*City Diversity Research* (2022) Volume 1 Issue 1, *pp.*1–10.
doi:10.24294/cd.v3i1.1951

# Will cultural diversity block the process of urbanization?

# ——- Empirical Study from the perspective of dialect

Shaoshuai, [1] liguangqin [1,2], caojianhua [1]
(1 School of urban and Regional Sciences, institute of Finance and economics, shanghai University of Finance and economics, shanghai 200433; 2 Jiyang college, zhejiang agriculture and Forestry University, hangzhou, zhejiang 311300)

**Abstract**: Based on the data samples of 276 cities at prefecture level and above in China from 2000 to 2012, using dialect diversity as a proxy to measure cultural diversity, using random effect model, system generalized moment estimation, two-stage least square method and other methods, this paper conducted an empirical investigation on the impact of cultural diversity on China's urbanization for the first time. It is found that dialect diversity has a significant negative impact on urbanization rate; considering the possibility of missing variables, the influence of dialect diversity on urbanization rate is still significantly negative; after using the historical immigration as the instrumental variable of dialect diversity, this negative influence still exists, but the degree of influence has decreased. Therefore, the cultural variables represented by dialects are an important factor affecting the process of urbanization.

**Keywords**: Cultural diversity; Dialect; urbanization; Tool variables; Mass migration

CLC classification No.: F29921 Docment identification code: A

Article No.: 1671-511x (2017) 05-0122-10

DOI: 10.13916/j.cnki. Issn1671-511x. 2017.05.014

## 1. Introduction

According to the data released by the National Bureau of statistics, china's urbanization rate has reached 561% in 2015. However, how to develop "new urbanization" and how to realize the "people-oriented" urbanization required by the plan are still in the stage of practice and exploration. Obviously, it is very important and necessary to clarify the root causes of the lagging development of urbanization in China. Although existing studies have explored the causes of China's lagging urbanization development from the economic, social and other dimensions, no literature has been found to investigate the root causes of China's lagging urbanization from the

[Receiving date]2017-06-10

[Fund project] one of the achievements of the National Natural Science Foundation Project (71373153, 71773075), the Shanghai dawn plan project (14SG32), and the key humanities and Social Sciences project of Zhejiang Provincial Universities (2014QN027)

[About the author] Shaoshuai, male, doctor of management, Professor of Institute of Finance and economics, Shanghai University of Finance and economics, research direction: Urban and regional economics

perspective of cultural diversity. China is a large country with a vast territory and many nationalities. There are great cultural differences and urbanization levels among different regions. Culture is the most fundamental ideological source rooted in a region and a nation, and urbanization is not only a process of continuous agglomeration of population, industry and economy, but also a process of mutual collision, exchange and integration between different cultures. Will a higher degree of cultural diversity produce greater "friction" to the urbanization process due to the increase of integration resistance between different cultures? This problem has not attracted the attention of existing research.

The research on the influencing factors of urbanization can be traced back to pandey[1]. Based on the cross-sectional data of Indian states, he found that industrialization has a positive impact on the urbanization rate, and the degree of crop planting has a negative impact on the urbanization rate, but the impact of economic development measured by average wage on the urbanization rate is not significant. Chang et al. [2] reached a similar conclusion in their research on China's urbanization rate. Moomaw et al. [3] investigated the impact of a series of relevant factors on urbanization, and the results showed that the per capita GDP, high industrialization level and high urbanization rate in export-oriented areas. Domestic scholars have also investigated the determinants of urbanization from the perspectives of land cost and manufacturing development level [4], economic development level and population migration [5], housing problems of new citizens [6], industrial structure and human capital [7-8], lagging reform of state-owned enterprises [9]. However, as Wan Guanghua and others [7] pointed out, the determinants of urbanization are often affected by urbanization and have a reverse causal relationship. However, the existing studies do not pay enough attention to endogenous problems, resulting in most of the current research conclusions can only prove that there is a correlation between relevant factors and urbanization, but can not accurately identify the causal relationship between these factors and urbanization. In addition, the current research is to investigate the root causes of regional differences in urbanization from an economic perspective, while ignoring the possible important impact of cultural factors on the process of urbanization.

This paper is the first empirical study on whether cultural diversity has a blocking effect on China's urbanization. We believe that the regional differences in urbanization rate are not only affected by some economic factors concerned by the existing research, but also affected by the regional internal cultural differences, which can be characterized by the diversity of dialects [10]. In other words, the cultural diversity represented by dialect diversity may have an impact on China's urbanization process that can not be ignored. In addition, we use the instrumental variables constructed by the index of population migration to identify the causal relationship between dialect diversity and urbanization, which can ensure the robustness of the research results to a great extent.

## 2. Hypothesis proposal and empirical strategy

## 2.1 Proposal of hypothesis

The traditional economic growth theory only regards labor, capital and other input factors as the source of economic growth, while the modern economic theory increasingly pays attention to the cultural factors behind the traditional economic growth factors. With the increasingly prominent role of cultural factors in the process of modern economic development, some foreign scholars have discussed the impact of culture on economic growth from the perspective of cultural factors such as religion, system and innovation, and some domestic scholars have also discussed the impact of culture on economic growth from the perspective of East Asian culture or Confucian culture. For example, gao Bo and zhangzhipeng pointed out that cultural capital is a key factor of production and an important explanatory variable that determines economic growth [11]. Further, based on the theory of cultural cost and cultural change, gao Bo analyzed the reasons for China's economic stagnation and economic growth since modern times, and believed that culture created conditions for institutional innovation and technological innovation, thus promoting economic growth [12]. Jiang Li [13] found that regional cultural differences can explain the differences in regional economic development through theoretical mechanism analysis and model deduction [13]. It can be found from the above literature that culture has a very important impact on economic growth, and economic growth and urbanization can be regarded as two important dimensions synchronously related in the process of China's economic development to a large extent [14]. It can be seen that since culture has an important impact on economic growth, it should also have an impact on urbanization that can not be ignored.

According to the existing literature, the cultural differences between regions in China can be characterized by language differences. It is reasonable for existing studies to measure cultural diversity through dialect diversity [10,15]. In essence, cities are places where people exchange information intensively, and the diversity of dialects represents the degree of language differences in a region. Cities naturally become places where various dialects communicate and collide with each other. Therefore, the process of urbanization is also accompanied by the process of mutual integration of various languages, which is bound to have an important impact on the development speed of urbanization. If there are more dialects in a region, the resistance of people to communicate with each other in the city may be greater, and the performance of communication will be reduced accordingly, which will lead to the difficulty of potential urban migrants to integrate into the urban environment culturally and reduce their willingness to migrate, which is not conducive to the rapid promotion of urbanization. To sum up, this paper puts forward the following hypotheses:

Under the condition that other conditions remain unchanged, the higher the dialect diversity of a region, the lower the urbanization rate.

## 2.2 Empirical strategies

According to the theoretical hypothesis, we build the following regression model to empirically test the hypothesis:

$$Urban_{it} = \alpha + \beta^{*}\ Div_{it} + \gamma^{*}\ X_{it} + \mu_{i} + \delta_{t} + \xi_{it} \qquad (1)$$

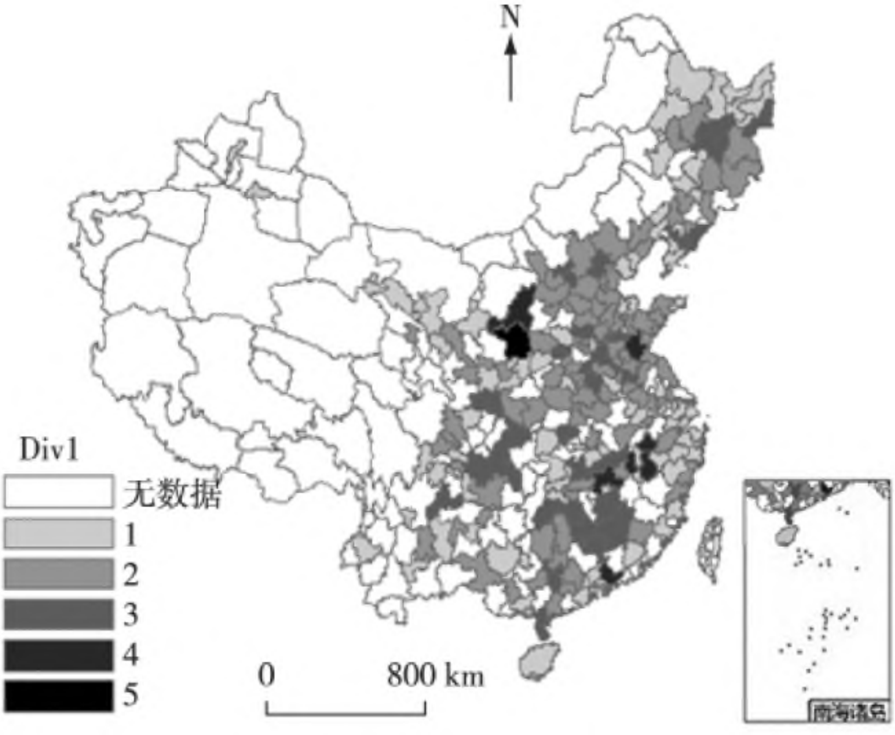
Where, i and t represent city and year respectively; urban represents urbanization rate; div stands for dialect diversity index; x represents a group of factors that affect the urbanization rate, including the degree of opening to the outside world, economic development level, financial self-sufficiency, human capital level, investment rate, industrial structure and other control variables; μ Indicates regional fixed effect; δ Indicates time fixed effect; ξ is random disturbance term; α, β, γ is the parameter to be estimated, where β For the core parameters concerned in this article, if $\beta<0$ and statistically significant, indicating that the hypothesis proposed in this paper is valid.

The first task of this paper is to select the urbanization index reasonably. Existing studies usually use the following two data sources when measuring urbanization at the urban level in China: the most common is the statistical data on non-agricultural population in the China Urban Statistical Yearbook, but this data only counts the registered residence population, and does not consider the urban permanent population in the rural floating population, so there is an obvious deviation, and it is only reported until 2008; another data source is the population statistics of the people's Republic of China by county and city issued by the Ministry of public security every year. This statistical data accurately reports the basic information of the registered temporary population and registered residence population nationwide, including the comprehensive information of the registered residence population and most of the information of the temporary population. The latest data of this data can be found until 2012. Obviously, the latter measures the urbanization rate more accurately and has higher credibility. Therefore, in this paper, the ratio of urban resident population to total population derived from the latter will be valued as real_urban, which will be taken as the core explained variable, while the ratio of non-agricultural population derived from the former will be valued as nominal urbanization rate (nomi_urban), which will be used in the robustness test as a substitute variable for the explained variable. According to the availability of data, the time span of the main research samples in this paper is 2000-2012, while the time span of the data samples for the robustness test is 2000-2008.

The cultural diversity measured by dialect diversity is the core explanatory variable of this paper. The dialect diversity index in this paper comes from the dialect data in the dialect database constructed by xuxianxiang et al. [10] and liuyuyun et [2]al. [15]. We briefly explain it as follows. Based on the administrative divisions in 1986, the dictionary of Chinese dialects makes a statistical survey of the Chinese dialects in various counties and cities in China. It shows that there are 17 dialects and 105 sub dialects in China. Based on the current cities at prefecture level and above as research samples, we find the number of dialects and sub dialects in all districts and counties of the city as the dialect diversity index (div1). The database matches 277 prefecture level cities, but Chaohu City has adjusted its administrative division in 2011, so it is excluded.

[2] The original data comes from the atlas of Chinese language and the dictionary of Chinese dialects.

The final research sample is 276 cities. Figure 1 shows the regional distribution of dialect diversity index in China. In addition, we also constructed another dialect diversity index (div2) for robustness test by using the population proportion of dialects (which will be explained later). Because the dialect diversity index is the cross-sectional data that does not change with time, and the explanatory variable of this paper is the panel data, if the fixed effect model is used to estimate, the dialect diversity will be automatically eliminated. Therefore, in the benchmark empirical analysis, we mainly use the random effect model and the system generalized moment estimation method to estimate the parameters.



| |
|---|
| No data |
| South China Sea Islands |

Fig. 1 Regional distribution of dialect diversity index in China

Source: the author uses arcgis software to exchange according to the dialect database.

A series of control factors selected in this paper are described as follows.

(1) Degree of opening to the outside world (FDI): a measure of the proportion of actually utilized foreign capital (FDI) in GDP converted into RMB at the average exchange rate. The higher the opening-up of a region, the more employment opportunities there will be in the city, and the urbanization rate will be relatively high. Therefore, it is expected that the coefficient sign will be positive. (2) Economic development level (PGDP): measured by per capita GDP. We convert the per capita GDP into the constant price series of 2000, and then take the natural logarithm, and expect its coefficient sign to be positive. (3) Fiscal self-sufficiency: it is measured by the ratio of revenue in the fiscal budget to expenditure in each year. Generally speaking,

the stronger the financial self-sufficiency, the stronger the urban public infrastructure and security capacity, so as to attract more residents to the city. Therefore, the coefficient sign is expected to be positive. (4) Human capital level (EDU): it is measured by the proportion of education expenses in financial expenditure. The level of human capital is often represented by the average number of years of education. However, at the urban level in China, only the data of per capita years of education in census years can be obtained, and the data of per capita years of education in each year can not be obtained. Generally speaking, the more the regional financial expenditure on education, the more conducive to the improvement of the regional education level, and the stronger the residents' willingness to enter the city. Therefore, the coefficient sign is expected to be positive. (5) Investment rate: it is measured by the proportion of fixed asset investment in GDP of the whole society. The higher the fixed asset investment rate in a region, the more perfect the infrastructure may be, and the greater the employment demand will be, attracting more residents to the city. Therefore, it is expected that the coefficient sign will be positive. (6) Industrial structure: it is measured by the employment proportion of the secondary industry and the tertiary industry. The existing literature usually uses the output ratio of the secondary industry and the tertiary industry to measure the industrial structure, but this paper believes that the greater impact on the urbanization rate is employment opportunities, so we use 2 The employment proportion of the three industries is used to measure the industrial structure, and the coefficient sign is expected to be positive. (7) The lag period of urbanization rate (l.real_u urban and l.nomi_u urban). Since we cannot fully control all the important variables affecting urbanization, we further introduce the urbanization rate of the first lag period into the model as the basic control variable to reduce the bias of the estimation results caused by the omission of variables.

# 3. Empirical results and discussion

## 3.1 Benchmark regression

Table 1 reports the baseline regression results for the model.

Table 1 Benchmark regression results

| | Explained variable: Real_ urban | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | *RE* | *RE* | *SYS-GMM* | *SYS-GMM* |
| *Div1* | -4.285 $^{***}$<br>(1.249) | -3.160 $^{***}$<br>(1.030) | -32.882 $^{***}$<br>(12.213) | -27.845$^{**}$<br>(11.765) |
| *L.Real_ urban* | | | 0.415 $^{***}$<br>(0.092) | 0.296 $^{***}$<br>(0.048) |
| *Fdi* | | 0.042<br>(0.099) | | 0.224<br>(0.308) |
| *Pgdp* | | 8.009 $^{***}$<br>(1.271) | | 25.341 $^{***}$<br>(9.461) |
| *Fiscal* | | 0.148 $^{***}$ | | 0.328$^{*}$ |

| | | | | |
|---|---|---|---|---|
| | | (0.036) | | (0.191) |

Table 1 contiued

| | Explained variable: Real_ urban | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | *RE* | *RE* | *SYS-GMM* | *SYS-GMM* |
| *Edu* | | 0.122 | | 0.587 |
| | | (0.116) | | (1.194) |
| *Invest* | | 0.051 $^{***}$ | | 0.177$^{**}$ |
| | | (0.015) | | (0.079) |
| *Struc* | | -12348$^{*}$ | | 5.414 |
| | | (7.437) | | (88.993) |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Not controlled | Control | Not controlled | Control |
| F inspection value | 200.250 | 235.560 | 49.074 | 10.812 |
| (P) | (0.000) | (0.000) | (0.000) | (0.000) |
| AR (2) inspection value | | | 0.521 | 1.241 |
| (P) | | | (0.603) | (0.215) |
| Hansen test value | | | 81.258 | 67.983 |
| (P) | | | (0.168) | (0.131) |
| Sample size | 3588 | 3588 | 3312 | 3312 |

Note: in the brackets below the coefficient are robust standard errors; $^{*}$, $^{**}$ and$^{***}$ Represent the significant level of 10%, 5% and 1% respectively; the following table is the same.

Among them, column (1) and column (2) use random effect model (RE) to estimate parameters, column (1) only considers dialect diversity index and time fixed effect, and uses clustering for each sample to obtain robust standard error. The results show that dialect diversity index has a significant negative impact on the real urbanization rate; after adding other control variables in column (2) and considering the time fixed effect and urban fixed effect, the dialect diversity index still has a significant negative impact on the real urbanization rate. The influence coefficients of the dialect diversity index in column (1) and column (2) on the real urbanization rate are -4285 and -3160 respectively, both of which are negative, indicating that the dialect diversity index has a significant negative impact on the real urbanization rate. On the basis of columns (1) and (2), columns (3) and (4) respectively add the real urbanization rate with a lag of one period as the control variable, so the model becomes a dynamic panel model. We use the system generalized moment estimation method (sys-gmm) specially suitable for estimating the dynamic panel model for parameter estimation. The results show that the dialect diversity index still has a significant negative effect on the urbanization rate, but compared with columns (1) and (2), the absolute value of the coefficient increases significantly, which indicates that the influence of dialect diversity on the real urbanization rate is underestimated to a certain extent in the model without considering the lag period of the real urbanization rate.

Since the results of the system generalized moment estimation method for the dynamic panel model are more reliable, we focus on the estimation results in column (4). The estimation results in column (4) show that, on average, for each additional dialect, the real urbanization rate will decrease by about 27845 percentage points, indicating that the cultural diversity represented by dialect diversity does have a negative impact on urbanization that can not be underestimated. The real urbanization rate lagging behind the first period also has a significant positive impact on the urbanization rate of the current period, indicating that the urbanization process has a significant "self reinforcing" effect; after adding the urbanization rate that lags behind the first stage, it is consistent with the expectation. The coefficient symbols of all control variables are positive. Among them, the coefficients of economic development level, investment rate and financial self-sufficiency are significantly positive. Although the regression coefficients of human capital level, openness and industrial structure are positive, they are not significant, indicating that human capital investment, openness and industrial structure have not played a significant role in promoting China's urbanization process.

### 3.2 Robustness test

The dialect diversity index (div1) used above may have some unreasonable measurement errors. For example, there are 1million people in a region with two dialects, of which only 10000 people speak one dialect and 990000 people speak the other dialect; another region also has 1million people and two dialects, with 500000 people speaking each dialect. Then, if we use the above language diversity index, the language diversity index of the two regions is 2, but the dialect influence of the two regions is obviously different. Therefore, we need to revise the language diversity index. In order to make the dialect diversity index reasonably reflect the influence of dialects, xuxianxiang et al. [16] calculated the ratio of the number of people using a certain dialect in a region to the number of people in the whole region when building the language diversity database, obtained the population weight (PIJ) of each dialect, and then calculated the revised dialect diversity index (div2) using the following formula:

$$Div2_i = 1 - \sum_{j=1}^{n} p_{i}j^{2} \qquad (2)$$

Where, is $Div2_i$ the language diversity index 2 of city I, is $P_{ij}$ the population weight of the j dialect used in city I, and N is the total number of dialects and sub dialects used in a region. The index shows that if the more people in a region speak a certain dialect, the $P_{ij}^2$ greater the value of div2, the smaller the coefficient of div2, indicating the smaller the linguistic diversity; on the contrary, the fewer people in a region speak a certain dialect, and the region has multiple dialects, the $p_{ij}^2$ smaller the value of, and the greater the coefficient of div2. Therefore, the coefficient is between 0 and 1. The larger the value, the more diverse the dialect is. Thus, the div2 of the above two regions are 002 and 05 respectively. Obviously, the revised dialect diversity index of the second region is larger, which indicates the rationality of this index.

In view of this, we replaced the dialect diversity index div1 with div2 to re estimate the parameters. The results in Table 2 show that the dialect diversity has a significant negative impact on the real urbanization rate whether the control variable is added or not. In column (4), the regression coefficients of the lag period of the real urbanization rate and other control variables are significantly positive (except the degree of opening up). The influence coefficient of dialect diversity on the real urbanization rate is -156, which is significant at the level of 5%, indicating that for every increase of one standard deviation (024) in dialect diversity index (div2), the real_urban rate will decrease by about 3775 (0242$^{*}$156) percentage points.

T

Table 2 Robustness test 1

| | Explained variable: Real_ urban | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | *RE* | *RE* | *SYS-GMM* | *SYS-GMM* |
| *Div1* | -16.691 $^{***}$ | -7.427$^{*}$ | -38.263+ н | -15.600$^{**}$ |
| | (4.417) | (3.987) | (12.789) | (7.844) |
| *L.Real_ urban* | | | 0.611 $^{***}$ | 0.418$^{***}$ |
| | | | (0.123) | (0.072) |
| *Fdi* | | 0.038 | | 0.054 |
| | | (0.099) | | (0.117) |
| *Pgdp* | | 8.098 $^{***}$ | | 18.570$^{***}$ |
| | | (1.282) | | (2.970) |
| *Fiscal* | | 0.147 $^{***}$ | | 0.117$^{***}$ |
| | | (0.036) | | (0.041) |
| *Edu* | | 0.118 | | 0.261$^{**}$ |
| | | (0.116) | | (0.115) |
| *Invest* | | 0.051 $^{***}$ | | 0.132 $^{***}$ |
| | | (0.015) | | (0.030) |
| *Struc* | | 12.847$^{*}$ | | 113.840 $^{***}$ |
| | | (7.430) | | (31.719) |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Not controlled | Control | Not controlled | Control |
| F inspection value | 201.640 | 230.930 | 112.081 | 40.201 |
| (P) | (0.000) | (0.000) | (0.000) | (0.000) |
| AR (2) inspection value | — | — | 1.251 | 0.87 |
| (P) | | | (0.212) | (0.385) |
| Hansen test value (P) | — | — | 212.970 | 224.524 |
| | | | (0.123) | (0.155) |
| Sample size | 3588 | 3588 | 3312 | 3312 |

Although the above urbanization rate data published by the Ministry of public security is more accurate, considering that the non-agricultural population data can reflect the traditional rural and urban residents' identity in the form of registered residence, we use the nominal urbanization rate measured by the proportion of non-agricultural population as the explained variable to test the robustness. Columns (1) and (2) of Table 3 are the regression results of div1 on the nominal urbanization rate, and columns (3) and (4) are the regression results of div2 on the nominal urbanization rate.

It can be seen that the coefficients of div1 and div2 are significantly negative no matter whether the first lag period of the explained variable is added or not, but the absolute value of the coefficient estimated by sys-gmm is smaller after the first lag period of urbanization is added, which indicates that the result without considering the first lag period of urbanization is overestimated; when considering that urbanization lags behind for a period of time, the nominal urbanization rate will decrease by about 3552% for every dialect (div1) added on average under other conditions unchanged; when div2 increases by one standard deviation (0242), the nominal urbanization rate will decrease by about 1099 (0242$^{*}$ 4558) percentage points. The above results show that whether the actual urbanization rate or the nominal urbanization rate is adopted, whether the dialect diversity index considering the population weight is adopted, or whether the urbanization lag period is introduced as the control variable, the cultural diversity represented by the dialect diversity shows a significant blocking effect on China's urbanization, which fully shows that the hypothesis we put forward is valid.

Table 3 Robustness test 2

| | Explained variable: Nomi_ urban | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | RE | SYS-GMM | RE | SYS-GMM |
| Div1 | -32.906$^{***}$ | -3.553$^{*}$ | | |
| | (1.460) | (1.993) | | |
| L.Real_ urban | | | -92.638$^{***}$ | -4.558$^{**}$ |
| | | | (30.964) | (2.245) |
| Fdi | | 0.842$^{***}$ | | 0.882$^{***}$ |
| | | (0.091) | | (0.066) |
| Pgdp | 0.163“ | 0.099$^{*}$ | 0.163$^{**}$ | 0.070 |
| | (0.066) | (0.059) | (0.066) | (6.823) |
| Fiscal | 0.262 | 3.205$^{**}$ | 0.262 | 2.648 |
| | (0.824) | (1.413) | (0.824) | (2.310) |
| Edu | 0.025 | 0.013 | 0.025 | 0.009 |
| | (0.050) | (0.117) | (0.050) | (8.492) |
| Invest | 0.008 | 0.013 | 0.008 | 0.013 |
| | (0.009) | (0.009) | (0.009) | (1.824) |
| Struc | 2.874 | 0.527 | 2.874 | 4.297 |
| | (3.900) | (4.686) | (3.900) | (3.639) |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Control | Control | Control | Control |
| F inspection value | 343.192 | 365.413 | 330.640 | 971.361 |
| (P) | (0.000) | (0.000) | (0.000) | (0.000) |
| AR (2) inspection value | | -0.416 | | -0.450 |
| (P) | | (0.677) | | (0.651) |
| Hansen test value | | 189.627 | | 244.821 |
| (P) | | (0.1119) | | (0.881) |

| | | | | |
|---|---|---|---|---|
| Sample size | 2484 | 2208 | 2484 | 2208 |

# 4. Further investigation based on instrumental variable method

As mentioned above, the biggest problem in the existing research on the determinants of urbanization may be the insufficient attention to endogenous issues [17]. In this paper, the cultural diversity represented by linguistic diversity will affect the urbanization process, but the urbanization process of a region is also accompanied by the collision and blending of various cultures, which makes people with different languages and cultural backgrounds gradually converge in the city. At the same time, people who speak different dialects are usually forced to use a higher frequency of Putonghua in cities, or learn the local popular dialects, so as to facilitate communication. This "devouring" effect on the original dialect and culture of urban immigrants may be more obvious in the next generation of urban immigrants. Therefore, there is likely to be a simultaneous causal relationship between culture (dialect) and urbanization, resulting in endogenous problems. In addition, the factors that affect urbanization are complex and diverse, so it is difficult to control them comprehensively. Therefore, the model in this paper also faces the endogenous problem caused by the omission of some important factors. Therefore, we need to pay special attention to the endogenous problem. In this section, we will use the instrumental variable method to conduct a more robust empirical study.

One of the reasons for the formation of Chinese dialects is that the languages of the original ethnic minority areas have gradually evolved into unique dialects [16]; second, due to the large-scale migration of population, the language will "migrate" to another place, and merge or assimilate with the local language to form a new dialect [17]; third, a certain region is isolated from the outside world, and the local language and the external language cannot evolve synchronously, thus forming a regional dialect [18]. Due to the low urbanization rate in ethnic minority areas, it is obviously inappropriate to use the relevant variables in ethnic minority areas as instrumental variables. Therefore, we believe that the most ideal instrumental variable should be the population migration. Specifically, we first identified the population migration in China's history. There were ten large-scale population migrations, but five domestic population migrations with short-term, forced and large-scale characteristics (see Table 4). These five major migrations have caused the population of the Central Plains, hebei, shaanxi, the southwest of the mountains, the two lakes, guangdong, guangdong, jiangxi and Fujian to migrate to the southeast coast, sichuan and other regions on a large scale, which has enriched the dialects in the relocated areas and gradually formed the current dialect distribution pattern. After consulting the specific historical data of five population migrations, we assigned the value of 1 to the area with more than one population migration, and 0 to the area without population migration. Specifically, if it can be identified that a certain city is the place of population migration, then the city will be assigned as 1; if we can't identify the specific immigration place of a certain population migration, but can roughly identify a province or a region of a province to which a certain population migrated, then we will assign all cities of the province or the region to 1; if there are no cities with population migration in the five major population migrations, all of them will be assigned as 0. We express this tool variable as tool.

Table 4 Five large-scale population migration events

| Event | Time of occurrence | Emigration site | Place of immigration |
|---|---|---|---|
| Yongjia rebellion [19]<br>An Shi rebellion [20]<br>Shame of Jingkang [21]<br>Hongdong dahuashu resettlement [22<br>Huguang filling Sichuan [23] | Late Western Jin Dynasty<br>Tianbao period of Tang Dynasty<br>Jingkang period of the Northern Song Dynasty and the early Ming Dynasty<br>Late Ming and early Qing Dynasty | Central plains<br>Henan, hebei, shaanxi<br>Zhongyuan<br>Southern Shanxi<br>Lianghu, liangguang, jiangxi, fujian | Jiangnan, hunan and Hubei<br>Now there is the area from Jingzhou in Hubei to Changde in Hunan; east into the Yangtze Huaihe River Basin and Taihu Lake Basin; west into Sichuan.<br>The southeast provinces, fujian, guangdong, southern Jiangsu, zhejiang, henan, hebei, shandong, anhui, jiangsu and other Central Plains regions Sichuan |

Data source: the author sorted it out according to relevant literature.

Table 5 reports the results of parameter estimation using the two-stage least square method (2SLS) based on the random effect model, in which the explained variable is the real_urban and tool is used as the dialect diversity tool variable. Columns (1) and (2) are the analysis results of div1's impact on the real urbanization rate, and columns (3) and (4) are the analysis results of div2's impact on the real urbanization rate.

Table 5 Estimation of instrumental variables (real_urban)

| | Phase II | | | |
|---|---|---|---|---|
| | Explained variable: Real_ urban | | | |
| | (1) | (2) | (3) | (4) |
| *Div1* | -4.293** | -2.52** | | |
| | (-2.433) | (-2.053) | | |
| *Div2* | | | -25.336** | -14.546** |
| | | | (-2.433) | (-2.047) |
| *Edu* | | 0.121* | | 0.119* |
| | | (1.721) | | (1.701) |
| *Fdi* | | 0042 | | 0033 |
| | | (0.473) | | (0.373) |
| *Fiscal* | | 0.148 *** | | 0.146*** |
| | | (6.984) | | (6.851) |
| *Invest* | | 0.051 *** | | 0.051*** |
| | | (4.444) | | (4.405) |
| *Pgdp* | | 8.068 *** | | 7.900*** |
| | | (9.761) | | (9392) |

| | | |
|---|---|---|
| *Struc* | 12.497$^{**}$ | 12.594$^{**}$ |
| | (-2.323) | (-2.339) |

Table 5 continued

| | Phase II | | | |
|---|---|---|---|---|
| | Explained variable: Real_ urban | | | |
| | (1) | (2) | (3) | (4) |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Control | Control | Control | Control |
| Sample size | 3588 | 3588 | 3588 | 3588 |
| Wald test value | 87.202 | 301.162 | 87.201 | 299.031 |
| (P) | (0.000) | (0.000) | (0.000) | (0.000) |
| First stage regression | | | | |
| | Explained variable: div1 | | Explained variable: div2 | |
| *Tool* | 1.760 $^{***}$ | 1.765 $^{***}$ | 0.298$^{***}$ | 0.306$^{***}$ |
| | (0.0231) | (0.0244) | (0.0101) | (0.0098) |
| Control variable | Consider | Consider | Consider | Consider |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Control | Control | Control | Control |
| Wald test value | 5793.000 | 5335.000 | 871.000 | 1061.000 |
| (p) | (0.000) | (0.000) | (0.000) | (0.000) |
| Sample size | 3588 | 3588 | 3588 | 3588 |

The results show that the Wald statistics estimated in the first stage of the four models are very significant, indicating that there is no problem of weak instrumental variables, and the regression coefficients of tool for div1 and div2 are significantly positive, indicating that the selection of instrumental variables is reasonable. In the second stage regression, from the regression coefficients of column (1) and column (3), the estimated results are overestimated to a certain extent when the control variables are not considered; from the regression coefficients of columns (2) and (4), when the control variable is added, the regression coefficients of dialect diversity index (div1 and div2) to the real urbanization rate are significantly negative; except for the degree of openness, other control variables passed the significance test of more than 10%. Under the condition that other factors remain unchanged, for div1, the urbanization rate will decrease by about 2521 percentage points for each additional dialect, while for div2, the urbanization rate will decrease by about 352 (0242$^{*}$ 14546) percentage points for each additional standard deviation (0242). This result is smaller than the coefficient of the benchmark regression, indicating that the estimation result of the benchmark regression is overestimated to a certain extent.

Table 6 reports the estimated results with the explanatory variable nomi_urban and 2SLS. Columns (1) and (2) are the analysis results of div1's impact on the nominal urbanization rate, and columns (3) and (4) are the analysis results of div2's impact on the nominal urbanization rate. The four regression groups are estimated by random effect model. The Wald statistics estimated in the first stage are very significant, indicating that there is no problem of weak instrumental variables. Moreover, the regression coefficients of tool for div1 and div2 are significantly positive, which also indicates that the selection of instrumental variables is reasonable. From the regression coefficients in columns (2) and (4), as far as dialect diversity index div1 is concerned, each additional dialect will reduce the nominal urbanization rate by about 3.113 percentage points; in terms of dialect diversity index div2, the nominal urbanization

Table 6 Estimation of instrumental variables (nomi_urban)

| | | Phase II | | |
|---|---|---|---|---|
| | | Explained variable: real_ *urban* | | |
| | (1) | (2) | (3) | (4) |
| *Div1* | -3.91$^{**}$ | -3.113 $^{***}$ | | |
| | (-2.263) | (-2.771) | | |
| *Div2* | | | -22.219$^{**}$ | -17.795 $^{***}$ |
| | | | (-2.221) | (-2.759) |
| *Edu* | | 0022 | | 0.022 |
| | | (0.732) | | (0.733) |
| *Fdi* | | 0.071$^{*}$ | | 0076$^{*}$ |
| | | (1.812) | | (1.915) |
| *Fiscal* | | 0.074 $^{***}$ | | 0.073 $^{***}$ |
| | | (7.972) | | (7.833) |
| *Invest* | | 0.018 $^{***}$ | | 0.018 $^{***}$ |
| | | (3.319) | | (3.288) |
| *Pgdp* | | 2.631 $^{***}$ | | 2.553$^{***}$ |
| | | (6.558) | | (6.291) |
| *Struc* | | 1.287 | | 1.157 |
| | | (0.538) | | (0.481) |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Control | Control | Control | Control |
| Sample size | 2268 | 2268 | 2268 | 2268 |
| Wald test value | 607.533 | 687.052 | 60.735 | 681.821 |
| (P) | (0.000) | (0.000) | (0.000) | (0.000) |
| | | First stage regression | | |
| | Explained variable: div1 | | Explained variable: div2 | |
| *Tool* | 1.773 $^{***}$ | 1.770 $^{***}$ | 0.312 $^{***}$ | 0.310 $^{***}$ |
| | (0.029) | (0.029) | (0.012) | (0.012) |
| Control variable | Consider | Consider | Consider | Consider |
| Time fixed effect | Control | Control | Control | Control |
| Urban fixed effect | Control | Control | Control | Control |
| Wald test value | 3642.000 | 3649.000 | 705.000 | 715.000 |
| (p) | (0.000) | (0.000) | (0.000) | (0.000) |
| Sample size | 2484 | 2484 | 2484 | 2484 |

rate will decrease by about 4306 (0.242* 17795) percentage points every time the dialect diversity increases by one standard deviation (0.242). Comparing the results in Table 5 and Table 6, it can be seen that in the estimated results using instrumental variables, the influence coefficient of dialect diversity index (div1 and div2) on the nominal urbanization rate is greater than that of the real urbanization rate, indicating that the cultural factors reflected by dialect diversity have a greater impact on farmers' transferring from agricultural to non-agricultural household, while a relatively small impact on the actual flow of farmers from rural to urban areas. Overall, the negative impact of dialect diversity on urbanization is robust and significant, which shows that the hypothesis proposed in this paper is valid.

## 5. Conclusion

The results show that the dialect diversity has a significant negative impact on the urbanization rate after controlling several related control variables; after controlling the lag of the explained variable for one period in the regression equation, this significant negative effect is still robust; in order to control the endogenous problems caused by reverse causality, we took whether there were immigrants in a region during the five large-scale population migrations in history as the indicator construction tool variable, and used the two-stage least square method to estimate the parameters. It was found that the above results were still stable. Therefore, the empirical results of this paper reveal that the cultural variables represented by dialects are an important factor affecting the process of urbanization.

In the process of promoting urbanization, we should fully consider the objective fact that urban migrants and potential urban migrants with various cultural and linguistic backgrounds will collide, communicate and integrate with each other in the city. In the process of formulating urbanization policies, we must avoid policies that are not conducive to cultural integration, but pay attention to the integration and guidance between different cultures, try to ensure that residents of various cultural backgrounds can enjoy their respective rights and obligations in the city and share the city's public services. Finally, it should be pointed out that the impact mechanism of cultural diversity on urbanization is relatively complex, which needs to be further explored in the follow-up study.

### References

[1] Pandey S.M.Nature and determinants of urbanization in a developing economy: the case of India[J]. Economic Development and Cultural Change,1977,25(2) : 265-278.

[2] Chang G.H.,Brada J.C.The paradox of China’s growing under-urbanization[J]. Economic Systems,2006,30(1) : 24-40.

[3] moomawr L,Shatter A M.Urbanization and economic development: a bias toward large cities? [J]. Journal of Urban Economics,1996,40(1) : 13-37.

[4] Zhang Tao, li Bo, buyongxiang, wu chaoming Manufacturing industry, land cost

and China's urban development -- a panel data model of the determinants of China's urbanization [J]. Financial research, 2007 (3): 10-24

[5] Wang duanyong, zhu Nong Regional differences in determinants of China's urbanization development [J]. China population, resources and environment, 2007 (1): 66-71

[6] Zhanghongming Some basic ideas on solving the housing problem of migrant workers [J]. Journal of East China Normal University: Philosophy and Social Sciences Edition, 2016 (6): 141-144

[7] Wan Guanghua, zhengsiqi, anett hofmann Determinants of urbanization level: Transnational regression model and analysis [J]. World economic journal, 2014 (4): 20-35

[8] Wang Xi, chenzhongfei Determinants of China's urbanization level: Based on international experience [J]. World economy, 2015 (6): 167-192

[9] Liuruiming, shi Lei Ownership basis of China's urbanization stagnation: theoretical and empirical evidence [J]. Economic research, 2015 (4): 107-121

[10] Xuxianxiang, liuyuyun, xiaozekai Dialect and economic growth [J]. Journal of economics, 2015 (2): 1-32

[11] Gao Bo, zhangzhipeng Cultural Capital: an explanation of the source of economic growth [J]. Journal of Nanjing University: philosophy, humanities and Social Sciences, 2004 (5): 102-112

[12] High wave China's economic growth: an analytical framework of cultural change [J]. Nanjing Social Sciences, 2007 (7): 17-25

[13] Jiang Li Cultural dimension and spatial economic growth [J]. Economist, 2010 (5): 81-87

[14] Zhu Konglai, li Jingjing, le Feifei An empirical study on the relationship between urbanization and economic growth in China [J]. Statistical research, 2011 (9): 80-87

[15] Liuyuyun, xuxianxiang, xiaozekai. The inverted U-shaped model of labor flow across dialects [J]. Economic research, 2015 (10): 134-146

[16] Fu Ailan. Chinese dialects and national languages [J]. Dialect, 2001 (3): 193-198

[17] Peizeren Population migration in Ming Dynasty and the formation of Northern Henan dialect [J]. Zhongzhou journal, 1988 (4): 102-106

[18] Mai Yun On the formation pattern of Chinese dialects from the emergence and development of Cantonese [J]. Dialect, 2009 (3): 219-232

[19] Hanshufeng During the northern and Southern Dynasties, the border Haozu of huaihan and Yibei [M]. Beijing: Social Science Literature Press, 2003

(Responsible editor: Yumin)