11

Determinants of Spatial Concentration and Local Productivity

The previous chapter discussed various approaches that measure the spatial concentration of economic activity. This line of research is part of a more comprehensive research program that has the ambition of answering the following fundamental questions. The first one, studied in chapter 10, can be formulated as follows: which industries are characterized by high spatial concentration, and how has this spatial concentration changed over time? The second question follows on naturally: what are the determinants underlying the spatial concentration, and are the corresponding explanatory variables consistent with those put forward by theoretical models?

We address these various issues in this chapter. In the first section, we present an approach that consists of regressing industry-specific indices of spatial concentration on a number of explanatory variables suggested by theoretical models, such as the intensity of increasing returns, the level of trade costs, or the importance of intermediate goods. Unfortunately, the selected explanatory variables are often not fully consistent with theory, while the results obtained may be given several conflicting interpretations.

With these shortcomings in mind, the second section introduces a markedly different approach, which focuses on the determinants of sectoral productivity, or growth in each geographical area under consideration, rather than studying solely the overall spatial concentration. This alternative approach makes better use of all available information and allows for a more rigorous interpretation of the results, which may be considered as the estimated specifications of simple theoretical models.

Introducing these two approaches serves to underscore some of the main difficulties encountered in empirical economic geography studies, namely missing variables and endogeneity. They will also allow us to bridge these approaches to the next two chapters, in which the empirical models used are more closely related to those presented in part II.

Before proceeding, the following comment is in order. The contributions discussed in this chapter often use the word "industry," while we have retained the word "sector" in previous chapters. For this reason, we will use industry and sector interchangeably.

11.1 The Determinants of Spatial Concentration

As seen in chapter 10, computing spatial concentration indices for a number of different industries is relatively easy when one has access to regional data, such as industry-specific regional employment. Several authors have taken up the ambitious task of understanding the *determinants* that underlie the values these spatial concentration indices can take.

11.1.1 The Framework

Kim (1995) may be viewed as a precursor in this field, and his work has inspired many researchers. His starting point was to regress a spatial concentration index on variables suggested by theory and, hence, expected to have a significant degree of explanatory power. Let $I_{s,t}$ be the index of spatial concentration for sector s at date t (across regions in a given country, for instance), and let $X_{s,t}$ be the vector of explanatory variables. The approach consists of estimating a vector of parameters β and two other sets of parameters γ_s and δ_t (one for each sector and one for each date, as discussed below in section 11.1.3) such that

$$I_{s,t} = X_{s,t}\beta + \gamma_s + \delta_t + \varepsilon_{s,t}, \tag{11.1}$$

where $\varepsilon_{s,t}$ is an error term. Kim (1995) considers two explanatory variables in the $X_{s,t}$ vector:

(i) the average size of firms in a specific sector at a given date and

(ii) the share of raw materials used in this sector.

These two variables are not as far-fetched as they may appear at first sight. In fact, they characterize the two main lines of research followed in explaining the spatial distribution of production, namely economic geography and standard trade theory.

As seen from chapter 2 onward, increasing returns appear to be necessary to account for the spatial concentration of economic activities, at least when space is homogeneous. An intuitive test is thus to verify whether the industries in which returns to scale are stronger do indeed correspond to those in which spatial concentration is greater. Unfortunately, data measuring the level of scale economies in a given industry are not available. Kim uses the average size of firms in each industry as a proxy: under zero profits, the stronger increasing returns are, the larger the size of plants. Specifically, finding a positive and statistically significant coefficient for a plant's average size would confirm the idea that increasing returns help to account for spatial concentration. When regressing the Gini indices computed on U.S. data in 1880, 1914, 1947, 1967, and 1987 for twenty industries, Kim finds that scale economies have a positive impact on the spatial concentration index. Several authors have tried to reproduce these results in the case of Europe. For example, Amiti (1999) adopts the same approach and also observes a positive correlation between scale economies and the spatial concentration of different industries. However, for reasons that will become clear below, the robustness of those results is questionable.

11.1.2 Omitted Variables

Economic theory typically focuses on one particular effect by controlling for (i.e., neutralizing) a large number of others that are at work in the real world. For instance, in order to isolate the trade-off between increasing returns and trade costs, we have assumed in part II that regions share the same technologies, endowments, and preferences. Such region-specific variables are sources of potential heterogeneity that could blur the trade-off we want to study. Yet these sources of heterogeneity are key variables in standard trade theories and, hence, could play an important role in shaping the spatial distribution of activity. Hence, while omitting these effects is legitimate from the theoretical standpoint, this approach makes little sense in empirical studies whose purpose is precisely to explain reality as well as possible, thus calling for the inclusion of as many relevant variables as necessary. Moreover, it is reasonable to believe that real-world patterns of activity are the outcome of the interplay between the main variables of economic geography and standard trade theory. The challenge is then to discriminate between these two approaches by determining which one accounts for the greater share of regional specialization or agglomeration. The set of economic geography variables must, therefore, be supplemented by control variables that account for the effects of regional heterogeneity. This is what Kim attempts, by using a country's share of raw materials as a control for natural endowments. The idea is that industries that are intensive in using raw material should be agglomerated because of their dependency on the supply of these inputs, and not because of increasing returns.

Unfortunately, this approach has severe limitations and there is little hope that it can provide convincing results. First, when attempting to account for the different degrees of spatial concentration observed across industries, the models presented in part II reveal that scale economies are not the only source of agglomeration. Several relevant explanatory variables are totally absent. Specifically, trade costs, which may vary substantially across goods, are missing (chapter 5). In the context of monopolistic competition, also absent is the elasticity of substitution that can vary substantially across industries and time (chapter 6). Another potential shortcoming is the fact that intermediate goods are not taken into account in the regression, a variable that has proved particularly important in some contexts (chapter 7). All the models presented in part II show that the list of omitted variables could be extended further. It is, therefore, somewhat naive to expect a single variable (here, the average size of firms in a given industry) to capture all these effects adequately. Obviously, if these effects happened to be distributed randomly across industries, their omission would have no impact as the error term's very function is to capture such effects in regression analyses. Unfortunately, such a strong assumption is rarely accurate in practice.

This problem goes under the general heading of *omitted variables*, and econometricians have long emphasized the biased estimates that can result. The omitted variable bias is not specific to a particular set of variables: it applies both to economic geography and other variables, which is the second main drawback of Kim's approach. In this respect, when he considers the share of raw materials in production and excludes any other explanatory variables, he makes other strong assumptions. For instance, absent from Kim's model are capital and labor intensities, two variables at the heart of Heckscher and Ohlin's theory. Along the same lines (and very relevant in modern economies), incorporating variables that distinguish between skilled and unskilled labor intensities might be an important addition. These variables are omitted in Kim (1995). Moreover, while controlling for factor intensities in production is undoubtedly important, it is misleading to study their role without taking into account how production factors are distributed across space. Indeed, factor intensities matter for spatial concentration in standard trade theory when the distribution of factors across regions is uneven.

11.1.3 Fixed Effects

A first solution to deal with omitted variables was implemented by Kim (1995) himself. It requires access to panel data across different industries and for different time periods, which allows one to use the method of *fixed effects*. In (11.1), γ_s *is an industry fixed effect, which is a dummy that equals 1 for all observations corresponding to sector s and 0 otherwise*. When a complete set of fixed effects is included in an econometric specification, they are perfectly collinear with the intercept (the constant term), so one of them must be dropped in the estimation. Thus, the excluded industry becomes the reference for the others. For instance, the fixed effect corresponding to each of the remaining industries measures, everything else being equal, the difference in spatial concentration between the sector under consideration and the excluded sector. The concentration of the latter (net of the impact of explanatory variables) is measured by the intercept. Similarly, γ_t is a *time* fixed effect; it takes a value of 1 for each observation corresponding to year t and a value of 0 otherwise. Again, one year dummy must be excluded to avoid collinearity.[1]

Fixed effects have the advantage of controlling for all variables that are constant over time but specific to each industry (the industry fixed effects) or constant across industries but proper to each time period (the time fixed effects), without the need for any data related to these variables. The major drawback of using fixed effects is that they allow for the estimation of the overall contribution of these variables but not for the estimation of the effect of each one separately.

In this context, β represents what is called the "between time and industry" effect. It captures the correlation between $I_{s,t}$ and $X_{s,t}$ across industry and time simultaneously, but not their cross-section correlation or their time correlation only. If $I_{s,t}$ varies only across industries and keeps the same value across time, β will be zero when the equation is estimated with industry fixed effects. Similarly, it would be also zero with time fixed effects if $I_{s,t}$ were to vary across time only. In other words, β is nonzero when $I_{s,t}$ and $X_{s,t}$ deviate in a correlated way once normalized by their industry averages (denoted $I_{s,\cdot}$ and $X_{s,\cdot}$) and by their time averages (denoted $I_{\cdot,t}$ and $X_{\cdot,t}$). In this case, estimating (11.1) or

$$I_{s,t} - I_{s,\cdot} - I_{\cdot,t} = (X_{s,t} - X_{s,\cdot} - X_{\cdot,t})\beta + \epsilon_{s,t}, \tag{11.2}$$

obtained from (11.1) by simple manipulations, where $\epsilon_{s,t}$ is a new error term, leads to the same estimate of β (under certain conditions on the distribution of errors). This new formulation illustrates the fact that the correct interpretation is in terms of correlated departures of $I_{s,t}$ and

[1] Alternatively, one could choose to exclude both the constant and an industry (respectively, time) dummy in order to keep all time (respectively, industry) dummies. Such choices merely amount to different normalizations.

$X_{s,t}$ *from their industry and time averages*. By contrast, β will likely b different when estimated without fixed effects through

$$I_{s,t} = X_{s,t}\beta + \nu_{s,t}$$

despite the fact that this formulation also yields (11.2).

In short, the introduction of sectoral fixed effects is equivalent t assuming that the omitted variables remain constant over time: a assumption that is much less extreme than supposing they have n impact at all. For instance, over a fairly short period, it is reasonabl to assume that differences in elasticities of substitution across indus tries barely vary, and are therefore controlled for by the fixed effects However, such an assumption becomes more problematic when the tim period is long. It becomes even more problematic when dealing wit variables that often exhibit significant variations. In this case, turning t fixed effects does not help much. Similarly, time fixed effects control fo macroeconomic-type shocks, provided these shocks affect all industrie in the same way. For example, an increase in growth on a national scal could, in a given year, temporarily increase the average size of all plant across all industries, thus affecting equally their spatial concentration Cyclical, macroeconomic effects of this nature do not provide any addi tional clues as to the determinants of spatial concentration: they ar absorbed by time fixed effects that leave the impact of the other vari ables unchanged. This makes time fixed effects just as useful as thei sectoral counterparts. Again, just as sectoral fixed effects are ineffectiv when the omitted variables are expected to vary over short time periods the same caveat holds for time fixed effects when omitted variables var across industries.

Furthermore, it is worth noting that fixed effects may be introduce to solve a problem that is specific to the determinants of spatial concen tration. As noted above, a number of indices (e.g., the Gini index) are no comparable across industries (see property 10.1). Keeping this in mind can we hope to infer anything about spatial concentration by comparin the different values these indices take across industries? For instance estimating (11.1) only makes sense if we have been careful in choosin an index of spatial concentration that allows for comparisons betwee industries. However, under a less careful choice of indices, we can stil rely on sectoral fixed effects to partly curb this problem. That said, cor recting a gross concentration index to make it comparable across indus tries, as proposed by Ellison and Glaeser (1997), requires a more comple transformation than the log-linear rescaling corresponding to the fixe effects strategy.

11.1.4 Additional Variables

Using fixed effects requires panel data. Moreover, the existence of omitted variables that vary in both spatial and temporal dimensions remains problematic. Intuitively, an alternative remedy for correcting omitted variable bias is to add new explanatory variables to the vector $X_{s,t}$ in (11.1). A few additional variables that immediately come to mind are the degree of increasing returns and input–output linkages, trade costs, and the extent of technological spillovers, as well as the structure of local labor markets in terms of skilled and unskilled workers.

In this respect, Rosenthal and Strange (2001) adopt one of the most exhaustive specifications to date. Their estimations are conducted on three different spatial scales (U.S. municipalities, counties, and states) and each specification uses Ellison and Glaeser's index as a dependent variable, which accounts for differences in industrial concentration (chapter 10). Given the absence of any available time dimension, Rosenthal and Strange include industry fixed effects, but only at a more aggregated level than when computing the dependent and explanatory variables. Under each of the three spatial scales, their estimates reveal that labor market structure (in terms of skilled versus unskilled labor) has the most robust effect on spatial concentration. The variable accounting for technological spillovers also proves robust, but only at the municipal level, which seems reasonable given the local impact spillovers are expected to have. At the state level, intermediate inputs and natural endowments increase spatial concentration, while trade costs reduce it. Using European data, Amiti (1999) finds that vertical linkages have a statistically significant impact on spatial concentration within Europe.[2]

This approach is still wanting on a number of fronts. First, the compilation of comprehensive databases that cover the whole set of missing variables is often out of reach. Second, a more fundamental problem remains. To date, nobody has been able to show in a theoretical model how any of the spatial concentration indices presented in chapter 10 vary with the explanatory variables. This issue's persistence is hardly surprising, for its solution is extremely involved and requires computing these variables. To better grasp the difficulty of this exercise, it should be stressed that even more modest endeavors continue to look like stumbling blocks: in many models, the analytical expressions of the endogenous variables that underpin spatial concentration indices

[2] See Combes and Overman (2004) for a comprehensive review of this literature, which also discusses studies that examine simple correlations between spatial concentration and a given factor without adding control variables or fixed effects.

cannot be determined. To make matters even more involved, this procedure should be implemented in the context of a model involving several regions and industries, characterized by specific technologies and factor endowments. With these new difficulties, the task at hand seems almost impossible.

To conclude, the following comments are in order. Choosing a spatial concentration index as a dependent variable leads to a substantial loss of information. Indeed, such indices are aggregated variables, while the disaggregated information required to estimate their value is available for each region and sector. It is, therefore, questionable to only study the spatial concentration of a sector within a country, while it seems possible, using the same data, to identify, in any region, the determinants of employment, productivity, or growth of each sector. Furthermore, it seems more promising to use theoretical models to derive functional forms linking these disaggregated variables to explanatory variables. Providing estimations that respect theoretical models down to the last detail is a very difficult task, and we will cover such attempts in chapters 12 and 13.

In the next section, we will focus on a third approach, which, while still a far cry from economic geography models, has the advantage of being easier to interpret, as well as providing relevant results, using more robust methods than those initially put forward by Kim. In so doing, we will encounter endogeneity problems that are recurrent in empirical economic geography. We will see that the solution of endogeneity problems will allow us to solve some of the omitted variable problems discussed above.

11.2 The Determinants of Local Productivity

Our objective is now to discuss some studies that aim at evaluating the impact of the main variables considered in economic geography using what are known as nonstructural or *reduced form* specifications. Such specifications are not directly associated with a particular theoretical model, but can be useful in uncovering new ideas regarding the forces that underlie agglomeration economies.

11.2.1 The Theoretical Background

The main ideas of economic geography can be grasped with the help of a fairly simple model. Specifically, we consider a firm j located in region r and operating in sector s, which uses labor in quantity l_j and other inputs, viewed as a composite, in quantity k_j. We assume that its production is given by a Cobb–Douglas function:

$$y_j = A_j(s_j l_j)^{\mu} k_j^{1-\mu}, \tag{11.3}$$

where A_j is a Hicks-neutral factor-augmenting technology level, and s_j is the efficiency level of workers; both are specific to the firm. This firm's profits are given by

$$\pi_j = \sum_b p_{jb} y_{jb} - w_j l_j - r_j k_j,$$

where y_{jb} is the quantity exported to region b, p_{jb} is the mill price set in region b net of the marginal cost of intermediate goods, w_j is the wage rate, and r_j is the cost of inputs other than labor and intermediate goods. This function may then be rewritten as

$$\pi_j = p_j y_j - w_j l_j - r_j k_j,$$

where

$$p_j = \sum_b p_{jb} \frac{y_{jb}}{y_j}$$

is the average unit value, net of the cost of intermediate inputs, of the good produced by the firm. Hence, $p_j y_j$ denotes the firm's value-added and not the value of its production. This change is made in order to match data. Applying the first-order conditions to the firm's profit-maximizing problem and rearranging terms yields the following two equations:

$$w_j = \mu p_j A_j s_j^{\mu} \left(\frac{k_j}{l_j}\right)^{1-\mu} \quad \text{and} \quad r_j = (1-\mu) p_j A_j s_j^{\mu} \left(\frac{k_j}{l_j}\right)^{-\mu}. \tag{11.4}$$

By plugging the second expression into the first, we obtain

$$w_j = \mu(1-\mu)^{(1-\mu)/\mu} s_j \left(\frac{p_j A_j}{r_j^{1-\mu}}\right)^{1/\mu}. \tag{11.5}$$

Equation (11.5) requires individual-level wage data, which has only been made available very recently. Previous work relied on average wage in region r and sector s, which takes the following form:

$$w_{rs} = \frac{\mu(1-\mu)^{(1-\mu)/\mu}}{n_{rs}} \sum_{j \in (rs)} s_j \left(\frac{p_j A_j}{r_j^{1-\mu}}\right)^{1/\mu}, \tag{11.6}$$

where n_{rs} is the number of firms in region r and sector s.

In which region is the marginal productivity of labor, which is equal to the equilibrium wage, the highest? Equation (11.5) shows that wages are directly proportional to workers' efficiency, as reflected by s_j. While this finding is not specific to economic geography, we will see that it is crucial to keep it in mind when studying interregional wage differences. Moreover, (11.5) takes into account the variables p_j and r_j, which capture the main agglomeration and dispersion forces described in part II. A greater p_j (be it because demand is high, competition is weak, or because intermediate goods are cheap) translates to a higher wage, which in turn contributes to a higher degree of agglomeration of workers in that region. Conversely, low demand or fiercer competition brings down wages in a region, thus encouraging workers to leave it. The presence of r_j in the wage equation captures the effects transmitted through other factor prices. For instance, if a number of new suppliers were to move closer to their customers (i.e., an increase in the supply of a given production factor), the price of the corresponding factor would decrease. This, in turn, would translate to an increase in wages. Conversely, when production factors have a low elasticity of supply (land being the typical example), prices for these factors will be higher in areas characterized by more concentrated economic activity, which pushes down the wage rate. The models presented in part II serve the exact purpose of delving into these mechanisms, giving them micro-foundations, while they are conveniently expressed here by the "black boxes" p_j and r_j.

So far we have refrained from introducing technological externalities. This choice was made to avoid imposing any ad hoc components, with the objective of isolating phenomena that are micro-founded and endogenous. Yet Marshall has stressed the potential importance of technological externalities, such as knowledge and learning spillovers. They are taken into account here through the term A_j. Intuitively, regions characterized by an easy circulation of information and/or endowed with a high concentration of skilled workers are likely to benefit from more productive technologies, thus implying higher wages, as shown by (11.5). On the other hand, one would expect a heavily congested transportation network, or the emergence of high levels of pollution in densely populated areas, to worsen productivity and to act as dispersion forces through the corresponding decline in wages.

In short, the *wage equation* (11.5), or its aggregated version (11.6), captures the full breadth of agglomeration and dispersion forces, even though the microeconomic foundations of the underlying model are kept deliberately vague. For example, a number of details have been glossed over, including consumer preferences or the assumptions regarding the mobility of goods and factors. Recall that our goal here is not to construct a fully fledged economic geography model, as in previous chapters. Rather, constructing a simple framework in which prices and costs depend on both region and sector characteristics provides a clear vantage point from which to better understand the empirical results presented below.

Given that wage data are often available on a local scale for a number of different industries, most of the existing works use wages as the dependent variable. However, when data related to value-added and capital stocks are available, the possibility of conducting similar estimations by using the average productivity of labor, or total factor productivity, should not be overlooked. Specifically, it follows from (11.3) and (11.4) that the average labor productivity is given by

$$\frac{p_j y_j}{l_j} = (1-\mu)^{(1-\mu)/\mu} s_j \left(\frac{p_j^{1-\mu} A_j}{r_j^{1-\mu}} \right)^{1/\mu} \qquad (11.7)$$

and the total factor productivity is given by

$$\frac{p_j y_j}{l_j^{\mu} k_j^{1-\mu}} = p_j A_j (s_j)^{\mu}. \qquad (11.8)$$

Observe that these two expressions are almost identical to (11.5) in that the left-hand side variables correspond to various productivity measures that are all linked to the same right-hand side variables: local input and output prices and the local levels of technology and labor efficiency.[3] Note also that the costs of inputs other than labor do not appear in (11.7) and (11.8).

11.2.2 The Econometric Analysis

One of the most important empirical questions in economic geography might read as follows: *is productivity higher in areas characterized by highly concentrated economic activity*, and if so, by how much? In other words, the first task is to uncover any existing correlation between the value of local productivity and the density of economic activities in the same region. A simple thought experiment is to consider the percentage change in productivity brought about by doubling employment or population density. Answering this type of question seems fairly straightforward. Specifically, we regress either the total factor productivity or, more often, the nominal wage on the employment (or population) density:[4]

$$\ln w_{rs} = \alpha + \beta \ln \text{den}_r + \varepsilon_{rs}, \qquad (11.9)$$

[3] It should be kept in mind that talking about productivity is a slight abuse of language because $p_j y_j$ is not the value of production but the value-added.

[4] In order to interpret the coefficient in terms of elasticity, we take the logarithm of all variables.

where ε_{rs} is an error term and $\text{den}_r = \text{emp}_r/\text{area}_r$ is the total number of employees in region r (emp_r) divided by its surface area (area_r). The estimated coefficient that results from this regression indicates that a 1% higher density implies a β% higher productivity (if β is positive). For a density twice as high, wages increase by $(2^{\beta} - 1) \times 100\%$.[5]

As with nearly all of the studies presented in section 11.1, a number of econometric problems arise. To begin, it is worth noting that estimating (11.9) is equivalent to estimating (11.6) under the following assumption:

$$\ln \frac{1}{n_{rs}} \sum_{j \in (rs)} s_j \left(\frac{p_j A_j}{r_j^{1-\mu}} \right)^{1/\mu} = \beta \ln \text{den}_r + \varepsilon_{rs}. \qquad (11.10)$$

Thus, the implicit assumption is that the density affects the wage level through the following variables:

(i) the local level of technology, A_j,

(ii) the output price, p_j,

(iii) the input prices other than labor, r_j, or

(iv) the local efficiency of labor, s_j.

However, we are not able to determine which variables are most affected. Furthermore, only the *net* effect of density is identified, leaving us in the dark as to whether the possible negative impact on some variables is compensated by the possible positive impact on others. That said, knowing the net effect is still of critical importance to public decision makers who might want to design policies that aim to concentrate or disperse activities. Once a given policy has been implemented, the present framework also allows for total net productivity gains or losses to be quantified.

When considering the sources of potential econometric bias, of chief concern are the potentially large number of omitted variables, an issue discussed above and which will be illustrated here using wage data. Before moving on, let us stress the main advantage of expressing all variables in logarithmic form. Aside from facilitating interpretation (the estimated coefficients become elasticities), taking logarithms brings residuals closer to the normal distribution (recall that, in regression analysis, a number of statistical tests assume that residuals are normally distributed).

A large fraction of regional differences in labor productivity stems not from the presence of local externalities but from the fact that some

[5] Consider two individuals located in regions 1 and 2, respectively, that differ only in terms of density. Then, (11.9) implies that their difference in productivity is such that $\log(w_2/w_1) = \beta \log(\text{den}_2/\text{den}_1)$. When $\text{den}_2/\text{den}_1 = k$, we have $w_2/w_1 = k^{\beta}$.

workers have a higher level of skill than others. Overlooking variables that account for differences in average regional skill levels is equivalent to assuming that labor skills are randomly distributed across regions and captured by the term ε_{rs}. Since this assumption is easily refuted empirically, it is standard practice to introduce control variables that capture workers' skills, qualifications, or academic achievement in the regression. It is straightforward to figure out what happens when these variables are omitted. If workers are more skilled in regions characterized by highly concentrated economic activity (which is generally the case), overlooking such variables overestimates the impact of density, because this variable also captures the influence of s_j.

Note that the variable w_{rs} we seek to explain depends on both the region r and the sector s, while the explanatory variable considered in (11.9) (density) varies across regions but not across sectors. Therefore, the literature usually also tries to control for the region's industrial mix, i.e., for the way in which local economic activity is distributed across a range of industries. Indeed, regions with the same density may have very different industries, or have the same industries but in very different proportions. For example, if the good is sold to a small number of industries, or if the factors used are industry specific, the industrial mix is crucial because it affects the level of productivity through the prices effects described above.[6] The industry's share in local economic activity is the first variable that is usually included in the specification:

$$\mathrm{spe}_{rs} = \frac{\mathrm{emp}_{rs}}{\mathrm{emp}_r},$$

where emp_{rs} is employment in sector s and region r. By measuring the relative size of sector s in the local economy, the specialization index allows us to capture the effects of *intraindustry externalities* (resulting from the concentration of this sector only) and to distinguish them from *interindustry externalities* (resulting from the concentration of the overall activity), which are likely to be apprehended by the density variable. Knowing the relative importance of these two types of externalities is a major issue for the design of regional development policies. Indeed, this knowledge would allow public decision makers to design policies that would either favor the concentration of a handful of industries, as in the case of the Italian industrial districts, or welcome any industry because all of them would benefit from the externalities generated by the others.

[6]Note, however, that the industrial mix is much less important when the good under consideration is sold to most local industries and/or is designed for final consumers. The industrial mix is also rather unimportant when the inputs used to produce the good come from many local industries and/or mainly consist of labor.

Some authors further extend the set of explanatory variables and consider other kinds of intraindustry and interindustry externalities. Regarding the former, the number of local plants in the sector is a variable that allows us to determine whether intraindustry externalities depend on the average size of plants in the local industry rather than on the total number of employees (already captured by spe_{rs}). As for the interindustry externalities, an "industrial diversity" variable is often added. For given density and size of an industry, such a variable aims at evaluating how the distribution of employment spreads over the other local sectors and, therefore, at determining whether the industry benefits from the others. The inverse of the Herfindhal index in terms of industries' shares in regional employment is often used:

$$\mathrm{div}_r = \left[\sum_s \left(\frac{\mathrm{emp}_{rs}}{\mathrm{emp}_r}\right)^2\right]^{-1}.$$

Finally, it might be worth including each regions' surface area area_r in the explanatory variables. Indeed, for a given density, the absolute size of a region may play an important role, as it accounts for the total population on which externalities are built.

Note that several specifications expressed in logarithms are formally equivalent. For example, estimating the model

$$\ln w_{rs} = \beta \ln \mathrm{den}_r + \eta \ln \mathrm{area}_r + \varepsilon_{rs} \tag{11.11}$$

is equivalent to estimating

$$\ln w_{rs} = \beta \ln \mathrm{emp}_{rs} + \varrho \ln \mathrm{area}_r + \varepsilon_{rs}, \tag{11.12}$$

since $\varrho = \eta - \beta$. Interpreting econometric results, therefore, warrants a degree of caution. For example, the effect of an increase in density for a given surface area (β in (11.11)) is tantamount to an increase in the employment level for a given surface area (β in (11.12)). However, if the density is held constant, an increase in surface area (η in (11.11)) is not equivalent to the same increase when the employment level is kept fixed (ϱ in (11.12)), since the former requires a proportional increase in employment for the density to remain the same.

More variables that should be controlled for are known under the general heading of *natural amenities* and *local public goods*. Natural amenities are benefits ranging from a favorable climate, a coast-line location, and the presence of lakes and mountains to any natural endowments in raw materials. However, it should be stressed that the level of some amenities is the outcome of public policies; think of leisure facilities (theaters, swimming pools, etc.) or public services (schools, hospitals, etc.).

Public goods are said to be local when their benefits are only reaped by local consumers, while the access costs of using these goods by more distant consumers are very high. Local public goods can also be used by firms. Transport infrastructures, research laboratories, and job training centers are just a few examples. What happens when these amenities and local public goods are not included in the regressions? Local public goods inflate the productivity of production factors, such as labor and intermediate goods. If these local public goods were randomly distributed across space, their omission would be taken into account by the error term. Unfortunately, the supply of local public goods is the outcome of specific policies and often greater in areas characterized by concentrated economic activity. In this case, the effect of density is overestimated, as the density variable also captures the positive effect of these (omitted) local inputs. As shown by Roback (1982), dealing with natural amenities is slightly more involved. To see this, assume that a region is endowed with such amenities which attract migrants, all else being equal. The inflow of this new population exerts an upward pressure on the demand for housing, thereby pushing up rents. Higher land rents induce firms to substitute other production factors, such as labor, for land. As the marginal productivity of labor decreases, land–labor substitution leads to a drop in wages. When natural amenities are more abundant in heavily populated regions (as is the case for leisure facilities), the effect of density is thus underestimated. The key point is that omitted variables such as these can bias estimates in both directions, thus leaving us in the dark as to the magnitude and direction of the bias.

In the spatial context, there is still another group of omitted variables. All the explanatory variables considered so far have been restricted to the geographic area r under consideration; none have taken into account effects, such as interindustry or intraindustry externalities, that could emanate from *neighboring areas*. In other words, the implicit assumption so far has been a complete absence of nonmarket interactions between areas. Everything is estimated under the presumption that no spillover effects exist *between* regions or that those are randomly distributed across regions. If, as suggested in chapter 5, distance has a negative impact on interregional interactions (via trade flows or knowledge transfers), such an assumption seems untenable. It is undoubtedly the main weakness of the approach presented so far. Very few attempts have been made to correct the resulting biased estimates. First of all, a market-potential variable defined as the sum of each region's density weighted by the inverse of its distance to this area can be introduced.[7] Another

[7] See chapter 12 for a detailed discussion of the concept of market potential.

approach consists of using techniques borrowed from spatial econometrics, by adding spatially lagged variables and accounting for the potential autocorrelation of the residuals.[8] In both cases, the objective of introducing such variables is to correct for an econometric bias, but they are often introduced in an ad hoc manner (for instance, functional forms for distance-decay effects are chosen arbitrarily) and might be difficult to interpret. We will see in the next chapters how the introduction of such variables may be better justified.

In a way, we find ourselves with the familiar quest of adding to our regressions a seemingly endless string of control variables. As in section 11.1, using fixed effects is an option. Namely, when a panel of industries in different regions is available, we may introduce region and industry fixed effects to control for omitted variables. For instance, we can evaluate the extent of interindustry externalities controlling for regional fixed effects, provided that at least two years of data are available, and making the reasonable assumption that amenity and public good endowments are constant during the short time period under consideration. In the same vein, industry fixed effects can be introduced. Indeed, in addition to controlling for missing sector-specific variables, they are necessary to capture differences in labor shares across different industries: replacing μ with μ_s in (11.6) implies in turn that the intercept α in (11.9) should be industry specific. As more and more data become available, we should even consider industry–time fixed effects in order to purge the model of business cycle effects that are specific to some regions.

11.2.3 Endogeneity Bias

The above approaches shed light on a more general problem that often plagues empirical studies in economic geography: the endogeneity of some explanatory variables. Formally, OLS estimates are biased when some explanatory variables are correlated with the residuals of the regression. These variables are then said to be *endogenous*. The presence of such a correlation can be tested with the help of appropriate statistical techniques, provided a sufficient number of exogenous variables are available. Using the density variable as an example, we first want to obtain some clues as to the nature of the endogeneity problem. To this end, assume that a given region experiences a shock observed by economic agents but overlooked by the econometrician. For example, a positive shock may stem from the decisions made by regional governments that lead to a higher local productivity; conversely, a hike in

[8] See Bailey and Gatrell (1995) for an introduction to these techniques, and Anselin et al. (2003) for a more advanced presentation.

the oil price is a negative shock for regions having several oil-intensive industries. Some of these shocks may randomly affect the productivity of all inputs and may, therefore, be assumed to be independently and identically distributed across regions. In this case, they would be completely absorbed by the residual term, ε_{rs}. However, in economic geography, shocks are often localized and thus have an impact on the location of agents, who are attracted by regions benefiting from positive shocks (generating wage increases) and repelled by those suffering from negative shocks. These relocations obviously have an impact on regions' levels of economic activity and, consequently, on their density of regional employment. In other words, the employment density is necessarily correlated with the residuals (it is positive in our example):

$$\mathrm{corr}(\ln \mathrm{den}_r, \varepsilon_{rs}) \neq 0.$$

Density is thus endogenous, which contradicts one of the assumptions underpinning the validity of the OLS estimator, biasing it upwards here. Endogeneity is often framed as a problem of *reverse causality*: the unobserved shock initially affects wages, and thus density, through the mobility of workers, and not the other way around as equation (11.9) implies. If, however, the production factors were to be nearly immobile, one would expect the endogeneity bias to be weaker. That said, even in the context of immobile production factors, a given shock may affect the level of regional employment via the creation and destruction of jobs. As a result, the employment density variable would again be endogenous.

We want to stress the difference between the endogeneity problem in econometrics and the choice of endogenous variables in economic models, namely those that are determined in equilibrium. As mentioned above, in econometrics, endogeneity arises when some explanatory variables and the residuals are correlated. Thus, variables that are endogenous in the economic sense are likely to be endogenous from an econometric point of view. Even explanatory variables that are not directly correlated with the residuals may be tied to other endogenous variables (via the equation system describing the equilibrium outcome) which are themselves correlated with the residuals. This need not be the case, however. One may come across situations in which variables are endogenous in the economic sense but exogenous from the econometric standpoint, and vice versa. It all depends on the economic interpretation of the residuals, the determination of which is therefore a crucial step in the specification of an econometric model. Assessing the degree of econometric endogeneity of a given explanatory variable is only possible once the source of the economic model's residuals has been clearly identified.

The endogeneity problem is not specific to economic geography; the issues it generates are encountered in many other fields of economics. The issues' pervasiveness has one clear benefit: a wide variety of techniques have been proposed to address them. The most common approach involves using what are known as *instrumental variable techniques*. This consists of finding variables, called *instruments*, that are correlated with the endogenous explanatory variables but *not* with the residuals. The first step is to regress the variable whose exogeneity is suspect on the chosen instrument(s). In the present context, we may regress the density of regional employment at date t on the region's density several decades earlier. Such an instrumental regression may, for instance, be expressed as follows:

$$\ln \mathrm{den}_r = \rho \ln \mathrm{den}_{r,t-150} + \nu_r,$$

where $\mathrm{den}_{r,t-150}$ is the region's density 150 years before the year of interest and ν_r is an error term. This provides us with a predicted value for the density given by $\widehat{\ln \mathrm{den}_r} = \hat{\rho} \ln \mathrm{den}_{r,t-150}$, where $\hat{\rho}$ is the OLS estimator for ρ. In the next step, the density in the initial regression (11.9) is replaced by its predicted value (the explanatory variable den_r is then said to be instrumented), which is uncorrelated with the residuals since the instrument is by definition exogenous:

$$\begin{aligned}\mathrm{corr}(\widehat{\ln \mathrm{den}_r}, \varepsilon_{rs}) &= \mathrm{corr}(\hat{\rho} \ln \mathrm{den}_{r,t-150}, \varepsilon_{rs}) \\ &= \mathrm{corr}(\ln \mathrm{den}_{r,t-150}, \varepsilon_{rs}) \\ &= 0.\end{aligned}$$

In this case, the OLS estimate of the equation

$$\ln w_{rs} = \alpha + \beta \widehat{\ln \mathrm{den}_r} + \varepsilon_{rs}$$

no longer suffers from endogeneity bias and provides an unbiased estimate of the effect of density (see Wooldridge (2002) for further details).

A few comments are in order. First, everything rests on the alleged exogeneity of the chosen instrument. Once again, both economic and econometric considerations must be taken into account. From the economic standpoint, in the density example, it is quite plausible that there is no correlation between past employment density and present-day productivity shocks. However, a time gap, be it 150 years or longer, is not necessarily a sufficient condition for exogeneity, because the source of a shock may be linked to unobserved factors that persist over time. Bearing this in mind, it is imperative to ponder all possible sources of endogeneity for both the explanatory variables and the possible instruments. Regardless of our confidence about the exogeneity of this or that variable, it is standard practice in econometrics to carry out *overidentification tests*, which can be interpreted as exogeneity tests for some of the proposed instruments. These tests are relatively straightforward, but they require the number of instruments to be greater than the number of instrumented variables. Regarding density, additional instruments might be given by past population levels at several different dates, or by former population growth rates. Other potential instruments may be based on regional skill endowments, as measured by the past regional levels of literacy or numbers of students.

Another advantage of using instrumental variable techniques is that it may address problems related to omitted variables. Indeed, as for reverse causality, they can also be framed in terms of a correlation between one or more explanatory variables and the residuals. To illustrate, let us assume we have omitted public infrastructure from our regression (whose effect is therefore captured by the residuals), and that such an infrastructure is more prevalent in dense areas. As a result, a positive correlation emerges between the residuals and one of the explanatory variables (the density again), generating the upward bias mentioned above. Given that the current level of public infrastructure can often be traced back only to recent governments' decisions, it should not be correlated with the population level several decades ago. In running the instrumental regression, any existing correlation between the current density and infrastructure is thus relegated to the residuals, which means that the new predicted value of density is free of omitted variable bias.[9]

Finally, it is worth noting that the endogeneity problem addressed in this subsection has been illustrated only for the density variable. Almost all other variables, discussed above and usually introduced in this type of regression, are, however, likely to be endogenous. For example, any variables related to the industrial structure are intimately linked to location decisions, which also leads to biased OLS estimators.

11.2.4 The Impact of Density on Wages

In practice, what is the extent of economies of density and of the biases arising from omitted variables and endogeneity? Results drawn from

[9]Note that the presence of omitted variables and the existence of reverse causality both bias OLS estimators by producing a correlation between one or more explanatory variables and the residuals. However, the source of this correlation is not the same. In the first case, the residuals are not random because they are correlated with omitted variables that are not random. In the second case, the residuals may be random *ex ante*, but their realizations, observed *ex post* by the agents, lead to decisions that affect the explanatory variables, thus making them correlated with the residuals.

Combes et al. (2008b) provide a useful starting point from which to address this question. This study estimates the magnitude of agglomeration economies on the basis of disaggregated French data, available at the individual level. Namely, the data set gives the location $r(i,t)$ and the sector $s(i,t)$ associated with each worker i at time t. Furthermore, it covers a time period spanning 1976 to 1998. The dependent variable is a worker's wage at a given date. The resulting specification bears some resemblance to the model presented above by assuming that the amount of efficient labor used in firm j at date t is expressed as follows (all variables now also depend on date t):

$$s_{j,t} l_{j,t} = \sum_{i \in (j,t)} s_{i,t} \ell_{i,t},$$

where $s_{i,t}$ is the efficiency of worker i at date t and $\ell_{i,t}$ is his supply of labor. In equilibrium, the first-order condition yields

$$w_{i,t} = \mu(1-\mu)^{(1-\mu)/\mu} s_{i,t} \left(\frac{p_{j,t} A_{j,t}}{r_{j,t}^{1-\mu}} \right)^{1/\mu},$$

so that the wage equation to be estimated is

$$\ln w_{i,t} = \theta_i + \lambda \,\text{age}_{i,t} + \mu(\text{age}_{i,t})^2 + X_{r(i,t),t}\beta + Z_{r(i,t)s(i,t),t}\phi + \gamma_{s(i,t)} + \delta_t + \epsilon_{i,t}, \quad (11.13)$$

where $\epsilon_{i,t}$ represents an individual-specific productivity shock, while the remaining four groups of variables explain the wage rate. More precisely, $X_{r(i,t),t}$ is a vector of variables associated with the worker's location $r(i,t)$ at date t, the aim of which is to capture interindustry externalities (density, surface area, and diversity); $Z_{r(i,t)s(i,t),t}$ is a vector of variables that capture intraindustry externalities (specialization and number of firms); $\gamma_{s(i,t)}$ and δ_t are industry and time fixed effects. Finally, the worker-specific variables, which depend directly on i and t, constitute the fourth group; they capture the impact of a given worker's skills $s_{i,t} \equiv \theta_i + \lambda \,\text{age}_{i,t} + \mu(\text{age}_{i,t})^2$, which is assumed to depend on a worker's fixed effect, θ_i, and her experience, which is reflected by her age and her age squared (note that μ is usually negative).

These last group of variables distinguishes estimations based on individual data from those using aggregate data. In particular, a specification that uses aggregate data explains the *average* wage $w_{rs,t}$ and includes as a covariate the *average* workers' average skills $Q_{rs,t}$ in sector s and region r at date t:

$$\ln w_{rs,t} = Q_{rs,t}\theta + X_{r,t}\beta + Z_{rs,t}\phi + \gamma_s + \delta_t + \epsilon \quad (11.14)$$

Typically, $Q_{rs,t}$ is assumed to depend on the average literacy, education, or skill levels of the employees in the local industry. This is to be compared with (11.13), in which θ_i is the efficiency pertaining to each worker, estimated as an individual-specific fixed effect. In other words, it need not depend only on the worker's education or skill level, as is the case when using aggregate data. It encompasses any effect specific to the worker that does not vary over time *whether it is observable* (i.e., available in the data set) *or not*. Access to data spanning several years allows for the introduction of such a fixed effect, the estimation of which is based on variations in a worker's wages over time and possibly across locations if she moves. This fixed effect does not, however, take into account time variations in an individual's skills. Thus, to complete the model, we add the worker's age and its square, the aim of which is to account for the large fraction of time fluctuations in an individual's skills (as shown by labor economists).

Ultimately, this type of estimation is much more general than models based on aggregate data on the following grounds:

(i) it exploits more information (e.g., using individual wages instead of average wages, and individual skills instead of average skills) and

(ii) the skill variables included in the model are no longer constrained to being proportional to other available explanatory variables.

Again, we use the density to illustrate the bias resulting from omitted variables and endogeneity. The most comprehensive estimation uses individual-specific data that include variables controlling for natural amenities, local public goods, and the market potential of neighboring areas; all variables that capture interindustry externalities are instrumented (as discussed above).[10] The elasticity of wage with respect to density is found to be 0.03, which means that *doubling the density of employment increases productivity by* $(2^{0.03} - 1) \times 100\% = 2.1\%$. When the endogenous explanatory variables are not instrumented, the same regression leads to a higher estimate of 0.037. Thus, failing to control for endogeneity would amount to overestimating agglomeration economies by more than 20%, which is still reasonable when compared with the larger bias caused by omitted variables. Let us now turn to this problem.

When working with aggregate data, the estimation of (11.14) shows that the impact of density on wages is 0.056 under the instrumented specification and 0.063 otherwise. Moreover, surface area is estimated to have an impact of 0.034, while in the context of individual data its impact is not statistically significant. This suggests that working with aggregated instead of individual data can be a very significant source of bias. This is because such data fails to capture differences in labor skills across regions accurately. Average skill levels taken into account as controls capture imperfectly, at best, real differences in skills across individuals. Adopting a fixed effect for each worker, together with age and its square, changes the estimated density by a factor of two, while the impact of the surface area disappears. The underlying reason is that workers are sorted across space according to their overall skills. Even when workers have identical observable skills (e.g., their levels of education or their qualifications), *the most efficient workers in terms of nonobservable characteristics* (e.g., their motivation or other psychological and cultural characteristics) *are located in the densest areas*. Therefore, overlooking or failing to adequately control for this selection of nonobservable skills across space (i.e., a problem of omitted variables) can lead to very inaccurate evaluations of agglomeration economies. This might bias estimations even more than edogeneity. For a public decision maker, the fact that doubling the density of economic activity increases factor productivity by either 2.1% or 4.5% makes a big difference.

Building on Combes et al., Mion and Naticchioni (forthcoming) study the spatial variation of wages in Italy. Also using individual data, their results corroborate what we have just seen, namely that the elasticity of wages with respect to density is largely explained by differences in worker skill levels (66% of the total variance), and that taking endogeneity into account reduces this elasticity by nearly 50%. Mion and Naticchioni also observe that the presence of skilled workers in the most populated areas can only be partly attributed to migration. More precisely, everything works as if the place of birth were a spatial sorting device. The authors' hypothesis is that the interregional distribution of skills is linked to the size of the cities as producers of knowledge, as suggested by Glaeser and Mare (2001) in the context of the United States. In this case, *the spatial selection of skills could be considered a dynamic process in which the largest cities play a crucial role*, in that the accumulation of skills occurs more rapidly in these areas than elsewhere. However, more research is called for before any definitive conclusion can be drawn on that important issue.

To put the above estimations into perspective, note that Ciccone and Hall (1996) and Ciccone (2002) have studied the impact of density, the former for the United States and the latter for the large EU countries. Both papers use instrumented wage equations and find that density has an estimated elasticity of approximately 0.04–0.05; they show that these

[10]The instruments used are lagged variables taken at dates distant enough to ensure their exogeneity. Combes et al. (2008c) confirm those results by considering either wages or total factor productivity as the dependent variable, and geological features as instruments.

estimates are barely affected by endogeneity bias. At first glance, this result is at odds with those obtained from French data. However, having noticed that differences in labor productivity were only controlled at the aggregate level in these two papers, it remains to be seen whether these estimates would be robust to omitted nonobservable characteristics.

11.2.5 Regional Dynamics

There are related, and sometimes older, branches of literature that have attempted to apply the same type of ideas to the analysis of regional economic dynamics. The underlying idea is readily grasped: rather than having an immediate impact on productivity, agglomeration economies could have a dynamic impact, thereby exerting an influence on regional growth. In other words, if $X_{rs,t}$ encompasses all local externalities (in logarithms), and if the logarithm of the marginal productivity of labor (11.6) is expressed as[11]

$$g_{rs,t} \equiv \ln\left[s_{rs,t}\left(\frac{p_{rs,t}A_{rs,t}}{\gamma_{rs,t}^{1-\mu}}\right)^{1/\mu}\right],$$

it is also customary to estimate

$$g_{rs,t} - g_{rs,t-k} = X_{rs,t-k}\beta + \varepsilon_{rs,t},$$

where k is the lagged effect of externalities, measured in years, whereas the assumption made until now was

$$g_{rs,t} = X_{rs,t}\beta + \varepsilon_{rs,t}.$$

Glaeser et al. (1992) and Henderson et al. (1995) set the groundwork for an alternative specification that has been often used since then. The specification involves choosing a different dependent variable, i.e., replacing change in productivity by change in employment levels. While the choice of this alternative dependent variable is alluring because relevant data are often available on a very fine spatial scale, the drawback is that the resulting specification strays from its theoretical foundations, generating new issues in the interpretation of the estimations. For example, it is possible for the growth in productivity to lead to a drop in regional employment, which is at odds with the assumptions underlying this alternative specification (see Combes et al. (2004) for further details).

Another important issue in the literature is how fast externalities vanish across time. Finding a cogent answer to this question has clear and direct implications for the optimal timing of regional policies. Henderson

[11] As discussed above, one could use similarly the average labor productivity or the total factor productivity.

(2003) tackles this problem by considering lagged externalities, for each of a number K of years, as explanatory variables. The model estimated is given by

$$g_{rs,t} = \sum_{k=0}^{K} X_{rs,t-k}\beta_k + \varepsilon_{rs,t}.$$

Interpreting this specification warrants caution, however, as we run the risk of mixing the influence of lagged density values on local externalities with some simple possible inertia of productivity over time. The existence of such an inertia is plausible because it takes time to adjust production factors and/or to set up new plants. With this in mind, Henderson (1997) provides what seems to be the most appropriate model for testing the dynamics of local externalities:[12]

$$g_{rs,t} = \sum_{k=1}^{K} \alpha_k g_{rs,t-k} + \sum_{k=0}^{K} X_{rs,t-k}\beta_k + \varepsilon_{rs,t}.$$

This specification has the benefit of testing the persistence of externalities across time, while simultaneously controlling for the inertia effects of the dependent variable. Moreover, econometric techniques developed in the context of dynamic panels, such as generalized methods of moments, allow one to address endogeneity issues without finding specific instrumental variables. Indeed, it can be shown that sufficiently lagged values of the variables in level are valid instruments for the variables in first difference that are endogenous, and vice versa. In other words, the model is first rewritten as follows:

$$\begin{aligned} g_{rs,t} - g_{rs,t-1} = {} & \sum_{k=1}^{K} \alpha_k (g_{rs,t-k} - g_{rs,t-k-1}) \\ & + \sum_{k=0}^{K} (X_{rs,t-k} - X_{rs,t-k-1})\beta_k + \varepsilon_{rs,t} - \varepsilon_{rs,t-1}. \end{aligned}$$

This specification also allows one to take into account the impact of region and industry fixed effects. Moreover, lagged values of $g_{rs,t-1}$ and of $X_{rs,t}$ are used as instruments whose validity can be checked by means of overidentification tests.[13]

Somewhat unexpectedly, Combes et al. (2004) find that the adjustment process shows greater inertia in the United States than in France, despite the lower mobility of French workers. Static externalities are found to be predominant in France (lagged values stop being significant after one

[12] This approach has been revisited by Combes et al. (2004) to allow for the simultaneous estimation of the dynamics of employment and of the number of firms.

[13] Arellano (2003) gives a detailed account of these techniques.

year), which is starkly at odds with the six- or seven-year lags found in Henderson (1997). Combes et al. also suggest that the elements conducive to the growth of existing firms (the intensive margin) are not necessarily the same as those that foment the creation of new firms (the extensive margin). More precisely, it appears that a large number of different-sized plants positively influences the growth of existing plants, whereas more new plants tend to be created where there are a small number of plants having a similar size. Finally, a large regional labor market with a small number of similar-sized industries would favor the growth of both new and existing firms. Hence, contrary to general beliefs, *a strategy that aims to diversify the local industrial structure is not necessarily a good strategy for boosting regional development.*

11.3 Concluding Remarks

While they can be alluring, simple regressions that rely on industry-specific characteristics to account for differences in spatial concentration give rise to a great many econometric and analytical problems that can be resolved imperfectly at best. Due caution needs to be exercised when running these regressions, paying particular attention to the potential for omitted variables and endogeneity biases. Despite their tenuous link with economic geography models, the other approaches discussed in this chapter lead to suggestively stylized facts about the magnitude of agglomeration economies and the regional structure of industries. Here also, we have encountered a number of econometric issues that are generic in empirical economic geography, namely omitted variables problems related to the imperfectly measured characteristics of the areas as well as endogeneity biases due to workers' and firms' endogenous location choices. Having said that, even when we account for a large number of explanatory variables and econometric issues, *agglomeration economies remain important*, thus inviting us to continue the exploration of the mystery of economic agglomeration.

The approaches we have covered are said to be *nonstructural* in the sense that they are not directly derived from a specific model, and do not have the aim of estimating the parameters of such a model (note that this did not preclude us from framing these nonstructural approaches within a general theoretical context). In the final two chapters of this book, the benefits of applying structural models will be presented in greater detail. As a preview, one such benefit is that structural models are more capable of capturing various types of interactions across regions, a task that is

not often accomplished in the literature presented in this chapter. Common to all fields of economics, these two types of approaches (structural versus nonstructural) should be seen as complementary. The former is helpful in identifying robust correlations between variables that lie at the heart of economic geography, which involves a large number of variables. The latter intends to validate particular theoretical models with greater rigor, but this is often at the cost of a loss of generality.

11.4 Related Literature

The idea of agglomeration economies dates back at least to Weber (1909), while the potential role of industrial diversity in fostering local development was first discussed by Jacobs (1969). Intraindustry externalities are also called *localization economies* (Hoover 1936) or Marshall–Arrow–Romer (MAR) externalities. Interindustry externalities are called *urbanization economies* or Jacobs externalities. This cornucopia is a major source of confusion. It was not until the work by Glaeser et al. (1992) and Henderson et al. (1995), which both deal with employment growth in American cities, that a new strand of research has begun to estimate more precisely the magnitude of agglomeration economies. It took several more years for the many difficulties associated with such estimations to be fully understood. A fairly comprehensive review of the literature is provided by Rosenthal and Strange (2004). The reader will find in Cingano and Schivardi (2004) an analysis of regional productivity growth in Italy. Focusing on Chinese cities, Au and Henderson (2006) consider the impact of city size on wages. The existence of a bell-shaped relationship is confirmed, with the striking result that a large number of Chinese cities are undersized, as all agglomeration economies are not being fully exploited.