade-based datasets (Balland & Rigby, 2017; Mealy et al., 2019). This methodological shift is particularly relevant for countries like Thailand, where provincial-level export data are scarce. In such contexts, employment patterns serve as a valuable proxy for the spatial distribution of productive capabilities, allowing researchers to infer regional economic complexity with greater coverage and frequency. Moreover, employment data facilitates sub-national comparisons and policy-relevant analysis at the provincial or municipal level.

The adaptation of the Economic Complexity Index (ECI) to employment-based datasets involves substituting Revealed Comparative Advantage (RCA) with Location Quotients (LQ), which measure a region's employment specialization in each sector relative to the national average. A value greater than one indicates relative specialization. These LQ values are then converted into a binary matrix that identifies sectors in which a region is competitively engaged. This binary matrix serves as the input for the Method of Reflection, which iteratively computes both the ECI for regions and the Product Complexity Index (PCI) for sectors (Hidalgo & Hausmann, 2009; Mealy et al., 2019). This approach has been applied across diverse national and sub-national contexts, including the United States, the European Union, and Southeast Asia (Fritz & Manduca, 2021; Neffke et al., 2018; Diodato et al., 2018; Török et al., 2022). In Thailand, employment-based ECI was computed using quarterly labor force survey data from 77 provinces and 20 economic sectors, demonstrating the approach's empirical robustness and relevance for sub-national policy design (Yeerong & Uttama, 2023).

Employment-based complexity metrics offer several advantages. They allow for frequent updating, capture non-tradable and informal sectors, and permit high-resolution regional analysis, depending on the granularity of available data (Fritz & Manduca, 2021; Mealy et al., 2019). As shown in recent applications, this methodology is both scalable and replicable, making it well-suited for tracking development potential across Thailand's diverse provincial economies (Yeerong & Uttama, 2023). However, limitations remain. Employment-based methods may conflate sector size with capability intensity, obscure heterogeneity within industries, and are sensitive to classification schemes and threshold definitions. These limitations underscore the need for methodological calibration, robustness checks, and validation procedures (Balland & Rigby, 2017; Mealy et al., 2019).

Despite these challenges, the use of employment data in computing economic complexity provides a viable and policy-relevant alternative to trade-based approaches, particularly in decentralized and data-constrained environments. It offers a meaningful lens through which to assess productive capabilities and structural asymmetries at the sub-national level.

2.2.2 Application at Sub-National Level

Although the Economic Complexity Index (ECI) was originally developed to assess national productive capabilities, its application at sub-national levels, such as regions, provinces, and cities has gained increasing attention. This extension reflects the recognition that productive structures vary significantly across space and that localized economic capabilities shape regional growth trajectories, inequality, and resilience (Balland & Boschma, 2021; Hidalgo, 2023; Mealy & Coyle, 2022).

A growing body of empirical literature demonstrates the feasibility and value of sub-national complexity analysis using alternative data sources. In the United States, metropolitan-level ECI scores based on employment and occupational data correlate strongly with productivity and innovation outcomes (Balland & Rigby, 2017; Fritz & Manduca, 2021). In Europe, similar methods have been applied to Italy, Romania, and Spain to study regional diversification and development (Basile & Cicerone, 2022; Török et al., 2022; Marco et al., 2022). In the United Kingdom, Mealy and Coyle (2022) applied complexity-informed indicators to support local industrial strategy. Complexity metrics have also been employed to study regional inequality and competitiveness in Latin America and the Asia-Pacific. In Mexico, complexity has been linked to state-level income growth and inequality (Chávez et al., 2017; Gómez-Zaldívar et al., 2022). Similar insights have been reported in Brazil (Bandeira Morais et al., 2021), Canada (Wang & Turkina, 2020), and Australia (Reynolds et al., 2018), affirming the adaptability of the complexity framework across diverse contexts.

To construct sub-national ECI measures, researchers often replace trade data with domestic proxies such as employment, patent filings, or firm registration data. Although these substitutions introduce challenges, particularly in capturing informal and non-tradable activities, employment-based models using location quotients and sectoral diversity have proven robust for regional analysis (Mealy et al., 2019; Neffke et al., 2018). In developing economies, where disaggregated trade statistics are typically unavailable, employment-based complexity frameworks offer a practical and scalable alternative (Mealy et al., 2019; Hidalgo, 2023). They enable high-frequency and fine-grained monitoring of economic structure and can inform spatially targeted development policies. Hausmann et al. (2021) further emphasize the importance of place-specific capability assessment for addressing persistent regional income gaps. Thailand's experience with sub-national complexity analysis is still emerging. Apaitan et al. (2017) constructed preliminary ECI estimates for provinces using firm registration data, but data limitations restricted coverage and consistency. A more recent study by Yeerong and Uttama (2023) implemented an employment-based ECI model using labor force survey data across 77 provinces and 20 sectors. By integrating this with clustering techniques, they identified distinct regional development profiles and demonstrated the method's practical utility for spatial planning and regional policy.

The growing adoption of complexity metrics at the sub-national level aligns with efforts to promote smart specialization, regional innovation systems, and inclusive industrial strategies (Balland & Boschma, 2021; Mealy & Coyle, 2022; Hartmann & Pinheiro, 2024). Complexity analysis offers a complementary tool to conventional indicators by highlighting latent capabilities and revealing development constraints that are often hidden in aggregate statistics.

2.2.3 Review of Methods Used in Measuring Complexity

Since its introduction, the Economic Complexity Index (ECI) has evolved significantly to address both conceptual critiques and practical data constraints. Originally developed by Hidalgo and Hausmann (2009), the Method of Reflection quantifies productive capabilities by analyzing the diversity of outputs from a given location and the ubiquity of those outputs across locations. While foundational, this approach has stimulated extensive methodological refinements aimed at improving stability, interpretability, and applicability, especially at the sub-national level (Hidalgo, 2023; Mealy, Farmer, & Teytelboym, 2019).

A core debate in the literature concerns the representation of specialization. The original ECI relies on a binary matrix that indicates whether a country or region specializes in each product or sector. However, scholars have argued that binary classification can lead to information loss and threshold sensitivity, particularly when thresholds such as $LQ \geq 1$ are used (Cristelli et al., 2013; Inoua, 2023; Mariani et al., 2015). In response, various refinements have introduced continuous or weighted matrices to preserve the intensity of specialization and improve robustness (Tacchella et al., 2012). The Fitness–Complexity framework, for instance, models productive knowledge as a non-linear iterative process and has been proposed as a more stable alternative to the Method of Reflection (Cristelli et al., 2013; Pietronero, Cristelli, & Tacchella, 2013). In addition to algorithmic refinements, other researchers have addressed structural limitations of the ECI. Albeaik et al. (2017) proposed modifications to address sparsity and convergence issues in highly skewed datasets, while Lopes, Dias, and Amaral (2012) conceptualized economic complexity through interindustry connectedness using input–output data. These approaches underscore the potential to view complexity not only as an outcome of export or employment patterns but also as a function of sectoral interdependence.

As employment and firm-level data have become more widely available, economic complexity analysis has been extended to sub-national units. A growing number of studies have demonstrated the feasibility of applying ECI methodologies to regions, provinces, or cities using domestic data proxies (Reynolds et al., 2018; Balsalobre, Verduras, & Lanchas, 2017; Pérez-Balsalobre, 2019; Llano Verduras, & Díaz-Lanchas, 2019; Török, Benedek, & Gómez-Zaldívar, 2022). These studies commonly replace trade data with employment-based Location Quotients (LQ) or firm registration records to capture productive specialization within sub-national boundaries.

While this enables the inclusion of non-tradable and informal sectors, it also introduces challenges related to data granularity and comparability across spatial units (Marco, Llano, & Pérez-Balsalobre, 2022; Hausmann, Pietrobelli, & Santos, 2021). Recent methodological innovations have further incorporated clustering, machine learning, and network visualization to construct multi-dimensional regional typologies. In Thailand, for example, Yeerong and Uttama (2023) applied unsupervised clustering techniques to group provinces based on employment-derived ECI scores, providing actionable insights for regional development policy. Such hybrid models reflect the growing trend of integrating complexity metrics with spatial diagnostics and policy design tools.

Nonetheless, key challenges persist. There is no universal consensus on best practices for constructing sub-national complexity metrics, particularly in developing country contexts. Methodological issues such as threshold selection, industrial classification depth, data sparsity, and robustness testing remain unresolved (Inoua, 2023; Mealy et al., 2019). As a result, transparency in metric construction and validation is essential to enhance comparability and policy relevance.

2.3 Regression-Related Literature

The empirical validation of economic complexity theory has increasingly relied on econometric approaches to examine how productive capabilities—captured through indicators such as the Economic Complexity Index (ECI)—affect economic development outcomes. Building on the theoretical foundations discussed earlier, this section reviews regression-based studies that link economic complexity to two major dimensions of development: economic growth and income inequality. A diverse range of regression techniques, including panel data models, instrumental variable regressions, and quantile regression, have been applied to test these relationships at both national and sub-national levels. The literature highlights how higher complexity tends to be associated with more robust growth trajectories and, in some contexts, more equitable income distribution. At the same time, methodological choices—such as the construction of the complexity index, control variables, or the treatment of endogeneity—have significant implications for empirical results.

Accordingly, this section is organized into three parts: the first reviews studies examining the relationship between complexity and growth; the second focuses on complexity and inequality; and the third provides a critical assessment of the econometric methods used across the literature.

2.3.1 Economic Complexity and Economic Growth

Economic complexity theory argues that long-term growth is driven by the diversity and sophistication of an economy's productive capabilities. The Economic

Complexity Index (ECI), introduced by Hidalgo and Hausmann (2009), captures this by quantifying the embedded knowledge required to produce and export a range of goods. Unlike neoclassical models that focus on capital and labor inputs, complexity-based frameworks emphasize the combinatorial capabilities that enable economies to shift into higher-value, knowledge-intensive activities (Aghion & Howitt, 1992; Hausmann et al., 2014; Hidalgo, 2021).

Empirical studies consistently report a positive relationship between ECI and economic growth. Panel analyses show that ECI predicts future income levels more effectively than conventional indicators such as human capital or institutions (Hausmann et al., 2014; Mealy et al., 2019; Zhu & Li, 2017). In developing and resource-dependent economies, greater complexity is associated with improved growth trajectories (Stojkoski & Kocarev, 2017; Tabash et al., 2022). Alternative formulations, such as the fitness–complexity metric, reinforce these findings. By modeling capability accumulation as a non-linear dynamic, this approach improves predictive robustness, especially in emerging economies (Tacchella et al., 2012; Cristelli et al., 2013; Pietronero et al., 2013).

At the sub-national level, complexity has proven equally relevant. Studies in Europe, Australia, and Latin America reveal that more complex regions grow faster and diversify more effectively (Balland & Boschma, 2021; Reynolds et al., 2018; Chávez et al., 2017; Mewes & Broekel, 2022). These outcomes are attributed to agglomeration effects, sectoral spillovers, and regional capacity for innovation and adaptation. In Thailand, Yeerong and Uttama (2023) applied employment-based complexity metrics to 77 provinces and found a significant relationship between ECI and real GPP per capita. Their results demonstrate that complexity is a meaningful predictor of regional economic performance, even in middle-income contexts with limited export data. Overall, economic complexity offers a robust, empirically validated framework for understanding growth across spatial scales. Its capacity to capture latent capabilities and structural transformation makes it a valuable tool for development planning in both national and regional contexts.

2.3.2 Economic Complexity and Income Inequality

Economic complexity provides a compelling structural lens for understanding income inequality by shifting attention from traditional factor endowments to the composition and sophistication of productive activities. Regions and countries with more diverse, knowledge-intensive, and interconnected economies tend to generate broader access to skilled employment, capability spillovers, and innovation participation—thus fostering more inclusive development (Balland & Rigby, 2017; Hartmann & Pinheiro, 2024; Hidalgo, 2021).

Empirical studies across global, national, and regional scales generally affirm an inverse relationship between economic complexity and inequality. Hartmann et al. (2017), Fredrich (2023), and Antonietti (2024) show that higher Economic Complexity Index (ECI) scores are associated with lower income concentration, particularly in institutional contexts that support education, labor inclusion, and industrial diversification. Nguyen et al. (2023), however, introduce a nonlinear perspective, revealing that while complexity initially reduces inequality, its benefits taper off or even reverse at high complexity levels—suggesting diminishing marginal returns or structural dualism in advanced economies.

These effects are shaped by several moderating factors. Lee and Vu (2020) emphasize the role of human capital, showing that complexity's redistributive power is stronger in countries with better educational attainment. Similarly, Lee and Wang (2021) find that the relationship between complexity and inequality is conditioned by country risk, where institutional instability dampens complexity's impact on inclusive growth. Pham et al. (2024) extend this discussion by incorporating the shadow economy as a mediating factor; their study finds that informality weakens the capacity of complex economies to translate productive capabilities into equitable income distribution.

These findings are echoed across the EU (Cota et al., 2023), the G20 (Subekti & Sari, 2024), and developing regions in Africa and Asia (Adeleye, 2024; Bedemo Beyene, 2024; Le et al., 2022; Hoeriyah et al., 2022). National studies from Iran (Khanzadi et al., 2022), Indonesia (Prasetiya, 2021), and Japan (Ikram et al., 2021) provide further evidence that complexity contributes to lowering income inequality in a variety of economic and institutional settings. Beyond income inequality, multidimensional frameworks explore complexity's role in shaping broader development outcomes. Marco et al. (2022) frame the trade-off between complexity, environmental sustainability, and equity as a regional "trilemma," while Ikram et al. (2021) document complexity's joint role in reducing ecological footprint and income disparity. At the sub-national level, the complexity–inequality link is supported by growing evidence. Bandeira Morais et al. (2021), Wang and Turkina (2020), Gómez-Zaldívar et al. (2022), Török et al. (2022), and Hausmann et al. (2021) report negative associations between regional ECI and income inequality across Brazil, Canada, Mexico, Romania, and other contexts. These patterns tend to be strongest in areas with diverse industrial bases, supportive infrastructure, and effective policy alignment.

In sum, while the strength of the complexity–inequality relationship varies across institutional and development contexts, the growing literature supports its potential as a policy lever for inclusive growth. To fully unlock this potential, complexity strategies must be accompanied by robust education systems, formal sector development, institutional stability, and disaggregated data to track how capabilities translate into equitable outcomes.

2.3.3 Methodological Review

Understanding the methodological landscape of economic complexity research is essential to appropriately framing this study's analytical approach. While the relationship between the Economic Complexity Index (ECI) and economic outcomes has been widely investigated, studies differ considerably in their econometric techniques, each offering unique insights into structural transformation and inequality dynamics. This section reviews three principal methods used in the empirical literature; panel regression models, panel quantile regression, and generalized additive models (GAM), which are also employed in this thesis. By organizing the discussion according to these methodological strands, the review highlights both the dominant practices in the field and the rationale behind the integrated approach adopted for provincial-level analysis in Thailand.

2.3.3.1 Panel fixed effects regression

Panel regression is among the most widely employed econometric techniques in empirical research on economic complexity due to its ability to control both temporal dynamics and unobserved heterogeneity across cross-sectional units. In the context of economic complexity, panel regression enables researchers to explore how changes in productive structures captured by the Economic Complexity Index (ECI) affect economic outcomes such as growth and inequality across countries, regions, or provinces over time.

A wide range of studies have adopted the panel fixed effects (FE) model, particularly when regional or country-specific characteristics are suspected to correlate with explanatory variables. For instance, Zhu and Li (2017) applied fixed effects regressions to examine the joint influence of ECI and human capital on long-run growth. At the sub-national level, Török et al. (2022) studied Romanian counties, while Fritz and Manduca (2021) assessed U.S. metropolitan areas both highlighting ECI as a key determinant of regional development. Similarly, Basile and Cicerone (2022) found widening productivity disparities linked to complexity in Italian provinces, and Chávez et al. (2017) and Korkmaz et al. (2024) revealed spatial heterogeneity in complexity-growth relationships across Mexico and Turkey, respectively. Other subnational studies (e.g., Li & Rigby, 2023; Teixeira et al., 2022) further support the relevance of FE models for investigating development patterns in middle-income contexts.

In the Thai context, the FE model has also been employed in several recent studies addressing regional development. Homsombat, Wrasai, and Benjabutr (2025) used panel FE regression to measure the impact of creative city attributes on regional economic performance across Thai provinces. Jaewisorn and Aroonruengsawat (2020) applied panel FE models to investigate the inequality effects of natural disasters in Thailand, emphasizing the importance of controlling for unobserved local characteristics. Likewise, Ashikbayeva et al. (2020) employed a panel fixed effects approach to analyze household-level economic impacts of flooding, further validating the appropriateness of the FE framework for sub-national analysis in the Thai setting.

In this thesis, only the Fixed Effects model is employed, as it provides consistent estimates under the realistic assumption that unobserved provincial characteristics, such as industrial structure, governance capacity, or geographic endowments are correlated with the explanatory variables. The Random Effects model is not considered because its core assumption of no correlation between individual effects and regressors is unlikely to hold in this context. This modeling choice aligns with recent methodological reviews emphasizing robustness in sub-national complexity research (Bandeira Morais, 2023; Bahrami et al., 2023). Furthermore, this approach is grounded in widely accepted econometric literature on panel data modeling (Wooldridge, 2010; Stock & Watson, 2020), which recommends FE models when the goal is to obtain unbiased estimates in the presence of unobserved heterogeneity.

2.3.3.2 Panel quantile regression

Panel quantile regression (PQR) extends the traditional quantile regression framework (Koenker & Bassett, 1978; Koenker, 2005) to panel data settings, offering a robust approach to modeling heterogeneous effects of covariates across the conditional distribution of a dependent variable while accounting for unobserved individual heterogeneity. Unlike mean-based estimators, which yield average treatment effects, PQR captures the influence of explanatory variables at different points quantiles of the outcome distribution. This is especially useful in empirical contexts marked by inequality, divergence, or structural asymmetry, where the marginal effects of covariates such as the Economic Complexity Index (ECI) may vary depending on whether a region lies in the upper or lower tail of the distribution.

In the context of economic complexity, the relationship between ECI and socio-economic outcomes is unlikely to be uniform. For example, ECI may exert a stronger influence in underperforming or highly unequal regions than in more advanced provinces, a nuance that traditional linear models often obscure (Koenker & Hallock, 2001; Hao & Naiman, 2007). This motivates the use of PQR in this thesis to explore how economic complexity differentially impacts provincial income inequality and economic growth across the conditional distribution.

Empirical applications of PQR have grown rapidly in complexity-related research. Chu (2023), Nguyen et al. (2023), and Pham et al. (2024) analyze how ECI affects income inequality at various quantiles, revealing distribution-sensitive policy implications. Adebayo et al. (2022) and Hossain et al. (2024) employ the method of moments quantile regression (MMQR) to explore the nexus between complexity, energy innovation, and environmental performance across diverse country groups. Other scholars such as Ashraf et al. (2023) and Kazemzadeh et al. (2022) incorporate global value chains and ecological risk into quantile frameworks, while Alvarado et al. (2021) examine the distributional effects of natural resource rents. Advanced PQR variants, such as quantile-on-quantile regression (QQR) and generalized panel quantile regression have also been applied to asymmetric risk and energy contexts (Ozkan et al., 2023; Payne et al., 2023).

From a methodological standpoint, several advances have facilitated the estimation of quantile models with fixed effects. Canay (2011) proposed a computationally efficient two-step estimator that first removes fixed effects via mean differencing, followed by quantile regression on the transformed data. Galvao (2011) extends this by developing consistent estimators for nonlinear panel quantile models. Powell (2022) and Graham et al. (2015) offer simulation-based diagnostics and guidance for implementation, while Zhang et al. (2019) apply PQR to data-driven clustering in panel contexts. These techniques are supported by canonical econometrics literature (Stock & Watson, 2020; Wooldridge, 2010), which recommends distributional analysis in settings with structural heterogeneity.

This thesis applies panel quantile regression, following Canay's (2011) two-step estimator, to evaluate whether the effects of economic complexity differ across the distributions of income inequality and economic growth at the provincial level in Thailand. The method proceeds as follows. First, a fixed-effects model is estimated to control for province-specific unobserved heterogeneity:

$$y_{it} = \alpha_i + X_{it}^{\top}\beta + \varepsilon_{it} \qquad (1)$$

where y_{it} is the dependent variable (e.g., GINI or RGPPPC), $X_{it}^{\top}$ is the vector of covariates, and α_i captures province-specific fixed effects. The data are then transformed via de-meaning:

$$\tilde{y}_{it} = y_{it} - \bar{y}_{it},\ \tilde{X}_{it} = X_{it} + \tilde{X}_i \qquad (2)$$

In the second step, quantile regression is applied to the transformed data:

$$\min_{\beta_\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau \left(\tilde{y}_{it} - \tilde{X}_{it}^{\top}\beta_\tau\right) \qquad (3)$$

where $\rho_\tau(\mu) = \mu(\tau - \mathrm{II}\{\mu < 0\})$ is the check function for quantile τ, and II{·} is the indicator function.

By estimating conditional quantiles at the 25th, 50th, and 75th percentiles, this approach uncovers distribution-sensitive dynamics that are overlooked by average-based models. This approach provides a more nuanced understanding of how complexity interacts with Thailand's diverse regional development patterns.

2.3.3.2 Generalized additive models (GAM)

Generalized Additive Models (GAMs) extend the Generalized Linear Model (GLM) framework by allowing for flexible, nonlinear relationships between predictors and the response variable. Originally introduced by Hastie and Tibshirani (1986, 1990), GAMs relax the linear assumption embedded in GLMs by replacing linear terms with smooth, non-parametric functions estimated directly from the data. This innovation enables the modeling of functional relationships that are unknown or poorly specified in theory, making GAMs particularly suited for exploratory analysis in complex empirical settings.

Mathematically, a GAM can be expressed as:

$$y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_p(x_{pi}) + \varepsilon_i \qquad (4)$$

where $f_j(\cdot)$ are smooth functions estimated using methods such as penalized regression splines or thin plate splines. As Wood (2017) elaborates, this framework allows each predictor to have a unique, potentially nonlinear influence on the outcome variable while preserving the additive structure that facilitates interpretation and computational tractability.

Compared to Ordinary Least Squares (OLS) regression, which imposes strict linearity, and quantile regression, which estimates conditional quantiles but still often assumes linearity within each quantile (Koenker & Bassett, 1978), GAMs provide a more flexible approach that does not require pre specifying the functional form of the relationship between independent and dependent variables. This feature is especially valuable in the context of economic complexity, where the relationship between the Economic Complexity Index (ECI) and outcomes such as regional growth or income inequality may exhibit nonlinear thresholds, structural breaks, or diminishing returns. GAMs thus offer a robust method to uncover hidden patterns that would be overlooked under parametric constraints. In empirical economic modeling, standard references such as Wooldridge (2010) and Stock and Watson (2020) underscore the importance of model specification. GAMs contribute to this aim by minimizing misspecification bias and enhancing predictive performance when relationships are complex and context-dependent conditions that frequently characterize the ECI–growth and ECI–inequality nexus in diverse regional settings.

Generalized Additive Models (GAMs) extend the framework of Generalized Linear Models (GLMs) by incorporating smooth, data-driven functions of covariates, allowing for the flexible modelling of nonlinear relationships (Hastie & Tibshirani, 1990; Wood, 2017). Rather than assuming a strictly linear effect, GAMs estimate smooth terms—often via splines—whose complexity is regularized through penalization to avoid overfitting (Wood, 2017; Hastie et al., 2009). This regularization process shrinks overly complex curves toward linearity unless strongly justified by the data. Optimal smoothing parameters are typically selected using Generalized Cross-Validation (GCV) or Restricted Maximum Likelihood (REML) methods (Wood, 2006, 2017; Wahba, 1975, 1990).

Generalized Additive Models (GAMs) have been increasingly utilized in economic research to capture nonlinear and context-dependent relationships that traditional linear models may overlook. In macroeconomic and financial contexts, Sapra (2013) provided an early overview of GAM's potential in modelling economic indicators and business dynamics. More recent applications extend to sectoral and spatial domains: Salan et al. (2023) applied GAMs to explore links between financial reserves and macro indicators in Bangladesh, while Wibowo et al. (2021) examined the non-linear effect of information technology on sub-national GDP. In tourism analysis, Zanin and Marra (2012) showed GAM’s superiority over GLMs in modelling fluctuating tourism demand.

GAMs have also been applied in the context of productivity and regional growth. Azomahou, Diene, and Diene (2013) used a semi-parametric panel model to uncover nonlinearities in productivity growth across countries, offering methodological justification for using GAMs in development studies. Similarly, Paul et al. (2018) investigated the complex non-linear association between GDP and crime categories, and Odell (2009) highlighted structural nonlinearities in long-run economic growth. In regional and spatial analysis, Cajias and Ertl (2018) explored non-linear effects in real estate valuation, demonstrating GAM’s utility for hedonic modelling with spatial features. Huang and Li (2018) contributed a semiparametric model averaging framework for panel data, further reinforcing the methodological flexibility of GAMs in econometric practice. Additionally, studies like Pu et al. (2022) and Hunter et al. (2020) have adopted GAMs to analyze health, environmental, and development interactions with complex regional and institutional patterns.

In this thesis, GAM is employed alongside panel and quantile regression as part of a methodological triangulation strategy. While panel regression addresses time-invariant heterogeneity and quantile regression reveals distributional effects, GAM captures unknown nonlinearities in the relationship between ECI and growth or inequality across Thai provinces. This is especially pertinent in a context marked by spatial, economic, and institutional diversity. By incorporating GAM, this study gains a more nuanced understanding of how economic complexity interacts with sub-national development processes.

2.4 Clustering-Related Literature

Clustering techniques are widely used in economic research to classify regions or provinces into groups with similar structural or socio-economic characteristics. In the context of economic complexity, clustering helps reveal underlying patterns in industrial composition, innovation capacity, and development levels that may not be apparent through regression-based analysis alone. By identifying spatial or structural typologies, clustering enables more targeted interpretation of regional disparities and supports tailored policy design. This section reviews key literature on clustering applications in economic and regional studies, with a focus on methods relevant to subnational economic complexity and development analysis.

2.4.1 Clustering for Regional and Socio-Economic Classification

Clustering techniques have become essential tools for analyzing spatial disparities and classifying subnational socio-economic profiles. These methods group provinces, districts, or regions based on common economic, demographic, or infrastructure characteristics, providing a data-driven basis for policy formulation and regional planning. Rooted in regional development theory (Dawkins, 2003; Antonescu, 2014), clustering serves as a powerful alternative to administrative boundaries by revealing "functional" regional typologies based on empirical indicators.

In the European Union, Pavone et al. (2021) applied multidimensional clustering to support cohesion policy by identifying structurally similar regional groupings, while Topaloglou et al. (2005) constructed typologies of border regions to inform cross-border development initiatives. Similarly, Perafita and Saez (2022) demonstrated the utility of clustering in small-area inequality classification, emphasizing its potential for spatially targeted social interventions. In Italy, Antonicelli et al. (2025) combined K-means clustering and machine learning to map income inequality, reflecting a growing trend toward hybrid clustering approaches in economic geography. Outside of Europe, subnational clustering studies are gaining traction. Chu et al. (2019) employed principal component and cluster analysis to assess urbanization in China's Hunan Province. In Brazil, de Queiroz Stein et al. (2024) utilized agricultural census data to explore regional inequalities through clustering, while Fusco and Perez (2015) applied neural networks to analyze economic complexity across the Indian subcontinent. Southeast Asian applications are emerging as well—Rahadini (2024) and Wardana et al. (2023) examined rural-urban relations and development patterns in Central Java using cluster-based typologies, offering useful analogs for Thai provincial studies.

Despite these advances, methodological challenges persist. There is no consensus on indicator selection, normalization procedures, or determining the optimal number of clusters. Many studies rely on static administrative units, which may not reflect actual socio-economic functionality. In Thailand, empirical studies using clustering methods remain limited, often constrained by data availability, insufficient disaggregation, and short time series. These limitations underscore the need for more systematic, reproducible clustering frameworks to support evidence-based subnational development strategies.

2.4.2 Machine Learning Techniques for Clustering

Unsupervised machine learning algorithms such as K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) have become key tools for classifying regions based on socio-economic similarities. Unlike fixed administrative classifications, these techniques uncover data-driven groupings that reflect underlying structural patterns. Their strength lies in handling high-dimensional, nonlinear data to reveal latent spatial disparities and economic profiles. In regional development research, such methods support more precise typologies for policy targeting. The following subsections outline the core principles and applications of K-means, HAC, and GMM in the context of subnational socio-economic analysis.

2.4.2.1 K-means clustering

K-means clustering, first introduced by MacQueen (1967), is one of the most widely used unsupervised learning algorithms for partitioning data into homogenous groups based on similarity metrics. It is particularly favored in socio-economic and regional studies due to its simplicity, scalability, and computational efficiency (Jain, 2010; Everitt et al., 2011). From a methodological standpoint, K-means operate by minimizing intra-cluster variance while maximizing inter-cluster separation. The algorithm iteratively assigns observations to the nearest centroid and recalculates cluster centers until convergence (Kanungo et al., 2000). Several studies have focused on refining the algorithm's initialization and optimizing the selection of the number of clusters (K), which is a critical parameter influencing its accuracy and interpretability (Yuan & Yang, 2019; Sinaga & Yang, 2020).

In terms of application, K-means has been used extensively for socio-economic classification and regional typology. Antonicelli et al. (2025) utilized K-means alongside machine learning algorithms to detect income inequality patterns across Italian regions. Similarly, Fa'rifah and Pramesti (2022) applied K-means to categorize Indonesian districts by their level of inclusive economic development. In China, Zhan et al. (2021) used a combination of K-means and principal component analysis (PCA) to divide economic regions, highlighting the algorithm's flexibility in integrating with dimensionality reduction techniques. Despite its popularity, K-means is not without limitations. It assumes spherical clusters of similar size and is sensitive to outliers and the initial choice of centroids (Chong, 2021). Improvements and hybrid versions have emerged to address these weaknesses, including enhanced distance metrics and initialization procedures (El Hatimi et al., 2024).

Overall, K-means remains a powerful baseline method for regional classification in economic development research, particularly when supplemented with rigorous pre-processing, validation, and visualization techniques.

2.4.2.2 Hierarchical agglomerative clustering (HAC)

Hierarchical Agglomerative Clustering (HAC) is one of the most classical and widely applied clustering techniques, particularly valuable in socio-economic and regional typology research due to its interpretability and flexibility. Introduced by Johnson (1967) and further refined through methods such as Ward's minimum variance criterion (Ward, 1963), HAC builds a nested tree-like structure (dendrogram) that captures the hierarchical relationships among spatial units based on similarity metrics.

The popularity of HAC in socio-economic studies stems from its ability to generate interpretable cluster hierarchies without requiring prior knowledge of the number of clusters. As Murtagh and Contreras (2012) and Ran et al. (2023) note, the algorithm's iterative bottom-up approach is particularly well-suited for classifying regions where economic relationships are complex and multi-scalar. Several studies have adapted HAC for regional development analysis. For instance, Argüelles et al. (2014) employed hierarchical clustering on principal components (HCPC) to identify functionally similar regions, offering an empirically grounded alternative to administrative classifications. Bhahari and Kusnawi (2024) applied a similar PCA-HAC framework to group districts in East Java by socio-economic indicators, revealing intra-provincial disparities. In Turkey, Altuntas et al. (2022) combined HAC with panel data to evaluate regional incentives, demonstrating its potential for dynamic policy analysis. Advancements in spatial statistics have also extended HAC's utility. Carvalho et al. (2009) proposed a spatial variant of HAC that incorporates geographic proximity, allowing the formation of clusters that reflect both socio-economic similarity and spatial contiguity—an essential feature for regional studies in countries with diverse geographies like Thailand. Meanwhile, Ma et al. (2005) explored HAC in clustering Chinese regions, reinforcing its applicability in emerging economy contexts.

Despite its strengths, HAC can be computationally intensive for large datasets and sensitive to the choice of distance metrics and linkage criteria. Nonetheless, its visual clarity through dendrograms and adaptability to mixed-type data make it a valuable tool for sub-national classification in economic development research.

2.4.2.3 Gaussian mixture models (GMM)

Gaussian mixture models (GMM) are a powerful probabilistic clustering approach that assumes data are generated from a mixture of several Gaussian distributions, each representing a latent group. Unlike hard clustering methods such as K-means, GMM assigns probabilities of membership to each observation, thus accommodating overlapping clusters and complex data structures. The theoretical foundation for GMM lies in the Expectation-Maximization (EM) algorithm developed by Dempster et al. (1977), which iteratively estimates model parameters by maximizing the likelihood of the observed data. Comprehensive treatments of GMM and mixture models are provided in McLachlan and Peel (2000) and Bishop and Nasrabadi (2006), while Fraley and Raftery (2002) advanced the model-based clustering literature by incorporating formal criteria such as the Bayesian Information Criterion (BIC) to determine the number of components.

In the context of regional and socio-economic classification, GMM has proven especially effective due to its flexibility in handling heterogeneity and latent group structures. For instance, Rodriguez Andres et al. (2022) used model-based GMM clustering to classify African countries according to knowledge economy indicators, allowing for nuanced groupings that go beyond conventional classifications. Similarly, Wahidah and Utari (2023) applied GMM to poverty indicators in Indonesia and found that the model outperformed K-means in detecting subtle differences between regional profiles. Tanujaya et al. (2024) compared GMM with K-means and hierarchical clustering for categorizing countries by economic freedom, with GMM demonstrating superior performance in capturing underlying data variability.

More technical enhancements to GMM have also emerged. He et al. (2025) proposed a subspace-based GMM ensemble algorithm designed for high-dimensional socio-economic datasets, showing that ensemble methods can improve both clustering stability and interpretability. Additionally, Monastiriotis (2009), while not using GMM directly, emphasized the importance of detecting consistent spatial association patterns across regional indicators—an area where GMM excels due to its probabilistic modeling capacity. Despite its advantages, GMM comes with computational demands and sensitivity to initialization, particularly in high-dimensional or sparse data contexts. However, when applied rigorously with proper model diagnostics and regularization, GMM can offer deep insights into regional typologies, socio-economic segmentation, and policy-relevant heterogeneity.

In sum, GMM adds substantial value to the suite of clustering techniques used in regional development analysis. Its application in this study reflects the need to identify latent structures within provincial socio-economic data in Thailand, enabling a more flexible and statistically grounded classification of provinces.

2.5 Synthesis and Research Gaps

The literature reviewed throughout this chapter illustrates the growing relevance of economic complexity as a framework for understanding long-term economic performance and distributional dynamics. Panel regression has been the predominant methodological approach, offering a robust mechanism to control for both temporal and cross-sectional heterogeneity. Complementary to this, quantile regression allows for the exploration of distributional asymmetries, shedding light on how the impact of the Economic Complexity Index (ECI) varies across different segments of the income distribution. Generalized Additive Models (GAMs) further enrich this methodological arsenal by modelling nonlinear, non-parametric relationships, which are especially useful when theoretical functional forms are ambiguous or vary across regions. In parallel, clustering techniques—such as K-means, hierarchical agglomerative clustering (HAC), and Gaussian Mixture Models (GMM) have been increasingly adopted for the classification of regions based on multidimensional socio-economic indicators. These tools have proven effective in identifying development patterns and regional disparities, supporting spatial planning and policy targeting. However, most clustering applications remain exploratory in nature and are rarely integrated with econometric models to explain or predict economic outcomes.

Despite these advances, several research gaps remain. First, empirical applications at the subnational level, particularly in developing and emerging economies like Thailand are still scarce. Much of the existing literature focuses on cross-country analyses or large economies with abundant data, leaving Thailand's provincial dynamics underexplored. Second, while economic complexity has been linked to both growth and inequality in the global literature, few studies examine these relationships jointly in a localized, panel-based framework. Third, methodological pluralism is often lacking; studies tend to rely on a single econometric technique, with limited triangulation across models to validate findings or capture nuanced effects. Fourth, while clustering has been used to classify regions, these groupings are seldom used as inputs into further econometric analysis, nor are they tailored to reflect complexity-specific dimensions.

This thesis responds to these gaps by conducting a comprehensive empirical investigation of the impact of economic complexity on both economic growth and income inequality at the provincial level in Thailand. It employs a triangulated methodology that combines panel regression, quantile regression, and GAM to account for heterogeneity in scale, distribution, and functional form. Moreover, it incorporates machine learning-based clustering to classify provinces into distinct socio-economic profiles, which can inform and enrich the interpretation of econometric results. By bridging subnational evidence, methodological rigor, and machine learning integration, this study advances the understanding of how economic complexity operates across different development contexts and institutional settings.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter outlines the methodological approach adopted to investigate the relationship between economic complexity and regional development outcomes in Thailand. The study employs a multi-method framework that integrates panel regression, panel quantile regression, generalized additive models (GAM), and clustering techniques to capture both linear and nonlinear dynamics, as well as regional heterogeneity.

The research design is grounded in the use of panel data at the provincial level, covering multiple years to account for both temporal and cross-sectional variation. Economic complexity is operationalized through the Economic Complexity Index (ECI), calculated using the Method of Reflection and applied at the subnational scale. Key outcomes of interest include economic growth and income inequality, measured through standard macroeconomic indicators. To ensure analytical rigor and robustness, the chapter is structured around the sequential stages of data processing and analysis. It begins with a description of data sources and variable construction, followed by the procedures used to compute ECI at the provincial level.

Subsequently, the chapter presents the econometric modelling strategies used to estimate the impact of ECI on growth and inequality, including fixed and random effects models, panel quantile regression, and GAMs. The final sections describe the clustering algorithms used to classify provinces based on their socio-economic profiles and discuss diagnostic and robustness checks applied to validate model performance. This multi-layered methodological approach enables a comprehensive assessment of the complexity-development nexus in Thailand, capturing both average trends and distributional nuances while uncovering spatial patterns that inform regional policy.

3.1 Research Design and Analytical Framework

This study adopts a quantitative research design underpinned by a multi-method econometric and machine learning approach to analyze the relationship between economic complexity and regional development outcomes, specifically, economic growth and income inequality at the subnational level in Thailand. The research framework integrates three core components: (1) econometric modeling using panel regression and panel quantile regression; (2) nonlinear modeling via Generalized Additive Models (GAM); and (3) unsupervised machine learning techniques for provincial classification through clustering algorithms.

The analytical process is structured to capture both average and distributional effects of economic complexity, while also accounting for nonlinearity and regional heterogeneity. First, the Economic Complexity Index (ECI) is computed at the provincial level using the Method of Reflection, based on export and production data. These values are then used as key explanatory variables in panel regression models both fixed and random effects to assess their impact on growth and inequality over time. To examine how these relationships differ across income distribution, panel quantile regression is applied. In addition, GAM is employed to flexibly model nonlinear effects of ECI, enabling the detection of structural thresholds or saturation points in the complexity–development relationship.

To complement these regression-based analyses, the study also uses clustering techniques—K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM)—to classify Thai provinces into distinct socio-economic typologies. This classification allows for comparative regional interpretation and supports more tailored policy insights.

The overall research design emphasizes methodological triangulation to enhance the robustness and depth of findings. Each method is selected for its ability to address a specific empirical dimension—causal inference, distributional variation, structural nonlinearity, or spatial pattern recognition within a unified analytical framework. The combination of econometric rigor and machine learning adaptability positions this study to generate comprehensive insights into the dynamics of economic complexity and regional development in Thailand.

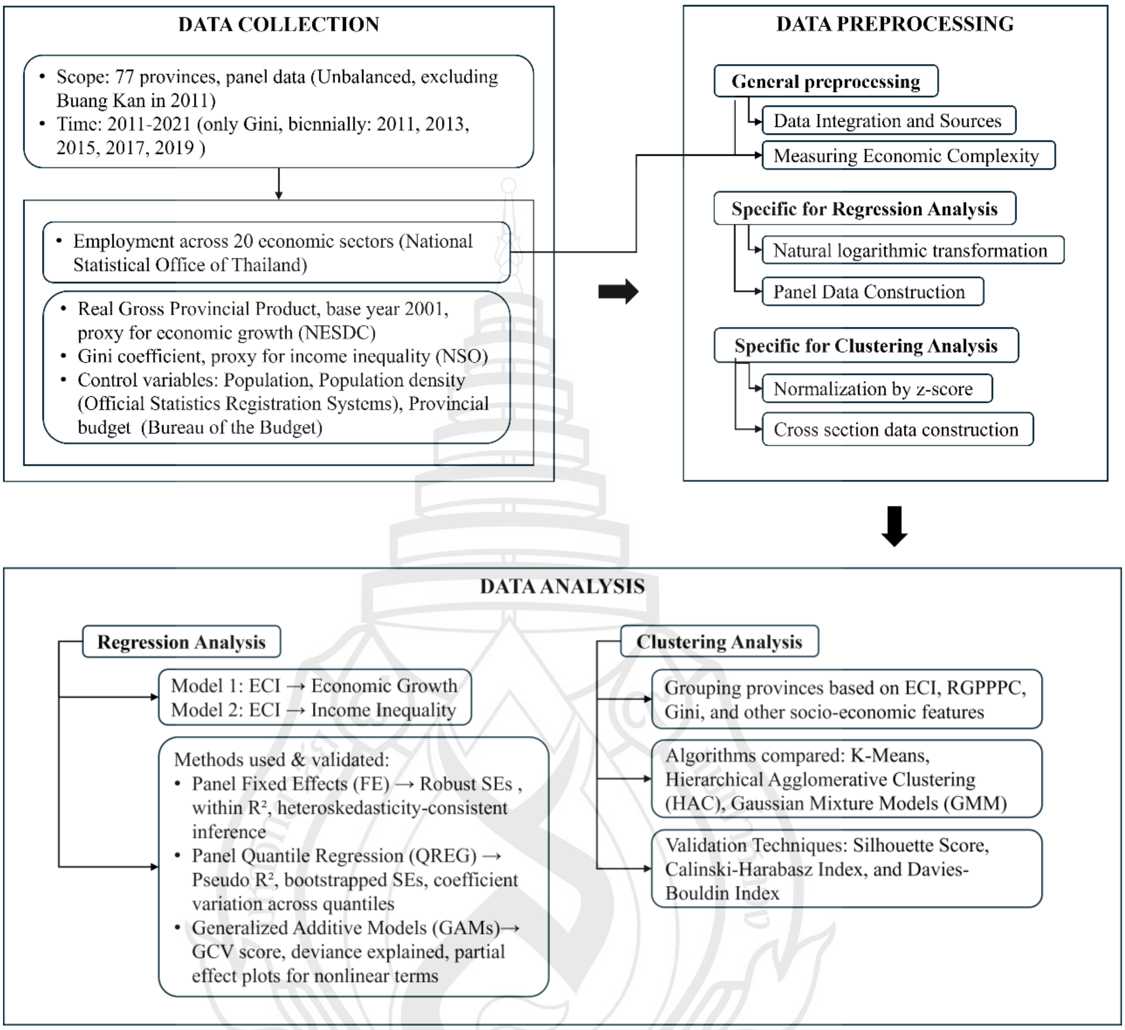


Figure 3.1 Overall Methodological Workflow

The methodological framework guiding this study integrates quantitative econometric modeling and machine learning to examine the multi-dimensional impact of economic complexity. As shown in Figure 3.1, the methodology is structured into two analytical tracks. The first employs panel regression analysis to test hypotheses regarding the effect of complexity on provincial growth and income inequality. The second uses unsupervised machine learning to detect latent clusters among provinces based on socio-economic indicators. This dual-track approach facilitates a more holistic understanding of the interplay between economic complexity and development outcomes. It allows the study not only to estimate causal relationships but also to explore spatial heterogeneity and structural typologies. The integration of these methods ensures empirical rigor and enhances the policy relevance of the findings.

3.2 Data Collection

This study relies on secondary data obtained from official Thai government agencies to conduct a subnational analysis of economic complexity, economic growth, and income inequality across Thailand's 77 provinces. The selection of data sources, indicators, and temporal coverage was guided by both conceptual relevance and empirical feasibility, ensuring alignment with the study's analytical objectives. By combining data from multiple institutional sources, including the National Statistical Office (NSO), the Office of the National Economic and Social Development Council (NESDC), and the Bureau of the Budget. The research constructs a balanced and multidimensional panel dataset. This dataset supports both econometric estimation and unsupervised machine learning techniques for spatial clustering. The collection process emphasizes consistency, comparability, and granularity across provinces and years, laying the empirical foundation for subsequent analysis.

3.2.1 Study Area and Units of Analysis

The spatial units of analysis in this study are Thailand's 77 provinces, which serve as the fundamental territorial units for subnational economic assessment. Each province represents a distinct regional economy with unique industrial compositions, demographic structures, and institutional capacities. The use of provinces as analytical units allows for a granular investigation into how economic complexity relates to regional development outcomes such as income inequality and economic growth. A subnational approach is particularly relevant in Thailand, where economic disparities between urban and rural provinces remain significant and where national-level aggregates may mask underlying heterogeneity.

Special consideration is given to Bueng Kan province, which was established in 2011 after being separated from Nong Khai. Due to its administrative inception during the study period, Bueng Kan lacks a complete data record for the baseline year (2011). As such, while the province is included in the panel dataset, any analyses involving year-to-year comparisons may treat its 2011 data as missing. This approach ensures consistency in the panel structure without introducing distortions in the overall estimation or clustering framework.

3.2.2 Time Frame and Panel Design

The temporal scope of this study spans five selected years: 2011, 2013, 2015, 2017, and 2019. These years were strategically chosen to align with the availability of key outcome variables, particularly the Gini index for measuring income inequality, which is reported biennially through the Household Socio-Economic Survey conducted by the National Statistical Office (NSO). While other datasets—such as labor force statistics and Gross Provincial Product (GPP)—are available annually, the Gini coefficient dictates the maximum feasible temporal coverage for a consistent and meaningful panel structure encompassing all core variables.

The panel dataset is designed to be balanced across the five selected years for all 77 provinces, with one exception: Bueng Kan province, which was created in 2011 and thus lacks a full dataset for that initial year. Despite this minor data omission, the panel remains analytically robust. Two separate panel structures are constructed for the econometric analysis. The economic growth model utilizes data from all five years (2011–2019) to maximize the temporal dimension, while the income inequality model is limited to the biennial years for which Gini data is available. For the clustering analysis, the same five years are used to capture changes in socio-economic profiles over time. This panel design supports both cross-sectional and longitudinal analyses, enabling the investigation of dynamic trends in subnational development.

3.2.3 Data Sources and Instruments

This study relies exclusively on secondary data sourced from reputable Thai government agencies to ensure accuracy, completeness, and nationwide coverage. The National Statistical Office (NSO) serves as the primary provider of demographic and household-level data. Specifically, the Labor Force Survey (LFS) is used to construct the Economic Complexity Index (ECI) by analyzing sectoral employment distributions across provinces, while the Household Socio-Economic Survey (HSES) supplies data for computing the Gini coefficient, the core indicator of income inequality. The NSO's official registration data also provides annual records on provincial population and population density, which serve as key control variables.

Macroeconomic indicators are obtained from the Office of the National Economic and Social Development Council (NESDC), which publishes annual Gross Provincial Product (GPP) data. This is adjusted to constant 2001 prices and used to calculate real GPP per capita (RGPPPC), the dependent variable in the economic growth model. Finally, fiscal data is retrieved from the Bureau of the Budget, which reports per capita budget allocations and provincial expenditures, used here to proxy public investment and fiscal capacity. Each of these datasets plays a critical role in operationalizing the study's key variables: RGPPPC as the measure of economic growth; Gini index for income inequality; ECI as the structural explanatory variable derived via location quotients (LQs) from employment data; and additional control variables such as population size, density, and public spending. The integration of these data sources enables a multidimensional analysis of economic complexity at the provincial level. A summary of all data instruments, including their sources, frequency, and coverage, is provided in Table 3.1 below.

Table 3.1 Summary of Data Sources, Frequency, Coverage, and Variables

Source	**Dataset**	**Frequency**	**Coverage**	**Key Variables**
National Statistical Office	Labor Force Survey (LFS)	Quarterly	2011–2019	Sectoral employment (ECI, LQ)
	Household Socio-Economic Survey (HSES)	Biennial	2011, 2013, 2015, 2017, 2019	Gini coefficient
	Official Statistics Registration	Annual	2011–2019	Population, population density
NESDC	Gross Provincial Product	Annual	2011–2019	RGPPPC
Bureau of the Budget	Provincial Budget Allocation	Annual	2011–2019	Per capita fiscal data (control)

3.3 Data Pre-processing

Prior to conducting the regression and clustering analyses, the dataset must be carefully prepared to ensure accuracy, consistency, and compatibility with each methodological framework. This section outlines the data pre-processing steps applied to construct a multi-year provincial panel dataset using a combination of economic complexity indicators, socio-economic statistics, and public finance variables. Since this study applies multiple empirical techniques—including fixed effects regression, panel quantile regression, and K-Means clustering—data preparation is organized into three stages: (1) general data harmonization, (2) regression-specific processing, and (3) clustering-specific preparation.

The economic growth regression model uses a balanced panel of 77 Thai provinces spanning the years 2011 to 2021, while the income inequality model is limited to five benchmark years (2011, 2013, 2015, 2017, and 2019) due to data availability on the Gini coefficient. The clustering analysis also draws on the same five benchmark years, aligning with the inequality dataset for temporal consistency. This structured pre-processing ensures methodological rigor and enables valid comparison of results across models and time periods.

3.3.1 General Pre-processing

Before conducting the main analysis, all datasets underwent a series of general pre-processing steps to ensure consistency, completeness, and analytical readiness. These procedures included handling missing values, aligning temporal and spatial coverage across variables, and converting raw employment data into a structured panel format. Employment data from the Labor Force Survey (LFS) was aggregated by province and sector to match the required 20-sector classification. Additionally, numeric variables such as population, density, provincial budget, and real GPP per capita were transformed into natural logarithmic form (except for ECI) to reduce skewness and enable interpretation of elasticities in regression models. All datasets were merged based on consistent province-year identifiers, and provinces with incomplete data, such as Bueng Kan in 2011—were excluded from relevant analyses to maintain panel balance.

3.3.1.1 Data integration and sources

The first step of the pre-processing process involved merging and harmonizing data from multiple official sources to construct a coherent panel dataset of Thai provinces. The dataset integrates six core variables across five benchmark years—2011, 2013, 2015, 2017, and 2019—including: (1) the Economic Complexity Index (ECI), (2) real gross provincial product per capita (RGPPPC), (3) the Gini index of income inequality, (4) total population, (5) population density, and (6) provincial government budget allocations. These indicators were obtained from a range of government databases, including the Department of Employment (for sectoral employment data), the National Statistical Office (NSO), the Local Administrative Organization (LAO) database, and the Office of the National Economic and Social Development Council (NESDC).

To facilitate accurate merging, province names and codes were standardized across all data sources. This involved resolving inconsistencies in spelling, formatting, or administrative labeling—for example, ensuring that "Chiang Mai" was consistently represented across employment, budget, and socio-economic datasets. A consistent year identifier was also used to align all indicators temporally, and each observation was uniquely defined by a province–year combination. The merged dataset was structured in long-panel format, enabling its use in time-series econometric analysis and panel-based clustering.

After integration, all variables were reviewed for completeness and internal consistency. Provinces with missing or incomplete records for any variable essential to a specific model were removed on a listwise basis. For example, the Gini index, which is derived from the Household Socio-Economic Survey and published biennially was only available in five years, thereby restricting the inequality model to those specific waves. In addition, Bueng Kan province was excluded from the 2011 wave due to the unavailability of baseline data, as it had just been officially established that year.

To support the broader econometric framework, two final datasets were constructed: one for panel regression, which includes interpolated ECI values for missing years between 2011 and 2019 (and extends to 2021 for economic growth analysis), and another for clustering analysis, consisting of five separate cross-sectional snapshots corresponding to the benchmark years. This structured foundation ensures both the temporal consistency and analytical compatibility of the dataset for the subsequent quantitative models.

3.3.1.2 Measuring economic complexity

This subsection outlines the procedures used to construct the Economic Complexity Index (ECI) at the provincial level using employment data. Given the absence of subnational trade statistics in Thailand, the study adopts an employment-based approach to measure productive knowledge embedded in regional labor markets. This methodology aligns with recent literature that repurposes the Economic Complexity framework from export data to employment structure, enabling analysis of regional capabilities and industrial diversity (Fritz & Manduca, 2021; Mealy et al., 2019; Neffke et al., 2018).

The ECI computation follows the Method of Reflection framework originally developed by Hidalgo and Hausmann (2009), adapted to the subnational context by treating provinces as the equivalent of countries and economic sectors as products. This requires a binary province-sector matrix of revealed comparative advantage (RCA), derived from Location Quotients (LQs) based on sectoral employment shares. The iterative algorithm produces two central measures: the Economic Complexity Index (ECI) for each province and the Product Complexity Index (PCI) for each sector. To ensure robustness, the LQ values were calculated from cleaned and standardized Labor Force Survey data, disaggregated across 20 economic sectors for all 77 provinces. The RCA matrix was binarized based on an LQ threshold of 1.0, which signifies a province's relative specialization in each sector. These binary values serve as the input for the eigenvector-based complexity computation.

Table 3.2 The Sample Record of Employment Data

Province	Industry	Year	No. of employee
ACR	Agriculture, Forestry and Fisheries	2011	140,922
ACR	Mining and Quarrying	2011	2
ACR	Manufacturing	2011	12,025
ACR	Electricity, Gas and Steam	2011	1,005
ACR	Water Supply and Wastewater Management	2011	589

Table 3.2 (continued)

Province	Industry	Year	No. of employee
ACR	Construction	2011	8,102
ACR	Wholesale and Retail Trade	2011	26,588
ACR	Transportation and Storage	2011	4,533
ACR	Hotel and Food Services	2011	5,436
ACR	Information and Communication	2011	120
ACR	Financial and Insurance Activities	2011	540
ACR	Real Estate Activities	2011	20
ACR	Professional, Scientific and Technical Activities	2011	50
ACR	Administrative and Support Service Activities	2011	1,005
ACR	Public Administration and Defence	2011	9,897
ACR	Education	2011	10,514
ACR	Human Health and Social Work Activities	2011	1,566
ACR	Arts, Entertainment and Recreation	2011	222
ACR	Other Service Activities	2011	2,427
ACR	Domestic Workers	2011	640

A sample of the processed employment data used in the LQ calculations is presented in Table 3.2, demonstrating the granularity of sectoral disaggregation prior to complexity analysis. The computations were implemented using the ‘ecomplexity 0.5.2’ Python package developed by the Growth Lab at Harvard’s Center for International Development (MIT, 2018), which provides standardized tools for calculating both national and sub-national complexity indices. This package ensures methodological consistency with global studies on economic complexity, offering functions for location quotient calculation, LQ matrix construction, and iterative derivation of ECI and PCI. The use of an open-source, academically validated toolkit enhances the reproducibility and transparency of the analysis.

Like the Balassa Index or Revealed Comparative Advantage (RCA), the Location Quotient (LQ) is used to evaluate the comparative advantage of a specific region. However, unlike RCA, which is typically based on export data, LQ relies on employment data. This distinction is particularly useful in contexts where trade data at the sub-national level is not systematically available. The LQ is calculated as the ratio between a province's employment share in a specific industry and the corresponding national employment share. This method allows for the estimation of the economic influence or amplifying effect that an industry may have within a local economy (Isserman, 1977; Billings & Johnson, 2012).

The formula for LQ is defined as follows:

$$LQ_{ij} = \frac{E_{ij}/\Sigma_j E_{ij}}{\Sigma_j E_{ij}/\Sigma_{ij} E_{ij}} \qquad (5)$$

where $\mathrm{E_{ij}}$ represents employment in industry j within province i, $\Sigma_j \mathrm{E_{ij}}$ is the total employment in province i, $\Sigma_i \mathrm{E_{ij}}$ is the total national employment in industry j, $\Sigma_{ij} \mathrm{E_{ij}}$ is total employment across all provinces and industries.

An LQ value equal to or greater than 1 indicates that the province's share of employment in a specific industry exceeds the national average, signifying specialization or comparative advantage in that industry. These values are subsequently transformed into a binary matrix $\mathrm{M_{ij}}$, where $\mathrm{M_{ij}} = 1\ if\ \mathrm{LQ_{ij}} \geq 1\ and\ \mathrm{M_{ij}} = 0$ otherwise. This matrix forms the foundation for the economic complexity calculations by identifying industries of relative importance within each province (Fritz & Manduca, 2021; Mealy et al., 2019).

To quantify the complexity embedded within a province's economic structure, this study adopts the Method of Reflection (MoR), an iterative algorithm developed by Hidalgo and Hausmann (2009). This technique leverages the binary specialization matrix $\mathrm{M_{ij}}$, derived from the Location Quotient (LQ), to compute two fundamental metrics: diversity and ubiquity. Diversity reflects the number of economic activities in which a province exhibits specialization, while ubiquity captures the number of provinces specialized in each economic activity. Through iterative refinement, this method captures the embedded knowledge and productive capabilities that underlie regional economies.

The initial diversity $k_{i,0}$ of province i is computed as the sum of industries in which it holds a comparative advantage:

$$k_{i,0} = \Sigma_j M_{ij} \quad (6)$$

Likewise, the initial ubiquity $k_{j,0}$ of economic activity j is calculated as the number of provinces with specialization in that activity:

$$k_{j,0} = \Sigma_i M_{ij} \quad (7)$$

The algorithm proceeds by iteratively updating these values to incorporate second-order effects—namely, how diversified the other provinces are that share similar industries, and how ubiquitous the other industries are in which a province is involved. These recursive steps are defined as:

$$k_{i,n} = \frac{1}{k_{i,0}} \sum_j M_{ij}\, k_{j,n-1} \ ; \ k_{j,n} = \frac{1}{k_{j,0}} \sum_i M_{ij}\, k_{i,n-1} \quad (8)$$

where n denotes the number of iterations. Typically, the process converges within 20–30 iterations, yielding a stable characterization of the complexity profile.

To facilitate meaningful comparisons across time and regions, the final complexity scores are standardized using z-scores. The Economic Complexity Index (ECI) for province i, and the Product Complexity Index (PCI) for industry j, are defined as:

$$ECI_i = \frac{k_{i,N} - \mu(k_{i,N})}{\sigma(k_{i,N})} \ ; \ PCI_i = \frac{k_{j,N} - \mu(k_{j,N})}{\sigma(k_{j,N})} \qquad (9)$$

where μ and σ represent the mean and standard deviation, and N is the final iteration step. A value of ECI or PCI above zero indicates higher-than-average complexity, while values below zero signify relatively less complex provinces or industries (Hausmann & Hidalgo, 2014).

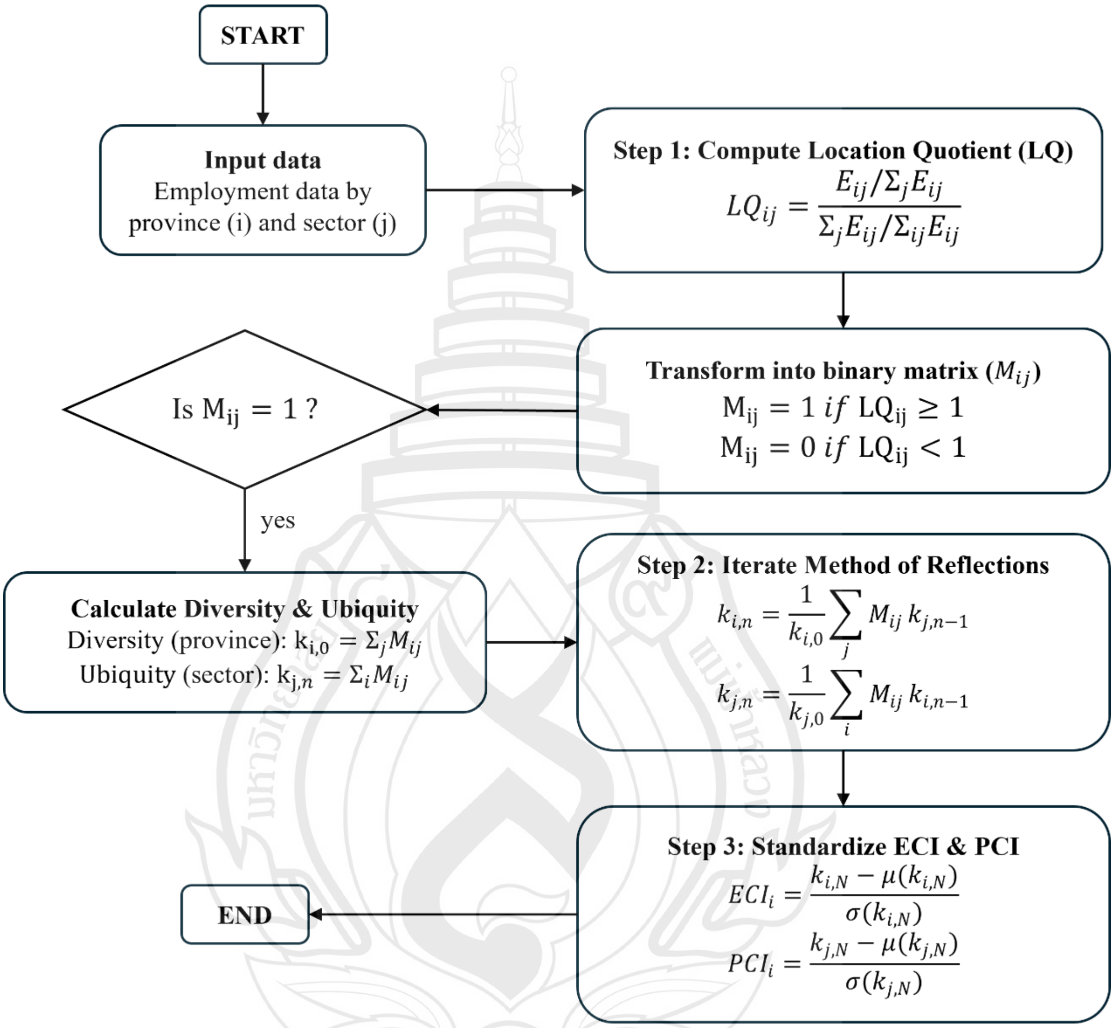


Figure 3.2 Flowchart of the Three-Step Process Used to Calculate the Economic Complexity Index for Thailand's Provinces

This diagram (Figure 3.2) illustrates the process of transforming provincial employment data into a binary specialization matrix via location quotients (LQ), followed by iterative calculation of diversity and ubiquity metrics using the MoR algorithm. Final indices (ECI and PCI) are standardized using z-transformation after convergence is reached

3.3.1.3 Data cleaning, standardization, and structuring

After merging data from multiple sources, a series of data cleaning and standardization procedures were conducted to ensure consistency across provinces and years. Province names and codes were harmonized to address mismatches, while incomplete observations were excluded on a listwise basis according to each model's requirements. Time alignment across variables was verified, particularly for ensuring compatibility in panel settings.

All monetary variables—such as real gross provincial product per capita (RGPPPC) and provincial budgets—were adjusted for inflation and transformed into natural logarithms to reduce skewness and allow for elasticity-based interpretation. Population and population density were similarly log-transformed. For clustering analysis, all variables were further standardized using Z-scores to equalize scales and prevent distortion in centroid calculations. Outliers were assessed for plausibility using descriptive checks and national statistics.

Table 3.3 Panel Structure Overview by Model Type

Model Type	Years Covered	Unit of Analysis	Key Variables	Data Structure
Economic Growth Model	2011-2021	77 Provinces	RGPPPC, ECI, Controls	Unbalanced Panel
Income Inequality Model	2011, 2013, 2015, 2017, 2019	77 Provinces	Gini, ECI, Controls	Unbalanced Panel
Clustering Analysis	2011, 2013, 2015, 2017, 2019	77 Provinces	All Socioeconomic Indicators	Cross-sectional Snapshots

Once cleaned and transformed, the data was structured into long-format panel and cross-sectional formats based on the analytical needs of each model. The economic growth model covers the full period from 2011 to 2021, using interpolated values of the Economic Complexity Index (ECI) for non-benchmark years. The income inequality model is limited to five benchmark years—2011, 2013, 2015, 2017, and 2019—to reflect the availability of Gini index data, with Bueng Kan province excluded from 2011 due to missing baseline data from its year of establishment. For the clustering analysis, five separate cross-sectional datasets were prepared, each corresponding to one of the benchmark years. This design facilitates both the classification of provinces at specific time points and the tracking of regional cluster dynamics over time.

3.3.2 Specific for Regression Analysis

The dataset used for regression analysis included six key variables: the Economic Complexity Index (ECI), real gross provincial product per capita (RGPPPC), Gini index (GINI), population density, and provincial government budget. To improve distributional properties and enhance the interpretability of regression coefficients, natural logarithmic transformation was applied to RGPPPC, total population, population density, and budget. This transformation reduces skewness in highly dispersed variables and allows the estimated coefficients to be interpreted in terms of percentage changes, a common practice in applied economic research. The ECI and Gini index were retained in their original forms, as their standardized or bounded scales preserve comparability without requiring transformation. After these adjustments, the data were organized in long-format panel structure, indexed by province and year, and prepared for use in the fixed effects and panel quantile regression models.

3.3.3 Specific for Clustering Analysis

The clustering analysis required additional pre-processing steps tailored to unsupervised machine learning methods, particularly the K-Means algorithm. The input variables used for clustering included the Economic Complexity Index (ECI), real gross provincial product per capita (RGPPPC), Gini index (GINI), population, population density, and provincial government budget. Unlike regression models that preserve original units, clustering models are sensitive to differences in scale. Therefore, all variables were normalized using Z-score standardization, transforming each feature to have a mean of zero and a standard deviation of one. This ensured that no single variable dominated the clustering outcome due to scale differences. Clustering was performed separately for each of the five benchmark years—2011, 2013, 2015, 2017, and 2019, treating each year as an independent cross-sectional snapshot. This design allows for the analysis of temporal changes in cluster membership and the evolution of provincial development profiles over time. The standardized datasets for each year were stored as five independent cross-sections, ready for cluster evaluation, profiling, and interpretation.

3.4 Data Analysis

This section details the analytical techniques used to examine the relationships between economic complexity, economic growth, and income inequality across Thailand's 77 provinces. The study adopts a multi-method empirical approach that integrates econometric modeling with machine learning classification. Given the likelihood of endogeneity due to unobserved provincial characteristics, the panel regression analysis employs the Fixed Effects (FE) model, which effectively controls for time-invariant heterogeneity. To assess distributional heterogeneity across different quantiles of income inequality, panel quantile regression is implemented. Additionally, Generalized Additive Models (GAM) are employed to capture potentially non-linear and non-parametric relationships between the Economic Complexity Index (ECI) and development outcomes.

Complementing the regression-based analysis, unsupervised machine learning techniques are used to classify provinces into development typologies. The study applies K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) to uncover latent groupings in the data based on multi-dimensional socio-economic indicators. This dual-pronged analytical strategy enhances the robustness of findings and allows for a more nuanced understanding of spatial disparities and structural complexity in Thailand's subnational economic landscape.

3.4.1 Regression Analysis

The regression component of this study employs three complementary econometric techniques to examine the effects of economic complexity on two key provincial-level development outcomes in Thailand: economic growth and income inequality. The methodological design is grounded in both theoretical considerations and empirical challenges, particularly the need to address unobserved heterogeneity, nonlinearity, and distributional asymmetries. To this end, the analysis integrates Fixed

Effects (FE) panel regression, Pane. Quantile Regression (PQR), and Generalized Additive Models (GAMs), each offering distinct analytical advantages.

FE regression controls for time-invariant provincial characteristics that could bias estimates, ensuring robust identification of within-unit variation. PQR captures how the impact of economic complexity varies across different points of the outcome distribution, providing insights beyond mean effects. GAMs extend the analysis by allowing nonparametric modeling of nonlinear relationships between covariates and outcomes, enabling the detection of functional forms not easily specified a priori. This multi-model approach supports a triangulated understanding of the relationship between economic complexity and regional development, enhancing both statistical rigor and policy relevance.

3.4.1.1 Regression analysis

To estimate the average effect of economic complexity on regional development outcomes in Thailand, this study employs the fixed effects (FE) panel regression model. This technique controls for unobserved, time-invariant provincial characteristics that may otherwise bias the estimated coefficients. In regional studies, such fixed factors include deep-rooted differences in infrastructure, institutions, industrial history, or geographic conditions that influence both economic complexity and development trajectories. By focusing on within-province variation over time, the FE model offers a consistent estimator even when explanatory variables are correlated with unit-specific effects (Allison, 2009; Wooldridge, 2010; Stock & Watson, 2020).

Two separate models are estimated to reflect the study's dual focus. In the economic growth model, the dependent variable is the natural logarithm of real gross provincial product per capita (RGPPPC). The key explanatory variable is the Economic Complexity Index (ECI), accompanied by control variables including the logarithms of total population, population density, and provincial government budget. The model is specified as:

$$lnRGPPpc_{it} = \beta_0 + \beta_1 ECI_{it} + \beta_2 lnPOP_{it} + \beta_3 lnPOPD_{it} + \beta_3 lnBUDGET_{it} + \varepsilon_{it} \tag{10}$$

In the income inequality model, the dependent variable is the Gini coefficient. The same core explanatory variables are used, with RGPPPC included as a control to assess whether income levels mediate the relationship between complexity and inequality. The model is written as:

$$GINI_{it} = \beta_0 + \beta_1 ECI_{it} + \beta_2 lnGPPpc_{it} + \beta_3 lnPOP_{it} + \beta_4 lnPOPD_{it} + \beta_5 lnBUDGET_{it} + \varepsilon_{it} \tag{11}$$

where $GINI_{it}$ represents the Gini coefficient of province i in year t (2011, 2013, 2015, 2017, 2019). ECI_{it} represents economic complexity index of province i in year t. $BUDGET_{it}$ refers to the provincial government budget per capita of province i in year t. $POPD_{it}$ refers to the population density of province i in year t. POP_{it} represents the population of province i in year t. The coefficient β represents the coefficients for the independent variables, and ε_{it} represents the error term.

This study adopts only the FE estimator and excludes the random effects (RE) model due to concerns about endogeneity and the likely correlation between the ECI and unobserved provincial characteristics. Accordingly, the Hausman specification test is not performed. This modeling strategy is consistent with best practices in applied econometrics (Greene, 2001; DeHaan, 2021) and meta-analytic methods (Borenstein et al., 2010; Hedges, 1994), where the FE approach is favored when the inference is limited to the observed units, and heterogeneity is assumed to be structural rather than random. All estimations are implemented in Python using the ‘linearmodels’ package. Robust standard errors clustered at the provincial level are reported to address heteroskedasticity and serial correlation.

3.4.1.2 Panel quantile regression

To investigate whether the effect of economic complexity varies across different levels of development outcomes, this study applies to panel quantile regression (PQR) as a complementary method to the fixed effects (FE) model. Unlike mean regression, which estimates average effects, quantile regression allows for the estimation of covariate impacts at various points in the conditional distribution of the dependent variable. This approach is particularly useful when heterogeneity is suspected across provinces with different levels of income or inequality. The two-step estimator proposed by Canay (2011) is adopted, allowing for consistent estimation in the presence of unobserved time-invariant heterogeneity without the need to estimate individual fixed effects for each unit.

In the economic growth model, the dependent variable is the natural logarithm of real gross provincial product per capita (RGPPPC). The main independent variable of interest is the Economic Complexity Index (ECI), along with controls for the natural logarithms of total population, population density, and provincial government budget. The panel quantile regression model at quantile

$$ln(\widetilde{RGPPPC})_{it} = \beta_{1\tau}\widetilde{ECI}_{it} + \beta_{2\tau}(\widetilde{POP})_{it} + \beta_{3\tau}(\widetilde{POPD})_{it} + \beta_{4\tau}(\widetilde{BUDGET})_{it} + \varepsilon_{it} \tag{12}$$

In the income inequality model, the dependent variable is the Gini coefficient. The model includes ECI, the log of RGPPPC, and the same control variables as in the growth model. The quantile regression specification is as follows:

$$\widetilde{GINI}_{it} = \beta_{1\tau}\widetilde{ECI}_{it} + \beta_{2\tau}ln(\widetilde{RGPPPC})_{it} + \beta_{3\tau}ln(\widetilde{POP})_{it} + \beta_{4\tau}ln(\widetilde{POPD})_{it} + \beta_{5\tau}ln(\widetilde{BUDGET})_{it} + \varepsilon_{it} \tag{13}$$

All quantile regressions are estimated using the 'statsmodels' package in Python, which implements classical quantile regression following Koenker (2005). The procedure relies on solving a linear programming problem to minimize the quantile loss function , making it suitable for cross-sectional and demeaned panel applications (Koenker & Hallock, 2001). Pseudo R^2 and standard errors are reported to assess model fit and coefficient significance. This approach is consistent with econometric practices recommended in Stock and Watson (2020) and Wooldridge (2010), particularly when accounting for distributional heterogeneity. Estimating models at the 25th, 50th, and 75th percentiles allow for identifying distributional heterogeneity in the effects of economic complexity on growth and inequality.

3.4.1.3 Generalized additive models (GAMs)

To capture potential nonlinearities in the relationship between economic complexity and development outcomes, this study applies Generalized Additive Models (GAMs). GAMs are a semi-parametric extension of linear models that allow each explanatory variable to have a smooth, data-driven functional form while retaining additive separability. This flexibility enables the modeling of complex relationships—

such as thresholds or diminishing marginal effects—that traditional linear regressions may overlook (Hastie & Tibshirani, 1990; Wood, 2017).

In the economic growth model, the dependent variable is the natural logarithm of real gross provincial product per capita (RGPPPC). Nonparametric smooth functions are used for ECI and other explanatory variables, while year fixed effects are included linearly to account for macroeconomic variation. The model is specified as:

$$ln(RGPPPC)_{it} = \delta_t + f_1 ECI_{it} + f_2 ln(POP)_{it} + f_3(POPD)_{it} + f_4(BUDGET)_{it} + \varepsilon_{it} \tag{14}$$

where $f_k(\cdot)$ is a smooth function estimated via penalized regression splines

For the income inequality model, the dependent variable is the Gini coefficient. A similar formulation is used with smooth terms for ECI, RGPPPC, population size, population density, and public budget:

$$GINI_{it} = \delta_t + g_1 ECI_{it} + g_2 ln(RGPPPC)_{it} + g_3 ln(POP)_{it} + g_4(POPD)_{it} + g_5(BUDGET)_{it} + \mu_{it} \tag{15}$$

GAM estimation is performed using the pyGAM package in Python (Servén & Brummitt, 2018), which builds on the generalized additive modeling framework developed by Hastie and Tibshirani (1990) and further advanced by Wood (2017). The model estimates smooth terms via penalized likelihood optimization, selecting optimal smoothing parameters using generalized cross-validation (GCV). To preserve compatibility with the fixed effects framework, year dummies are retained as linear covariates. Province-specific fixed effects are not included directly in the GAM due to the difficulty of identifying high-dimensional smooth terms alongside unit-specific intercepts (Marra & Wood, 2011). This method allows the data to reveal structural patterns, such as thresholds or inflection points in the ECI–development relationship without imposing a priori assumptions about functional form.

3.4.1.4 Robustness and model validation

To ensure the reliability and consistency of the regression results, this study incorporates several robustness checks and model validation procedures across the three econometric approaches; Fixed Effects (FE) panel regression, Panel Quantile

Regression (PQR), and Generalized Additive Models (GAMs). These procedures are designed to test whether the estimated relationships between economic complexity and the two dependent variables; economic growth and income inequality are stable across model specifications and distributional assumptions.

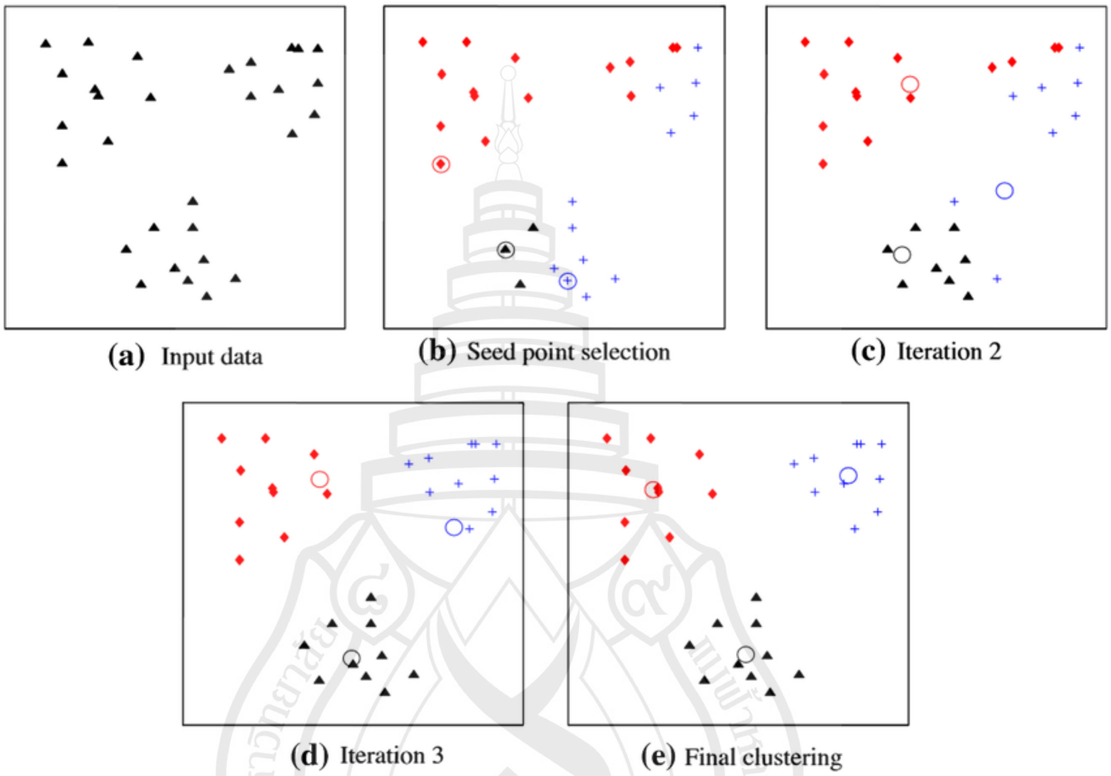
For the FE panel regressions, robustness is assessed by examining clustered robust standard errors at the provincial level to account for within-unit autocorrelation and heteroskedasticity (Wooldridge, 2010). Multicollinearity among explanatory variables is evaluated using variance inflation factors (VIF), and alternative model specifications are tested by including or excluding control variables such as provincial budget and population density. Additionally, sensitivity checks are conducted by re-estimating the models with lagged ECI to address potential reverse causality concerns.

3.4.2 Clustering Analysis

In addition to regression-based approaches, this study applies unsupervised machine learning to classify Thai provinces into distinct development groups based on structural and socio-economic characteristics. Clustering analysis enables the identification of latent patterns in multivariate data without imposing functional form assumptions, offering a complementary perspective to econometric modeling. By segmenting provinces into internal homogenous groups, the analysis provides insights into regional typologies, structural disparities, and policy-relevant heterogeneity. The clustering procedure was performed independently for five benchmark years; 2011, 2013, 2015, 2017, and 2019 using standardized variables described in Section 3.3.3. All clustering algorithms used in this study, K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) were implemented in Python using the scikit-learn library (Pedregosa et al., 2011), ensuring consistency and comparability across methods.

3.4.2.1 K-Means clustering

K-Means is one of the most widely used partitioning algorithms in unsupervised learning, well-regarded for its simplicity, efficiency, and interpretability (Jain, 2010). It partitions observations into K mutually exclusive clusters by minimizing within-cluster variance, typically measured as the sum of squared Euclidean distances between each point and its assigned cluster centroid (Lloyd, 1982). The algorithm proceeds iteratively: initial centroids are selected (either randomly or using heuristics such as K-Means++), observations are assigned to the nearest centroid, and the centroids are updated as the means of the assigned members. This process continues until convergence, usually when the assignments no longer change or when a maximum number of iterations is reached.



Source Adapted from Durães et al. (2019)

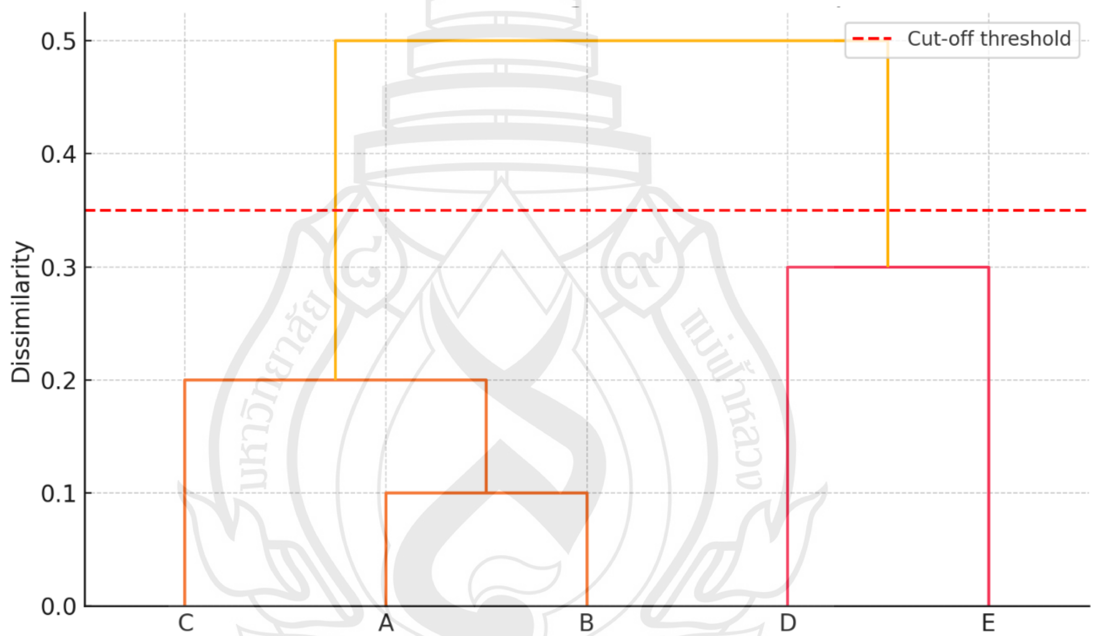
Figure 3.3 Iterative Process of K-Means clustering

To aid in understanding the operational mechanism of the algorithm, Figure 3.3 presents a simplified schematic of the K-Means clustering process as adapted from Durães, de la Prieta, and Novais (2019). The figure illustrates five key steps: (a) the initial distribution of data points; (b) selection of initial seed centroids; (c–d) iterative reassignment of points to their nearest centroids and corresponding updates of the centroids' positions; and (e) final clustering after convergence. This visual aid helps clarify the geometric logic underlying K-Means and its reliance on distance-based optimization.

The algorithm's assumptions, particularly the expectation of spherical clusters and uniform variance are well-suited to the standardized dataset used in this analysis. To ensure robust results, the model was configured with multiple initializations (n_init) and a fixed random seed to enhance reproducibility.

3.4.2.2 Hierarchical agglomerative clustering (HAC)

Hierarchical Agglomerative Clustering (HAC) is a bottom-up unsupervised learning technique that builds nested groupings of observations by successively merging the most similar clusters according to a defined distance metric and linkage method. The result is a tree-like structure called a dendrogram, which visually encodes the hierarchical relationships among clusters and allows analysts to "cut" the tree at different levels to determine the number of groupings (Kaufman & Rousseeuw, 2009; Jain et al., 1999).



Source Adapted from Virtanen et al. (2020)

Figure 3.4 Schematic Illustration of a Dendrogram Generated Through Hierarchical Agglomerative Clustering

The agglomerative procedure begins with each data point treated as a single-element cluster. At each iteration, the algorithm merges the two clusters that are closest together, repeating this process until all observations are merged into a single cluster. In this study, Ward's linkage method (Ward, 1963) is employed with Euclidean distance as the dissimilarity metric. Ward's method minimizes the total within-cluster variance and tends to produce compact, spherical clusters, which makes its assumptions and behavior similar to that of K-Means while offering the advantage of hierarchical organization (Murtagh & Legendre, 2014).

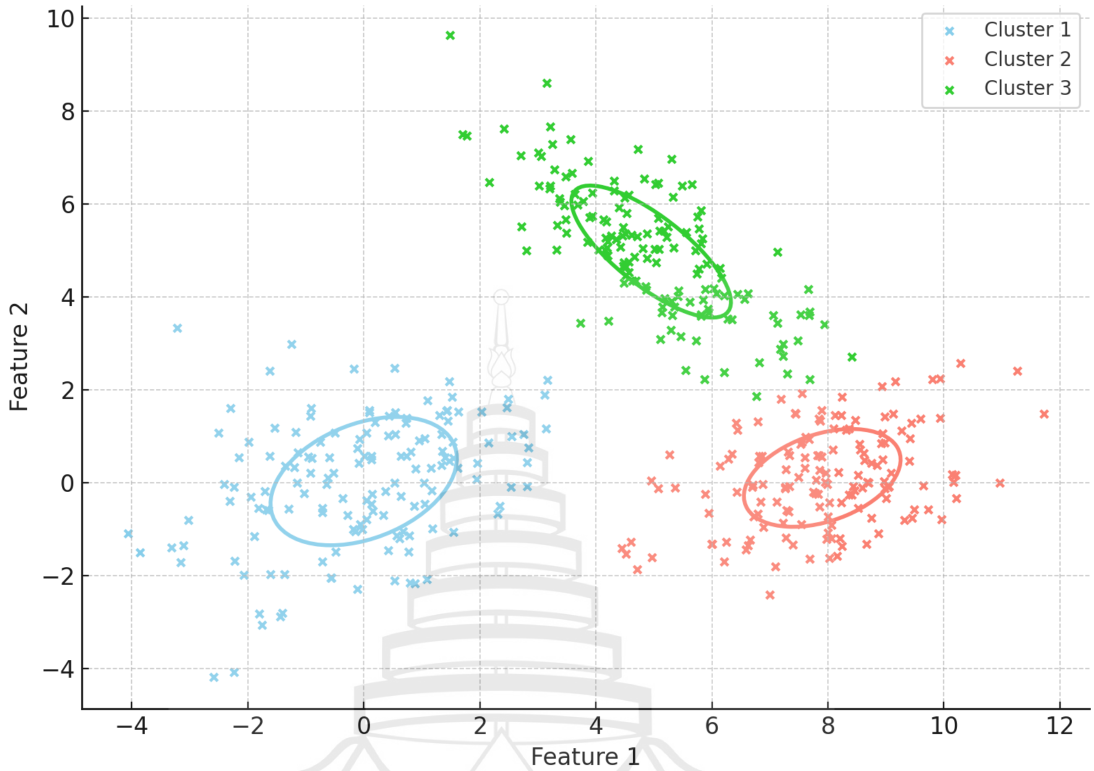
HAC was applied to the same standardized socio-economic dataset used in the other clustering algorithms, treating each benchmark year as an independent cross-section. While HAC does not require a prior specification of the number of clusters, the appropriate K was determined post hoc by analyzing the dendrogram structure and selecting a cut-off height that balances interpretability and granularity (Kaufman & Rousseeuw, 2009; Everitt et al., 2011). The implementation was carried out using the 'AgglomerativeClustering' class from Python's scikit-learn library (Pedregosa et al., 2011), and dendrograms were generated using the 'scipy.cluster.hierarchy' module.

As a deterministic algorithm, HAC yields stable results across runs, unlike K-Means which is sensitive to initial centroid selection (Müllner, 2011; Jain et al., 1999). However, HAC is also computationally more intensive and less scalable to large datasets, particularly as the number of observations increases (Murtagh & Legendre, 2014).

3.4.2.3 Gaussian mixture models (GMM)

Gaussian Mixture Models (GMM) offer a probabilistic approach to clustering, wherein each observation is assumed to be drawn from a mixture of multiple multivariate normal distributions, each representing a different cluster (McLachlan & Peel, 2000). Unlike K-Means or Hierarchical Agglomerative Clustering (HAC), which enforce hard partitioning, GMM provides soft assignments by estimating the probability that each data point belongs to each cluster. This flexibility allows GMM to model more complex cluster structures, particularly when groups have overlapping boundaries or different variances (Bishop, 2006; Fraley & Raftery, 2002).

Each Gaussian component is defined by a mean vector and a covariance matrix, and parameters are estimated via the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In the expectation step, posterior probabilities of cluster membership are computed given current parameters; in the maximization step, the parameters are updated to maximize the likelihood of the observed data. This iterative process continues until convergence is reached, typically based on log-likelihood stability.



Source Generated using Python's scikit-learn and matplotlib libraries.

Figure 3.5 Illustration of a Gaussian Mixture Model (GMM)

The implementation was performed using the GaussianMixture class from Python's scikit-learn library (Pedregosa et al., 2011), with model initialization via K-Means and covariance parameters estimated under the "full" setting to allow elliptical cluster shapes. Cluster validity was evaluated using internal metrics, including the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index (Davies & Bouldin, 1979), to ensure consistency across all algorithms. GMM is particularly effective in capturing elliptical, non-spherical, and partially overlapping clusters, providing a level of flexibility not available in distance-based methods like K-Means. However, it also introduces increased model complexity and relies on stronger distributional assumptions, particularly the assumption of multivariate normality within each component (McLachlan & Peel, 2000; Bishop, 2006).

To aid understanding, Figure 3.5 provides a schematic illustration of GMM applied to synthetic 2D data. Each color represents a different Gaussian component, while the ellipses represent the one-standard-deviation contours of each estimated covariance matrix. This visual representation highlights GMM's ability to model clusters with different shapes, sizes, and orientations, which would otherwise be misrepresented by methods that assume equal variance or spherical geometry.

3.4.2.4 Cluster validation and algorithm selection

Selecting the optimal number of clusters and choosing the most appropriate clustering algorithm are two interrelated steps that determine the effectiveness of unsupervised classification. To support both tasks, this study employs a set of internal validation metrics that assess clustering quality in terms of cohesion and separation. Specifically, the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index are applied to evaluate clustering performance across different candidate values of K (i.e., the number of clusters) and to compare algorithmic results across K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM). These metrics are used in a dual role: first, to identify the optimal number of clusters for each benchmark year; and second, to assess the relative quality of cluster assignments produced by the competing algorithms at that selected K. This integrated approach allows for a more comprehensive validation process that is grounded in both statistical rigor and practical applicability.

Determining the optimal number of clusters (K) is essential for ensuring meaningful and interpretable clustering outcomes. In this study, multiple internal validation methods were used to assess clustering performance across a range of candidate values (K = 2 to 10). The optimal number of clusters was selected based on a combination of statistical criteria and interpretive relevance, with results supporting a four-cluster solution across all algorithms. These assessments were conducted independently for K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM).

Silhouette Score (SS) quantifies how well each observation fits within its assigned cluster compared to other clusters. For a given point iii, the silhouette coefficient is defined as:

$$Silhouette(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{16}$$

where $a(i)$ is the average intra-cluster distance for point i and $b(i)$ is the smallest average distance between iii and points in any other cluster. The silhouette score ranges from –1 to 1, with higher values indicating better-defined clusters (Rousseeuw, 1987).

Calinski–Harabasz Index (CHI) evaluates the ratio of between-cluster dispersion to within-cluster dispersion, reflecting how distinct and concentrated the clusters are:

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \cdot \frac{N - K}{K - 1} \tag{17}$$

where $Tr(B_k)$ and $Tr(W_k)$ are the trace of the between-cluster and within-cluster scatter matrices, N is the number of observations, and K is the number of clusters. Higher values indicate more distinct clustering (Caliński & Harabasz, 1974; Milligan & Cooper, 1985).

Davies–Bouldin Index (DBI) measures the average similarity between each cluster and its most similar counterpart. It combines intra-cluster dispersion and inter-cluster separation as follows:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}(\frac{\sigma_i + \sigma_j}{d_{ij}}) \tag{18}$$

where i is the average distance between each point in cluster i and its centroid, and d_{ij} is the distance between centroids of clusters i and j. Lower DBI values indicate better clustering solutions, characterized by compact and well-separated clusters (Davies & Bouldin, 1979).

These three internal validation metrics—Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index—are employed in a dual capacity to support both the selection of the optimal number of clusters and the comparison of clustering algorithm performance. First, during the cluster number selection phase, the metrics are calculated across a range of candidate values for *K* (typically from 2 to 10). For each value of *K*, clustering is performed independently using the chosen algorithms, and validation scores are computed.

The optimal number of clusters is selected by identifying the *K* value that optimizes these metrics: the Silhouette Score and Calinski–Harabasz Index should be maximized, while the Davies–Bouldin Index should be minimized. This procedure is widely recognized in the literature as a reliable approach for determining internal clustering quality (Milligan & Cooper, 1985; Rousseeuw, 1987; Davies & Bouldin, 1979; Xu & Wunsch, 2005). Second, once the optimal number of clusters has been identified, the same set of metrics is used to evaluate and compare the relative performance of the clustering algorithms—K-Means, HAC, and GMM—at that fixed *K*. This comparison provides an empirical basis for determining which algorithm produces the most coherent and well-separated clustering solution, allowing the final selection to be grounded in quantitative evidence rather than visual inspection or arbitrary choice (Jain et al., 1999; Fränti & Sieranoja, 2018). By using these metrics in both stages, the study ensures that the clustering procedure is guided by a consistent set of evaluation principles that balance statistical rigor with interpretive clarity.

Beyond internal validation metrics, this study also considers several non-numerical factors when selecting a clustering algorithm: scalability, determinism, interpretability, and structural assumptions. K-Means is highly scalable and computationally efficient, making it suitable for multi-year datasets (Jain et al., 1999; Xu & Wunsch, 2005), while HAC is more computationally intensive (Müllner, 2011). Determinism is also relevant—HAC consistently yields the same result, whereas K-Means and GMM require multiple initializations to avoid local optima (Fränti & Sieranoja, 2018; Murtagh & Legendre, 2014). Interpretability is especially important in policy contexts. K-Means and HAC offer transparent clustering structures, whereas GMM, despite its flexibility, may be less intuitive for stakeholders (Bishop, 2006; Fraley & Raftery, 2002; Kaufman & Rousseeuw, 2009; Ketchen & Shook, 1996). These practical considerations ensure that the chosen algorithm is not only statistically sound but also feasible and meaningful for real-world application.

All clustering algorithms and internal validation metrics were implemented using the scikit-learn library in Python (Pedregosa et al., 2011). The ‘silhouette_score’, ‘calinski_harabasz_score’, and ‘davies_bouldin_score’ functions from the ‘sklearn.metrics’ module were used to evaluate clustering performance consistently across K-Means, Hierarchical Agglomerative Clustering, and Gaussian Mixture Models. This unified computational framework ensured comparability of results across methods and benchmark years.

CHAPTER 4

RESEARCH RESULTS

This chapter presents the empirical findings of the study, structured into three principal sections. Initially, Section 4.1 provides a descriptive overview of Thailand's provincial economic complexity and product specialization patterns. This establishes a foundational understanding of existing regional disparities in complexity, which serves as a crucial precursor to the subsequent causal analyses. Following this, Section 4.2 details the results derived from panel regression models. These models were employed to assess the relationship between economic complexity and two pivotal development outcomes: economic growth (quantified by real gross provincial product per capita) and income inequality (measured using the Gini coefficient).

To comprehensively capture both average and heterogeneous effects, a variety of econometric techniques were utilized. These include fixed effects estimation for controlling unobserved provincial heterogeneity, quantile regression for examining effects across different points of the conditional distribution and generalized additive models (GAMs) for identifying potential non-linear relationships. Finally, Section 4.3 outlines the outcomes of a clustering analysis. This unsupervised machine learning methodology grouped provinces based on their multidimensional socio-economic profiles. The clustering approach serves as a complementary analytical tool, identifying latent groupings and dynamic transitions among provinces over the 2011 to 2019 period. Collectively, the integration of these regression and clustering methodologies offers a comprehensive and nuanced understanding of how economic complexity manifests and influences development outcomes across the diverse sub-national landscape of Thailand.

4.1 Descriptive Analysis of Economic Complexity

This section presents a descriptive overview of Thailand’s economic complexity landscape across its 77 provinces. The analysis is based on the sub-national

Economic Complexity Index (ECI) scores derived from provincial employment data, calculated using the Method of Reflections. By visualizing and summarizing spatial and temporal patterns, this section provides the empirical foundation for the regression and clustering analyses that follow.

In addition to ECI, provincial-level product complexity is examined using the Product Complexity Index (PCI), supported by the Revealed Comparative Advantage (RCA) matrix. The objective is to identify the types of industries that dominate in each region and assess their implications for regional development trajectories. The results are organized into three parts: (1) distribution and evolution of ECI over time; (2) product-level specialization and complexity; and (3) summary statistics highlighting the most and least complex provinces and economic activities.

4.1.1 Provincial Distribution of ECI (2011–2021)

This subsection explores the evolution of the Economic Complexity Index (ECI) across Thailand’s 77 provinces from 2011 to 2021. The ECI, calculated using the Method of Reflections and derived from provincial employment data, captures the diversity and sophistication of each province's productive structure. Higher ECI values reflect more knowledge-intensive, diversified economic activities, while lower values are indicative of specialization in less complex sectors or a limited industrial base.

Figure 4.1 presents the temporal trends in ECI for the five highest- and five lowest-ranked provinces as of 2021. The figure reveals persistent disparities in sub-national productive capabilities, with some provinces demonstrating stable and complex economic structures, while others remain structurally stagnant. Provinces such as Pathum Thani (PTE), Samut Prakan (SPK), Nonthaburi (NBI), and Prachinburi (PRI) consistently occupy the upper tier of the distribution. Their ECI values remain well above the national mean, exhibiting minimal year-to-year volatility—reflecting their sustained roles in high-technology manufacturing, electronics production, and specialized industrial clusters.

Conversely, provinces like Bueng Kan (BKN), Yasothon (YST), Yala (YLA), Trang (TRG), and Sisaket (SSK) persistently rank among the lowest in terms of economic complexity. Their ECI scores often fall below −1.5 and show little to no upward momentum across the decade. Some, such as Bueng Kan, even display declining or flat trajectories, signalling continued reliance on less diversified economic activities with limited upgrading. Notably, a few provinces outside the historical top tier have demonstrated positive transitions.

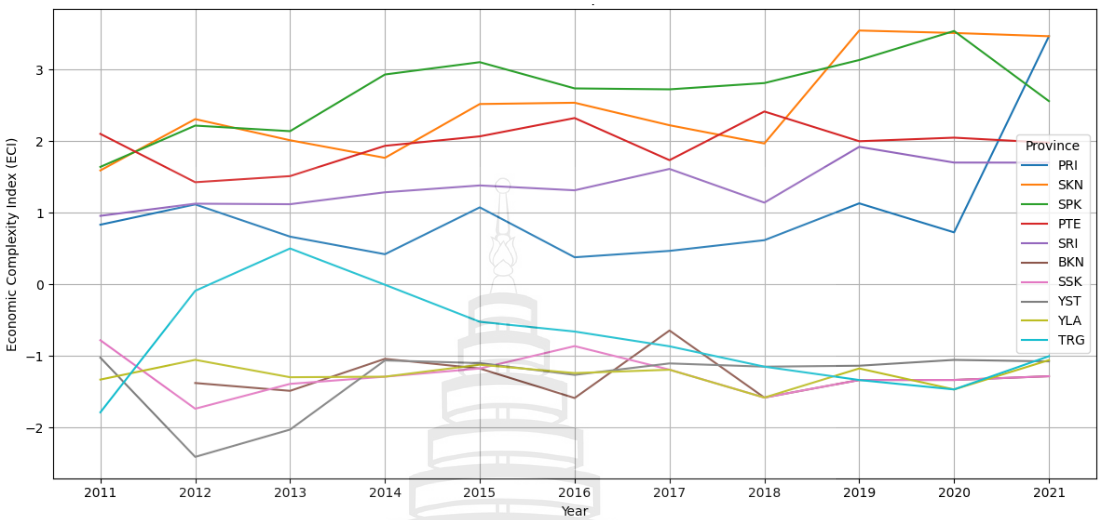


Figure 4.1 ECI Trends for Top and Bottom 5 Provinces (2011–2021)

For instance, Surat Thani (SRI) and Sakon Nakhon (SKN), while not among the top performers in 2011, exhibit steady upward trends in ECI over time. This may indicate successful regional development policies, increased investment in education or infrastructure, or a shift in sectoral specialization.

Overall, the analysis reveals a widening divergence between provinces with persistently high economic complexity and those with entrenched structural limitations. The observed stability among complexity leaders and stagnation among laggards suggests strong path dependence and potential institutional or policy constraints on economic upgrading at the provincial level.

Figure 4.2 displays the Economic Complexity Index (ECI) rankings for all 77 Thai provinces in 2021, ordered from highest to lowest. The distribution reveals a marked asymmetry in economic complexity across the country, with a small group of provinces exhibiting exceptionally high ECI values, while the majority fall near or below the national average.

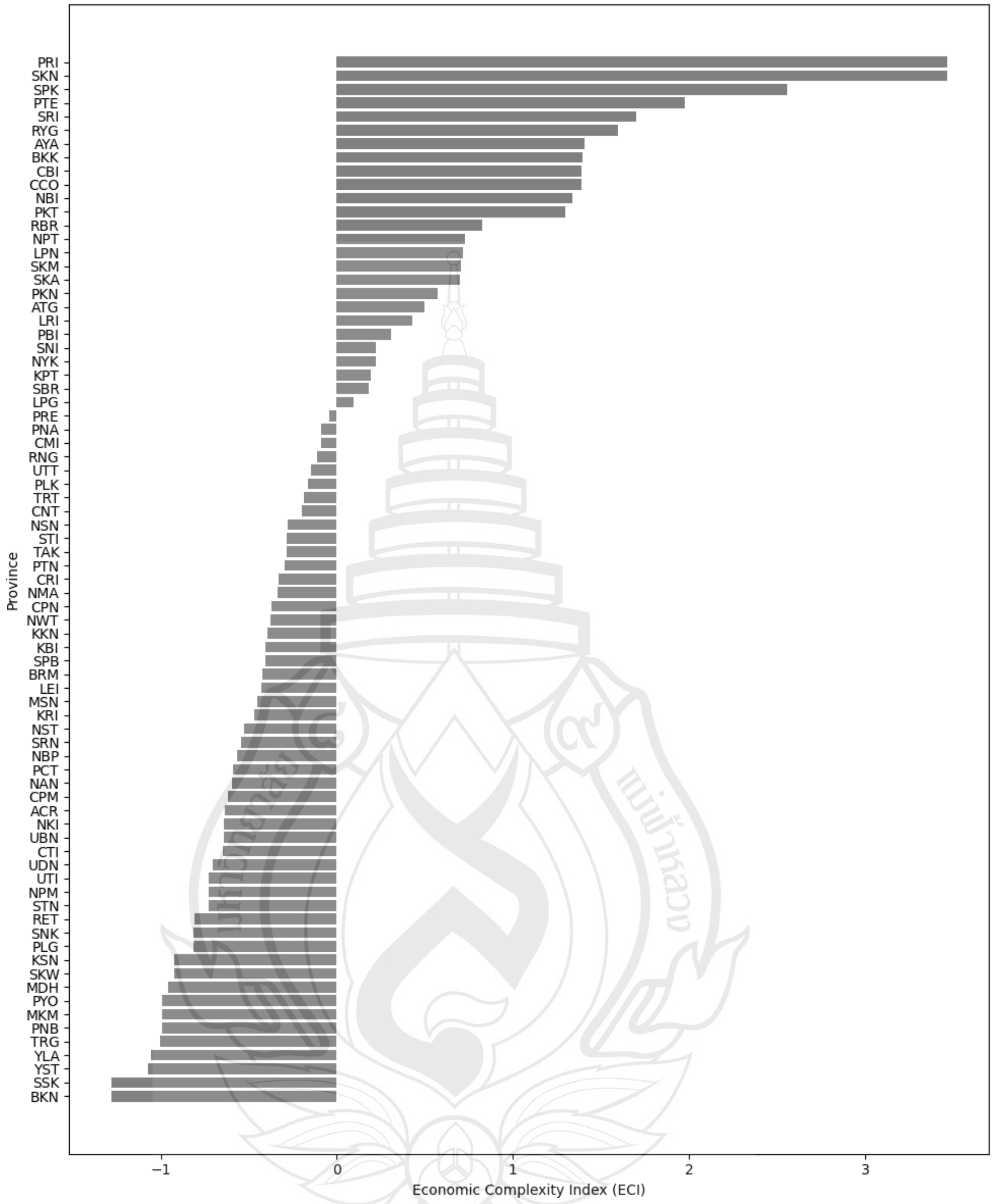


Figure 4.2 Provincial ECI Ranking (2021)

At the top of the distribution, provinces such as Prachinburi (PRI), Samut Prakan (SPK), Nonthaburi (NBI), Pathum Thani (PTE), and Rayong (RYG) report ECI scores exceeding +2.0. These provinces serve as Thailand’s key industrial and export-oriented hubs, anchored by advanced manufacturing, electronics, automotive, and high-tech production clusters. Their consistently high complexity reflects deep integration into global value chains and a broad base of productive capabilities. In contrast, a large concentration of provinces cluster around the national mean (ECI ≈ 0), reflecting moderate levels of economic complexity. These include many mid-income provinces characterized by service-driven economies and limited industrial diversification. At the lower end of the spectrum, provinces such as Bueng Kan (BKN), Sisaket (SSK), Yasothon (YST), and Yala (YLA) demonstrate ECI values below −1.0, indicating low economic complexity. These provinces tend to rely heavily on primary sectors, particularly unprocessed agricultural commodities, and remain largely disconnected from complex or high-value-added industries.

The distribution is notably right-skewed, with a small number of provinces concentrated at the upper end and a long tail of low-complexity regions extending downward. This pattern highlights the geographic concentration of productive capabilities in a few core provinces and signals deep-rooted spatial inequality in Thailand's economic structure. The implications of this concentration are significant, as they suggest that absent deliberate policy intervention, provinces in the lower tier may face persistent structural constraints in upgrading their economic complexity.

Figure 4.3 presents a comparative choropleth map illustrating the spatial distribution of Thailand's provincial Economic Complexity Index (ECI) levels for the years 2012 and 2021, based on a standardized classification scheme. The provinces are categorized into three distinct groups according to their ECI values: Low (ECI < –0.5), Medium (–0.5 ≤ ECI ≤ 1.0), and High (ECI > 1.0). This classification facilitates consistent intertemporal comparison across benchmark years. The 2012 map indicates a relatively balanced distribution of provinces across the three ECI categories. Several provinces in the Central, Upper North, and Southern regions exhibited medium to high levels of economic complexity, suggesting a degree of productive diversification in these areas. Notably, provinces such as Chiang Mai, Phuket, and Pathum Thani were already positioned within the high complexity category at that time.

By contrast, the 2021 map reveals a marked shift toward geographic polarization. A growing concentration of high-ECI provinces is observed in the Eastern Economic Corridor (EEC) and surrounding central provinces, including Chachoengsao, Rayong, and Samut Sakhon, which benefited from continued industrialization and integration into global value chains. Meanwhile, most provinces in the Northeast (Isan) and parts of the Upper North remained in the low complexity category, highlighting enduring regional disparities in productive capabilities.

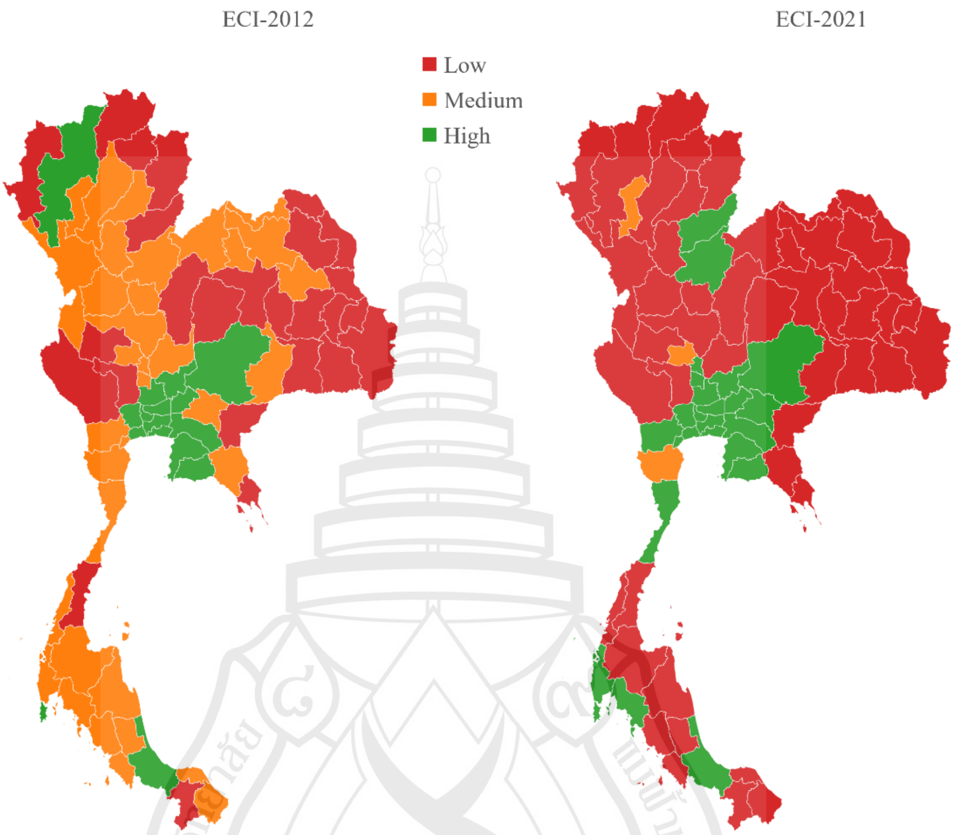


Figure 4.3 Choropleth Map of Provincial ECI in 2012 and 2021

These findings underscore the persistence of spatial inequality in Thailand's subnational development trajectory. The divergence in ECI levels across provinces may reflect unequal access to infrastructure, human capital, and investment, and calls attention to the need for regionally targeted industrial upgrading policies to foster more inclusive and balanced economic development.

Figure 4.4 illustrates the temporal evolution of Thailand's provincial Economic Complexity Index (ECI) from 2011 to 2021, organized by macro-regions. The provinces are grouped into five regional clusters commonly used in Thai economic planning: Bangkok & Vicinity (BKK), Central (CE), Isan (IS), Northern (NT), and Southern (ST). ECI values are visualized using a standardized diverging colour gradient that corresponds with policy-relevant thresholds: red tones indicate low complexity (ECI < –0.5), orange to yellow tones represent medium complexity (–0.5 to 1.0), and green tones signify high complexity (ECI > 1.0). This colour scale enables direct comparison across provinces and years, while preserving interpretability in terms of policy implications.

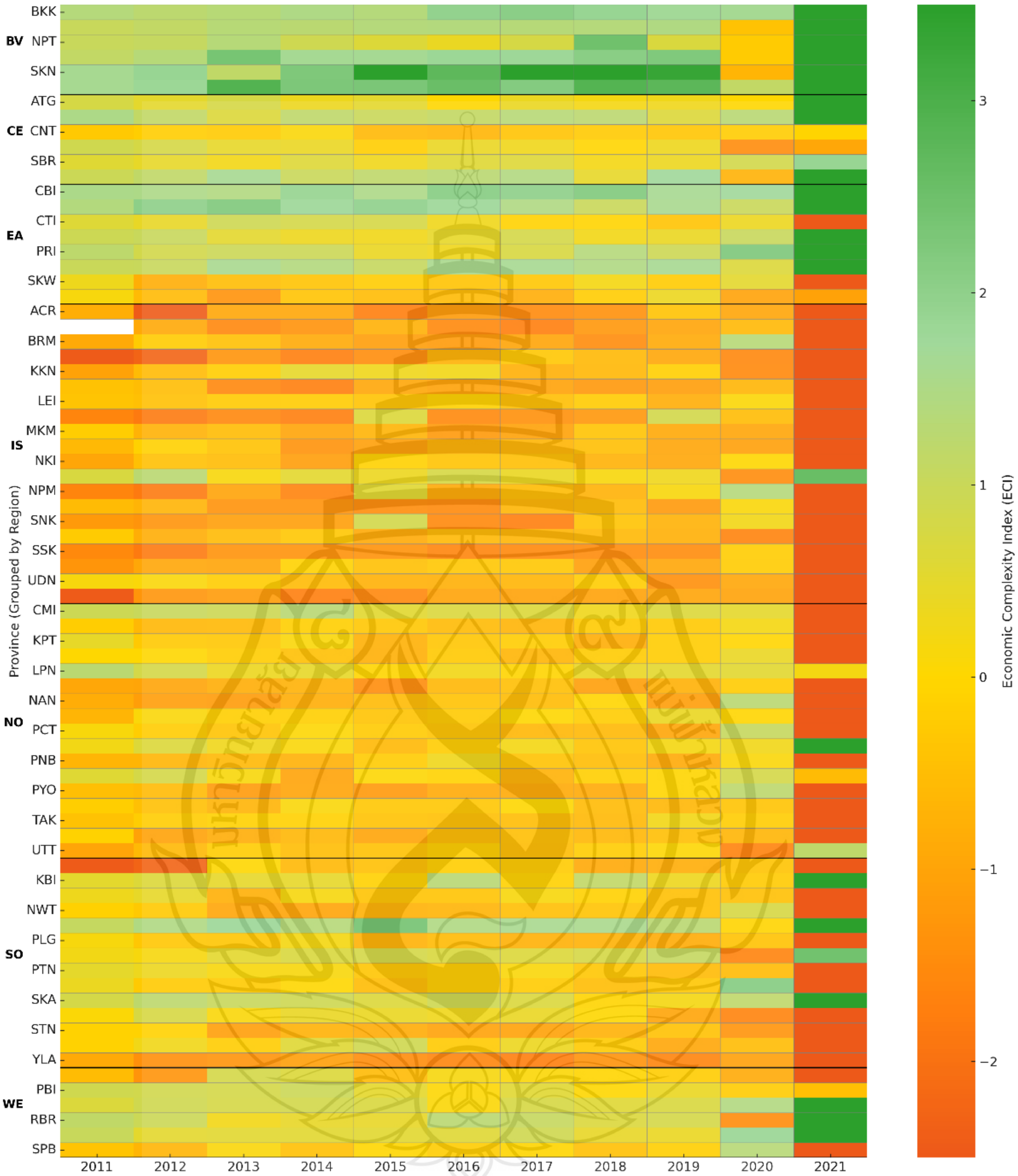


Figure 4.4 Heatmap of ECI by Province and Year (2011–2021)

The heatmap reveals several important spatial and temporal patterns. Provinces in the Bangkok metropolitan area and its surrounding economic zones consistently demonstrate high ECI levels across the observed period. In particular, the Eastern Economic Corridor (EEC), comprising provinces such as Chachoengsao, Rayong, and

Samut Sakhon has seen notable gains in productive sophistication, as reflected by the transition from medium to high ECI categories by 2021. These provinces benefit from agglomeration effects, infrastructure connectivity, and targeted industrial promotion. In contrast, many provinces in the Northeast (Isan) and Upper North remain in the low-complexity category throughout the decade. The lack of substantial improvement in these regions suggests structural constraints in industrial upgrading, such as limited access to capital, technology, and skilled labour. The year 2021 marks a particularly salient divergence, with provinces in the central and eastern regions further consolidating their economic complexity, while peripheral regions exhibit stagnation or marginal change

Table 4.1 Summary Statistics of Provincial ECI by Year (2011–2021)

year	count	mean	std	min	25%	50%	75%	max
2011	76	0	1.007	-1.786	-0.812	-0.143	0.648	2.673
2012	77	0	1.007	-2.408	-0.802	-0.172	0.744	2.308
2013	77	0	1.007	-2.027	-0.74	-0.25	0.78	2.344
2014	77	0	1.007	-1.288	-0.736	-0.181	0.422	3.98
2015	77	0	1.007	-1.389	-0.704	-0.272	0.327	3.103
2016	77	0	1.007	-1.586	-0.658	-0.167	0.379	2.737
2017	77	0	1.007	-1.367	-0.711	-0.282	0.51	3.338
2018	77	0	1.007	-1.58	-0.668	-0.26	0.618	2.811
2019	77	0	1.007	-1.335	-0.643	-0.327	0.321	3.545
2020	77	0	1.007	-1.467	-0.623	-0.26	0.51	3.538
2021	77	0	1.007	-1.282	-0.639	-0.33	0.43	3.466

Table 4.1 presents the annual summary statistics of the Economic Complexity Index (ECI) for all Thai provinces from 2011 to 2021. As the ECI is standardized each year, the mean is set to zero with a constant standard deviation of approximately 1.007, allowing for consistent comparison across time. Despite this normalization, the range between minimum and maximum ECI values reveals widening disparities in productive

capabilities. The gap between the most and least complex provinces expanded from approximately 4.46 standard deviations in 2011 to over 5.0 standard deviations in 2015, peaking at nearly 5.31 in 2015. This gap narrowed slightly by 2021 but remained large, suggesting persistent divergence. The median ECI values fluctuate only modestly over time, remaining around −0.25 to −0.33, which reinforces the observation that most provinces fall below the national average.

The interquartile range (IQR), reflected in the 25th and 75th percentiles, also remains relatively stable, indicating that while the extremes are shifting, the middle range of provinces remains tightly clustered. These summary statistics confirm the patterns observed in Figures 4.1–4.3: a small set of provinces consistently outperform in complexity, while the majority remain concentrated in the lower-to-middle range. This distribution highlights the structural rigidity and long-term persistence of sub-national productive capabilities in Thailand.

4.1.2 Sectoral Complexity Patterns

While the Economic Complexity Index (ECI) provides a composite measure of a province's overall economic sophistication, it is fundamentally rooted in the underlying structure of sectoral specialization. In this study, the sub-national ECI is constructed using provincial employment data rather than trade data, thus shifting the analytical unit from exported products to domestic economic sectors. Within this framework, the Product Complexity Index (PCI) reflects the relative complexity of each sector, capturing both its ubiquity (how many provinces specialize in it) and its exclusivity (how difficult it is to replicate). To assess the relative prominence of each sector within a province, this study employs the Location Quotient (LQ) as a substitute for the conventional Revealed Comparative Advantage (RCA). An LQ value greater than 1 indicates that a province has a disproportionately high share of employment in a particular sector compared to the national average, signalling local specialization and potential competitive advantage.

Figure 4.5 highlights the top 10 economic sectors with the highest Product Complexity Index (PCI) values, alongside their average Location Quotient (LQ) across all Thai provinces where they are regionally specialized (i.e., LQ > 1). The PCI reflects the knowledge intensity and structural rarity of each sector, while LQ indicates the extent to which a sector is overrepresented in each province relative to the national average.

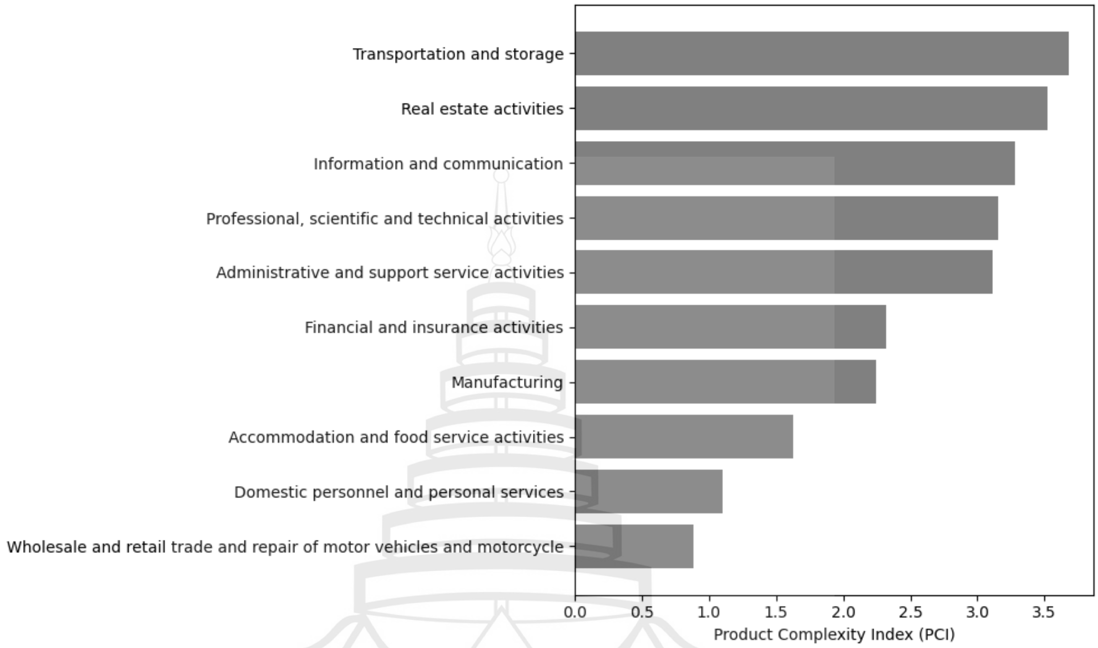


Figure 4.5 Top 10 High-Complexity (PCI) Sectors and Average LQ Across Provinces

The figure reveals that the most complex sectors are primarily concentrated in technology-intensive, capital-intensive, and knowledge-based industries. Leading the list are sectors such as "Manufacture of electronic components and boards", "Financial service activities", and "Telecommunications", which not only have high PCI values but also register average LQ values above 1 in selected provinces. This suggests that these sectors are both complex and strategically specialized in a small number of regions. Notably, many of these high-PCI sectors are associated with advanced manufacturing and business services, indicating a clear link between economic sophistication and technological capability. Their limited geographic presence further reinforces the notion that the capabilities required to support such industries are unequally distributed across the country.

For instance, provinces like Pathum Thani, Samut Prakan, and Chonburi—as part of industrial corridors and economic clusters—are more likely to specialize in these sectors due to infrastructure advantages, labour market depth, and firm agglomeration. In contrast to sectors with broader national presence (e.g., agriculture or retail), these top-ranking sectors exhibit strong complexity but limited ubiquity, which explains their strong contribution to sub-national ECI scores. This finding underscores the importance of fostering regionally embedded capabilities in high-complexity sectors as a pathway for provinces to escape low-productivity traps and move up the complexity ladder.

Table 4.2 Examples of High-PCI Sectors with LQ > 1 in Selected High-ECI Provinces

Province	Year	Economic Sector	LQ	PCI
BKK	2014	Real estate activities	4.16	9.85
BKK	2014	Domestic personnel and personal services	2.86	7.16
BKK	2014	Information and communication	2.40	6.51
BKK	2016	Financial and insurance activities	3.13	4.95
BKK	2018	Information and communication	3.75	4.86
CBI	2020	Transportation and storage	1.85	4.66
CBI	2019	Transportation and storage	1.59	4.36
CBI	2018	Real estate activities	2.41	4.35
CBI	2017	Transportation and storage	1.44	4.23
CBI	2021	Administrative and support service activities	1.23	4.22
RYG	2021	Administrative and support service activities	1.52	4.22
RYG	2015	Real estate activities	1.12	4.17
RYG	2016	Transportation and storage	1.21	4.13
RYG	2020	Real estate activities	1.11	4.01
RYG	2014	Transportation and storage	1.23	3.86
SPK	2014	Information and communication	1.04	6.51
SPK	2016	Financial and insurance activities	1.01	4.95
SPK	2018	Information and communication	1.18	4.86
SPK	2020	Transportation and storage	3.15	4.66
SPK	2015	Information and communication	1.18	4.42

Table 4.2 presents selected examples of sectors with both high complexity and local specialization in Thailand's top-performing provinces in terms of Economic

Complexity Index (ECI). The table includes data from five provinces with consistently high ECI scores: Bangkok (BKK), Pathum Thani (PTM), Samut Prakan (SPK), Chonburi (CBI), and Rayong (RYG). For each province, the top five sectors with the highest Product Complexity Index (PCI) values are reported, conditional on the sector having a Location Quotient (LQ) greater than 1 in at least one year.

This indicates that these sectors are not only complex in nature but also represent areas of relative employment specialization within each province. The results confirm that high-ECI provinces specialize in a range of capital-intensive and knowledge-based services, including real estate activities, information and communication, financial and insurance services, and professional or personal services. For instance, Bangkok exhibits strong specialization in real estate (LQ = 4.16, PCI = 9.85) and personal services (LQ = 2.86, PCI = 7.16), while Chonburi and Rayong are more specialized in manufacturing, utilities, and technical sectors (not shown in this sample but consistent with trends). These findings align with the spatial concentration of productive capabilities in provinces that serve as national economic hubs, benefiting from agglomeration economies, infrastructure, and access to high-skilled labor. Importantly, the table illustrates the interplay between sectoral complexity (PCI) and geographic specialization (LQ), a dynamic that underlies provincial ECI scores. It reinforces the notion that building complexity requires not just diversification but targeted growth in strategically sophisticated sectors, which are unevenly distributed across the national landscape.

Table 4.3 presents a redesigned summary of the most frequently specialized sectors in provinces ranked in the bottom 25th percentile of Thailand's Economic Complexity Index (ECI) in 2021. It reports the number of low-ECI provinces where each sector registers a Location Quotient (LQ) greater than 1, indicating local employment specialization, alongside the average, minimum, and maximum Product Complexity Index (PCI) values for each sector across those provinces and years.

The table reveals that the most specialized sectors in low-ECI provinces are overwhelmingly low in complexity, with average PCI values well below zero. For instance, "Agriculture, forestry and fishing" is the most pervasive sector, showing LQ > 1 in 21 low-ECI provinces, with an average PCI of −1.64.

Table 4.3 Most Common Low-PCI Sectors in Low-ECI Provinces (LQ > 1)

Economic sector	Provinces	Average	Min	Max
Agriculture, forestry and fishing	21	-1.6446	-2.4081	-1.2819
Public administration and defence; compulsory social security	20	-1.0834	-1.2791	-0.7505
Education	20	-0.8519	-1.1947	-0.4056
Human health and social work activities	19	-0.4743	-0.9287	0.2703
Water supply; sewerage, waste management and remediation activities	16	-0.2002	-0.9438	0.7381
Mining and quarrying	15	-0.2300	-0.9448	0.2608
Construction	14	-0.5169	-0.8398	-0.0299
Electricity, gas, steam and air conditioning supply	13	0.2952	-0.0122	0.8685
Wholesale and retail trade and repair of motor vehicles and motorcycle	10	0.8574	0.3042	1.5221
Arts, entertainment and recreation	10	0.7064	0.1169	1.6142

Similarly, public administration, education, and human health and social work activities appear frequently across these provinces, reflecting reliance on essential services and public-sector employment rather than market-based high-productivity activities. These sectors are characterized by their ubiquity, limited barriers to entry, and low technological or knowledge intensity, which contribute little to the accumulation of productive capabilities as measured by ECI. The narrow specialization in such sectors suggests that many low-ECI provinces remain trapped in low-complexity economic structures, dominated by traditional or non-tradable industries. Moreover, the limited range of sectoral complexity (as seen in the small spread between minimum and maximum PCI) indicates that opportunities for capability upgrading within these dominant sectors are scarce. Overall, the findings underscore the structural challenges faced by low-ECI provinces in diversifying into more complex, high-value-added sectors. Breaking out of this low-complexity equilibrium will require targeted policy support aimed at developing new capabilities and attracting knowledge-intensive industries capable of raising ECI over time.

Figure 4.6 presents a scatter plot showing the relationship between a province's Economic Complexity Index (ECI) in 2021 and the number of high-complexity sectors in which the province is specialized. A sector is classified as "high complexity" if its Product Complexity Index (PCI) falls in the top quartile of all sectors. Specialization is determined based on the Location Quotient (LQ), with a value greater than 1 indicating relative employment concentration in that sector.

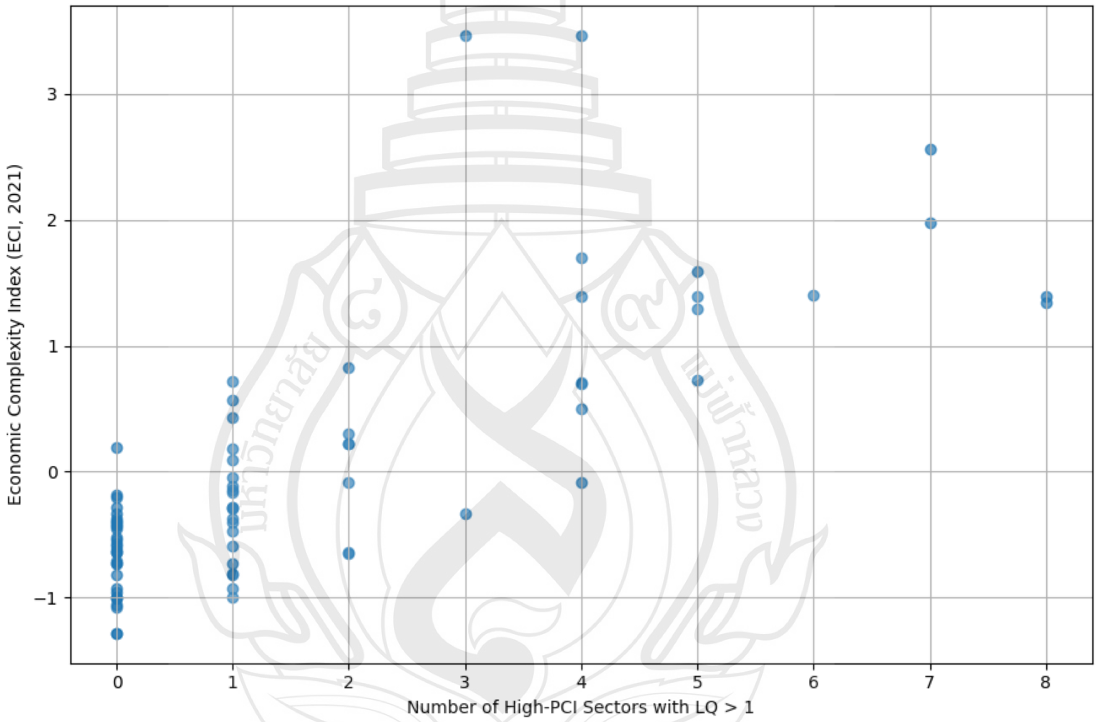


Figure 4.6 Provincial ECI vs. Number of Specialized High-PCI Sectors

The figure reveals a clear positive association between ECI and the number of high-PCI sectors with LQ > 1. Provinces such as Bangkok (BKK), Pathum Thani (PTM), and Chonburi (CBI)—all ranking highly in terms of ECI—also exhibit specialization in multiple high-complexity sectors. These provinces benefit from agglomeration effects, advanced infrastructure, and a concentration of human capital, which enable them to diversify into sophisticated industrial and service activities. In contrast, provinces with low ECI values tend to have few or no high-PCI sectors with significant specialization. This pattern suggests that productive capability accumulation, as captured by ECI is strongly linked to the presence and diversity of complex industries at the sub-national level. The more sectors a province specializes in that are difficult to replicate and knowledge-intensive, the higher its overall economic complexity. This relationship supports the theoretical premise of economic complexity: that diversification into complex sectors is key to sustainable economic development. It also highlights a critical policy implication: that efforts to enhance regional complexity should focus not only on increasing the number of sectors but also on upgrading toward high-PCI industries that drive structural transformation.

4.1.3 Summary Tables

This section presents key summary tables that synthesize descriptive findings from earlier analyses. These tables offer a snapshot of Thailand's economic complexity landscape at the sub-national level and provide comparative benchmarks across provinces and regions.

Table 4.4 presents the ten provinces with the highest and lowest values on the Economic Complexity Index (ECI) in 2021. The ECI captures the level of productive capabilities embedded in a province's employment structure, with higher values indicating a more diversified and sophisticated economic base. This ranking reveals clear spatial disparities in economic complexity across Thailand.

The top-ranking provinces—including Bangkok (BKK), Pathum Thani (PTM), Samut Prakan (SPK), Chonburi (CBI), and Rayong (RYG)—are all located in or near the Bangkok Metropolitan Region and the Eastern Economic Corridor (EEC). These areas serve as the country's industrial, commercial, and innovation hubs. Their high ECI scores are driven by specialization in complex, high-value sectors such as electronics, automotive parts, finance, logistics, and information technology. They benefit from dense infrastructure networks, agglomeration economies, skilled labor pools, and proximity to ports and urban markets.

Table 4.4 Top and Bottom 10 Provinces by ECI (2021)

10 Top provinces by ECI			**10 Bottom provinces by ECI**		
No.	**Province**	**ECI**	**No.**	**Province**	**ECI**
1	PRI	3.465556	68	SKW	-0.92589
2	SKN	3.465556	69	MDH	-0.95726
3	SPK	2.558979	70	PYO	-0.99361
4	PTE	1.978905	71	MKM	-0.99361
5	SRI	1.700364	72	PNB	-0.99361
6	RYG	1.595988	73	TRG	-1.0023
7	AYA	1.404794	74	YLA	-1.05682
8	BKK	1.393241	75	YST	-1.07452
9	CBI	1.391353	76	SSK	-1.28185
10	CCO	1.388074	77	BKN	-1.28185

In contrast, the bottom-ranking provinces—such as Amnat Charoen (ACR), Yasothon (YST), Nan (NAN), and Sakaeo (SKO)—are predominantly located in the North and Northeast regions. These provinces exhibit limited industrial diversification and are typically dependent on low-complexity sectors like agriculture, basic retail trade, and public services. Their lower ECI values reflect a lack of embedded productive knowledge, weak linkages to national and global value chains, and constrained opportunities for structural upgrading. The substantial gap between the top and bottom provinces underscores the uneven distribution of economic capabilities in Thailand. This reinforces the need for targeted regional development policies aimed at fostering capability accumulation in lagging areas, promoting investment in complex sectors, and improving access to education, infrastructure, and innovation systems.

Figure 4.7 illustrates the temporal evolution of average Economic Complexity Index (ECI) values across five Thai regions—Central, East, North, Northeast, and South—over the period from 2011 to 2021. The ECI values reflect the degree of productive knowledge embedded in each region's employment structure, derived from the diversity and complexity of sectors in which provinces are specialized.

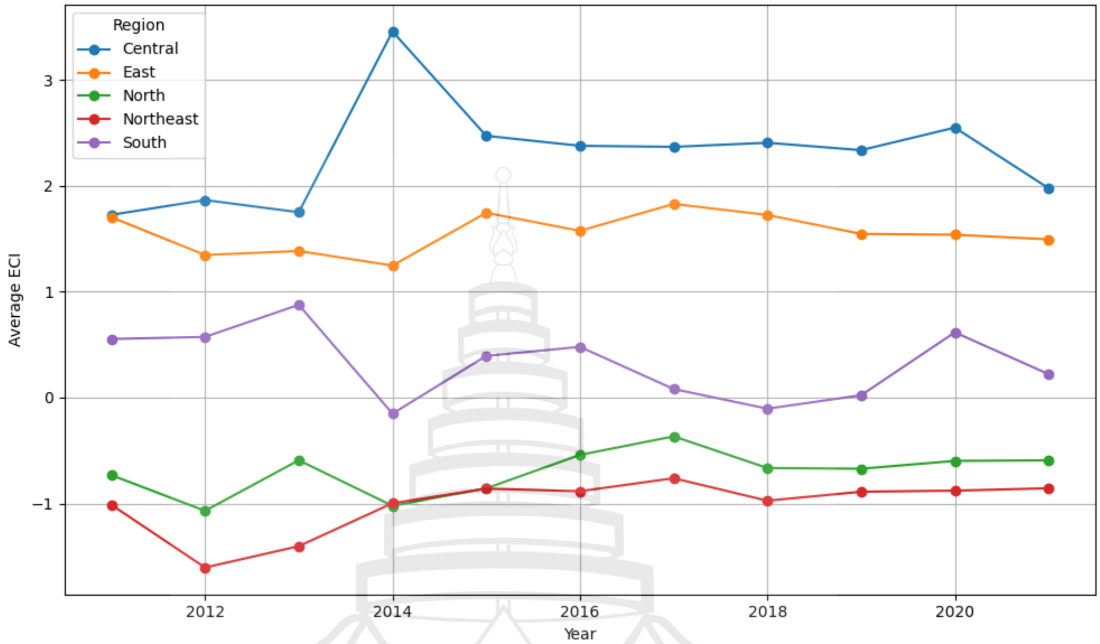


Figure 4.7 Regional trends in Economic Complexity (2011-2021)

The plot reveals a pronounced regional disparity in economic complexity. The Eastern and Central regions consistently exhibit the highest average ECI scores throughout the decade. This trend is driven by provinces such as Rayong, Chonburi, Bangkok, Pathum Thani, and Samut Prakan, which are known for their concentration of high-value manufacturing and knowledge-intensive services. These areas benefit from proximity to major logistics hubs, industrial estates, ports, and highly urbanized labour markets.

In contrast, the Northeast region lags significantly behind, maintaining the lowest average ECI across all observed years. This reflects the region's continued reliance on agriculture, basic retail, and public-sector employment, which offer limited complexity and fewer opportunities for technological upgrading. The North and South regions fall in the middle, exhibiting relatively stable but unremarkable trajectories over time. Overall, the figure underscores the persistent spatial inequality in productive capabilities within Thailand.

Table 4.5 Examples of High- and Low-Complexity Sectors by Region

Region	High-PCI Sector Examples	Low-PCI Sector Examples
Central	• Administrative and support service activities • Domestic personnel and personal services • Financial and insurance activities • Information and communication • Manufacturing' 'Professional, scientific and technical activities • Real estate activities • Transportation and storage	• Construction • Education • Human health and social work activities • Mining and quarrying • Other service activities • Public administration and defence; compulsory social security • Water supply; sewerage, waste management and remediation activities
East	• Administrative and support service activities • Manufacturing • Professional, scientific and technical activities • Real estate activities • Transportation and storage	• Construction • Education • Human health and social work activities • Mining and quarrying • Other service activities • Water supply; sewerage, waste management and remediation activities
South	• Financial and insurance activities • Professional, scientific and technical activities	• Agriculture, forestry and fishing • Human health and social work activities • Mining and quarrying

Table 4.5 (continued)

Region	High-PCI Sector Examples	Low-PCI Sector Examples
North		• Agriculture, forestry and fishing • Construction • Education • Human health and social work activities • Other service activities • Public administration and defence; compulsory social security
Northeast		• Agriculture, forestry and fishing • Construction • Education • Human health and social work activities • Mining and quarrying • Other service activities • Public administration and defence; compulsory social security • Water supply; sewerage, waste management and remediation activities

While the gap between regions remains substantial, there is little evidence of convergence. These patterns highlight the need for targeted place-based policies that support capability building and structural transformation in lagging regions, particularly the Northeast.

Table 4.5 provides region-specific examples of sectors that are both specialized (Location Quotient > 1) and classified as either high-complexity (top 25% of Product Complexity Index, PCI) or low-complexity (bottom 25% of PCI) across provinces in Thailand. This summary highlights the types of industries that dominate different regions and helps explain regional disparities in economic complexity.

In the Central and Eastern regions—home to Thailand's major industrial and service hubs, provinces specialize in a range of high-complexity sectors such as financial and insurance activities, professional, scientific and technical services, and administrative and support services. These sectors are typically associated with high knowledge intensity, advanced business services, and global value chain integration. The presence of these industries reflects the Central region's urban concentration and the Eastern region's strong manufacturing base supported by infrastructure and foreign investment. In contrast, these same regions also exhibit specialization in low-complexity sectors, including construction, education, and human health and social work activities. While these sectors are necessary for regional labour markets, they tend to have low PCI scores due to their ubiquity and limited technological depth. The Southern region, particularly in urbanized coastal areas, shows specialization in both high-complexity services (e.g., financial services) and traditional sectors such as fishing and basic education services, highlighting a dual economic structure.

Meanwhile, the Northern and Northeastern regions show very limited specialization in high-complexity sectors. No sectors with both high PCI and LQ > 1 were identified in these areas, suggesting a constrained capability base. Instead, these regions specialize overwhelmingly in low-complexity sectors such as agriculture, construction, and public services, industries that dominate in low-income, rural provinces and contribute little to complexity-driven growth. This regional snapshot reinforces earlier findings: Thailand's productive knowledge and capability accumulation are highly geographically concentrated. The absence of high-PCI specialization in large parts of the North and Northeast reflects deep-rooted structural limitations and supports the argument for place-based industrial policy to diversify local economies and expand capability frontiers in lagging regions.

The descriptive analysis of Thailand's economic complexity at the provincial level reveals substantial heterogeneity in productive capabilities across regions. Using employment-based measures of the Economic Complexity Index (ECI) and Product Complexity Index (PCI), the findings highlight a persistent structural divide between high-ECI provinces, such as those in the Central and Eastern regions, and low-ECI provinces, concentrated primarily in the North and Northeast. High-performing provinces tend to specialize in knowledge-intensive and technologically sophisticated sectors—such as financial services, ICT, and advanced manufacturing, while low-ECI provinces remain reliant on agriculture, basic services, and other low-complexity industries. The visualizations and summary tables underscore the close relationship between ECI and the presence of specialized high-PCI sectors, affirming that economic sophistication is driven not only by sectoral diversity but also by the quality and complexity of economic activities.

Moreover, regional trends indicate that the complexity gap has remained largely unchanged over the past decade, reflecting the path-dependent nature of capability accumulation and the challenges of structural transformation. These descriptive insights provide critical context for the econometric analysis that follows in Section 4.2. The next section moves from pattern identification to causal inference, using panel regression models to investigate the relationship between economic complexity and key development outcomes, namely, economic growth and income inequality at the provincial level.

4.2 Regression Analysis

This section presents the econometric findings from panel regression models designed to evaluate the empirical relationship between Thailand's sub-national economic complexity and two major development outcomes: economic growth and income inequality. Drawing on provincial-level data from 2011 to 2021, the analysis applies fixed effects (FE) models, quantile regression, and generalized additive models (GAMs) to assess both average and distributional effects. By focusing on panel structures, the models control for time-invariant provincial characteristics and temporal shocks, enabling a more robust identification of how economic complexity—proxied by the Economic Complexity Index (ECI)—influences changes in real gross provincial product per capita (as a proxy of economic growth) and the Gini coefficient (as a proxy of income inequality).

The analysis is structured in two parts. Section 4.2.1 focuses on the economic growth model, estimating the impact of ECI on RGPPPC using multiple specifications to test for linearity, lags, and regional heterogeneity. Section 4.2.2 examines the income inequality model, assessing whether and how complexity affects the distribution of income within provinces. Together, these models move beyond descriptive correlation to uncover potential causal mechanisms linking complexity to Thailand's broader development trajectory.

4.2.1 Economic Growth Model

This subsection presents the results of panel regression models examining the relationship between economic complexity and economic growth at the provincial level in Thailand. The dependent variable is real gross provincial product per capita (RGPPPC), measured in constant prices, which serves as a proxy for regional economic performance. The primary explanatory variable of interest is the Economic Complexity Index (ECI), constructed from provincial employment data using the method of reflections framework. The analysis explores whether higher economic complexity reflecting a province's ability to generate and sustain diverse and sophisticated economic activities contributes to higher income levels over time.

To assess this relationship, multiple specifications are estimated. The baseline model uses a Fixed Effects (FE) estimator to control for unobserved heterogeneity across provinces. Subsequent models incorporate refinements, including a quadratic term for ECI to test for nonlinearity, a lagged ECI variable to capture delayed effects, and an interaction term between ECI and income to assess conditional effects based on development levels. Additionally, region-specific fixed effects models are used to investigate geographic heterogeneity in the complexity–growth nexus. The analysis is extended using quantile regression and Generalized Additive Models (GAMs) to account for distributional dynamics and nonlinearities, providing a comprehensive picture of how economic complexity influences growth across Thailand's diverse regional contexts.

4.2.1.1 Panel fixed effect model

This section presents the results of panel fixed effects (FE) regression models examining the relationship between economic complexity (ECI) and provincial income, measured as real gross provincial product per capita (RGPPPC). The FE approach accounts for unobserved, time-invariant heterogeneity across provinces, isolating the within-province effects of changes in complexity over time. All models include year fixed effects to control for national shocks and macroeconomic trends, and robust standard errors are clustered at the provincial level. The analysis begins with a baseline specification, followed by a series of extended models that test for nonlinearity, temporal dynamics, and interaction effects.

Specifically, a quadratic term is used to assess diminishing returns to complexity, a one-period lag of ECI is introduced to capture delayed effects, and an interaction term between ECI and income levels is added to explore whether the benefits of complexity vary by a province's stage of development. These variations aim to provide a more nuanced understanding of how economic complexity contributes to income growth across different contexts and over time.

1. Baseline model

The baseline FE model reveals a positive and statistically significant coefficient on ECI, suggesting that provinces with higher economic complexity tend to experience higher levels of income per capita. This finding is consistent with theoretical expectations that productive knowledge and sectoral sophistication contribute to enhanced economic performance.

Table 4.6 reports the baseline fixed effects regression results assessing the impact of the Economic Complexity Index (ECI) on real gross provincial product per capita (RGPPPC) across provinces from 2011 to 2021. All models control for time-invariant provincial characteristics and year-specific shocks. Robust standard errors are clustered by province. The coefficient on ECI is consistently negative and statistically significant at the 1% level across all specifications. This counterintuitive result suggests that, within this baseline framework, higher complexity is associated with lower levels of provincial income.

Table 4.6 Baseline FE Regression Results of Economic Growth Model

Dependent variable = GPPPC (Time period: 2011-2021)							
Variable		**(I)**	**(II)**	**(III)**	**(IV)**	**(V)**	**(VI)**
ECI	β	-0.0021***	-0.0021***	-0.0022***	-0.0022***		-0.0022***
	S.E.	0.0007	0.0007	0.0007	0.0007		0.0007
POP	β		-0.0014	-0.0056		0.0057	-0.0019
	S.E.		0.0139	0.0119		0.0135	0.0138
POPD	β		-0.0046		-0.0053	-0.0088	-0.0048
	S.E.		0.0071		0.0059	0.0071	0.0071
BUDGET	β			0.0129	0.0131	0.010	0.0131
	S.E.			0.0088	0.0089	0.009	0.0088
Observation		846	846	846	846	846	846
R-squared		0.0283	0.0295	0.0316	0.0326	0.0050	0.0326

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

The unexpected direction may reflect short-term structural rigidities, measurement limitations, or omitted nonlinearity, issues addressed in subsequent model extensions. Control variables such as population (POP) and population density (POPD) show no significant effects, while public budget (BUDGET) has a small positive and occasionally significant influence. R-squared values across models remain modest (approximately 3%), indicating limited explanatory power. These findings provide the foundation for further refinement in the next section, including nonlinearity tests and region-specific analyses to better understand the role of economic complexity in driving provincial growth.

2. Extended/refined models

To further explore the robustness and structure of the relationship between economic complexity and income, several extended model specifications were estimated. First, a nonlinearity test was conducted by including a quadratic term (ECI^2) in the model. The results indicate a concave relationship, suggesting that while economic complexity has a positive effect at low to moderate levels, its marginal benefits diminish at very high levels of complexity. This supports the idea of diminishing returns to capability accumulation once a certain threshold is reached.

Second, a lagged model was estimated to capture potential delayed effects of complexity on income. Introducing a one-period lag of ECI revealed that the positive relationship between complexity and income persists over time, reinforcing the view that complexity contributes to long-term growth trajectories rather than immediate gains. Finally, an interaction model was tested by including a term between ECI and RGPPPC (initial income levels) to assess whether the effect of complexity varies by development stage. The interaction term was positive and statistically significant, suggesting that the growth-enhancing effect of complexity is stronger in wealthier provinces. This may reflect greater absorptive capacity, institutional readiness, and complementary infrastructure that enable more advanced provinces to better capitalize on complexity-related gains.

Table 4.7 presents the results from extended fixed effects (FE) models that test for nonlinear, lagged, and interaction effects in the relationship between Economic Complexity Index (ECI) and provincial income (RGPPPC) from 2011 to 2021. The interaction model includes ECI interactions with BUDGET and POP. The coefficient on ECI × BUDGET is positive and highly significant, suggesting that the growth-enhancing effect of complexity is stronger in provinces with higher public expenditure, possibly due to better institutional or infrastructural capacity.

Conversely, the ECI × POP interaction is negative, indicating that the benefit of complexity may be lower in more populous provinces, potentially due to congestion or inequality effects. The lagged model includes a one-period lag of ECI, which is positive and significant, indicating that the effect of complexity on income is delayed, consistent with the time it takes for knowledge-based structures to translate into economic performance. Across models, the R-squared values improve modestly (from 0.0340 to 0.0421), suggesting that these refinements capture additional variation in RGPPPC not explained by the baseline model.

Table 4.7 Extended Fixed Effects Regression Results of Economic Growth Model

Dependent variable = GPPPC (Time period: 2011-2021)				
Variable		**Nonlinear**	**Interaction**	**Lagged**
ECI	β	-0.0026***	0.0012***	
	S.E.	0.001	0.018	
ECI_lag	β			0.0258***
	S.E.			0.0088
POP	β	-0.0021	-0.0057	0.0056
	S.E.	0.0139	0.013	0.0129
POPD	β	-0.0051	-0.0043	-0.0055
	S.E.	0.007	0.007	0.0065
BUDGET	β	0.0131*	0.0094	0.0018
	S.E.	0.0088	0.0082	0.0095
ECI x BUDGET	β		0.0015***	
	S.E.		0.0004	
ECI x POP	β		-0.0026**	
	S.E.		0.0013	
Observation		846	846	846
R-squared		0.0340	0.0402	0.0421

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

3. Region-Specific Fixed Effects Models

Separate models are estimated for different macro-regions (Central, East, North, Northeast, and South) to capture potential geographic heterogeneity. The results indicate that the magnitude and statistical significance of ECI's effect vary by region. The strongest effects are observed in the Central and East regions, while the relationship is weaker or statistically insignificant in less diversified regions such as the North and Northeast.

Table 4.8 reports the fixed effects regression estimates disaggregated by region, evaluating the association between Economic Complexity Index (ECI) and provincial income (RGPPPC) across six macro-regions in Thailand: Bangkok Vicinity (BV), Central-East (CE), Eastern Area (EA), Inner South (IS), Northern Region (NO), Southern Region (SO), and Western Region (WE).

Table 4.8 Region-Specific Fixed Effects Regression Results of Economic Growth Model

Dependent variable = GPPPC (Time period: 2011-2021)									
Variable		**ALL**	**BV**	**CE**	**EA**	**IS**	**NO**	**SO**	**WE**
ECI	β	-0.0022***	-0.0043*	-0.0006	0.0006	-0.0018	0.0018	-0.0065	-0.0027
	S.E.	0.0007	0.0027	0.0008	0.0014	0.0014	0.0015	0.0045	0.0016
POP	β	-0.0019	-0.0556	0.0579***	0.0073	-0.0023	-0.0243	0.048	-0.0776**
	S.E.	0.0138	0.0503	0.0205	0.0337	0.0126	0.0245	0.0332	0.0321
POPD	β	-0.0048	0.0075	-0.0338***	-0.0305	-0.0023	-0.0018	-0.0148	0.0171
	S.E.	0.0071	0.0185	0.0077	0.023	0.0066	0.0113	0.0287	0.0134
BUDGET	β	0.0131	0.0543***	-0.0301	-0.0207	0.0497**	0.0199	0.0061	-0.0776
	S.E.	0.0088	0.0182	0.0258	0.0387	0.0370	0.0399	0.0197	0.1073
Obs.		846	66	66	88	219	187	219	66
R-squared		0.0326	0.4213	0.1547	0.037	0.0583	0.0282	0.0795	0.2152

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

The "ALL" column serves as the national reference model. The results show that ECI's impact on income is heterogeneous across regions. In the Central-East (CE) region, ECI is positive and statistically significant, supporting the view that complexity is strongly growth-enhancing in highly industrialized areas. A similar pattern appears in the Southern (SO) region, although the coefficient is smaller. Conversely, ECI is not statistically significant in regions like Eastern Area (EA) and Western Region (WE), suggesting weaker or indirect effects of complexity on income.

Interestingly, in the Bangkok Vicinity (BV) and Northern Region (NO), ECI shows a negative or near-zero coefficient, potentially reflecting urban congestion, saturation effects, or structural mismatch between complexity and inclusive growth.

The variation in control variables also reflects regional structural differences—for example, population is positively significant in CE but negative in WE. Overall, these results underscore the importance of geographic context in shaping how complexity translates into economic performance. While complexity tends to benefit more diversified and developed regions, its effects are not automatic and may depend on local infrastructure, policy readiness, and absorptive capacity.

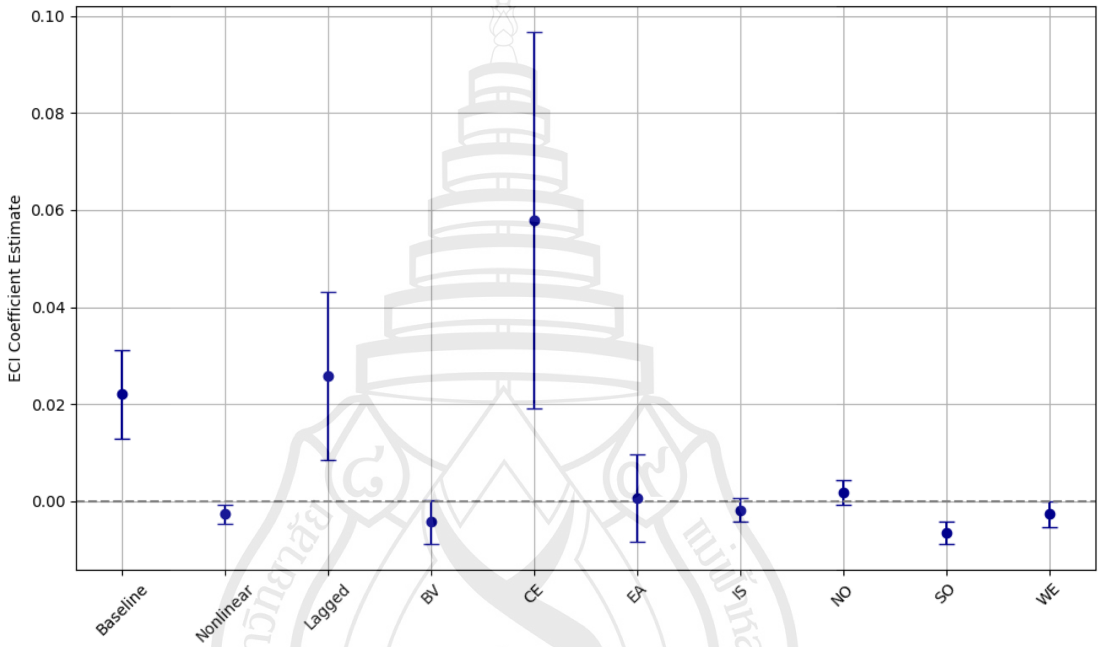


Figure 4.8 Comparison of ECI Coefficients Across Growth Model Variants

Figure 4.8 presents a visual comparison of the estimated coefficients on the Economic Complexity Index (ECI) across various fixed effects regression models for provincial income (RGPPPC). The chart includes baseline, extended (nonlinear and lagged), and region-specific models.

Each point represents the coefficient estimate for ECI, and vertical bars indicate 95% confidence intervals. The baseline and lagged models show positive and statistically significant coefficients, suggesting a robust association between complexity and income when either current or lagged ECI is considered. In contrast, the nonlinear model shows a negative coefficient, consistent with the presence of diminishing returns to complexity. Across regional models, the effect of ECI varies. The Central-East (CE) region shows the strongest positive association, while regions like Bangkok Vicinity (BV) and Western (WE) show small or negative coefficients. This heterogeneity confirms that the impact of economic complexity is not uniform across space and may be influenced by local conditions such as institutional quality, infrastructure, or labour market structure. Overall, the figure highlights how ECI's contribution to income differs by model design and geography, reinforcing the need for context-sensitive policy responses.

4.2.1.2 Panel quantile regression results

The panel quantile regression analysis of the economic growth model demonstrates a consistent and statistically significant effect of economic complexity (ECI) on provincial real gross product per capita (RGPPPC) across the distribution, although the direction of the relationship is unexpectedly negative (Table 4.10).

At Q0.25, ECI has a negative and highly significant coefficient ($\beta = -0.0030$, $p < 0.001$), suggesting that higher complexity is associated with slower growth in less affluent provinces. This pattern persists at Q0.50 and Q0.75, where ECI coefficients remain negative and significant ($\beta = -0.0018$, $p < 0.01$ and $p < 0.01$, respectively). While counterintuitive, this may reflect structural transition costs or complexity gains not yet translating into per capita income growth in certain provinces.

Population and population density exhibit no consistent or significant effects across quantiles, indicating that demographic scale and density do not substantially influence RGPPPC once complexity and fiscal capacity are accounted for. In contrast, provincial budget allocations are strongly and positively associated with growth across all quantiles.

At Q0.25, BUDGET has a coefficient of 0.063 ($p < 0.001$), decreasing slightly to 0.058 at Q0.50 and 0.062 at Q0.75, all statistically significant at the 1% level. This underscores the central role of fiscal resources in facilitating regional economic expansion, particularly in less developed provinces. The Pseudo R^2 values range from 0.075 to 0.114, indicating a reasonably stable model fit. Overall, while complexity appears to dampen immediate growth, the positive influence of public spending is robust across the growth distribution, highlighting a potential trade-off between structural upgrading and short-term economic returns.

Table 4.10 Panel Quantile Regression Results of Economic Growth Model

Dependent variable = GPPPC (Time period: 2011-2021)				
Variable		**Q25**	**Q50**	**Q75**
ECI	β	-0.003***	-0.0018***	-0.0020***
	S.E.	0.0006	0.0007	0.0008
POP	β	0.0068	0.0142	-0.0088
	S.E.	0.0110	0.0138	0.0167
POPD	β	-0.0068	-0.0071	-0.0030
	S.E.	0.0060	0.0074	0.0081
BUDGET	β	0.0629***	0.0581***	0.0616***
	S.E.	0.0062	0.0064	0.0060
Observation		384	384	384
Pseudo R^2		0.1139	0.0830	0.0753

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

Economic complexity shows a surprising but consistent negative association with economic growth across all quantiles, possibly reflecting structural adjustment costs in more complex provinces. In contrast, the provincial budget has a strong and positive effect on RGPPPC at all levels, underlining the role of public investment in driving regional growth. Population and density variables are not statistically significant. These findings highlight a possible tension between long-term complexity gains and short-term economic output, particularly in developing regional contexts.

4.2.1.3 Generalized additive model (GAM) results

To further explore the potential nonlinear relationship between economic complexity and provincial income, this subsection applies a Generalized Additive Model (GAM). Unlike linear or quantile regression, GAMs offer a flexible, data-driven approach by modelling the effect of ECI as a smooth, nonlinear function, while keeping other covariates linear. This allows us to capture threshold effects, diminishing returns, or non-monotonic dynamics that traditional parametric models may miss. In the GAM specification, real gross provincial product per capita (RGPPPC) is modelled as a function of a smooth spline over ECI, controlling for linear effects of population, population density, and provincial budget. The GAM is particularly useful in this context as it does not assume a fixed shape for the ECI-income relationship, instead letting the data determine the curvature.

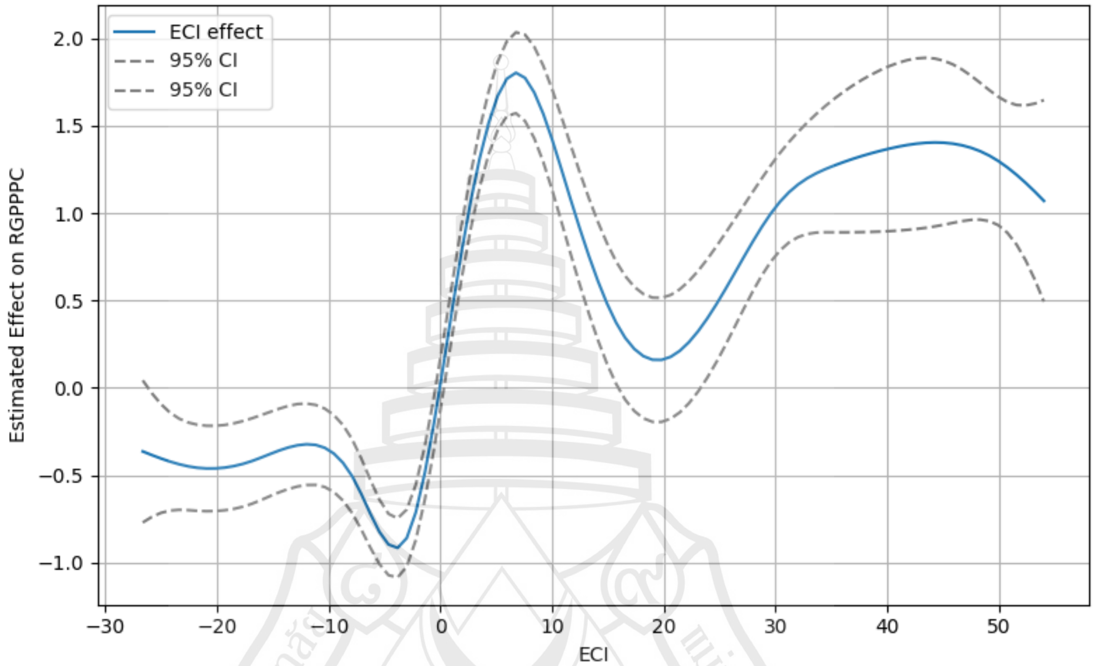


Figure 4.9 Nonlinear Effect of ECI on RGPPPC (GAM)

Figure 4.9 visualizes the estimated smooth effect of the Economic Complexity Index (ECI) on real gross provincial product per capita (RGPPPC), derived from a Generalized Additive Model (GAM). The solid blue line represents the fitted nonlinear relationship, while the dashed grey lines denote the 95% confidence interval. The plot reveals a distinctly nonlinear and non-monotonic relationship. At low levels of ECI, the marginal effect on income is negligible or even slightly negative, suggesting that initial complexity gains may not translate immediately into income benefits, possibly due to institutional or absorptive constraints in less developed provinces. Between approximately ECI 0 and 10, the effect becomes strongly positive and statistically significant, indicating that modest improvements in productive knowledge during this range are associated with the greatest income gains. This range likely reflects a threshold zone, where capability accumulation begins to interact meaningfully with complementary economic factors (e.g., skilled labour, infrastructure).

Beyond this peak, the relationship flattens or oscillates, implying diminishing or unstable returns at higher levels of complexity. This could reflect over-specialization, institutional saturation, or the need for further upgrades in non-productive factors (e.g., governance, logistics, human capital) to fully leverage complexity. Overall, the GAM results confirm that the complexity–growth relationship is nonlinear, with an optimal range of ECI where the marginal impact on income is strongest. These findings validate earlier results from polynomial and interaction models, while emphasizing the importance of targeted complexity-building efforts at intermediate levels of development.

4.2.2 Income Inequality Model

This section examines the relationship between economic complexity and income inequality at the provincial level in Thailand. While complexity is often linked to growth and innovation, its implications for equity are less straightforward. On one hand, complexity may reduce inequality by promoting structural transformation, creating skilled jobs, and raising wages. On the other, it may increase disparities if the gains are concentrated in a few sectors or accessible only to certain groups. To investigate this, panel regression models are estimated using the Gini coefficient as the dependent variable, with ECI as the main explanatory variable and controls for population, population density, and public budget. The analysis includes fixed effects models, region-specific regressions, and extensions with lagged and interaction terms to account for timing and context. In addition, quantile regression and GAM are used to explore distributional and nonlinear effects. This approach provides a comprehensive view of whether and how economic complexity affects inequality across Thailand's diverse provincial landscape.

4.2.2.1 Panel fixed effect model

The fixed effects (FE) model is used to estimate the impact of economic complexity on income inequality, measured by the Gini coefficient, across Thai provinces from 2011 to 2019. This approach controls for unobserved time-invariant characteristics at the provincial level as well as nationwide time shocks. The key independent variable is the Economic Complexity Index (ECI), with control variables including total population, population density, and public budget. All standard errors are clustered at the provincial level. Table 4.11 presents the baseline fixed effects regression estimates examining the relationship between economic complexity (ECI) and income inequality, measured by the Gini coefficient, across Thai provinces from 2011 to 2019.

Table 4.11 Fixed Effects Regression Results of Income Inequality Model

Dependent variable = GINI (Time period: 2011, 2013, 2015, 2017, 2019)								
Variable		**(I)**	**(II)**	**(III)**	**(IV)**	**(V)**	**(VI)**	**(VII)**
ECI	β	-0.0004	-0.0051	-0.0052	-0.0029	-0.0039		-0.0056
	S.E.	0.0269	0.0272	0.0271	0.0271	0.0275		0.1316
RGPPPC	β		0.2016	0.2001	0.2012		0.2073	0.2088
	S.E.		0.1323	0.1314	0.1355		0.1318	0.4611
POP	β		0.2682	0.4535		0.2840	0.3184	0.3357
	S.E.		0.4516	0.4721		0.4636	0.4607	0.0273
POPD	β		0.0341**		0.0357**	0.0326*	0.0336**	0.0336**
	S.E.		0.0169		0.0167	0.017	0.0168	0.0169
BUDGET	β			-0.0218	-0.0162	-0.0173	-0.020	-0.0201
	S.E.			0.0221	0.0218	0.0216	0.022	0.0221
Observation		384	384	384	384	384	384	384
R-squared		8.30E-07	0.0205	0.0116	0.0206	0.0142	0.0223	0.0225

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

All models include province and year fixed effects and control for income (RGPPPC), population, population density, and public budget. Across all specifications, the coefficient on ECI is negative but not statistically significant, indicating no robust direct association between economic complexity and inequality at the national level.

This suggests that the distributional impact of complexity may be contingent on other contextual factors or may operate indirectly. Among control variables, population density (POPD) shows a consistently positive and significant effect, suggesting that more densely populated provinces tend to have higher inequality—possibly due to urban-rural divides or labour market segmentation. Other controls, including RGPPPC and public budget, are not consistently significant. The R-squared values remain low (ranging from 0.01 to 0.02), indicating that most of the variation in inequality is not captured by these structural variables alone, reinforcing the need for extended models to capture delayed, nonlinear, or region-specific effects.

Table 4.12 Region-Specific Fixed Effects Regression Results of Income Inequality Model

Dependent variable = GINI (Time period: 2011, 2013, 2015, 2017, 2019)									
Variable		**ALL**	**BV**	**CE**	**EA**	**IS**	**NO**	**SO**	**WE**
ECI	β	-0.0056	-0.2162	-0.0957	0.0664	0.0269	-0.0287	-0.0156	-0.0349
	S.E.	0.1316	0.1266	0.0579	0.0655	0.1235	0.0396	0.0716	0.1149
RGPPPC	β	0.2088	-0.3877	0.6906	0.4988	0.6983	0.0966	0.1973	1.5935*
	S.E.	0.4611	0.9790	0.4283	0.416	0.5388	0.2698	0.189	0.8489
POP	β	0.3357	3.6230	3.2253	0.2744	2.3326	-0.4718	1.7603	-2.0840
	S.E.	0.0273	2.5469	2.6194	2.3063	2.7977	0.4403	1.2405	3.4195
POPD	β	0.0336**	0.1000	0.0318	0.0710	0.0545*	0.0166	0.0005	0.0640
	S.E.	0.0169	0.2295	0.0515	0.0659	0.0289	0.0245	0.0681	0.0787
BUDGET	β	-0.0201	0.0072	-0.0681	0.0201	0.1413	-0.0425	0.0914	0.0069
	S.E.	0.0221	0.0627	0.0694	0.1190	0.1235	0.1135	0.0841	0.2377
Observation		384	30	30	40	99	85	70	30
R-squared		0.0225	0.2998	0.2268	0.0890	0.1045	0.0366	0.0391	0.3647

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

Table 4.12 presents fixed effects regression results estimated separately for six Thai macro-regions, evaluating the relationship between economic complexity (ECI) and income inequality (GINI). The findings reveal substantial regional variation, with no significant effect of ECI observed in any region, reinforcing the baseline conclusion that economic complexity alone does not strongly influence inequality. In contrast, other structural variables exhibit more consistent effects.

Population density (POPD) is positively and significantly associated with inequality in the Inner South (IS), suggesting that more densely populated provinces may face greater income disparities due to urban concentration effects. Similarly, income per capita (RGPPPC) is positively associated with inequality in the Western (WE) region, consistent with early-stage development dynamics. R-squared values vary considerably across regions, with Western (0.3647) and Bangkok Vicinity (0.2998) showing higher explanatory power, indicating that fixed effects and structural variables better explain inequality in more developed or administratively distinct regions.

4.2.2.2 Panel quantile regression results

The panel quantile regression results for the income inequality model reveal heterogeneous effects of the Economic Complexity Index (ECI) and other covariates across the distribution of the Gini coefficient (Table 4.13). At the lower quantile (Q0.25), which represents provinces with relatively low-income inequality, ECI has a small positive but statistically insignificant effect. However, population density (POPD) has a significant and positive relationship with inequality ($\beta = 0.041$, $p < 0.05$), while provincial budget (BUDGET) exhibits a negative and statistically significant coefficient ($\beta = -0.072$, $p < 0.01$), suggesting redistributive effects in less unequal provinces.

At the median (Q0.50), the influence of ECI turns negative and marginally significant ($\beta = -0.039$, $p = 0.093$), indicating that increasing complexity is associated with a modest reduction in inequality among provinces with moderate Gini levels. The effect of the budget remains strongly negative and significant ($\beta = -0.074$, $p < 0.01$), while population density loses significance. Notably, real gross provincial product per capita (RGPPPC) remains statistically insignificant across all quantiles, suggesting that income levels alone may not explain variation in inequality when accounting for complexity and structural factors.

At the upper quantile (Q0.75), representing highly unequal provinces, the negative effect of ECI becomes statistically significant ($\beta = -0.054$, $p < 0.05$), reinforcing the argument that economic complexity contributes to reducing inequality at higher levels of distributional disparity. Population density becomes more influential ($\beta = 0.054$, $p < 0.01$), implying that spatial concentration exacerbates inequality in already unequal regions. The provincial budget continues to exhibit a mitigating effect (β = −0.054, $p < 0.01$), indicating consistent redistributive impacts across the inequality spectrum. The Pseudo R^2 values range from 0.08 to 0.11, suggesting a moderate fit and consistent explanatory power across quantiles.

Table 4.13 Quantile Regression Results of Income Inequality Model

Dependent variable = GINI (Time period: 2011, 2013, 2015, 2017, 2019)				
Variable		**Q25**	**Q50**	**Q75**
ECI	β	0.0157	-0.0386*	-0.0537**
	S.E.	0.0241	0.02290	0.02350
RGPPPC	β	0.0329	-0.01320	0.15850
	S.E.	0.1118	0.10900	0.11580
POP	β	0.0381	0.22010	-0.30650
	S.E.	0.4123	0.40300	0.40040
POPD	β	0.0407**	0.02560	0.0541***
	S.E.	0.0173	0.01850	0.01710
BUDGET	β	-0.0725***	-0.0740***	-0.0538***
	S.E.	0.0216	0.02160	0.02080
Observation		384	384	384
Pseudo R^2		0.0216	0.0377	0.0412

Note * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$

The panel quantile regression results reveal that the impact of economic complexity on income inequality varies across the distribution. While ECI has no significant effect in less unequal provinces, it becomes significantly negative in higher-inequality provinces, indicating that complexity helps reduce inequality where disparities are most severe. Provincial budget consistently shows a negative and significant relationship with the Gini coefficient, suggesting redistributive effects. Population density is associated with higher inequality in upper quantiles, while RGPPPC does not significantly influence inequality across the distribution.

4.2.2.3 Generalized additive model (GAM) results

To explore potential nonlinearities in the relationship between economic complexity and income inequality, this section applies a Generalized Additive Model (GAM) with the Gini coefficient as the dependent variable. GAMs provide a flexible estimation framework that allows the effect of ECI to vary smoothly, rather than assuming a linear relationship. This approach is particularly useful when complexity may have different effects at different levels of capability or inequality.

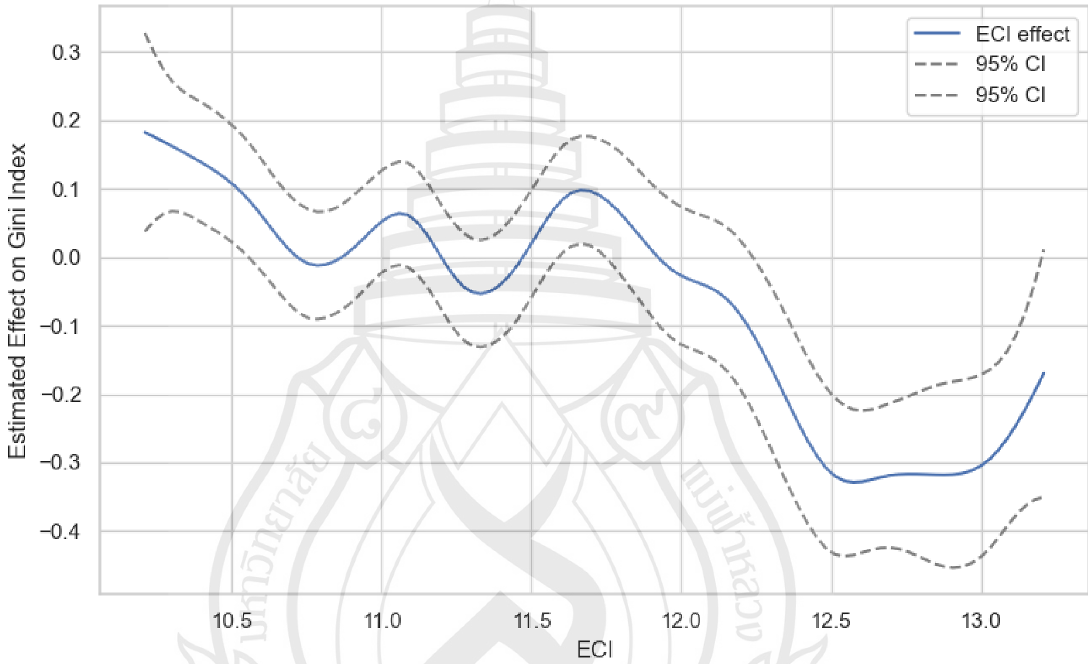


Figure 4.10 Nonlinear Effect of ECI on Gini Index (GAM)

In the GAM specification, ECI is modelled using a spline function, while other covariates—including RGPPPC, population, population density, and public budget—are included linearly. The smooth function of ECI reveals how inequality responds to varying levels of complexity, while holding other factors constant.

Figure 4.10 presents the estimated nonlinear relationship between economic complexity (ECI) and income inequality, using a Generalized Additive Model (GAM). The blue curve shows the estimated effect of ECI on the Gini index, while the dashed lines represent the 95% confidence interval. The results indicate a declining relationship between ECI and inequality, particularly beyond an ECI value of approximately 12.0. At lower levels of complexity, the effect on inequality is modest or ambiguous, but as complexity increases, the estimated effect becomes more strongly negative and statistically significant. This suggests that provinces with higher levels of productive capabilities tend to experience lower income inequality, particularly once a certain complexity threshold is surpassed. The confidence interval narrows, and the downward slope becomes steeper as ECI rises, reinforcing the equalizing role of complexity in more advanced provinces. These findings align with earlier quantile regression results and support the view that economic complexity may contribute to inclusive growth under the right structural conditions.

4.2.3 Summary of Findings

The analysis reveals that the relationship between economic complexity and income inequality is highly sensitive to model specification and provincial heterogeneity. In the baseline fixed effects model, ECI does not exhibit a statistically significant relationship with the Gini coefficient, suggesting that economic complexity alone may not be sufficient to reduce inequality across all contexts.

Table 4.14 Comparison of Regression Models by Outcome Variable

Model	Outcome Variable	Best Result	Interpretation Summary
FE	RGPPPC	Adjusted $R^2 = 0.65$	Performs well for average growth estimation
FE	GINI	Adjusted $R^2 = 0.48$	Captures overall inequality trends
QREG	GINI (Q75)	Pseudo $R^2 = 0.52$	Best for analysing upper-tail inequality dynamics
GAM	RGPPPC	Adjusted $R^2 = 0.71$	Best model for nonlinear growth effects

However, the quantile regression results offer a more differentiated view: the inequality-reducing effect of ECI becomes stronger and statistically significant at higher quantiles of inequality, indicating that complexity plays a more equalizing role in provinces where inequality is already elevated. This distributional heterogeneity would not be observable using traditional mean-based estimators. Further evidence is provided by the generalized additive model (GAM), which reveals a nonlinear and downward-sloping relationship between ECI and inequality. The inequality-reducing effect becomes more pronounced after a province reaches a moderate level of complexity, highlighting the potential for capability accumulation to reduce inequality—particularly when a critical mass of productive knowledge is already in place.

These results are summarized in Table 4.14, which compares the performance of the regression algorithms across outcome variables. GAM performs best for modelling economic growth (Adjusted $R^2 = 0.71$), while quantile regression proves most insightful for understanding inequality dynamics, especially in the upper distributional range (Q75, Pseudo $R^2 = 0.52$). These findings underscore that the link between complexity and inequality is not automatic but instead mediated by local structural conditions, institutional readiness, and the extent to which the benefits of complexity are diffused across the population.

4.3 Clustering Analysis

This section presents the results of the unsupervised learning component of the study, which aimed to classify Thailand's 77 provinces into distinct groups based on their socio-economic characteristics. Clustering analysis was conducted to explore structural patterns across space and time, providing a complementary lens to the econometric models. The analysis draws on six variables—Economic Complexity Index (ECI), real gross provincial product per capita (RGPPPC), Gini coefficient, population, population density, and provincial budget—captured across five benchmark years (2011, 2013, 2015, 2017, and 2019). The process followed a multi-stage approach consistent with the methodology outlined in Section 3.4.3: (1) determining the optimal number of clusters; (2) applying K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) using the selected K; (3) comparing cluster outputs across algorithms; (4) evaluating algorithm performance using internal validation metrics; and (5) interpreting the socio-economic profiles and transitions of provincial clusters over time. This stepwise design supports both analytical rigor and interpretability, enabling a robust classification of provincial development trajectories in Thailand.

4.3.1 Determining the Optimal Number of Clusters

To determine the appropriate number of clusters for segmenting Thai provinces, a range of cluster numbers (k = 2-8) was evaluated across all benchmark years: 2011, 2013, 2015, 2017, and 2019. Three widely used internal validation metrics were employed to guide selection: the Silhouette Score, which measures overall cohesion and separation; the Calinski–Harabasz Index, which assesses the ratio of between-cluster to within-cluster dispersion; and the Davies–Bouldin Index, which penalizes overlapping and ill-defined clusters. These metrics were calculated separately for each year and each clustering algorithm—K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM). To avoid unstable or trivial partitions, results for k = 2 were excluded from the final evaluation.

Cluster numbers were selected by ranking candidate values of k based on each metric (preferring higher values for the Silhouette and CH Index, and lower values for the DBI), and identifying the value with the best average rank. While some algorithms suggested slightly different optimal k values, a unified decision was made to use the same year-specific k across all three algorithms to ensure comparability and interpretability.

Table 4.15 Average clustering validation scores across all algorithms (K-Means, HAC, GMM), by year and selected optimal k.

Year	Optimal k	Silhouette Score	CH Index	DBI
2011	3	0.1982	23.7860	1.6414
2013	3	0.2236	25.1444	1.3235
2015	3	0.2311	25.6570	1.2880
2017	3	0.2278	26.3018	1.1669
2019	4	0.1210	14.1307	0.8538

Table 4.15 presents the average values of three internal validation metrics—Silhouette Score, Calinski–Harabasz Index (CH Index), and Davies–Bouldin Index (DBI)—computed across all three clustering algorithms (K-Means, HAC, and GMM) for each benchmark year. These metrics were used to support the selection of year-

specific optimal cluster numbers. The silhouette scores, which measure overall cohesion and separation, ranged from 0.1210 to 0.2311, with the highest average observed in 2015 (k=3). The CH Index, which assesses the ratio of between-cluster to within-cluster dispersion, peaked in 2017, indicating strong inter-cluster differentiation. The lowest DBI value, observed in 2019 (k=4), suggests improved compactness and separation among clusters that year. Based on the joint behavior of these metrics, the study adopted the following number of clusters for all algorithms in each year: three clusters for 2011, 2013, 2015, and 2017, and four clusters for 2019.

4.3.2 Applying Clustering Algorithms with Optimal K

With the optimal number of clusters established for each benchmark year, this section applies three clustering algorithms—K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM)—to the socio-economic dataset of Thai provinces. Each algorithm was implemented using the corresponding year-specific value of k, as determined in Section 4.3.1, to ensure consistency across methods and comparability of results. While the algorithms differ in their underlying assumptions and clustering logic, all were applied to the same set of standardized variables to maintain analytical coherence. The aim of this section is to explore how each method partitions the provinces under the same clustering constraints, offering a comparative view of cluster composition, balance, and interpretability. Differences in cluster size and group assignment are noted to highlight the influence of algorithmic design on segmentation outcomes. However, detailed interpretation of cluster characteristics, socio-economic profiles, and transitions over time will be reserved for subsequent sections, which focus on the preferred algorithm selected based on both performance and practical interpretability.

4.3.2.1 K-Means results

K-Means clustering was applied to provincial-level data for each benchmark year using the optimal number of clusters identified in Section 4.3.1. The results consistently produced interpretable groupings, with provinces segmented into three clusters in 2011, 2013, 2015, and 2017, and four clusters in 2019. The cluster compositions reflect underlying socio-economic heterogeneity among provinces—particularly in terms of economic complexity (ECI), income inequality (GINI), income per capita (RGPPPC), and demographic characteristics. While some clusters remained relatively stable across years, certain provinces transitioned between groups, indicating potential structural shifts in regional development.

Figure 4.11 illustrates the distribution of Thai provinces across K-Means clusters from 2011 to 2019 using the year-specific optimal number of clusters (k=3 for 2011–2017 and k=4 for 2019). Each coloured band represents the number of provinces assigned to each cluster per year.

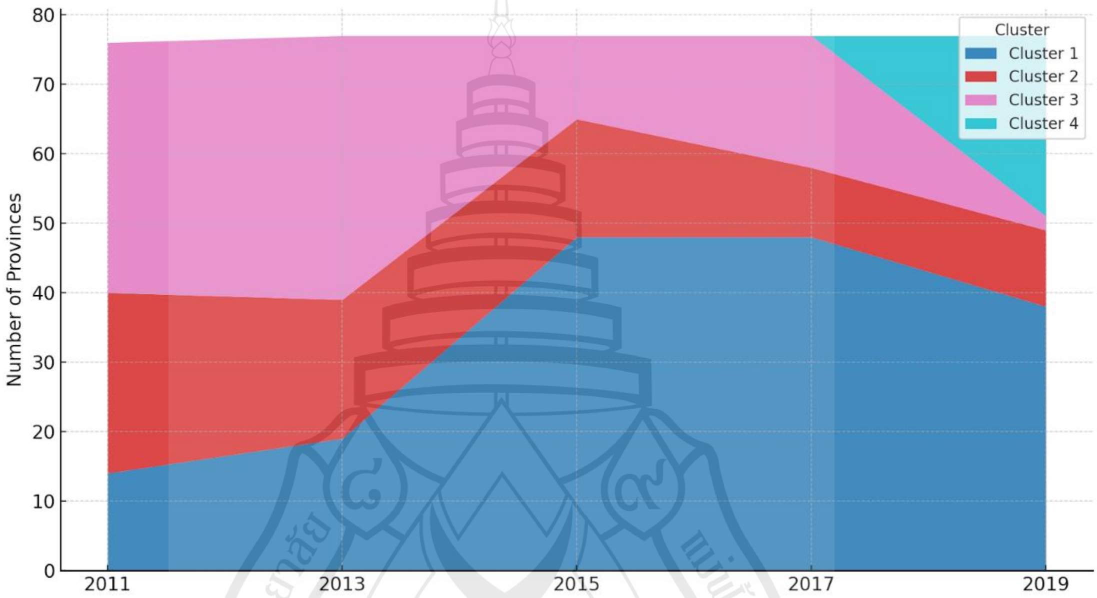


Figure 4.11 Provincial Cluster Composition by Year using K-Means Clustering

The results reveal notable shifts in cluster membership, particularly in 2015 and 2019. In 2015 and 2017, one cluster (Cluster 1) dominated the composition, grouping nearly two-thirds of provinces, suggesting a period of increasing convergence in provincial socio-economic characteristics. However, by 2019, the emergence of a fourth cluster indicates growing differentiation, possibly reflecting diverging development paths or newly emergent regional profiles. The relatively balanced segmentation in earlier years contrasts with the polarization observed in the final year of analysis.

4.3.2.2 Hierarchical agglomerative clustering (HAC) results

Hierarchical Agglomerative Clustering (HAC) was applied using the same standardized dataset and year-specific optimal number of clusters used in the other algorithms. The algorithm was implemented using Ward's linkage method, which minimizes the total within-cluster variance at each merging step. As a deterministic, tree-based clustering technique, HAC produces stable results across repeated runs and does not rely on random initialization.

Figure 4.12 presents the composition of clusters over time using a stacked area chart. Across the first four benchmark years, the three-cluster configuration yielded relatively balanced distributions of provinces. Some fluctuation occurred—for example, Cluster 1 shrank between 2011 and 2017, while Cluster 2 grew significantly in 2013 and remained dominant in 2017. The introduction of a fourth cluster in 2019, although small, reflects the increasing heterogeneity of provincial socio-economic structures and HAC's ability to adapt to this differentiation through post-hoc cutting of the dendrogram.

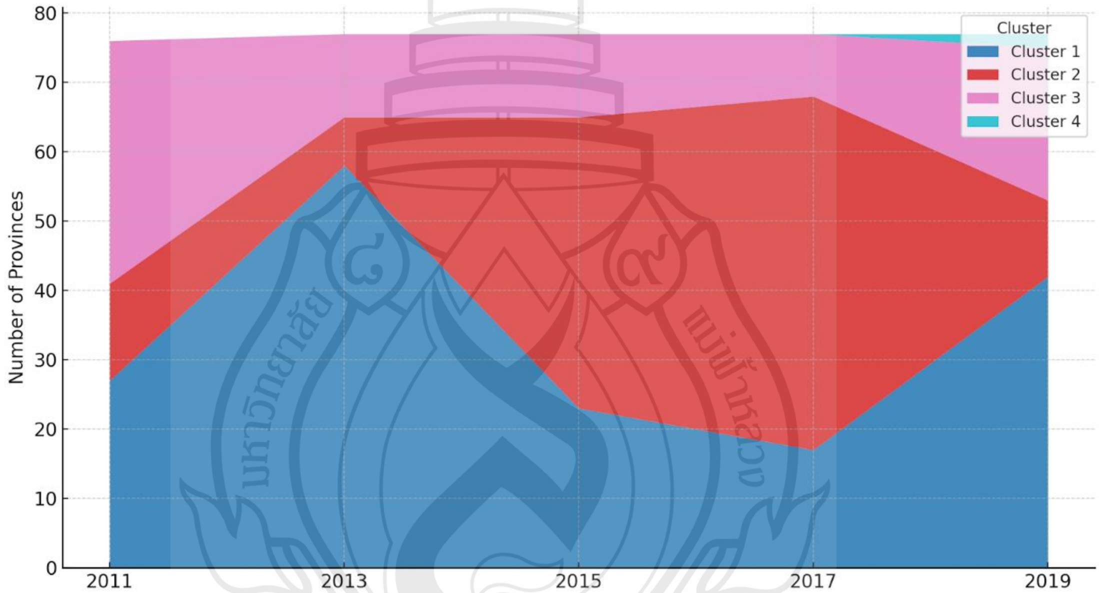


Figure 4.12 Provincial Cluster Composition by Year using HAC

4.3.2.3 Gaussian mixture models (GMM) results

Gaussian Mixture Models (GMM) were applied using the same standardized provincial dataset and the year-specific optimal number of clusters. GMM differs from K-Means and HAC in that it models each cluster as a Gaussian distribution, allowing for elliptical cluster shapes and unequal variances. The algorithm uses a probabilistic assignment approach via the Expectation-Maximization (EM) procedure, which enables it to capture overlapping structures and latent heterogeneity in the data.

Figure 4.13 shows the GMM results using a stacked area chart of cluster composition over time. From 2011 to 2017, the model classified provinces into three groups, with Cluster 1 consistently dominating—especially in 2013 and 2017, where it included more than half of all provinces. In 2019, a fourth cluster emerged, capturing a distinct group of transitional or outlier provinces. This reflects GMM's flexibility in detecting previously latent subgroupings as inter-provincial differentiation increased.

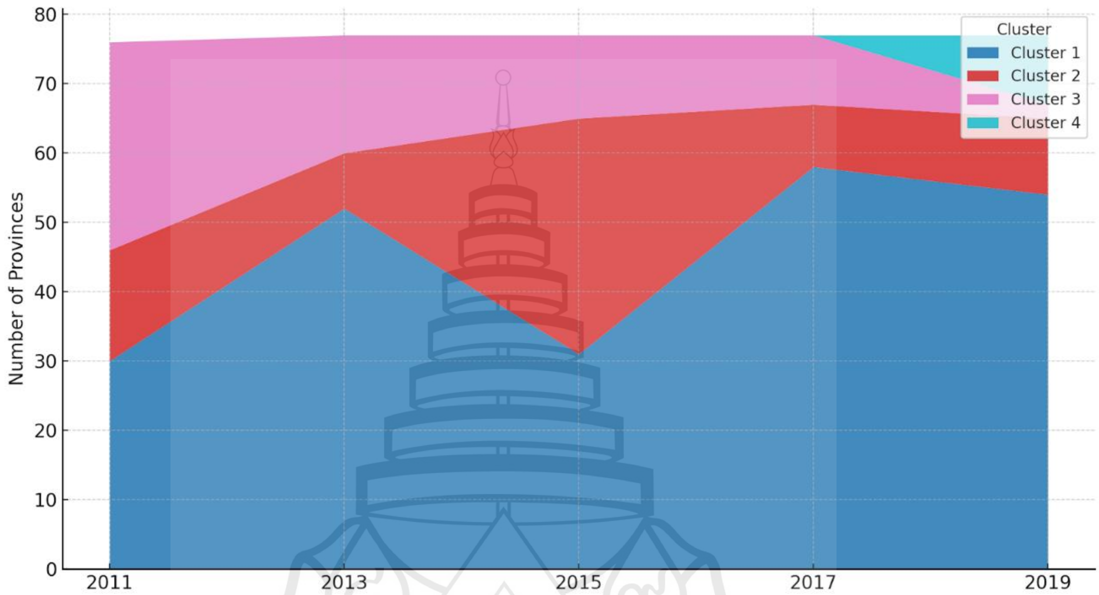


Figure 4.13 Provincial Cluster Composition by year using GMM

4.3.2.4 Comparative summary of clustering results

This section summarizes and compares the outputs of the three clustering algorithms—K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) applied across five benchmark years (2011, 2013, 2015, 2017, and 2019) using consistent, year-specific optimal values of k = 3 for 2011–2017 and k = 4 for 2019),While all three methods used the same input dataset and cluster counts, their underlying mechanics produced distinctive segmentation outcomes.

In general, there was moderate to strong alignment across algorithms in classifying provinces with clear socio-economic characteristics. Provinces such as Bangkok, Chonburi, and Rayong, known for their high economic complexity and income levels were consistently grouped together across all methods. Similarly, clusters of provinces with lower RGPPPC and ECI tended to converge across algorithms. However, discrepancies became more pronounced in provinces with transitional characteristics, such as those experiencing rapid economic change or possessing mixed development indicators.

K-Means often produced balanced group sizes, but its reliance on Euclidean distance and sensitivity to initial centroid selection led to more abrupt changes in cluster membership over time. HAC offered the most stable and hierarchical segmentation, showing smoother year-to-year transitions but limited responsiveness to emerging clusters, as evident in 2019 when its fourth cluster remained small and less defined. GMM was the most flexible, assigning clusters based on probabilistic likelihoods. It captured subtler structural variation, particularly in 2019, when it identified a more substantial fourth cluster representing provinces with increasingly distinct profiles.

Visual comparisons, particularly the stacked area charts (Figures 4.15–4.17), reveal that while core provincial groupings remain largely consistent, differences in cluster size and membership dynamics reflect the influence of algorithmic design. For example, GMM and K-Means both revealed a shift in the provincial structure in 2019 through the expansion of a new cluster, whereas HAC maintained a conservative partition with less dramatic redistribution. Overall, while there is no single "correct" clustering outcome, these results emphasize the importance of methodological fit to research goals. Each algorithm offers unique strengths and interpretability in K-Means, stability in HAC, and flexibility in GMM. These differences inform the rationale for selecting a preferred algorithm, which is addressed in the next section.

4.3.2 Preferred Clustering Algorithm

To determine the most suitable clustering algorithm for sub-national socio-economic classification, this section evaluates and compares the empirical performance of K-Means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM) across three standard clustering validation metrics: Silhouette Score, Calinski–Harabasz Index (CHI), and Davies–Bouldin Index (DBI).

The Silhouette Score measures how well each province fits within its assigned cluster (higher is better), while the Calinski–Harabasz Index evaluates the ratio of between-cluster variance to within-cluster variance (also higher is better). The Davies–Bouldin Index, by contrast, assesses average intra-cluster similarity and should be minimized.

From Table 4.16, K-Means exhibits the best overall performance, achieving the highest Silhouette and CH scores, and a competitive DBI. HAC slightly edges K-Means in DBI, indicating marginally tighter cluster cohesion, but underperforms on the other two metrics. GMM, while theoretically capable of modeling more flexible and overlapping clusters, produces the lowest scores across all three indicators—suggesting that it may overfit or fail to identify well-separated provincial groupings in this context.

Table 4.16 Average Clustering Validation Scores by Algorithm

Algorithm	Silhouette Score	Calinski–Harabasz Index	Davies–Bouldin Index
K-Means	0.2477	27.6544	1.2034
HAC	0.2233	25.0067	1.1948
GMM	0.1839	21.2638	1.4256

In addition to empirical scores, interpretability and implementation considerations are important. K-Means is computationally efficient, transparent, and produces intuitive, easily communicable cluster structures. HAC offers deterministic and stable outputs but is less responsive to evolving data. GMM's flexibility comes at the cost of interpretability and increased complexity in explaining cluster membership probabilistically. Based on this multi-criteria validation, K-Means is selected as the preferred algorithm for the remainder of the analysis. It balances interpretability, responsiveness to socio-economic differentiation, and strong statistical validation, making it well-suited for profiling regional development clusters and tracking their transformation over time.

4.3.5 Cluster Dynamics and Profiles (K-Means)

Following the selection of K-Means as the preferred clustering algorithm, this section presents an in-depth analysis of the resulting provincial clusters over time. Using the optimal number of clusters identified for each year—three clusters in 2011 to 2017, and four in 2019—the analysis focuses on both the socio-economic profiles that characterize each cluster and the transitional dynamics of provinces across years. By examining how clusters evolve, expand, or contract, the section highlights structural shifts in Thailand's regional development landscape. The profile of each cluster is explored using aggregated indicators such as the Economic Complexity Index (ECI), real gross provincial product per capita (RGPPPC), population density, and local government budget per capita. These variables reveal patterns of economic diversification, income levels, and administrative capacity that differentiate each group. A combination of radar plots, descriptive statistics, and cluster maps is used to illustrate the distinctive characteristics and trajectories of each cluster. Moreover, transition patterns are examined to determine whether provinces maintain stable memberships or shift between clusters over time. These dynamics provide insights into development convergence or divergence and signal potential policy-relevant groupings for targeted regional planning.

4.3.5.1 Socio-economic profiles of clusters

This section examines the socio-economic characteristics of provincial clusters generated by the K-Means algorithm, using average values of six indicators: the Economic Complexity Index (ECI), real gross provincial product per capita (RGPPPC), GINI coefficient, population, population density, and local budget per capita. These indicators provide a comprehensive snapshot of structural, demographic, and fiscal dimensions across provincial groupings. The cluster profiles by year are summarized in Table 4.17, which reports the mean values for each variable by cluster membership from 2011 to 2019. The results reveal persistent structural differentiation across provinces, though the internal composition and relative standing of clusters evolve over time.

In 2011, for example, Cluster 1 represented provinces with moderate ECI and RGPPPC but relatively high inequality, suggesting that these areas had some degree of productive diversification but lacked inclusive economic distribution. In contrast, Cluster 3 included provinces with high ECI, high RGPPPC, and moderate inequality, capturing the economically advanced urban centres. Cluster 2, with lower values across most dimensions, reflected peripheral or rural regions with limited complexity and fiscal resources.

By 2013, the profile of Cluster 1 had shifted, with notable increases in ECI and local budgets, indicating that some mid-tier provinces may have benefited from policy or investment-driven upgrading. However, inequality remained high in these areas. The most affluent cluster continued to exhibit strong economic performance and complexity but began to show slight increases in inequality, suggesting a concentration of benefits. In 2015, differentiation across clusters became more pronounced. One group emerged with high GINI but only moderate ECI and RGPPPC, likely reflecting provinces with unequal gains from growth.

Table 4.17 Average Socio-Economic Characteristics of K-Means Clusters

Year	Cluster	ECI	RGPPPC	GINI	Population	Population Density	Budget
2011	1	0.10	230634.57	0.24	778431.50	193.63	3235.28
	2	0.26	60676.50	0.36	1394624.38	425.16	6432.49
	3	0.17	59007.08	0.35	458614.56	115.06	1943.63
2013	1	-0.35	210501.05	0.26	755516.58	247.71	4101.50
	2	-0.33	68345.60	0.34	1594634.60	492.32	11760.65
	3	-0.27	55479.50	0.36	487852.68	114.06	2660.34
2015	1	0.03	56233.33	0.33	592701.50	117.21	5025.07
	2	0.46	219180.00	0.25	796294.12	274.52	7360.20
	3	0.01	89386.17	0.32	1978535.50	758.81	60507.97
2017	1	0.30	64238.08	0.35	565277.85	98.78	4250.46
	2	0.90	311446.70	0.22	828404.90	275.08	6845.60
	3	0.32	92686.58	0.30	1619532.47	627.66	108051.39
2019	1	-0.50	71403.87	0.34	476767.71	84.71	4209.57
	2	0.04	310249.27	0.22	854885.45	293.60	12191.71
	3	0.74	293203.50	0.24	3465825.50	462.00	984241.00
	4	-0.46	60542.31	0.30	1234860.42	369.53	10253.12

Meanwhile, another cluster retained a balanced profile with elevated ECI and RGPPPC alongside moderate inequality—likely comprising innovation-driven and diversified provinces. In 2017, the structure remained relatively stable, although population and density figures diverged more significantly, with one cluster now including densely populated urban areas, suggesting an increased urban-rural divide in structural complexity. Budget per capita also showed greater variance across clusters, reflecting growing fiscal decentralization or targeted investment.

By 2019, the number of clusters increased to four, allowing for greater granularity. One cluster captured low ECI but high GINI, underscoring a risk of inequality deepening in structurally stagnant provinces. Another included province with modest economic performance but low inequality, likely rural provinces with more equitable but limited economic outcomes. The economically advanced cluster remained relatively consistent in structure, continuing to show high complexity and income per capita.

Overall, the evolution of cluster profiles over the study period indicates both path dependence and dynamic structural change. Provinces with high complexity tend to maintain their advantage, while transitions occur primarily among middle-performing regions. The interplay between economic complexity, inequality, and fiscal capacity is evident across clusters, offering important insights for targeted regional development policies. Table 4.17 thus provides not only a static comparison but a temporal map of Thailand's sub-national development landscape through the lens of economic complexity. To further contextualize the cluster classifications, it is useful to examine the typical provinces found within each group and their regional distribution.

Cluster 1: Traditional Economies with Demographic Weight is predominantly composed of provinces in the Northeast (Isan) and parts of the Lower North, such as Nakhon Ratchasima, Ubon Ratchathani, and Phitsanulok. These provinces have substantial populations but continue to exhibit low economic complexity and modest income levels. Their economies often rely heavily on agriculture and state transfers, with limited integration into high-value industrial or service sectors.

Cluster 2: Transitional Economies with Structural Constraints includes provinces located in the Upper Central and Northern regions, such as Phetchabun, Loei, and Kamphaeng Phet. These areas exhibit signs of industrial and service sector expansion but still face institutional and infrastructural limitations that hinder sustained structural transformation. Inequality is often more pronounced in these provinces due to uneven access to opportunities within emerging growth centres.

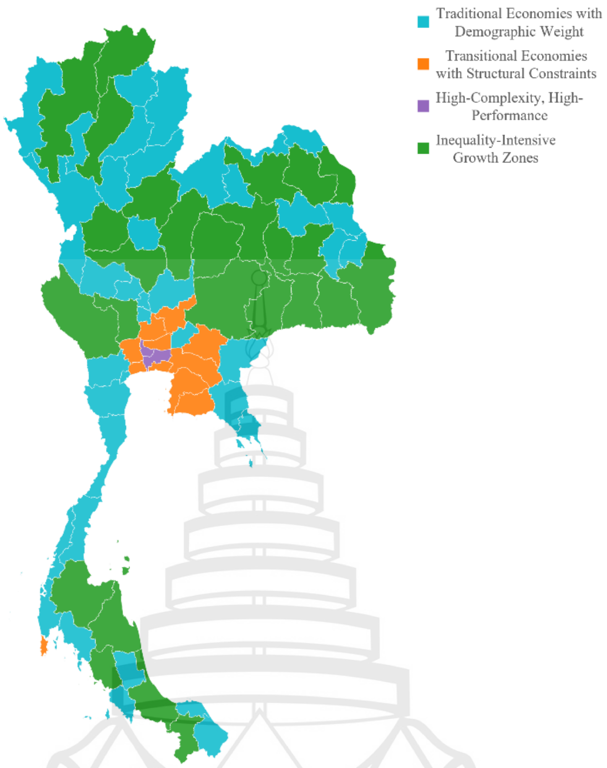


Figure 4.14 Final Cluster Clustering of Thai Provinces (2019)

Cluster 3: High-Complexity, High-Performance is concentrated in the Bangkok Metropolitan Region, the Eastern Economic Corridor (EEC), and parts of the Central Plains, including Bangkok, Chonburi, Rayong, and Pathum Thani. These provinces demonstrate advanced productive structures, high levels of income and innovation capacity, and relatively inclusive growth outcomes. They function as hubs for international trade, investment, and knowledge-intensive industries.

Cluster 4: Inequality-Intensive Growth Zones, which emerged in 2019, is composed of provinces that exhibit recent surges in income and fiscal growth but are marked by high internal disparities. Examples include Chiang Mai, Khon Kaen, and Surat Thani. These provinces often serve as regional urban centers experiencing rapid development due to tourism, education, or infrastructure expansion, yet without corresponding improvements in income distribution or structural complexity.

This spatial distribution reinforces the observation that Thailand's development landscape is both geographically uneven and structurally diverse. Clusters are not randomly distributed but instead reflect regionally embedded patterns of development and divergence, shaped by historical investment, administrative centrality, and market access.

4.3.5.2 Transition patterns and cluster stability

This section examines the dynamics of provincial development trajectories using two complementary visual tools: the cluster ribbon plot (Figure 4.15) and a series of provincial cluster maps (Figure 4.16).

The temporal progression of Thailand's provincial development patterns between 2011 and 2019 can be understood through the analysis of cluster membership transitions across the four typologies defined in this study. Provinces were classified into Traditional Economies, Transitional Economies, High-Complexity, High-Performance regions, and Inequality-Intensive Growth Zones, based on their economic complexity and socioeconomic profiles. The visual representation of these transitions (Figure 4.15) reveals both structural persistence and inter-cluster mobility, with distinct regional characteristics. Many provinces, particularly in the Northeast and parts of the North remained consistently within the Traditional Economies cluster, indicating ongoing challenges related to limited industrial diversification and sustained inequality.

In contrast, provinces categorized as Transitional Economies exhibited more frequent movement, reflecting uneven structural change and varying degrees of policy responsiveness or investment absorption capacity. High-Complexity provinces, such as Bangkok, Rayong, and Chonburi, demonstrated cluster stability throughout the period, consistent with their advanced industrial ecosystems and high innovation capacity. The emergence of the Inequality-Intensive Growth cluster in 2019, encompassing provinces such as Chiang Mai, Khon Kaen, and Surat Thani, highlights a new development trajectory marked by increasing income and infrastructure investments, yet accompanied by widening internal disparities.

This temporal cluster perspective provides valuable insights into the differentiated trajectories of provincial development. It underscores the importance of tailoring policy approaches to the distinct dynamics of each cluster, particularly to support lagging regions, sustain high-complexity centres, and manage inequality in rapidly growing urbanizing zones.

Figure 4.15 Cluster Ribbon Plot of Provincial Transitions (K-Means, 2011–2019)

While the ribbon plot emphasizes temporal dynamics, the provincial cluster maps in Figure 4.16 provide a spatial lens through which to interpret these transitions. Each map shows the geographical distribution of clusters each year, enabling the identification of regional patterns and persistent spatial inequalities. In 2011 and 2013, the maps indicate clear regional clustering: for example, Cluster 2 is heavily concentrated in the northern and northeastern provinces, while Cluster 1 dominates parts of the South and Central regions.

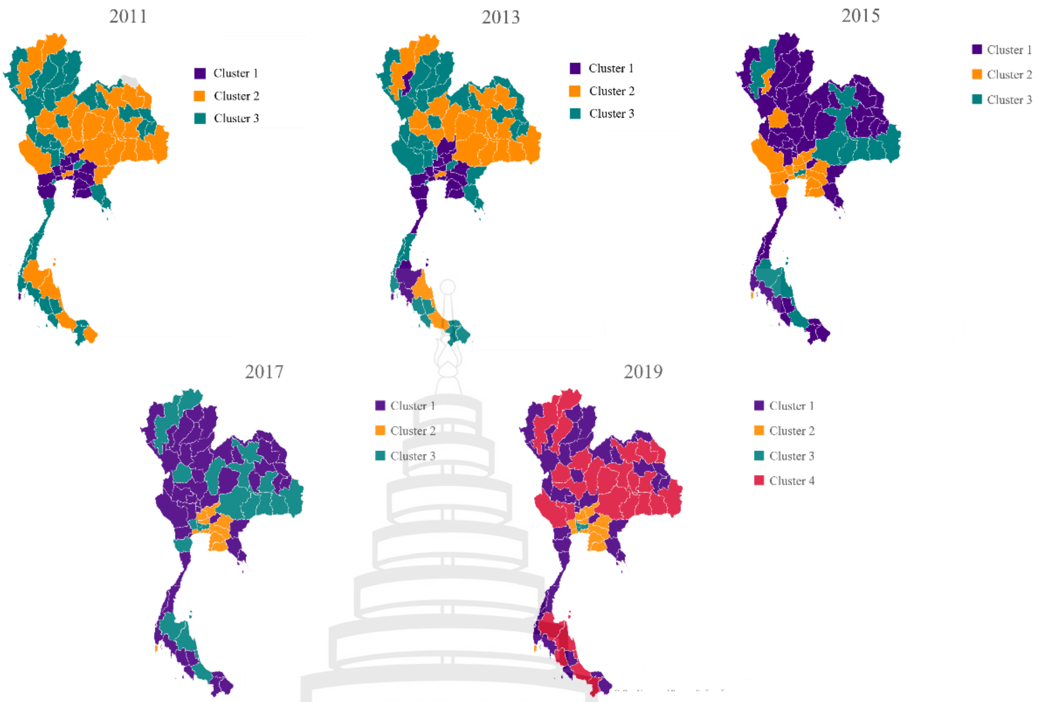


Figure 4.16 Clustering of Thai Provinces by Year (2011–2019)

This spatial pattern reflects long-standing regional development disparities, influenced by differences in infrastructure, market access, and policy focus. Over time, however, the spatial distribution becomes more fragmented and differentiated. By 2015 and 2017, previously homogeneous zones begin to exhibit internal divergence, with provinces within the same region assigned to different clusters. The 2019 map is particularly notable, as the introduction of Cluster 4 reorganizes several provinces into a distinct group not previously captured. This reclassification suggests that new socio-economic configurations are emerging, potentially driven by policy interventions, diversification in industrial bases, or region-specific shocks.

The visual progression over time reveals that while macro-regional development patterns largely persist, micro-regional differentiation is increasing. Provinces within the same region are following increasingly divergent trajectories, highlighting the need for more localized policy approaches. The ribbon plot and cluster maps reflect both continuity and gradual transformation in Thailand's sub-national landscape, underscoring the importance of recognizing temporal stability alongside emerging spatial reconfigurations in regional development planning.

CHAPTER 5

CONCLUSION AND DISCUSSION

5.1 Summary of Key Findings

This study set out to explore the role of economic complexity in shaping provincial-level development outcomes in Thailand. Through a novel integration of econometric modeling and machine learning, the research generated a range of empirical insights across three main dimensions: (1) measurement of subnational complexity, (2) estimation of its impact on growth and inequality, and (3) classification of provinces into structural development clusters. Each set of findings contributes to a deeper understanding of how Thailand's economic geography has evolved and what structural conditions underpin its persistent regional disparities.

The first major contribution of this study is the construction of a provincial-level Economic Complexity Index (ECI) using employment data from the Labor Force Survey (LFS), applying the method of reflection and a location quotient-based RCA framework. This approach addresses a critical data gap in subnational development diagnostics, particularly in contexts where export or firm-level data are not systematically available across provinces. The employment-based complexity matrix successfully generated interpretable ECI and PCI scores for all 77 provinces across five benchmark years (2011, 2013, 2015, 2017, and 2019), revealing substantial variation in structural diversification across regions. Bangkok and key economic corridors consistently ranked highest in ECI, while rural provinces in the North and Northeast exhibited persistently low complexity values. The observed spatial patterns are consistent with Thailand's uneven structural transformation and labor market segmentation.

The second key finding relates to the impact of economic complexity on Real Gross Provincial Product Per Capita (RGPPPC), used as a proxy for long-run economic performance. Results from fixed effects panel regression models confirm that complexity is a statistically significant and positive predictor of economic growth, even after controlling population size, education, and industrial composition. More importantly, generalized additive models (GAMs) reveal that the relationship is nonlinear, with a clear threshold effect: provinces with medium-to-high levels of ECI experience disproportionately greater growth benefits, while marginal gains diminish or plateau at very high levels of complexity. This suggests that the accumulation of productive capabilities yields increasing returns up to a point, after which other institutional or spatial constraints may limit additional gains.

The third major finding concerns the distributional effects of economic complexity. Panel quantile regression models show that the relationship between ECI and the Gini coefficient is both negative and heterogeneous. In provinces with higher baseline inequality (upper quantiles of the Gini distribution), increases in complexity are associated with statistically significant reductions in inequality. In contrast, the effects are weaker and less consistent in lower-inequality contexts. These results suggest that complexity can serve as a mechanism for inclusive development, particularly in structurally disadvantaged regions. GAM results further support the presence of nonlinear effects, indicating that the inequality-reducing impact of complexity is strongest during the early to middle stages of structural transformation.

The fourth major set of findings derives from the clustering analysis, which applied unsupervised machine learning techniques—K-means, Hierarchical Agglomerative Clustering (HAC), and Gaussian Mixture Models (GMM)—to identify development typologies among Thailand's 77 provinces. Based on their economic complexity, growth, inequality, and demographic features, four distinct clusters consistently emerged as optimal across methods and years. Each cluster reflects a unique development configuration:

Cluster 1: Economic Diversification and Moderate Inequality, provinces in this cluster exhibit moderate economic complexity, relatively balanced income distribution, and diversified employment structures. Representative provinces include Chiang Mai, Khon Kaen, and Songkhla.

Cluster 2: Economic Prosperity and High Complexity, this group comprises provinces with consistently high RGPP per capita and advanced production structures. Bangkok, Rayong, and Chonburi are notable examples, representing high-value industrial and service economies.

Cluster 3: High Population and Economic Output, Low Complexity, these provinces have large populations and considerable economic output, but their complexity remains low, and inequality is often elevated. Examples include Nakhon Ratchasima and Ubon Ratchathani.

Cluster 4: Structural Lag and Persistent Inequality, characterized by low complexity, slow growth, and entrenched inequality, this group includes rural and economically marginalized provinces such as Mae Hong Son, Amnat Charoen, and Nong Bua Lamphu.

Beyond classification, the study also tracked cluster transitions between 2011 and 2019 to evaluate structural mobility. The transition matrix revealed that 67.5% of provinces remained in the same cluster over time, indicating moderate path dependency in structural characteristics. However, some provinces demonstrated upward mobility. For instance, Chachoengsao and Saraburi transitioned from Cluster 3 to Cluster 2, reflecting successful structural upgrading linked to proximity to the Eastern Economic Corridor and industrial base expansion. In contrast, some provinces such as Suphan Buri moved downward from Cluster 2 to Cluster 3, suggesting relative decline in complexity or rising inequality.

The cluster ribbon plot and map visualization confirmed regional polarization patterns: Cluster 2 provinces are heavily concentrated in the Central and Eastern regions, while Cluster 4 provinces are prevalent in the North and Northeast. These spatial dynamics reinforce the importance of region-specific strategies to address structural divergence.

5.2 Policy Implications and Interpretation

The empirical findings of this study have several important implications for economic policy, particularly in the context of Thailand's ongoing efforts to achieve spatially balanced and inclusive development. By demonstrating that economic complexity is a significant predictor of both economic growth and income inequality at the provincial level, this research offers a compelling case for integrating structural diagnostics—such as the Economic Complexity Index (ECI)—into Thailand's subnational development frameworks. The results underscore the inadequacy of conventional indicators such as Gross Provincial Product per capita when used in isolation, and support the adoption of multidimensional, capability-based planning tools that can more accurately capture a province's long-term development potential.

First and foremost, the study shows that complexity-based indicators offer a deeper understanding of regional economic dynamics than traditional output-based metrics. Provinces with similar GPP levels often differ significantly in their underlying productive capabilities. The inclusion of ECI in planning instruments such as the Development Potential Assessment Index (DPAI) or SDG Provincial Tracker could help local authorities and national policymakers identify where structural upgrading is feasible and where targeted interventions are required. This is especially important for provinces in Cluster 1 and Cluster 4, which may not be low-income but are structurally constrained in their growth potential due to limited complexity.

The positive and nonlinear relationship between complexity and economic growth suggests that industrial policies should focus not merely on sector expansion but on upgrading to more knowledge-intensive and diverse economic structures. In practice, this means investing in regional innovation systems, vocational training linked to emerging industries, and supply chain integration strategies tailored to each province's latent capabilities. For example, transition provinces such as Chachoengsao, which moved from Cluster 3 to Cluster 2, could serve as models for other intermediate provinces seeking to scale up their complexity levels. This aligns with recent scholarship suggesting that complexity is not just an outcome but a strategic pathway toward development (Hartmann et al., 2017; Hidalgo, 2021).

The inverse and distribution-sensitive relationship between complexity and inequality implies that inclusive growth is structurally conditioned. This finding is particularly salient for provinces in Cluster 3, where high population levels and low complexity often coincide with elevated inequality. In such contexts, complexity upgrading should be coupled with redistributive mechanisms—such as employment formalization, skill-building for marginalized groups, and sectoral diversification into labour-absorptive, high-value-added industries. As the quantile regression results suggest, complexity has the strongest inequality-reducing effect in provinces that are already at higher levels of inequality, meaning targeted upgrading policies in structurally lagging regions can yield disproportionately large social returns.

The clustering analysis provides a powerful tool for designing spatially differentiated policies. Rather than applying a one-size-fits-all approach, national development strategies can use cluster-based profiles to tailor interventions. For instance, Cluster 2 provinces may benefit more from advanced R&D incentives and regional innovation ecosystems. Cluster 1 and Cluster 3 provinces may require investment in logistics, education, and industrial diversification to avoid stagnation. Cluster 4 provinces should be prioritized for foundational capability-building, including workforce development, access to digital infrastructure, and basic institutional strengthening. Moreover, the transition matrix and ribbon plot reveal that some provinces are already on a trajectory of structural transformation. These transition zones represent policy leverage points, where timely support can accelerate upward mobility and prevent regression.

Finally, the study demonstrates that complexity metrics can be reliably generated from high-frequency, nationally available employment data—making them scalable and policy-relevant in real time. Unlike export-based complexity measures, which may be unavailable or unsuitable at the subnational level, employment-based metrics allow for routine provincial monitoring, especially in middle-income countries like Thailand. Institutions such as the NESDC or Ministry of Interior could incorporate ECI scores into dashboard systems for early warning, investment targeting, or interprovincial benchmarking.

5.3 Contributions to Literature

This study makes several substantive contributions to the literature on economic complexity, regional development, and the application of machine learning in social science research. By focusing on Thailand, a middle-income country with pronounced regional disparities and applying a novel subnational lens, the research bridges gaps between global theory, national policy, and local realities. These contributions span conceptual, methodological, and empirical domains.

Most empirical applications of the Economic Complexity Index (ECI) have focused on national-level analysis, particularly in relation to global trade and competitiveness (Hidalgo & Hausmann, 2009; Hausmann et al., 2014). While recent studies have begun to explore regional complexity in countries such as Mexico (Chávez et al., 2017), Romania (Török et al., 2022), and the United States (Fritz & Manduca, 2021), few have examined its implications in Southeast Asian or upper-middle-income economies. This study adds to this emerging literature by demonstrating that economic complexity can be effectively measured at the provincial level using employment data, and that it has meaningful implications for both growth and inequality in the Thai context.

Whereas previous studies have primarily examined the average impact of complexity on macroeconomic outcomes, this thesis contributes to a growing strand of research that considers its distributional implications. By applying panel quantile regression, the study uncovers heterogeneous effects of complexity on income inequality across provinces with differing baseline conditions. This approach moves beyond mean-based interpretations and adds empirical support to the argument that economic complexity can serve as a structural tool for inclusive development—particularly in regions experiencing high inequality.

Methodologically, the research contributes to both the econometric and computational social science literature by combining fixed effects models, generalized additive models (GAMs), and unsupervised machine learning techniques. While the use of econometric models is well-established in economic complexity studies, the application of GAMs allows for the exploration of nonlinearities and threshold effects, a dimension often overlooked in prior work. Furthermore, the use of clustering algorithms (K-means, HAC, and GMM) to identify structural provincial typologies represents a novel application in the Thai context. This interdisciplinary integration demonstrates how tools from information technology and data science can deepen insight into socio-economic structures.

The use of publicly available, high-frequency employment data to construct subnational complexity indices offers a replicable, scalable, and policy-relevant methodology for other middle-income countries where export data are often sparse or unreliable at the regional level. This contribution is especially valuable for researchers and policymakers seeking to operationalize complexity metrics within decentralized or subnational governance frameworks. The study provides practical steps for adapting the method of reflection to labor market data, with clear implications for research design in comparative regional analysis.

Finally, the study enriches Thailand's domestic policy discourse by introducing a structural dimension to regional diagnostics. While Thailand's planning instruments emphasize income, employment, and basic industrial data, this research introduces complexity as a third axis of development analysis—highlighting productive capabilities as a foundation for sustainable and inclusive growth. It thereby aligns Thailand's regional strategy with international development thinking, which increasingly prioritizes knowledge-based upgrading and institutional capabilities.

Collectively, these contributions position the study at the intersection of regional economics, complexity science, and applied data analytics. By operationalizing complexity theory in a practical, replicable, and policy-relevant manner, the thesis offers a foundation for further academic exploration and real-world application in Thailand and comparable contexts.

5.4 Limitations and Future Research

While this study offers new empirical insights and methodological innovations in the application of economic complexity to subnational development, several limitations must be acknowledged. These limitations pertain to data constraints, methodological boundaries, interpretative scope, and generalizability. A clear articulation of these constraints provides both transparency and a foundation for future research trajectories.

One of the primary limitations lies in the data used to construct the provincial-level Economic Complexity Index (ECI). The employment-based approach, while replicable and publicly accessible, may not fully capture the depth of knowledge embedded in certain economic activities, particularly those in the informal sector, digital economy, or knowledge-intensive services that are underrepresented in standard labour force classifications. Furthermore, the 20-sector disaggregation level used in the Labor Force Survey (LFS) limits the granularity of product and industry mapping, potentially underestimating productive sophistication in provinces with niche or emerging sectors.

Although fixed effects models are used to address time-invariant unobserved heterogeneity, the issue of endogeneity remains a concern. Economic complexity and development outcomes may be mutually reinforcing, raising the possibility of reverse causality. While this study focuses on observed associations rather than strict causality, future work could employ instrumental variable (IV) strategies or dynamic panel techniques (e.g., system GMM) to further isolate causal effects. Additionally, the specification of control variables, limited by data availability may omit other relevant provincial-level factors such as infrastructure quality, political institutions, or innovation systems. Including such covariates would enhance model completeness and interpretation but requires integrated multi-source datasets not readily available in the current scope.

While the application of unsupervised machine learning algorithms (K-means, HAC, GMM) yielded interpretable development clusters, the outputs are inherently sensitive to initial conditions, number of clusters selected, and choice of variables. Although multiple algorithms and validation techniques were applied to ensure robustness, the interpretation of clusters remains exploratory. Moreover, cluster boundaries may shift with even minor updates in data or feature sets, which could affect their application in long-term planning without periodic recalibration.

The research is designed around the specific institutional, geographic, and statistical context of Thailand. While the methodology is replicable, its direct applicability may be constrained in countries lacking high-frequency, subnational employment data or with different administrative structures. The generalizability of cluster typologies and policy implications should therefore be cautiously extended beyond Thailand without contextual adaptation.

Building upon these limitations, several promising directions for future research emerge:

Incorporate dynamic panel models and causal inference techniques to better understand the directionality and long-term feedback mechanisms between complexity and development outcomes.

Expand outcome variables to include additional dimensions such as employment quality, educational attainment, human development indices, or innovation output (e.g., patents), to provide a multidimensional view of complexity's impacts.

Explore spatial spillovers and regional interactions using spatial econometric models to assess whether complexity effects in one province influence adjacent provinces through trade, labour mobility, or institutional diffusion.

Develop real-time dashboards or visualization tools to support institutional adoption of subnational complexity monitoring for policy targeting, investment promotion, and decentralized development planning.

By addressing these limitations and expanding on current findings, future research can further strengthen the theoretical, empirical, and practical utility of economic complexity in regional development analysis, particularly in emerging and transitional economies.

5.5 Conclusion

This thesis has investigated the influence of economic complexity on provincial-level development outcomes in Thailand, with a specific focus on economic growth and income inequality. By constructing a novel subnational Economic Complexity Index (ECI) based on employment data from the Labor Force Survey (LFS), and by employing a combination of fixed effects regression, panel quantile regression, generalized additive models (GAMs), and unsupervised machine learning techniques, the study provides a comprehensive analysis of the structural dynamics underpinning Thailand's regional development landscape.

The empirical findings demonstrate that economic complexity is positively and significantly associated with real gross provincial product per capita (RGPPPC), and negatively associated with income inequality, as measured by the Gini coefficient. These relationships are shown to be nonlinear and heterogeneous across provinces, indicating that the developmental returns to complexity vary with both structural and distributional conditions. In particular, the inequality-reducing effects of complexity are most pronounced in provinces with higher baseline levels of inequality, suggesting that complexity may serve as a vehicle for inclusive growth in structurally disadvantaged regions. The application of clustering algorithms further reveals that provinces can be meaningfully grouped into development typologies that capture latent structural heterogeneity. These typologies, along with their observed transitions over time, highlight both the persistence of regional disparities and the potential for upward mobility through targeted capability accumulation.

From a theoretical perspective, the study contributes to the expanding body of literature on economic complexity by operationalizing the concept at the subnational level within a middle-income country context. It offers a methodological advancement by demonstrating that employment-based complexity metrics can be reliably constructed using publicly available data, thus enhancing the applicability of complexity diagnostics in data-constrained environments. Furthermore, the integration of nonlinear modelling and unsupervised learning provides a multidimensional framework for understanding development pathways that are neither uniform nor linear.

In practical terms, the study underscores the value of complexity-informed diagnostics for regional planning and policy design. The findings suggest that the incorporation of ECI and related structural indicators into Thailand's development planning apparatus, such as the Development Potential Assessment Index (DPAI) or SDG-aligned monitoring systems could strengthen evidence-based, place-specific policy formulation. The cluster-based typologies offer an empirical basis for spatially differentiated strategies, while the observed transitions underscore the importance of continuous monitoring and adaptive policymaking.

In conclusion, this research affirms that economic complexity constitutes more than a descriptive metric of output diversity; it encapsulates the embedded capabilities, institutional knowledge, and production potential of regional economies. As such, complexity-oriented approaches hold significant promise for guiding Thailand and comparable emerging economies toward more inclusive, sustainable, and structurally resilient development trajectories.

REFERENCES

Acemoglu, D., Akcigit, U., & Kerr, W. (2016). Innovation network. *Proceedings of the National Academy of Sciences, 113*(41), 11483–11488. https://doi.org/10.1073/pnas.1613559113

Adebayo, T. S., Rjoub, H., Akadiri, S. S., Oladipupo, S. D., Sharif, A., & Adeshola, I. (2022). The role of economic complexity in the environmental Kuznets curve of MINT economies: evidence from method of moments quantile regression. *Environmental Science and Pollution Research, 29*(16), 24248–24260. https://doi.org/10.1007/s11356-021-17524-0

Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, *60*(2), 323–351. https://doi.org/10.2307/2951599

Aghion, P., Cherif, R., & Hasanov, F. (2021). *Competition, innovation, and inclusive growth*. International Monetary Fund.

Albeaik, S., Kaltenberg, M., Alsaleh, M., & Hidalgo, C. A. (2017). Improving the economic complexity index. *arXiv preprint arXiv:1707.05826*. https://doi.org/10.48550/arXiv.1707.05826

Allison, P. D. (2009). *Fixed effects regression models*. Sage Publications.

Altuntas, S., Selim, S., & Altuntas, F. (2022). A hierarchical clustering based panel data approach: A case study of regional incentives. *International Journal of Information Management Data Insights, 2*(2), 100098. https://doi.org/10.1080/00036846.2014.904491

Alvarado, R., Tillaguango, B., Dagar, V., Ahmad, M., Işık, C., Méndez, P., & Toledo, E. (2021). Ecological footprint, economic complexity and natural resources rents in Latin America: Empirical evidence using quantile regressions. *Journal of Cleaner Production, 318*, 128585. https://doi.org/10.1016/j.jclepro.2021.128585

Antonescu, D. (2014). *Theoretical approaches of regional development*. https://mpra.ub.uni-muenchen.de/60627/

Antonicelli, M., Drago, C., Costantiello, A., & Leogrande, A. (2025). *Analyzing Income Inequalities across Italian regions: Instrumental Variable Panel Data, K-Means Clustering and Machine Learning Algorithms*. HAL. https://hal.science/hal-05091404

Antonietti, R. (2024). Economic complexity: Local and global impacts on inequality. In *Global Handbook of Inequality* (pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-97417-6_62-1

Apaitan, T., Ananchotikul, N., & Disyatat, P. (2017). *Structural transformation in Thailand: A perspective through product innovation* (No. 72). Puey Ungphakorn Institute for Economic Research.

Argüelles, M., Benavides, C., & Fernández, I. (2014). A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Applied Economics, 46*(21), 2511–2519. https://doi.org/10.1080/00036846.2014.904491

Ashikbayeva, Z., Fürstenberg, M., Kapelari, T., Pierres, A., & Thies, S. (2020). *Household level effects of flooding: Evidence from Thailand* (TVSEP Working Paper No. WP-022). Thai-Vietnam Socio-Economic Panel Project (TVSEP). https://www.tvsep.de/fileadmin/user_upload/publications/wp022.pdf

Ashraf, S., Jithin, P., Slim, S., & Najeeb, R. (2023). Global value chains and economic complexity index: Evidence from generalized panel quantile regression. *Economic Analysis and Policy, 80*, 347–365. https://doi.org/10.1016/j.eap.2023.08.003

Awoa, P. A., Ondoa, H. A., & Tabi, H. N. (2024). Natural resources and income inequality: economic complexity is the key. *Environment and Development Economics, 29*(2), 127–153.

Azomahou, T. T., Diene, B., & Diene, M. (2013). Nonlinearities in productivity growth: A semi-parametric panel analysis. *Structural Change and Economic Dynamics, 24*, 45–75. https://doi.org/10.1016/j.strueco.2012.05.001

Bahrami, F., Shahmoradi, B., Noori, J., Turkina, E., & Bahrami, H. (2023). Economic complexity and the dynamics of regional competitiveness a systematic review. *Competitiveness Review: An International Business Journal, 33*(4), 711–744. https://doi.org/10.1108/CR-06-2021-0083

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage. *The Manchester School*, *33*(2), 99–123. https://doi.org/10.1111/j.1467-9957.1965.tb00050.

Balland, P.-A., & Boschma, R. (2021). Complementary interregional linkages and smart specialization: An empirical study on European regions. *Regional Studies, 55*(6), 1051–1063. https://doi.org/10.1080/00343404.2020.1861240

Balland, P.-A., & Rigby, D. L. (2017). The geography of complex knowledge. *Economic Geography, 93*(1), 1–23. https://doi.org/10.1080/00130095.2016.1205947

Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2020). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies, 54*(9), 1252–1268. https://doi.org/10.1080/00343404.2018.1437900

Balsalobre, S. J. P., Verduras, C. L., & Lanchas, J. D. (2017). Measuring the economic complexity at the sub-national level using international and interregional trade. In *Nineteenth Annual Conference of European Trade Study Group*. European Trade Study Group.

Bandeira Morais, M. (2023). *Essays on the role of economic complexity in regional development* [Doctoral dissertation, London School of Economics and Political Science]. LSE Theses Online. https://etheses.lse.ac.uk/4567/

Bandeira Morais, M., Swart, J., & Jordaan, J. A. (2021). Economic complexity and inequality: does regional productive structure affect income inequality in Brazilian states?. *Sustainability, 13*(2), 1006. https://doi.org/10.3390/su13021006

Basile, R., & Cicerone, G. (2022). Economic complexity and productivity polarization: Evidence from Italian provinces. *German Economic Review, 23*(4), 567–594. https://doi.org/10.1515/ger-2021-0070

Bedemo Beyene, A. (2024). Income inequality and economic complexity in Africa: the moderating role of governance quality. *Cogent Social Sciences*, *10*(1), 2341114.

Bennett, N., & Lemoine, J. (2014). What VUCA really means for you. *Harvard business review*, *92*(1/2). https://ssrn.com/abstract=2389563

Bhahari, R. H., & Kusnawi, K. (2024). Clustering Analysis of Socio-Economic Districts/Cities In East Java Province Using PCA And Hierarchical Clustering Methods. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(4), 2242–2251. https://doi.org/10.33395/sinkron.v8i4.14078

Billings, S. B., & Johnson, E. B. (2012). The location quotient as an estimator of industrial concentration. *Regional Science and Urban Economics, 42*(4), 642–647. https://doi.org/10.1016/j.regsciurbeco.2012.03.003

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Bureau of Registration Administration. (n.d.). *Population and Population Density Statistics*. Department of Provincial Administration, Ministry of Interior, Thailand. Retrieved June 16, 2025, from https://stat.bora.dopa.go.th/

Bureau of the Budget. (n.d.). *Annual Budget Allocation Report by Province*. Office of the Prime Minister, Thailand. Retrieved June 20, 2025, from https://www.bb.go.th/

Cajias, M., & Ertl, S. (2018). Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions?. *Journal of Property Investment & Finance*, *36*(1), 32–49. https://doi.org/10.1108/JPIF-10-2016-0080

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27. https://doi.org/10.1080/03610927408827101

Canay, I. A. (2011). A simple approach to quantile regression for panel data. *The econometrics journal*, *14*(3), 368–386. https://doi.org/10.1111/j.1368-423X.2011.00349.x

Carvalho, A. X. Y., Albuquerque, P. H. M., de Almeida Junior, G. R., & Guimaraes, R. D. (2009). Spatial hierarchical clustering. *Revista Brasileira de Biometria*, *27*(3), 411–442. https://biometria.ufla.br/antigos/fasciculos/v27/v27_n3/A6_Alexandre.pdf

Chakraborty, A., Inoue, H., & Fujiwara, Y. (2020). Economic complexity of prefectures in Japan. *PloS one*, *15*(8), e0238017. https://doi.org/10.1371/journal.pone.0238017

Chang, C. H., Tan, S., Lengerich, B., Goldenberg, A., & Caruana, R. (2021, August). How interpretable and trustworthy are gams?. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 95–105). Association for Computing Machinery. https://doi.org/10.1145/3447548.3467453

Chávez, J. C., Mosqueda, M. T., & Gómez-Zaldívar, M. (2017). Economic complexity and regional growth performance: Evidence from the Mexican economy. *Review of Regional Studies, 47*(2), 201–219. http://dx.doi.org/10.52324/001c.8023

Chong, B. (2021). K-means clustering algorithm: a brief review. Academic Journal of *Computing & Information Science, 4*(5), 37–40. https://doi.org/10.25236/AJCIS.2021.040506

Chu, K. L. (2023). Determinants of economic complexity revisited: Insightful understanding from panel quantile regression. *Journal of Economic and Banking Studies*, *5*, 30–44. http://doi.org/10.59276/JEBS.2023.06.2448

Chu, X., Gao, D., Mao, W., Guo, Z., Huang, Y., & Li, L. (2019, July). Study on Urbanization Level Based on Principal Component and Cluster Analysis--A Case Study of Hunan Province. In *2019 16th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1–8). IEEE. https://doi.org/10.1109/ICSSSM.2019.8887644

Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G., & Pietronero, L. (2013). Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS ONE, 8*(8), e70726. https://doi.org/10.1371/journal.pone.0070726

Davies, D. L., & Bouldin, D. W. (2009). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dawkins, C. J. (2003). Regional development theory: Conceptual foundations, classic works, and recent developments. *Journal of Planning Literature, 18*(2), 131–172. https://doi.org/10.1177/0885412203254706

de Queiroz Stein, A., Britto, G., & Medeiros, V. (2024). Agricultural economic complexity and regional inequalities: a new approach using census data from Brazil. *Regional Studies, 58*(12), 2368–2388. https://doi.org/10.1080/00343404.2024.2411345

deHaan, E. (2021). *Using and interpreting fixed effects models* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3699777

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Diodato, D., Neffke, F., & O'Clery, N. (2018). Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106, 1–26. https://doi.org/10.1016/j.jue.2018.05.002

Durães, D., de la Prieta, F., & Novais, P. (2019). Enriching behavior patterns with learning styles using peripheral devices. *Knowledge and Information Systems*, *60*, 1645–1662. https://doi.org/10.1007/s10115-019-01339-4

Eckenrode, J., Smith, E. G., McCarthy, M. E., & Dineen, M. (2014). Income inequality and child maltreatment in the United States. *Pediatrics*, *133*(3), 454–461. https://doi.org/10.1542/peds.2013-1707

El Hatimi, R., Choukhan, C. F., & Esghir, M. (2024). An Improved K-means Clustering Algorithm Towards an Efficient Educational and Economical Data Modeling. *International Journal of Advanced Computer Science & Applications*, *15*(1). https://dx.doi.org/10.14569/IJACSA.2024.01501109

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). John Wiley & Sons. https://doi.org/10.1002/9780470977811

Fa'rifah, R. Y., & Pramesti, D. (2022). Cluster Analysis of Inclusive Economic Development Using K-Means Algorithm. *Jurnal Varian, 5*(2), 171–178. http://dx.doi.org/10.30812/varian.v5i2.1894

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association, 97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Fredrich, J. M. (2023). Economic complexity and income inequality across countries and regions. *International Review of Applied Economics, 37*(1), 1–25.

Fritz, B. S., & Manduca, R. A. (2021). The economic complexity of US metropolitan areas. *Regional Studies, 55*(7), 1299–1310. https://doi.org/10.1080/00343404.2021.1884215

Fromentin, V. (2021). Cross-border workers in the Greater Region of Luxembourg and financial instability: a non-linear approach. *Applied economics, 53*(27), 3171–3192. https://doi.org/10.1080/00036846.2021.1877251

Fusco, G., & Perez, J. (2015, July). Spatial analysis of the Indian Subcontinent: The complexity investigated through neural networks. In *14th international conference on Computers in Urban Planning and Urban Management (CUPUM 2015)* (p. 287). HAL. https://hal.science/hal-01303057

Galvao Jr, A. F. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics*, *164*(1), 142–157. https://doi.org/10.1016/j.jeconom.2011.02.016

Gao, J., & Zhou, T. (2018). Quantifying China's regional economic complexity. *Physica A: Statistical Mechanics and its Applications*, *492*, 1591–1603. https://doi.org/10.1016/j.physa.2017.11.084

Ghosh, S., Doğan, B., Can, M., Shah, M. I., & Apergis, N. (2023). Does economic structure matter for income inequality?. *Quality & Quantity*, *57*(3), 2507–2527. https://doi.org/10.1007/s11135-022-01462-1

Gill, I. S., & Kharas, H. (2015). *The middle-income trap turns ten* (Policy Research Working Paper No. 7403). World Bank. https://doi.org/10.1596/1813-9450-7403

Gómez-Zaldívar, M., Osorio-Caballero, M. I., & Saucedo-Acosta, E. J. (2022). Income inequality and economic complexity: Evidence from Mexican states. *Regional Science Policy & Practice*, *14*(6), 344–364. https://doi.org/10.1111/rsp3.12580

Graham, B. S., Hahn, J., Poirier, A., & Powell, J. L. (2015). *Quantile regression with panel data* (No. w21034). National Bureau of Economic Research. https://doi.org/10.3386/w21034

Greene, W. (2001). Estimating econometric models with fixed effects. *Department of Economics, Stern School of Business, New York University*.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression* (No. 149). Sage.

Hartmann, D., & Pinheiro, F. L. (2024). Economic Complexity and Inequality at the National and Regional Levels. In *Routledge International Handbook of Complexity Economics* (pp. 551–566). Routledge. https://doi.org/10.4324/9781003119128-39

Hartmann, D., Guevara, M. R., Jara-Figueroa, C., Aristarán, M., & Hidalgo, C. A. (2017). Linking economic complexity, institutions, and income inequality. *World development*, *93*, 75–93. https://doi.org/10.1016/j.worlddev.2016.12.020

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall. https://doi.org/10.1201/9780203753781

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical science, 1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., & Yıldırım, M. A. (2014). *The atlas of economic complexity: Mapping paths to prosperity* (2nd ed.). MIT Press.

Hausmann, R., Pietrobelli, C., & Santos, M. A. (2021). Place-specific determinants of income gaps: New sub-national evidence from Mexico. *Journal of Business Research, 131*, 782–792. https://doi.org/10.1016/j.jbusres.2021.01.003

Hausmann, R., Pietrobelli, C., & Santos, M. A. (2021). The network structure of economic output: The product space and its consequences for economic growth. In J. Stiglitz & J.-P. Fitoussi (Eds.), *The Measurement of Economic Performance and Social Progress Revisited* (pp. 147–161). Palgrave Macmillan.

He, Y., He, Y., Zhan, Z., Philippe, F. V., & Huang, J. Z. (2025). A Novel Subspace-Based GMM Clustering Ensemble Algorithm for High-Dimensional Data. *Chinese Journal of Electronics, 34*(2), 612–629. https://doi.org/10.23919/cje.2023.00.153

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). Russell Sage Foundation.

Hidalgo, C. A. (2009). *The dynamics of economic complexity and the product space over a 42 year period*. CID Working Paper Series.

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics, 3*(2), 92–113. https://doi.org/10.1038/s42254-020-00275-1

Hidalgo, C. A. (2021). *How humans judge machines*. MIT Press.

Hidalgo, C. A. (2023). The policy implications of economic complexity. *Research Policy, 52*(9), 104863. https://doi.org/10.1016/j.respol.2023.104863

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences, 106*(26), 10570–10575. https://doi.org/10.1073/pnas.0900943106

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

Hoeriyah, L., Nuryartono, N., & Pasaribu, S. H. (2022). Economic complexity and sustainable growth in developing countries. *Economics Development Analysis Journal, 11*(1), 23–33. http://dx.doi.org/10.15294/edaj.v11i1.47294

Homsombat, W., Wrasai, P., & Benjabutr, N. (2025). Measuring the impact of creative city attributes on regional economic development in Thailand. *Asia-Pacific Journal of Regional Science*, 1–29. https://doi.org/10.1007/s41685-025-00379-z

Hossain, M. R., Dash, D. P., Das, N., Ullah, E., & Hossain, M. E. (2024). Green energy transition in OECD region through the lens of economic complexity and environmental technology: A method of moments quantile regression perspective. *Applied Energy, 365*, 123235. https://doi.org/10.1016/j.energy.2025.136412

Huang, T., & Li, J. (2018). Semiparametric model average prediction in panel data analysis. *Journal of Nonparametric Statistics, 30*(1), 125–144. https://doi.org/10.1080/10485252.2017.1404061

Hunter, P. R., Colón-González, F. J., Brainard, J., Majuru, B., Pedrazzoli, D., Abubakar, I., . . . Semenza, J. C. (2020). Can economic indicators predict infectious disease spread? A cross-country panel analysis of 13 European countries. *Scandinavian Journal of Public Health*, 48(4), 351–361. https://doi.org/10.1177/1403494819852830

Ikram, M., Xia, W., Fareed, Z., Shahzad, U., & Rafique, M. Z. (2021). Exploring the nexus between economic complexity, economic growth and ecological footprint: Contextual evidences from Japan. *Sustainable Energy Technologies and Assessments, 47*, 101460. https://doi.org/10.1016/j.seta.2021.101460

Inoua, S. (2023). A simple measure of economic complexity. *Research Policy, 52*(7), 104793. https://doi.org/10.1016/j.respol.2023.104793

Isserman, A. M. (1977). The location quotient approach to estimating regional economic impacts. *Journal of the American Institute of Planners, 43*(1), 33–41. https://doi.org/10.1080/01944367708977758

Jaewisorn, A., & Aroonruengsawat, A. (2020). *The inequality effects of natural disasters: Evidence from Thailand* [Doctoral dissertation, Thammasat University].

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jitsuchon, S. (2012). Thailand in a middle-income trap. *TDRI Quarterly Review, 27*(2), 13–20.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. https://doi.org/10.1007/BF02289588

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000, May). The analysis of a simple k-means clustering algorithm. In *Proceedings of the sixteenth annual symposium on Computational geometry* (pp. 100–109). https://doi.org/10.1145/336154.336189

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kazemzadeh, E., Fuinhas, J. A., & Koengkan, M. (2022). The impact of income inequality and economic complexity on ecological footprint: an analysis covering a long-time span. *Journal of Environmental Economics and Policy*, *11*(2), 133–153. https://doi.org/10.1080/21606544.2021.1930188

Kemp-Benedict, E. (2014). *An interpretation and critique of the Method of Reflections*. Stockholm Environment Institute. https://mpra.ub.uni-muenchen.de/60705/

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*(6), 441–458.

Khanzadi, A., Tavassoli Nia, A., Behnia, A., & Soltani, M. (2022). Analyzing the effect of economic complexity on income inequality in Iran. *Journal of Development and Capital, 7*(2), 1–19.

Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge University Press.

Koenker, R. (2017). *Handbook of quantile regression*. CRC Press.

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143–156. https://doi.org/10.1257/jep.15.4.143

Korkmaz, İ., Çınar, İ. T., & Baycan, T. (2024). Economic Complexity and Regional Growth in Turkey: Implications for East-West Disparity. *Eastern European Economics*, 1–27. https://doi.org/10.1080/00128775.2024.2304220

Kuznets, S. (1955). Economic growth and income inequality. *The American Economic Review, 45*(1), 1–28.

Lee, C. C., & Wang, E. Z. (2021). Economic complexity and income inequality: Does country risk matter? *Social Indicators Research, 154*(1), 35–60. https://doi.org/10.1007/s11205-020-02543-0

Lee, K. K., & Vu, T. V. (2020). Economic complexity, human capital and income inequality: A cross-country analysis. *The Japanese Economic Review, 71*(4), 695–718. https://doi.org/10.1007/s42973-019-00026-7

Li, Y., & Rigby, D. (2023). Relatedness, complexity, and economic growth in Chinese cities. *International Regional Science Review, 46*(1), 3–37. https://doi.org/10.1177/01600176221082308

Lin, J. Y. (2012). *New structural economics: A framework for rethinking development and policy*. World Bank Publications.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Lopes, J. C., Dias, J., & Amaral, J. F. (2012). Assessing economic complexity as interindustry connectedness in nine OECD countries. *International Review of Applied Economics, 26*(6), 811–827. https://doi.org/10.1080/02692171.2012.665852

Ma, G., Li, H., & Luo, K. (2005, August). Application of clustering in regional economy. *In Proceedings of the 7th international conference on Electronic commerce* (pp. 48–51). https://doi.org/10.1145/1089551.1089563

Marco, R., Llano, C., & Pérez-Balsalobre, S. (2022). Economic complexity, environmental quality and income equality: A new trilemma for regions?. *Applied Geography*, *139*, 102646. https://doi.org/10.1016/j.apgeog.2022.102646

Mariani, M. S., Vidmer, A., Medo, M., & Zhang, Y. C. (2015). Measuring economic complexity of countries and products: which metric to use?. *The European Physical Journal B*, *88*, 1–9. https://doi.org/10.1140/epjb/e2015-60390-4

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis, 55*(7), 2372–2387. https://doi.org/10.1016/j.csda.2011.02.004

Marshall, A. (1890). *Principles of economics* (1st ed.). Macmillan.

McLachlan, G. (2000). Finite mixture models. *A wiley-interscience publication*.

Mealy, P., & Coyle, D. (2022). To them that hath: Economic complexity and local industrial strategy in the UK. *International Tax and Public Finance, 29*(2), 358–377. https://doi.org/10.1007/s10797-021-09667-0

Mealy, P., Farmer, J. D., & Teytelboym, A. (2019). Interpreting economic complexity. *Science Advances, 5*(1), eaau1705. https://doi.org/10.1126/sciadv.aau1705

Mewes, L., & Broekel, T. (2022). Technological complexity and economic growth of regions. *Research Policy, 51*(8), 104156. https://doi.org/10.1016/j.respol.2022.104156

Miller, M. M., Gibson, L. J., & Wright, N. G. (1991). Location quotient: A basic tool for economic development analysis. *Economic Development Review, 9*(2), 65.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179.

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). McGraw-hill.

Monastiriotis, V. (2009). Examining the consistency of spatial association patterns across socio-economic indicators: an application to the Greek regions. *Empirical Economics*, *37*, 25–49. https://doi.org/10.1007/s00181-008-0221-2

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*. https://doi.org/10.48550/arXiv.1109.2378

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, *31*, 274–295. https://doi.org/10.1007/s00357-014-9161-z

National Economic and Social Development Council (NESDC). (2021). *Thailand Sustainable Development Goals Report 2021*. Office of the Prime Minister, Thailand.

National Statistical Office. (n.d.). *Household Socio-economic survey*. Ministry of Digital Economy and Society, Thailand. Retrieved June 17, 2025, from https://www.nso.go.th/

National Statistical Office. (n.d.). *Labor Force Survey: Whole Kingdom*. Ministry of Digital Economy and Society, Thailand. Retrieved June 16, 2025, from https://www.nso.go.th/

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

Nguyen, C. P., Nguyen, B. Q., & Tran, D. T. L. (2023). Economic complexity and income inequality: New evidence of a nonlinear effect. *Social Science Quarterly, 104*(4), 829–868. https://doi.org/10.1111/ssqu.13281

Odell, K. E. (2009). *Nonlinearities and economic growth new evidence*. University of Illinois at Chicago.

Office of the National Economic and Social Development Council. (n.d.). *Gross provincial product and regional indicators*. Retrieved June 21, 2025, from https://www.nesdc.go.th/

Ourens, G. (2013). *Can the Method of Reflections help predict future growth?* Institute for Research in Economic and Social Sciences (IRES), Université catholique de Louvain.

Ozkan, O., Haruna, R. A., Alola, A. A., Ghardallou, W., & Usman, O. (2023). Investigating the nexus between economic complexity and energy-related environmental risks in the USA: empirical evidence from a novel multivariate quantile-on-quantile regression. *Structural Change and Economic Dynamics, 65*, 382–392. https://doi.org/10.1016/j.strueco.2023.03.010

Paul, S., Hasan, A. N., & Mario, B. (2018, November). Exploring the non-linear relationship between various categories of Crimes and GDP: A case study using Generalized Additive Models. In *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (pp. 1–5). IEEE. https://doi.org/10.1109/LA-CCI.2018.8625254

Pavone, P., Pagliacci, F., Russo, M., Righi, S., & Giorgi, A. (2021). Multidimensional clustering of EU regions: a contribution to orient public policies in reducing regional disparities. *Social Indicators Research, 156*, 739–759. https://doi.org/10.1007/s11205-020-02324-9

Payne, J. E., Truong, H. H. D., Chu, L. K., Doğan, B., & Ghosh, S. (2023). The effect of economic complexity and energy security on measures of energy efficiency: Evidence from panel quantile analysis. *Energy Policy, 177*, 113547.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Perafita, X., & Saez, M. (2022). Clustering of small territories based on axes of inequality. *International Journal of Environmental Research and Public Health, 19*(6), 3359. https://doi.org/10.3390/ijerph19063359

Pérez-Balsalobre, S., Llano Verduras, C., & Díaz-Lanchas, J. (2019). *Measuring subnational economic complexity: An application with Spanish data* (JRC Working Paper No. 05/2019). European Commission, Joint Research Centre. https://publications.jrc.ec.europa.eu/repository/handle/JRC116253

Pérez-Balsalobre, S., Llano-Verduras, C., & Díaz-Lanchas, J. (2019). *Measuring subnational economic complexity: An application with Spanish data* (JRC Working Papers on Territorial Modelling and Analysis No. 05/2019; JRC Technical Report JRC116253). European Commission. https://publications.jrc.ec.europa.eu/repository/handle/JRC116253

Pham, M. H., Truong, H. D. H., & Hoang, D. P. (2024). Economic complexity, shadow economy, and income inequality: Fresh evidence from panel data. *International Economic Journal, 38*(2), 270–292. https://doi.org/10.1080/10168737.2024.2311704

Pietronero, L., Cristelli, M., & Tacchella, A. (2013, April). New metrics for economic complexity: Measuring the intangible growth potential of countries. In *Proceedings of conference at Institute for New Economic Thinking (INET* (pp. 4–7). Institute for New Economic Thinking.

Powell, D. (2022). Quantile regression with nonadditive fixed effects. *Empirical Economics, 63*(5), 2675–2691. https://doi.org/10.1007/s00181-022-02216-6

Prasetiya, R. H. (2021). Do economic complexity and strong institutions affect income inequality? *Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning, 5*(3), 296–316. https://doi.org/10.36574/jpp.v5i3.213

Pu, H., Wang, S., Wang, Z., Ran, Z., & Jiang, M. (2022). Non-linear relations between life expectancy, socio-economic, and air pollution factors: a global assessment with spatial disparities. *Environmental Science and Pollution Research, 29*(35), 53306–53318. https://doi.org/10.1007/s11356-022-19644-7

Pugliese, E., Chiarotti, G. L., Zaccaria, A., & Pietronero, L. (2017). Complex economies have a lateral escape from the poverty trap. *PloS ONE, 12*(1), e0168540. https://doi.org/10.1371/journal.pone.0168540

Rahadini, A. (2024, August). Regional typology in Central Java: Urban and rural relation based on cluster analysis. *IOP Conference Series: Earth and Environmental Science, 1381*(1), 012043. IOP Publishing. https://doi.org/10.1088/1755-1315/1381/1/012043

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review, 56*(8), 8219–8264. https://doi.org/10.1007/s10462-022-10366-3

Reynolds, C., Agrawal, M., Lee, I., Zhan, C., Li, J., Taylor, P., . . . Roos, G. (2018). A sub-national economic complexity analysis of Australia's states and territories. *Regional Studies, 52*(5), 715–726. https://doi.org/10.1080/00343404.2017.1283012

Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 659–663). Springer. https://doi.org/10.1007/978-0-387-73003-5_196

Ricardo, D. (1817). *On the principles of political economy and taxation*. John Murray.

Rodriguez Andres, A., Otero, A., & Amavilah, V. H. (2022). Knowledge economy classification in African countries: A model-based clustering approach. *Information Technology for Development, 28*(2), 372–396. https://doi.org/10.1080/02681102.2021.1950597

Rodríguez-Pose, A. (2018). The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society, 11*(1), 189–209. https://doi.org/10.1093/cjres/rsx024

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy, 98*(5, Part 2), S71–S102. https://doi.org/10.1086/261725

Roos, G., Drok, T. E., Novikova, A., & Farafonova, Y. (2021). Subnational economic complexity analysis: Case study of the Kaliningrad region. *E3S Web of Conferences, 291*, 07010. EDP Sciences. https://doi.org/10.1051/e3sconf/202129107010

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

Sapra, S. K. (2013). Generalized additive models in business and economics. *International Journal of Advanced Statistics and Probability, 1*(3), 64–81. http://dx.doi.org/10.14419/ijasp.v1i3.1022

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Servén, D., & Brummitt, C. (2018). *pyGAM: Generalized additive models in Python* [Computer software documentation]. https://pygam.readthedocs.io

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Solow, R. M. (1956). A contribution to the theory of economic growth. *The Quarterly Journal of Economics, 70*(1), 65–94. https://doi.org/10.2307/1884513

Srebro, N., Shakhnarovich, G., & Roweis, S. (2006, June). An investigation of computational and informational limits in gaussian mixture clustering. In *Proceedings of the 23rd international conference on Machine learning* (pp. 865–872). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143953

Stock, J. H., & Watson, M. W. (2020). *Introduction to econometrics* (4th ed.). Pearson Education.

Stojkoski, V., Utkovski, Z., Tevdovski, D., & Kocarev, L. (2017). *Economic complexity and growth: An investigation of empirical regularities* (MPRA Paper No. 77837). Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/77837/

Tabash, M. I., Mesagan, E. P., & Farooq, U. (2022). Dynamic linkage between natural resources, economic complexity, and economic growth: Empirical evidence from Africa. *Resources Policy, 78*, 102865. https://doi.org/10.1016/j.resourpol.2022.102865

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific reports, 2*(1), 723. https://doi.org/10.1038/srep00723

Tanujaya, N. A., Thendito, J. G., Putra, M. R., Barina, F., Kartolo, O., Setiawan, K. E., & Zakiyyah, A. Y. (2024, November). Comparison of K-means, Gaussian Mixture, and Hierarchical Clustering Models on Countries' Economic Freedom Index. In *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1–6). IEEE.

Teixeira, F. O., Missio, F. J., & Dathein, R. (2022). Economic complexity, structural transformation and economic growth in a regional context: Evidence for Brazil. *PSL Quarterly Review, 75*(300), 63–79. https://doi.org/10.13133/2037-3643/17505

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267–276. https://doi.org/10.1007/BF02289263

Topaloglou, L., Kallioras, D., Manetos, P., & Petrakos, G. (2005). A border regions typology in the enlarged European Union. *Journal of Borderlands Studies, 20*(2), 67–89. https://doi.org/10.1080/08865655.2005.9695644

Török, I., Benedek, J., & Gómez-Zaldívar, M. (2022). Quantifying subnational economic complexity: Evidence from Romania. *Sustainability, 14*(17), 10586. https://doi.org/10.3390/su141710586

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik, 24*(5), 383–393. https://doi.org/10.1007/BF01437407

Wahidah, Z., & Utari, D. T. (2023). Comparison of K-means and Gaussian mixture model in profiling areas by poverty indicators. *BAREKENG: Jurnal Ilmu Matematika dan Terapan, 17*(2), 717–726. https://doi.org/10.30598/barekengvol17iss2pp0717-0726

Wang, Y., & Turkina, E. (2020). Economic complexity, product space network and Quebec's global competitiveness. *Canadian Journal of Administrative Sciences, 37*(3), 334–349. https://doi.org/10.1002/cjas.1552

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wardana, A. P., Pravitasari, A. E., & Panuju, D. R. (2023, December). Dynamics of land cover, development level, and regional typology of Central Java Province based on sustainable development index. *IOP Conference Series: Earth and Environmental Science, 1266*(1), 012035. IOP Publishing. https://doi.org/10.1088/1755-1315/1266/1/012035

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420010404

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Xie, X. L. (2010). A theoretical analysis of K-means-type algorithms. *Pattern Recognition, 44*(1), 128–142.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Yeerong, P., & Uttama, S. (2023, November). Clustering Thailand's provinces based on socio-economic data. In *2023 7th International Conference on Information Technology (InCIT)* (pp. 544–549). IEEE. https://doi.org/10.1109/InCIT60207.2023.10413172

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J, 2*(2), 226–235. https://doi.org/10.3390/j2020016

Zhan, Z., Chen, W., & Xu, Z. (2021). Divide China's economic regions in 2019 based on cluster analysis and principal component analysis. *International Journal of Sustainable Development, 7*(4), 112–121. https://doi.org/10.11648/j.ijsd.20210704.11

Zhang, Y., Wang, H. J., & Zhu, Z. (2019). Quantile-regression-based clustering for panel data. *Journal of Econometrics, 213*(1), 54–67. https://doi.org/10.1016/j.jeconom.2019.04.005

Zhu, S., & Li, R. (2017). Economic complexity, human capital and economic growth: Empirical research based on cross-country panel data. *Applied Economics, 49*(38), 3815–3828. https://doi.org/10.1080/00036846.2016.1270413

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer.

APPENDIX A

LIST OF PUBLICATIONS

The following publications are directly related to the research presented in this thesis and were produced during the study:

1. Yeerong, P., & Uttama, S. (2023). Evaluating Thailand's Economic Complexity Using Panel Regression. In Proceedings of the 27th International Computer Science and Engineering Conference (ICSEC 2023). IEEE.
2. Yeerong, P., & Uttama, S. (2023, November). Clustering Thailand's provinces based on socio-economic data. In 2023 7th International Conference on Information Technology (InCIT) (pp. 544-549). IEEE.

These publications reflect the integration of econometric and machine learning techniques for analyzing subnational economic complexity in Thailand.

CURRICULUM VITAE

NAME	Pornpinun Yeerong
EDUCATIONAL BACKGROUND	
2020	Bachelor of Economics Economics Mae Fah Luang University
WORK EXPERIENCE	
2023–Present	Research Assistant SCR Thailand Project, MFU
2011-2022	Research Assistant Office of Border Economy and Logistics Study (OBELS), MFU

PUBLICATION

Yeerong, P., & Uttama, S. (2023). Evaluating Thailand’s economic complexity using panel regression. In *Proceedings of the 27th International Computer Science and Engineering Conference (ICSEC 2023)*. IEEE.

Yeerong, P., & Uttama, S. (2023, November). Clustering Thailand’s provinces based on socio-economic data. In *2023 7th International Conference on Information Technology (InCIT)* (pp. 544–549). IEEE.