Can Foreign Direct Investment Increase the Productivity of Domestic Firms? Identifying FDI Spillovers from Borders of Chinese Dialect Zones

Sen Ma[1]

Department of Economics

University of Illinois at Urbana-Champaign

11/22/2017

Abstract

Many developing countries adopt policies to attract FDI, hoping multinational firms yield knowledge spillovers and raise the productivity of domestic firms. This study identifies positive productivity spillovers from FDI, employing discontinuous increases in investment from Hong Kong, Macau, and Taiwan (HMT) to mainland China at the geographical borders of Chinese linguistic dialect zones. Using a spatial regression discontinuity design, I find that the share of HMT firms in a location just inside a region that speaks the same dialects as HMT is 5-7 percentage points (20 percent) higher than a location just outside. Using this discontinuous increase in HMT investment across the dialect borders as exogenous variation, I find positive local horizontal productivity spillovers to domestic firms, especially in industries where HMT firms are more productive than domestic firms. A 1 percentage point increase in the HMT firm share raises the productivity of domestic firms in the same location and industry by 1.7 to 2.8 percent.

Keywords: FDI, Dialect, Spillover

JEL codes*:* F21, F23, O19, O24

[1] This article was previously circulated as "The Effects of Cultural Similarity on Foreign Direct Investment and Productivity of Domestic Firms: Identification from Borders of Chinese Dialect Zones". I am indebted to my thesis committee members: Richard Akresh, Benjamin Marx, Daniel McMillen, and Adam Osman for their numerous invaluable comments and support. I thank Jianfeng Xu for his great help in the process of data collection. I further thank Brian Kovak, Marti Mestieri, Andrea Lassmann, Lixing Li, Ming Lu, and Yi Lu for useful comments. I benefited from comments of participants at the annual meetings of the ASSA in Chicago, the European Trade Study Group in Helsinki, The Chinese Economist Association annual meeting in Sacramento, the CCER Summer Institute 2017 in Yantai, and student seminars at the University of Illinois at Urbana-Champaign.
Contact information: University of Illinois at Urbana-Champaign, Department of Economics, 214 David Kinley Hall, 1407 W. Gregory Urbana, Illinois 61801, USA. Email: senma2@illinois.edu.

1. Introduction

Many developing countries rely on foreign direct investment (FDI) to provide much-needed capital, technology, and management skills. In addition to the direct benefits of FDI, policy makers also hope that domestic firms can benefit indirectly from the presence of foreign firms. Domestic firms can learn from foreign firms through directly observing the foreign firms, establishing business relations with the foreign firms or human capital transfer as domestic employees move from the foreign to the domestic firms. Therefore, numerous empirical studies have been conducted in different countries to estimate the productivity spillovers from FDI to domestic firms.[2]

Even though many estimates of productivity spillovers have been produced in previous studies, it is widely acknowledged that finding a good strategy to identify a causal relationship is challenging. Previous studies usually follow the empirical framework pioneered by Aitken and Harrison (1999) and further developed by Javorcik (2004). In this framework, the main variation comes from variation across industries in the presence of foreign firms. Spillovers are estimated through establishing a relationship between a measure of the productivity (usually TFP) of domestic firms and a measure of the presence of foreign firms in the same industry as the domestic firms. However, when foreign firms make decisions to enter the market of the host countries, the productivity of domestic firms is a factor they will take into consideration. As a

[2] The theoretical framework of the analysis is based on Rodriguez-Clare (1996). For a summary of empirical findings, please refer to the meta-analyses conducted by Gorg and Strobl (2001), Harvranek and Irsova (2011), and Irsova and Havranek (2013). The authors conclude that previous empirical studies reach the following conclusions: First, horizontal spillovers are on average zero. Second, spillovers to domestic suppliers through backward linkages are on average positive. Third, FDI generates small spillovers to domestic buyers through forward linkages. Finally, the signs and magnitudes of spillovers depend systematically on the characteristics of the domestic economy and foreign investors.

result, the distribution of foreign firms across industries is the result of foreign firms' optimal decisions, and therefore is not exogenous. [3]

In this study, instead of using variation across industries, I use variation in FDI due to the unique cultural ties between the destination locations and the source regions of FDI to identify the causal effect of FDI on the productivity of domestic firms. Exogenous variation comes from discontinuous changes in investment from Hong Kong, Macau and Taiwan (abbreviated as HMT) to mainland China at the geographical borders of Chinese dialect zones. Language borders are used by Egger and Lassmann (2015) to identify the causal effect of common language on the patterns of international trade. This study uses a similar strategy and is the first study that identifies the casual effect of cultural ties[4], represented by common dialect on the destination choice of FDI using discontinuous variation at language borders. I use maps from the "Language Atlas of China" to divide China geographically into regions where people speak different Chinese dialects. A region speaking the same dialects as HMT implies unique cultural ties originated from common cultural origin and thus expects to attract more investment from HMT as well. Figure 1 shows the two Chinese dialect zones studied in this research. Regions within the Cantonese dialect zone speak the same dialect as Hong Kong and Macau.[5] Similarly, regions within the Min dialect zone speak the same dialect as the majority of Taiwanese.[6] The Geographical borders of the two dialect zones can generate discontinuous variation in the

[3] Lu et al. (2017) use changes in Chinese FDI regulation policies as the identification strategy to address this endogeneity problem. The authors find negative national level horizontal productivity spillovers to domestic firms. The magnitude of spillovers is much larger than those estimated from models that do not address potential endogeneity problems, indicating significant bias in an OLS framework.

[4] The effects of cultural ties represent the aggregate effects of cultural similarity and unique social networks based on cultural linkages.

[5] In China, people use the same written language, therefore the major linguistic difference is from spoken language.

[6] Shen Zhen, which is a city on the dialect border but, is a special case in the sample because the composition of population is mainly migrants. As a robustness check, I tried to remove Shen Zhen from the sample and find that the major results are similar.

distribution of population from different cultural groups. Thus, variation at the borders can be used to identify the causal effects of cultural ties on the destination choice of FDI. Another characteristic of the dialect borders is that they are independent from the administrative borders in many places, which makes it possible to estimate the net effects of cultural borders, excluding the effects of administrative borders.

In the first stage, I estimate a discontinuous increase in HMT investment at the geographical borders of Chinese dialect zones using a spatial regression discontinuity design. I find that the share of firms from HMT in zip codes located just inside the common dialect zones (speaking the same dialects as HMT) in mainland China is 5 to 7 percentage points (20 percent) higher than zip codes just outside. I also find that the magnitude of the discontinuous increase in HMT investment at the common dialect borders varies by industries. Larger discontinuous increases are identified in industries where the entry of FDI is not regulated by government policy and industries where HMT firms have higher productivity. These findings are consistent with the hypothesis that the entry of HMT firms is encouraged by a lower entry cost and a larger advantage in productivity.

In the second stage, I use speaking the same dialects as HMT around the dialect borders as the instrument for the potentially endogenous share of HMT investment in the empirical framework of Aitken and Harrison (1999) to identify horizontal spillovers. I find that in industries where HMT firms are more productive than domestic firms, [7] the increase in the share of HMT firms leads to the improvement in the productivity of domestic firms in the same industry and location (horizontal spillovers). In terms of magnitude, a 1 percentage point

[7] Industries where HMT firms are more productive are represented by manufacturing industries and steal industries. Industries where HMT firms have lower productivity are represented by food processing industries and cultural industries.

increase in the share of HMT firms across the dialect borders raises the productivity of domestic firms, measured in TFP following Brandt et al. (2012) and Ackerberg et al. (2015), in the same location and industry by around 1.7 to 2.8 percent. However, if the empirical models are estimated without instruments, horizontal spillovers are estimated to be negative in this sample. This is probably because HMT firms can selectively enter industries where the productivity of domestic firms is low, leading to a negative correlation between the presence of HMT firms and the productivity of domestic firms at the industry level. It is only after correcting for the endogenous entry problem using instruments that I identify the actual spillovers to be positive. The positive horizontal spillovers appear to be inconsistent with the average results, which show zero or negative horizontal spillovers, in the literature (Irsova and Havranek, 2013). However, spillovers captured in this specific identification strategy is a very local spillover effect, which is often estimated to be positive (Halpern and Murakozy, 2007; Xu and Sheng, 2012; Lu et al., 2017), because positive knowledge spillovers are more likely to happen when the domestic firms are located close to the foreign firms.

Finally, I conduct additional tests to address the identification assumptions of this empirical strategy. Identification of the empirical models requires two important assumptions. The first assumption requires that factors other than dialect should change continuously at the borders such that the discontinuous change in the share of HMT firms is only caused by the variation in dialect. I conduct the following four tests to support this assumption. First, observable non-economic covariates, such as demographic and geographic characteristics change continuously at the dialect borders. Second, the share of foreign firms from countries other than HMT does not discontinuously change at the borders, which indicates zip codes that speak the same dialects as HMT do not generally attract more FDI, instead, they attract more FDI only

from HMT. Third, the share of HMT firms does not discontinuously change at a placebo dialect border (Wu dialect border) that is correlated with neither Cantonese nor Min culture, which confirms that dialect borders, without specific cultural meanings attached to them, do not generate direct effects on economic outcomes. Fourth, the share of HMT firms does not discontinuously change at placebo dialect borders generated by moving the actual dialect borders arbitrarily inward or outward by 100 kilometers. The second assumption requires that common dialect affects the productivity of domestic firms only through attracting more investment from HMT. This argument is supported by three additional tests. First, the productivity of domestic firms, which are from industries that are not influenced by HMT investment, does not discontinuously change at the dialect borders. Second, the positive spillovers are only identified in industries where HMT firms are more productive than domestic firms. In industries where HMT firms are less productive than domestic firms, the productivity of domestic firms are very similar across the dialect borders, Third, the productivity of domestic firms also does not change discontinuously at placebo dialect borders.

In addition to the literature on estimating productivity spillovers from FDI, this study also contributes to the literature on cultural/linguistic similarity and economic exchange. Previous studies have shown that common language or culture affects economic activities, especially the patterns of international trade (Rauch and Trindade, 2002; Chiswick, 2008; Melitz, 2008; Felbermayr and Toubal, 2010; Egger and Lassmann, 2012; Falck et al., 2012; Sauter, 2012; Melitz and Toubal, 2014; Egger and Lassmann, 2015). The possible relationship between cultural similarity and the destination choice of FDI is studied by Guiso et al. (2009) and Kim et al. (2015), although it is hard to empirically single out culture as the primary factor by excluding other possible mechanisms. Thus, by studying variation across language borders in a spatial

regression discontinuity design, this study provides more reliable estimates to the literature on the effects of cultural similarity and destination choice FDI.

The rest of this article is organized as follows. Section 2 provides background about Chinese dialect borders and investment from HMT to mainland China. I highlight evidence showing that the geographical distribution of people who speak a specific dialect changes discontinuously at the dialect borders. Section 3 discusses the data used to calculate HMT investment and productivity spillovers. Section 4 discusses empirical strategies. Section 5 reports empirical results and discusses possible mechanisms. Section 6 discusses limitations and concludes.

2. Background

2.1 Dialect Zones in China from "Language Atlas of China"

In this section, I describe how to use the dialect zones from "Language Atlas of China" to generate the measure of cultural ties. The borders of the dialect zones have two characteristics. First, they are independent from administrative borders in many places. Therefore, I can identify the effects of culture net of the effects of economic policy. Second, the geographical distribution of population from different cultural groups change discontinuously at these borders, which satisfies the assumption to use a spatial regression discontinuity design.

In this study, cultural ties are measured by whether a location in mainland China speaks the same Chinese dialect as Hong Kong, Macau and Taiwan (HMT). The same dialect does not only indicate lower communication cost, but also implies similar cultural origins of the population. People who speak different dialects are so distinct that dialect significantly affects all kinds of economic activities. For example, Chen et al. (2014) shows that the ability to speak Shanghainese generates positive returns in the labor market in Shanghai.

I use maps from "Language Atlas of China"[8] to define locations that speak different dialects in mainland China and the geographical borders of those dialect zones.[9] "Language Atlas of China" are maps published by the Chinese Academy of Social Science and show the geographical distribution of major Chinese dialects. The maps are constructed by linguists and anthropologists based on their field works, in which they collect data on the dialect spoken by the majority of the population at each sampled Chinese village. Figure 3, which is a map of Jiangyin county,[10] illustrates how geographical borders between different dialects are drawn in these maps. From this map, we can see that the county is separated into multiple villages. Linguists and anthropologists have qualitative information on the dialect used by the majority of the population in these villages.[11] Different dialects are visually represented in the figure by different notations. Then the borders of dialect zones are drawn between villages that speak different dialects.

An interesting feature of the dialect borders is that they do not always coincide with the current county level administrative borders,[12] [13]which usually determine differences in economic policies. Figure 4 shows an example of the relationship between dialect borders and

[8] I use the maps published in 1987 (Wurm et al., 1987) to define dialect borders. A new version of the maps is published in 2008. In the new version of the map, it is mentioned that the Min and Cantonese dialect borders did not undergo significant changes from the older version of the maps.

[9] I am using the definition of major dialect groups to define cultural groups. Major dialect groups can be further divided into more detailed cultural subgroups. This characteristic is relevant in the analysis of the Min dialect zone. The Min dialect zone can be further divided into the southern Min dialect zone and Northern Min dialect zone. Because the southern Min dialect group is more similar to Taiwan, we should expect that most of the effects are driven by the southern side of the Min dialect border. Therefore, I conducted a robustness check by excluding the northern part of the Min dialect border and find that major results are not altered if we only use the southern part of the Min dialect border.

[10] This county is not in the analytical sample. I use this example to illustrate how the borders are generally constructed.

[11] I have data on the final borders drawn by linguists and anthropologists, but do not have village level data.

[12] The dialect borders may be determined by the location of historical administrative borders (He,2003; Hu and Yao, 2011). Since historical administrative borders change over time due to strategic concerns (Gao and Long, 2014) while cultural borders are more stable, we observe the discrepancy between dialect (cultural) borders and modern administrative borders in many places today.

[13] Figure A2.1 shows that the dialect borders also do not coincide with major rivers around this area.

administrative borders. The dark black line, which indicates a dialect border, passes through several counties, separating them into regions that speak different dialects. The points in this figure indicate the centroid of zip codes. My identification strategy will compare zip codes within the same county on two sides of the dialect border. Thus, any policy differentiations at the county level will be controlled by including county fixed effects.

Using the dialect borders defined by "Language Atlas of China" to study discontinuous changes in investment from HMT requires an important identification assumption that the geographical distribution of people who speak a specific dialect and belong to a certain cultural group changes discontinuously at the dialect borders.[14] However, this assumption is ambiguously supported by the literature of anthropology and linguistics. Some early studies claim that the changes in population who speak a certain dialect at the borders are "sharp". In some cases, when we move from villages on one side of the border to villages on the other side, the language people speak drastically changes (He,2003). However, later studies note that changes in some places are more continuous due to cultural communication between villages on two sides of the cultural borders (Simmons et al., 2006). Unfortunately, this assumption can't be directly tested because there are no comprehensive data to describe the geographical distribution of Chinese people by what dialect they speak. However, this assumption can be indirectly tested by studying the geographical distribution of Chinese people by their surnames. People who are from different cultural groups have drastically different distribution of surnames. Therefore, the distribution of representative surnames can be used to speculate the distribution of dialect groups.

To study the geographical distribution of representative surnames, I use the results from a study by Chinese linguists (Du and Yuan, 1993) to define representative surnames for each dialect

[14] Several recent studies have documented that cultural borders can generate discontinuous effects on economic outcomes (Egger and Lassmann, 2015; Becker et al., 2016; Lowes, 2017).

group. In this study, the authors list the 10 most common surnames (ranked by population share) for each Chinese dialect zone. For the two dialect zones that are relevant (Min and Cantonese [Yue]), we do find that the distribution of surnames is drastically different from other dialect zones. For example, "Lin (林)" is the second largest surname for the Min dialect zone and "Liang (梁)" is the second largest surname for the Cantonese dialect zone. However, these two surnames are not ranked as top ten for any other Chinese dialect zones. I use the three most common surnames as the representative surnames for Min (Chen [陈], Lin [林] and Huang [黄]) and Cantonese (Yue) dialect zones (Chen [陈], Liang [梁] and Li [李]). The empirical results are similar if I use these surnames separately.

Using data from the 2005 Chinese population census, I document discontinuous changes in the geographical distribution of representative surnames around the dialect borders. I calculate the population share of representative surnames in each county.[15] In Figure 5, the population share of representative surnames are plotted by distance to the dialect borders. Figure 5.1 shows the population share of representative surnames of the Min dialect group at the border of the Min dialect zone. The horizontal axis denotes the distance to the dialect borders with negative values indicating the location is inside the dialect border. The vertical axis denotes the population share of representative surnames. We can identify a significant discontinuous drop when moving from inside the border to outside. Similarly, I plot the population share of representative surnames of the Cantonese (Yue) dialect group at the border of the Cantonese (Yue) dialect zone in Figure 5.2. I still find a discontinuous drop even though it is smaller in magnitude than that of the Min dialect

[15] The most detailed geographical information in Chinese population census is at the county level. Therefore, it is not possible to test discontinuous changes in geographical distribution of representative surnames at the zip code level due to data limitation. However, the analysis at the county level does convince the idea that dialect (culture) does change discontinuously at the dialect borders defined by "Language Atlas of China".

zone. Both exercises support the assumption that geographical distribution of population that are from different dialect groups does discontinuously change at the dialect borders defined in "Language Atlas of China".

2.2 Investment from HMT to mainland China

In this study, cultural ties are linked to investment from HMT to mainland China. A large proportion of FDI in China comes from regions that are part of China, but became colonies of western countries or Japan, such as Hong Kong, Macau and Taiwan (HMT). Investment from HMT receives regulations in the same way as other types of FDI. Entry of HMT investment into certain industries is regulated due to strategic concerns.

Since 1978, when China initiated the open-door policy and became more integrated into the world economy, the country witnessed rapid growth in Foreign Direct Investment (FDI). An important feature of FDI in China is that a large proportion of investment comes from regions that are part of China, but were historically separated from China and became colonies of Britain, Portugal and Japan. Thus, these regions established different economic systems and experienced different paths of economic development. At the moment when China began opening up, these regions had well-established market institutions and were far ahead of mainland China in economic development. Figure 2 shows how the share of investment from Hong Kong and Taiwan[16] among all FDI in mainland China changes across time since the 1990s. The share of Hong Kong investment was around 30 to 40 percent until 2005 and then increased to around 60 percent. The share of Taiwan investment was around 10 percent in the 1990s, but has been declining over time.[17]

[16] The share of Macau investment is very small and thus not emphasized in this study.

[17] Calculating investment from HMT suffers from the problem of "Round trip FDI" in mainland China. Some domestic investors from mainland China will move their assets to Hong Kong before they start an investment

Unique cultural ties with mainland China are believed to explain the large share of Hong Kong and Taiwan investment in previous studies (Zhang, 2005). In an underdeveloped market environment, cultural ties help investors reduce transaction cost (Dai et al., 2016), build up trust and protect property rights.

An important benefit of using investment from HMT as a case to study the effects of culture and FDI is that the economic influences of the dialect borders are not likely to be correlated with unobservable local characteristics. This is because the dialect borders carried no economic meanings until HMT historically became colonies of western countries (or Japan) and took very different routes of economic development than mainland China. It is only after the divergence due to colonization that regions speaking Cantonese or the Min dialect can take advantage of their economically more developed cultural brothers. As a result of colonization, HMT becoming economically more developed is independent from any local economic factors in mainland China.

Also, the economic influences of the dialect borders do not represent the long-run effects of trading with HMT. This is because China followed central planned economy and was closed from the world economy when the economy of HMT took off. It is only after the 1980s that HMT started to generate economic impact on mainland China through FDI.

program in mainland China. Then, their investment will be counted as investment from Hong Kong and their firms can benefit from FDI encouraging policies. I think the identification strategy of this study helps to solve the problem, because the true investments from HMT are more likely to be affected by the dialect borders than "Round trip FDI" that are from all parts of mainland China. Therefore, the discontinuous gap at the dialect borders is more likely to represent differences in true HMT investment instead of differences in "Round trip FDI". As an additional test, I estimate the discontinuous gap in the investment from HMT only at the Min dialect border because investment from Taiwan is less likely to be affected by "Round trip FDI" than investment from Hong Kong. The results are shown in Appendix section 9. I find that the magnitude of the discontinuous increase is similar to (even slightly larger than) the baseline model (using both the Min and the Cantonese dialect borders). Therefore, the effects of "Round trip" FDI in the baseline model should not be very large.

In addition to the discontinuous changes at the dialect borders, I also explore heterogeneous effects of the dialect borders across industries by Chinese FDI regulation policies, which restrict FDI from entering certain industries. Lu et al. (2017) shows that changes in these FDI regulation policies indeed affect inflow of FDI into regulated industries.[18] The major goal of FDI entry-regulation policy is to protect domestic firms in the same industry from competition from foreign investment. In 1997, the central government of China published the "Catalog for the Guidance of Foreign Investment Industries," which became the government guidelines for FDI regulations. Specifically, the catalog classifies products into four categories: (1) FDI was supported, (2) FDI was permitted, (3) FDI was restricted and (4) FDI was prohibited. The catalog of products listed as restricted to FDI underwent changes upon China's accession to WTO in 2001, which generated a new catalog. Combining both new and old catalogs, I check at the two-digit industry level whether any goods produced by a specific industry in a specific year is listed on the catalogs. If any goods produced by a two-digit industry are listed as either "restricted to FDI" or "forbidden to FDI", I treat that two-digit industry as being regulated in that specific year. Thus, I create an industry-year dummy variable "regulation", which indicates whether a specific industry receives FDI entry-regulation policy in a specific year. Summary statistics in Table 1 show that about 33 percent of firms are from industries that receive entry-regulation.[19]

3. Data

Data on HMT investment is calculated from firm level data from the Chinese Industrial Census from 1998 to 2006. The census collects data on all manufacturing firms with sales above

[18] Lu et al. (2017) focuses on changes in these regulation policies across time, while this research only focuses on industrial variation. Because this study is conducted at a very small geographical level (zip code level), changes in regulation policies only happen in a small proportion of the sample and therefore do not generate enough power to identify the effects on domestic firms. For the same reason, this study analyzes at the two-digit industry level instead of four. If I analyze at the four-digit industry level, each zip code will have on average less than two firms. Thus, spillovers cannot be identified at the four-digit level.

[19] Appendix section 4 shows that the geographical distribution of the share of regulated industries is continuous across the dialect borders.

5 million RMB. One limitation of the data is that firms from HMT are put into one category. Ideally, I should address changes of HMT firms at the Cantonese border as well as changes of Taiwan firms at the Min dialect border. However, given the data structure, it is impossible to separate Hong Kong and Macau firms from Taiwan firms. Therefore, I make an additional assumption that investment from Taiwan is not affected by the Cantonese border and investment from Hong Kong and Macau is not affected by the Min dialect border.[20] Then, I combine the analysis of the two borders into one framework and address the changes in investment from Hong Kong, Macau and Taiwan at the Cantonese and the Min dialect borders. If the assumptions were true, the changes at the Cantonese border reflect changes in investment from Hong Kong and Macau while changes at the Min dialect border reflect changes in investment from Taiwan. Firm level variables such as output, value added, employment and real capital are calculated following the framework and deflators provided by Brandt et al. (2012).

Then, the firm level data is aggregated into zip code level.[21][22] Road distance from each zip code to both dialect borders are calculated. I use the smaller distance among the two to define the distance from a specific zip code to the nearest dialect border. For each zip code, I calculate

[20] This assumption can't be directly tested. However, its validity is strengthened by the placebo tests introduced later in this paper, which shows that investment from other foreign countries don't change discontinuously across the borders. Thus, I also expect that investment from a region that is not related to the specific dialect shouldn't be affected by the border. Under this assumption, the measurement error of the dependent variable will be independent from common dialect. Therefore, the coefficient on speaking the same dialect will not be biased.

[21] I use the geographical locations of current zip codes in the analysis. There could be some changes in geographical locations of some zip codes across years, but I don't have information to track the changes in zip codes across years. As a robustness check, I use the geographical locations of zip codes at 2005 (Data provided by Michigan China Data Center) and find similar empirical results. (The 2005 version has less zip codes and therefore significantly reduces the sample size).

[22] The geographical distribution of zip codes passes the RD density test by Cattaneo, et al. (2016) (T-value 0.81; P-value 0.42) at the dialect borders. Therefore, there is no evidence of manipulating the distribution of zip codes according to the location of dialect borders.

the distance to Hong Kong, if the nearest border is the Cantonese border and the distance to Taipei if the nearest border is the Min dialect border.

4. Empirical Strategies

This section shows the empirical strategies to identify spillovers from FDI. First, I use regression discontinuity design to identify the discontinuous increase in HMT investment at the dialect borders. Second, I show how the discontinuous increase in HMT investment varies by regulation status of industries and productivity of HMT firms. Third, I use the discontinuous increase in HMT investment at the dialect borders as an instrument to identify horizontal spillovers from FDI. Finally, I discuss additional tests to check the validity of identification assumptions.

4.1. Estimate the effects of common dialect on investment from HMT (zip code level analysis)

First, I use the spatial regression discontinuity design to estimate the discontinuous increase in investment from HMT at the borders of the two dialect zones.

The dependent variable is constructed by aggregating firm level data and calculating the share of HMT firms among all firms in zip code *i* and year *t*. The employment share of HMT firms is calculated using the following equation:

$$HMT\_share_{it,employment} = \frac{\sum_{f\in\Omega_{it}} HMT_{fit}\times Employment_{fit}}{\sum_{f\in\Omega_{it}} Employment_{fit}}, \quad (1)$$

where $Employment_{fit}$ measures the total employment of firm f in zip code *i* and year *t*; HMT_{fit} measures the HMT equity share of firm f and Ω_{it} is the set of all firms in zip code i and year t. The output share of HMT firms is calculated in a similar way using equation (2) as an alternative measure of the share of HMT firms. Table 1 shows that the average share of HMT firms in the sample is about 26 percent.

$$HMT\_share_{it,output} = \frac{\sum_{f\in\Omega_{it}} HMT_{fit}\times Output_{fit}}{\sum_{f\in\Omega_{it}} Output_{fit}}. \quad (2)$$

The empirical model of the spatial regression discontinuity design is specified as:

$$HMT\_share_{it} = \beta_0 + \beta_1 T_i + f(D_i) + T_i \times f(D_i) + f(DC_i) + c_j + \eta_t + \epsilon_{it}, \quad (3)$$

where i denotes zip code, *t* denotes year and *j* denotes county. HMT_share_{it} denotes the share of HMT firms in zip code *i* and year *t*. T_i is a dummy variable indicating whether zip code *i* is located inside the common dialect zones. D_i is the road distance from zip code *i* to the nearest dialect border. $f(D_i)$ are polynomials of the road distance D_i. $f(DC_i)$ are polynomials of the distance from zip code *i* to Hong Kong (if the nearest dialect border is Cantonese)[23] or Taipei (if the nearest dialect border is Min dialect). c_j are county fixed effects and η_t are year fixed effects. Thus, β_1 captures the discontinuous changes in the share of HMT firms at the dialect borders.

I estimate the polynomial model specified as equation (3) with a bandwidth of 40 kilometers, because 40-kilometer is approximately the size of one county in terms of road distance in China. Within the bandwidth, about 63 percent of the zip codes are inside the common dialect borders. As robustness checks, I also estimate a local linear regression model with bandwidth optimally chosen. The model is specified as equation (4):

$$HMT\_share_{it} = \theta_0 + \theta_1 T_i + \theta_2 D_i + \theta_3 T_i \times D_i + f(DC_i) + c_j + \eta_t + \epsilon_{it}, \quad (4)$$

in which linear models are used to model distance to the dialect borders and optimal bandwidth is chosen by the cross-validation method proposed by Imbens and Lemieux (2008).

4.2. Heterogeneous effects across industries by industrial characteristics

Next, I study the heterogeneous effects of common dialect on HMT investment across industries to strengthen my identification strategy. First, I compare those industries that are under FDI entry regulation with those industries that are not under regulation and expect to find that the

[23] Given that the share of Macau firms is very small compared to the share of Hong Kong firms and Macau and Hong Kong are geographically close, I only control distance to Hong Kong when evaluating zip codes located near the Cantonese dialect border.

discontinuous increase in HMT investment is larger in magnitude when the industries are not under regulation because of the lower entry cost. Second, I compare those industries where HMT firms are more productive with those industries where HMT firms are less productive and expect to find that the discontinuous gap is larger in those industries where the productivity of HMT firms is higher.

To test the heterogeneous effects by regulation, I estimate a model following spatial regression discontinuity design at the firm level with interactions between dialect borders and regulations:

$$HMT_{fikt} = \delta_0 + \delta_1 T_i + \delta_2 R_{kt} \times T_i + \delta_3 R_{kt} + f(D_i) + T_i \times f(D_i) + f(DC_i) + Z_{fikt} + c_j + \eta_t + \delta_k + \epsilon_{it}\,, \quad (5)$$

where the dependent variable HMT_{fikt} denotes the HMT equity share of firm f in zip code *i*, industry *k* and year *t*. R_{kt} is a dummy variable indicating whether industry *k* in year *t* receives FDI entry-regulation or not.[24] T_i, $f(D_i)$ and $f(DC_i)$ are defined in the same way as equation (3). Z_{fikt} are firm level control variables. c_j, η_t and δ_k are county, year and industry fixed effects respectively. The parameter of interest is δ_2 and I expect δ_2 to be negative, indicating that the discontinuous increase in HMT investment at the dialect borders is smaller in industries that receive FDI entry-regulations. This hypothesis indicates that the discontinuous increase in HMT investment at the dialect borders responds to a policy that specifically targets FDI. Therefore, the discontinuous increase at the dialect borders is more likely to be caused by FDI instead of other unobservable factors.

[24] The Chinese government lists goods that are either "restricted for FDI" or "forbidden to FDI". If a good that is produced by industry j is listed as either "restricted" or "forbidden" in year t, I code R_{jt} as 1. Goods that are listed as "forbidden" do not change much over time. Goods listed as "restricted" saw some changes upon China's accession to WTO.

Then, to test heterogeneous effects by the productivity of HMT firms, I estimate a model similar to model (5) with interactions between dialect borders and the productivity of HMT firms:

$$HMT_{fikt} = \rho_0 + \rho_1 T_i + \rho_2 HMTP_k \times T_i + f(D_i) + T_i \times f(D_i) + f(DC_i) + Z_{fikt} + c_j + \eta_t + \delta_k + \epsilon_{it} \ , \quad (6)$$

where $HMTP_k$ is the average TFP (defined in section 4.2) of HMT firms in industry *k* calculated using all HMT firms operating in mainland China.[25] The parameter of interest is ρ_2 and I expect ρ_2 to be positive, indicating that in those industries where the productivity of HMT firms is higher, HMT firms are more likely to enter the market. Therefore, the discontinuous increase at the dialect borders will also become larger in magnitude.

4.3. Identify horizontal spillovers from HMT investment

Third, I use the discontinuous increase in the share of HMT firms at the dialect borders as the exogenous variation to estimate the effects of HMT investment on the productivity of domestic firms in the same industry and location in an instrumental variable framework.

The empirical model is estimated using a sample of all domestic firms. A firm is defined as a domestic firm if foreign (including HMT) equity share equals to zero.[26]

The dependent variable is the total factor productivity (TFP) of each firm. To calculate TFP of a specific firm, I firstly follow the method proposed by Brandt et al. (2012). TFP is calculated as:

$$\ln(TFP_{ft}) = (q_{ft} - \bar{q_t}) - \widetilde{S_{ft}}(l_{ft} - \bar{l_t}) - (1 - \widetilde{S_{ft}})(k_{ft} - \bar{k_t}), \quad (7)$$

[25] Using data on all HMT firms operating in mainland China, I calculate the average TFP of HMT firms for each industry. Then this measure at the industry level is applied to the regions around dialect borders to measure the advantage of HMT firms in each industry.

[26] The main empirical results are similar if domestic firm is defined as foreign equity share smaller than 10 percent.

where q_{ft}, l_{ft} and k_{ft} are logarithms of firm f's value added, labor and capital in year t. $\bar{q}_t$, $\bar{l}_t$ and $\bar{k}_t$ are industry average logarithms of value added, labor and capital in year *t*. Labor is weighted by $\widetilde{S_{ft}}$, which denotes the share of wage in total value added, while capital is weighted by $(1-\widetilde{S_{ft}})$. $\widetilde{S_{ft}}$ is calculated as $\widetilde{S_{ft}} = (S_{ft} + \bar{S}_t)/2$, where S_{ft} is firm f's share of wage in total value added in year t and $\bar{S}_t$ is industry average share of wage in total value added in year t. In this specification, each firm is compared with a hypothetical average firm in the industry and productivity deviation from the average firm is captured by TFP. Table 1 shows that TFP estimated following this method has a mean close to zero and a standard deviation around 1.

Since the share of labor in the production function ($\widetilde{S_{ft}}$) is endogenously chosen by firms, TFP measured as equation (7) may be biased. Therefore, I also follow the framework developed by Olley and Pakes (1996) and Ackerberg et al. (2015) to estimate an unbiased measure of $\widetilde{S_{ft}}$ using non-parametric methods. Appendix section 1 shows the detailed procedures to get this alternative measure of TFP.

Then, I estimate spillovers from HMT investment to the TFP of domestic firms following the empirical model developed by Aitken and Harrison (1999). The baseline model has the following form:

$$\ln(TFP_{fikt}) = \gamma_0 + \gamma_1 HMT\_share\_Horizontal_{ikt} + Z_{fikt} + c_j + \eta_t + \delta_k + \epsilon_{it}, \quad (8)$$

where the dependent variable $\ln(TFP_{fikt})$ is the TFP of firm f in zip code i, industry k and year t and the key independent variable $HMT\_share\_Horizontal_{ikt}$ measures the presence of HMT firms in the same industry and location as firm f. Thus, γ_1 captures horizontal spillovers from HMT firms to domestic firms. Following the literature, $HMT\_share\_Horizontal_{ikt}$ is measured by the output share of HMT firms in each industry, zip code and year:

$$HMT\_share\_Horizontal_{ikt} = \frac{\sum_{m\in\Omega_{ikt}} HMT_m \times Output_m}{\sum_{m\in\Omega_{ikt}} Output_m} \ , \quad (9)$$

where HMT_m denotes the HMT equity share of firm m; $Output_m$ denotes total output of firm m; Ω_{ikt} denotes all firms in zip code i, industry k and year t. In short, the presence of HMT firms is measure by HMT equity share weighted total output over total output. Z_{fikt} are firm level control variables. c_j, η_t and δ_k are county, year and industry fixed effects respectively. Specifically, I control for the following firm level variables: logarithm of total output, labor capital ratio, logarithm of total export, the number of years since the firm was established and the output share of other foreign firms in the same location and industry. Total output and labor capital ratio controls for the effect of economy of scale on productivity; total export controls for the effect of trade on productivity, because regions from the common dialect zones may have more opportunities to export to HMT; number of years in business controls for the effect of firms' experience in the industry on productivity; output share of other foreign firms controls for spillovers from other types of FDI. [27]

To solve the problem that HMT firms endogenously choose the locations and industries that they enter, I instrument the presence of HMT investment using the discontinuous increase at the dialect borders. The discontinuous changes in HMT investment at the dialect borders is modeled in a similar way to equation (3) but at the firm level:

$$HMT\_share\_Horizontal_{ikt} = \alpha_0 + \alpha_1 T_i + f(D_i) + T_i \times f(D_i) + f(DC_i) + Z_{fikt} + c_j + \eta_t + \delta_k + v_{it} \ , \quad (10)$$

where the dependent variable $HMT\_share\_Horizontal_{ikt}$ is the measure of the presence of HMT firms in zip code i, industry k and year t specified as the key independent variable in

[27] Adding these control variables does not affect the magnitude of the main results significantly.

equation (8); T_i is a dummy variable indicating whether zip code *i* is located inside the common dialect zones; D_i is the road distance from zip code *i* to the nearest dialect borders; $f(D_i)$ are polynomials of the road distance D_i. The other controls are the same as equation (8). In this model, α_1 captures the discontinuous increase in the presence of HMT firms at the dialect borders.

Next, in the second stage, I use the predicted presence of HMT firms from equation (10) in place of the actual presence of HMT firms in equation (8):

$$\ln(TFP_{fikt}) = \sigma_0 + \sigma_1 \widehat{HMT\_share\_Horizontal}_{ikt} + f(D_i) + T_i \times f(D_i) + f(DC_i) + Z_{fikt} + c_j + \eta_t + \delta_k + u_{it} . \quad (11)$$

The coefficient of interest, σ_1 captures the changes in the productivity of domestic firms due to the discontinuous increase in HMT investment at the dialect borders.

4.4. Identification assumptions and additional tests

Identification of these empirical models requires two important assumptions and this section will discuss tests of their validity. First, factors other than dialect should change continuously at the borders such that the discontinuous changes in the share of HMT firms is only caused by variation in dialect. Second, common dialect affects the productivity of domestic firms only through increasing investment from HMT. I will conduct five tests to check the validity of these assumptions.

First, I check whether observable non-economic covariates change continuously at the dialect borders. I check two sets of variables that can be acquired at the zip code level in China: demographic variables from the population census 2010 and geographic variables from corresponding maps. Specifically, I check the geographical distribution of the following

variables at the dialect borders: total population, the share of people under 14, the share of people above 65, the share of people who hold local *hukou*,[28] elevation and slope.

Second, it is possible that people on one side of the dialect borders attract more investment because of differences in investment opportunities. To address this concern, I use the share of foreign firms from countries other than HMT (for example Japan and Korea) as the dependent variables in equation (3) and (4). I expect to find no significant discontinuity, indicating that regions inside the borders attract more FDI only from HMT.

Third, language borders can generate unobservable variation other than culture. To address this concern, I analyze changes in the share of HMT firms at a placebo dialect border that is correlated with neither Cantonese nor Min dialect. I choose the Wu dialect border around Shanghai (shown in Figure 11) to conduct this analysis because regions at the Wu dialect border is geographically close and economically comparable to regions at the Cantonese and Min dialect borders. Continuous changes at the Wu dialect border indicate that the language border itself, without the specific cultural meaning attached, does not generate direct effects on the outcome variables.

Fourth, I create additional placebo dialect borders by moving the actual dialect borders inward and outward arbitrarily by 100 kilometers. I expect to find continuous changes in both the share of HMT firms and the productivity of domestic firms at these placebo dialect borders.

Finally, culture may exert direct effects on productivity instead of affecting productivity through attracting more investment from HMT. Thus, I conduct another placebo test and check whether the productivity of domestic firms, which are from industries that receive the least

[28] People who hold local *hukou* in China indicates those people who are registered as local people. The difference between total population and people of local *hukou* implies people who migrated into the zip code. Therefore, the population share of local *hukou* can be used to capture the degree of in-migration.

influence from HMT, changes discontinuously at the borders. Similarity in the productivity of these firms indicates that the productivity of firms might be similar without investment from HMT.

5. Empirical Results

5.1. Results from graphs

Before reporting detailed estimation results from the empirical models, I first show graphical evidence of discontinuous changes in the share of HMT firms at the dialect borders. Figure 6 shows the geographical distribution of the share of HMT firms by distance to the dialect borders. This graph shows the local linear fit of the data with optimal bandwidth chosen to be 30 kilometers.[29] In this figure, the horizontal axis indicates distance to the dialect borders with negative value indicating inside the common dialect borders (common dialect with HMT). The vertical axis denotes the employment share of HMT firms. From Figure 6, we can clearly identify a significant discontinuous drop in the share of HMT firms when moving from inside the borders to outside the borders.

A potential problem of drawing conclusions only from Figure 6 is that the discontinuous changes at the borders are a mixture of cultural effects and policy effects because part of the dialect borders coincide with the administrative borders. To exclude the effects from policy variation at the administrative borders, I standardize the outcome variables by subtracting county mean and dividing by county standard deviation.[30] This process will remove policy effects that are common to every zip code from the same county and allow the author to conduct within-county comparisons. Figure 7 shows the discontinuous changes in the share of HMT firms with

[29] Appendix section 11 shows that the discontinuous increase in the share of HMT firms at the dialect borders is robust in magnitude when we change the bandwidth from 20 km to 60 km.
[30] A county is a larger geographical area than a zip code. Each county may contain multiple zip codes.

standardized outcome variables. Figure 7.1 shows the distribution of employment share and Figure 7.2 shows the distribution of output share respectively.

From Figure 7.1, we can identify a significant drop in the standardized share of HMT firms at the dialect borders when moving from inside the borders to outside. Due to the process of standardization, the value of points that are farther away from the dialect borders become close to zero.[31]This is because the distribution of the average share of HMT firms at the county level is downward sloping. If a location is farther away from the dialect border, the location is more likely to be from a county where a large proportion of zip codes are inside (outside) the common dialect borders and therefore the average share of HMT firms is high (low). Only from those locations that are very close to the dialect borders, can we observe a significant drop in the share of HMT firms across the dialect borders. Figure 7.2 shows similar empirical patterns when I use the output share as an alternative measure of the share of HMT firms.

In Figure 8, I use the fourth-degree polynomials of distance to the dialect borders to fit the share of HMT firms (standardized). Similar to Figure 8, I observe a discontinuous drop in the share of HMT firms at the borders of the common dialect zones.

As one of the placebo tests, Figure 9 investigates whether we can observe similar discontinuous changes in the share of foreign firms from regions other than HMT at the dialect borders. Using the same specification as Figure 7, Figure 9 shows no discontinuous changes in the share of firms from other foreign countries, indicating that common dialect increases FDI through attracting more investment only from HMT.

[31] If cultural borders had no effects at all, the graph of the standardized distribution should look like a flat line. All points should fluctuate around 0 due to standardization.

5.2. Main results

This section shows that common dialect, which represents unique cultural ties, increases investment from HMT to mainland China by around 5 to 7 percentage points at the dialect borders. The increase in the investment from HMT also leads to the improvement in the productivity of domestic firms from the same industry and location, especially in industries where the productivity of HMT firms is higher than domestic firms.

Table 2 reports the effects of common dialect on investment from HMT estimated using the spatial regression discontinuity design (equation [3] and [4]) at the zip code level. I report the coefficients on the dummy variable indicating common dialect (β_1) in the table, which captures the discontinuous changes in the outcome variables at the dialect borders. I use employment share as the dependent variable in Panel A and output share in Panel B respectively. Column (1) to Column (4), which are estimated using equation (3), represent separate estimations with the first to fourth degree polynomials of distance to the dialect borders as control variables. The bandwidth is chosen to be 40 kilometers, which is about the size of one county in China. Column (5) is estimated using a local linear model specified as equation (4). The optimal bandwidth chosen by cross-validation is 30 kilometers. I also include county fixed effects, year fixed effects, and polynomials of distance to Hong Kong (or Taipei) as control variables in all specifications

Panel A of Table 2 shows that common dialect discontinuously increases the employment share of HMT firms by around 5 to 7 percentage points at the borders of the dialect zones, which is about a 20 percent increase. The effects are similar in magnitude across all different specifications, even though less statistically significant when higher degree polynomials are

included as control variables.[32] Panel B shows that the output share of HMT firms increases by around 4 to 7 percentage points at the dialect borders.

Next, in Table 3, I show the heterogeneous effects of common dialect by industries. Panel A shows the effects by FDI regulation and Panel B shows the effects by the productivity of HMT firms. The models are estimated at the firm level using equation (5) and (6). The dependent variable is the HMT equity share of a specific firm. I report the coefficients on the dummy variable indicating common dialect (δ_1 and ρ_1) as well as the coefficients on the interaction terms ($\delta_2 \; and \; \rho_2$). Similar to Table 1, Column (1) to Column (4) represent separate estimations with the first to fourth degree polynomials of distance to the dialect borders as control variables. Column (5) is estimated with a local linear model where the bandwidth is chosen to be 30 kilometers.

From panel A of Table 3, we can conclude that the discontinuous increase in HMT equity share at the dialect borders is smaller when the entry cost of HMT firms, approximated by entry-regulation, is higher. First, the coefficients on the common dialect dummy variable are generally positive, indicating that common dialect discontinuously increases the HMT equity share of industries that are not under entry-regulation. Second, the coefficients on the interaction term between the common dialect dummy variable and the entry-regulation dummy variable are negative, indicating that the HMT equity share does not increase as much at the borders when the entry cost is higher due to the entry-regulation policy.

From panel B of Table 3, we can conclude that the discontinuous increase in HMT equity share at the dialect borders is larger in industries where the productivity of HMT firms is higher. This finding is supported by the positive and statistically significant coefficients on the

[32] The AIC information criterion reported in the table strictly prefers the local linear model, in which the discontinuous gap at the dialect borders is more efficiently estimated.

interaction term between common dialect and the productivity of HMT firms at the industry level. When HMT firms have an advantage in specific industries, they are more likely to enter those industries and therefore lead to a larger discontinuous increase across the dialect borders.

With the discontinuous increase in the share of HMT firms across the dialect borders well established, I move on to use the discontinuous increase in HMT investment across the dialect borders as the exogenous variation to identify spillovers to domestic firms.

Before showing the results from the full model with instruments, I firstly estimate the reduced form model, which captures the direct effects of common dialects on the productivity of domestic firms across the dialect borders. The models are estimated at the firm level using equation (10) with the TFP of domestic firms as the dependent variable. Since knowledge spillovers only happen when HMT firms have an advantage in productivity over domestic firms, I categorize industries based on the productivity of HMT firms relative to domestic firms[33] and estimate the model separately for the two groups of industries. Table 4 shows the estimation results. As expected, Column (3) and (4) shows that in industries where HMT firms are more productive than domestic firms, common dialect increases the productivity of domestic firms at the dialect borders. In contrast, Column (1) and (2) shows that in industries where HMT firms are less productive than domestic firms, the productivity of domestic firms is very similar across the dialect borders. The findings indicate that cultural ties increase the productivity of domestic firms, but only in industries where HMT firms have a productivity advantage over domestic firms. Therefore, the effects of cultural ties on the productivity of domestic firms is more likely

[33] Using the sample of firms from all over China, I calculate the average productivity of HMT firms and domestic firms for each two-digit industry and then categorize industries into industries where HMT firms are more productive and industries where domestic firms are more productive.

to be through attracting more investment from HMT instead of through the direct effects of different cultures.

Combining the discontinuous increase in the share of HMT firms and the discontinuous increase in the productivity of domestic firms across the dialect borders, I estimate the magnitude of horizontal spillovers using equation (11) [34][35]. In this specification, the discontinuous increase in the share of HMT firms at the dialect borders is used as an instrument for the presence of HMT firms. In terms of model selection, according to the results from Table 2, the local linear model shows the largest statistical power to identify the discontinuous increase at the dialect borders. Therefore, I use the local linear model as the first stage to avoid potential problems of week instruments. Column (1) and (2) use a sample of all domestic firms within 30 km of the dialect borders. Column (3) and (4) use a sample of domestic firms from industries where the productivity of HMT firms is higher than domestic firms. In Column (1) and (3), I use TFP estimated following Brandt et al. (2012) as the dependent variable, while in Column (2) and (4), I use TFP estimated non-parametrically following Ackerberg et al. (2015). All models also include the following control variables: the output share of other foreign firms, age of the firms, logarithm of total output, labor capital ratio and logarithm of total export.[36]

In Table 5, the horizontal-spillover coefficients σ_1 in equation (11) are estimated to be positive in all models but only statistically significant in the sample of domestic firms that are from industries where HMT firms are more productive than domestic firms. In industries where

[34]In addition to horizontal spillovers, industrial variation in regulation policies interacted with spatial discontinuity in dialects make it possible to also estimate vertical spillovers (spillovers to domestic suppliers and buyers). In the full model with both horizontal and vertical spillovers estimated, vertical spillovers are not precisely estimated (negative but statistically insignificant) due to very weak first stage results. The results of vertical spillovers are shown in Appendix section 6.

[35] We can also estimate the spillovers on other foreign firms. However, the sample size on other firms are too small to get a reliable estimator.

[36] The major results are not significantly affected by including these additional control variables.

HMT firms have an advantage in productivity over domestic firms, we find positive local horizontal spillovers: a 1 percentage point increase in the share of HMT firms raises the productivity of domestic firms in the same zip code and industry by around 1.7 (estimated following Ackerberg et al. [2015]) to 2.8 (estimated following Brandt et al. [2012]) percent. The spillovers are not statistically significant in the whole sample, because there are no productivity spillovers in industries where the productivity of HMT firms is lower than domestic firms.

The identification strategy employed in this study significantly affects the empirical estimates of the horizontal spillovers. Table 6 shows that if the models in Table 5 are estimated without using instruments, the coefficients on horizontal spillovers are negative in most specifications. This negative correlation can be driven by the selective entry of HMT firms into industries where the productivity of domestic firms is low. It is only after correcting for the endogenous entry problem using instruments that I identify the actual spillovers to be positive

In terms of magnitude, the horizontal spillovers estimated in this study is much larger than previous studies that do not solve the endogenous entry problem of FDI. However, the magnitude of effects is in line with the local horizontal spillovers estimated in Lu et al. (2017) (6.644), in which the authors use changes in regulation policy to identify the causal effect of FDI on the productivity of domestic firms. My estimates strengthen the argument that the OLS estimates tend to be biased and a proper identification strategy is needed when estimating the spillovers from FDI.

5.3 Mechanisms of the positive horizontal spillovers

This section discusses possible mechanisms that are driving the positive horizontal spillovers from HMT investment to domestic firms. The horizontal spillovers estimated in this study seem to be inconsistent with the average results from the meta-analysis (Irsova and

Havranek, 2013), which concludes that horizontal spillovers are on average zero. However, the authors also indicate that "the sign and magnitude of spillovers depend systematically on the characteristics of domestic economy and foreign investors". Therefore, the sign of spillovers seems to be quite sensitive to the economic context and several mechanisms, namely cultural similarity, local effects, and crowding-out effects, can explain why positive horizontal spillovers are reasonable in this specific context.

First, when the source country of FDI is similar to the host country in terms of culture, domestic firms are likely to adopt foreign technology more easily (Crespo and Fontoura, 2007). This argument applies perfectly to the context of this study, where the host regions are culturally very similar to HMT. Common culture plays the role of a mediating factor, which can reduce the cost of knowledge spillovers. Thus, firms in regions that are culturally similar to HMT are more likely to benefit from the presence of HMT investment.

Second, in this study, I estimate local spillovers in a very small geographical region, therefore the results can be very different from national spillovers estimated in the literature. As shown by previous studies, distance matters when estimating horizontal spillovers. For example, Halpern and Murakozy (2007) find no evidence of horizontal spillovers at the country level. However, when the authors take distance into account, they find positive horizontal spillovers to domestic firms that are close to foreign firms. In the context of China, Xu and Sheng (2012) and Lu et al. (2017) also find that horizontal spillovers turn from negative to positive if the analysis is restricted from the national level to the regional level. The authors argue that positive horizontal spillovers work through knowledge spillovers and labor pooling, which are more likely to happen when domestic firms are located close to foreign firms. In contrast, negative horizontal

spillovers are usually due to competitions in the product market, which is more integrated at the national level.

Finally, investment from HMT could have crowded out less productive domestic firms from the same industry. Thus, the remaining domestic firms could be relatively more productive. It is established in the literature that domestic firms can be crowded out of market due to the presence of FDI (Kosova, 2010). I also find similar empirical patterns in this study. Table 7 shows the estimation results of equation (3) and (4) with the share of domestic firms as the dependent variables. We can conclude that the share of domestic firms decreases at the dialect borders as the share of HMT firms increases.[37] The decrease in the share of domestic firms may not directly indicate crowding-out effect because it is possible that the decrease in the share of domestic firms is merely caused by the increase in the level of HMT firms, while the level of domestic firms stays the same. As an additional evidence of crowding out effect, Appendix section 5 shows the estimation results of equation (4) with the level of total employment and output (instead of share) as the dependent variables. I find that the level of domestic firms in terms of total employment and output decreases at the dialect borders. At the same time, the level of HMT firms increases by a similar amount and the level of other foreign firms stays the same. Therefore, there is indeed suggestive evidence on the crowding-out of domestic firms. However, due to the cross-sectional nature of my empirical strategy, I cannot track the productivity of those domestic firms that are crowded out of the market.

5.4 Additional Tests

To show the validity of the identification assumptions specified in section 4, I conduct several additional tests. First, I show that observable non-economic covariates, such as demographic and

[37] Appendix section 12 shows that the decrease in the share of domestic firms is larger in those industries that do not receive FDI entry regulation and those industries where the productivity of domestic firms is higher.

geographic characteristic change continuously at the dialect borders. Second, I show that investment from foreign countries other than HMT does not change discontinuously at the borders. Third, I find that investment from HMT does not change discontinuously at unrelated dialect borders. Fourth, investment from HMT does not change discontinuously when the actual dialect borders are arbitrarily moved. Fifth, the productivity of domestic firms from industries without investment from HMT are similar across the dialect borders. The evidence jointly supports the main identification assumptions of this study.

Figure 9 shows the geographical distribution of demographic and geographic characteristics of zip codes across the dialect borders.[38] From figure (a), I conclude that the total population are similar across the dialect borders. Therefore, zip codes inside the borders do not contain more population centers than zip codes outside the borders. Figure (b) and figure (c) jointly show that the population structure in terms of age is balanced across the dialect borders. Figure (d) shows that the proportion of local people is similar across the borders, indicating that zip codes inside the borders do not attract more in-migrant workers than zip codes outside the borders. Finally, figure (e) and figure (f) jointly show that the geographical conditions that may affect the productivity of firms, such as elevation and slope, are balanced across the dialect borders.[39]

Table 8 shows the effects of the dialect borders on investment from other foreign countries. I report the estimation results of equation (3) and (4) with the share of other foreign firms as the dependent variables. Model specifications are the same as Table 2. As expected, I do not find statistically significant discontinuity at the dialect borders for most of the models. The

[38] The geographical distribution of average temperature and precipitation are shown in Appendix section 13, even though the data on temperature and precipitation are not as detailed as other variables.

magnitude of coefficients is also very close to zero. Thus, I can claim that regions inside the dialect border do not generally attract more FDI. Instead, they attract more FDI only from HMT.

Table 9 shows the effects of an alternative dialect border, which is unrelated to Min and Cantonese culture, on HMT investment. This exercise shows that language border itself does not generate any effects on investment from HMT. Table 9 shows the estimation results of equation (3) and (4) with the dialect border specified as the Wu dialect border (location shown in Figure 11).[40] In most model specifications, the Wu dialect border generates no statistically significant effect on investment from HMT. The magnitude of coefficients is close to zero and much smaller than those of Table 2.

Table 10 and Table 11 show the results from creating hypothetical placebo dialect borders by moving the actual dialect borders outward or inward by 100 kilometers. Table 10 replicates the results of Table 2 by estimating equation (3) and (4) using the placebo dialect borders. As expected, none of these hypothetical dialect borders generate any discontinuous changes in the employment share of firms from HMT. The estimated discontinuous increase in the share of HMT firms at the placebo dialect borders are not only statistically insignificant but also very small in magnitude. Table 11 estimates the discontinuous changes in the TFP of domestic firms at the placebo dialect borders. I estimate equation (10) with TFP of domestic firms as the dependent variables. Similar to the results from Table 9, none of the placebo dialect borders generate statistically significant discontinuous changes in the TFP of domestic firms. Also, the discontinuous increase in the TFP of domestic firms is not higher in those industries where the productivity of HMT firms is higher. These results suggest that investment from HMT

[40] A figure showing the geographical distribution of HMT investment across the Wu dialect border is shown in Appendix section 3.

and the TFP of domestic firms change continuously when we move away from the dialect borders.

Finally, Table 12 supports the assumption that the productivity of domestic firms had been similar across the dialect borders without investment from HMT. To investigate this hypothetical question, I estimate the direct effect of common dialect borders on TFP using a sample of domestic firms from industries that receive the least influence from HMT. I report the coefficients on the common dialect dummy variable in the table. I find that when the share of HMT firms in the industries is zero percent, the difference in TFP across the dialect borders is very close to zero. Therefore, the productivity of domestic firms should have been similar across the borders if there were no investment from HMT. As the share of HMT firms increases from zero to four percentage points, we gradually start to observe an increase in the productivity gap across the dialect borders. Yet the difference in TFP is still too small to be statistically identified as different from zero.

6. Conclusions and Discussions

In this study, I show that foreign investment from Hong Kong, Macau and Taiwan (HMT) to mainland China generates positive local horizontal spillovers to raise the productivity of domestic firms in industries where HMT firms are more productive than domestic firms. The causal effect is identified through exploring discontinuous changes in HMT investment at the borders of Chinese dialect zones. My empirical estimates show that the horizontal spillovers in previous studies might have been wrongly estimated in sign or significantly underestimated in magnitude. Due to potential endogeneity concerns, better empirical designs are required in the literature of estimating spillovers from FDI to domestic firms.

This study finds that spillovers are larger in magnitude when foreign firms are culturally similar to the domestic firms. This finding implies that policy makers should not only design policies to attract FDI, but also consider matching foreign firms with domestic firms that are similar in certain aspects (for example culture) such that domestic firms can benefit more from the spillovers from foreign firms.

Even though some domestic firms could have benefited from spillovers from HMT firms, I also find evidence showing that some domestic firms might have been crowded out of the market due to the entry of HMT firms. Therefore, the entry of HMT firms does not benefit all domestic firms. Thus, the overall welfare effects on domestic firms is mixed

One major limitation of this study is external validity. The identification strategy only allows the estimation of spillovers from HMT firms. However, the spillovers from HMT firms maybe systematically different from the spillovers from other foreign firms as shown by Lin et al. (2009) and Du et al. (2012). Both studies find that non-HMT foreign firms (primarily from OECD countries) generate positive spillovers, yet HMT firms generate negative (close to zero) spillovers in China. The difference between HMT and non-HMT firms could be caused by biased estimates due to endogeneity concerns. However, I am not able to conclude that the unbiased spillovers from HMT firms are also different from non-HMT firms, because the identification strategy of this study does not allow the estimation of non-HMT firms. Therefore, we should be very cautious when trying to generalize the results of this study to investment from other foreign countries. Also, in this study, the destination location is culturally very similar to the source region of FDI, which may actually drive the positive spillovers. Therefore, the results may not be applied to a situation where the destination location is not similar to the source region.

In this study, I mainly focus on the analysis of horizontal spillovers from HMT investment. However, as shown by Melitz and Toubal (2014), spillovers are more likely to happen across industries, especially to suppliers of the foreign firms. Spillovers across industries are unlikely to be precisely identified from the empirical setting of this study, because I rely on regional variation (across the dialect border) in HMT investment to identify spillovers. It is unlikely that suppliers and buyers of a specific firm is restricted to the area where the firm is located.[41] Also, in my sample, many zip codes specialize in certain industries. Therefore, I don't observe enough cases of spillovers to other industries in the same geographical location. Thus, unable to precisely identify vertical spillovers is another important limitation of this strategy.

Finally, investment from HMT can generate spillovers to domestic firms across the dialect borders. As a result, domestic firms from zip codes that do not speak the same dialect as HMT, which are the control group of this study, might also be affected by spillovers to some degree. Thus, this study may underestimate and get a lower bound of the true effect of horizontal spillovers assuming that horizontal spillovers are also positive across the dialect borders.

[41] Results of vertical spillovers are shown in Appendix section 6.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, 2015, "Identification Properties of Recent Production Function Estimators." *Econometrica*, 83(6): 2411-2451.

Aitken, Brian J., and Ann E. Harrison, 1999, "Do Domestic Firms Benefit from Direct Foreign Investment? Evidence from Venezuela." *American economic review*, 89(3), 605-618.

Becker, Sascha O., Katrin Boeckh, Christa Hainz, and Ludger Woessmann, 2016, "The Empire is Dead, Long Live the Empire! Long-run Persistence of Trust and Corruption in the Bureaucracy." *The Economic Journal* 126(590), 40-74.

Brandt, Loren, Johannes Van Biesebroeck, and Yifan Zhang, 2012, "Creative Accounting or Creative Destruction? Firm-level Productivity Growth in Chines Manufacturing." *Journal of Development Economics*, 97(2), 339-351.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, 2016, "rddensity: Manipulation Testing Based on Density Discontinuity." *Stata Journal* (ii): 1-18.

Chen, Zhao, Ming Lu, and Le Xu, 2014, "Returns to Dialect: Identity Exposure through Language in the Chinese Labor Market." China Economic Review, 30, 27-43.

Chiswick, Barry, 2009, "The Economics of Language for Immigrants: An Introduction and Overview." *The Education of Language Minority Immigrants in the United States*, 72-91.

Crespo, Nuno, and Maria Paula Fontoura, 2007, "Determinant Factors of FDI Spillovers–what do We Really Know?" *World Development*, 35(3),410-425.

Dai, Yi Yi, Jin Li Xiao, and Yue Pan, 2016. “Can ‘Local Accent’ Reduce Agency Cost? Study Based on the Perspective of Dialects”, *Economic Research*, 51(012), 147-160. (Chinese)

Du, ruofu and Yi Da Yuan, 1993, “Evolution of Chinese Surnames and Distribution of Chinese Surnames across Dialect Zones”, *Chinese Social Science*, 4, 177-190 (Chinese).

Egger, Peter H., and Andrea Lassmann, 2012, "The Language Effect in International Trade: A Meta-analysis." *Economics Letters*, 116(2), 221-224.

Egger, Peter H., and Andrea Lassmann, 2015, "The Causal Impact of Common Native Language on International Trade: Evidence from a Spatial Regression Discontinuity Design." *Economic Journal*, 125(584), 699-745.

Falck, Oliver, Stephan Heblich, Alfred Lameli, and Jens Südekum, 2012, "Dialects, Cultural Identity, and Economic Exchange." *Joural of Urban Economics*, 72(2), 225-239.

Felbermayr, Gabriel J., and Farid Toubal, 2010, "Cultural Proximity and Trade." *European Economic Review*, 54(2), 279-293.

Gao, Xiang, and Cheryl Xiaoning Long, 2014,"Cultural Border, Administrative Border, and Regional Economic Development: Evidence from Chinese Cities." *China Economic Review*, 31, 247-264.

Gorg, Holger, and Eric Strobl, 2001,"Multinational Companies and Productivity Spillovers: A Meta-analysis." *Economic Journal*, 111(475),723-739.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, 2009,"Cultural Biases in Economic Exchange?" *Quarterly Journal of Economics*, 124(3), 1095-1131.

Halpern, László, and Balázs Muraközy, 2007, "Does Distance Matter in Spillover?" *Economics of Transition*, 15(4), 781-805.

Havranek, Tomas, and Zuzana Irsova, 2011, "Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect is." *Journal of International Economics*, 85(2): 234-244.

He Deng Song, 2003, *Chinese Linguistic Geography*, Shanghai Education Press (Chinese).

Hu, A Xiang, and Le Yao, 2011 “Discussion of Factors that Influence Cultural Zones in Jiangsu Province”, *Journal of Huaiyin Teachers College (Social Science)*, 33(3), 334-345 (Chinese).

Imbens, Guido W., and Thomas Lemieux, 2008, "Regression Discontinuity Designs: A Guide to Practice." *Journal of econometrics*, 142(2), 615-635.

Iršová, Zuzana, and Tomáš Havránek, 2013, "Determinants of Horizontal Spillovers from FDI: Evidence from a Large Meta-analysis." *World Development*, 42, 1-15.

Javorcik, Beata Smarzynska, 2004, "Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers through Backward Linkages." *American Economic Review* 94(3), 605-627.

Kim, Moonhawk, Amy H. Liu, Kim-Lee Tuxhorn, David S. Brown, and David Leblang, 2015, "Lingua Mercatoria: Language and Foreign Direct Investment." *International Studies Quarterly*, 59(2), 330-343.

Kosova, Renata, 2010, "Do Foreign Firms Crowd Out Domestic Frms? Evidence from the Czech Republic." *The Review of Economics and Statistics*, 92(4), 861-881.

Lin, Ping, Zhuomin Liu, and Yifan Zhang, 2009,"Do Chinese Domestic Firms Benefit from FDI Inflow?: Evidence of Horizontal and Vertical Spillovers." *China Economic Review*, 20 (4), 677-691.

Lowes, Sara, 2017. “Kinship Systems, Gender Norms, and Household Bargaining: Evidence from the Matrilineal Belt.” *Unpublished Manuscript.*

Lu, Yi, Zhigang Tao, and Lianming Zhu, 2017, "Identifying FDI Spillovers." *Journal of International Economics*, 107, 75-90.

Melitz, Jacques, 2008, "Language and Foreign Trade." *European Economic Review*, 52(4), 667-699.

Melitz, Jacques, and Farid Toubal, 2014, "Native Language, Spoken Language, Translation and Trade." *Journal of International Economics*, 93(2), 351-363.

Olley, G. Steven, and Ariel Pakes, 1992, "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica*, 64(6), 1263-1297.

Rauch, James E., and Vitor Trindade, 2002, "Ethnic Chinese Networks in International Trade." *Review of Economics and Statistics*, 84(1), 116-130.

Rodriguez-Clare, Andres, 1996, "Multinationals, Linkages, and Economic Development." *American Economic Review*, 86(4), 852-873.

Sauter, Nicolas, 2012, "Talking Trade: Language Barriers in Intra-Canadian commerce." *Empirical Economics* 42(1), 301-323.

Simmons Richard AanNess, Rujie Shi, and Qian Gu, 2006, *Chinse Dialect Geography: Distinguishing Mandarin and Wu in Their Boundary Region*, Shanghai Education Press (Chinese).

Wurm, Stephen A., Rong Li, and Theo Baumann, 1987. Language atlas of China. Australian Acad. of the Humanities; Longman Group (Far East).

Xu, Xinpeng, and Yu Sheng, 2012, "Are FDI Spillovers Regional? Firm-level Evidence from China." Journal of Asian Economics, 23(3), 244-258.

Zhang, Kevin Honglin, 2005, "Why does so Much FDI from Hong Kong and Taiwan Go to Mainland China?" *China Economic Review*, 16(3), 293-307.

Table 1: Summary Statistics

Variables	Observations	Mean	Standard Deviation
Panel A: Zip-code-level variables (Bandwidth: 40km):			
Common dialect	9456	0.62	0.48
Employment share of HMT firms	9456	0.26	0.35
Output share of HMT firms	9456	0.25	0.34
Employment share of other foreign firms	9456	0.09	0.20
Output share of other foreign firms	9456	0.10	0.22
Distance to dialect borders (km, absolute value)	9456	20.73	11.00
Panel B: Firm-level variables (All firms):			
Common dialect	77531	0.68	0.47
FDI regulation	77531	0.33	0.47
HMT equity share	77531	0.27	0.43
Panel C: Firm-level variables (Domestic firms):			
Ln(TFP)	33589	0.08	1.04
Ln(TFP) (Non-parametric)	33589	-0.01	0.87
Output share of HMT firms in the same industry and location	33589	0.08	0.19

Note: This Table shows the summary statistics of major variables. HMT refers to Hong Kong, Macau and Taiwan. Panel A shows variables used in zip-code-level analysis. Panel B shows variables used in Firm-level analysis. Panel C shows variables used in the estimation of spillovers of domestic firms. The sampling bandwidth of Panel A and Panel B are 40 km. The sampling bandwidth of Panel C is 30km.

Table 2: The Effects of Common Dialect on the Share of Firms from Hong Kong, Macau, and Taiwan (HMT)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent variable: Employment share					
Common Dialect	0.055***	0.075**	0.048	0.067	0.070***
	(0.021)	(0.033)	(0.047)	(0.063)	(0.024)
Optimal Bandwidth					30
AIC	733	725	727	726	42
Panel B: Dependent variable: Output share					
Common Dialect	0.046**	0.056*	0.065	0.044	0.054**
	(0.022)	(0.034)	(0.050)	(0.065)	(0.024)
Optimal Bandwidth					30
AIC	1164	1160	1163	1164	350
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				
Observations	9456	9456	9456	9456	7067

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of HMT firms. Models are estimated following the spatial regression discontinuity design using a sample of zip codes within 40 kilometers of the dialect borders. The dependent variable is the total employment (output) of HMT firms over total employment (output) of all firms for each zip code. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Column (1) to column (4) estimate model (3) by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with optimal bandwidth chosen using cross-validation. Panel A uses the employment share as the dependent variable. Panel B uses the output share as the dependent variable respectively. AIC refers to Akaike information criterion for each model.

Table 3: Heterogeneous Effects of Common Dialect on the Share of HMT Firms by Industry (FDI Regulation and Productivity of HMT firms)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Dependent Variable: HMT equity share					
Panel A: By whether the industry is under FDI entry-regulation					
Common Dialect	0.065*	0.085	0.56	0.057	0.095**
	(0.036)	(0.054)	(0.074)	(0.075)	(0.039)
Common Dialect*FDI Regulation	-0.053*	-0.054*	-0.055*	-0.055*	-0.036
	(0.029)	(0.029)	(0.029)	(0.029)	(0.028)
Bandwidth					30
Observations	77537	77537	77537	77537	58060
Panel B: By the productivity of HMT firms for each industry					
Common Dialect	0.12***	0.13**	0.12	0.13	0.18***
	(0.049)	(0.07)	(0.08)	(0.08)	(0.05)
Common Dialect*Produ -ctivity of HMT firms	0.33***	0.33***	0.33***	0.33***	0.40***
	(0.10)	(0.10)	(0.10)	(0.11)	(0.11)
Bandwidth					30
Observations	77466	77466	77466	77466	58000
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei				

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of HMT firms in different industries. Panel A compares the effects in industries under FDI entry-regulation with industries not under regulation. Panel B compares the effects in industries where the productivity of HMT firms is high with industries where the productivity of HMT firms is low. All models are estimated following spatial regression discontinuity design using a sample of firms within 40 kilometers of the dialect borders [model (5) and model (6)]. The dependent variable is HMT equity share for each firm. Column (1) to column (4) estimate the models by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. In panel A, I report the coefficients on the common dialect dummy variable and the coefficients on the interaction between common dialect and whether the industry is under FDI regulation. Similarly, in panel B, I report the coefficients on the common dialect dummy variable and the interaction between common dialect and the productivity of HMT firms for each industry.

Table 4: The Effects of Common Dialect on the Productivity of Domestic Firms

	(1)	(2)	(3)	(4)
	HMT firms less productive than domestic firms		HMT firms more productive than domestic firms	
	Ln(TFP)	Ln(TFP) Non-parametric	Ln(TFP)	Ln(TFP) Non-parametric
Common dialect	-0.061	-0.027	0.26***	0.12**
	(0.076)	(0.054)	(0.079)	(0.054)
Observations	11580	11580	22009	22009
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, the share of other foreign firms in the same zip code and industry, Age of the firm, capital-labor ratio, log(output), log(export)			

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the productivity of domestic firms by industries. Industries are categorized by whether an average HMT firm from the industry is more productive than an average domestic firm from the industry. All models estimate the discontinuous changes in the TFP of domestic firms at the common dialect borders using spatial regression discontinuity design at the firm level. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. All models use a sample of all domestic firms within 30 km of the common dialect borders. Column (1) and (2) use a sample of domestic firms from industries where the productivity of HMT firms is lower than domestic firms. Column (3) and (4) use a sample of domestic firms from industries where the productivity of HMT firms is higher than domestic firms. Column (1) and (3) use TFP calculated following the framework of Brandt et al. (2012). Column (2) and (4) use an alternative measure of TFP calculated using non-parametric method following Ackerberg et al. (2015) as the dependent variables.

Table 5: Horizontal Spillovers from HMT Firms to Domestic Firms (GMM IV Estimation)

	(1)	(2)	(3)	(4)
	All Firms		HMT firms more productive than domestic firms	
	Ln(TFP)	Ln(TFP) Non-parametric	Ln(TFP)	Ln(TFP) Non-parametric
Panel A: Second Stage:				
Horizontal Spillovers	1.62	1.15	2.78**	1.69**
	(1.59)	(1.20)	(1.28)	(0.86)
Panel B: First Stage:				
Common Dialect	0.051*	0.051*	0.091***	0.091***
	(0.030)	(0.030)	(0.035)	(0.035)
Observations	33589	33589	22009	22009
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, the share of other foreign firms in the same zip code and industry, Firms' age, capital-labor ratio, log(output), log(export).			

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the horizontal spillovers from HMT investment to domestic firms in the same zip code and industry. Column (1) and (2) use a sample of all domestic firms within 30 km of the dialect borders. Column (3) and (4) use a sample of domestic firms from industries where the productivity of HMT firms is higher than domestic firms. All models are estimated using two stage GMM IV method, where the share of HMT firms is instrumented by the common dialect dummy variable in the spatial regression discontinuity design [model (9)]. The first stage is estimated using the local linear version of the spatial regression discontinuity design with bandwidth chosen to be 30 km. The coefficients on the common dialect dummy variable in the first stage are shown in panel 2 of the table. In panel 1, I report the coefficients from the second stage on the share of HMT firms in the same industry and zip code, which is interpreted as horizontal spillovers. Column (1) and (3) use TFP calculated following the framework of Brandt et al. (2012). Column (2) and (4) use an alternative measure of TFP calculated using non-parametric method following Ackerberg et al. (2015) as the dependent variables.

Table 6: Horizontal Spillovers Estimated without Instruments

	(1)	(2)	(3)	(4)
	All Firms		HMT firms more productive than domestic firms	
	Ln(TFP)	Ln(TFP) Non-parametric	Ln(TFP)	Ln(TFP) Non-parametric
Horizontal Spillover	-0.086	-0.065*	-0.12	-0.10*
	(0.058)	(0.039)	(0.074)	(0.052)
Observations	33589	33589	22009	22009
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, Firms' years in business, capital-labor ratio, log(output), whether exporter.			

Notes: Robust standard errors clustered at the zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the horizontal spillovers to domestic firms estimated without using the identification strategy at the dialect borders. Column (1) and (2) use a sample of all domestic firms within 30 km of the dialect borders. Column (3) and (4) use a sample of domestic firms from industries where the productivity of HMT firms is higher than domestic firms. Column (1) and (3) use TFP calculated following the framework of Brandt et al. (2012) as the dependent variable. Column (2) and (4) use TFP calculated using non-parametric method following Ackerberg et al. (2015) as the dependent variable.

Table 7: The Effects of Common Dialect on the Share of Domestic Firms (Crowding-out Effect)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent Variable: Employment share					
Common Dialect	-0.064***	-0.063*	-0.011	-0.082	-0.085***
	(0.023)	(0.034)	(0.050)	(0.067)	(0.026)
Bandwidth					30
Panel B: Dependent Variable: Output share					
Common Dialect	-0.056**	-0.056	0.013	-0.018	-0.082***
	(0.025)	(0.038)	(0.055)	(0.038)	(0.028)
Bandwidth					30
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				
Observations	9456	9456	9456	9456	7067

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of domestic firms to show the possible crowding-out effects of HMT investment. Column (1) to column (4) estimate the model of spatial regression discontinuity design [model (3)] by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Panel A uses the employment share of domestic firms as the dependent variable. Panel B uses the output share as the dependent variable.

Table 8: The Effects of Common Dialect on the Share of Other Foreign Firms (Additional Tests)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent Variable: Employment share:					
Common Dialect	0.001	-0.01	-0.037	0.015	0.0003
	(0.014)	(0.020)	(0.027)	(0.034)	(0.016)
Optimal Bandwidth					30
Panel B: Dependent Variable: Output share:					
Common Dialect	0.0004	0.0006	-0.078**	-0.025	0.010
	(0.016)	(0.022)	(0.031)	(0.040)	(0.017)
Optimal Bandwidth					30
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				
Observations	9456	9456	9456	9456	7067

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of Common Dialect on the share of other foreign firms, which serves as a falsification test. Column (1) to column (4) estimate the model of spatial regression discontinuity design [model (3)] by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with optimal bandwidth chosen to be 30 km. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Panel A uses the employment share of other foreign firms as the dependent variable. Panel B uses the output share as the dependent variable.

Table 9: The Effects of Wu Dialect Border (placebo dialect border) on the Share of HMT Firms (Additional Tests)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent Variable: Employment share					
Common Dialect	-0.0016	-0.0003	-0.0038	-0.0030	-0.0010
	(0.0081)	(0.012)	(0.0015)	(0.019)	(0.0091)
Bandwidth					30
Panel B: Dependent Variable: Output share					
Common Dialect	0.0084	0.0093	0.018	0.0043	0.0083
	(0.0078)	(0.011)	(0.015)	(0.018)	(0.0089)
Bandwidth					30
Control variables	County fixed effects, year fixed effects, distance to Shanghai				
Observations	10127	10127	10127	10127	7902

Notes: Robust standard errors clustered at the zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of Wu dialect border on the share of HMT firms. Because Wu culture is not related to Hong Kong or Taiwan culture, I expect to observe no discontinuous changes in HMT investment at the Wu dialect border. Column (1) to column (4) estimate the model of spatial regression discontinuity design [model (3)] by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Panel A uses the employment share of HMT firms as the dependent variable. Panel B uses the output share as the dependent variable respectively.

Table 10: The Effects of Hypothetical Placebo Dialect Borders on the Employment Share of Firms from HMT (Additional Tests)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Moving the actual dialect border out by 100km					
Common Dialect	0.0065	-0.021	-0.044	-0.032	-0.0043
	(0.025)	(0.030)	(0.037)	(0.047)	(0.028)
Bandwidth					30
Observations	3444	3444	3444	3444	2722
Panel B: Moving the actual dialect border in by 100km					
Common Dialect	-0.0011	-0.034	-0.011	-0.014	0.0061
	(0.039)	(0.053)	(0.064)	(0.083)	(0.045)
Bandwidth					30
Observations	3004	3004	3004	3004	2160
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of hypothetical placebo dialect borders on the employment share of HMT firms. Panel A moves the actual dialect borders outward by 100km. Panel B moves the actual dialect borders inward by 100km. Column (1) to column (4) estimate the model of spatial regression discontinuity design [model (3)] by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Panel A uses employment share of HMT firms as the dependent variable. Panel B uses output share as the dependent variable respectively.

Table 11: The Effects of Hypothetical Placebo Dialect Borders on the Productivity of Domestic Firms (Additional Tests)

	(1)	(2)	(3)	(4)
	All firms		HMT firms more productive than domestic firms	
	Ln(TFP)	Ln(TFP) Non-parametric	Ln(TFP)	Ln(TFP) Non-parametric
Panel A: Moving the actual dialect border out by 100km				
Common dialect	0.092	0.091	0.062	0.059
	(0.13)	(0.10)	(0.11)	(0.083)
Observations	20296	20296	9129	9129
Panel B: Moving the actual dialect border in by 100km				
Common dialect	0.008	-0.029	0.058	-0.027
	(0.071)	(0.054)	(0.11)	(0.074)
Observations	34305	34305	15201	15201
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, the share of other foreign firms in the same zip code and industry, Age of the firm, capital-labor ratio, log(output), log(export)			

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 15%, 10%, 5% and 1% level respectively. This table shows the effects of hypothetical placebo dialect borders on the productivity of domestic firms. All models estimate the discontinuous changes in the TFP of domestic firms at the hypothetical dialect borders using spatial regression discontinuity design at the firm level. The coefficients that capture the discontinuous increase in the outcome variables at the hypothetical dialect borders are shown in the table. Panel A moves the actual dialect borders outward by 100km. Panel B moves the actual dialect borders inward by 100km. Column (1) and (2) use a sample of all domestic firms within 30 km of the hypothetical dialect borders. Column (3) and (4) use a sample of domestic firms from industries where the productivity of HMT firms is higher than domestic firms. Column (1) and (3) use TFP calculated following the framework of Brandt et al. (2012). Column (2) and (4) use an alternative measure of TFP calculated using non-parametric method following Ackerberg et al. (2015) as the dependent variables.

Table 12: The Effects of Common Dialect on the Productivity of Domestic Firms from Industries with Low HMT Investment (Additional Tests)

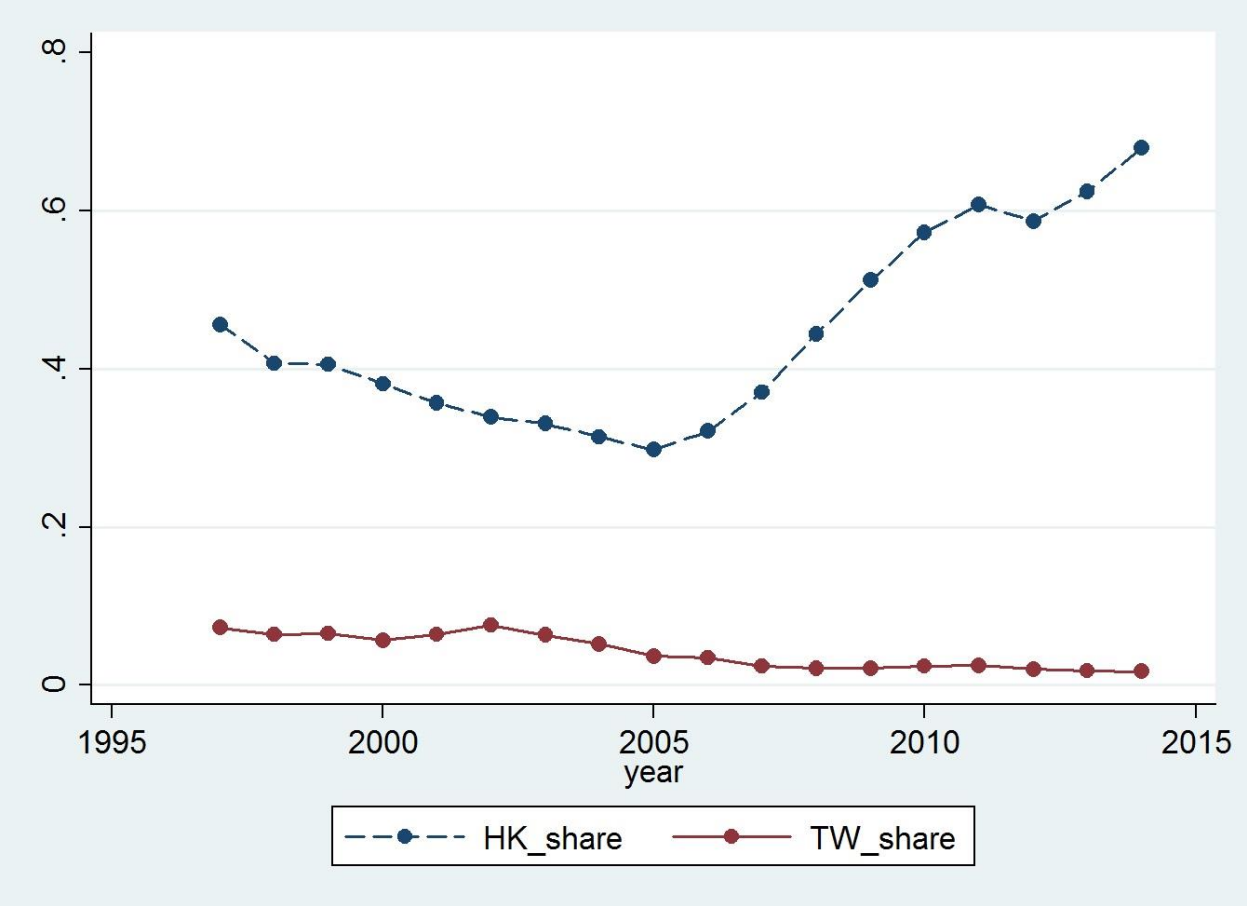
	(1)	(2)	(3)	(4)	(5)
	HMT share==0	HMT share<0.01	HMT share<0.02	HMT share<0.03	HMT share<0.04
Dependent Variable: ln(TFP)					
Common dialect	-0.013	0.023	0.046	0.063	0.058
	(0.054)	(0.053)	(0.052)	(0.052)	(0.051)
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, Firms 'age, ln(output), capital-labor ratio, ln(export).				
Observations	27447	28414	28828	29412	30005

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 15%, 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the productivity of domestic firms from industries with low HMT investment. All models estimate discontinuous changes in the TFP of domestic firms at the common dialect borders using spatial regression discontinuity design at the firm level. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. All models use a sample of all domestic firms within 30 km of the common dialect borders. Column (1) to column (5) show the results on industries with different share of HMT investment (ranging from 0 to 4 percent).

Figure 1 : The Cantonese Dialect Zone and the Min Dialect Zone

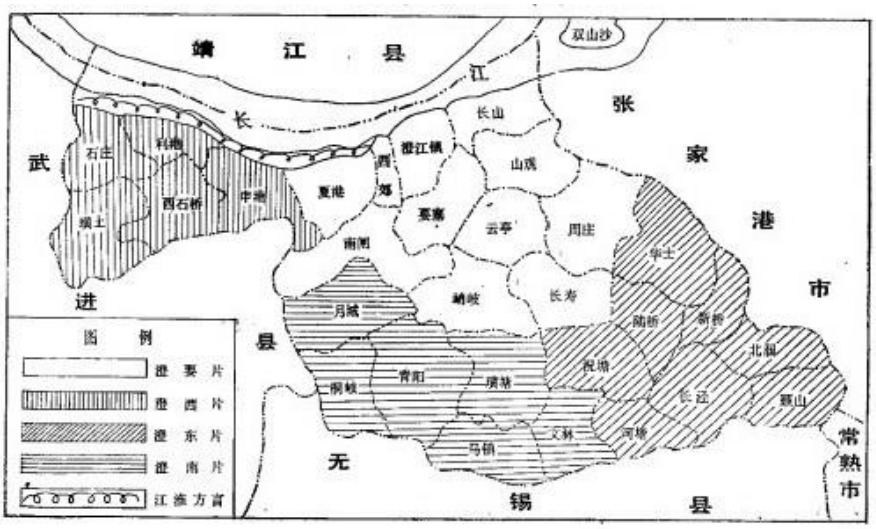
Notes: This figure shows the geographical location of the two dialect zones investigated in this study: the Cantonese (Yue) dialect zone and the Min dialect zone. Data source: Language Atlas of China

Figure 2: The Share of Investment from Hong Kong and Taiwan



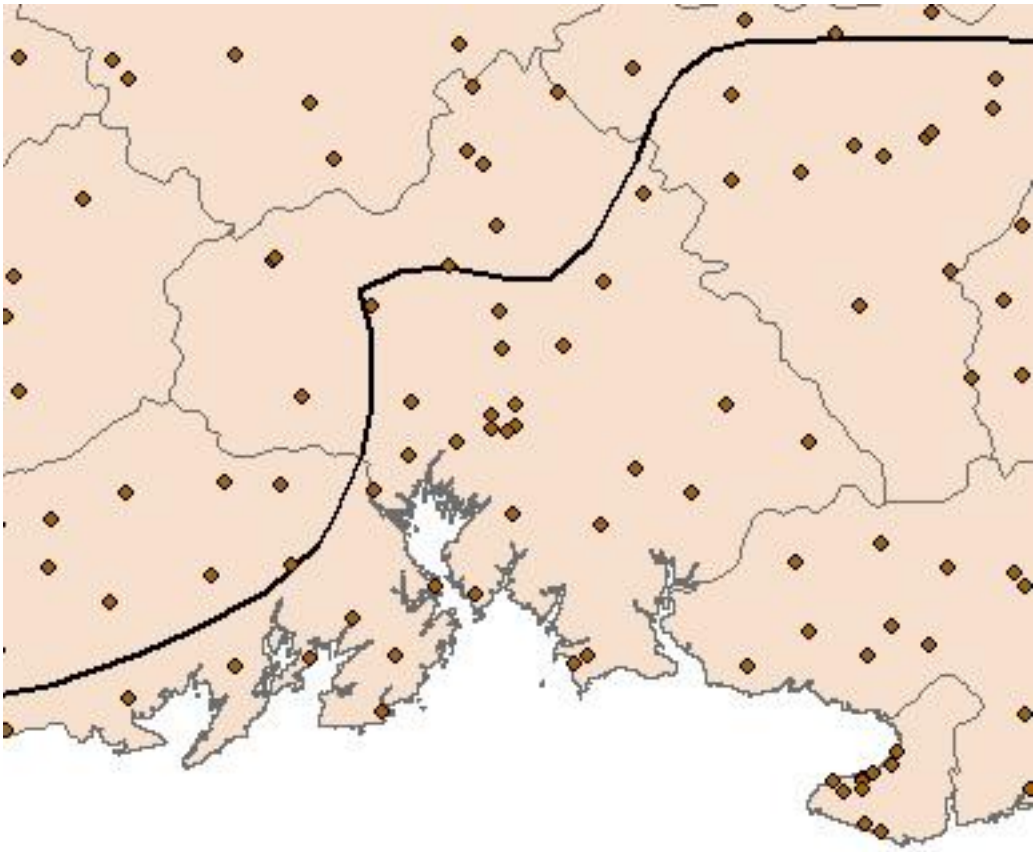
Notes: This figure shows the share of Hong Kong and Taiwan investment to mainland China among all foreign investments. Dashed line indicates the share of Hong Kong investment while solid line indicates the share of Tai wan investment. Data source: China statistical year books.

Figure 3: An Example of a Dialect Border



Notes: This figure shows an example of how dialect border is constructed from knowledge on major dialect used by villages of China. Different notations on the map indicates different types of dialects. For example, the white section denotes Chengyao pian and the vertical dashed line indicates Chengxi pian. (The example county is not in Cantonese or Min dialect zone)

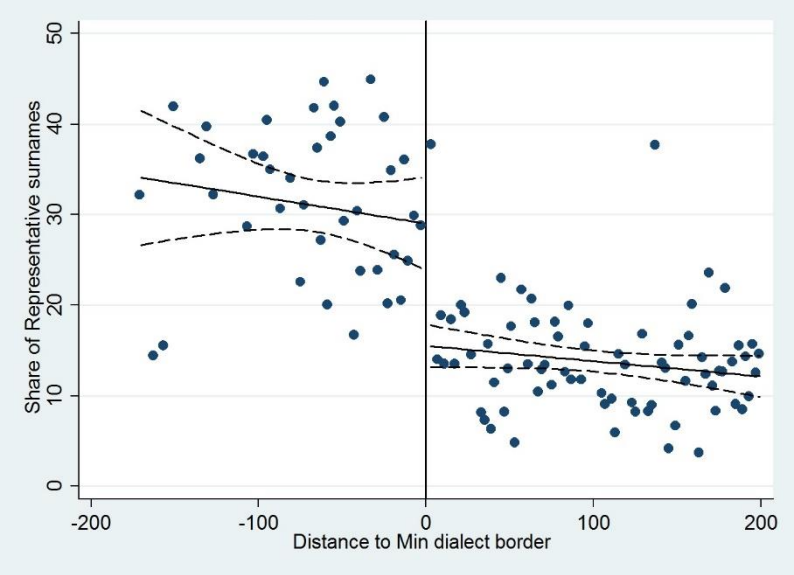
Figure 4: Dialect Borders and Administrative County Borders



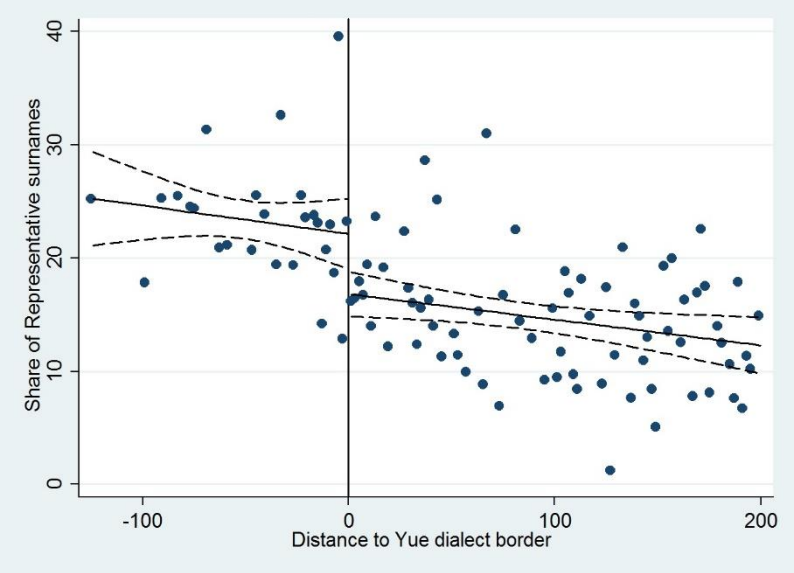
Notes: This figure shows an example of the relationship between dialect borders and county level administrative borders. The solid black line denotes the Cantonese dialect border. The shallow grey lines denote administrative county borders. The points refer to the centroids of zip codes. We can see that an administrative county could be separated into two different dialect zones.

Figure 5. The Geographical Distribution of the Share of Representative Surnames

5.1 Min dialect border

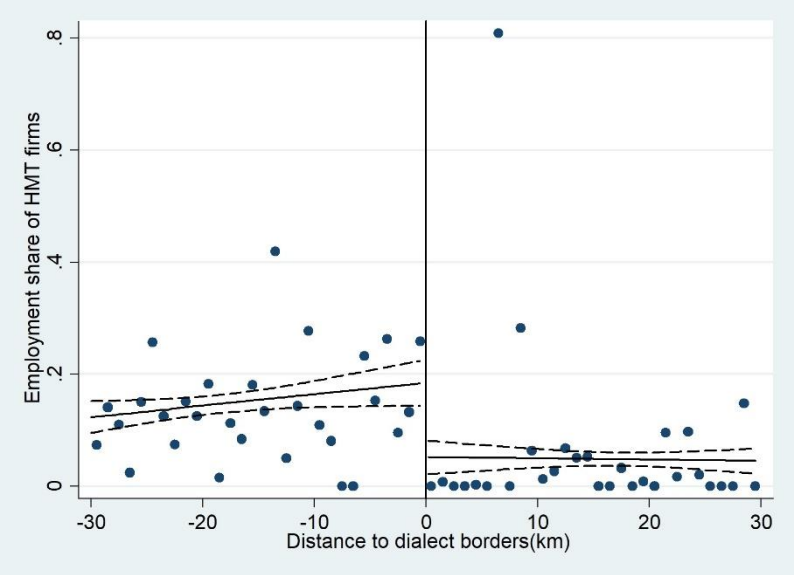


5.2 Cantonese (Yue) dialect border



Notes: This figure shows the discontinuous changes in the population share of representative surnames at the dialect borders. Figure 5.1 shows the population share of representative surnames (3 most common surnames) for the Min cultural group (Chen [陈], Lin [林] and Huang [黄]) at the border of Min dialect zone. Figure 5.2 shows the population share of representative surnames for the Cantonese cultural group (Chen [陈], Liang [梁] and Li [李]) at the border of Cantonese dialect zone. Horizontal axis shows the distance to dialect borders with negative value indicating inside the borders. Vertical axis shows average population share of representative surnames for each given distance. Data source: Calculated from the 2005 Chinese population census.

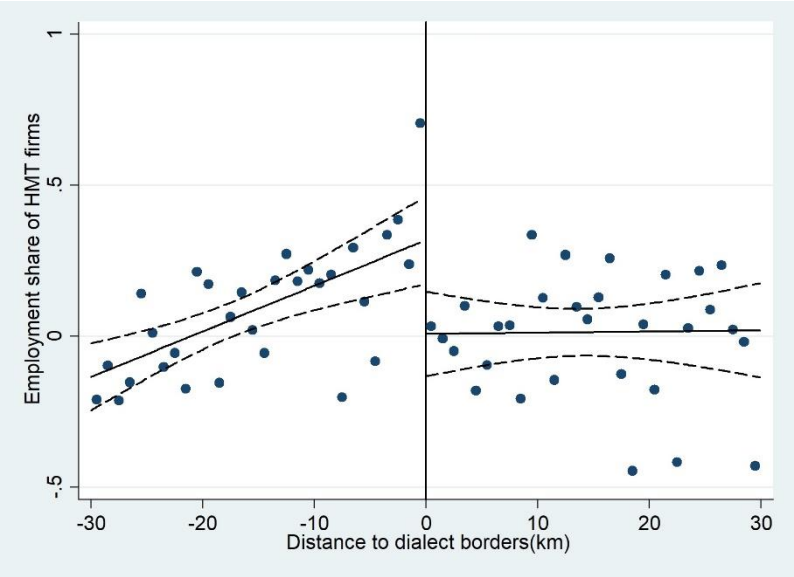
Figure 6: Discontinuity in the Share of HMT Firms at the Borders of the Dialect Zones



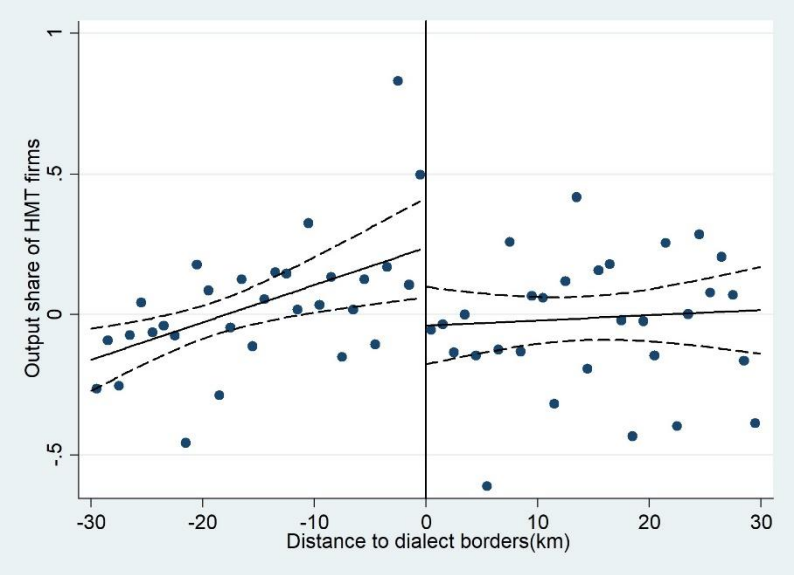
Notes: This figure shows the employment share of HMT firms among all firms by distance to the dialect borders. All zip codes within 30 kilometers are included in the analysis. The horizontal axis denotes distance to the dialect borders with negative value indicating the zip code is located inside the border (the same dialect as HMT). The vertical axis denotes the average share of HMT firms (total employment of HMT firms over total employment of all firms) for a given distance.

Figure 7: Discontinuity in the Share of HMT Firms at the Borders of the Dialect Zones (Standardization at the county level)

7.1. Employment Share



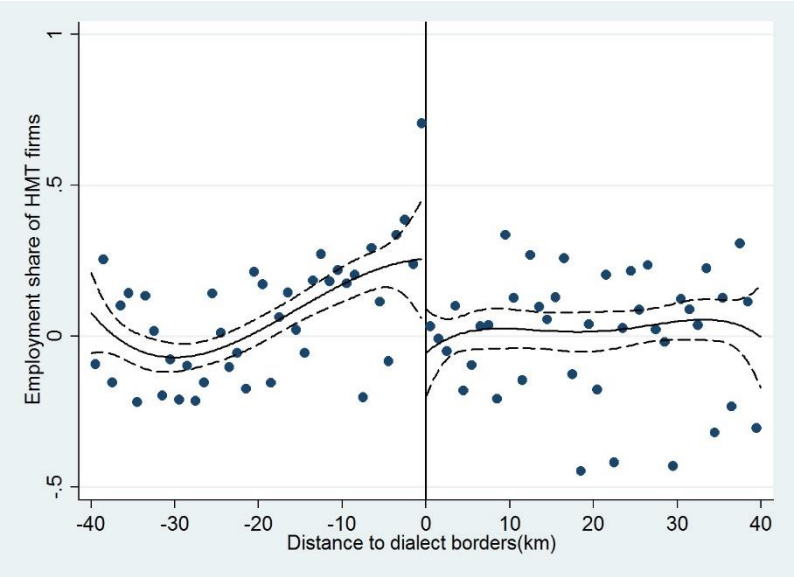
7.2. Output Share



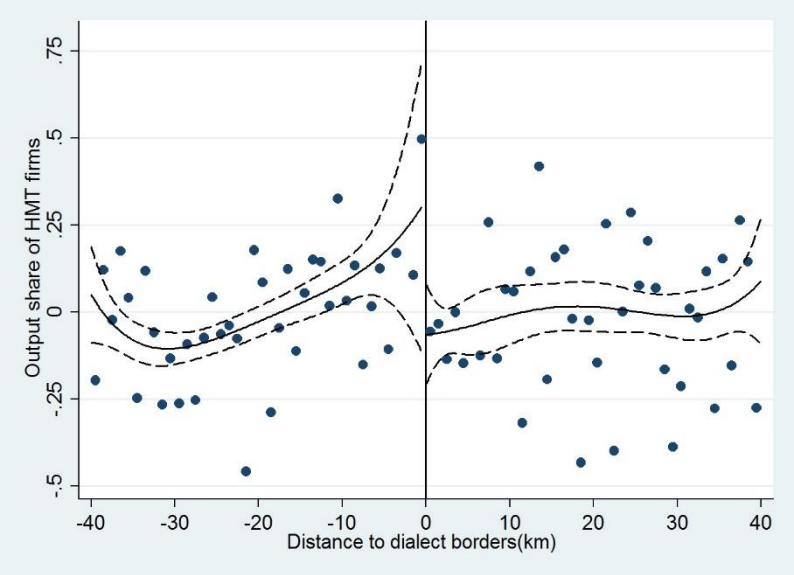
Notes: These figures show the share of HMT firms among all firms by distance to the dialect borders. All zip codes within 30 kilometers are included in the analysis. The horizontal axis denotes distance to the dialect borders with negative value indicating that the zip code is located inside the border (same dialect as HMT).The vertical axis denotes the average share of HMT firms for a given distance. Figure 7.1 shows the distribution of employment share; Figure 7.2 shows the distribution of output share. The share of HMT firms is standardized at the county level by subtracting county mean and dividing by county standard deviation.

Figure 8: Discontinuity in the Share of HMT firms at the Borders of the Dialect Zones (Non-linear)

8.1. Employment Share



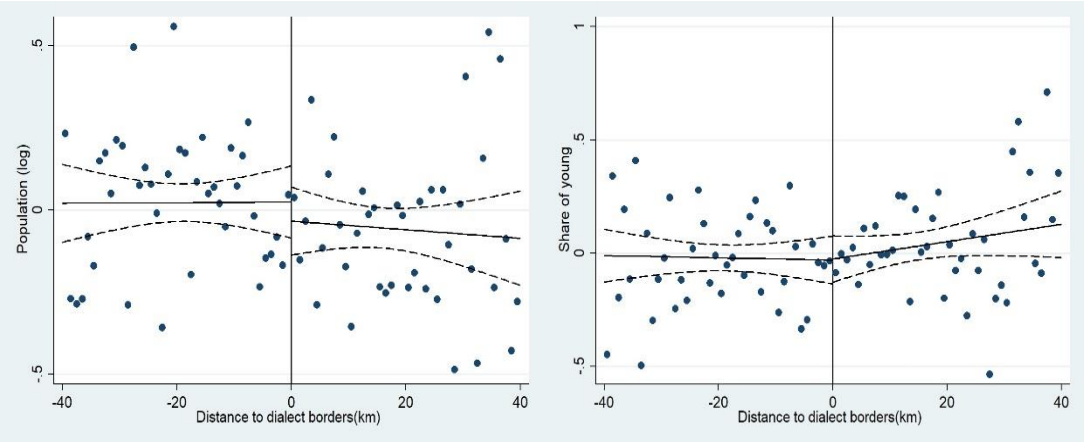
8.2 Output Share



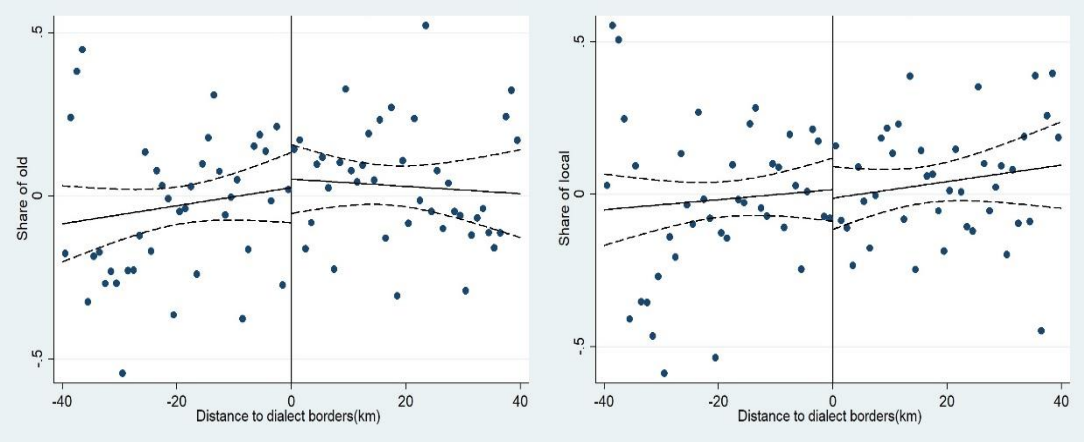
Notes: These figures show the share of HMT firms among all firms by distance to the dialect borders. All zip codes within 40 kilometers are included in the analysis. The horizontal axis denotes the distance to the dialect borders with negative value indicating that the zip code is located inside the border (same dialect as HMT). The vertical axis denotes the average share of HMT firms for a given distance. The model is fitted using the fourth-degree polynomials of distance to the borders. Figure 8.1 shows the distribution of employment share; Figure 8.2 shows the distribution output share. The share of HMT firms is standardized at the county level by subtracting county mean and dividing by county standard deviation.

Figure 9: The Geographical Distribution of Demographic and Geographic Variables across the Dialect Borders (Additional Tests)

9.1 Demographic Variables

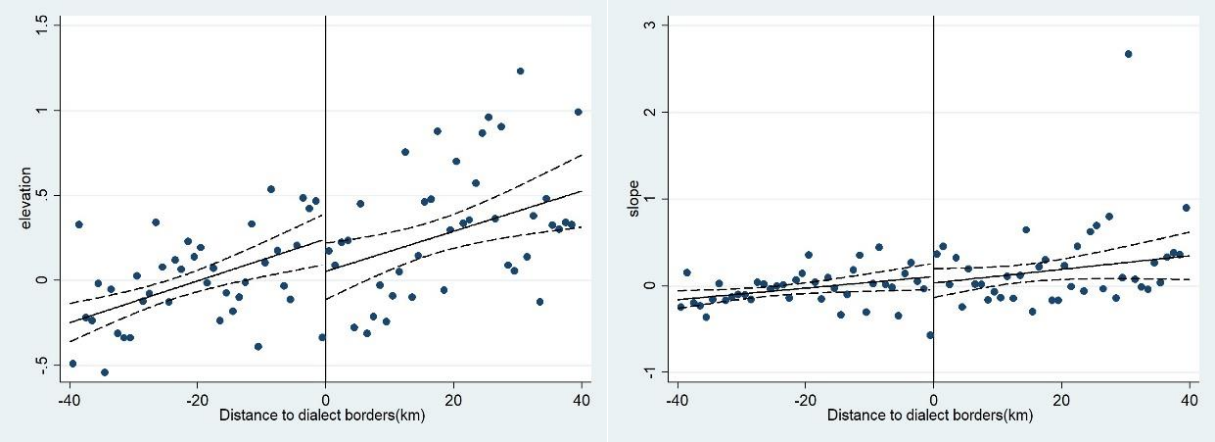


(a) Population **(b) Share of young people**



(c) Share of old people **(d) Share of local people**

9.2 Geographical Variables

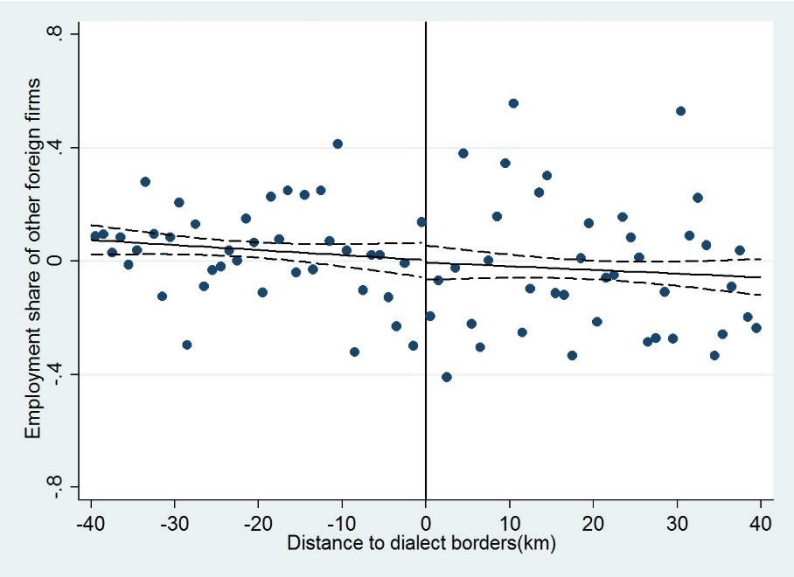


(e) Elevation **(f) Slope**

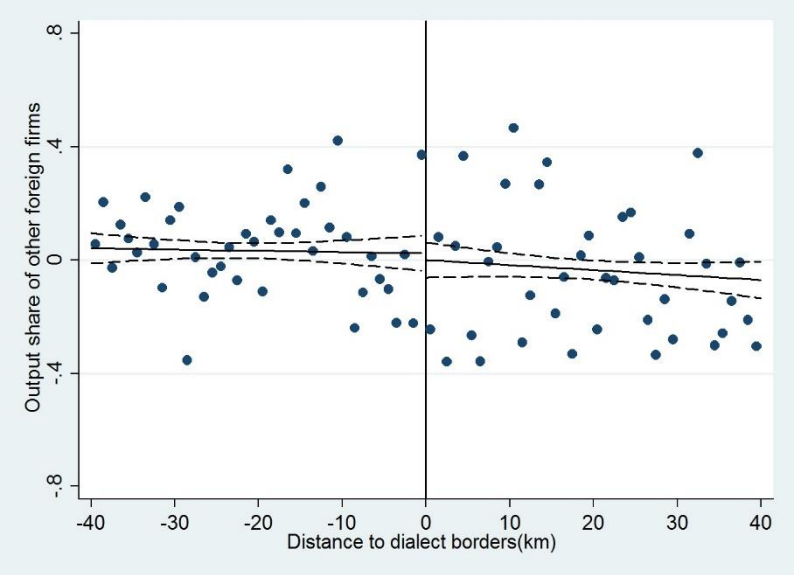
Notes: These figures show the geographical distribution of demographic and geographic variables by distance to the dialect borders. All zip codes within 40 kilometers are included in the analysis. The horizontal axis denotes distance to the dialect borders with negative value indicating the zip code is located inside the border (the same dialect as HMT). The vertical axis denotes outcome variables: (a) logarithm of total population; (b) the share of people who are under 14 years old; (c) the share of people who are above 65 years old; (d) the share of people who have local *hukou* (is a local resident) (People who do not have local *hukou* but are included in the census are in-migrants); (e) elevation of the zip code; (f) slope of the zip code. All outcome variables are standardized at the county level by subtracting county mean and dividing by county standard deviation. Demographic data are from aggregated data of Chinese population census 2010. Geographic variables are calculated by the author.

Figure 10: The Geographical Distribution of the Share of Foreign Firms from Other Countries at the Borders of the Dialect zones (Additional Tests)

11.1 Employment Share



11.2 Output Share



Notes: These figures show the share of foreign firms from regions other than HMT among all firms by distance to the dialect borders. These figures serve as a falsification test. The horizontal axis denotes the distance to the dialect borders with negative value indicating the zip code is located inside the borders (same dialect as HMT). The vertical axis denotes the average share of other foreign firms for a given distance. Figure 11.1 shows the distribution of employment share; Figure 11.2 shows the distribution of output share. The share of other foreign firms is standardized at the county level by subtracting county mean and dividing by county standard deviation.

Figure 11: Illustration of the Placebo Dialect Border (Wu dialect border)

Notes: This figure shows the geographical location of the placebo dialect border (Wu dialect border) relative to the Cantonese and Min dialect border. The common border between Wu and Min dialect zones is excluded from the placebo analysis.

Online Appendix Materials

1. TFP Estimation using Non-Parametric Method

Following Ackerberg et al. (2015), I use the following procedures to non-parametrically estimate labor and capital share (β_L and β_k) in the production function. Then, I replace $\widetilde{S_{ft}}$ with β_L and $1-\widetilde{S_{ft}}$ with β_k in equation (7) to get TFP.

Suppose empirical production function is specified as $y = \beta_0 + \beta_l \times l + \beta_k \times k + \omega + \varepsilon$, where y denotes the logarithm of value added, l denotes the logarithm of labor input and k denotes the logarithm of capital input. ω denotes TFP which is unobservable to the researcher. ε denotes random productivity shock.

I use the following procedures to get $\widehat{\beta_l}$ and $\widehat{\beta_k}$:

(1) Non-parametrically regress y on l, k and intermediary input and get predicted $\hat{y} = \hat{\varphi}(l, k, input)$;

(2) Then TFP ω can be estimated as $\hat{\omega} = \hat{\varphi}(l, k, input) - \beta_0 - \beta_l \times l - \beta_k \times k$.

(3) Assuming for each firm, $\omega_t = \rho\omega_{t-1} + u_t$, where u_t denotes exogenous productivity shock, given parameters ρ, β_0, β_l and β_k, $\hat{\omega}_t$, $\hat{\omega}_{t-1}$ and $\widehat{u_t}$ can be calculated from the data.

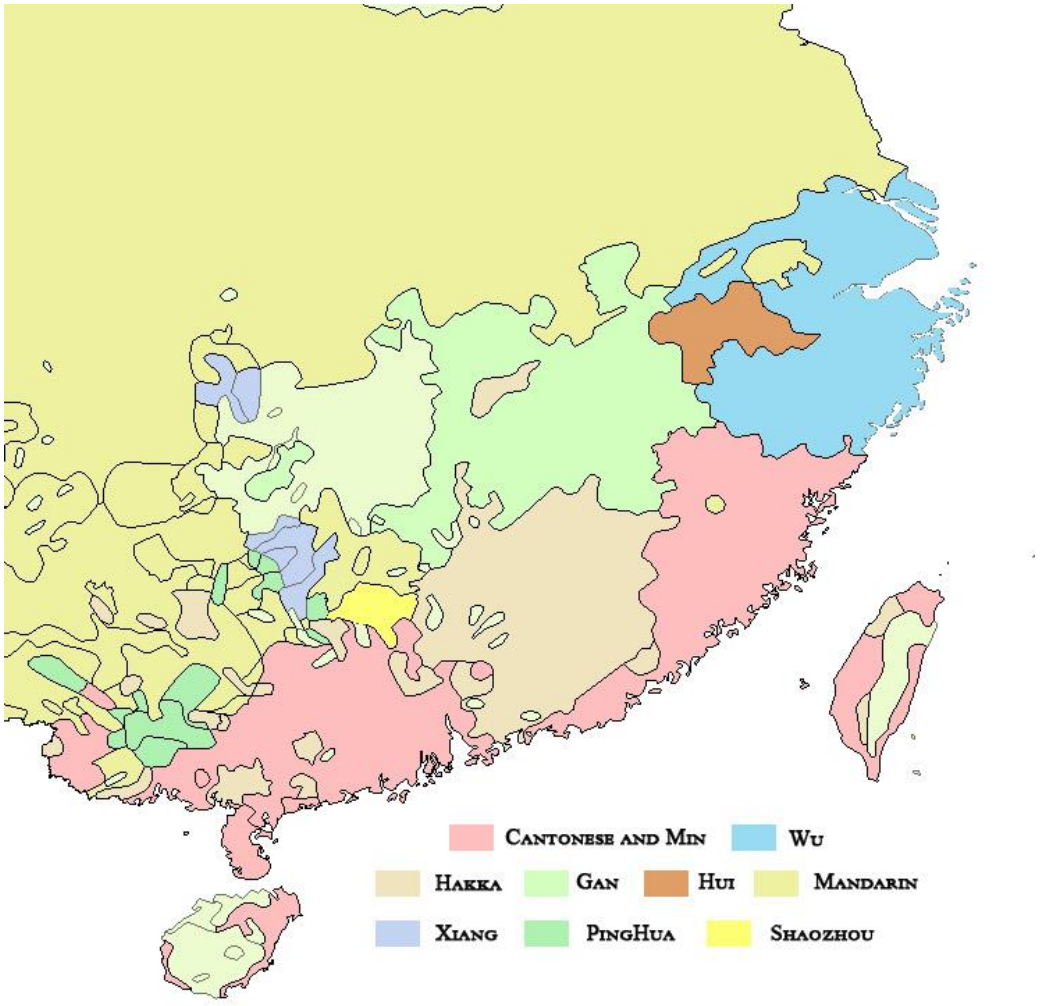
(4) Using the moment condition that $\widehat{u_t}$ is orthogonal to 1,l_{t-1}, k_t and $\hat{\varphi}(l, k, input)_{t-1}$ and GMM method to estimate $\widehat{\beta_l}$ and $\widehat{\beta_k}$.

2. Other Maps of Dialect Zones

Figure A1: Map of Dialect zones with Major Rivers:

Notes: This figure shows the location of the two dialect zones investigated by this study and major rivers in this area.

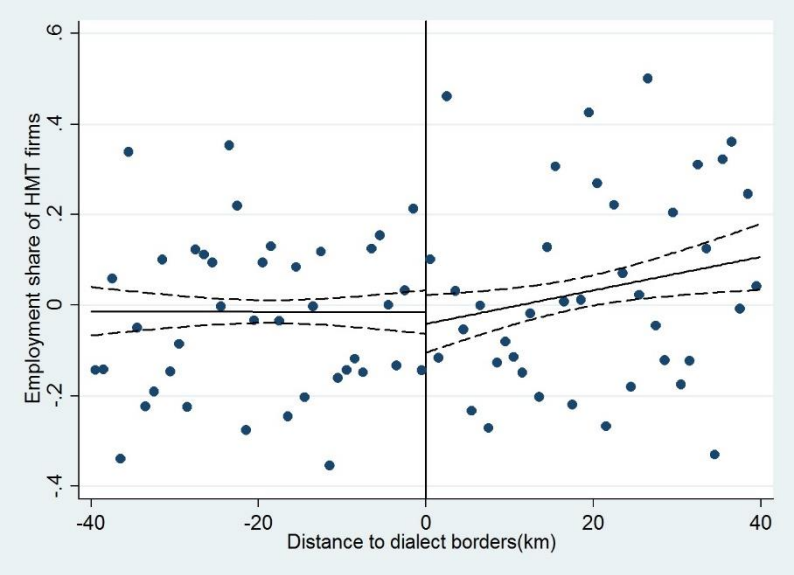
Figure A2: Map of Other Dialect Zones around this Area



Notes: This figure shows the location of other Chinese dialect zones around this area.

3. The Employment Share of HMT Firms at the Border of the Wu Dialect Zone (Placebo Test)

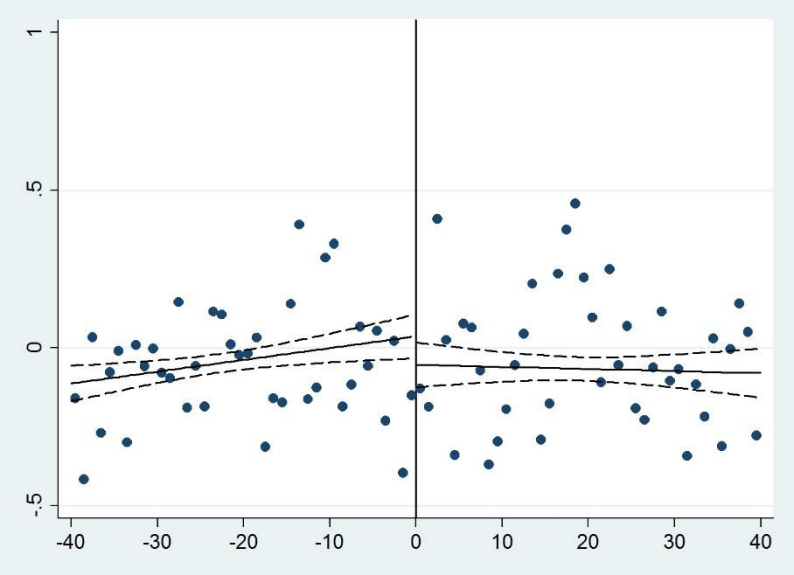
Figure A3: The Employment Share of HMT Firms at the Border of the Wu Dialect Zone



Notes: This figure shows the share of HMT firms among all firms at each zip code by distance to the Wu dialect border. All zip codes within 40 kilometers are included in the analysis. The horizontal axis denotes distance to the Wu dialect border with negative value indicating the zip code is located inside the border. Vertical axis denotes average share of HMT firms for a given distance. The share of HMT firms is standardized by subtracting county mean and dividing by county standard deviation.

4. The Share of Regulated Industries at the Dialect Borders

Figure A4: The Share of Regulated Industries at the Dialect Borders



Notes: When studying the effects of common dialect by industrial entry regulation, one concern is that the intensity of regulation may change discontinuously at the borders. To address this issue, I plot the geographical distribution of the employment share of regulated industries among all industries. Figure A4 shows the employment share of regulated industries among all industries at each zip code by distance to common dialect borders. All zip codes within 40 kilometers are included in the analysis. The horizontal axis denotes distance to the common dialect borders with negative value indicating the zip code is located inside the borders. The vertical axis denotes the average share of regulated industries for a given distance. The Share of regulated industries is standardized by subtracting county mean and dividing by county standard deviation. From this figure, I find no evidence to suggest that regions inside the borders receive more regulation than regions outside the borders.

5. The Effects of Common Dialect on the Level of Employment and Output by Firm Type

Table A1: The Effects of Common Dialect on the Level of Employment and Output by Firm Type

	(1)	(2)	(3)
	HMT	Domestic	Other foreign
Panel A: Dependent Variable: ln (Total Employment)			
Common Dialect	0.44	-0.46*	-0.18
	(0.30)	(0.24)	(0.27)
Panel B: Dependent Variable: ln (Total Output)			
Common Dialect	0.57**	-0.90***	-0.39
	(0.24)	(0.34)	(0.49)
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei		
Observations	7067	7067	7067

Notes: Robust standard errors clustered at the zip code level are shown in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on total employment and output by types of firms. Column (1) to (3) are estimated using the local linear model [model (4)] and bandwidth is chosen to be 30km. Column (1) shows the effects on HMT firms. Column (2) shows the effects on Domestic firms. Column (3) shows the effects on other foreign firms. Panel A uses the logarithm of total employment of a zip code as the dependent variable. Panel B uses the logarithm of total output of a zip code as the dependent variable

6. Identify both Horizontal and Vertical Spillovers

In addition to the horizontal spillovers estimated in section 4.3, industrial variation in regulation interacted with dialect borders allows me to also estimate vertical spillovers (spillovers to domestic suppliers and buyers).

The baseline model that include both horizontal and vertical spillovers can be specified as:

$$\ln(TFP_{fikt}) = \mu_0 + \mu_1 HMT\_share\_Horizontal_{ikt} + \mu_2 HMT\_share\_Backward_{ikt} + \mu_3 HMT\_share\_Forward_{ikt} + Z_{fikt} + c_j + \eta_t + \delta_k + \epsilon_{it}, \quad \text{（A6.1）}$$

where $HMT\_share\_Horizontal_{ikt}$ captures the presence of HMT firms in the same industry and location as firm f; $HMT\_share\_Backward_{ikt}$ captures the presence of HMT firms in industries that are supplied by industry *k* (downstream industries of industry *k*) in zip code *i* and year *t*; $HMT\_share\_Forward_{ikt}$ captures the presence of HMT firms in industries that supply to industry *k* (upstream industries of industry *k*) in zip code *i* and year *t*. As a result, μ_1 captures horizontal spillovers; μ_2 captures spillovers from HMT firms to domestic suppliers in the same zip code (backward linkage); while μ_3 captures spillovers from HMT firms to domestic buyers in the same zip code (forward linkage).

$HMT\_share\_Backward_{ikt}$ can be defined as a weight average of HMT output share for all downstream industries of industry k:

$$HMT\_share\_Backward_{ikt} = \sum_{u\ if\ u\neq k} \alpha_{uk} \frac{\sum_{p\in\Omega_{iut}} HMT_p \times Output_p}{\sum_{p\in\Omega_{iut}} Output_p}, \quad \text{(A6.2)}$$

where u denotes all downstream industries of industry *k* and weight α_{uk} is the proportion of industry *k*'s output supplied to industry *u* taken from the 2002 input-output table at the two-digit industry level. The HMT output share is calculated in the same way as equation (9), which is HMT equity weighted total output over total output.

Similarly, $HMT\_share\_Forward_{ikt}$ is defined as a weighted average of HMT output share (excluding export) for all upstream industries of industry k:

$$HMT\_share\_Forward_{ikt} = \sum_{v\ if\ v\neq k} \theta_{vk} \frac{\sum_{q\in\Omega_{ivt}} HMT_q \times (Output_q - Export_q)}{\sum_{q\in\Omega_{ivt}} (Output_q - Export_q)}, \quad \text{(A6. 3)}$$

where v denotes all upstream industries of industry *k* and weight θ_{vk} is the proportion of industry *k*'s input supplied by industry v taken from the 2002 input-output table at the two-digit industry level. The HMT output share is calculated in a similar way to equation (A6.2) except that export need to be excluded when calculating linkage to domestic upstream industries.

To solve the endogeneity problem in equation (A6.1), I use dialect borders and dialect borders interacted with regulation policies as the instruments for $HMT\_share\_Horizontal_{ikt}$, $HMT\_share\_Backward_{ikt}$ and $HMT\_share\_Forward_{ikt}$. Specifically I use the following four instruments: T_i, $R_{kt} \times T_i$, $\sum_{u\ if\ u\neq k} \alpha_{uk} R_{ut} \times T_i$ and $\sum_{v\ if\ u\neq k} \theta_{vk} R_{vt} \times T_i$, where T_i indicates whether location *i* is in the common dialect area and R_{kt} indicates whether industry *k* is regulated in year *t*. Section 4.2 shows that entry-regulation does generate heterogeneous effects across industries. Therefore $R_{kt} \times T_i$ provides additional information to identify the coefficients.

$\sum_{u\ if\ u\neq k} \alpha_{uk} R_{ut} \times T_i$ measures the regulation status of all downstream industries of industry k interacted with common dialect, which helps to identify the backward linkage parameter μ_2; $\sum_{v\ if\ u\neq k} \theta_{vk} R_{vt} \times T_i$ measures the regulation status of all upstream industries of industry k interacted with common dialect, which helps to identify the forward linkage parameter μ_3 . In the full model, I also include additional control variables as follows: R_{kt} measures the regulation status of industry k;$\sum_{u\ if\ u\neq k} \alpha_{uk} R_{ut}$ measures the regulation status of industry k's downstream industries and $\sum_{u\ if\ u\neq k} \theta_{uk} R_{ut}$ measure the regulation status of industry k's upstream industries.

Results estimated using equation (A6.1) with instruments are shown as Column (1) and (2) of Table A2. I report the estimated results of μ_1 (horizontal spillovers), μ_2(backward linkages) and μ_3 (forward linkages) from equation (A6.1). The coefficients representing horizontal spillovers are estimated to be positive and the coefficients representing both backward and forward linkages are estimated to be negative. Yet all coefficients are imprecisely estimated and statistically insignificant. This is mainly because we have a very weak first stage when predicting the presence of HMT firms in downstream and upstream industries.

Table A3 shows the results from the first stage. I find that even though the interaction between dialect and regulation can strongly predict the presence of HMT firms in the same industry, the interactions between common dialect and regulation in downstream and upstream industries are not strong instruments for the presence of HMT firms in downstream and upstream industries. Also, the coefficients on the interaction between common dialect and upstream regulation has a positive effect on the share of the presence of HMT firm in upstream industries, which is not consistent with expectation. Therefore, using the empirical framework and the sample of this research, I do not have the power to clearly identify vertical spillovers using the interaction between common dialect and regulation as instruments.

Table A2: Horizontal and Vertical Spillovers from HMT Firms to Domestic Firms (GMM IV estimation)

	(1)	(4)
	Ln(TFP)	Ln(TFP) Non-parametric
Horizontal Spillovers	4.00	2.74
	(5.35)	(3.93)
Backward Spillovers	-10.35	-12.31
	（29.98）	(22.42)
Forward Spillovers	-17.09	-7.48
	(21.18)	(14.69)
First Stage (Common Dialect)	Table A3	Table A3
Observations	33589	33589
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, Firms' age, capital-labor ratio, log(output), log(export). Additional control variables for column (3) and (4): Regulation status of industry *k*, downstream industries of industry *k* and upstream industries of industry *k*.	

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively.This table shows horizontal and vertical spillovers from HMT investment to domestic firms in the same zip code and industry. Column (1) and (2) jointly estimate horizontal and vertical spillovers using equation (A6.1) with instruments. Local linear model is estimated using 30km as the optimal bandwidth. Column (1) uses TFP calculated following the framework of Brandt et al. (2012) and Column (2) uses TFP calculated using non-parametric method following Ackerberg et al. (2015) as the dependent variable.

Table A3: First-Stage Results of Table A2

	(1)	(2)	(3)
	Horizontal HMT share	Downstream HMT share	Upstream HMT share
Common Dialect	0.0005	0.0045	0.0054
	(0.04)	(0.0037)	(0.014)
Common Dialect×	-0.098**	-0.0075	-0.022
Regulation	（0.039）	(0.009)	(0.014)
Common Dialect×	0.027	-0.015	0.013
Downstream Regulation	(0.080)	(0.022)	(0.024)
Common Dialect×	0.025***	0.029	0.030
Upstream Regulation	(0.054)	(0.022)	(0.022)
Observations	33589	33589	33589

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This stable shows first-stage results of Table A2. Only coefficients on the instrumental variables in the first stage are shown in the table.

7. Results of Table 11 with Low HMT Share in Own, Down-stream and Up-stream industries:

This section shows an extension of the placebo test conducted in Table 11. In Table 11, I conduct a placebo test using industries with low HMT investment share, assuming that domestic firms from these industries are not affected by investment from HMT. However, domestic firms can still be affected by the presence of HMT firms in downstream and upstream industries, even though I do not find statistically significant vertical spillovers. Therefore, I conduct an additional placebo test using industries with low HMT presence in own, downstream and upstream industries.

Table A4 shows the estimation results using the same empirical specification as Table `. The model is estimated using firms from industries which are chosen such that the share of HMT firms in these industries, the share of HMT firms in the downstream industries of these industries and the share of HMT firms in the upstream industries of these industries are smaller than 5 percent. 5 percent is chosen to maintain a reasonable sample size. Then, TFP is regressed on the dummy variable indicating common dialect. As a result, I find no discontinuous changes in the productivity of domestic firms across the borders in these industries, indicating that the productivity of domestic firms is the same if they do not receive influence from HMT firms.

Table A4: The Effects of Common Dialect on the Productivity of Domestic Firms

	Dependent Variable: ln(TFP)
	HMT share<0.05 Backward Share <0.05 Forward Share<0.05
Common dialect	-0.076
	(0.062)
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei, Firms' age, log(output), capital-labor ratio, log(export).
Observations	16919

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the productivity of domestic firms from industries with low HMT presence in own, downstream and upstream industries. The dependent variable is the TFP of domestic firms. The table shows the coefficient on the dummy variable indicating common dialect zone.

8. The Effects of Common Dialect on Total employment and Output

One of the major identification assumptions of this study is that factors other than dialect should change continuously at the dialect borders. One potential concern is that the degree of economic development might be very different across the border. Therefore, I conduct an additional placebo test using measures of the total size of the industrial sector (for example total industrial output and total employment) as the dependent variable and expect to show that there are no discontinuous changes in the total scale of industrial sector across borders. However, this is not a clean placebo test, because investment from HMT can affect economic growth and thus generate differentiation in economic development in the long-run.

Table A5 reports the results of zip code level analysis (Model (3) and (4)) with total industrial employment and output of zip code *i* as the dependent variable. Table A5 shows that zip codes inside the common dialect zones tend to have higher total industrial employment and output, but the differences are statistically insignificant. Moreover, when the bandwidth becomes smaller, the difference in total size of industrial sector also shrinks to close to zero. Therefore, I do not find discontinuous changes in the total size of industrial sector at the dialect borders.

Table A5: Common Dialect on Total Industrial Production

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent Variable: ln(Employment)					
Common Dialect	0.06	0.22	0.27	0.44	0.024
	(0.17)	(0.25)	(0.35)	(0.48)	(0.20)
Optimal Bandwidth					30
Panel B: Dependent Variable: ln(Output)					
Common Dialect	-0.079	0.25	0.41	0.71	-0.008
	(0.19)	(0.29)	(0.41)	(0.57)	(0.22)
Optimal Bandwidth					30
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				
Observations	9456	9456	9456	9456	7067

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on total output and employment. Column (1) to (4) are estimated using equation (3), Column (5) is estimated using equation (4) with bandwidth shown in the table. From column (1) to (4), first to fourth degree polynomials of distance to the borders are used as control variables. Estimated coefficients β_1 , which are the measure of discontinuous changes at the borders, are reported in the table. Panel A uses logarithm of total employment as the dependent variable. Panel B uses logarithm of total output as the dependent variable respectively.

9. Results of Table 1 with only the Min Dialect Border:

Table A6: Common Dialect on the Share of HMT Firms

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Panel A: Dependent Variable: Employment share					
Common Dialect	0.065	0.073	0.080	0.12	0.078*
	(0.043)	(0.056)	(0.072)	(0.097)	(0.046)
Optimal Bandwidth					30
Panel B: Dependent Variable: Output share					
Common Dialect	0.057	0.084	0.13*	0.13	0.075*
	(0.042)	(0.056)	(0.072)	(0.093)	(0.045)
Optimal Bandwidth					30
Control variables	County fixed effects, year fixed effects, distance to Hong Kong or Taipei				
Observations	2212	2212	2212	2212	1632

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of HMT firms using only variation at the Min dialect border. Models are estimated following the spatial regression discontinuity design using a sample of zip codes within 40 kilometers of the Min dialect border. The dependent variable is the total employment (output) of HMT firms over total employment (output) of all firms for each zip code. The coefficients that capture the discontinuous increase in the outcome variables at the common dialect borders are shown in the table. Column (1) to column (4) estimate model (3) by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with optimal bandwidth chosen using cross validation. Panel A uses employment share as the dependent variable. Panel B uses output share as the dependent variable respectively.

10. Results from a Difference-in-Differences Identification Strategy:

This section shows estimation results of spillover effects using difference-in-differences (DID) identification strategy. The regression discontinuity design (RD) used in the main text reduces endogeneity concern through comparing locations that are geographically very close. Therefore, the cost of applying the RD design is a small sample size that affects the precision of estimation. As mentions in the main text, the vertical spillover effects are not precisely estimated due to the small sample size around the dialect borders. Thus, this section re-estimate spillover effects using DID identification strategy, which requires different identification assumptions and can include more observations to increase the power of estimation.

The DID identification strategy explores variation in dialect and industrial level entry-regulation policies. The empirical model of the first stage can be specified as follows:

$$HMT_{fikt} = \beta_0 + \beta_1 T_i + \beta_2 R_{kt} \times T_i + \beta_3 R_{kt} + Z_{fikt} + \eta_t + \delta_k + \epsilon_{it} \quad (A10.1) \ ,$$

where T_i denotes whether location *i* speaks the same dialect as Hong Kong, Macau and Taiwan (HMT), R_{kt} denotes whether industry *k* receives entry-regulation policy in year *t*. Z_{fikt} denotes firm level control variables. η_t and δ_k are year and industry fixed effects. Model (A10.1) corresponds to Model (5) in the RD design. There are several major differences. First, the analysis is not restricted to a bandwidth of 40km. All zip codes are included in the analysis to increase power of estimation at a cost of reducing cleanness of identification, because the treatment group and control group become less comparable. Second, distance to the dialect borders are no longer controlled in the model. Finally, county fixed effects are removed because we no longer explore variation within the same county when using the DID strategy.

The estimation results of Model (A10.1) are shown in Table A7. I find that the coefficient on common dialect is positive and statistically significant and the coefficient on the interaction term between common dialect and entry-regulation is negative and statistically significant. These findings indicate that speaking the same dialect increases investment from HMT and the effect is larger in unregulated industries. The findings are qualitatively consistent with the conclusions from the RD design. However, in terms of magnitude, the coefficient on the common dialect dummy variable in the DID design (16 percent) is much larger the RD design (5 to 7 percent). The difference in magnitude indicates that other unobservable factors start to affect investment when we move away from the dialect borders. The DID strategy incorporates locations that are not closely comparable with each other in unobservable characteristics. Therefore, the DID strategy suffers more from the endogeneity problems even though it can increase the power of estimation.

Then, I use the results from Model (A10.1) to estimate horizontal and vertical spillovers following equation (A6.1). Horizontal, backward and forward spillovers are instrumented by four instruments: the common dialect dummy variable, common dialect interacted with entry-regulation of industry *k*, common dialect interacted with entry-regulation of downstream industries of industry *k* and common dialect interacted with entry-regulation of upstream industries of industry *k*. The estimation results are shown in Table A8. Most coefficients are still not statistically precisely estimated, indicating that common dialect interacted with FDI entry-

regulation still do not generate enough variation to clearly identify all vertical spillovers even though the sample size is significantly enlarged when all zip codes are included into the analysis.

Table A7: Common Dialect and FDI Entry Regulation on HMT equity share (DID specification)

	Dependent Variable: HMT equity share
Common Dialect (β_1)	0.16***
	(0.0077)
Common Dialect*FDI Regulation (β_2)	-0.015***
	(0.0051)
Observations	686584
Control variables	Year fixed effects, industry fixed effects, Firms 'age, log(output), capital-labor ratio, log(export).

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect and entry-regulation policy on investment from HMT using DID specification. The model I follow is Model A10.1. All zip codes from the eight southern Chinese provinces are included in the sample (Zhejiang, Fujian, Guangdong, Guangxi, Hunan, Jiangxi, Jiangsu, Anhui).

Table A8: Spillovers on the Productivity of Domestic Firms (DID specification)

	(1) Ln(TFP)	(2) Ln(TFP) Non-parametric	(3) Ln(TFP)
Horizontal Spillover	1.28*** (0.22)	1.00*** (0.15)	9.74 (7.70)
Backward Spillover			-54.06 (36.11)
Forward Spillover			29.03* (16.40)
Observations	509704	509704	509704
Control variables	Year fixed effects, industry fixed effects, Firms 'age, log(output), capital-labor ratio, log(export).		

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the horizontal and vertical spillovers of investment from HMT on the productivity of domestic firms. The model is estimated using equation (A6.1). Endogenous variables are instrumented using equation (A10.1) (DID specification).

11.The Effects of Common Dialect on the Share of HMT Firms with Different Bandwidth

Table A9: The Effects of Common Dialect on the Share of HMT Firms (Different bandwidth):

	(1)	(2)	(3)
	1st degree	2nd degree	3rd degree
Dependent Variables: HMT Employment Share			
Border specifications:			
<20km of dialect borders	0.045	0.057	0.129*
	(0030)	(0.048)	(0.068)
<30km of dialect borders	0.070***	0.040	0.061
	(0.024)	(0.039)	(0.055)
<50km of dialect borders	0.052***	0.045	0.104**
	(0.019)	(0.029)	(0.041)
<60km of dialect borders	0.043**	0.056**	0.068*
	(0.018)	(0.027)	(0.036)

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the employment share of HMT firms with various definitions of bandwidth (from 20 to 60km). Column (1) to (3) are estimated using equation (3) with various degree of polynomials of distance to the borders as control variables. Estimated coefficients β_1 , which are the measure of discontinuous changes at the border, are reported in the table.

12. Additional Effects on Crowding-out

This section shows additional evidence on crowding-out of domestic firms. Table A10 estimate equation (5) and (6) with the dependent variable changed to domestic equity share of a specific firm. From Table A10, we find that the coefficients on the common dialect dummy variable are generally negative, the coefficients on the interaction between common dialect and FDI regulation are positive and the coefficients on the interaction between common dialect and the productivity of HMT firms are negative. Thus, we can conclude that the share of domestic firms decreases at the dialect borders and the discontinuous decrease is larger in magnitude in industries not under FDI entry-regulation and in industries in which HMT firms are more productive. Because the effects on domestic firms are exactly in contrary to the effects on HMT firms in signs, we interpret this evidence as showing that domestic firms are crowed out of the market by the entry of HMT firms at the borders of the dialect zones.

Similarly, in Table A11 I analyze the effects of common dialect on the share of other foreign firms by industries. Most of the estimates are statistically insignificant, indicating that other foreign firms are not significantly affected by the entry of HMT firms.

Table A10: Heterogeneous Effects of Common Dialect on the Share of Domestic Firms by Industry (FDI Regulation and Productivity of HMT firms)

	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Dependent Variable: Domestic equity share					
Panel A: By whether the industry is under FDI entry-regulation					
Common Dialect	-0.085**	-0.12**	-0.073	-0.042	-0.12***
	(0.040)	(0.057)	(0.074)	(0.082)	(0.044)
Common Dialect*FDI Regulation	0.065*	0.067*	0.068*	0.070*	0.044
	(0.038)	(0.038)	(0.038)	(0.038)	(0.035)
Bandwidth					30
Observations	77537	77537	77537	77537	58060
Panel B: By the productivity of HMT firms for each industry					
Common Dialect	-0.14***	-0.18***	-0.14*	-0.12	-0.21***
	(0.053)	(0.069)	(0.084)	(0.089)	(0.053)
Common Dialect*Produ-ctivity of HMT firms	-0.34***	-0.35***	-0.34***	-0.35***	-0.41***
	(0.12)	(0.12)	(0.12)	(0.13)	(0.13)
Bandwidth					30
Observations	77466	77466	77466	77466	58000
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei				

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of Domestic firms in different industries. Panel A compares the effects in industries under FDI entry-regulation with industries not under regulation. Panel B compares the effects in industries where HMT firms are good at producing (high productivity) with industries where HMT firms are not good at producing (low productivity). All models are estimated following spatial regression discontinuity design using a sample of firms within 40 kilometers of the dialect borders [model (5) and model (6)]. The dependent variable is HMT equity share for each firm. Column (1) to column (4) estimate the models by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. In panel A, I report the coefficients on the common dialect dummy variable and the coefficients on the interaction between common dialect and whether the industry is under FDI regulation. Similarly, in panel B, I report the coefficients on the common dialect dummy variable and the interaction between common dialect and the productivity of HMT firms for each industry.

Table A11: Heterogeneous Effects of Common Dialect on the Share of Other Foreign Firms by Industry (FDI Regulation and Productivity of HMT firms)

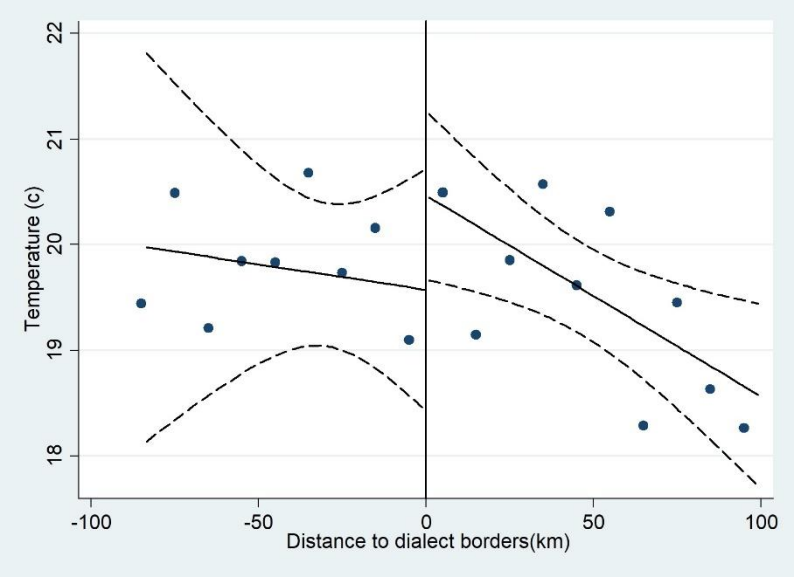
	(1)	(2)	(3)	(4)	(5)
	1st degree	2nd degree	3rd degree	4th degree	Local linear
Dependent Variable: Other foreign firms' equity share					
Panel A: By whether the industry is under FDI entry-regulation					
Common Dialect	0.020	0.035*	0.016	-0.015	0.028*
	(0.013)	(0.021)	(0.027)	(0.029)	(0.016)
Common Dialect*FDI	-0.012	-0.013	-0.013	-0.015	-0.0072
Regulation	(0.011)	(0.012)	(0.012)	(0.012)	(0.011)
Bandwidth					30
Observations	77537	77537	77537	77537	58060
Panel B: By the productivity of HMT firms for each industry					
Common Dialect	0.019	0.035	0.016	-0.014	0.029**
	(0.013)	(0.022)	(0.028)	(0.032)	(0.014)
Common Dialect*Produ	0.014	0.018	0.016	0.016	0.012
-ctivity of HMT firms	(0.036)	(0.036)	(0.036)	(0.0370	(0.035)
Bandwidth					30
Observations	77466	77466	77466	77466	58000
Control variables	County fixed effects, year fixed effects, industry fixed effects, distance to Hong Kong or Taipei				

Notes: Robust standard errors clustered at zip code level are in parenthesis. *, ** and *** indicate statistical significance at 10%, 5% and 1% level respectively. This table shows the effects of common dialect on the share of firms from other countries in different industries. Panel A compares the effects in industries under FDI entry-regulation with industries not under regulation. Panel B compares the effects in industries where HMT firms are good at producing (high productivity) with industries where HMT firms are not good at producing (low productivity). All models are estimated following spatial regression discontinuity design using a sample of firms within 40 kilometers of the dialect borders [model (5) and model (6)]. The dependent variable is HMT equity share for each firm. Column (1) to column (4) estimate the models by controlling the first to fourth degree polynomials of distance to the dialect borders respectively. Column (5) estimate a local linear model with bandwidth chosen to be 30 km. In panel A, I report the coefficients on the common dialect dummy variable and the coefficients on the interaction between common dialect and whether the industry is under FDI regulation. Similarly, in panel B, I report the coefficients on the common dialect dummy variable and the interaction between common dialect and the productivity of HMT firms for each industry.

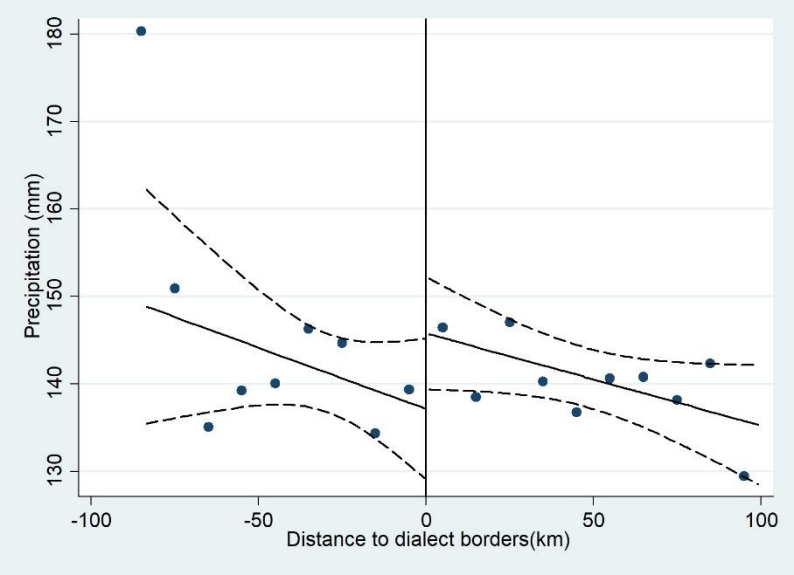
13. The Geographical Distribution of Temperature and Precipitation across the Dialect Borders

Figure A5. The Geographical Distribution of Average Temperature and Precipitation across the Dialect Borders

Temperature



Precipitation



Notes: These figures show the geographical distribution of average temperature and precipitation by distance to the dialect borders. The data are from Terrestrial Air Temperature 1900-2010 Gridded Monthly Time Series and Terrestrial Precipitation 1900-2010 Gridded Monthly Time Series. The resolution is at 0.5 by 0.5 degree (about 55km).