WIDER Working Paper 2024/38

Inequality and institutional outcomes in Viet Nam

A combined principal components and clustering analysis

Thu K. Hoang,[1,*] Klarizze Anne M. Puzon,[2] Hoai Thi Thu Dang,[3] and Rachel M. Gisselquist[4]

June 2024

United Nations University World Institute for Development Economics Research

wider.unu.edu

Abstract: Better understanding of inequality, including its relationship to governance and other key outcomes, is relevant both to academic researchers and to policy-makers. Nevertheless, efforts to establish causal relationships empirically remain hampered by the quality and availability of data, especially for Global South countries at the sub-national level. This paper draws on newly available data on income inequality in Viet Nam at the provincial level to show how unsupervised learning techniques might be used as tools in consideration of the relationship between inequality and governance. While previous empirical work in this area has largely used standard techniques such as regression analysis aimed at establishing causal relationships, this is often hampered by the quality and availability of data. Adopting a different approach, this paper applies K-means clustering and principal components analysis (PCA) to show how unsupervised learning techniques can provide relevant insight into structures and patterns in data. Using PCA, it identifies two groupings of provinces based on similarities in institutional quality measures. K-means analysis points to similar relative inequality levels but substantially different absolute inequality and income levels, suggesting two broad 'types' of provinces. The results are suggestive of the positive impact of initial inequality on institutions and that better quality of institutions might reduce inequality for some groups. In general, increased incomes might imply improved inequality and institutional quality outcomes in some cases. A final section considers key limits to such analysis, alongside extensions and further applications.

Key words: relative inequality, absolute inequality, institutional quality, K-means cluster, principal components analysis

JEL classification: C38, D02, D31

Acknowledgements: This study was supported by the Novo Nordisk Foundation Grant NNF19SA0060072. We are grateful to Finn Tarp for thoughtful comments. The authors have no competing interests to declare that are relevant to the content of this article.

[1] Economics Development Department, Academy of Policy and Development, Ministry of Planning and Investment (MPI), Hanoi, Vietnam; [2] Sveriges Lantbruksuniversite, CERE-SLU, Umea, Sweden; [3] Department of Sectoral Policy Studies, Central Institute for Economic Management, MPI, Hanoi, Vietnam; [4] UNU-WIDER, Helsinki, Finland; * corresponding author: Thu K. Hoang, thuhoangkim@gmail.com

This study has been prepared within the project The impacts of inequality on growth, human development, and governance—@EQUAL, supported by the Novo Nordisk Foundation Grant NNF19SA0060072.

ISSN 1798-7237 ISBN 978-92-9267-500-4

https://doi.org/10.35188/UNU-WIDER/2024/500-4

Typescript prepared by Ayesha Chari.

United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland and Sweden, as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

As underscored by the Sustainable Development Goals, there is now broad agreement that reducing inequalities across and within countries is a major global challenge. Understanding inequality, both its patterns and trends and its influences and implications, is relevant not only to academic researchers but also to policy-makers. Yet even basic facts remain to some extent contested. This stems in part from divergent measurement choices. One striking example can be seen in discussion of global income inequality trends: while many studies observe a decline in global income inequality over the past four decades using the standard (relative) Gini coefficient, an increasing trend is, in fact, observed when an absolute inequality measure is considered (Niño-Zarazúa et al. 2017).

Data availability and quality also influence substantially the ability to test core theoretical hypotheses. For instance, there is a significant theoretical literature pointing to the influence of inequality on governance institutions and vice versa (Acemoglu and Robinson 2006; Ansell and Samuels 2010; Boix, 2003; Chong and Gradstein 2007; Haggard and Kaufman 2012; Kotschy and Sunde 2017; Savoia et al. 2010), but efforts to establish causal relationships empirically remain hampered by the quality and availability of data (Ferreira et al. 2022). While cross-country data are challenging, sub-national level data tend to be even more so, especially in Global South countries.

In this paper, we draw on newly available data on income inequality in Viet Nam at the provincial level, alongside a unique provincial-level dataset on governance, to explore how 'unsupervised learning techniques' can be used in analysing social science phenomena. In particular, we use K-means clustering and principal components analysis (PCA). While other empirical work on inequality and governance primarily has adopted supervised learning techniques such as regression analysis, aimed at establishing causal relationships, such analysis is hampered by the data available. We show how unsupervised learning techniques aimed at providing insight into structures and patterns in data, might also be informative—even when available data do not allow for rigorous causal analysis.

Viet Nam provides an interesting case study for this exercise. Over the past decades, it has shown both substantial economic growth and notable declines in poverty (Nguyen and Pham 2018). Although this has largely been a story of 'growth with equity' (Benjamin et al. 2017), some concerns have arisen over rising inequality (World Bank 2014). Research has documented, for instance, both rising horizontal inequality (Dang 2018) and increasing spatial concentration of poverty rates over time (Lanjouw et al. 2016).

As illustrated in Figure 1, Viet Nam's (relative) income inequality as measured by the Gini coefficient remains moderate overall, with some generally moderate variation across provinces. Larger variation in absolute income inequality, as measured by the absolute Gini, however, can be observed, generally correlated with higher variation in income. A question is whether there is an empirical relationship between such variation in inequality and governance quality. While causal analysis of this relationship is challenging given the available data, the use of unsupervised learning techniques offers some insight. Focusing on the period 2011–20, using PCA we identify two groupings of provinces based on similarities in average governance quality measures over the period. K-means analysis shows these have similar relative inequality levels, but substantially different absolute inequality and income levels, suggesting two broad 'types' of provinces. As discussed in Section 3, the current analysis is conducted using 10-year average data due to limitations in time-series PCA methodology as well as data challenges; an extension for future work would be to incorporate time-series analysis.

The next section of this paper introduces and compares PCA and K-means clustering as data science tools. Section 3 then discusses the measures and data used in this analysis, while Section 4 draws on these data to present a brief overview of income inequality and governance quality across provinces in Viet Nam using standard descriptive statistics. The paper then turns to application of PCA and K-means clustering approaches and considers the results of these analyses. A final section concludes.

2 Principal component analysis and clustering as data science tools

When analysing economic phenomena, data science has recently offered novel technical tools. The type of analysis can be divided into two categories: unsupervised learning and supervised learning (Athey 2019; Athey and Imbens 2019). Supervised learning involves identifying causal relationships, where input data (independent variables) are put into a model along with output data (dependent variables). The outcome of supervised learning is training the mathematical model such that it can predict output when it is given new observations. Examples of supervised learning tasks include standard regressions.

In unsupervised learning, by contrast, only input data are provided into the model, thus it does not deal directly with identifying causal relationships. In general, the goal of unsupervised learning is to discover underlying structures or hidden patterns in each dataset (Kleinberg et al. 2017; Mullainathan and Spiess 2017). Clustering is an example of unsupervised learning, where the goal is to partition and group similar observations to uncover trends. Another example is dimensionality reduction or feature extraction. It reduces complexity in the data and uncovers the important features that explain the most variance (Jolliffe and Cadima 2016).

This paper focuses on the use of select unsupervised learning techniques: K-means clustering and principal component analysis (PCA), two of the most commonly used methods in the social sciences (Fonseca 2013). While both are unsupervised learning techniques, they are used for different purposes and have different outcomes.

Clustering is a tool used to group similar observations together using a given similarity measure. The outcome of clustering is to partition the dataset into distinct groups ('clusters'), where instances within a cluster are more similar to each other than they are to other clusters. One of the most often used is K-means clustering. K-means clustering algorithm defines the cluster centroids as the average or the mean of all the observations in the cluster. It uses the Euclidean distance to calculate the similarity between an observation and the cluster centroid. The repetitive iteration used by the algorithm assigns clusters and designates observations to a centroid that has the smallest distance from them.

PCA, on the other hand, is a dimensionality reduction technique. PCA's objective is to discover the most important variables that explain the most variance in the dataset. That is, it is used to reduce the number of variables to be considered in a dataset. It then projects the data onto a new, lower-dimensional space that captures the most integral information. The new features that are generated by PCA are called principal components. These are linear combinations of the original variables (Jolliffe and Cadima 2016).

Clustering can be useful in identifying patterns and trends in economic data, grouping similar economic units, and understanding the underlying economic structures. One common use of clustering in economics is in the analysis of consumer data. Consumer segmentation can be done based on purchasing habits and demographics, which can then prove beneficial in targeted

marketing campaigns (Jolliffe and Cadima 2016). Another application is the identification of spatial clusters of economic activity (Crone 2005). For instance, clustering can be used to identify geographic regions where similar industries are concentrated or to identify characteristics of urban areas.

Similarly, PCA has been used to identify the most relevant variables that could explain economic phenomena better. In financial economics, PCA helps discover patterns in financial data, such as stock prices or exchange rates, that can be used to predict future market trends. For example, researchers have used PCA to identify patterns in stock prices that can be used to develop more accurate stock trading algorithms (Ghorbani and Chong 2020). In relation to the quality of institutions, Coppedge et al. (2008) used PCA of 11 datasets to show that democracy indices measured different phenomena. Variance on the two characteristics of democracy that Robert Dahl outlined in *Polyarchy*—contestation and inclusiveness—makes up approximately 75% of what polity, freedom house, and other indices of democracy have been measuring. More recently, Magyar (2022) has used PCA of data from 17 advanced democracies to show that the most important factors differentiating party systems are the size of the two largest parties and competition between them, whereas standard party system typologies consider mainly the first.

In brief, unsupervised learning techniques, in particular PCA and K-means clustering, can be useful tools for social science analysis, providing insight into underlying data structures that would be difficult to discover through other means. While PCA, and to a lesser extent K-means analysis, has been applied to selected social science topics, there is considerable room for further applications. In considering the use of these tools, it is worth noting also their differences. As Table 1 summarizes, K-means clustering is a categorization tool whereas PCA is about feature extraction, that is, identifying the most useful features that best describe existing data.

Table 1: Comparison of K-means clustering and PCA analyses

Characteristics	K-means	PCA
Definition	• Partition or categorize *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest average (*cluster centroid*) • To assign an observation to a specific cluster, the goal is to minimize the Euclidean distance between each data point and a centroid	• Identifies the 'principal components', reducing the dimensionality of big data sets; it transforms a large set of variables into a smaller one, which still contains most of the information that provides a representation of the large set • Overall, it takes into account correlations between variables; it drops the least important variables and keeps the most integral features that best describe the dataset (i.e. those that provide the most variance)
Use	Cluster analysis: group observations where members of the same group have similar characteristics and are different from observations in other groups	Dimension reduction: extract the most important variables that best explain the nature of the dataset
Type of data required	• Continuous, non-binary • Variable must be normalized (i.e. transform data into similar units)	• Continuous, non-binary • Variable units must be normalized

Source: authors' compilation.

In this paper, we use both of these techniques to consider inequality and its relationship with governance quality in Viet Nam. We first use PCA to consider underlying patterns in indicators of governance quality across Viet Nam's provinces, identifying two new institutional indices with different characteristics. We then form two groups with K-means analysis by using these two new institutional indices as clustering variables, and we compare their descriptive statistics in terms of

relative and absolute inequality to consider whether these two clusters are different in terms of inequality.

3 Data sources and computations

We gathered data from 63 provinces in Viet Nam. Inequality is measured using the Gini index, absolute Gini index, ratio of income of the highest income quintile group and the lowest one (group income ratio), and the absolute income gap of the highest income quintile group and the lowest one (group income gap). Among these four indicators of inequality, the Gini index presents the inequality of the whole income distribution while the two latter indicators focus on the inequality among two ends of the distribution. Income and inequality indicators are computed by Viet Nam's General Statistics Office based on the Viet Nam Households Living Standard Survey (VHLSS) 2010, 2012, 2014, 2016, 2018, and 2020. The General Statistics Office calculates the Gini index using the VHLSS of more than 30,000 households every even year. Values range from zero to one, with values nearest one indicating higher inequality. As the Gini coefficients are released biannually, we averaged the available data in the period to create comparable measures in our study.[1]

One challenge associated with time-dependent analyses involving relative Ginis is that income inequality is quantified in relation to the mean income. This suggests that the time trend of relative inequality indicators are expected to mirror the trends seen in the income or GDP process from which they are derived. This implies that using absolute Ginis may be better suitable for research that include the consideration of time (Bandyopadhyay 2018). In this research, both relative and absolute Gini coefficients will be applied to discover the insight structures and patterns and for comparison. The absolute Gini coefficient (A_t) was generated by multiplying the relative Gini coefficient by provincial income levels following Bandyopadhyay (2018):

$$A_t = G_t \times \mu_t \tag{1}$$

where μ_t is the mean income of income distribution Y_t and G_t is the relative Gini coefficient where

$$G_t = G(Y_t) = \frac{1}{2n^2\mu_t}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_{it} - y_{jt}\right|$$

In Section 4, we briefly discuss patterns and trends in these data across the 63 provinces. In the analysis presented in Section 5, we use averages over the decade 2011–20. The primary reason for this is the absence of established statistical theory that supports PCA analysis with time-series data (Zhang and Tong 2022). An additional challenge is that the inequality and governance data have gaps and missing years that are not consistent with one another.[2]

Provincial income or the gross regional domestic product (GRDP) is the final result of production performed by locally residential production units. At the level of provinces under the central government of Viet Nam, GRDP is calculated by the production approach. Accordingly, GRDP

[1] As we do not have access to the VHLSS, we obtained data for this study from Viet Nam's General Statistics Office upon request.

[2] An alternative approach would have been to use two sets of 5-year averages, but the four resultant clusters in the subsequent cluster analysis then would not be directly related and comparable.

is the sum of the value added at basic price of all economic activities plus taxes on products less subsidies on products. For the analysis conducted in this paper, provincial income is denominated in Vietnamese currency (Vietnamese Dong or VND).

We also used governance quality data from the Viet Nam Provincial Governance and Public Administration Performance Index (PAPI). This index was developed through a collaborative effort between the Centre for Community Support Development Studies (CECODES), operating under the Viet Nam Union of Science and Technology Associations (VUSTA), and the United Nations Development Programme (UNDP) in Viet Nam (see CECODES et al. 2024). The PAPI measures and benchmarks citizens' experiences and perception on policy-making, policy implementation, and the monitoring of public service delivery across all 63 provincial governments in Viet Nam to advocate for effective and responsive governance. The dimensions are specifically tailored to Viet Nam's national and local level contexts. PAPI is based on annual surveys of approximately 16,000 individuals. From 2009 to 2020, the PAPI index has effectively captured and represented the experiences of 146,233 citizens with diversified demographic backgrounds. The index's sampling approach adheres to international state-of-the-art methodological standards, namely probability proportional to size and random selection of respondents. The questionnaires are collected via face-to-face interviews lasting between 45 and 60 minutes.

Currently, the survey includes 550 substantive questions of about 120 indicators, divided into 28 sub-dimensions converging to eight PAPI dimensions:[3]

1. Participation at local levels
2. Transparency of local decision-making
3. Vertical accountability
4. Control of corruption in the public sector
5. Public administrative procedures
6. Public service delivery
7. Environmental governance (added since 2018)
8. E-governance (added since 2018)

In this study, only the mean of the six first sub-dimensions of PAPI was calculated for the appropriate comparison of the research time period.

4 Description of provincial inequality in Viet Nam

In this section, we compare absolute and relative inequality across the 63 provinces in Viet Nam, averaged across the decade 2011–20.

Let us first look at the overall mean values in Table 2. We find that the average Gini coefficient is at 0.38 with a minimum value of 0.31 and maximum value of 0.47 for all the provinces. This implies that, on the aggregate, relative income inequality in Viet Nam is at a medium level. On the other hand, mean absolute inequality is about 24,000 (in billion VND), with a large range from a minimum value of 2,473 to a maximum value of 27,3376. The mean group income ratio is at 7.3

[3] There have been some modifications to the PAPI sub-indices. From 2011 to 2017, only six PAPI sub-indices were recorded. Since 2018, there have been eight sub-indices of PAPI, with the addition of two dimensions.

with a minimum value of 5.3 and maximum value of 9.9 and the mean group income gap is 5,481 billion VND. Provincial income was approximately 63,000 billion VND on average.

Table 2: Average GRDP and inequality indicators (all years).

Variable	Observation	Mean	SD	Min	Max
Mean GRDP (billion VND)	63	62,753.301	11,5914.87	6,004.187	76,5760.75
Gini index	63	0.3810	0.032	0.319	0.47
Absolute Gini (billion VND)	63	23,624.046	43,336.823	2,473.725	27,3376.59
Group income ratio	63	7.366667	0.99174	5.3	9.9
Group income gap	63	5,481.619	1,585.948	3,241	10,322

Note: GRDP, gross regional domestic product; VND, Vietnamese Dong; SD, standard deviation.

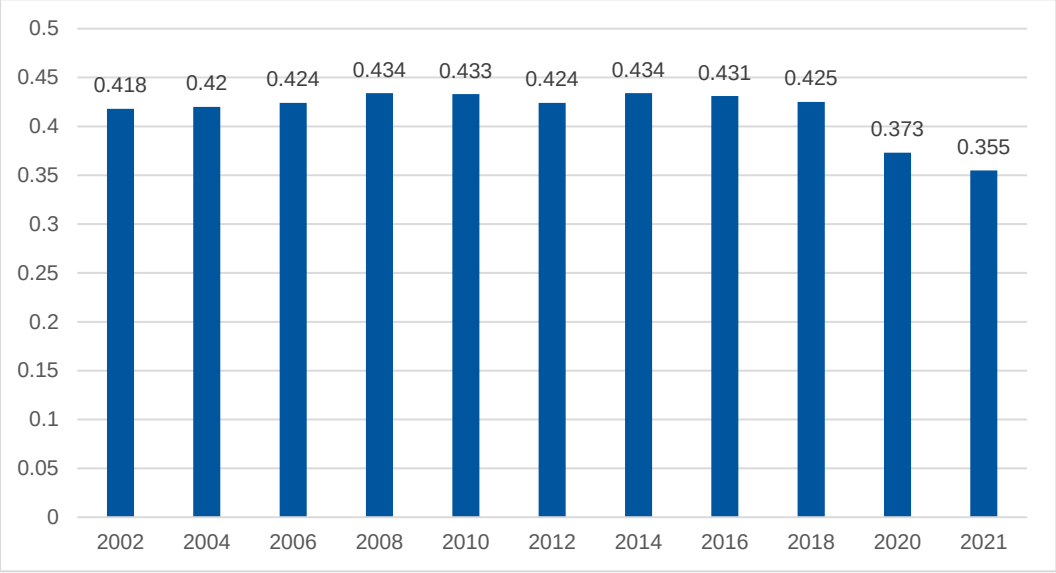
Source: authors' calculation.

As our absolute measure for provincial income inequality is measured in billion VND and not inflation-adjusted US dollars, we can only compare the aggregate relative inequality levels of Viet Nam with the relative inequality outcomes at a more global level.

The analysis of Figure 1 reveals that the relative Gini coefficient in the whole country was relatively stable around 0.375–0.431 during the research period, maintaining a moderate level. Niño-Zarazúa et al. (2017) indicate that the relative Gini coefficient for East Asia and the Pacific in 2005 and 2010 was around 0.5–0.6. This value range is higher than the average value that we obtained for Viet Nam, which is around 0.38. This may indicate that, relative to the East Asia and Pacific regions, Viet Nam has slightly lower relative income inequality estimates.

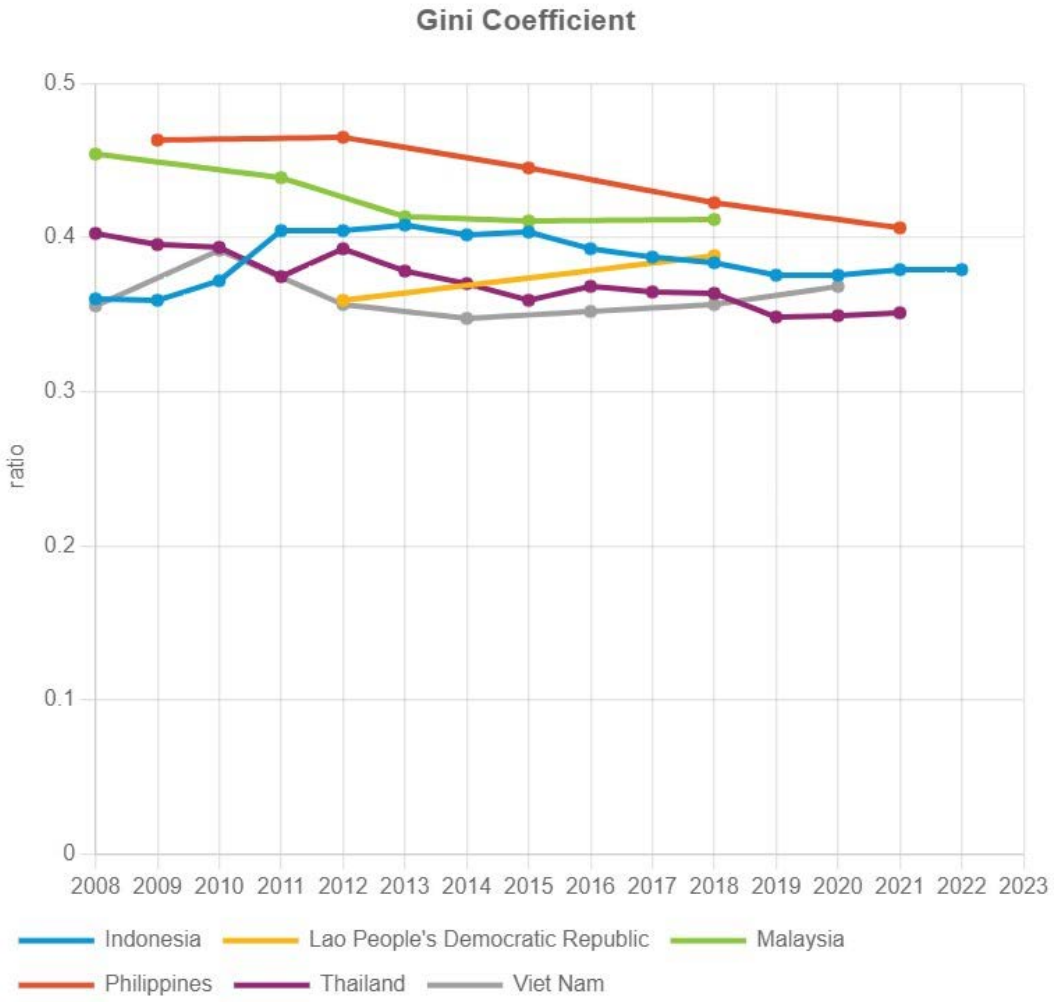
Again, our calculations are consistent with those recently found by the World Bank (2022), which indicate relative Gini coefficients ranging from 0.35 to 0.4 for 2010 to 2020. Using information from Chancel et al. (2022), these relative inequality values of 0.35–0.5 (with an average of 0.38) are similar to those observed in 2014 (i.e. the available data year that all countries have in common) for Thailand (0.37), Indonesia (0.4), and Malaysia (0.41). This is notable in that these geographic neighbours of Viet Nam have similar economic development levels, but vary in terms of governance structures (Figure 2).

Figure1: Relative Gini coefficient in Viet Nam for the period 2002–21



Source: authors' computation based on data from Viet Nam's General Statistics Office.

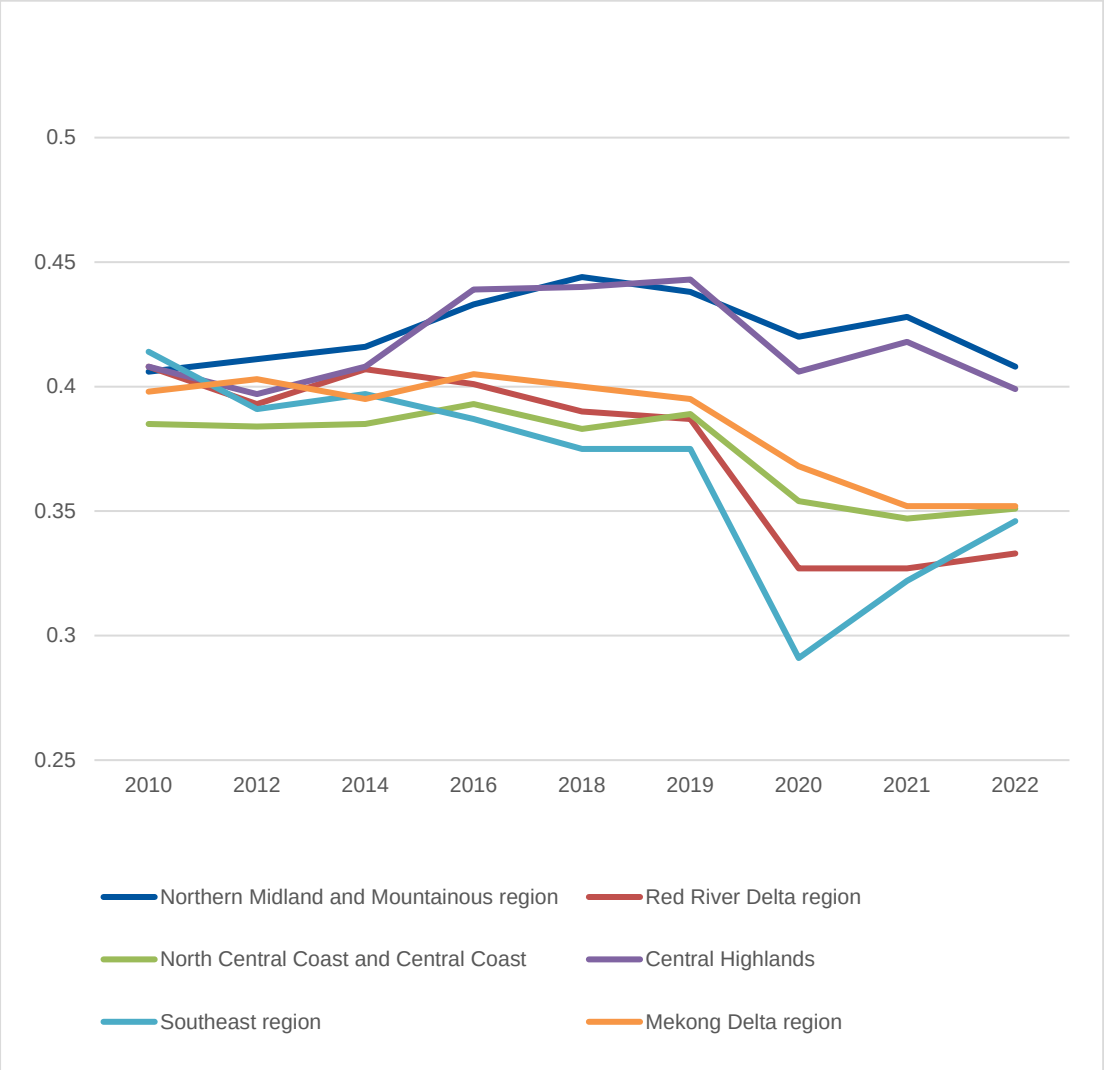
Figure 2: Relative Gini coefficient of Viet Nam and some other ASEAN countries for the period 2008–22



Source: authors' computation based on data from ADB key indicators (see ADB 2023).

Although the overall pattern of income inequality throughout the whole country seems to be rather consistent over the decade, it is probable that differences in inequality levels between mountainous/highland regions and delta areas have been more pronounced (Figure 3). Among the six socio-economic regions, the Northern Midland and Mountainous region as well as the Central Highlands are those with the highest relative inequality level. The Southeast region has demonstrated the most notable progress in the realm of inequality reduction. This region, along with the Red River Delta region, has the greatest level of equality.

Figure 3: Relative Inequality in the six socio-economic regions in Viet Nam



Source: authors' computation based on data from Viet Nam's General Statistics Office.

We now look at provincial outcomes in detail in Table 3.

Table 3 Inequality outcomes and income levels at the extremes

Relative inequality: Lowest			Absolute inequality: Lowest			Income levels: Lowest		
Binh Thuan	0.319	NCCCC	Bac Kan	2,474	NMM	Bac Kan	6,004	NMM
Thai Binh	0.324	RRD	Lai Chau	3,718	NMM	Lai Chau	8,647	NMM
Hung Yen	0.325	RRD	Dien Bien	4,127	NMM	Cao Bang	8,841	NMM
Hai Duong	0.327	RRD	Cao Bang	4,155	NMM	Dien Bien	9,509	NMM
Ha Nam	0.337	RRD	Kon Tum	4,386	CH	Kon Tum	11,159	CH
Relative inequality: Highest			Absolute inequality: Highest			Income levels: Highest		
Kien Giang	0.432	MD	Dong Nai	54,320	SE	Dong Nai	157,449	SE
Dien Bien	0.434	NMM	Binh Duong	71,417	SE	Binh Duong	189,939	SE
Ba Ria–Vung Tau	0.435	SE	Ba Ria–Vung Tau	104,000	SE	Ba Ria–Vung Tau	239,781	SE
Lao Cai	0.437	NMM	Ha Noi	210,000	RRD	Ha Noi	524,723	RRD
Cao Bang	0.470	NMM	Ho Chi Minh City	273,000	SE	Ho Chi Minh City	765,761	SE

Note: NCCCC, North Central Coast and Central Coast; NMM, Northern Midland and Mountainous region; RRD, Red River Delta; CH, Central Highlands; MD, Mekong Delta region; SE, Southeast region.

Source: authors' calculations based on data from Viet Nam's General Statistics Office.

A few observations can be noted here.

- While the regional capitals Ho Chi Minh City and Hanoi have the highest provincial income levels, they are also characterized by the highest absolute inequality outcomes. The income level as well as absolute inequality of these two cities are much higher than that of other provinces. GRDP of these two cities account for about 32% of the gross domestic product (GDP) of Viet Nam. Therefore, these two cities may influence the whole income distribution of Viet Nam.
- Results for Ba Ria–Vung Tau are more striking. It is the third richest province in Viet Nam, on average, but it is among the top three provinces for having the worst outcomes in terms of relative and absolute inequality. This might be explained by the fact that a majority of income of this province is from crude oil exploitation.
- The five provinces with the lowest absolute inequality are also those that are the poorest while the five provinces with highest absolute inequality are the richest. We observe that at the opposite, the third poorest province—Cao Bang—has the highest relative income inequality.
- While provinces like Bac Kan and Lai Chau have among the lowest absolute inequality levels, they are among the poorest. These provinces are located in mountainous and remote regions of Viet Nam, and therefore they face many geographical constrains for development. The data show clear regional dimensions in development outcomes for Viet Nam.
- When considering relative inequality, the delta regions are doing better than the mountainous areas. However, when absolute inequality is taken into account, the scenario is the opposite. The mountainous/highland areas have lower inequality than the Southeast and delta regions.
- The Southeast region has been doing the best in terms of relative inequality reduction, but remains among the highest absolute inequality areas.

In summary, we see in these descriptive statistics that increased incomes, even in regional centres, do not necessarily imply better inequality outcomes for Viet Nam's provinces.

5 PCA and K-means analysis

5.1 Use of PCA

As before, we begin with PCA of the six PAPI institutional measures to consider underlying patterns in institutional quality across Viet Nam's provinces.

We find in the descriptive statistics (Table 4) that the mean of these six variables ranges from 5 to 7. Their minimum and maximum values range from 4.6 to 7.6. On the aggregate, the 63 provinces in Viet Nam have a highest average value for the fifth dimension, 'public administrative procedures'. Mean values are lowest for the first dimension, 'Participation at local levels'.

Table 4: Summary of PAPI institutional measures

Variable	Observation	Mean	SD	Min	Max
1: Participation at local levels	63	5.17	0.339	4.622	6.028
2: Transparency of local decision-making	63	5.57	0.324	4.959	6.231
3: Vertical accountability	63	5.312	0.27	4.829	6.053
4: Control of corruption in the public sector	63	6.205	0.408	5.158	7.084
5: Public administrative procedures	63	7.09	0.149	6.792	7.437
6: Public service delivery	63	6.973	0.25	6.481	7.636

Source: authors' calculation.

PCA of the six PAPI measures identifies two new components (or 'indices'). These two components, Index 1 and Index 2, are uncorrelated with each other and explain 50% and 20% of the variance in our data, respectively, and can be used to represent the original six institutional variables. Index 1 explains most of the variance in the dataset, whereas Index 2 represents what Index 1 is not able to capture. Table 5 describes the eigenvectors or coefficients representing the impact or weight of each of the six PAPI variables on each component.

Table 5: PCA results for PAPI institutional measures (eigenvectors)

Variable	Index 1	Index 2
1: Participation at local levels	0.475	−0.368
2: Transparency of local decision-making	0.522	−0.169
3: Vertical accountability	0.509	−0.131
4: Control of corruption in the public sector	0.217	0.594
5: Public administrative procedures	0.417	0.255
6: Public service delivery	0.148	0.633

Source: authors' calculation.

In other words, the two indexes can be described as follows:

$$\text{Index } 1 = 0.47 * \text{PAPI\_1} + 0.52 * \text{PAPI\_2} + 0.50 * \text{PAPI\_3} + 0.21 * \text{PAPI\_4} + 0.41 * \text{PAPI\_5} + 0.14\text{\_PAPI} * 6$$

$$\text{Index } 2 = -0.37 * \text{PAPI\_1} - 0.17 * \text{PAPI\_2} - 0.13 * \text{PAPI\_3} + 0.59 * \text{PAPI\_4} + 0.25 * \text{PAPI\_5} + 0.63\text{\_PAPI} * 6$$

Index 1 is mostly affected by measures 1: Participation at local levels, 2: Transparency of local decision-making, 3: Vertical accountability, and 5: Public administrative procedures, whereas Index 2 is mostly affected by measures 4: Control of corruption in the public sector and 6: Public service delivery.

As Table 6 summarizes, on average, Index 1 and Index 2 both have mean values of zero. Index 1 ranges from −3.2 to 4.3, whereas Index 2 ranges from −2.0 to 2.8.

Table 6: Index 1 and 2 descriptive statistics

Variable	Observation	Mean	SD	Min	Max
Index 1	63	0	1.726	−3.244	4.303
Index 2	63	0	1.09	−2.017	2.805

Source: authors' calculation.

We now use these two new institutional variables (Index 1 and Index 2) to categorize the 63 provinces in Viet Nam into two groups using K-means as a clustering algorithm based on averages.

5.2 Use of K-means clustering analysis

We begin by calculating the distance between each data point (i.e. one observation per province) and a centroid, to assign it to a category or cluster. Each observation is assigned to the cluster where the distance is lowest. Using Index 1 and Index 2 as clustering variables, we form two clusters. Table 7 describes summary statistics of two indexes and six PAPI indicators of two clusters. Cluster 1 shows that the average Index 1 is 1.5, whereas for Cluster 2 it is −1.3. This implies that the two provincial groupings, Cluster 1 and Cluster 2 significantly differ in terms of Index 1. They do not, however, differ in terms of Index 2 as both clusters have an average value of zero. Therefore, K-means grouping were mainly determined by stark differences of the two provincial clusters for Index 1. Recall that Index 1 mostly describes variation in terms of *participation at local levels, transparency in local decision-making, vertical accountability,* and *public administrative procedures*, whereas Index 2 relates to *control of public sector corruption* and *public service delivery* (Table 5). Table 7 also shows that the difference in means of these four PAPI indicators of two clusters seems to be higher than that of the other two indicators, but is relatively small. In other words, the greatest institutional differences across these two clusters relate to (better versus worse) ratings in terms of participation, transparency in local decision-making, vertical accountability, and public administrative procedures, but not to ratings on control of public sector corruption and public service delivery. Cluster 2 has a negative average and worse overall ratings than Cluster 1.

Table 7: Average PAPI scores in Clusters 1 and 2

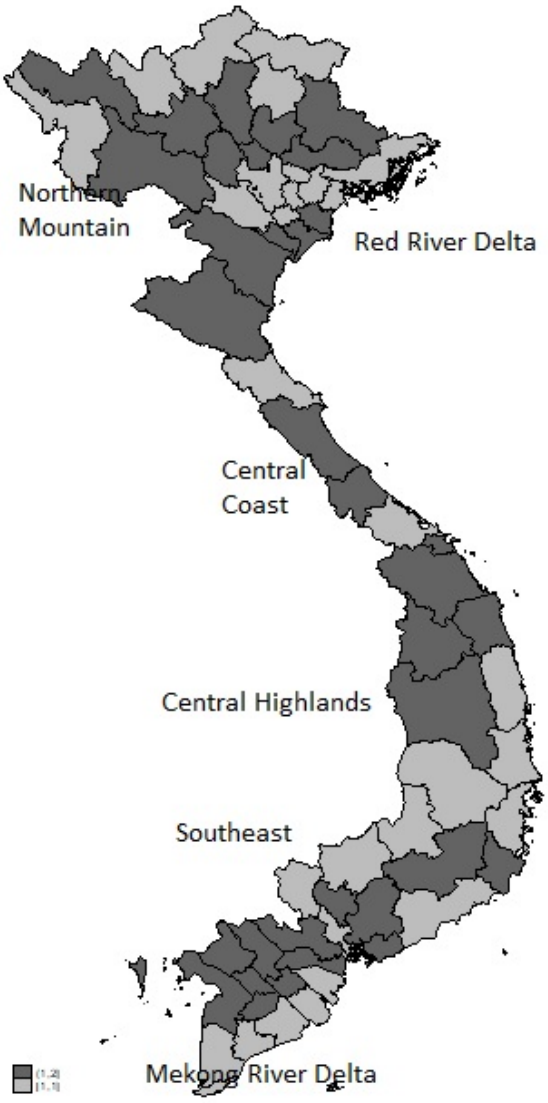
Variable	Observation	Mean	SD	Min	Max
Cluster 1					
Index 1	*29*	*1.526*	*1.075*	*0.25*	*4.303*
Index 2	*29*	*−0.008*	*1.039*	*−1.623*	*2.364*
1: Participation at local levels	29	5.43	0.227	5.091	6.028
2: Transparency of local decision-making	29	5.833	0.189	5.541	6.231
3: Vertical accountability	29	5.507	0.24	5.022	6.053
4: Control of corruption in the public sector	29	6.35	0.334	5.89	7.084
5: Public administrative procedures	29	7.183	0.119	7.005	7.437
6: Public service delivery	29	7.027	0.231	6.486	7.636
Cluster 2					
Index 1	*34*	*−1.302*	*0.908*	*−3.244*	*0.001*
Index 2	*34*	*0.007*	*1.147*	*−2.017*	*2.805*
1: Participation at local levels	34	4.948	0.25	4.622	5.443
2: Transparency of local decision-making	34	5.347	0.233	4.959	5.713
3: Vertical accountability	34	5.146	0.161	4.829	5.522
4: Control of corruption in the public sector	34	6.082	0.428	5.158	6.974
5: Public administrative procedures	34	7.01	0.124	6.792	7.277
6: Public service delivery	34	6.928	0.259	6.481	7.434

Source: authors' calculation.

Considering specific provinces (Table 8), we can see that the regional centres Hanoi and Ho Chi Minh City are in Cluster 2, which is the cluster with poorer institutions. This reflects that most PAPI indicators of these two centres have lower scores than the average of all provinces. This might be because these two centres are at higher development stages and people tend to have a higher expectation on different dimensions of governance. Because of a stark difference of these two cities compared with other provinces/cities of Viet Nam, their presence in Cluster 2 may significantly drive the results.

Table 8 Provinces by Clusters 1 and 2

Cluster 1	Cluster 2
Bac Kan	Ha Noi
Tuyen Quang	Ha Giang
Lao Cai	Cao Bang
Son La	Dien Bien
Hoa Binh	Lai Chau
Thai Nguyen	Yen Bai
Lang Son	Quang Ninh
Bac Giang	Hai Phong
Phu Tho	Hung Yen
Vinh Phuc	Thua Thien-Hue
Bac Ninh	Quang Nam
Hai Duong	Quang Ngai
Thai Binh	Phu Yen
Ha Nam	Khanh Hoa
Nam Dinh	Ninh Thuan
Ninh Binh	Binh Thuan
Thanh Hoa	Kon Tum
Nghe An	Gia Lai
Ha Tinh	Dak Lak
Quang Binh	Dak Nong
Quang Tri	Lam Dong
Da Nang	Tay Ninh
Binh Dinh	Binh Duong
Binh Phuoc	Dong Nai
Ba Ria–Vung Tau	Ho Chi Minh City
Long An	Tien Giang
Ben Tre	Tra Vinh
Dong Thap	Vinh Long
Can Tho	An Giang
	Kien Giang
	Hau Giang
	Soc Trang
	Bac Lieu
	Ca Mau



Note: the map on the right is the geographic location of the 63 provinces in the two clusters listed on the left. Provinces in Cluster 1 are coloured dark grey; provinces in Cluster 2 are light grey. This is for visual presentation of the two clusters for institutional performance of Viet Nam's provinces/cities only, not for the purpose of mapping the whole of Viet Nam.

Source: authors' compilation. This map was created using STATA14. GIS shapefiles were downloaded from the iGISmap website at: https://www.igismap.com/vietnam-shapefile-download-country-boundaryline-polygon/.

We now look at the summary statistics of inequality and income indicators of two clusters to see whether there is any correlation between institutional quality and income inequality. Table 9

presents the summary statistics of the two clusters on the different inequality indicators and income level with and without Hanoi and Ho Chi Minh City in Cluster 2. The means of GRDP in the table show that having Hanoi and Ho Chi Minh City in Cluster 2 makes a big difference. It can be seen that Cluster 1 (better institutional quality group) has a mean GRDP of 45,464 billion VND, which is much lower than the mean GRDP of Cluster 2 at 77,499 billion VND. However, when Hanoi and Ho Chi Minh City are taken out of Cluster 2, it shows that Cluster 2 has a mean GRDP of 42,015 billion VND, lower than that of Cluster 1.

Table 9: Descriptive statistics on inequality and income level for Cluster 1, Cluster 2, and Cluster 2 excluding Hanoi and Ho Chi Minh City

Variables	Observation	Mean	SD	Min	Max
Cluster 1					
Mean GRDP (billion VND)	29	45,464.835	42,244.601	6,004.187	239,780.7
Gini index	29	0.375	0.028	0.324	0.437
Absolute Gini (billion VND)	29	17,237.668	18,131.222	2,473.725	10,4304.61
Group income ratio	29	7.196552	0.783528	6	8.4
Group income gap	29	5,328.138	1,358.477	3,241	8,918
2010 Gini index	29	0.3782255	0.025628	0.32817	0.44078
2010 Absolute Gini (billion VND)	29	437.9172	126.1363	284.7	757.9
2010 Group income ratio	29	6.965517	0.7784031	6	9
2010 Group income gap (billion VND)	29	2,207.172	612.4075	1,478	3,732
Cluster 2					
Mean GRDP (billion VND)	34	77,499.345	15,2456.44	8,647.454	76,5760.75
Gini index	34	0.386	0.034	.319	0.47
Absolute Gini (billion VND)	34	2,9071.251	56,419.582	37,18.405	27,3376.59
Group income ratio	34	7.511765	1.131308	5.3	9.9
Group income gap	34	5,612.529	1,766.743	3,266	10,322
2010 Gini index	34	0.3959135	0.0359179	0.32668	0.46552
2010 Absolute Gini (billion VND)	34	504.9618	233.2925	195.1	1,299
2010 Group income ratio	34	7.205882	1.008431	5	9
2010 Group income gap (billion VND)	34	2,481.118	1,066.183	1,154	6,034
Cluster 2 excluding Hanoi and Ho Chi Minh City					
Mean GRDP (billion VND)	32	42,015.44	41,353.67	8,647.454	189,938.8
Gini index	32	0.386875	0.034422	0.319	0.47
Absolute Gini (billion VND)	32	15,786.15	15,044.64	3,718.405	71,416.97
Group income ratio	32	7.5125	1.133294	5.3	9.9
Group income gap	32	5,366.813	1,503.283	3,266	10,322
2010 Gini index	32	0.394294	0.03616	0.32668	0.46552
2010 Absolute Gini (billion VND)	32	471.4281	193.8163	195.1	1,299
2010 group income ratio	32	7.15625	0.987319	5	9
2010 group income gap (billion VND)	32	2,331.313	891.1325	1,154	6,034

Source: authors' calculation.

In terms of relative inequality, the two clusters have almost the same values of Gini index at about 0.38 whether or not they have Hanoi and Ho Chi Minh City in the sample. This is coincidentally the same as the overall average of the 63 provinces. However, in terms of group income ratio, Cluster 1 tends to have a slightly lower mean of group income ratio. It means that the better institutional group tends to have a lower income ratio between the highest income quintile and lowest one and vice versa. It seems that institutional quality might have a positive relationship with inequality between the high end and low end of the income distribution. In other words, better institutions might benefit the poor.

In terms of absolute inequality, by contrast, those in Cluster 1 have substantially lower mean absolute inequality levels than those in Cluster 2 (17,237 compared with 29,071). Given that absolute inequality is calculated using income, it is not surprising that there are also differences in mean GRDP across Clusters 1 and 2, with Cluster 2 being richer than Cluster 1. This is indeed shown when taking Hanoi and Ho Chi Minh City out of the sample: the mean GRDP of Cluster 2 is reduced by half and therefore the mean absolute Gini is also reduced (to 15,786), which is slightly lower than the mean absolute Gini of Cluster 1 (17,237). This suggests that better institutional quality might have a positive relationship with income level but a negative relationship with absolute inequality. However, this relationship is not clear when considering the group income gap; the mean of this variable is almost the same for both clusters in Table 9. This suggests that higher absolute inequality might be concurrent with higher income for the whole population but not necessarily for the highest and lowest income groups.

Table 9 also presents some statistics for initial level of inequality (both absolute and relative inequality indicators in 2010, the beginning of the study period) to show possible correlation with institutional quality. The results suggest consistently that the better institutional quality group (Cluster 1) has lower initial inequality across all indicators compared with Cluster 2, both with and without Hanoi and Ho Chi Minh City. This seems to show that initial inequality, whatever its measurements, has positive correlations with institutional quality in Viet Nam.

In sum, Table 9 shows that the results are quite sensitive to the presence of Hanoi and Ho Chi Minh City in the sample. We suggest that provinces in the institutionally 'better' cluster (Cluster 1), when compared with those in the 'worse' cluster (Cluster 2), are poorer, with similar levels of relative inequality but lower levels of absolute inequality. However, when Ho Chi Minh City and Hanoi are removed from the sample, the pattern of this relationship changes towards better institutions accompanying better income, but also higher absolute inequality for the whole distribution and lower inequality for the income ratio between the highest income and lowest income quintile. The latter results seem to be broadly consistent with the existing research finding that better institutional quality goes along with both higher income and lower inequality levels (Chong and Calderón 2000; Chong and Gradstein 2007). A question for future research concerns what causal mechanisms may underlie such empirical relationships. One hypothesis is that this relationship may be driven in part by the PAPI data being based on citizen perceptions of institutional quality; therefore, it can explain the difference of the results with and without Hanoi and Ho Chi Minh City. Research in other contexts shows that individuals' fairness perceptions, linked with inequality, may influence their satisfaction with democracy (Saxton 2021). Similarly, it may be that citizens in wealthier regions have higher expectations of government and thus are more likely to rate critically provincial institutional quality.

6 Conclusion

We have provided descriptive statistics of the average distribution of income in 63 provinces in Viet Nam for the 2010s decade. While Viet Nam has moderate levels of relative inequality, our observations imply that economic growth measured as having increased income levels and economic growth do not necessarily translate to lower absolute inequality in developing countries like Viet Nam. However, this does lower the income ratio between the highest income quintile and the lowest one. It means that institutional quality and income growth does benefit the poor. In addition, this paper shows that unsupervised learning techniques (e.g., cluster analyses and PCA) are helpful in producing quantitative categories of data in which there is no initial structure.

The results also show that better quality institutions tend to have a lower level of initial inequality. In other words, lower inequality might be supportive of the development of better institutions, particularly in terms of participation, transparency in local decision-making, vertical accountability, and public administrative produces. It is also shown that in general, increased incomes might accompany better institutional quality and inequality outcomes, although not in all cases.

Finally, this analysis implies that there exist many measures of inequality that researchers and policy-makers alike ought to consider: from the relative Gini coefficient to absolute measures, to those considering the share of income of the poorest of the poor. When combined in the present analyses, these different measures may provide further depth into our understanding of how income is distributed. In addition, the group income ratio and gap could also be used to provide further policy insights. It is necessary to remember, however, that each inequality measure and underlying data set has its limitations and strengths. Consequently, they should be interpreted based on this. The complementarities in the information these measures provide will aid in finding a clearer picture of the implications of inequality, especially for an emerging economy like Viet Nam.

Given limits of the unsupervised learning techniques applied in this paper, further research could go deeper into the causal relationship of inequality and institutional quality and the underlining mechanism with different dimensions of governance. This type of analysis would need additional data but can bring further evidence and insights on the relations between inequality and institution.

References

Acemoglu, D., and J.A. Robinson (2006). *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511510809

ADB (2023). ADB Key Indicators Database. Available at: https://kidb.adb.org/explore?filter[year]=2009%2C2010%2C2011%2C2012%2C2013%2C2014%2C2015%2C2016%2C2017%2C2018%2C2019%2C2020%2C2021%2C2022%2C2023%2C2024%2C2008&filter[indicator_id]=2100020&filter[economy_code]=CAM%2CINO%2CLAO%2CMAL%2CPHI%2CTHA%2CVIE&showR (accessed May 2024).

Ansell, B.W., and D.S. Samuels (2010). 'Inequality and Democratization: A Contractarian Approach'. *Comparative Political Studies*, 43(12): 1543–74. https://doi.org/10.1177/0010414010376915

Athey, S. (2019). 'The Impact of Machine Learning on Economics'. In A. Ajay, G. Joshua, and G. Avi (eds), *The Economics of Artificial Intelligence* (pp. 507–52). Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0021***

Athey, S., and G.W. Imbens (2019). 'Machine Learning Methods That Economists Should Know About'. *Annual Review of Economics*, 11(1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bandyopadhyay, S. (2018). 'The Absolute Gini Is a More Reliable Measure of Inequality for Time Dependent Analyses (Compared with the Relative Gini)' .*Economics Letters Journal*, 162: 135–39. https://doi.org/10.1016/j.econlet.2017.07.012

Benjamin, D., L. Brandt, and B. McCaig (2017). 'Growth with Equity: Income Inequality in Vietnam, 2002–14'. *The Journal of Economic Inequality*, 15(1): 25–46. https://doi.org/10.1007/s10888-016-9341-7

Boix, C. (2003). *Democracy and Redistribution*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804960

CECODES, VUSTA, and UNDP (2024). PAPI Data. Available at: https://papi.org.vn/eng/papi-data/ (accessed May 2024).

Chancel, L., T. Piketty, E. Saez, G. Zucman et al. (2022). *World Inequality Report 2022*. Paris: World Inequality Lab. Available at: https://wir2022.wid.world/ (accessed May 2024).

Chong, A., and C. Calderón (2000). 'Institutional Quality and Income Distribution'. *Economic Development and Cultural Change*, 48(4): 761–86. https://doi.org/10.1086/452476

Chong, A., and M. Gradstein (2007). 'Inequality and Institutions'. *The Review of Economics and Statistics*, 89(3): 454–65. https://doi.org/10.1162/rest.89.3.454

Coppedge, M., A. Alvarez, and C. Maldonado (2008). 'Two Persistent Dimensions of Democracy: Contestation and Inclusiveness'. *The Journal of Politics*, 70(3): 632–47. https://doi.org/10.1017/S0022381608080663

Crone, T.M. (2005). 'An Alternative Definition of Economic Regions in the United States Based on Similarities in State Business Cycles'. *The Review of Economics and Statistics*, 87(4): 617–26. https://doi.org/10.1162/003465305775098224

Dang, T.T.H. (2018). 'Does Horizontal Inequality Matter in Vietnam?'. *Social Indicators Research*, 145: 943–56. https://doi.org/10.1007/s11205-018-1896-1

Ferreira, I.A., R.M. Gisselquist, and F. Tarp (2022). 'On the Impact of Inequality on Growth, Human Development, and Governance'. *International Studies Review*, 24(1). https://doi.org/10.1093/isr/viab058

Fonseca, J.R.S. (2013). 'Clustering in the Field of Social Sciences: That Is Your Choice'. *International Journal of Social Research Methodology: Theory & Practice*, 16(5): 403–28. https://doi.org/10.1080/13645579.2012.716973

Ghorbani, M., and E.K.P. Chong (2020). 'Stock Price Prediction Using Principal Components'. *PLOS One*, 15(3): Article e0230124. https://doi.org/10.1371/journal.pone.0230124

Haggard, S., and R.R. Kaufman (2012). 'Inequality and Regime Change: Democratic Transitions and the Stability of Democratic Rule'. *American Political Science Review*, 106(3): 495–516. https://doi.org/10.1017/S0003055412000287

Jolliffe, I.T., and J. Cadima (2016). 'Principal Component Analysis: A Review and Recent Developments. *Philos Trans A Math Phys Eng Sci*, 374(2065): Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). 'Human Decisions and Machine Predictions'. *The Quarterly Journal of Economics*, 133(1): 237–93. https://doi.org/10.1093/qje/qjx032

Kotschy, R., and U. Sunde (2017). 'Democracy, Inequality, and Institutional Quality'. *European Economic Review*, 91: 209–28. https://doi.org/10.1016/j.euroecorev.2016.10.006

Lanjouw, P., M. Marra, and C. Nguyen (2016). 'Vietnam's Evolving Poverty Index Map: Patterns and Implications for Policy'. *Social Indicators Research*, 1–26. https://doi.org/10.1007/s11205-016-1355-9

Magyar, Z.B. (2022). 'What Makes Party Systems Different? A Principal Component Analysis of 17 Advanced Democracies, 1970–2013'. *Political Analysis*, 30(2): 250–68. https://doi.org/10.1017/pan.2021.21

Mullainathan, S., and J. Spiess (2017). 'Machine Learning: An Applied Econometric Approach'. *The Journal of Economic Perspectives*, 31(2): 87–106. https://doi.org/10.1257/jep.31.2.87

Nguyen, C.V., and N.M. Pham (2018). 'Economic Growth, Inequality, and Poverty in Vietnam'. *Asian-Pacific Economic Literature*, 32(1): 45–58. https://doi.org/10.1111/apel.12219

Niño-Zarazúa, M., L. Roope, and F. Tarp (2017). 'Global Inequality: Relatively Lower, Absolutely Higher'. *Review of Income and Wealth*, 63(4): 661–84. https://doi.org/10.1111/roiw.12240

Savoia, A., J. Easaw, and A. McKay (2010). 'Inequality, Democracy, and Institutions: A Critical Review of Recent Research'. *World Development*, 38(2): 142–54. https://doi.org/10.1016/j.worlddev.2009.10.009

Saxton, G.W. (2021). 'Governance Quality, Fairness Perceptions, and Satisfaction with Democracy in Latin America'. *Latin American Politics and Society*, 63(2): 122–45. https://doi.org/10.1017/lap.2021.8

World Bank (2014). *Taking Stock: An Update on Vietnam's Recent Economic Developments (July 2014)—Key Findings*. Available at: https://www.worldbank.org/en/news/feature/2014/07/08/key-findings-of-taking-stock-an-update-on-vietnams-recent-economic-developments-july-2014 (accessed May 2024).

World Bank (2022). *From the Last Mile to the Next Mile: 2022 Vietnam Poverty and Equity Assessment*. Washington, DC: World Bank. Available at: https://www.worldbank.org/en/country/vietnam/publication/2022-vietnam-poverty-and-equity-assessment-report (accessed May 2024).

Zhang, X., and H. Tong (2022). 'Asymptotic Theory of Principal Component Analysis for Time Series Data with Cautionary Comments'. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(2): 543–65. https://doi.org/10.1111/rssa.12793