

Linguistic clustering and aggregate productive efficiency in Indonesia

Alexandre Repkine

*Department of Economics, Konkuk University, Gwangjin-Gu,
Seoul, Republic of Korea*

Abstract

Purpose – The purpose of this study is to explore the link between aggregate production efficiency and the extent of linguistic clustering in Indonesia.

Design/methodology/approach – The author draws on the stochastic frontier model and applies it to the data on Indonesian provinces to compute the effects of various determinants on these provinces' aggregate production efficiency. The key determinant is the spatial index of linguistic clustering that the author believes has never been applied before in this context.

Findings – Linguistic clustering is an important determinant of aggregate production efficiency. Linguistic diversity is positively associated with productive efficiency if members of a specific linguistic group are not clustered beyond a certain level.

Originality/value – To the best of the author's knowledge, this is the first study that links the spatial index of linguistic clustering (because of Massey and Danton) to production efficiency. In other words, the contribution of this study is to introduce a geographical dimension to the mainstream analysis of the association between ethnic diversity and economic performance.

Keywords Linguistic clustering, Linguistic fragmentation, Aggregate production efficiency, Conflict, Social cohesion

Paper type Research paper

1. Introduction

Linguistic differences affect economic outcomes. [Greenberg \(1956\)](#) introduced a measure of linguistic diversity as a function of shares of the country's population who speak different languages, suggesting that highly fragmented societies are likely to be fraught with conflicts inhibiting these societies' ability to use scarce resources efficiently. A seminal contribution of [Easterly and Levine \(1997\)](#) confirms this view. [Alesina and La Ferrara \(2005\)](#) provide a comprehensive overview of research regarding the relationship between linguistic fragmentation and economic performance and conclude that the sign of the association is mostly negative. A recent literature survey by [Ginsburgh and Weber \(2020\)](#) concludes that "linguistic fragmentation has an adverse impact on economic development and growth."

© Alexandre Repkine. Published in *Applied Economic Analysis*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence maybe seen at <http://creativecommons.org/licenses/by/4.0/legalcode>

The author thank the three anonymous referees for their helpful comments that have helped to substantially improve the quality of this manuscript.



One of the major reasons why fragmented societies are performing worse compared to their more homogeneous counterparts appears to be that fragmentation creates some sort of conflict. Thus, [Easterly and Levine \(1997\)](#) conclude that ethno-linguistic interest groups engage in rent-seeking behavior that hampers the provision of public goods. [Fishman \(1968\)](#) praises the ability of more homogeneous societies to keep “primordial passions” under control. [Nettle \(2000\)](#) concludes that “linguistic fragmentation leads to social division, conflict, factionalism, and corruption.” [Desmet et al. \(2012\)](#) find a strong link between linguistic fragmentation and civil conflict.

Several studies, however, have indicated that social fragmentation *per se* is not necessarily causing social conflict. Thus, [Horowitz \(1985\)](#) argues that conflicts are less likely to arise in either highly homogeneous or highly heterogeneous societies. [Esteban and Ray \(1994\)](#) show that it is polarization, not diversity, that is generating social tensions. [Montalvo and Reynal-Querol \(2005\)](#) arrive at a similar conclusion.

Both diversity and polarization measures discussed above are invariant with respect to the spatial distribution of the various groups’ members. Yet, geographical distribution of the various social groups is likely to be important for economic performance, as a number of studies suggest that these patterns affect the extent of social cohesion. Thus, [Cutler and Glaeser \(1997\)](#) find a negative effect of racial segregation on the economic performance of the black US citizens blaming segregation for a lack of social cohesion. [Rothwell \(2012\)](#) argues that the extent of different ethnicities’ spatial segregation is important to the formation of social cohesion. [Sturgis et al. \(2014\)](#) conclude that social cohesion of London communities is affected by these communities’ spatial segregation.

The theoretical framework of [Esteban et al. \(2012\)](#) links conflict intensity to the social groups’ diversity, polarization and cohesion. Based on this theory, we contribute to the empirical literature on the association between linguistic heterogeneity and economic performance by using the extent of spatial linguistic segregation as a proxy for social cohesion and explicitly incorporating this measure into an empirical model linking aggregate productive efficiency to linguistic fragmentation and social cohesion. Our contribution is, thus, not to modify or extend the theory of [Esteban et al. \(2012\)](#) but to use it as a foundation of our empirical framework.

We use a rich data set on 26 Indonesian provinces for the period between 1983 and 2007. We chose Indonesia, as it is one of the world’s most linguistically diverse economies with our database containing detailed information on 399 living languages currently spoken there. The period between 1983 and 2007 because the data on the provincial-level capital stock was available for this period.

We propose to measure linguistic segregation by an index of *absolute clustering* developed by [Massey and Denton \(1988\)](#) that increases in the extent to which speakers of the same language tend to reside in close geographical proximity to each other. To compute this index, we use a unique [World Language Mapping System \(2004\)](#).

We find that linguistic clustering is an important determinant of the aggregate productive efficiency. We formulate and test three hypotheses related to the link between linguistic clustering and productive efficiency. We find that the two are positively associated with each other and that linguistic diversity is positively associated with the productive efficiency unless the competing linguistic groups are clustered beyond a certain threshold level.

This paper is organized as follows. We review existing literature on the estimation of technical efficiency levels and their determinants in Section 2. In Section 3, we derive a theoretical basis for our empirical work based on the model of [Esteban et al. \(2012\)](#). Section 4 discusses the index of absolute linguistic clustering. In Section 5, we describe our data sources. In Section 6, we present our empirical results. Section 7 discusses, and in Section 8, we provide general summary and conclusions.

2. The estimation of technical efficiency levels and their determinants

The stochastic frontier analysis framework (SFA) we use in this study was introduced more than 40 years ago independently by [Aigner *et al.* \(1977\)](#) and [Meeusen and van den Broeck \(1977\)](#), spawning thousands of both applied and theoretical contributions. The main feature of the SFA paradigm is a composite error term in the formulation of the stochastic production frontier of the form $\varepsilon_{it} = v_{it} - u_{it}$ that represents deviations of the actually observed output level y_{it} from the parametric frontier $f(x_{it})$, where x_{it} is a vector of inputs for a decision-making unit (DMU) i at time t . The DMU's inefficient use of productive resources is captured by an asymmetric component $u_{it} \geq 0$ that is often assumed to be distributed half-normally or exponentially, although gamma and other one-sided distributions have been suggested in the literature (a comprehensive review in [Greene, 2008](#)).

While the contributions of [Aigner *et al.* \(1977\)](#) and [Meeusen and van den Broeck \(1977\)](#) allowed for the estimation of average levels of technical (in)efficiency in a sample of DMUs, the problem of evaluating the *individual* levels of technical inefficiency remained unsolved until [Jondrow *et al.* \(1982\)](#) suggested estimating the individual inefficiency components u_{it} as their expectations conditional on the estimated values of the composite error term $\hat{\varepsilon}_{it}$, i.e. $\hat{u}_{it} = E(u_{it}|\hat{\varepsilon}_{it})$.

The SFA was not the only paradigm within which the attempts were made to estimate producers' technical efficiency levels. Non-parametric techniques combining the data envelopment analysis approach developed by [Charnes *et al.* \(1978\)](#) with the idea of a radial distance function by [Farrell \(1957\)](#) became another mainstream approach to the measurement of productive efficiency. However, we are focusing on the discussion of the SFA paradigm, as this is the approach assumed in this study.

In more than 40 years that have passed since the introduction of the SFA, numerous extensions have been suggested. For instance, a panel data approach to the analysis of productive efficiency was initiated by the contributions of [Pitt and Lee \(1981\)](#) and [Schmidt and Sickles \(1984\)](#). Time dependency of the inefficiency levels was analyzed by [Amsler *et al.* \(2014\)](#) with the use of statistical copulas. [Kumbhakar *et al.* \(2014\)](#) in a panel data setting models both time-invariant and time-varying inefficiency levels. A comprehensive discussion of the various approaches to the estimation of technical inefficiency levels within the SFA approach can be found in, for example, [Sickles and Zelenyuk \(2019\)](#).

Identifying a set of determinants that affect levels of technical efficiency is practically important for policymaking. [Kumbhakar *et al.* \(1991\)](#) and [Battese and Coelli \(1995\)](#) are the two earlier studies that model the influence of exogenous determinants on the level of technical inefficiency by allowing the inefficiency components $u_{it} \sim f(\mu_{it})$ with the DMU-specific mean μ_{it} of some asymmetric distribution pdf $f(\bullet)$ be a function $\mu_{it} = z'_{it}\gamma$ of a vector of inefficiency determinants z_{it} . Evaluating the possible effects produced by the inefficiency determinants is not only important from the point of view of policymaking, as ignoring it may lead to the biased and inconsistent estimates of the production function parameters, as argued in [Wang and Schmidt \(2002\)](#).

While it appears tempting to estimate the technical inefficiency components u_{it} separately first and regress them on a vector of possible determinants in the second stage, this two-stage approach is seriously flawed, as argued in [Battese and Coelli \(1995\)](#) and [Wang and Schmidt \(2002\)](#). As a result, it appears to have become common practice to estimate technical inefficiency levels simultaneously with the effects of the inefficiency determinants, which is the approach we pursue in this study by maximizing a likelihood function in one step.

As correctly noticed by [Lovell \(1993\)](#), "[...] unfortunately, economic theory does not supply a theoretical model of the determinants of efficiency." Given the data availability

constraints and the existing empirical literature, we focused on several channels through which macroeconomic variables may affect aggregate productive efficiency, that is, labor market rigidities, human capital, the extent of urbanization, public spending and trade openness.

Gonzalez and Miles-Touya (2012) analyze the Spanish labor market and conclude that labor market rigidities are positively related to economic inefficiency. Beeson and Husted (1989) in the US context find that differences in the labor market characteristics affect technical efficiency levels. Some studies suggest that the generally available data on the labor market can be used as proxies for the extent of the labor market rigidities. Micco and Pages (2006) analyze the data on more than 60 countries and find that labor market rigidities result in decreased job turnover and employment rates.

There is plenty of empirical evidence that increased levels of human capital stock increase the extent of productive efficiency. Thus, Lich *et al.* (2022) conduct a meta-analysis of 268 food-crop farming studies, suggesting a positive link between human capital and the farms' technical efficiency levels. Lee *et al.* (2002) find that promoting the creation and fostering of human capital is associated with higher levels of aggregate technical efficiency. Repkine (2014) analyzed the performance of African economies in terms of their aggregate production efficiency levels and found that they are affected by the level of human capital. Gheit (2022) finds a positive link between manufacturing productivity and the percentage of hours worked by the college educated workers.

A number of studies argue in favor of a positive link between the rates of urbanization and aggregate productive efficiency. In the Chinese context, Zhao *et al.* (2022) demonstrate that urbanization rates are positively linked to the levels of technical efficiency in the Chinese agriculture. Interesting insights are provided by Bannister and Stolp (1995), who find a positive association between the extent of industrialization and productive efficiency in the Mexican states. Bertinelli and Zou (2008) examine a cross-section of countries and find empirical support for a positive link between more urbanization and higher rates of human capital accumulation that, as argued above, are likely to lead to increases in productive efficiency.

Public spending is also likely to be related to the level of firms' technical efficiency, as it affects the overall institutional environment. Thus, Auci *et al.* (2021) find evidence of a strong association between the size and the type of public expenditure and the aggregate technical efficiency levels of the European economies. In a study of Mexican municipalities, Becerra-Ornelas and Nunez (2019) examine the effect of public spending on technical efficiency of the local production and find a negative association between the two irrespectively of the expenditure type.

Engagement in the inter-regional trade is likely to increase the extent of competitive pressure and foster the exchange of ideas, positively affecting the regional productive efficiency. Hart *et al.* (2015) analyze economic performance of 16 EU member states and provide evidence of a positive long-run association between trade openness and technical efficiency. Chortareas *et al.* (2003) find a positive impact of trade openness on the aggregate technical efficiency in the OECD economies. Tsekeris and Papaioannou (2017) find a positive effect of the interregional trade openness on the regional technical efficiency levels in Greece with similar conclusions reached in the case of Turkey by Demir and Mahmud (1998).

While the importance of analyzing the effects of ethnic diversity and polarization on economic performance has long been recognized in the literature (Ginsburgh and Weber, 2020), explicitly including the measures of ethnic heterogeneity into the list of determinants of productive efficiency is less widespread. In particular, the spatial dimension of ethnic and language heterogeneity appears to have been overlooked in the existing literature. The main

contribution of this study is to emphasize the importance of spatial clustering of the speakers of different languages to the study of the determinants of productive efficiency.

3. Theoretical background

We draw on the theory of [Esteban et al. \(2012\)](#) that links the intensity of conflict to social diversity, polarization and cohesion. Consider a society of N individuals that belong to m linguistic groups with N_i , $i = 1, m$ being the number of individuals in each group and $n_i \equiv \frac{N_i}{N}$ the individual group shares. Linguistic cleavages between these groups result in conflicts. The winning group in such a conflict obtains a prize that has a private component with per capita value μ and a public component with per capita population-normalized value π .

Conflict entails using group-specific resources R_i , $i = 1, m$ with $R_N \equiv \sum_{i=1}^m R_i$ being the total

amount of such resources used to maintain the inter-group conflict. The per capita amount of resources spent in conflict is given by $\rho \equiv \frac{R_N}{N}$. The per capita income equivalent of resources spent on conflict is given by a function $c(\rho)$ that is assumed to be increasing and strictly convex, that is, $c'(\rho) > 0$ and $c''(\rho) > 0$, assuming $c(\rho)$ to be twice differentiable. The income cost of entering a conflict is made to be zero, that is, $c'(0) = 0$.

Finally, every individual is assumed to have a utility function that places weight 1 on the individual payoffs, while $\alpha > 0$ is the weight placed on the aggregate of all payoffs of the other members of his or her group. In other words, every individual, say j , has an extended utility function W_j of the form $W_j = U_j + \alpha U_{-j}$ where U_j is the personal payoff of individual j , while U_{-j} is the aggregate of the payoffs of all other individuals in the group ([Sen, 1966](#)). [Esteban et al. \(2012\)](#) refer to α as a measure of *social cohesion*, as the higher levels of α reflect the extent to which the collective benefit relates to the individual benefit. The value of $\alpha = 0$ represents an exclusively self-centered behavior, while $\alpha > 1$ corresponds to altruism.

From an individual perspective, the income equivalent of resources ρ spent to participate in conflict is equal to the product of ρ and its shadow price $c'(\rho)$. Intensity of conflict is then defined to be $\sigma \equiv \frac{c'(\rho)\rho}{\pi + \mu}$. We consider the case of a prize contested in the inter-group conflicts being entirely public, that is, $\mu = 0$ and $\lambda \equiv \frac{\pi}{\pi + \mu} = 1$. It follows then from *P1* in [Esteban, Mayoral and Ray \(2012\)](#) that the intensity of conflict σ is given by:

$$\sigma = \alpha P + \frac{1 - \alpha}{N} G \quad (1)$$

where P is the linguistic polarization index, G is an index of linguistic diversity and α as before is the weight accruing to the aggregate of all payoffs of the other members of the social group in an extended utility function along the lines of [Sen \(1966\)](#). The linguistic

diversity index G is computed as is [Greenberg's \(1956\)](#) A-index, that is, $G = \sum_{i=1}^m n_i(1 - n_i)$,

while linguistic polarization index is defined according to [Montalvo and Reynal-Querol \(2005\)](#) as $P = \sum_{i=1}^m n_i^2(1 - n_i)$. While the difference between G and P may seem unimportant,

[Montalvo and Reynal-Querol \(2005\)](#) demonstrate that it is social polarization rather than diversity *per se* that is a more powerful determinant of conflict, with [Esteban and Ray \(2011\)](#) providing a theoretical basis for that claim.

It is worthwhile noticing that linguistic polarization P enters the list of the determinants of conflict intensity σ in [equation \(1\)](#) only as a multiple of social cohesion α . Another interesting implication of [equation \(1\)](#) is that, while more linguistic polarization is conducive to a higher conflict intensity as long as some social cohesion is present, that is, if $\alpha \neq 0$, more linguistic diversity is not necessarily so as the marginal effect of G on σ becomes negative in those societies where the conflicting groups are altruistic, that is, where $\alpha > 1$.

We additionally assume that intensity of the inter-group conflict σ is positively linked to productive efficiency via a monotonically increasing function H , that is:

$$E = H[\sigma] = H \left(\alpha P + \frac{1 - \alpha}{N} G \right) \quad (2)$$

where E is the level of aggregate productive efficiency and $\frac{dE}{d\sigma} > 0$. As the conflicting groups in [Esteban et al. \(2012\)](#) have to procure resources R_i , $i = 1, m$ that they use to engage in conflict, this need creates incentives for the conflicting groups to produce more efficiently as the supply of R_i would be limited in case productive resources were wasted. Some studies such as [Caraballo and Buitrago \(2019\)](#) pursue this line of reasoning to argue that the inter-group conflicts result in increased productivity.

In this study, we endogenize the extent of social cohesion α among the speakers of a particular language by positing that it is an increasing function of the extent to which these speakers are geographically clustered, that is, $\alpha = \alpha(ALC)$, $\frac{\partial \alpha}{\partial ALC} > 0$ where ALC is an *index of absolute linguistic clustering* developed by [Massey and Denton \(1988\)](#) and discussed in the next section. As discussed in the Introduction, a number of studies argue that the extent of geographical segregation along racial and ethnic cleavages is affecting the extent of social cohesion. As ethnicity and language are closely related, we believe geographical segregation, or clustering, along the linguistic lines is likely to affect social cohesion as well. Second, given an interpretation of social cohesion given by [Esteban et al. \(2012\)](#) as the extent to which “within-group monitoring [...] manages to overcome the free-rider problem,” we find it natural to assume that such monitoring will be easier in case speakers of the same language reside in compact geographical clusters.

[Equation \(2\)](#) bears several testable hypotheses:

H1. The marginal effect of linguistic clustering on productive efficiency is positive.

Indeed, it follows from [equation \(2\)](#) that $\frac{dE}{dALC} = \frac{dE}{d\sigma} \frac{\partial \sigma}{\partial \alpha} \frac{d\alpha}{dALC}$. The effect of social cohesion on conflict intensity is given by $\frac{\partial \sigma}{\partial \alpha} = P - \frac{G}{N}$. As $G \in [0, 1]$ and N is on the order of 240 million, the term $\frac{G}{N} \approx 0$. It follows that $\frac{dE}{dALC} = \frac{dE}{d\sigma} \left(P - \frac{G}{N} \right) \frac{d\alpha}{dALC} \approx \frac{dE}{d\sigma} P \frac{d\alpha}{dALC} > 0$, as $\frac{d\alpha}{dALC} > 0$ and $\frac{dE}{d\sigma} > 0$.

H2. The association between linguistic diversity G and productive efficiency is positive unless the extent of linguistic clustering exceeds a certain threshold in which case the sign of the effect becomes negative.

The marginal effect of linguistic diversity G on productive efficiency is given by $\frac{dE}{dG} = \frac{dE}{d\sigma} \frac{\partial \sigma}{\partial G} = \frac{dE}{d\sigma} \frac{1 - \alpha}{N}$. As by assumption $\frac{dE}{d\sigma} > 0$, it follows that $\frac{dE}{dG} > 0$ if $\alpha < 1$, that is, if the linguistic groups are not altruistic, to borrow the terminology of [Esteban et al. \(2012\)](#). In case, however, the extent of within-group cohesion $\alpha > 1$, linguistically more diverse societies will be using their productive resources less efficiently so that $\frac{dE}{dG} < 0$ if $\alpha > 1$. As

by assumption $\frac{d\alpha}{dALC} > 0$, $\frac{dE}{dG} > 0$ if $ALC < ALC^*$, and $\frac{dE}{dG} < 0$ if $ALC > ALC^*$ where ALC^* is some threshold level of absolute linguistic clustering:

- H3. $\frac{\partial E}{\partial (ALC \times P)} > 0$ or linguistic polarization is positively associated with productive efficiency with the magnitude of the effect being stronger in linguistically more clustered cases.

We test these three hypotheses in Section 4 after we define and discuss the index of absolute linguistic clustering (ALC) and describe and summarize the data in the next two sections.

4. The index of absolute linguistic clustering

Based on the original analysis of the “checkerboard problem” by [Geary \(1954\)](#), [Dacey \(1968\)](#) and [White \(1983\)](#), [Massey and Denton \(1988\)](#) suggest an index that measures the extent to which members of the same group live in contiguous areas. We compute [Massey and Denton’s \(1988\)](#) formula of the absolute clustering index to the Indonesian provinces and refer to it as an ALC index. Higher values of this index correspond to a higher tendency of the speakers of a specific language to live in enclaves, with the index itself varying between zero and unity.

Suppose that the area of interest is divided into K equally sized geographical parcels. Denote N the total number of speakers of, for example, Javanese in this area. Denote by x_i , $i = 1, \dots, K$ the number of Javanese speakers in each of the K parcels with the total number of Javanese

speakers being $X = \sum_{i=1}^K x_i$. Let the total population of parcel $i = 1, \dots, K$ be equal to t_i . Finally, let

c_{ij} , $i, j = 1, \dots, K$ be a decreasing function of the geographical distance between parcels i and j . The index of absolute clustering for the Javanese speakers in the province of interest is then defined as follows:

$$ALC = \frac{\sum_{i=1}^K \frac{x_i}{X} \sum_{j=1}^K c_{ij} x_j - \frac{X}{K^2} \sum_{i=1}^K \sum_{j=1}^K c_{ij}}{\sum_{i=1}^K \frac{x_i}{X} \sum_{j=1}^K c_{ij} t_j - \frac{X}{N^2} \sum_{i=1}^K \sum_{j=1}^K c_{ij}} \quad (3)$$

We chose the number of geographical parcels K to be such that the area of each such parcel be equal to 1 km^2 . The linguistic dataset for the computation of ALC is obtained from the [World Language Mapping System \(2004\)](#). We compute the ALC indices for each language spoken in the 26 Indonesian provinces. The province-specific indices of absolute linguistic clustering are computed as a weighted average of the language-specific ALC indices using the speaker shares as relative weights. Following [Massey and Denton \(1988\)](#), we define $c_{ij} = e^{-\beta d_{ij}}$ where d_{ij} is the geographical distance between parcels i and j , and β is a decay parameter that we set to $\beta = 0.1$.

[Table 1](#) provides summary statistics for the indices of absolute linguistic clustering, diversity and polarization we computed. Linguistic diversity index G is computed as the A index from [Greenberg \(1956\)](#). Linguistic polarization index P is calculated according to [Montalvo and Reynal-Querol \(2005\)](#).

Despite the amazing amount of linguistic diversity in Indonesia, the country appears to be relatively clustered in the sense that most of Indonesia’s languages are spoken by a median of six thousand speakers in a country of 260 million people. The abundance of

					Aggregate productive efficiency
Region	Province	Absolute linguistic clustering <i>ALC</i> (%)	Linguistic diversity <i>G</i> (%)	Linguistic polarization <i>P</i> (%)	
Sumatra	Aceh	93.21	39.33	57.99	133
	Sumatera Utara	92.62	87.49	42.66	
	Sumatera Barat	91.66	17.06	30.56	
	Riau	92.29	11.99	22.87	
	Jambi	83.69	41.84	65.83	
	Bengkulu	88.83	78.50	61.96	
	Sumatera Selatan	91.66	61.92	74.67	
	Lampung	88.91	20.54	38.41	
Kalimantan	Kalimantan Barat	93.23	82.64	53.45	133
	Kalimantan Tengah	96.61	68.08	76.57	
	Kalimantan Timur	93.07	86.59	42.23	
	Kalimantan Selatan	94.01	16.87	30.66	
Java and Bali	Jakarta Raya	99.99	0.06	0.11	
	Java Barat	89.25	31.56	58.91	
	Java Tengah	92.87	1.76	3.53	
	Java Timur	64.32	34.94	67.53	
	Bali	98.23	1.52	3.04	
	Yogyakarta	99.99	0	0	
Sulawesi	Sulawesi Selatan	76.62	67.52	71.14	133
	Sulawesi Tengah	95.71	76.35	57.54	
	Sulawesi Tenggara	92.52	83.97	50.39	
	Sulawesi Utara	86.45	74.66	60.88	
Nusa Tenggara	Nusa Tenggara Timur	91.3	94.65	18.76	
	Nusa Tenggara Barat	97.31	46.79	71.91	
Maluku	Maluku	94.16	89.72	32.26	
Irian Jaya	Irian Jaya	95.00	85.07	42.20	
Average		91.29	50.05	43.69	

Notes: The indices of absolute linguistic clustering, diversity and polarization are province-specific. The province-specific values of the index of absolute linguistic clustering are computed as weighted averages of the language-specific indices of absolute linguistic clustering with the speaker shares serving as weights

Source: Author's own calculations

languages spoken by a small number of people together with a significant clustering of the more populous dominant languages such as Javanese and Malay provides for the rather high levels of the absolute linguistic clustering index.

5. Data sources and summary

The World Language Mapping System (WLMS, 2004), version 17, is the source of our data on the language-related variables. This database maps the world's languages listed by the Ethnologue (Simons and Fennig, 2018) into their respective geographical areas. The WLMS database contains information on 399 languages in Indonesia. For the purpose of visualizing the data, we used the QGIS mapping software. We wrote an application that computes the

values of absolute linguistic clustering index according to [equation \(3\)](#) for each language. The values of ALC_i at a province level are computed as a weighted average of this index for all languages spoken in the province using speaker shares as weights.

The data on real gross domestic product (GDP), population and capital stock for the period between 1983 and 2007 at the provincial level are provided by [Kataoka \(2013\)](#). The values of real GDP and capital stock are provided in the constant prices of 2001. We also used Indonesian statistical yearbooks as an alternative source of the real GDP data for a robustness check.

Indonesian statistical yearbooks report the data on regional GDP in the constant prices of different years depending on the time period. Thus, for the period between 1983 and 1992, the base year is 1983, changing to 2000 for the period between 1994 and 2001. For 2002–2007, the base year is 2000. We compute GDP deflator values by using the corresponding real GDP data. For the deflator data for 1994, we use the data on real GDP growth rate by province from the Indonesian statistical yearbooks. For the year of 1993, we impute the values of provincial GDP deflator from the data set of [Kataoka \(2013\)](#). The resulting series of real provincial GDP in Indonesia is computed in constant prices of 2002.

The data on economic indicators used as controls are taken from the Indonesian statistical yearbooks. [Table 2](#) lists summary statistics for the real GDP, real GDP per capita, capital stock and population. The values of level and per-capita real GDP as well as the values of capital stock are given in the constant prices of 2001.

Expectedly, the capital region of Jakarta earns highest per-capita incomes, while the residents of Nusa Tenggara Timur earn the least. The Indonesian population is highly concentrated in the three provinces of the Java island. Capital stock exceeds the value of a provincial real GDP by an average factor of 2.12.

6. Empirical framework and results

6.1 Basic empirical framework

The basis of our empirical analysis is a stochastic frontier framework developed by [Meeusen and van den Broeck \(1977\)](#) independently of [Aigner et al. \(1977\)](#). We use a true random effects model by [Greene \(2005\)](#) within this framework assuming a Cobb-Douglas aggregate production function:

$$\begin{aligned} \ln Y_{it} &= \delta_0 + \delta_i + \delta_1 \ln K_{it} + \delta_2 \ln L_{it} + \delta_3 \text{YEAR}_t + v_{it} - u_{it} \\ i &= 1..N, \quad t = 1..T \\ v_{it} &\sim N(0, \sigma_v^2) \\ u_{it} &\sim \left| N(\mu_{it}, \sigma_u^2) \right| \\ \mu_{it} &= \beta_0 + \beta_1' Z_{it} + \beta_2' M_i \end{aligned} \tag{4}$$

Table 2.
Basic summary
statistics for the
Indonesian
provinces, 1983–2007

Variable	No. of observations	Mean	Median	SD	Minimum	Maximum
Real GDP, trillion Rupiahs	650	44.9	16.3	64.3	1.782	339
Real GDP per capita, million Rupiahs	650	6.763	4.285	7.14	1.093	36.733
Real GDP per capita, 2001 US\$	650	\$675	\$427	\$713	\$109	\$3,668
Capital stock, trillion Rupiahs	650	95.1	29.3	16.9	2.537	1,154
Population, million people	650	7.298	3.300	10.1	0.899	49.553

Sources: [Kataoka \(2013\)](#). Monetary values are in constant prices of 2001

where $N = 26$ is the number of provinces and $T = 25$ is the number of years, K_{it} is the capital stock, and L_{it} is provincial population, and $YEAR_t$ are the year dummies.

In the composite error term $\varepsilon_{it} = v_{it} - u_{it}$ the symmetric components v_{it} are independently and identically distributed normal random shocks, that is, $v_{it} \sim N(0, \sigma_v^2)$, and δ_i are province-specific time-invariant random terms that capture cross-province heterogeneity. The assumption maintained within the random-effects setting is that δ_i are uncorrelated with any of the independent variables. The term $u_{it} > 0$ represents technical inefficiency and is assumed to follow a truncated normal distribution, that is, $u_{it} \sim |N(\mu_{it}, \sigma_u^2)|$. The mean μ_{it} of the inefficiency term u_{it} is itself a function of determinants Z_{it} and M_i where Z_{it} is a vector of labor market, schooling, agricultural and budgetary characteristics, and $\vec{M}_i$ is a province-specific, time-invariant vector of linguistic characteristics.

As the determinants of the level of technical inefficiency contain both time-varying determinants Z_{it} , and their time-invariant counterparts M_i , it appears unreasonable to estimate equation (4) using a fixed-effect approach, Pesaran and Zhou (2018), which motivates our using a random-effects model.

6.2 Productive efficiency estimates

We estimate the parameters in equation (4) in a single step using the *sfp* command in Stata for Greene's (2005) true random effects model with inefficiency determinants. Parameter $\lambda = \frac{\sigma_u}{\sigma_v}$ represents the importance of inefficiency effects and is estimated to be $\lambda = 1.36$, suggesting a significant presence of inefficient productive behavior. Table 3 reports our estimates of provincial technical inefficiency.

None of the Indonesian provinces during the period of study has operated on the best practice frontier. While average productive efficiency appears to have been slowly growing in the period before the Asian financial crisis of 1997, it was dropping as slowly in the period after 1998.

6.3 Linguistic clustering, heterogeneity and aggregate productive efficiency

We model the means of aggregate productive efficiency in the Indonesian provinces to be functions of the time-varying determinants Z_{it} and the time-invariant determinants M_i . The vector M_i includes those variable combinations that are implied by the theory we discussed in Section 3, in particular equation (2), namely, $M_i = (ALC_i, P_i \times ALC_i, G_i, G_i \times ALC_i)$ where ALC_i is the index of absolute linguistic clustering, G_i is the index of linguistic diversity and P_i is the index of linguistic polarization. Vector Z_{it} includes labor market characteristics such as the log of population $Log(POP_{it})$, labor force participation ratio $LFPR_{it}$ and the ratio of the number of working people to the number of those economically active WTA_{it} . Z_{it} includes schooling measures such as the ratio of teachers to pupils in primary schools $PRIM_{it}$, the ratio of professors to students in private universities $UPRIV_{it}$ and the ratio of professor to students in state universities $USTATE_{it}$. As agriculture is important to the Indonesian economy, we included the following measures in Z_{it} : the ratio of the forests to the provinces' total areas $FORSHARE_{it}$, the percentage of households employed in the fishery sector $FISHARE_{it}$, the yield rate of rice paddy fields $PADDY_{it}$, cassava yield rates $CASSAVA_{it}$ and the yield rates of sweet potatoes $SWEET_{it}$. Finally, Z_{it} contains $BUDGET_{it}$, the ratio of the provincial government's budget balance to the provincial regional GDP, and $TROPEN_{it}$, the ratio of the freight in tons loaded and unloaded in the provinces' ports that either originates abroad or is destined to be shipped abroad to the total weight of the loaded and unloaded cargo.

Table 3.
Aggregate
productive efficiency
in the Indonesian
provinces, 1983–2007

Region	Province	1983–1987 (%)	1988–1993 (%)	1994–1998 (%)	1999–2003 (%)	2004–2007 (%)
Sumatra	Aceh	89.08	96.43	94.50	80.51	66.32
	Sumatera Utara	91.01	92.49	95.23	95.78	95.89
	Sumatera Barat	92.18	95.39	94.58	93.00	93.60
	Riau	97.24	94.21	92.34	93.81	84.85
	Jambi	93.86	90.72	91.11	95.63	97.32
	Bengkulu	97.12	90.42	85.35	92.02	96.81
	Sumatera Selatan	97.51	93.84	91.84	89.39	90.98
	Lampung	92.58	96.55	93.17	91.40	90.23
Kalimantan	Kalimantan Barat	80.57	93.66	96.47	95.42	93.20
	Kalimantan Tengah	91.11	96.23	95.79	89.16	85.77
	Kalimantan Timur	96.30	92.81	90.53	95.76	89.47
	Kalimantan Selatan	95.87	94.46	93.63	92.72	94.67
Java and Bali	Jakarta Raya	95.81	92.86	93.69	91.32	94.61
	Java Barat	92.03	92.19	91.14	96.09	96.39
	Java Tengah	93.29	93.95	92.99	94.28	96.30
	Java Timur	94.38	93.49	94.41	94.41	95.96
	Bali	84.82	90.45	94.83	96.35	96.74
	Yogyakarta	92.60	93.32	95.53	95.74	94.55
Sulawesi	Sulawesi Selatan	86.20	93.08	94.61	96.63	96.11
	Sulawesi Tengah	95.47	92.19	91.13	95.06	96.46
	Sulawesi Tenggara	91.62	95.07	95.16	91.90	90.84
	Sulawesi Utara	82.43	86.58	95.20	97.43	94.98
Nusa Tenggara	Nusa Tenggara Timur	97.00	94.70	91.81	93.27	94.09
	Nusa Tenggara Barat	92.16	89.55	89.13	95.47	96.83
Maluku	Maluku	85.29	96.17	92.89	64.06	74.45
Irian Jaya	Irian Jaya	76.28	84.00	95.29	97.25	81.02
Average		91.30	92.88	93.17	92.46	91.48

Source: Author's own calculations.

Table 4 displays a series of specifications of equation (4) with the nested sets of controls where we do not report coefficient estimates on the province and year dummies for the sake of brevity.

The direct effect of the index of absolute linguistic clustering on the aggregate productive efficiency is consistently estimated to be positive. The sign of the association between linguistic diversity G_i and productive efficiency, however, depends on the extent of the linguistic clustering ALC_i . Indeed, the most complete specification S4.6 implies that the marginal effect of linguistic diversity G_i on technical efficiency is equal to $\frac{\partial E_{it}}{\partial G_i} = 44.053 - 46.826 \times ALC_i$, where E_{it} are the estimated scores of technical efficiency.

Dependent variable (%)	(S4.1)	(S4.2)	(S4.3)	(S4.4) Aggregate productive efficiency	(S4.5)	(S4.6)
<i>ALC</i>	26.137 (9.999)***	24.491 (11.034)**	26.772 (9.519)***	27.168 (9.852)***	37.729 (1.118)***	37.355 (12.307)***
<i>G</i>	31.154 (12.356)**	28.745 (13.222)**	29.549 (12.412)**	30.591 (11.605)***	43.952 (12.285)	44.053 (15.462)***
<i>G × ALC</i>	-33.175 (13.161)**	-30.603 (14.083)**	-31.441 (13.221)**	-32.552 (12.360)***	-46.722 (13.079)***	-46.826 (16.461)***
<i>P × ALC</i>	0.652 (0.299)**	0.646 (0.374)*	0.861 (0.294)***	0.826 (0.363)**	1.284 (0.338)***	1.223 (0.420)***
<i>TROPEN</i>	0.006 (0.018)	0.021 (0.019)	0.017 (0.023)	0.018 (0.021)	0.012 (0.020)	0.008 (0.017)
<i>Log(POP)</i>	0.049 (0.045)	0.056 (0.067)	0.119 (0.069)*	0.103 (0.069)	0.090 (0.085)	0.069 (0.093)
<i>Log(AREA)</i>	-0.056 (0.075)	-0.064 (0.106)	-0.168 (0.111)	-0.143 (0.115)	-0.132 (0.150)	-0.104 (0.149)
<i>LFP</i>		-0.002 (0.001)**	-0.003 (0.001)**	-0.003 (0.001)***	-0.004 (0.001)***	-0.004 (0.001)***
<i>WTA</i>			0.001 (0.002)	0.001 (0.002)	0.003 (0.002)	0.003 (0.002)
<i>PRIM</i>				0.033 (0.678)	0.033 (0.400)	0.039 (0.814)
<i>UPRIV</i>				-0.063 (0.058)	-0.039 (0.065)	-0.024 (0.067)
<i>USTATE</i>				-0.001 (0.014)	0.001 (0.020)	-0.001 (0.024)
<i>FORSHARE</i>					-0.043 (0.063)	-0.073 (0.055)
<i>FISHARE</i>					-0.030 (0.156)	0.002 (0.159)
<i>PADDY</i>					0.005 (0.003)**	0.006 (0.003)**
<i>CASSAVA</i>					0 (0)	0 (0)
<i>SWEET</i>					0.002 (0.0005)***	0.002 (0.0005)***
<i>BUDGET</i>					0.002 (0.0004)***	-0.442 (1.014)
<i>Year dummies</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Constant</i>	-24.076 (9.528)**	-22.392 (10.514)**	-24.558 (9.126)***	-24.946 (9.348)***	-35.512 (9.528)***	-35.135 (11.845)***
<i>Observations</i>	525	434	375	374	367	356

Notes: Bootstrapped standard errors in parentheses. Significance levels are defined as *** $p < 0.01$; ** $p < 0.05$ and * $p < 0.1$. G stands for linguistic diversity; P = Linguistic polarization; ALC = Index of absolute linguistic clustering; POP = Population; LFP = Labor force participation rate; WTA = Ratio of the number of workers to the number of economically active population; PRIM = Teacher to pupil ratio in primary schools; UPRIV = Professors to students ratio in private universities and USTATE = Professors to students ratio in state universities. FORSHARE is the ratio of the forests to the provinces' total areas; FISHARE = The percentage of households employed in the fishery sector; PADDY = Yield rate of rice paddy fields; CASSAVA = Cassava yield rates; SWEET = Sweet potato yield rates; BUDGET is the ratio of the provincial government's budget balance to the value of regional GDP and TROPEN is the ratio of the freight in tons loaded and unloaded in the provinces' ports that either originates abroad or is destined to be shipped abroad to the total weight of the loaded and unloaded cargo

Source: Author's own calculations

Table 4.
Aggregate
productive efficiency
and linguistic
clustering, diversity
and polarization in
the Indonesian
provinces, 1983–2007

The total marginal effect of linguistic diversity on productive efficiency becomes negative for a high enough extent of linguistic clustering in all specifications. The effect of the interaction between linguistic polarization P_i and linguistic clustering ALC_i is positive and statistically significant in all specifications, suggesting that linguistic polarization is positively associated with productive efficiency given a non-negligible extent of linguistic clustering.

In [Table 5](#), we look at how sensitive our estimation results are to the choice of the sampling period and the source of data. We replace real GDP provided by [Kataoka \(2013\)](#) and provincial population with the real GDP and economically active population obtained from the Indonesian statistical yearbooks. At the same time, we add dummies controlling for income, the Asian financial crisis of 1997 and 1998 and an island dummy. Thus, the *MEDINC* dummy equals 1 if a province's per capita income is less than the median income for the country. Dummy *CRISIS* = 1 for the years of 1997 and 1998. Dummy *JAKYOG* = 1 for the provinces of Jakarta Raya and Yogyakarta, as both regions are classified as Special Regions by the Indonesian Government.

Qualitatively the results in [Table 5](#) are the same with the results reported in [Table 4](#). We conclude that the marginal effects of linguistic clustering, diversity and polarization on the aggregate productive efficiency are robust with respect to the source of data and the choice of controls.

7. Discussion

Our empirical results suggest that the extent of absolute linguistic clustering ALC_i is an important determinant of productive efficiency along with linguistic diversity and linguistic polarization. To our knowledge, ours is the first attempt to incorporate the absolute linguistic clustering index into a list of the determinants of economic performance by linking it to the extent of social cohesion. While the quantification of social cohesion in [Esteban et al. \(2012\)](#) is based on the answers to certain questions in the 2005 *World Values Survey*, ours is a function of the geographical distribution of the speakers of different languages, which makes this study a contribution to the literature on the economics of language as well.

Our empirical findings confirm the three hypotheses we formulated in Section 3. Indeed, given the empirical results presented in [Tables 4](#) and [5](#), our findings can be summarized as follows:

$$\left\{ \begin{array}{l} \frac{\partial E}{\partial ALC} > 0 \\ \frac{\partial E}{\partial G} > 0 \text{ if } ALC \leq ALC^* \\ \frac{\partial E}{\partial G} < 0 \text{ if } ALC > ALC^* \\ \frac{\partial E}{\partial (ALC \times P)} > 0 \end{array} \right. \quad (5)$$

where E is the level of aggregate production efficiency, G is the index of linguistic diversity, P is the index of linguistic polarization, ALC is the index of absolute linguistic clustering and ALC^* is a certain threshold level of it.

The first inequality in [equation \(5\)](#) says that linguistically more clustered societies are also producing more efficiently, which is the assertion of *H1* in Section 3. *H2* says that the

Dependent variable (%)	(S5.1)	(S5.2)	(S5.3)	Aggregate productive efficiency		(S5.4)	(S5.5)	(S5.6)
	All years	Income below median	Crisis of 1997–1998	Income and crisis		Income and crisis	Jakarta and Yogyakarta	Income, crisis and special regions
ALC	50.107 (12.459)***	53.068 (12.858)***	50.555 (14.106)***	53.602 (11.610)***		53.602 (13.900)***	50.017 (12.048)***	53.602 (13.900)***
G	66.745 (15.935)***	69.425 (16.102)***	66.764 (17.480)***	69.475 (15.177)***		69.475 (17.110)***	66.637 (15.615)***	69.475 (17.110)***
G × ALC	-71.029 (16.978)***	-73.870 (17.147)***	-71.053 (18.616)***	-73.926 (16.167)***		-73.926 (18.216)***	-70.917 (16.619)***	-73.926 (18.216)***
P × ALC	1.037 (0.481)**	1.102 (0.465)**	1.095 (0.603)*	1.168 (0.478)**		1.168 (0.525)**	1.035 (0.438)**	1.168 (0.525)**
TROPEN	0.018 (0.023)	0.023 (0.024)	0.022 (0.027)	0.027 (0.024)		0.027 (0.024)	0.017 (0.023)	0.027 (0.024)
Log(POP)	-0.114 (0.121)	-0.098 (0.114)	-0.098 (0.155)	-0.098 (0.124)		-0.098 (0.117)	-0.113 (0.105)	-0.098 (0.117)
Log(AREA)	0.228 (0.209)	0.172 (0.199)	0.201 (0.285)	0.141 (0.227)		0.229 (0.177)	0.229 (0.177)	0.141 (0.213)
LFPR	-0.003 (0.001)**	-0.003 (0.001)**	-0.003 (0.001)**	-0.003 (0.001)**		-0.003 (0.001)**	-0.003 (0.001)**	-0.003 (0.001)**
WTA	0.004 (0.003)*	0.004 (0.003)*	0.005 (0.002)**	0.004 (0.002)**		0.005 (0.003)*	0.005 (0.003)*	0.004 (0.002)*
PRM	0.095 (1.133)	0.081 (0.535)	0.096 (0.969)	0.082 (0.722)		0.082 (1.112)	0.095 (0.962)	0.082 (1.112)
UPRIV	-0.019 (0.099)	0.021 (0.060)	-0.016 (0.081)	-0.017 (0.090)		-0.017 (0.086)	-0.018 (0.075)	-0.017 (0.086)
USTATE	-0.009 (0.046)	-0.001 (0.039)	-0.001 (0.051)	-0.001 (0.033)		-0.001 (0.029)	-0.001 (0.029)	-0.001 (0.032)
FORSHARE	-0.012 (0.064)	-0.008 (0.061)	-0.010 (0.083)	-0.005 (0.060)		-0.005 (0.073)	-0.011 (0.084)	-0.005 (0.073)
FISHARE	-0.076 (0.214)	-0.035 (0.237)	-0.083 (0.164)	-0.042 (0.181)		-0.042 (0.185)	-0.076 (0.137)	-0.042 (0.185)
PADDY	0.009 (0.003)***	0.008 (0.003)**	0.009 (0.004)**	0.007 (0.003)**		0.007 (0.003)**	0.009 (0.004)**	0.007 (0.003)**
CASSAVA	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.00005 (0.0004)		0.0001 (0.0004)	0.0001 (0.0004)	0.0001 (0.0004)
SWEET	0.002 (0.0005)***	0.001 (0.0004)	0.002 (0.0005)***	0.001 (0.0005)**		0.001 (0.0004)***	0.002 (0.0004)***	0.001 (0.0004)***
BUDGET	-0.476 (0.503)	-0.357 (0.883)	-0.476 (0.398)	-0.355 (0.512)		-0.355 (0.741)	-0.476 (0.412)	-0.355 (0.741)
MEDINC		-0.049 (0.012)***		-0.049 (0.012)***		-0.049 (0.010)**		-0.049 (0.010)**
CRISIS			-0.015 (0.014)	-0.016 (0.013)		-0.016 (0.014)		-0.016 (0.014)
JAKYOG						-2.053 (1.208)*	-1.496 (1.358)	-2.053 (1.208)*
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Province dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	-48.249 (11.780)***	-50.521 (12.203)***	-48.700 (13.355)***	-51.052 (11.603)***		-51.052 (13.278)***	-48.249 (11.157)***	-51.052 (13.278)***
Observations	356	356	356	356		356	356	356

Notes: Bootstrapped standard errors in parentheses. Significance levels are defined as *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.1$. G stands for linguistic diversity; P = Linguistic polarization; ALC = Index of absolute linguistic clustering; POP = Population; LFPR = Labor force participation rate; WTA = Ratio of the number of workers to the number of economically active population; PRM = Teacher to pupil ratio in primary schools; UPRIV = Professors to students ratio in private universities and USTATE = Professors to students ratio in state universities; FORSHARE = The ratio of the forests to the provinces' total areas; FISHARE = The percentage of households employed in the fishery sector; PADDY = Yield rate of rice paddy fields; CASSAVA = Cassava yield rates; SWEET = Sweet potato yield rates; BUDGET is the ratio of the provincial government's budget balance to the value of regional GDP and TROPEN is the ratio of the freight in tons loaded and unloaded in the provinces' ports that either originates abroad or is destined to be shipped abroad to the total weight of the loaded and unloaded cargo. MEDINC = 1 if the province's income is less than the country's median income. CRISIS = 1 for the crisis years of 1997 and 1998. JAKYOG = 1 for the regions of Jakarta and Yogyakarta

Source: Author's own calculations

Table 5.
Robustness checks

sign of the marginal effect of linguistic diversity on productive efficiency is positive if the extent of linguistic clustering falls short of some threshold level ALC^* and is negative otherwise. The second and third inequalities in [equation \(5\)](#) confirm *H2*. Finally, the third inequality in [equation \(5\)](#) is a confirmation of *H3*, as we find a positive association between linguistic polarization and productive efficiency with the effect getting stronger if different language speakers are geographically more clustered.

In Section 3, we argued in favor of approximating the extent of within-group social cohesion by the index of absolute linguistic clustering on the basis of the ease of monitoring argument advanced by [Esteban et al. \(2012\)](#). We posit that such monitoring would be easier in those linguistic groups whose members tend to live in geographical clusters. We believe a theoretical framework is needed that would link the extent of geographical clustering of the competing groups to within-group social cohesion.

In our study, we relied on the assumption that a more intense conflict between linguistic groups creates incentives to produce more efficiently, as participation in a conflict requires resources whose supply would be limited in case the productive resources are wasted. *H1* says that conflicts between linguistic groups lead to a more efficient use of productive resources if these groups are internally more cohesive. The results of our study imply that the inter-group conflict has characteristics of competition, as the latter leads to increases in productive efficiency.

H2 is more difficult to interpret. In their empirical work, [Esteban et al. \(2012\)](#) find that the value of their social cohesion parameter α is greater than unity, suggesting altruistic behavior. They do not seem to explain, however, why altruistic behavior in more diverse societies is likely to decrease the intensity of conflict (see page 1320). Our findings are similar in the sense that, once a certain threshold level of linguistic clustering is exceeded, linguistic diversity starts producing a negative effect on the intensity of conflict and hence productive efficiency. However, we do not have a theory of why such a threshold would exist and how higher levels of within-group cohesion would make more linguistic diversity detrimental to productive efficiency.

Finally, our empirical findings concur with the theoretical predictions of [Esteban et al. \(2012\)](#), who say that linguistic polarization becomes a source of the inter-group conflict only when some extent of social cohesion is present. As we estimate a positive coefficient on the interaction term between the index of absolute linguistic clustering and linguistic polarization, our results confirm *H3* from Section 3. We believe this finding underscores the need to take into account the extent of social cohesion when analyzing the effects of polarization on the intensity of conflict or economic performance.

Our empirical results strongly suggest that linguistic clustering is an important determinant of the aggregate productive efficiency. Our study is one of the first steps toward incorporating the extent of geographical clustering of the competing groups' members into a more general framework that would link geographical clustering and economic performance in case the inter-group cleavages are defined along more general, not necessarily linguistic, lines.

8. Conclusion

To the best of the author's knowledge, this study is the first attempt to incorporate the spatial dimension of linguistic diversity into a productive efficiency framework by postulating a positive association between the degree of social cohesion within the linguistic groups and an extent to which these groups' members are spatially clustered. At a most basic level, our empirical results support the idea that increased cohesion between members

of the same linguistic group contributes to these groups' ability to use scarce production inputs more efficiently.

While the link between the extent of social cohesion and production efficiency makes intuitive sense and is supported by both the theories discussed in Section 3 and the empirical studies such as Audibert (1997), the relationship between social cohesion and spatial clustering is in need of a stronger theoretical support. Studies like Cutler and Glaeser (1997), Rothwell (2012) and Sturgis *et al.* (2014), among others, provide support for the claim that social cohesion is likely to be affected by the extent of the social group members' spatial segregation, but they stop short of discussing a theory behind the link. Yet, this link is important from the point of view of the government policymaking.

Edin *et al.* (2003), for instance, examine the causality between living in an ethnic enclave and economic outcomes in the context of immigration in Sweden. They identify several channels through which spatial clustering of the members of the same ethnic group may translate into a higher degree of social cohesion, two of which are most relevant in the context of this study. First, the network effects also identified by Portes (1987) and Lazear (1999) allow enclave residents to apply their existing skills because of the absence of language or cultural barriers. Second, enclave residents find it easier to acquire the new survival and job skills from the fellow enclave residents.

At the same time, residing in an ethnic enclave makes it difficult to benefit from skill complementarities with the other ethnicities, which produces a negative effect on economic performance. Geographical clustering then increases cohesion *within* ethnic groups while decreasing cohesion *between* distinct groups, implying the existence of an optimal extent of spatial clustering. The search for such optimum, however, would be difficult to undertake without a theory linking spatial clustering to social cohesion. In our opinion, efforts aimed at creating such a theory would be well justified.

References

- Aigner, D., Lovell, C.A.K. and Schmidt, P. (1977), "Formulation and estimation of stochastic frontier production function models", *Journal of Econometrics*, Vol. 6 No. 1, pp. 21-37.
- Alesina, A. and La Ferrara, E. (2005), "Ethnic diversity and economic performance", *Journal of Economic Literature*, Vol. 43 No. 3, pp. 762-800.
- Amsler, C., Prokhorov, A. and Schmidt, P. (2014), "Using copulas to model time dependence in stochastic frontier models", *Econometric Reviews*, Vol. 33 Nos 5/6, pp. 497-522.
- Auci, S., Castellucci, L. and Coromaldi, M. (2021), "How does public spending affect technical efficiency? Some evidence from 15 European countries", *Bulletin of Economic Research*, Vol. 73 No. 1, pp. 108-130.
- Audibert, M. (1997), "Technical inefficiency effects among paddy farmers in the villages of the 'office Du niger', Mali, west Africa", *Journal of Productivity Analysis*, Vol. 8 No. 4, pp. 379-394.
- Bannister, G.J. and Stolp, C. (1995), "Regional concentration and efficiency in Mexican manufacturing", *European Journal of Operational Research*, Vol. 80 No. 3, pp. 672-690.
- Battese, G.E. and Coelli, T.J. (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data", *Empirical Economics*, Vol. 20 No. 2, pp. 325-332.
- Becerra-Ornelas, A.U. and Nunez, H.M. (2019), "The technical efficiency of local economies in Mexico: a failure of decentralized public spending", *The Annals of Regional Science*, Vol. 62 No. 2, pp. 247-264.
- Beeson, E. and Husted, S. (1989), "Patterns and determinants of productive efficiency in state manufacturing", *Journal of Regional Science*, Vol. 29 No. 1, pp. 15-28.

- Bertinelli, L. and Zou, B. (2008), "Does urbanization foster human capital accumulation?", *The Journal of Developing Areas*, Vol. 41 No. 2, pp. 171-184.
- Caraballo, Á. and Buitrago, E.M. (2019), "Ethnolinguistic diversity and education. A successful pairing", *Sustainability*, Vol. 11 No. 23, p. 6625.
- Charnes, A., Cooper, W.W. and Rhodes, E. (1978), "Measuring efficiency of decision making units", *European Journal of Operational Research*, Vol. 2 No. 6, pp. 429-444.
- Chortareas, G.E., Desli, E. and Pelagidis, T. (2003), "Trade openness and aggregate productive efficiency", *European Research Studies*, Vol. 6 Nos 1/2, pp. 188-199.
- Cutler, D.M. and Glaeser, E.L. (1997), "Are ghettos good or bad?", *The Quarterly Journal of Economics*, Vol. 112 No. 3, pp. 827-872.
- Dacey, M.F. (1968), "A review on measures of contiguity for two and K-Color maps", *Spatial Analysis: A Reader in Statistical Geography*, Prentice Hall, Marble.
- Demir, N. and Mahmud, S. (1998), "Regional technical efficiency differentials in the Turkish agriculture: a note", *Indian Economic Review*, Vol. 33 No. 2, pp. 197-206.
- Desmet, K., Ortuño-Ortín, I. and Wacziarg, R. (2012), "The political economy of linguistic cleavages", *Journal of Development Economics*, Vol. 97 No. 2, pp. 322-338.
- Easterly, W. and Levine, R. (1997), "Africa's growth tragedy: policy and ethnic divisions", *The Quarterly Journal of Economics*, Vol. 112 No. 4, pp. 1203-1250.
- Edin, P.A., Fredriksson, P. and Åslund, O. (2003), "Ethnic enclaves and the economic success of immigrants—evidence from a natural experiment", *The Quarterly Journal of Economics*, Vol. 118 No. 1, pp. 329-357.
- Esteban, J.M. and Ray, D. (1994), "On the measurement of polarization", *Econometrica*, Vol. 62 No. 4, pp. 819-851.
- Esteban, J.M. and Ray, D. (2011), "Linking conflict to inequality and polarization", *American Economic Review*, Vol. 101 No. 4, pp. 1345-1374.
- Esteban, J., Mayoral, L. and Ray, D. (2012), "Ethnicity and conflict: an empirical study", *American Economic Review*, Vol. 102 No. 4, pp. 1310-1342.
- Farrell, M.J. (1957), "The measurement of productive efficiency of production", *Journal of the Royal Statistical Society*, Vol. 120 No. 3, pp. 253-281.
- Fishman, J. (1968), "Some contrasts between linguistically homogeneous and linguistically heterogeneous polities", *Language Problems of Developing Nations*, Wiley, New York, NY.
- Geary, M.C. (1954), "The contiguity ratio and statistical mapping", *The Incorporated Statistician*, Vol. 5 No. 3, pp. 115-127.
- Gheit, S. (2022), "A stochastic frontier analysis of the human capital effects on the manufacturing industries' technical efficiency in the United States", *Athens Journal of Business and Economics*, Vol. 8 No. 3, pp. 215-238.
- Ginsburgh, V. and Weber, S. (2020), "The economics of language", *Journal of Economic Literature*, Vol. 58 No. 2, pp. 348-404.
- Gonzalez, X. and Miles-Touya, D. (2012), "Labor market rigidities and economic efficiency: evidence from Spain", *Labour Economics*, Vol. 19 No. 6, pp. 833-845.
- Greenberg, J. (1956), "The measurement of linguistic diversity", *Language*, Vol. 32 No. 1, pp. 109-115.
- Greene, W. (2005), "Fixed and random effects in stochastic frontier models", *Journal of Productivity Analysis*, Vol. 23 No. 1, pp. 7-32.
- Greene, W.H. (2008), "The econometric approach to efficiency analysis", *The Measurement of Productive Efficiency and Productivity Growth*, Oxford University Press, New York, NY, pp. 92-251.
- Hart, J., Miljkovic, D. and Shaik, S. (2015), "The impact of trade openness on technical efficiency in the agricultural sector of the European Union", *Applied Economics*, Vol. 47 No. 12, pp. 1230-1247.

- Horowitz, D.L. (1985), *Ethnic Groups in Conflict*, University of CA Press, Berkeley, CA.
- Jondrow, J., Knox Lovell, C.A., Materov, I.S. and Schmidt, P. (1982), "On the estimation of technical inefficiency in the stochastic frontier production function model", *Journal of Econometrics*, Vol. 19 Nos 2/3, pp. 233-238.
- Kataoka, M. (2013), "Capital stock estimates by province and interprovincial distribution in Indonesia", *Asian Economic Journal*, Vol. 27 No. 4, pp. 409-428.
- Kumbhakar, S.C., Ghosh, S. and McGuckin, J.T. (1991), "A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms", *Journal of Business and Economic Statistics*, Vol. 9 No. 3, pp. 279-286.
- Kumbhakar, S.C., Lien, G. and Hardaker, J.B. (2014), "Technical efficiency in competing panel data models: a study of Norwegian grain farming", *Journal of Productivity Analysis*, Vol. 41 No. 2, pp. 321-337.
- Lazear, E.P. (1999), "Culture and language", *Journal of Political Economy*, Vol. 107 No. S6, pp. S95-126.
- Lee, C., Adkins, R.L., Moomaw and A., Savvides. (2002), "Institutions, freedom, and technical efficiency", *Southern Economic Journal*, Vol. 68 No. 1, pp. 92-108.
- Lich, H.-K., Tiet, T., Nguyen, T.-T. and Tuan, N.-A. (2022), "Impact of human capital on technical efficiency in sustainable food crop production: a meta-analysis", *International Journal of Agricultural Sustainability*, Vol. 20 No. 4, pp. 521-542.
- Lovell, C.A.K. (1993), "Production frontiers and production efficiency", *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, Oxford.
- Massey, D.S. and Denton, N.A. (1988), "The dimensions of residential segregation", *Social Forces*, Vol. 67 No. 2, pp. 281-315.
- Meeusen, W. and van den Broeck, J. (1977), "Efficiency estimation from Cobb-Douglas production functions with composed error", *International Economic Review*, Vol. 18 No. 2, pp. 435-444.
- Micco, A. and Pages, C. (2006), "The economic effects of employment protection: evidence from international industry-level data", Discussion Paper, Institute of Labor Economics, Bonn.
- Montalvo, J.G. and Reynal-Querol, M. (2005), "Ethnic diversity and economic development", *Journal of Development Economics*, Vol. 76 No. 2, pp. 293-323.
- Nettle, D. (2000), "Linguistic fragmentation and the wealth of nations: the Fishman-Pool hypothesis reexamined", *Economic Development and Cultural Change*, Vol. 48 No. 2, pp. 335-348.
- Pesaran, M.H. and Zhou, Q. (2018), "Estimation of Time-Invariant effects in static panel data", *Econometric Reviews*, Vol. 37 No. 10, pp. 1137-1171.
- Pitt, M.M. and Lee, L.-F. (1981), "The measurement and sources of technical inefficiency in the Indonesian weaving industry", *Journal of Development Economics*, Vol. 9 No. 1, pp. 43-64.
- Portes, A. (1987), "The social origins of the Cuban enclave economy of Miami", *Sociological Perspectives*, Vol. 30 No. 4, pp. 340-372.
- Repkine, A. (2014), "Ethnic diversity, political stability and productive efficiency: empirical evidence from the African countries", *South African Journal of Economics*, Vol. 82 No. 3, pp. 315-333.
- Rothwell, J. (2012), "The effects of racial segregation on trust and volunteering in US cities", *Urban Studies*, Vol. 49 No. 10, pp. 2109-2136.
- Schmidt, P. and Sickles, R.C. (1984), "Production frontiers and panel data", *Journal of Business and Economic Statistics*, Vol. 2 No. 4, pp. 367-374.
- Sen, A.K. (1966), "Labour allocation in a cooperative enterprise", *The Review of Economic Studies*, Vol. 33 No. 4, pp. 361-371.
- Sickles, R.C. and Zelenyuk, V. (2019), *Measurement of Productivity and Efficiency: Theory and Practice*, Cambridge University Press, New York, NY.
- Simons, G.F. and Fennig, C. (2018), *Ethnologue: Languages of the World*, 21st ed, SIL International, Dallas, TX.

- Sturgis, P., Brunton-Smith, I., Kuha, J. and Jackson, J. (2014), "Ethnic diversity, segregation and the social cohesion of neighbourhoods in London", *Ethnic and Racial Studies*, Vol. 37 No. 8, pp. 1286-1309.
- Tsekeris, T. and Papaioannou, S. (2017), "Regional determinants of technical efficiency: evidence from the Greek economy", *Regional Studies*, Vol. 52 No. 10, pp. 1398-1409.
- Wang, H.-J. and Schmidt, P. (2002), "One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels", *Journal of Productivity Analysis*, Vol. 18 No. 2, pp. 219-244.
- White, M.J. (1983), "The measurement of spatial segregation", *American Journal of Sociology*, Vol. 88 No. 5, pp. 1008-1018.
- World Language Mapping System (2004), available at: <https://mdl.library.utoronto.ca/collections/geospatial-data/world-language-mapping-system>
- Zhao, Z., et al. (2022), "The impact of the urbanization process on agricultural technical efficiency in northeast China", *Sustainability*, Vol. 14 No. 19, pp. 1-20.

Further reading

- Arjomand, M., Emami, K. and Salimi, F. (2016), "Growth and productivity; the role of budget deficit in the MENA selected countries", *Procedia Economics and Finance*, Vol. 36, pp. 345-352.
- Becker, G.S. and Murphy, K.M. (1992), "The division of labor, coordination costs, and knowledge", *The Quarterly Journal of Economics*, Vol. 107 No. 4, pp. 1137-1160.
- Collier, P. and Hoeffler, A. (1998), "On economic causes of civil war", *Oxford Economic Papers*, Vol. 50 No. 4, pp. 563-573.
- Gradstein, M. and Justman, M. (2002), "Education, social cohesion, and economic growth", *American Economic Review*, Vol. 92 No. 4, pp. 1192-1204.
- Green, A. and Mayes, D. (1991), "Technical inefficiency in manufacturing industries", *The Economic Journal*, Vol. 101 No. 406, pp. 523-538.
- Hopenhayn, H. and Rogerson, R. (1993), "Job turnover and policy evaluation: a general equilibrium analysis", *Journal of Political Economy*, Vol. 101 No. 5, pp. 373-394.
- Lazear, E. (1990), "Job security provisions and employment", *The Quarterly Journal of Economics*, Vol. 105 No. 3, pp. 699-726.

Corresponding author

Alexandre Repkine can be contacted at: repkine@konkuk.ac.kr

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgroupublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com