Asia Pacific Academy of Science Pte. Ltd.

City Diversity 2024, 5(1), 1959.
https://doi.org/10.54517/cd.v5i1.1959

Article

Dialect diversity and market integration: From the perspective of city circle

Congming Ding*, Zhenlin Ji, Yu Lei, Zhenqiao Liang

School of Public Affairs, Chongqing University, Chongqing 400044, China
*** Corresponding author:** Congming Ding, ding@126.com

CITATION

Ding C, Ji Z, Lei Y, Liang Z. Dialect diversity and market integration: From the perspective of city circle. *City Diversity*. 2024; 5(1): 1959. https://doi.org/10.54517/cd.v5i1.1959

ARTICLE INFO

Received: 13 April 2024
Accepted: 4 May 2024
Available online: 5 June 2024

Abstract: China, as a transitional economy, faces a high level of market segmentation among administrative regions, which lowers the efficiency of resource allocation and the total factor productivity (TFP) rate. The literature has focused on the negative effects of local protectionism and administrative division on the formation of market integration in the economic growth process. However, considering that administrative districts such as prefectures usually overlap with cultural regions in China, the effects of local protectionism and administrative division on market segmentation may be overestimated because cultural diversity may also be negatively related to market integration. More importantly, diversity of dialect tends to increase the cost of communication, making it a barrier to labor migration and decreasing the level of generalized trust among people. As a result, it may have adverse effects on the market integration process. Recently, more empirical works have explored the relationship between cultural diversity, which they usually measure as the number of dialects and amount of economic growth in the region, and have generally shown consistent results. For example, a study has shown that dialect diversity has adverse effects on GDP per capita. Another finds that dialect diversity and not genetic differences can explain regional disparities in China to a large extent. Similarly, Scholars indicates that dialect diversity has adverse effects on the productivity of neighboring counties. Nevertheless, to the best of our knowledge, few works reveal the impact of dialect diversity on the level of market segmentation among regions in China. Taking a somewhat different approach, we directly focus on the effects of dialect diversity on market segmentation. Empirically, to estimate the causal effects of dialect diversity on market segmentation, we randomly build the synthetic metropolitan area as the fundamental analysis unit in which a core prefecture borders several other governorates. Consequently, within the artificial metropolitan area, the number of dialects and amount of market segmentation can be measured. Given that the synthetic metropolitan area does not belong to any particular administrative district, the differences in market segmentation between synthetic metropolitan areas are attributed to variations in dialect and other economic or geographic factors rather than the administrative division between areas. Based on the method developed, this paper uses the seven categories of retail prices in prefectures in 2016 to calculate the market segmentation index of each synthetic metropolitan area, which it takes as the dependent variable. Furthermore, this paper constructs a dialect diversity index for each synthetic metropolitan area, which it takes as the key independent variable. The results show that diversity of dialect is a critical factor in lowering the amount of market integration in China. The findings are robust to various checks. Furthermore, this paper takes the number of local theatrical genres as an instrumental variable of dialect diversity. The instrumented estimations show that a one-dialect increase in the synthetic metropolitan area increases the amount of market segmentation by about 2.42%. The amount of market segmentation in the synthetic metropolitan area, which has the average number of dialects, is 8.23% higher than in areas with only one dialect. The empirical results imply that it is essential to weaken local protectionism and enhance cultural integration between regions to decrease market segmentation. This paper makes three contributions to the literature. First, it enriches the broad interpretations of the causes of market segmentation from the dialect diversity viewpoint. Second, it directly

estimates the effects of dialect diversity on market segmentation and determines the long-term effects of cultural factors, providing new cultural economics evidence from China. Third, this paper contributes to the literature analyzing the underlying mechanisms behind dialect diversity and growth, suggesting that market segmentation is another mechanism used to understand this causal relationship.

Keywords: diversity of dialect; market integration; synthetic metropolitan area

1. Introduction

Under the background of high-speed economic growth turning to high-quality development, it is necessary to establish an integrated market with full and free flow of factors to improve the efficiency of resource allocation and the speed of technological expansion. Considering the wide diversity of Chinese society, economic development shows a high degree of imbalance and inadequacy. Domestic market segmentation has seriously hindered the further improvement of factor allocation efficiency and total factor growth rate (Zheng and Li, 2003). Traditional studies emphasize the adverse effects of administrative segmentation, local protectionism, registered residence system, judicial independence and other factors on market integration (Yin and CAI, 2001; Zhou, 2004; Chen et al., 2007; Liu, 2012; Lu et al., 2011; Chen and Li, 2013). Indeed, the above factors have stimulated the internal driving force of "beggar thy neighbor" of local governments to varying degrees, and promoted the deepening of "local protectionism". However, combined with the fact that China's traditional administrative divisions and cultural regions are highly overlapping, the process of administrative segmentation hindering the formation of market integration also includes the potential impact of cultural diversity. And more importantly, with the improvement of transportation infrastructure and the weakening of institutional and policy barriers, the impact of deep-seated cultural diversity on market efficiency will become increasingly prominent.

In recent years, the research on the impact of cultural diversity, especially dialect diversity, on economic performance has achieved fruitful results in China. Gao and Long (2016) linked language and culture areas with provincial administrative divisions, and found that cities with different mainstream cultures in the province have a relatively backward level of economic development. Liu et al. (2015) revealed the migration mode of labor migration across dialect areas. Xu et al. (2015) pointed out that language diversity has a negative impact on per capita output. Zhao and Lin (2017) emphasized that linguistic (cultural) diversity rather than genetic diversity is an important factor in regional economic development differences. Dai et al. (2016) studied the impact of dialect consistency on reducing the agency cost of enterprises. Li et al. (2017) pointed out that language diversity has inhibited the opening to the outside world. Liu et al. (2017) emphasized the impact of dialect differences on the productivity differences of neighboring counties and cities. Lin and Zhao (2017) pointed out the impact of cultural differences on technology diffusion. Pan et al. (2017) believe that cultural diversity measured by the number of dialects promotes enterprise innovation. Different from the existing studies, this paper directly focuses on the impact of dialect diversity on market integration. The diversity of language increases the cost of communication, hinders the cross regional mobility of labor, and reduces

the level of social trust among people. This means that the diversity of dialects will increase the degree of regional market segmentation, which is not conducive to the formation of an integrated market.

Generally speaking, there are two difficulties in accurately identifying the causal relationship between language diversity and market segmentation. (1) How to separate the "cultural effect" that leads to market segmentation from the "administrative division effect". Throughout the history of regional development in China, the formation and evolution of dialects are closely related to geographical endowment, historical immigration and other factors, resulting in a high degree of overlap between the geographical distribution and administrative distribution of dialects. Cause and effect identification in measurement needs to break through the existing administrative divisions. Therefore, this paper no longer uses the established single province or prefecture level city as the basic analysis unit, but takes "city circle" as the basic measurement unit. A city circle is a geographical unit composed of a central prefecture level city and other prefecture level cities bordering it. In the urban circle, there are many dialects and market segmentation, but they are not a strict administrative division. Through this division, the overlap of language and administrative divisions can be effectively avoided. (2) Endogenous problem. First, there is a causal relationship between market segmentation and dialect diversity. On the one hand, market segmentation reduces the opportunities for cross regional language exchange and integration, and accelerates the solidification of dialects; on the other hand, dialects also aggravate market segmentation by increasing communication costs and emphasizing identity characteristics. Secondly, the estimation results face potential missing variables bias. For example, factors such as geographical endowment, traffic conditions, and ethnic "large and small communities" may not only strengthen the formation of dialects, but also be related to the degree of market segmentation. Based on this, this paper uses two empirical strategies to solve the above difficulties: on the one hand, it estimates the instrumental variables. This paper uses "local opera types" as the instrumental variable of dialect diversity. In traditional China, local operas are usually performed in dialect, and the audience is mainly limited to the dialect area. However, there is no obvious evidence that local operas are directly related to market segmentation, which means that local operas have the attribute of good instrumental variables; on the other hand, this paper uses the method of Nunn and Wantchekon (2011) for reference to measure the impact of unmeasurable factors on estimation errors, so as to prove the robustness of the empirical results in this paper. Overall, the empirical results show that the market segmentation index of the urban circle increases by 242% on average for each increase in dialect diversity. Compared with the city circle with only one dialect, the degree of market integration of the city circle with an average number of dialects is nearly 823% lower.

This paper is mainly complementary to the following important literatures. First, this paper provides a new interpretation of the causes of market segmentation from the perspective of dialect diversity. Early studies emphasized such factors as local protectionism (Yin and Cai, 2001; Zhou, 2004), development strategy (Lin and Liu, 2004), opening to the outside world (Chen et al., 2007; Zhang et al., 2010), subsidies to state-owned enterprises (Liu, 2012), registered residence system (Lu et al., 2011), judicial independence (Chen and Li, 2013). But as this article says, ignoring the

overlapping characteristics of administrative divisions and dialect distribution may underestimate the impact of dialect diversity on market segmentation.

Second, it complements the existing domestic literature on dialect diversity, market segmentation and resource mismatch. The closest research to this paper is the empirical analysis of dialect and resource mismatch by Liu et al. (2017). But this article is different from it in the following two aspects. First, in terms of research focus, this paper directly uses the price difference of seven categories of commodities to calculate the market segmentation degree between adjacent prefecture level cities, while Liu et al. (2017) used the labor productivity difference of industrial enterprises to measure the degree of resource mismatch. Comparatively speaking, using the price difference of commodity market as the measurement index is closer to the concept definition of market segmentation (Gui et al., 2006; Chen et al., 2007), and the measurement method is more direct. Liu et al. (2017) used the labor productivity difference measurement to focus on the degree of market resource mismatch, and market segmentation may be one of the many reasons for resource mismatch. Second, in terms of empirical strategies, this paper adopts the "circle drawing" method to form the basic unit of empirical analysis urban circle, so as to strip away the "language effect" and "administrative division effect" in market segmentation, while Liu et al. (2017) used the identification strategy of "catching the right" comparison between adjacent counties. The identification strategy of "catching the right" is more ingenious, while the "circle" method is more general, which is more suitable for empirical analysis within the framework of general causal identification. Of course, the ultimate goal of "catching the right" or "drawing a circle" is to separate the language effect from the administrative division effect. From this perspective, this study complements the research of Liu et al. (2017). In addition, there are also important differences between this paper and Gao and Long (2016) in terms of research intention. Gao and Long (2016) emphasized that administrative division blocks the connection between cultural areas, resulting in adverse effects on the economic development of isolated sub cultural areas. Therefore, this is actually a further verification that administrative division plays a negative regulatory role between cultural diversity and economic development. In contrast, this paper does not discuss the relationship between administrative divisions, cultural differences and economic growth, but directly analyzes the impact of cultural diversity on market segmentation. In order to obtain the causal effect of cultural differences on market segmentation, administrative division is only one of the interference factors that we need to consider and eliminate. In other words, this study emphasizes the lasting impact of cultural diversity, which to some extent provides empirical evidence from China for the economic literature on the lasting impact of culture (Alesina and Giuliano, 2015).

Thirdly, it enriches the domestic literature foundation on the internal mechanism of cultural diversity and economic development. At present, an existing literature focuses on analyzing the internal mechanism of dialect diversity affecting regional economic development, which mainly includes the discussion of labor mobility (Liu Zet al., 2015), technological change (Lin and Zhao, 2017), and this paper provides another possible internal transmission mechanism for the relationship between cultural diversity and economic development from the perspective of market segmentation.

The structure of the article is arranged as follows: the second part is the research

background and theoretical hypothesis; the third part is the explanation of research methods, dialect diversity and market segmentation index structure; the fourth part is empirical analysis and result explanation; the fifth part is robustness test; the last part is the conclusion of this paper.

2. Background and hypothesis

Generally speaking, the characteristics of highly overlapping dialect areas and administrative divisions depend on many factors, such as geographical endowment conditions, historical immigration and political control. This section gives a basic explanation of the highly overlapping historical background and the internal mechanism of dialect diversity affecting market integration.

2.1. High overlap of dialect, culture and administrative division

Firstly, geographical endowment is one of the important factors in the formation of historical administrative divisions. As recorded in the book of rites, the king system, "there are different systems in wide valleys and large rivers, and people's livelihood is different from that in the meantime.". In the traditional society with underdeveloped transportation, mountains and rivers are the direct basis for delimiting geographical boundaries. For example, it is recorded in the new Tang Dynasty geography annals that "however, the world was initially determined, and there were many prefectures and prefectures. In the first year of Taizong, it was ordered to merge the provinces, and because of the convenience of mountains and rivers, the world was divided into ten roads". The "ten roads" are basically divided by geographical mountains and rivers such as the Yellow River, qinling Mountains Huaihe River and Yangtze River. For example, "Hebei Road" refers to the area to the north of the Yellow River, and "Lingnan road" refers to the area to the south of the guide ridge. Most of these administrative divisions are named directly by their geographical location. After the song and Yuan Dynasties, the administrative division experienced changes such as "road" and "province", but this setting method of dividing the administrative units according to the mountains and rivers had a far-reaching impact on later generations. Once the basic pattern of the above division is formed, climate folklore and cultural diversity will be dependent on geographical endowment and gradually differentiated. Considering the relocation of traditional local society, culture, customs and language will gradually strengthen themselves, and multiculturalism will gradually expand in geographical space with its geographical environment as the basic "carrier".

Secondly, historical immigration is also an important reason for the high overlap between administrative divisions and dialect areas. You and Zhou (1985) pointed out that the main driving mechanism for the differentiation of Chinese language into dialects came from immigrants, and the immigrant areas formed by previous immigrants in history gradually formed later administrative areas. In ancient China, administrative regions were often set up because of immigrants, or immigrants were resettled according to administrative regions, which directly led to the high overlap between cultural regions and administrative regions. As one of the important representations of culture (Sapir, 2011), cultural geography is mainly represented by dialect geography and religious geography. Unlike Europe, there is no obvious

geographical distribution difference in the influence of religion in China, so cultural geography is mainly represented by dialect geography (Zhou, 2013). Considering that the research literature has been able to subdivide language categories into townships, and the high theoretical correlation and geographical overlap between language regions and cultural regions, dialect regions often become one of the best proxy variables of cultural regions (Gao and Long, 2014).

Finally, the need for political control makes administrative divisions and dialect divisions highly overlapping, but not completely corresponding. In particular, within large provincial administrative regions, there are often a variety of small dialects, or large language regions often cover several provincial administrative regions. This feature is mainly based on the consideration of reducing the cost of political control. Although it is convenient to divide administrative regions completely according to mountains and rivers, it has important disadvantages for the centralized central government, which is mainly manifested in the rising risk of separatist rule. The military level of traditional society was relatively low, and mountains, rivers and plateaus became one of the natural conditions for separatism. In order to avoid this kind of separatist regime, the administrative regions and cultural regions were set up in a dog's teeth pattern in the Yuan Dynasty. For example, "Shaanxi province crosses the Qinling Mountains and has the Hanzhong Basin, Hunan and Hubei Province are the main part of Hubei Province and Guangxi province crosses the Nanling Mountains, and Jiangxi Province also crosses the Nanling mountains and has Guangdong Province" (Zhou, 2013). This provincial system is also the predecessor of today's provincial system. Based on the above reasons, Gao and Long (2016) believe that the division of cultural areas by provincial administrative divisions will significantly reduce the economic output of non-mainstream cultural cities in the province. In turn, this means that the multiculturalism in a provincial area will have a negative impact on the integrated market within the province. Administrative division and dialect division are highly overlapped but not exactly the same. They reinforce each other and cause and effect each other, which makes it difficult to accurately identify the "administrative division effect" and "language effect" in market integration.

2.2 How dialect affects market segmentation

If the function of language is to realize the communication and exchange of information, the function of dialect hinders the wider communication and exchange to some extent, thus hindering the formation of the integrated market. The adverse effects of dialects on market integration mainly play a role through the following channels:

(1) Strengthen the identity within the group. Language is an explicit representation of a nation and culture. As a human capital that accompanies a lifetime, it is the most explicit and fastest identity symbol (Gao and Long, 2016). Linguistic diversity reflects cultural conflict, which will reduce social identity and increase psychological distance (Zhang et al., 2012). Alesina and Ferrara (2000) pointed out that the greater the proportion of similar individuals in the population, the greater the positive effect that the individual obtained, and vice versa. McPherson et al. (2001) put forward more clearly from the perspective of organizational concept that homogeneity affects the expansion of personal social network, and people are more

willing to interact with people who are "similar to me". Differences in individual characteristics (such as culture and race) will be subconsciously marked as "not my race". Such cultural differences will increase communication costs, and similar cultural backgrounds are more likely to cause interaction and communication. Chen et al. (2014) stressed that dialects mainly affect individual income through identity rather than communication costs. Through the study of Shanghai's labor market, it is found that auditory comprehension has no significant effect on income growth, because local residents can understand Mandarin, and communication is not a problem. The key is that workers who are proficient in using Shanghai dialect have stronger ties with the local people, strengthening their identity.

(2) Reduce the level of social trust. The identity brought by dialect further reduces the level of general social trust, and then hinders the formation of market integration. In fact, dialects can become an effective "screening mechanism". People can identify different "identities" of individuals through dialects and give them different levels of trust. Pendakur (2002) pointed out that dialect is an important mechanism of identity recognition, which affects the psychological distance between people, and then affects social trust. Huang and Liu Chang (2017) investigated the impact of dialects on social trust. The empirical results show that among strangers, using different dialects will reduce the level of trust between them. Falck et al. (2012) found that the dialect areas in Germany have hardly changed much in the past century, and the influence of cultural segmentation has been strengthened. Even if the geographical span is small, the labor force is still unwilling to flow to areas with unfamiliar cultural environment. He believes that trust breeds economic communication, which is then strengthened by factors such as culture, religion and genes. Alesina and Ferrara (2002) clearly pointed out that the key mechanism for income inequality and ethnic diversity to affect economic development is to affect the level of trust.

(3) It affects the cross regional flow of production factors and technologies. Li and Meng (2014) found that in the process of cross regional mobility, the labor force tends to move to work in places with low communication barriers in Putonghua and common cultural background. Therefore, the differences in dialects constitute one of the most important factors affecting labor mobility. Considering the complementary effect and identity effect brought about by dialect diversity, diversity itself may have an inflection point for labor mobility. Workers are not willing to carry out cross regional mobility. In a large dialect area, even if there are some differences, the complementary effect may occupy a dominant position (Liu et al., 2015). This means that the dialect will strengthen the labor flow within the large dialect area and hinder the formation of an integrated labor market, thus forming a separate regime of the labor market. In terms of technology diffusion, Lin and Zhao (2017) used dialects to measure cultural differences and found that cultural differences reduced the speed of technology diffusion through institutional mediation.

Based on the above reasons, this paper proposes the following hypotheses:

Hypothesis 1: under the condition of keeping other factors unchanged, the more dialects in a region, the higher the degree of market segmentation. It should be emphasized that the dialect diversity hypothesis in this paper is highly complementary to the existing studies on cultural and ethnic diversity. On the one hand, diversity studies generally show that cultural differences caused by diversity such as race and

demographic characteristics may lead to weak social relations, mistrust and the decline of collaboration efficiency (Parrotta et al., 2012; Ashrafand, 2013); on the other hand, the research also shows that diversity can produce complementary effects, which can lead to the improvement of output. Diversity such as age and educational background will produce knowledge spillover effect, innovation and creativity, form cognitive differences among individuals and form technological complementarity (Berliant and Fujita, 2008). The diversity of birthplaces has a significant promoting effect on enterprise performance, employee wage level and economic growth (Trax et al., 2015; Alesina et al., 2016). This study mainly emphasizes the negative impact of dialect diversity on regional economic integration, so it is a useful supplement to the first literature.

3. Research design

3.1. Method description

According to the above research hypothesis, this paper adopts the following linear model in the empirical research:

$$lnseg_i = \alpha + \beta dia_i + \psi X_i + \varepsilon_i \quad (1)$$

Among them, the explained variable $lnseg_i$ represents the index of the market segmentation degree of the ith "city circle", and its natural logarithm is taken to participate in the regression analysis; the key explanatory variables represent $dia_i$the dialect types of the ith "city circle"; it refers to $X_i$other control variables, including institutional factors such as the proportion of state-owned enterprises' employees, the degree of opening to the outside world, the ability of government intervention, fiscal decentralization, variables reflecting the level of economic development such as the GDP and population of the urban circle, and whether they belong to coastal areas, the longitude and latitude of the urban circle, the number of local administrative units within the urban circle and other characteristic factors $\varepsilon_i$within the urban circle; it is an error term to control the factors that have an impact on market segmentation but are difficult to capture. To measure the degree of dialect diversity affecting market segmentation, this paper expects that with the increase of dialect types, the market segmentation degree of urban circle will be significantly improved.

This paper takes Beijing city circle as an example to illustrate the construction method of city circle. Specifically, Beijing, Zhangjiakou, Baoding, Langfang, Tianjin and Chengde constitute an urban circle centered on Beijing, and on this basis, the market segmentation indicators and dialect diversity indicators of the urban circle are constructed. Similar to the division of inter provincial circles by Lu and Chen (2006), this paper uses an administrative region one level lower than the provincial administrative unit to construct an urban circle. Each city circle has a core city, and other municipal administrative units are included in the city circle according to the standard of "whether it borders on the core city" (Core cities include cities under the jurisdiction of provinces, municipalities directly under the central government, prefecture level cities, autonomous regions and regions, that is, all administrative regions one level lower than provincial administrative units are included. The reason is that the border cannot be lack of region.). The core cities of this paper include all

prefecture level cities, autonomous prefectures, municipalities directly under the central government, provincial cities and regions except Hong Kong, macao, taiwan, tibet and Hainan, with a total of 332 measurement units (Hong Kong, Macao, Taiwan and Hainan are excluded because the concept of "border" is not applicable to these regions, and Hong Kong, Macao and Taiwan have strong economic and administrative particularities. Tibet is excluded because of a large number of omissions in its economic data. However, for the integrity of the data of the urban circle, if some cities or regions in Tibet border on core cities, they are also included in the urban circle, but the prefecture and municipal administrative units under the jurisdiction of Tibet are not counted as core cities.).

This paper uses the city circle as the measurement unit, mainly based on the following three considerations: first, it is necessary to measure the impact of dialect diversity on market segmentation. If the province is used as the basic analysis unit, the geographical area of the province is large, which will lead to the convergence of the number of dialects in each province and the lack of variation of key explanatory variables; second, the administrative division system is one of the important factors causing market segmentation. If provinces are used as the analysis unit, it will be difficult for us to eliminate the noise caused by the administrative division itself; third, this paper cannot use prefecture level cities as the analysis unit, because measuring the degree of market segmentation within prefecture level cities requires complete county-level data, which obviously cannot be fully obtained. And like the use of prefecture level city and provincial data, the use of a single administrative unit will not be able to eliminate the noise of administrative divisions. To sum up, in order to eliminate the market segmentation effect caused by administrative divisions, this paper needs to construct a geographical unit that is not based on administrative regions, that is, a "synthetic virtual" economic activity unit. Therefore, the urban circle constructed in this paper meets the above basic conditions. There are dialect diversity and administrative division differences within the urban circle, but the urban circle itself is not a natural administrative division unit. In this way, the difference of market segmentation between urban circles will mainly come from the economic and cultural differences within the urban circle, rather than the administrative segmentation of the urban circle itself. Aiming at the problem of administrative division within the urban circle, this paper controls the number of prefecture level cities in the urban circle and the number of cross provincial administrative units in the regression equation to peel off the "administrative division effect" in the market segmentation, which is consistent with the idea of Liu et al. (2017) to distinguish the dialect effect from the system and policy effect in the resource mismatch.

The key explanatory variable of this paper is dialect diversity index. Similar to the method used by Xu et al. (2015) to measure the diversity of Chinese dialects, this paper measures the dialect diversity indicators including minority dialects. The diversity data of Chinese dialects comes from the Great Dictionary of Chinese dialects (Xu and Miyata, 1999). The data of minority dialects comes from the Chinese Language Atlas (Chinese Academy of Social Sciences, 2012). Based on the two, this paper sorts out the dialect diversity data of all administrative regions in China. Combined with the above two data sources, this paper divides the language level of China into four levels: dialect category → dialect category → dialect large → dialect

small. Among them, there are 9 kinds of dialects, namely, the language family to which the dialect belongs; there are 80 kinds of dialects; on the basis of dialect subclasses, dialect blockbusters are further divided into 173 types according to regional differences and language differences; there are 248 dialects, including Chinese dialects, minority dialects and sub dialects. In the benchmark model, dialect diversity is characterized by the number of dialect sub categories. For the sake of conservatism, this paper also provides empirical results of dialect categories, large dialects and small dialects.

In order to construct the dialect types of urban circle, this paper matches the dialect data of 1986 to the urban circle of 2016. Some counties and cities have changed their administrative divisions, such as renaming, revoking, merging, revoking counties into districts, revoking counties into cities, and establishing new counties. Finally, the dialect attributes of a total of 2318 county-level administrative units are matched to 332 urban circles. At the same time, in the construction of dialect indicators in urban circle, this paper also distinguishes between absolute indicators and relative indicators.

(1) Absolute indicators: slightly different from the Chinese dialect tree of Liu et al. (2015), this paper considers the linguistic diversity of all language families in China (except Hong Kong, Macao, Taiwan, Tibet and Hainan). As mentioned above, the absolute number of dialects in the urban circle is calculated according to "dialect category → dialect sub category → dialect large area → dialect small area". The descriptive statistical results are shown in **Table 1**.

Table 1. Descriptive statistics of dialect diversity in urban circle.

Variable name	Variable meaning	Observed value	Ave.	Min.	Max.	Standard deviation
dia1	Dialect category	332	213	1	5	134
dia2	Dialect subclass	332	466	1	19	333
dia3	Dialect blockbuster	332	773	1	24	455
dia4	Dialect fragment	332	943	1	37	58
city_ num	Number of administrative districts	332	634	2	14	181
div_ p	Dialect dispersion	332	043	0	083	023
div_ s	Dialect distance	332	059	0	15	04

Note: the data in this table are calculated by the author; the number of administrative districts of each urban circle includes the core city itself.

(2) Relative indicators: the regional differences of dialects are not only related to the types of languages used in the region, but also closely related to the number of language users and the degree of similarity. In order to consider the relative differences of languages, this paper uses the methods of Xu et al. (2015) and Liu et al. (2015) for reference to calculate the relative indicators of the diversity of the two dialects.

First, consider the diversity index of population differences in language use: dialect dispersion, expressed in, and the calculation div_pformula is: represents the proportion of $div_p_i = 1 - \sum_{j=1}^{N} S_{ji}^2$the S_{ji}population using dialect category J in urban circle I, and N represents the number of dialects in urban circle I.div_p The value of is 0–1. The greater the value, the greater the probability that people in the urban circle speak different dialects, that is, the higher the diversity of dialects.

Secondly, considering the diversity index of language similarity: dialect distance, expressed in, the calculation formula div_s is:, and is the proportion $div\_s_i = \sum_{j=1}^{J}\sum_{k=1}^{K} S_{ji} \times S_{ki} \times d_{jk}$ of S_{ji} dialect S_{ki} sub category J and dialect sub category K in urban circle I, and is the d_{jk} distance between the two dialect sub categories J and K. The indicators of dialect d_{jk} distance are as follows: compare the dialects with other dialects in the urban circle. If they are the same dialect category, the dialect difference is 0; if it belongs to different dialect subcategories under the same dialect category, it is 1; if it belongs to different dialect categories, it is 2. The div_s greater the value of, the greater the difference of languages in the urban circle, that is, the more diverse the dialects. See **Table 1** for the descriptive statistical results of the above relative indicators.

3.2. Market segmentation and other control variables

The explanatory variable of this paper is market segmentation index. Considering the inherent defects of the "production law", "Trade Law" and "professional index method" (Lu and Chen, 2009), this paper continues the idea of Gui et al. (2006) and uses the "price method" to build the 2016 urban circle market segmentation index. Generally speaking, when there is no flow barrier in the market, commodity and factor prices will gradually converge. Considering the transaction cost, the relative prices of the two places will only fluctuate in a reasonable range, rather than strictly tending to 1. At this point, the degree of market segmentation between the two places can be measured by the price volatility of the two places. This paper uses the price difference between adjacent prefecture level cities to calculate the degree of market segmentation, and then average the indicators of each core city and its adjacent prefecture level cities to calculate the average degree of market segmentation of the city circle (Please ask the author for a detailed description of the structure of the market segmentation index of the city circle.).

Based on the existing literature, this paper also includes a series of other control variables that may affect the degree of market segmentation. Specifically, Chen Min et al. (2007) believe that market segmentation is affected by the degree of government intervention. This paper uses the proportion of regional fiscal expenditure in GDP to measure the degree of government intervention. Fan and Zhang (2010) found that China's fiscal decentralization system promotes local governments to market segmentation. In this paper, the degree of decentralization is measured by the ratio of per capita fiscal expenditure in urban areas to national fiscal expenditure. Liu (2012) and Lu (2009) pointed out that the proportion of employees and openness of state-owned enterprises are positively correlated with the degree of market segmentation. Therefore, we will also consider the impact of these two factors. The openness is measured by the proportion of regional total import and export to regional GDP. At the same time, considering the impact of administrative division, this paper uses the number of prefecture level cities in each urban circle as the control variable. The market segmentation degree of urban circle is not only affected by government factors, but also closely related to geographical location. Therefore, this paper uses the longitude and latitude of the administrative center of the core city in the urban circle to express the geographical location differences.

4. Analysis of empirical results

4.1. Benchmark model

Table 2 reports the results of the baseline estimation of the regression equation (1). The estimation results in column (1) show that the estimation coefficient of the number of dialects is positive and significant at the level of 1%. Specifically, for each additional dialect category, the market segmentation will increase by 0.61% on average. Column (2) adds the government factors that affect the market segmentation, including the proportion of state-owned workers, opening to the outside world, government intervention and fiscal decentralization, and controls the attribute variables of the urban circle. The estimation results again show that the more dialects there are, the more serious the market segmentation is. The estimation coefficient rises to 106% and is still significantly different from zero at the 1% level. In addition, the control variable coefficient is consistent with the existing literature. The proportion coefficient of employees in state-owned enterprises is positive at the significance level of 1%, which indicates that the higher the proportion of state-owned enterprises in the region, the local governments are more motivated to carry out local protection and market segmentation out of the need for hidden subsidies to state-owned enterprises. The degree of opening to the outside world is positive at the significance level of 1%, which indicates that opening to the outside world may encourage local countries to sacrifice domestic trade and intensify domestic market segmentation (Chen et al., 2007). The estimation coefficients of government intervention, fiscal decentralization and the number of prefecture level cities in the urban circle are not significant. The possible reason is that the sample division of the urban circle constructed in this paper takes "whether it borders on the core cities" as the inclusion standard of local level cities. This method breaks the traditional division of administrative regions and weakens the direct impact of administrative power on market segmentation, which also confirms that the identification strategy of "drawing circles" to form urban circles can effectively reduce the noise brought by the administrative system. In column (3), longitude and latitude are further added, and geographic variables such as regional fixed effect and coastal or not are added (The core cities in coastal areas include Putian, Maoming, Qinhuangdao, Beihai, Zhuhai, Shanwei, Tianjin and other 45 prefecture level cities.). The results again show that the market segmentation effect brought by dialect diversity is significantly positive. The longitude estimation coefficient shows that the closer the relative position of the urban circle is to the East, the higher the degree of market integration is. The results are significant at the level of 1%.

It should be noted that the market segmentation index in columns (1)–(3) is measured according to the average of the market segmentation indexes of each group in the urban circle, and there may be errors in a single calculation method. Therefore, this paper further uses the GDP, population and administrative area of the urban circle to weight the market segmentation index (Area weighting calculation method: $nseg_{area} = corecity - city_i / \sum_{j=1}^{N}(corecity - city_j)$Among them, $corecity - city_j$ it represents the core city of the urban circle and its neighboring city I Area of. N represents the number of prefecture level cities bordering the core cities in the urban circle. Population and weighted calculation method are similar. The area, population

and GDP data of cities at various levels are from the 2016 China urban statistical yearbook.). As shown in columns (4)–(6) of **Table 2**, the key explanatory variables are still significant. In other words, after considering the possible measurement bias, the basic conclusion of this paper is still valid (In addition, due to the special administrative status of municipalities directly under the central government, the measurement of market segmentation in the urban circle may be biased (Gui et al., 2006; Zhao et al., 2009). Therefore, this paper also excluded the samples of municipalities directly under the central government and tested the robustness of the above basic conclusions again. The empirical results are consistent with the basic conclusions in **Table 2**. Limited by space, it is not presented in the text.). In general, the benchmark regression results in **Table 2** preliminarily verify the hypothesis of this paper: the more complex the types of dialects, the higher the degree of market segmentation of the urban circle, and the less conducive to the formation of market integration. Take column (3) as an example. If one dialect is added to the urban circle, the market segmentation will increase by about 11%. The average of dialect sub category indicators is 466, the minimum value is 1, and the maximum value is 19. This means that when other conditions remain unchanged, if the diversity of dialects within the urban circle is eliminated, the degree of market segmentation within the urban circle can be reduced by about 198% at most (Value 19 8% calculation method: every unit of dialect diversity decreases, the degree of market segmentation will decrease by 11%。 Given other conditions unchanged, if language diversity is eliminated, the degree of market segmentation can be reduced at most $1.1\% \times (19 - 1) = 19.8\%$).

Table 2. Benchmark regression: dialect diversity and market segmentation.

	Mean value			**GDP weighted**	**Population weighted**	**Area weighting**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Dialect subclass	0.0061*** (0.0020)	0.0106*** (0.0025)	0.0110*** (0.0028)	0.0111*** (0.0029)	0.0107*** (0.0029)	0.0080*** (0.0028)
Proportion of employees in state−owned enterprises		0.2621*** (0.0632)	0.1860*** (0.0693)	0.1742** (0.0725)	0.1935*** (0.0726)	0.1523** (0.0726)
Opening to the outside world		0.1364*** (0.0401)	0.0914** (0.0423)	0.1009** (0.0442)	0.1052** (0.0443)	0.0748* (0.0429)
Government intervention		−0.1163 (0.1172)	−0.0879 (0.1156)	−0.0887 (0.1210)	−0.1046 (0.1212)	−0.0709 (0.1173)
Fiscal Decentralization		0.0020 (0.0092)	0.0029 (0.0090)	0.0005 (0.0094)	0.0019 (0.0094)	0.0034 (0.0091)
Number of prefecture level cities		−0.0034 (0.0037)	−0.0008 (0.0038)	−0.0031 (0.0039)	−0.0027 (0.0039)	−0.0008 (0.0038)
Longitude			−0.0026** (0.0010)	−0.0029*** (0.0011)	−0.0028*** (0.0011)	−0.0028*** (0.0010)
Latitude			0.0008 (0.0011)	0.0011 (0.0012)	0.0008 (0.0012)	0.0002 (0.0012)
Provincial variables	No	Yes	Yes	Yes	Yes	Yes
Coastal or not	No	No	Yes	Yes	Yes	Yes
Number of samples	301	301	301	301	301	301
R^2	0.0296	0.1157	0.1553	0.1419	0.1462	0.1296

Note: a. The standard error of robustness of estimated coefficient in brackets, the same below; b. *, **, ***, respectively indicate that the variables are significant at the level of 1%, 5% and 10%, the same below; c. Limited by space, the definition, calculation method and data source of control variables are

not presented in the text; d. In the regression analysis, the core cities near the border and the samples of the virtual city circle formed around these core cities are deleted. Therefore, the sample size of the benchmark model is 301.

4.2. Robustness test

This section conducts a series of robustness tests based on **Table 2** to test the robustness of the estimation results.

(1) Language level: language attribution requires a large number of vocabulary, pronunciation, intonation and other language characteristics, but there are many language branches with different details. There are "degree" differences in the quantitative study of language. If only a single language standard is used as the dialect diversity index, there may be measurement errors. To test the robustness of the estimation results, columns (1)–(3) of **Table 3** respectively use "dialect category", "dialect large area" and "dialect small area" to replace the key explanatory variable "dialect sub category" in **Table 2**. The estimation results of the number of dialects classified according to different levels again show that the language diversity index is still significantly positive, and the estimation coefficient is about 06% to 18%.

(2) Relative indicators of language diversity: considering the influence of language use population weight and language similarity on (Calculation method for the population of each dialect in the urban circle: using the population data of each county in the 2016 China population and Employment Statistical Yearbook, match the dialect types of each county in the urban circle with the county population. If it is a single dialect County, all the county population will be used as the population of this dialect; if it is a multi-dialect County, the county population will be given to each dialect equally, and finally the population of various dialects in the urban circle will be counted.) language differences, this paper constructs an indicator "dialect dispersion" considering language use population differences and an indicator "dialect distance" considering language similarity. The regression results in columns (4)–(5) of **Table 3** show that the greater the probability of people using different languages, the higher the degree of dialect differentiation, and the more serious the market segmentation will be. This shows that even after considering the factors such as the population size and dialect similarity, the more complex dialect diversity will still significantly improve the degree of regional market segmentation. The estimated effect is increased to about 5%. The improvement of estimation coefficient results from the change of measurement index.

Table 3. Robustness test: different measurement methods based on the number of dialects.

	Dialect category	**Dialect blockbuster**	**Dialect fragment**	**Dialect dispersion**	**Dialect distance**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Linguistic diversity indicators	0.0179*** (0.0061)	0.0068*** (0.0021)	0.0064*** (0.0016)	0.0525* (0.0312)	0.0537** (0.0230)
Control variable	Yes	Yes	Yes	Yes	Yes
Number of samples	301	301	301	301	301
R^2	0.1359	0.1426	0.157	0.1187	0.1266

Note: a. The control variables are completely consistent with the benchmark model, the same below; b. Dialect levels mainly include: dialect category → dialect subclass → dialect large area → dialect small area. The results of dialect subclass have been reported in **Table 2**, and are not repeated here.

(3) Re measurement of market segmentation degree of urban circle: for the purpose of robustness test, this paper further considers the market segmentation degree between non-core cities in the calculation of existing market segmentation (We sincerely thank the anonymous reviewers for their valuable comments.). In fact, there is a high degree of consistency between the market segmentation degree of the urban circle obtained according to the above calculation method and the original market segmentation measurement indicators, and the correlation coefficient between the two is about 086. The regression results in **Table 4** show that, on average, the degree of market segmentation increases by about 1% to 12% for each increase in the number of dialect categories in the urban circle. Compared with the benchmark model (11%), there is no significant difference between the two. In general, the empirical results support the basic conclusions of this paper, and the estimated coefficient has good robustness.

Table 4. Robustness test: different measurement methods based on the degree of market segmentation.

	Mean value	**GDP weighted**	**Population weighted**	**Area weighting**
	(1)	**(2)**	**(3)**	**(4)**
Dialect subclass	0.0115***(0.0032)	0.0118***(0.0031)	0.0112***(0.0032)	0.0098***(0.0033)
Control variable	Yes	Yes	Yes	Yes
Number of samples	301	301	301	301
R^2	0.1754	0.1703	0.1661	0.1663

Note: the explanatory variable in this table is the index that considers the market segmentation degree between non-core cities in the urban circle.

(4) Potential impact of other characteristics in the urban circle: the difference in characteristics between the core cities in the urban circle and other cities will also affect the flow of resource elements between regions, and then have a potential impact on the market segmentation degree of the urban circle. Combined with the research background, we divide the control variables that need to be added into the following two categories: the first category, administrative division factors (We sincerely thank the anonymous reviewers for their valuable comments. Controlling the administrative division factors within the urban circle is the key to identify the causal effect of dialect diversity on market segmentation. We use the number of inter provincial administrative units in the urban circle, the number of inland level cities in the urban circle and the number of county (District) level units under the core city to describe the administrative segmentation within the urban circle. On average, the more administrative units (such as prefecture level cities) across the city circle, the higher the degree of administrative division. Indeed, considering the potential measurement bias, we also conducted a robustness test in the section of instrumental variables. The empirical results show that the estimated coefficients of dialect diversity variables do not fluctuate greatly, which indirectly proves that the above indicators can effectively measure the administrative zoning effect within the urban circle.). It includes the number of inter provincial administrative units in the city circle, the administrative area of the core city and the number of administrative units at the county (District) level under its jurisdiction. The second is the economic difference factor. The income gap of urban circle and the proportion of core city GDP in urban circle are used as

indicators to measure economic differences. Overall, the empirical results in **Table 5** further support the robustness of the basic conclusions. On average, the degree of market segmentation increases by about 0.9% to 12% for each increase in the number of dialect categories in the urban circle. It can be seen that after considering the potential impact of the differences in the characteristics of the cities in the above urban circle, the marginal impact of dialect diversity on market segmentation is still very close to the estimated results of the benchmark model.

5. Causal identification

Although the relative indicators of dialects and different measurement methods of market segmentation in the urban circle are considered in the above robustness test, the measurement results still have the possibility of estimation bias. In this section, we try to use three strategies to deal with endogenous problems. Firstly, the article further excludes the potential impact of geographical terrain, ethnic diversity and traffic conditions on the market segmentation of the urban circle. Secondly, using the method of Nunn and Wantchekon (2011) for reference, the effect of unobservable factors is estimated by using observed factors. Finally, we use "local opera types" as the instrumental variables of dialect diversity to further identify cause and effect.

Table 5. Robustness test: potential impact of other characteristics in the urban circle.

	Mean value	**GDP weighted**	**Population weighted**	**Area weighting**
	(1)	**(2)**	**(3)**	**(4)**
Dialect subclass	0.0116*** (0.0029)	0.0119*** (0.0031)	0.0115*** (0.0031)	0.0087*** (0.0030)
Number of cross provinces within the city circle	0.0260*** (0.0092)	0.0258*** (0.0096)	0.0263*** (0.0096)	0.0245*** (0.0093)
Income gap in urban circle	0.0113 (0.0238)	0.0004 (0.0250)	0.0056 (0.0250)	0.0057 (0.0242)
Proportion of core city GDP	−0.0707 (0.0662)	−0.0597 (0.0694)	−0.0539 (0.0694)	−0.0819 (0.0670)
Counties under core cities	−0.0007	−0.0013	−0.0015	−0.0013
Quantity (area)	(0.0015)	(0.0016)	(0.0016)	(0.0016)
Administrative division of core cities	0.0000	−0.0000	−0.0000	0.0000
The measure of area	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Control variable	Yes	Yes	Yes	Yes
Number of samples	301	301	301	301
R^2	0.2009	0.1838	0.1906	0.1784

5.1. Possible missing variables

In order to identify the influence of dialect diversity itself more "cleanly", the possible missing variables are considered from the following three aspects: geographical factors, ethnic diversity factors and traffic convenience.

(1) Geographical factors. The formation of dialects is closely related to the terrain. There has always been a saying among the people that "ten miles of different sounds, hundreds of miles of different words", and the complex terrain will bring geographical barriers, thus increasing people's transaction costs. **Table 2** the positive correlation between language diversity and the degree of market segmentation in regression is likely to come from geographical isolation rather than the influence of dialect diversity

itself. In order to rule out this possibility, we control the geographical factors. This paper adds the average altitude difference of the urban circle to represent the (The average altitude difference of the urban circle is constructed as the mean of the altitude difference between the core city and its adjacent cities.) geographical factors. When the altitude difference of a region is larger, it is easier to form geographical isolation. This isolation is not only reflected in the diversity of language, but also in the hindrance in the process of economic exchanges.

Column (1) of **Table 6** reports the estimated results of controlling geographical factors. After controlling the average altitude of the area, the coefficient of the average altitude difference is significantly positive. This shows that geographical isolation will indeed have a negative impact on market integration. However, even if geographical factors are controlled, the estimated effect of dialect diversity is still significantly positive, and the estimated coefficient is 123%, slightly increasing. It should be noted that this paper also uses whether the cities in the urban circle belong to the same topographic area and the same main watershed as the proxy variable to measure the geographical characteristics of the urban circle. The empirical results also support the above conclusions (We sincerely thank the anonymous reviewers for their comments. Due to space limitations, the empirical results are not presented in the text.).

Table 6. Possible missing variables: terrain, ethnic composition, transportation infrastructure.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Dialect subclass	0.0123*** (0.0028)	0.0066** (0.0032)	0.0144*** (0.0043)	0.0105*** (0.0028)	0.0096*** (0.0033)
Average altitude difference	0.0092** (0.0041)				0.0089** (0.0041)
Number of ethnic minorities		0.0070*** (0.0024)			0.0029 (0.0024)
Grade highway per capita				−0.0199** (0.0088)	−0.0114 (0.0090)
Control variable	Yes	Yes	Yes	Yes	Yes
Number of samples	301	301	217	301	301
R^2	0.1835	0.1790	0.1655	0.1700	0.1730

Note: the altitude data comes from the “contour map” online query website, and the link is http:// haiba qhdi. Com.

(2) Ethnic diversity. China is a multi-ethnic country. Ethnic diversity has created language diversity. The distribution of dialects is affected by ethnic migration, mobility and integration. At the same time, the distribution of ethnic minorities in China is characterized by “large-scale mixed living and small-scale settlement”, which is generally dominated by agricultural and handicraft industries. The market economy is underdeveloped, and the economic development status is quite different from that of the Han inhabited areas. This means that market segmentation does not necessarily come from linguistic diversity, but may also come from ethnic diversity. Column (2) of table 6 controls the number of ethnic groups with more than 10,000 people in the urban circle. Compared with the benchmark results in **Table 2**, the coefficient of dialect diversity decreased to 0.66%, but it is still significantly positive. The estimated coefficient of ethnic diversity is significantly positive, and the estimated effect is 0.7%. The dialect effect in the urban circle is almost the same as that of ethnic minorities. Further, in column (3) of **Table 6**, we exclude the sample of urban circle where the core city is inhabited by ethnic minorities. The estimated coefficient of key

explanatory variables is consistent with the expectation and has increased significantly. This shows that although ethnic diversity improves the degree of regional market segmentation, the effect of dialect diversity on market segmentation is still significant after controlling ethnic diversity.

(3) Traffic infrastructure impact. The negative impact of dialect diversity on market integration may be due to the backwardness of transportation infrastructure caused by geographical blockade, resulting in regional market segmentation. Therefore, the impact of traffic convenience is considered in column (4) of **Table 6**. After controlling the average mileage of regional grade roads per capita, the dialect diversity coefficient is almost unchanged and still significantly positive. The traffic convenience index is significantly negative, which shows that although the traffic infrastructure can promote the rapid integration of language and culture and promote the development of market integration, this promotion can not completely eliminate the blocking effect of dialect diversity on the integrated market. It should be noted that we also searched the "Baidu map" for the most convenient traffic route mileage between the two cities in the urban circle, and counted the average mileage between other cities in the circle and the core cities, so as to measure the traffic condition of each urban circle. The empirical results also support the basic conclusions of this paper.

In column (5) of **Table 6**, we control geographical factors, ethnic diversity and traffic factors at the same time. The empirical results show that the estimated coefficient of dialect diversity is still significantly positive. Based on the above estimation results, after adding the possible missing variables, the impact of dialect diversity on the degree of market segmentation is still robust.

5.2. Influence of unobservable factors

Are there unknown or unmeasurable missing variables that cause bias in the above regression results? For the sake of robustness, we need to further analyze the unmeasurable factors that may be omitted. According to Altonji et al. (2005) to estimate the effect of unobservable factors by using the coefficient changes estimated by the controlled observable factors.

Specifically, the following two regressions are considered: one is the regression with only constrained control variables, and the other is the regression equation with all control variables; the estimated coefficient in the first regression is recorded as (r stands for constrained), $\beta^R$and the estimated coefficient in the second regression is recorded as (f stands $\beta^F$for all), and the ratio is $|\beta^F/(\beta^R-\beta^F)|$calculated according to the formula:. The meaning of the formula is very intuitive. First of all, $\beta^R-\beta^F$the smaller the β^Rdenominator β^Fvalue, that is, the closer the sum value is, it means that after controlling all the observable factors, the change of the estimated coefficient is very small compared with that before control, which means that the change of the estimated coefficient is very limited by adding the known control variables, which also means that the influence of the unobservable factors should be much greater than the factors we have controlled in order to make the estimated coefficient produce large errors. Secondly, the larger the molecule, the greater the effect of unobservable factors. In conclusion, $|\beta^F/(\beta^R-\beta^F)|$ the greater the value of, the less likely the unobservable factors will have an impact on the regression results.

Similar to Nunn and Wantchekon (2011), this paper considers two groups of constrained control variables: one group does not add any control variables, and the other group adds the control of the basic characteristics of the urban circle, including only the per capita GDP and the number of administrative districts of the circle. In addition, we also consider two groups of full control variables: the first group of full control variables is consistent with the benchmark model, and the second group adds possible missing variables based on the first group, including altitude, ethnic diversity and transportation convenience. These four groups of regression coefficients are classified according to the constrained variable group and the fully controlled variable group, and the two combinations are used to calculate the ratio value. The estimation results are reported in **Table 7**.

Table 7. Using observable factors to evaluate the impact of unobservable factors.

Constrained control group	Full control variable group	Mean value	GDP weighted	Population weighted	Area weighting
No control variable	All control variables of benchmark regression	2.24	2.02	2.1	2.42
No control variable	All control variables of baseline regression, altitude, highway, ethnic diversity	2.74	2.33	2.51	3.24
Per capita GDP, number of districts	All control variables of benchmark regression	2.89	2.71	2.89	3.81
Per capita GDP, number of districts	All control variables of baseline regression, altitude, highway, ethnic diversity	4	3.5	4.04	7.56

None of the 16 ratios reported in **Table 7** is less than 1 (When the ratio value is greater than 1, that is, the required unobservable factor is more than 1 times of the observable factor. At this time, the estimation effect is not affected by the unobservable factor (Altonji et al., 2005; Nunn and Wantchekon, 2011).). The ratio range is 202–756 with an average of 322. This means that if the missing unobservable factors are to make the regression results seriously biased, the impact of unobservable factors required is at least 202 times that of the observable factors that have been controlled, and on average, more than 322 times. Obviously, according to the above calculation, this paper believes that the estimated effect of dialect diversity is unlikely to have about three times the influence of unobservable factors.

5.3. Tool variable method

In this section, tool variables are further used for causal identification. In this paper, the number of local operas is used as the instrumental variable and the two-stage least square regression is carried out.

Dialect is the most prominent feature of local cultural diversity. A local drama, folk art, ballad, riddle and other literary and artistic forms can be expressed only with dialect as a tool. The relationship between local operas and dialects is very close. The variety of dialects is one of the main reasons for the enrichment and diversification of local operas (You and Zhou, 1985). Because local operas are sung in dialect, and their audience is mostly limited to the dialect area, the number of local operas is largely affected by the diversity of dialects. At the same time, there is no obvious correlation between the number of local operas and the degree of market segmentation. Therefore, the number of local operas is an ideal choice of instrumental variables. According to

the distribution of Chinese operas, this paper calculates the number of local operas in each city circle (The distribution of operas comes from the manual of Chinese operas.).

In terms of the correlation between instrumental variables and endogenous variables (the number of dialect categories), the first stage regression results in **Table 8** report the F statistic. It can be seen that the F statistic is greater than 10, and the original assumption of "weak instrumental variables" is rejected according to the rule of thumb, which means that this paper will not face the problem of weak instrumental variables when using "the number of local opera categories" as the instrumental variable for two-stage estimation.

In terms of the exogenous nature of the instrumental variables, in order to verify their effectiveness, the instrumental variables are included in Equation (1) in **Table 8** for testing. The estimated results show that the coefficient of local opera types is not significant, which indicates that there is no correlation between local opera types and market segmentation, and provides indirect supporting evidence for the exogenous conditions of this instrumental variable. However, it is generally accepted that it is impossible to directly test the exclusive constraints of instrumental variables. Whether instrumental variables will affect the explained variables through other channels depends on the corresponding qualitative discussion to exclude them one by one. Based on this, we further test the exogenous conditions of the instrumental variables used in this paper from the following two aspects.

Table 8. Estimation of instrumental variables.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
	OLS	**Two stage estimation results**			**Approximate exogenous tool variables: ltz method**		
	Full sample	**Full sample**	**Excluding minority language areas**	**Exclude samples including municipalities directly under the central government**	**Full sample**	**Excluding minority language areas**	**Exclude samples including municipalities directly under the central government**
Dialect subclass	0.0107*** (0.003)	0.0242* (0.015)	0.0225* (0.013)	0.0263* (0.016)	0.0229* (0.0128)	0.0214** (0.0108)	0.0250* (0.0130)
Number of cross−cultural areas in the city circle	0.0030 (0.008)	−0.0081 (0.011)	−0.0055 (0.007)	0.0046 (0.008)	−0.0075 (0.057)	−0.0054 (0.010)	0.0044 (0.020)
Types of local operas	−0.0006 (0.001)						
Income gap in urban circle	0.0022 (0.023)	−0.0252 (0.026)	−0.0225 (0.018)	0.0065 (0.019)			
Proportion of core city GDP	−0.0783 (0.062)	−0.2011** (0.081)	−0.0894* (0.054)	−0.0530 (0.052)			
Control variable	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of samples	301	301	217	278	301	217	278
		Phase I estimation results					
		Explained variable: number of dialect sub categories in urban circle					
Types of local		0.0789***	0.0937***	0.0778*** (0.023)			

operas		(0.021)	(0.020)	
Control variable	Yes	Yes	Yes	Yes
Number of samples	301	301	217	278
R−squared	0.1565	0.6112	0.6819	0.5852
F−statistic	—	16.38	21.04	13.96

On the one hand, local operas should be excluded from influencing market segmentation through other non-dialect channels. Local operas are easily influenced by regional economic and cultural factors, which may also affect market segmentation. If the above factors (economic and cultural factors) are left in the disturbance item, it may cause the disturbance item to be related to the types of operas, so that the exogenous cannot be satisfied. Empirically, we control a series of economic and cultural factors in each urban circle in the two-stage regression, so as to block the correlation of opera types through cultural, economic factors and disturbance items.

On the other hand, based on Conley et al. (2012) in the empirical framework of IV estimation under the condition of loosening the exogenous instrumental variables, we re tested the robustness of the estimation results. Traditionally, the test of the exogenous conditions of instrumental variables relies on qualitative discussion. After finding out all other possible channels through which instrumental variables affect the explained variables, they are excluded one by one. The difference is that Conley et al. (2012) assume that the instrumental variables are close to the approximate exogenous, so as to investigate the change trend of the estimation coefficient of endogenous variables under different degrees of exogenous approximation. Lin and Zhao (2017) earlier used this method to test the robustness of their instrumental variable estimation results in their research on dialect and technology diffusion (The strict validity of the exogenous conditions of the instrumental variables can also be qualitatively discussed. In cultural geography, language is often regarded as one of the important representations of culture, and language differences are also the direct embodiment of cultural differences. For example, Zhou (2013) believes that the differences in geographical distribution of culture are mainly manifested in language and religion. Unlike Europe, there is no obvious difference in the geographical influence of religion in China, so cultural geography is mainly manifested in dialect geography. In recent empirical studies of cultural economics, most literatures often use the number of dialects as the proxy variable of cultural diversity in the region (Xu et al., 2015; Liu et al., 2015; Gao and Long, 2016; Zhao and Lin, 2017). Given the above conclusions, the number of dialects used in this paper can be regarded as measuring the language differences of each urban circle, as well as the cultural differences of each urban circle. Therefore, the use of "local opera types" as the instrumental variable of dialect quantity can be regarded as the use of "local opera types" as the instrumental variable of regional cultural diversity. Therefore, the argument that the instrumental variable "local opera types" affect market segmentation through other cultural factors is no longer logical. In fact, the empirical results in **Table 8** also support the above conclusion in terms of empirical evidence. After controlling the general cultural characteristics of the urban circle, compared with the estimation results of the benchmark instrumental variables (0.023–0.026), the estimation coefficients of the

endogenous variables are almost the same. This means that in the benchmark regression model, other cultural factors that may exist are not "omitted" in the disturbance term.).

Columns (2)–(4) of **Table 8** report the conventional two-stage estimation results. Column (2) is full sample regression. The first stage estimation results show that there is a high correlation between the number of local opera types and the diversity of dialects. The second stage estimation coefficient is 242%, which is about twice the OLS estimation results. The result means that the market segmentation degree of the city circle with the average number of dialect types is 823% higher than that of the city circle with a single dialect area. The market segmentation degree of the urban circle with the number of dialect sub categories in the top 5% is about 2952% higher than that of the latter 5%. The estimated results are highly significant both economically and statistically. Column (3) and column (4) are the regression of excluding minority language areas and excluding the samples of urban circle including municipalities directly under the central government, and the estimated coefficient is still significantly positive. Columns (5)–(7) are Conley et al The estimation results under the local approximate zero method (ltz) proposed by (2012) are still robust in the case of approximate exogenous instrumental variables.

6. Concluding comments

This paper reinterprets the integration of China's domestic market from the perspective of dialect diversity. In the traditional research of domestic market integration and regional market segmentation, the research on administrative decentralization, registered residence, opening to the outside world and other factors has achieved fruitful research results. However, considering the wide diversity of Chinese society, economic development shows a high degree of imbalance and inadequacy. Cultural diversity and dialect diversity still have an important impact on the formation of market integration. Due to the interweaving of administrative divisions and dialect diversity in the historical development, it is difficult to identify the exact cause and effect. This overlap means that empirical studies tend to overestimate the administrative division effect in market segmentation and underestimate the influence of cultural diversity factors such as dialect differences.

In order to eliminate the noise caused by the overlap of administrative areas and dialect areas, this paper breaks the traditional administrative division, and takes a city as the core, and divides the neighboring cities into a city circle. Through such a city circle structure, we can effectively break through the definition of the existing administrative division. In a delineated urban circle, the 2016 consumer goods market index is used to measure the degree of market segmentation of the urban circle, and different levels of language diversity are used to measure the impact of dialect diversity on market integration. The result of the benchmark model shows that the dialect diversity in the urban circle has a negative impact on the formation of the market integration of the urban circle. Learn from Altonji et al (2005) and Nunn and Wantchekon (2011), this paper calculates the influence effect of possible missing variables. The results show that changing the significance of the estimated results in this paper requires about three times the influence effect of the existing control

variables, which further proves the robustness of the conclusions of this paper. Finally, the paper further uses the instrumental variable estimation, considering the characteristics that local operas are mainly performed in local dialects and face the local dialect people, this paper uses the types of local operas as the instrumental variables of dialects. The results show that the effect of dialect diversity on market segmentation rises to 242%. According to this estimation result, compared with the city circle with a single dialect area, the market segmentation degree of the city circle with an average number of dialect types has increased by nearly 823%, and the market segmentation degree of the city circle with the number of dialect types in the top 5% is about 2952% higher than that of the city circle with the last 5%.

The conclusions of this paper have important implications for the construction of an integrated domestic market and urban pattern dominated by urban agglomerations. The shift from high-speed growth to high-quality development requires the establishment of an integrated domestic market in which factors can flow freely. An integrated market needs to take into account the unity and diversity of cultures. It is of great significance for the construction of the integrated market to strengthen the cultural communication and exchange between regions and break the local departmentalism.

Author contributions: Conceptualization, CD and JZ; methodology, LY; software, LZ; validation, CD, JZ and LY; formal analysis, LZ; investigation, CD; resources, JZ; data curation, LY; writing—original draft preparation, LZ; writing—review and editing, CD; visualization, JZ; supervision, LY; project administration, LZ; funding acquisition, CD. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Sapir E. Sapir on language, culture and personality. Commercial press; 2011.
2. Chen M, Gui Q, Lu M, et al. Economic opening and domestic market integration. China: Linking Markets for Growth. 2007. doi: 10.22459/clmg.08.2007.18
3. Chen G, and Li S. Judicial independence and market segmentation—A Study on the exchange of judges in different places as an experiment. Economic research. 2013; 9.
4. Dai Y, Xiao J, Pan Y. Can “local accent” reduce the agency cost of the company?—a study from the perspective of dialect. Economic research. 2016; 12.
5. Fan Z, Zhang J. Fiscal decentralization, transfer payment and domestic market integration. Economic research. 2010; 3.
6. Gui Q, chen M, Lu M. China’s domestic commodity market tends to be divided or integrated: Analysis Based on the relative price method. World economy. 2006; 2.
7. Gao X, Long X. Will the cultural division caused by provincial administrative divisions affect the regional economy. Economics (quarterly). 2016; 2.
8. Huang J, Liu C. Dialect and social trust. Financial research. 2017; 7.
9. Lin Y, Liu P. Local protection and market segmentation: from the perspective of development strategy. Working paper of China Economic Research Center, peking University; 2004.
10. Lu M, Chen Z. Market integration and industrial agglomeration in China’s regional economic development. Shanghai People’s publishing house; 2006.
11. Lu M, Chen Z. Economic growth by market segmentation why is economic opening likely to intensify local protection. Economic research. 2009; 3.

12. Li G, Cao J, Shao S. language diversity and regional differences in China's opening up. World economy. 2017; 3.
13. Li Q, Meng L. Dialects, putonghua and regional labor mobility in China. Journal of economics. 2014; 4.
14. LiuY, Xu X, Xiao Z. The inverted U-shaped model of labor flow across dialects. Economic research. 2015; 10.
15. LiuY, Dai T, Xu X. Chinese dialects, market segmentation and resource mismatch. Economics (quarterly). 2017; 4.
16. Liu R. State owned enterprises, implicit subsidies and market segmentation: theoretical and empirical evidence. Management world. 2012; 4.
17. Lin J, Zhao Z. Invisible barriers to balanced development: dialect, system and technology diffusion. Economic research. 2017; 9.
18. Pan Y, Xiao J Dai Y. Cultural diversity and enterprise innovation: a study from the perspective of dialect. Financial research. 2017; 10.
19. Yin W, Cai W. Causes and governance of local market segmentation in China. Economic research. 2017; 6.
20. Xu X, Liu Y, Xiao Z. Dialect and economic growth. Journal of economics. 2015; 2.
21. You R, Zhou Z. Dialects and Chinese Culture. Fudan Journal (SOCIAL SCIENCE EDITION). 1985; 3.
22. Chinese Academy of Social Sciences. Atlas of Chinese language. Commercial press; 2012.
23. Zhang J, Zhang P, Huang T. Does Market Segmentation Promote the Export of Chinese Enterprises. Economic research. 2010; 8.
24. Zhang H, Feng Y. Cultural Typology and Analysis of Chinese Regional Cultural Psychological Types. Yinshan Journal (SOCIAL SCIENCE EDITION). 2010; 1.
25. Zheng Y, Li G. Efficiency Loss of Local Segmentation in China. China Social Sciences. 2003; 1.
26. Zhou L. Incentive and cooperation of government officials in the promotion game—Also on the reasons for the long-standing problems of local protectionism and redundant construction in China. Economic research. 2004; 6.
27. Zhou Z. Transformation of Paradigms—The Process of Evolving Geopolitical and Political Geography in Geographical Administrative Regions. Journal Of Central China Normal University. 2013; 1.
28. Zhao Q, Xiong X. Comparative Analysis of The Segmentation Degree of China's Three Major Markets: Time Trend and Regional Differences. World Economy. 2009; 6.
29. Zhao Z, Lin J. Cultural Hypothesis of Economic Development Gap: From Gene to Language. Management World. 2017; 1.
30. Alesina A, Giuliano P. Culture and Institutions. Journal Of Economic Literature. 2015; 53(4): 898-944. doi: 10.1257/Jel.53.4.898
31. Alesina A, Harnoss J, Rapoport H. Birthplace Diversity and Economic Prosperity. Journal of Economic Growth. 2016; 21(2): 101-138. doi: 10.1007/S10887-016-9127-6
32. Alesina A, La Ferrara E. Participation in Heterogeneous Communities*. Quarterly Journal of Economics. 2000; 115(3): 847-904. doi: 10.1162/003355300554935
33. Alesina A, E. La Ferrara E. Who Trusts Others? Journal of Public Economics. 2002; 85(2): 207-234. doi: 10.1016/S0047-2727(01)00084-6
34. Altonji JG, Elder TE, Taber CR. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. Journal of Political Economy. 2005; 113(1): 151-184. doi: 10.1086/426036
35. Ashraf Q, Galor O. The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development. American Economic Review. 2013; 103(1): 1-46. doi: 10.1257/aer.103.1.1
36. Berliant M, Fujita M. Knowledge Creation as A Square Dance on The Hilbert Cube*. International Economic Review. 2008; 49(4): 1251-1295. doi: 10.1111/j.1468-2354.2008.00512.x
37. Chen Z, Lu M, Xu L. Returns to dialect. China Economic Review. 2014; 30: 27-43. doi: 10.1016/j.chieco.2014.05.006
38. Conley TG, Hansen CB, Rossi PE. Plausibly Exogenous. Review of Economics and Statistics. 2012; 94(1): 260-272. doi: 10.1162/REST_a_00139
39. Gao X, Long CX. Cultural border, administrative border, and regional economic development: Evidence from Chinese cities. China Economic Review. 2014; 31: 247-264. doi: 10.1016/j.chieco.2014.10.002
40. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology. 2001; 27(1): 415-444. doi: 10.1146/annurev.soc.27.1.415

41. Nunn N, Wantchekon L. The Slave Trade and the Origins of Mistrust in Africa. American Economic Review. 2011; 101(7): 3221-3252. doi: 10.1257/aer.101.7.3221
42. Pendakur K, Pendakur R. Language as Both Human Capital and Ethnicity. International Migration Review. 2002; 36(1): 147-177. doi: 10.1111/j.1747-7379.2002.tb00075.x
43. Trax M, Brunow S, Suedekum J. Cultural diversity and plant-level productivity. Regional Science and Urban Economics. 2015; 53: 85-96. doi: 10.1016/j.regsciurbeco.2015.05.004