Location of manufacturing FDI in Hungary: How important are inter-company relationships?

Gábor Békés*
Central European University

February 8, 2005

Abstract

Contributing to the new economic geography literature, this study focuses on location decisions of foreign investors within one country. The paper sets up a simple model with monopolistic competition with a direct application of input-output linkages, and estimates the determinants of location choice using econometric models with discrete dependent variables. In addition to classic attraction variables such as low wages, developed infrastructure and proximity to international borders, effects of access to final consumers and corporate partners are investigated. Various specifications are tested both in discrete choice and in count data modelling frameworks to see robustness of results. A tax report based dataset of Hungarian firms between 1992 and 2002 is employed, and determinants of entry by all the new-born companies in manufacturing with a foreign ownership are analysed.

JEL classification: F23, R3, R12, C35

Keywords: economic geography, industrial location, FDI, regional policy, discrete choice models

1 Introduction

Location decisions of foreign firms have been and will be of crucial interest in Central and Eastern Europe. Over the past fifteen years, foreign direct investment, (FDI) has been a key catalyst of economic transformation, and economic policy focused on attracting firms to locate anywhere in the country. As a side effect to development, spatial inequality has widened. Now, having joined the European Union, the key policy consideration will be the attraction of foreign investments in less developed regions. This new emphasis will render research on location decisions a special importance.

Location of industry and more specifically that of the foreign-owned manufacturing has long been in the limelight of economic research. The interest

*Budapest, Nádor utca 7., Hungary. E-mail: cphbeg01@phd.ceu.hu

in intra-national location choice has been inspired by facts of concentration of activity and a marked difference between developed and underdeveloped regions even within one country.

There are many reasons for the concentration of production such as the attractiveness of servicing a large market, the proximity to suppliers of intermediate goods or various forms of technological spill-overs. Of course, agglomeration forces do not prevail without boundaries, there are dispersion forces in action, too. First and foremost, high wages will make certain wage-sensitive industries incapable to offset rising costs. These companies will at some point opt to locate in the other region. Although they will face much higher transaction costs when selling to the larger (and richer) region, production costs will be much lower in the other region.

Most models of new economic geography (or NEG) aim at uncovering the essential reasons behind both agglomeration and dispersion of economic activity[1].

In Central and Eastern Europe (CEE) rapid changes and restructuring in manufacturing have taken place since 1990. Thus, countries like Hungary offer a laboratory experiment to study the geographic properties of a large number of new investments by firms entering a region previously closed to foreigners. As for the development, foreign direct investment was instrumental in transforming the industry of transition economies. In Hungary, for example, the stock of FDI reached 40% of GDP by 2003.

Spatial inequality is a well-known feature in Hungary and it is visible in many aspects of economic life: GDP per capita, wages, number of cars or telephone lines. Strong differences may also be observed in terms of new foreign investments in manufacturing. Table 1. reports some key figures on the twenty counties. During the 1993-2002 period, Budapest attracted 1.18 new firms per 1000 inhabitants while the same measure is less then 0.2 for the worst performing two counties.

[Table 1 about here]

There was a marked increase in spatial inequality throughout the nineties. For example, ratio of the GDP per capita in the two richest counties to the one in the two least developed counties rose from 2.3 in 1993 to 2.9 in 2000. This inequality is also reflected in the number of new firms. Considering all (similar-sized) counties but the large capital region of Budapest, new investments were often concentrated in a few areas. During the period 1992-2002, the top 3 counties (out of 20) were consistently responsible for 35%-40% of new foreign firm establishments. Further, as reported in Table 1, richer counties (Vas, Győr-Sopron-Moson, or Fejér) managed to increase their share of FDI over time.

[1] An excellent survey of key hypotheses emerging from models of new economic geography and their mixed empirical support can be found in Head and Mayer (2004).

Agglomeration of investments and a spatial polarization have also been visible phenomena in many sectors. For example, manufacturing of electronic devices by firms such as Flextronics in Central and Eastern Europe can be found in a fairly narrow band from north Poland through the Czech Republic, West Slovakia, West and Central Hungary down to North Slovenia and Croatia.[2].

The paper focuses on business-to-business linkages in order to explain some of the agglomeration forces in work. Supplier contacts are known to have been an important factor in the region. In Hungary, suppliers to the Suzuki car plant are mostly settled in three counties, within a close proximity to Suzuki. Also, several recent investments in North-Hungary in the motor vehicle sector are linked to the deployment of Peugeut and car manufacturers in the South part of neighboring Slovakia.

With emergence of studies testing NEG model predictions or comparing some key models, the gap between theory and empirics have become narrower. In this spirit, this paper looks at foreign direct investments in Hungary and investigates why the presence of a firm has an impact on investment by another. The paper contributes to existing literature on corporate location choice in three ways. First, it relates a new economic geography model with input-output linkages to data. Second, rarely rich datasets of corporate tax returns and annual labor surveys are used to get location, sales and wage data. Third, investigation is carried out on a small-to-medium sized European economy that has just gone through economic transition involving almost unprecedented rapid market liberalisation. Fourth, various econometric specifications of both discrete choice and count data models will be applied to provide robustness of results that may be crucial when working with firm level data.

The paper is organised as follows. Section two summarizes the related literature analysing results of firm location in general and FDI location in particular. This is followed by a presentation of the theoretical background of location choice in section three. Datasets and variables are described in section four discussing advantages and pitfalls of micro datasets as well as explaining the creation of the access variables. Section five present the econometric methodologies along with all the results and robustness checks. Section six concludes.

2 Related literature

Over the past few decades there has been a renewed interest in location theory of firms - within NEG frameworks, and it emerged that "new firms have a high propensity to settle at places where economic activities are already established" as posited in Ottaviano et al. (2003, p.7.). New economic ge-

[2]For details see Barta (2003).

ography models have been employed to explain location of overall economic activity by Krugman (1991), industrial clusters in Fujita et al. (1999, Ch. 16.) or location of various manufacturing sectors by Midelfart-Knarvik et al. (2000).

Both theoretical and empirical work in this field are centered around two key determinants of location: agglomeration externalities and market access. These notions will play a central role in this study and hence, it is worth giving a brief account of the key ideas.

Agglomeration externalities were first emphasised by Marshall, and formalisation of most such externalities may be found in Fujita et al. (1999, Ch. 16.). In most of the early models, labour migration was essential for agglomeration forces to work: an increased population generated greater demand inviting more firms to settle in a larger city, and this allowed for a lower import bill and hence, lower living costs in general. However, labour migration is rather low in Europe, even in the long run. Thus, another agglomeration force was required to explain the desire to co-locate in spite of low migration propensities. This explains why the incorporation of inter-company sales or in other words, input-output (I-O) linkages were so important for empirical work. These linkages try to capture trading costs between firms explicitly thus, provide a motive for co-location.

Of course, there are other well know reasons for agglomeration, drivers of industrial clustering. One such reason that makes worth locating close to one another is the potential of knowledge spillover. This is true for human as well as physical capital. The attraction to work close to other people is noted in Marshall (1920) and the importance of face to face communication is discussed in Leamer and Stolper (2000). As for firms, proximity allows to exchange inventions while technology spillovers help increase productivity using other firms' knowledge. Another such agglomeration force is labour pooling: firms enjoy the presence of a larger set of labour where the specific knowledge required by the firm, may just be fished out easily (as in Amiti and Pissarides (2001)).

Barry et al. (2003) emphasised that it is also difficult to disentangle the agglomeration and the demonstration effect because of a reputation effect that maks it optimal to mimic each others' location decisions. In their empirical study, demonstration is considered to be a part of co-location externalities that is not explained by agglomeration variables such as R&D intensity (spillover), excess job turnover (labour market thickness). Of course, information sharing and demonstration effects are closely interrelated.

As for the access to markets, the key idea that firm location depends on the proximity of demand was introduced a long ago, and Harris (1954) devised the simplest aggregate market-potential function. Market potential has been first investigated in an international context; proximity to key markets and suppliers has been explicitly featured in empirical works explaining overall economic activity or per capita income. Redding and Venables (2004) argue that a country's wage level (proxied by per capita income) is dependent on its capacity to reach export markets and to manage to get hold of the necessary intermediate goods cheaply. Head and Mayer (2005) look at Japanese investments carried out in the European Union. Results show that apart from a very important market potential measure, a number of traditional explanatory variables (e.g. taxes) and agglomeration variables turn to be significant as well.

There have been several papers dealing with location decisions of foreign investors within one country. Crozet et al. (2004) study location of FDI in France using model of oligopolistic competition model to simulate corporate choice of location. They find that firms of the same nationality like to group together, locations close to home country are chosen more frequently, and some industries (like car plants) have a strong tendency to agglomerate. Similarly, a study by Head and Ries (2001) looks at Japanese investments in the US and finds that firms belonging to the same *keiritsu* tend to settle close to each other.

Some studies considered countries of similar size and population to Hungary. Barrios et al. (2003) look at multinationals' location choice in Ireland with special interest in the role of agglomeration forces as well as state support. They find that agglomeration forces contributed substantially to location choices but proximity to major ports and airports was also helpful. More importantly, they find evidence that higher public incentives in designated areas have increased the probability of multinational investment. Figueiredo et al. (2002*a*) took Portuguese data to look at, among other factors, the home field advantage.

Anecdotal evidence confirms that agglomeration forces are active in transition economies of Central and Eastern Europe. The presence of industrial clusters is an easy to spot feature of new manufacturing base in the region, including the motor vehicle cluster in North-West Hungary, West Slovakia or South-West of the Czech Republic. Also, there is some evidence showing that large multinationals lured in their usual suppliers[3]. Results with data on these economies have just started to emerge of late. Disdier and Mayer (2004) compare French investment in the Western and Eastern part of Europe. They find that location choice is positively influenced by local demand and proximity to France increases the probability of a given country being chosen. Cieslik (2003) uses a Poisson model on 50 Polish regions to find that proximity of key export targets, industry and service agglomeration, and road network are the key magnets for foreign investment.

As for Hungary, Fazekas (2003) looks at the concentration of FDI from a labour market point of view to study what impact capital inflow had on the regional structure of the country. The paper finds that concentration

[3]The latest example is Hyundai motors in Slovakia where eight other Korean firms announced to follow Hyundai.

pattern of foreign-owned enterprises is just marginally higher than that of the domestically owned ones. However, FEs are concentrated in a different pattern, being located closely to the Western border. The approach of this paper is somewhat different to Fazekas (2003) in that it investigates the agglomeration patterns of foreign firms only.

3 Theoretical framework

The theoretical framework aims to emphasise business to business relations as a key driver of location decisions. The main relationship between any two firms is a potential of supplier-buyer link, i.e. one firm's output is the intermediate good of another. Modelling and measuring this potential will be in the centre of this analysis. The model is using the "classic" ingredients of new economic geography based on the monopolistic competition of Dixit and Stiglitz (1977) and first presented by Krugman (1991). One key aspect of firm-to-firm relationship here is related to input-output linkages that were introduced by Krugman and Venables (1995) in order to model the fact that firms sell goods not only to consumers but other firms as well. This paper follows the concise display of multi-country and multi-industry model in Fujita et al. (1999, Chapter 15A).

There is monopolistic competition in all sectors producing a range of differentiated goods. The paper focuses on manufacturing and the agriculture sector, which has been present in many similar models, will be overlooked. True, a dispersion force will be lost but in the lack of large-scale migartion, wages and local consumer demand should be strong enough to foster agglomeration.

There are $r = 1...R$ regions, $j = 1...J$ industries, with n_r^j firms producing a variety each of industry j in region r.

Profit for each firm depends on firm- and industry-level characteristics. Firm-level characteristics such as technology advantage over industry peers and quality of management are unobserved. However, these features are assumed to be independent from the choice of location. Another determinant of a given firm's profit depends on such industry features as (average) technology, skill requirements, transaction costs and location of markets. These are indeed region-dependent factors. Thus, profit for firm i in industry j and region r will come from these two terms and we assume additive separability:

$$\pi_i(^j_r) = \pi_i^* + \Pi_r^j \tag{1}$$

Since the focus is on location choice, the assumption of additive separability allows for working with Π_r^j only. Assume now that fixed cost of starting a new business is the same in all regions, and the cost of capital is unchanged through space as well - this can be considered as one key difference between national and international models. Firms pay taxes and receive investment support. However, in Hungary, local economic policy is not defined by counties but determined at the settlement level, and regional tax incentives are relative novelty, so it was assumed that region specific state intervention is zero. The profit is simply:

$$\Pi_r^j = (p_r^j x_r^j - mc_r^j x_r^j) \tag{2}$$

The representative consumer draws utility from consuming a composite manufacturing good: $U = C_M^{\mu}$. However, the consumer enjoys several manufacturing goods, and the composite good consumed comes from a constant elasticity of substitution (CES) subfunction of the available varieties.

$$C_M = \left(\int c_i^{1-1/\sigma} di \right)^{1/(1-1/\sigma)} \tag{3}$$

The elasticity of substitution between goods is measured by σ^j. Theoretically it measures to what extent goods are close to each other, i.e. whether consumers are easily willing to replace one with another. If it is small, products differ, in case of $\sigma^j = \infty$, the products are homogenous, and the market structure is identical to perfect competition.

As it is the case in models following the Dixit-Stiglitz tradition, profit maximisation yields a price that equals marginal cost and a markup, Φ^j:

$$p_r^j = mc_r^j \Phi^j \tag{4}$$

In our case, the markup depends on the elasticity of substitution. Assuming that firms have the same size, and there are N_r^j firms in region r, it can be shown that the markup is:

$$\Phi_r^j = \frac{\sigma^j}{\sigma^j - 1 + (\sigma^j - 1)/N_r^j} \tag{5}$$

Indeed, if two products are close substitutes, the market power to set prices should be small, hence the low markup. It is assumed that there is a large enough number of firms, hence: $\Phi_r^j (= \Phi^j) \simeq \sigma^j/(\sigma^j - 1)$, i.e. the markup is not dependent on consumption. This assumption is crucial for it yields that mill-pricing is optimal.

Firms use a set of goods produced by firms in other industries that are aggregated by a CES subutility function into a composite good. The intermediate good price index, G_r^j denoting the minimum cost of purchasing a unit of this composite good, is a key variable in this setup for firms benefit from supplier proximity. If a greater quantity of necessary intermediate goods is produced locally, less transportation cost will have to be paid. Hence, production costs will be lower, too. This creates a forward linkage. Here, the intermediate price index is weighted average (with n_l^j being the number of relevant firms[4]) of f.o.b. prices $\left(p_l^j \tau_{l\_r}^j\right)$ that already include an iceberg type transport cost, $\tau_{l\_r}^j \geq 1$: (i.e. for the home region only $\tau_{l\_l}^j = 1$).

$$G_r^j = \left[\sum_{l=1}^{R} n_l^j \left(p_l^j \tau_{l\_r}^j\right)^{1-\sigma^j} \right]^{\frac{1}{1-\sigma^j}} \tag{6}$$

This way of incorporating the price index implies the love of variety effect. The intermediate price index for a firm in industry j of region r is G_r^j, where io_{ij} is the input-output coefficient, i.e the share of industry j in all output used by industry i. In a small country, industry buys goods and services from abroad and the import coefficient, io_{j*}, for each industry gives the share of a composite imported good (priced G_W^j). Since data come from a complete national input-output (I-O) table, $\sum_{i=1}^{J} io_{ij} + io_{j*} = 1$, $\forall j \in J$.

$$GP_r^j = \left[\prod_{i=1}^{J} \left(G_r^i\right)^{io_{ij}} \right] \left(G_W^i\right)^{io_{i*}} \tag{7}$$

Firms sell their product to consumers and firms who use other firms' output as their input. This latter gives rise to a system of input-ouput linkages - a key agglomeration force.

Now the marginal cost function of a representative firm in industry j and region r may be defined as follows:

$$mc_r^j = w_r^{a_j} (GP_r^j)^{\mu_j} (\mathbf{b}_r^j)^{\delta_j} \tag{8}$$

where w_r is the nominal wage, GP_r^j is the composite price index of intermediate goods and $\mathbf{b_r}$ is a vector of other location dependent non-wage factors of the locally consumed production such as communication infrastructure.

Let us define q_{rl}^j, as demand in a region l, for a unit of industry j output, produced in region r. Demand can be derived from the CES utility:

$$q_{rl}^j = (p_r^j)^{-\sigma^l} (\tau_{l\_r}^j)^{1-\sigma^l} E_l^j (G_l^j)^{\sigma^l - 1}$$

Expenditure on the $j - th$ industrial goods for a given region (E_l^j) comes from two sources: consumers (who spend a μ_l^j fraction of their income on l region, j industry goods) and other firms coming from all industries.

$$E_l^j = \mu_l^j INC_l + \sum_{i=1}^{J} io^{ji} X_r^i \tag{9}$$

[4]Later, the number of firms may be replaced with volume of output.

In equilibrium, X_r^j, the supply of an industry j in region r will be equal to demand from Hungary and the rest of the world.

$$X_r^j = \sum_{l=1}^{R} q_{rl}^j + QW_r^j \tag{10}$$

where QW represents foreign demand.

Unlike in Fujita et al. (1999), this paper does not intend to end up with a set of equations and simulate results. Instead of a general equilibrium approach we need to be "short sighted" and consider a partial equilibrium without dynamic effects of an investment. In the long run equilibrium, prices are adjusted taking externalities into account. For example, wages or land prices will reflect benefits of agglomeration and lower prices in one region will only signal poorer circumstances. In the short run, disequilibria may exist and entry of firms (bidding up wages and input prices) shall be considered as a force to bring prices closer to their equilibrium value. The main goal of this exercise is to obtain a corporate profit function that will be linked to the settlement decisions of firms in the empirical work.

So, the profit function can now be rewritten:

$$\Pi_r^j = mc_r^j(\Phi^j - 1)\left[\sum_{l=1}^{R}(mc_r^j\Phi^j)^{-\sigma_j}(\tau_{l\_r}^u)^{1-\sigma_j}E_l^j(G_l^j)^{\sigma_j-1} + (\tau_{x\_r}^u)^{-1}QW_r^j\right] \tag{11}$$

where $\Psi^j := (\Phi^j - 1)(\Phi^j)^{1-\sigma_j}$ is a monotonically decreasing function of the industry specific elasticity of substitution, σ_j. Note, that this measure is industry-dependent only, and hence will be empirically irrelevant .

Let us define the aggregate demand variable, AD_r^j as

$$AD_r^j := \left[\sum_{l=1}^{R}\left((\tau_{l\_r}^u)^{1-\sigma_j}\left(\mu_l^j INC_l + \sum_{j=1}^{j} io^{ji}X_l^j\right)(G_l^j)^{\sigma_j-1}\right) + (\tau_{x\_r}^u)^{-1}QW_r^j\right] \tag{12}$$

Note that the way demand is set up creates a backward linkage: firms want to be close to their markets and potential customers.

So, the profit function is:

$$\Pi_r^j = \left(w_r^{a_j}(GP_r^j)^{\mu_j}(\mathbf{b}_r^j)^{\delta_j}\right)^{1-\sigma_j}\Psi^j\left[AD_r^j\right] \tag{13}$$

The profit function captures both key notions formerly introduced. Access to markets is incorporated both for firms and for final consumers. Agglomeration economies will be captured by some the $\mathbf{b}_r^j$ variables as well as some of the access variables. The way demand is set up allows the introduction of some of the key business to business relationships.

4 Description of the data and variables

An important contribution of this paper is the application of dataset that includes all firms throughout their lifetime including the year of entry and exit. Thus, final corporate decisions may be looked at instead of announcement of investment projects that may or may not have been realised. Furthermore, instead of estimations and aggregations, this dataset allows for creation of output variables based on the actual firm-level sales data only.

4.1 Datasets

There are two key datasets in the study. The corporate dataset used here, is based on annual balance sheet data submitted to the Tax Authority (APEH). This version comes from the Magyar Nemzeti Bank. In addition to this, a large employer-employee dataset is used that comes from annual Labour Market Survey (LMS) data compiled by the Ministry of Labour, containing employment data on a sample of some 140.000 employees per year. Employees are picked independently of their employers and the large sample size allows for annual coverage of all industries in almost all counties.

The APEH dataset contains information on *all* registered, double entry book-keeping firms. Data include industry code, size of employment, share of foreign ownership and a county code. Data are available annualy for the 1992-2002 period. The number of corporations varies year to year, rising from 57,862 in 1992 to 184,703 in 2002. The dataset was improved by the Economics Department of the Magyar Nemzeti Bank as well as the CEU Labor Project. (For details on the data, see the Appendix.)

In its tax report, each company reports a sales figure that can be picked up from its balance sheet attached to the earnings report. Sales data for a firm i operating in industry j registered in region r at time t is denoted by: $x^j_{r(t)}(i)$.

The year of *firm birth* equals the year of first appearance in the dataset, i.e. the first year of submitting a report to the Tax Authority. For this is compulsory, there should be little error in measuring the entry date. *Foreign ownership* is defined whenever the foreign share in equity capital passes a 10% threshold. For foreign companies defined this way, the average foreign share is very high and results are quite robust to raising the threshold to 25%. Also, whenever foreign ownership is low at the beginning, in most cases it will rise substantially after the first few years.

Overall, the dataset is composed of 5350 location settlements by firms with foreign ownership, but only 4557 may be certainly considered as new investment rather than foreign acquisition, and this paper deals with new investments only. Industries are grouped in sectors according to two-digit NACE codes. With merging some industries (e.g. clothing and leather), and excluding food production, there remain 15 sectors; Table 2. reports the main characteristics. As for changes through time, the furniture (and other misc.) industry fared the best and textiles the worst.

[Table 2 about here]

There are some coefficients that are not estimated but taken from other sources: Input-output table comes from Hungarian Statistics Office's publication on 1998 data (KSH (2001)). This is the only I-O table available for the time period used. However, the assumption that input requirements per sector have not greatly changed in a decade seems acceptable. The data indeed show that production is specialized, about half the value of output comes from purchasing goods and services from other producers. Out of domestic input, some 40% comes from buying goods, 55% from market services (including construction) and 5% from non-market services.

4.2 Variables for estimation

The profit function for the econometric model is created as follows. Consider the profit function (13), where profit is determined by labour costs, aggregate demand, intermediate good prices, and other cost factors, and take logs to get a linear relationship.

$$\ln \Pi_r^j = a_j \ln(w_r) + \mu_j \ln(GP_r^j) + \delta_j(1-\sigma_j)\ln(\mathbf{b}_r^j) + \Psi^j + \left[AD_r^j\right] \qquad (14)$$

Aggregate demand will be measured by final good demand, access to foreign demand and corporate market access. Final good demand is proxied by purchasing power of consumers that is measured by the variable INC, which is decomposed into the number of inhabitants, Pop and income per capita, IPC.

The intermediate good price index (GP_r^j) cannot be measured directly, so it will have to be proxied by supplier access variables. Given the market structure, the intermediate price index will be negatively correlated with the supply of these goods.

Access to foreign markets influencing both demand and intermediate good prices, is measured by a single foreign access variable (FMA_r). This takes into account that export is a crucial determinant of the revenue of Hungarian firms and the average import share reached 34% for manufacturing. Of course, by the theory, direct market access to foreign firms and customers should be taken into account.[5] This paper proxies access to foreign markets

[5] Amiti and Javorcik (2003) face the same challenge for Chinese subsidiaries of multinational firms that typically produce a great deal of their output for foreign markets. In their paper, access to foreign markets is proxied by the tariff rate but European free trade in manufactured goods makes this unnecessary.

by taking into account the distance to the key export borders.[6]

The vector of cost factors ($\mathbf{b_{r(t)}}$) includes some basic features of development that are not industry specific. A more developed county should yield lower transaction costs and hence, marginal cost of production. We use three such measure and look for a positive relationship between development and location choice. *Road network (Road_size)* measures the size of national road network within the given county. There has been little change in the size of the network throughout the observed period, the total size rose by about 3%. *Telephone network (Tel_size)* shows the number of telephone lines within the given county. This measures the number landline stations per county. This is a frequently used variable to proxy development of the infrastructure and thus, non-transportation linked transaction costs. Note however, that as a result of widespread use of mobile phones, this measure may have turned to be poor proxy of late. *College/University students (Edu_size)* represents number of students enrolled in higher education at institutions within the given county. This is should proxy the abundance of management and R&D knowledge in the county.

In addition to measures of development, population density (*Density*) indicates the size of population divided by size of the area of the county will pick up an agglomeration externality: it may be cheaper to sell products when people are close to each other. However, a negative sign would suggest that this urbanization effect is outweighed by higher land prices.

As for the labour market, $wage_{r(t)}$ measures the local wage. Wage variables were calculated from the LMS data and reflect (gross) labour costs that should be expected by a firm looking to settle in the given county.

Finally, we need to introduce the time dimension that has been so far overlooked. Explanatory variables are lagged one year for two reasons. The economic rationale (see "time-to-build" models) is that firms may be assumed to spend a year between investment decision and actual functioning (that is picked up by the data). The econometric support stems from a requirement to try to avoid endogeneity, and lagging will free the model of simultaneity bias. We also need assume that firms at time t considering values of explanatory variables at time $t-1$, pick a county independently of each other. Agglomeration works as firms locate close to other firms that had settled previously, but there is *no* strategic interaction between firms settling at time t. This is a necessary assumption for using the logit model.

For parsimonious notation, let me introduce the variable $ACC^{j}_{(m)r}$ that includes all access variables. Note that since Ψ^{j} is not county dependent, it shall be dropped. As a result, our expected profit function for a firm i is:

[6] We used distnace to the borders: West/Austria: Hegyeshalom; South/Croatia: Letenye; North/Slovakia: Komárom, East/Ukraine: Záhony ,Airport: Ferihegy/Budapest.

$$\pi^j_{r(t)}(i) = \alpha_1 wage^j_{r(t-1)} + \alpha_2 INC_{r(t-1)} + \boldsymbol{\beta_1'}\mathbf{ACC}^j_{r(t-1)} + \boldsymbol{\gamma} b^j_{r(t-1)} + \zeta^j_{r(t)}(i) \tag{15}$$

where the error term, $\zeta^j_{r(t)}(i)$ includes all the non-observed variables. Table 3 reports basic data on all our variables.

[Table 3 about here]

4.3 From firm data to access variables

All access variables to be tested in forthcoming subsections are based on output figures per county and sector ($Y^j_{r(t)}$). These numbers are determined by aggregating sales figures from the balance sheet data for all the relevant firms i (in industry j and region r, time t): $Y^j_r = \sum_i x^j_{r(t)}(i)$.

Corporate access measures proximity to firms that may be relevant for a new company, and the access variable is sum of output by firms weighted by distance and share in inter-company trade. Two corporate access variables are used: $ACCloc^j_r$ for access to local (internal or within county) firms and another one, $ACCnat^j_r$, for non-local (external or outside the county) firms. The reason for such dichotomy comes from the suspicion that the effect of distance is not linear, and firms clustered in one city or in a few cities close to each other, enjoy special agglomeration effects.[7]

$$ACC^j_r = \lambda_1 ACCloc^j_r + \lambda_2 ACCnat^j_r = \sum_j^K co^{ij} \left(\lambda_1(Y^j_r) + \lambda_2(\sum_{l \neq r}^R \frac{Y^j_r}{\tau^j_{l\_r}}) \right) \tag{16}$$

where co^{ij} includes all the trade coefficients, thus reflecting on supply and purchase links as well. In addition to these two measures, there is a third access variable, $BAloc^j_r$ that picks up access to business services, as a special determinant of production costs.

Unit transport costs are estimated by assuming a very simple relationship:

$$\tau^j_{l\_p} = dist_{l\_p} * V^j \tag{17}$$

i.e. it depends on the distance and on the cost of transporting one dollar worth of good by one kilometer. All data refer to distance by car, thus the road network that is crucial for transportation of goods is indeed taken into account of.

In reality we know little about coefficients of the relationship above. Studies with international data make use of the availability of cross-regional

[7] In a somewhat similar setup, Amiti and Javorcik (2003) created such aggregate access variables.

(i.e. international) figures for trade. This allows explicitly to estimate transportation costs. Here, it is assumed that shipping a good to 200km costs twice the amount it does for 100km. Note, that this is higher than some estimates for international shipment costs (e.g. Hummels (2000)). However, our variable includes all costs related to doing business.

The value of a typical package of industrial output $V^j = (\$/kg)^j$ on 1km comes from the World Bank database on international freight costs. True, these figures are based on more developed market data, and aggregation will mask many features. However, it helps correct for the fact that it is cheaper to ship €100 worth of laptop PC than the same value of steel.

There are various ways to measure distance between counties ($dist_{l\_p}$), and here a simple method is chosen. First, using the KSH "T-STAR" database on settlements, the most important city per county is picked (i.e. with the largest number of manufacturing plants). Note that picking the key city was straightforward for in all but one case, the largest city was at least twice the size of the second. Second, distance between any two counties is defined by measuring the road distance between the representative cities. It is assumed that goods are transported by trucks only, and that vehicles move at the same speed and costs are indifferent to road quality.

5 Estimation methods and results

First, conditional logit (CL) models will be estimated to study the influence of input-output linkages, labour market conditions and market access on investment decisions in Hungary. A key achievement that allows for such a structure to be used here is the Random Utility Maximisation framework of McFadden (1974). In this framework, firms are assumed to make decisions maximising expected profits, but given the scarcity of information and errors made by analysts, the maximisation procedure per se is less than perfect. Thus, profit (or utility for consumers) is a random function of explanatory variables. [8]

5.1 Conditional logit model

The methodology widely applied in spatial probability choice modelling is the conditional logit model based on Carlton (1983). Decision probabilities are modelled in a partial equilibrium setting with agents pursuing profit maximization behavior. Thus, they maximise a profit function like (15) subject to uncertainty. Apart from the observed characteristics of firms, such as sector and location (entering the profit equation), unobserved locational characteristics, measurement errors or improper maximization will determine actual profits. Note, that we do not observe either derived or actual

[8] For details, see Maddala (1983), Train (2003, Chapter 3.)

profits, but perceive locational decisions of firms. The explained variable is the location choice of firms so the choice variable is 1 if the investment took place in that particular county and 0 for the remaining 19 counties.

Taking all potential effects into account, a firm i (where $i \in \{1, ..., N\}$) of sector j (where $j \in \{1, ..., J\}$) locating in region r (where $r \in \{1, ..., R\}$) will attain a profit level dependent on various industry and region dependent variables. Importantly, not all of these variables matter, as the choice of region is independent on individual firm or industry characteristics. Thus, if agents maximise expected utility in this partial equilibrium setting, the number of firms in a region is related to the expected profit, as laid down in the profit function.

The profit equation (15) in parsimonious form for a firm i in industry j and region r is:

$$\pi^j_{r(t)}(i) = \gamma' b_{r(t)} + \lambda' d^j_{r(t)} + \varepsilon^j_{r(t)}(i) \tag{18}$$

In order to be able to use results of McFadden (1974), we need to assume that the error term, $\varepsilon^j_{r(t)}(i)$, is independently distributed across r and i, and has a type I extreme value (or Gumbel) distribution. The error term reflects unobserved terms as well those that depend on individual firms. A crucial assumption is that unobserved characteristics do not cause correlation, i.e. errors are independent of each other. In other words, independence here requires that the error for one alternative provides no additional information about the error for another one. It is likely that this assumption would not hold very well for the data but the generality of the CL model allows for a detailed investigation. (For details and some remedies, see section 5.4 .)

For every spatial option, the investor will compare expected profits and choose region r, provided that the following condition is fulfilled for $\forall l \neq r$:

$$prob[\pi_{ij}(r) < \pi_{ij}(l)] = prob[\varepsilon_{irj} < \varepsilon_{ilj} + A_r - A_l + \gamma' b_r + \lambda' d^j_r - \gamma' b_l - \lambda' d^j_l] \tag{19}$$

If this is the case, it can be posited that the investor's probability of selecting location r, provided she opted to invest in sector j is:

$$P^j_r = P_{r|j} = \frac{\exp(\gamma' b_r + \lambda' d^j_r)}{\sum_{l=1}^{R} \exp(\gamma' b_l + \lambda' d^j_l)} \tag{20}$$

Estimation is carried out by maximising the log-likelihood:

$$\log L = \sum_{j=1}^{J} \sum_{r=1}^{R} n^j_r \log P^j_r \tag{21}$$

where n^j_r denotes the number of investments carried out in sector j of region r.

5.2 Results with the conditional logit

5.2.1 Basic results

Basic results are reported in Table 4.(specifications CL1, CL2, CL3). To measure consumer demand, two variables were created: $INC_{r(t)} = IPC_{r(t)} * Pop_{r(t)}$, i.e. total income ($INC_{r(t)}$) is taken as income per capita ($IPC_{r(t)}$) multiplied by size of population ($Pop_{r(t)}$). For most specifications all these variables enter with the expected positive sign. Both access variables, described in (16) enter with the expected positive sign suggesting that proximity to both local businesses and firms nationwide induce more firms to settle. For access variables depend on the number and size of firms (+), transport costs(-) and road distance (-), affecting these would lead more entry by firms. Business access is also a significant determinant of firm location and so are low wages.

Note that coefficients are approximations of the elasticity of the probability of choosing a particular county for the average investor.[9] For example, considering the most basic setup of specifications, a 10% increase in the access variable (or 10% rise in the number of average firm) would raise the probability of choosing that county by 1%.

[Table 4 about here]

Specification CL2 includes non-industry dependent variables of $\mathbf{b_{r(t)}}$ such as the size of telephone or road network, both being positively related to firm location. This confirms generally held views that better infrastructure is key to attract FDI. The agglomeration variable of population density enters with a significantly positive coefficient, too. When adding average (trade weighted) distance from major markets (CL3 specification) some variables may loose significance. This stems from the fact that in Hungary many development related variables would increase as one travels from East to West, and hence, they are correlated with the distance from the Austrian border (that is by far the most important direction of trade).

5.2.2 Supplier and market access

However, since the input-output table is not symmetric (i.e. textile manufacturing uses a great deal of cotton, but cotton production uses little textile input), one should proxy G and the AD variable separately. Thus, instead of total access variables, four variables were created. Corporate demand may be proxied by a local and a national (all regions except for the local

[9] It can be shown that true coefficients are (1-p*) times the estimated figures, where p* is the average probability of choosing a region. Here, p*=1/20=0.05. Remember, that figures must be taken with care for the logit estimation is carried out with a normalization of the variance of the error term.

one) industry dependent market access variables (local: $MAloc_r^j$, national: $MAnat_r^j$).

$$MA_r^j = \lambda_1 MAloc_r^j + \lambda_2 MAnat_r^j = \sum_j^K co^{ij} \left(\lambda_1 (Y_r^j) + \lambda_2 (\sum_{l \neq r}^R \frac{Y_r^j}{\tau_{l\_r}^j}) \right) \quad (22)$$

In a similar spirit, the intermediate good price index is proxied by four supplier access variables:

$$SA_r^j = \vartheta_1 SAloc_r^j + \vartheta_2 SAnat_r^j = \sum_i^K co^{ji} \left(\vartheta_1 (X_r^j) + \vartheta_2 (\sum_{l \neq r}^R \frac{X_r^j}{1 + \tau_{l\_r}^j}) \right) \quad (23)$$

Note, that although supplier and market access variables are compiled in a similar fashion, they measure different types of variables. The market access is about demand, while the supplier access is just a proxy to (intermediate goods) prices.

According to specifications (CL4 and CL5 in Table 4), local supplier access and national (non-local) market access have a positive impact on location. This is true regardless of inclusion of the two other access variables: local market access and national (non-local) supplier access. These two would always enter with a negative sign, although frequently being insignificant. As one may have suspected, there is indeed a strong correlation between $SAloc$ and $MAloc$, and $SAnat$ and $MAnat$.

As for other access variables, access to business services ($BAloc$) does seem to induce more firms to enter. The border measures are overwhelmingly significant with the expected negative sign in any specification. For a small and open economy this is not surprising. Most governments emphasise the construction of major East-West or North South corridors and the importance of this notion is confirmed by the strong significance of our road distance to borders parameter.

However, positive coefficient of the road network variable suggest that building roads within a county will foster FDI inflow and industrial clustering as well.

5.2.3 Supplier, market and own industry access

One possible reason for correlation between access variables is the fact that own industry output influences both the supplier and the market access variable strongly. This stems from the structure of commerce between firms: companies trade the most with other companies in the very same industry.[10]

[10] This feature makes the use of models with two sectors, such as upstream and downstream industries, impossible.

On average, intra-industry trade amounts to one third of total inter-company sales, and this exacerbates correlation between the $MAloc$ and $SAloc$ variables.[11] To remedy this, a new variable, $IPloc$ is introduced that measures own industry output only. (This of course is also true for the non-local (national) variables.)

Results with these access variables are reported in Table 5. Some key results are encouraging. Most importantly, the access to own industry output ($IPloc$) is strongly significant through all specification and so is the national (external) market access variable ($MAnat$), or the local (internal) supplier access ($SAloc$). These suggest that input-output linkages are important determinants of location choice.

The result that the national market access is always significant and positive suggests that firms would want to settle closer to non-own industry suppliers. The local non-own industry supplier access variable is negative but highly non-significant. This suggests that supplier considerations may matter on the own-industry level, but not really on the non-own industry level. (Hence, a steelmaker would try to settle close to its potential buyers, but a carmaker disregards the location of steel-makers.)

[Table 5 about here]

However, some of the results that are robust through specifications remain puzzling: for example the national (external) supplier access ($SAnat$) is negative and almost always significant and so is the access to national own industry output ($IPnat$). True, the correlation between supplier and market access variables is strong and this may have remained a problem despite previous efforts[12]. Note that this is not a unique problem of this study, several previous empirical works with both supplier and market access variables faced this correlation issue (Redding and Venables (2004)).

As for other variables in this setup, consumer demand is a positive determinant, when proxied by two variables, both relative wealthiness and size of

[11] To see this, consider an economy consisting of three sectors only: $industry1$, $industry2$ and $industry3$. For example, supplier access for $industry2$ (shown by the superscript 2) is measured as a weighted average of industry output in the given country, with the weights being the coefficients of the input-output table. The input-output table has nine coefficients, with io_{12} being share of $industry1$ sales to $industry2$. Accordingly, (local) supplier access is constructed as follows: $SAloc^{(2)} = Y^{(1)} * io_{12} + Y^{(2)} * io_{22} + Y^{(3)} * io_{32}$, while market access for the same industry is constructed like: $MAloc^{2} = Y^{(1)} * io_{21} + Y^{(2)} * io_{22} + Y^{(3)} * io_{23}$. It is easy to see that $MAloc$ and $SAloc$ includes a common term: $Y^2 * io_{22}$, i.e. the output of own industry times the own industry coefficient. The new variable, $IPloc^{(2)}$ is simply $IPloc^{(2)} = Y^{(2)} * io_{22}$, and access variables can now be reconstructed as: $SAloc^{(2)} = Y^{(1)} * io_{12} + Y^{(3)} * io_{32}$ and $MAloc^{(2)} = Y^{(1)} * io_{21} + Y^{(3)} * io_{23}$.

[12] One potential reason for such result may be non-linearity in the data. To see this, I first looked at the access variables (in logs) and found that their distribution has a one-peaked distribution that looks not very different from a lognormal distribution. Second, quadratic terms were included to capture some sort of a hidden effect. It turned out that some quadratic variables were significant but they had no influence on any other variable.

county increases the likelihood of investment. As expected, a lower distance from the key external borders for exporters attracts FDI.

This structure allows to study agglomeration and industrial clustering more explicitly. Industries like to cluster for other reason than input-output linkages as it is supported by the strong significance of the industrial output variable of the actual sector ($IPloc$). One must remember that it is impossible to separate the key motives, such as labour pooling, knowledge spillover or a decrease in business costs due to information sharing. Despite our effort to filter out co-location due to supplier linkages, these problems remain important.

The simplest possible measure for agglomeration is population density. Its sign is not straightforward. On the one hand, a more dense area allows for lower transportation costs within the county, but on the other, it may lead to lower land prices and hence lower the cost of the investment. Results mostly support that a denser area is better for FDI.

Another measure of these forces may be the size of own industry output in the given region. However, our $IPloc$ variable does already capture intra-industry sales, both for inputs and outputs. Indeed, business to business relationships are important for same industry firms, but actual effects may hardly be disentangled.

As for access to national (external) output of the own industry ($IPnat$), the negative sign is likely to suggest some sort of a competition effect outweighing any agglomeration effect. Note, that although this is a feature that has been left out of the theoretical model with constant markups, other models of NEG (detailed in Baldwin et al. (2003)) would nevertheless have them. A negative sign may indicate such dispersion force: i.e. it is good to have similar firms close, but presence of too many firms in the neighborhood leads to market competition and under monopolistic competition, more varieties imply lower profits. Results suggest that within a county, agglomeration and input-output linkages are more important (as captured by a strongly positive $IPloc$ variable), while market crowding outweighs these positive externalities for other counties in proximity (negative $IPnat$).

5.2.4 The labour market

Previous empirical work suggests that one has to approach the impact of labour market on location choice with great care. The theoretical prediction of the wage coefficient is clear, wages are positively related to costs and hence a negative sign would suggest that high wages deter firm location. However, the empirical evidence is mixed with a slight leaning towards the opposite sign.[13]

[13] For example, in Figueiredo et al. (2002*b*) local wage has the expected sign, while in Holl (2004), the wage coefficient is insignificant. There are various explanations. For example Figueiredo et al. (2002*a*) argue that firms consider the wage level as a determinant

Results suggest that in basic specifications (such as CL3 or CL6), the plain wage variable ($wage_{r(t)}$) enters significantly with the expected negative sign confirming predictions of the theory. However, for several other cases, such as specifications CL7 in Table 5 or CLFE1 in Table 6, $wage_{r(t)}$ does indeed loose its significance.

There may be several explanations, this paper underlines just two such reasons. First, various industries use different types of labour in terms of skills, and hence, the industry mix of a region may or may not influence the aggregate wage variable. Second, individual industries use different type of labour in different shares. The share of blue-collar workers may vary a great deal among sectors and their wage may differ greatly depending on how skilled they are. Thus, in econometric models like those of this paper, wages may well reflect an "industry bias" as well as a "skill bias". An insignificant or a positive coefficient may just imply that investors are bringing in superior technology and hence, require more skilled and educated (i.e. more expensive) sort of labour reflected in higher wages.

Some further labour market variables were created in order to understand how important these issues are. Using the same annual labour survey, $wage\_ind^j_{r(t)}$ is generated by averaging wages of employees in a given county and industrial sector. For every employee there is a description of the job, put within one of the following categories: blue-collar, office and managerial. This allows to create three further variables for the average industry wage for these categories: $wage\_bc^j_{r(t)}$, $wage\_off^j_{r(t)}$, $wage\_man^j_{r(t)}$. Wages for blue-collar workers were available for almost all industries and regions. and hence, $wage\_bc^j_{r(t)}$ is first used together with $wage\_ind^j_{r(t)}$. Specifications CL8 or CLFE3 suggest that both variables enter significantly with $wage\_ind^j_{r(t)}$ suggesting the theoretical effect of wage costs and the positive sign of $wage\_bc^j_{r(t)}$ pointing to the skill bias important for blue-collar workers. For a somewhat smaller sample, other types of wages are shown to have not exhibited this same skill-bias (specification CL9).

5.3 Comparisons and robustness checks

The first issue of robustness is related to counties: would one area have a great influence on results? There are two suspects. First, the capital city Budapest has almost 18% of total population and a greater share of GDP. Second, Pest county, hosting satellite cities to Budapest, may have a special attraction feature. To check this, regressions were run leaving out one county at a time: results have hardly changed.[14]

So far we have pooled data for both years and industries. It is interesting to see to what extent coefficients change through periods in time and across

to locate in a cheaper country like Portugal (or even more so, Hungary) but within the country, wage has no effect.

[14] Results are available from the author on request.

groups of industries. [15]

To see how variables evolved through time, a basic regression was run for three periods: 1993-1995, 1996-1998 and 1999-2002. Many coefficients, including those related to the input-output linkages changed little through time. For example, access to own industry output remained a strong determinant throughout the period. However, some variables had different impact as time passed by: per capita income became less and less important an advantage, while high wages have become a more important deterrent by the latest period.

Regression results would be different if run sector by sector. Robustness of previous results were first checked by running regressions - leaving out one industry at a time. Results varied marginally only. Second, some industries were grouped into two categories: light industry (e.g. textile, clothing, etc.) and electronics/equipment (inc. electric machinery, audio-video manufacturing, etc.), and regressions were run for one group at a time. As expected, results vary substantially. Equipment and electronics industries prefer wealthier and more developed sites. Most interestingly, the national own industry output variable ($IPnat$) turned to be significantly positive for the second group, suggesting that competition was stronger for lower value added sectors.

5.4 Non-independent errors

The conditional logit modelling has some important limitations. An important restriction for CL models is

$$p_j(y_j)/p_h(y_h) = \exp((y_j - y_h)\beta) \tag{24}$$

so that "relative probabilities for any two alternatives depend only on the attributes of those two alternatives" (Wooldridge (2002, p. 501)). This is called the assumption of Independence of Irrelevant Alternatives (IIA). In our case, this posits that all locations are considered similar (having controlled for explanatory variables) by the decision making agent, yielding independent errors across individuals and choices. When IIA is assumed, an investor will look at all regions as equally potential places for investment. Thus complex choice scenarios cannot be included. Indeed unobserved site characteristics (such as actual geography) may well give way to correlation across choices.

To check whether the IIA assumption is strong enough, Hausman tests were run (Hausman and McFadden (1984)) for all NUTS2 regions. Results (reported in Table 6.) show that the IIA assumptions almost always fail

[15] Note that comparison within a logistic framework is not directly possible. In a logit regression, the variance of the error term cannot be estimated together with parameters and as thus, the variance term is normalized to one. As a result, a difference in values may only be due to a difference in the variance of the error term.

at the 1% level, suggesting that a more complex structure should be used. As is frequent for such exercise, asymptotic assumptions of the Hausman test fail for some occasions. To remedy this, the generalized Hausman test was applied and once again, IIA assumptions were rejected six times out of seven. There is no theoretical support for having seven regions, so an alternative structure with three larger regions (West, Central, East) was drawn, and the tests were run only to indicate that IIA fails universally for such tree-structure.

[Table 6 about here]

5.4.1 Conditional logit with fixed effects

One possible way to control for violations of the IIA assumption is to introduce dummy variables for each individual choice as suggested by Train (1986). Fixed effects are thus added to pick up possible level shifts caused by some omitted variables such as economic policy. County level as well as NUTS2 region level dummies are introduced to the key equations and results are reported in Table 7.As a result, (18) would become:

$$\pi^j_{r(t)}(i) = \delta_r + \gamma' b_{r(t)} + \lambda' d^j_{r(t)} + \varepsilon^j_{r(t)}(i) \tag{25}$$

where δ_r are location specific dummy variables.

First, the county fixed effects were introduced. Since many explanatory variables, which depend upon location only, change little over time, some will most likely loose significance in due course. As for access variables, sign and significance proved to be robust to these fixed effects. The wage coefficient remains negative but losses its significance. Fortunately, the industry specific wage variables turn to be significant. Second, fixed effects were introduced for the seven "NUTS2" regions. In general, results are similar to the previous ones. Here, the number of collage/university students becomes a factor of choice among counties with similar overall location.

[Table 7 about here]

With fixed effects most county feature variables indeed lost significance. It is interesting to note that, both the size of telephone network and distance from the key export borders remain a positive determinant of location choice even with county or region fixed effects introduced.

5.4.2 Nested logit

The nested logit model uses the same profit function as the conditional logit (15) works with a decision tree. The firm now first picks a region out of upper level alternatives u, and then chooses a county within the already selected region, out of lower level alternatives, r. Importantly, no assumption on a two-step decision making is necessary. It is enough to believe that certain counties are competing more closely than others.

Location probability in a county r, depends on probability of location in a region (u, upper level alternatives) times the probability of location in a county (m, lower level alternatives) in the given region:

$$\Pr_{ur} = \Pr_{r|u} * \Pr_{u} \tag{26}$$

$$\Pr_{r|u}^{NNNL} = \exp(\beta' Z_{ur}) / \sum_{n \in u} \exp(\beta' Z_{un}) \tag{27}$$

where Z explains the choice of an upper level (region) alternative in the conditional logit case $\beta' Z = \gamma' b_r + \lambda' d_r^j$.

$$\Pr u = \exp(\alpha' W_u + \xi_u IV_u) / \sum_{m} \exp(\alpha' W_m + \xi_m IV_m) \tag{28}$$

In this last equation, the inclusive value, $IV = \ln(\sum_{n \in u} \exp(\beta' Z_{un}))$, will tell us if the nest helps. From Maddala (1983), we know that $0 \leq \xi \leq 1$ and when $\forall\ \xi_m = 1$, the NL collapses to CL, while if $\forall\ \xi_m = 0$, the upper nest matters only i.e. firms choose a county randomly within the selected region.

It is important to stick to the RUM framework here as well, so a random utility maximization consistent nested logit had to be applied (Heiss (2002)) As a result, deterministic utilities must be scaled by the inverse of the IV_m parameters (ξ_m) in the conditional utility. This implies different scaling of the utilities across nests but allows the interpretation of $\beta' Z$ as RUM model.

$$\Pr_{r|u}^{RUMNL} = \exp(\beta' Z_{ur} / \xi_m) / \sum_{n \in u} \exp(\beta' Z_{un}) \tag{29}$$

There are two natural nests: the seven NUTS2 regions and three broad geographical areas: East, Central and West. Results, reported in Table 8, provide solid support for many of our previous results. According to specification NL1, the basic variables: per capita income, size, local and national corporate access, business services access and wages, enter all significantly and with the expected sign. With disaggregated variables (specifications NL2-NL5), own industry output remains one of the best performing variables along with national (external) market access. Better local (internal) supplier access remains a point of attraction, too. National (external) access to suppliers and the own industry remain to enter with the negative sign. Other explanatory variables loose or gain significance depending on the nest.

Specification test of the nested logit model is based on the values of the inclusive value parameters. The LR test of homoskedasticity (all values equal one) is clearly rejected for all specification. No single IV_m is ever close to the unity, suggesting that all parts of the nest is well warranted. However, greater than unity figures may indicate some specification problem. [16]

[Table 8 about here]

5.5 Count data approach

A great advantage of CL approach is its direct link with random utility maximisation. However, there may be several specification problems with the conditional logit model. The IIA assumption fails and the choice of a certain nested logit specification may seem somewhat arbitrary. Thus, one can apply count data models to see robustness of results. This comes with an additional advantage: the easy inclusion of time dummies. Indeed, during transition, there may have been important changes over time - such as shifts in policy - affecting regions differently.

5.5.1 Poisson models

In an effort to check robustness of CLM, count data models are used in this section, with the dependent variable representing the number or frequency of a particular event, in our case, the number of investments in a particular county. In these models, coefficients explain why $x\%$ more projects took place in county A relative to county B.

Define $n^j_{r(t)}$ as number of FDI investments in industry j, region r and time t. The explanatory variables are exactly the same as used in the previous sections.

$$Pr(Y^j_{r(t)} = n) = exp(-\lambda)\lambda n/n! \tag{30}$$

Importantly, Figueiredo et al. (2004) shows that the conditional logit equation as well as the Poisson model may stem from the same random utility maximisation model when firm-level characteristics are treated in a discrete fashion (such as operation in an industry). Alternative to the CL model, we can assume that $n^j_{r(t)}$ is the explained variable and $n^j_{r(t)}$ are independently Poisson distributed with

$$n^j_{r(t)} = \lambda^j_{r(t)} = \exp(\sum_j a^j d^j + \gamma' b_{r(t)} + \lambda' d^j_{r(t)}) \tag{31}$$

where d^j are dummy variables indicating if a firm is in industry j. [17]

[16] Note that greater than one inclusive values may suggest that the random utility model is inappropriate. Train (2003, Chapter 4.) discusses studies that prove that for several cases, RUM may well be consistent with IV values above one.

[17] Moreover, Figueiredo et al. (2004, p. 203.) shows that the Poisson concentrated log likelihood is "identical to the conditional logit likelihood with some constraints."

For every year, firm entry data were aggregated by industry and county, and Poisson regressions were run with the same set of explanatory variables used at logistic regressions. As expected, results were generally - but not always confirmed. Own industry output, once again proved to be one of the best performing variables with a coefficient close to 0.2, along with national (external) market access. However, supplier access variables swapped signs compared to logistic regressions. Other explanatory variables, such as distance from borders performed well, with even the number of college students making a difference. In a Poisson model context, the road network was unimportant though.

[Table 9 about here]

5.5.2 Negative binomial

The Poisson model has the advantage of being closely related to the conditional logit, but it assumes that the conditional variance of the dependent variable, λ equals the conditional mean of λ. However, equidispersion is rare property in reality, and for most cases, the variance is larger than the mean. Overdispersion may be treated, but in a more general, negative binomial model that allows to test the null hypothesis of equidispersion.[18] Given their easy applicability, no wonder that both the Poisson and the negative binomial model have been used in location research. (e.g. Basile (2004)).

The negative binomial distribution may be considered as a generalized Poisson, where the mean does not equal the variance. This deviation is represented with a dispersion parameter, α. The case with $\alpha = 0$ corresponds to equidispersion, and in that case the model collapses into a Poisson model.

Specification tests (LR test with one sided χ^2 statistics) suggested that the Poisson model is misspecified. However, results, reported as specification CNT5 and CNT6, suggest significance and in many cases even the magnitude of coefficients for the negative binomial are identical to those of the Poisson model despite the failure of the LR test. This robustness is not unusual in the literature, for example Smith and Florida (1994) finds a similar pattern for Poisson, negative binomial and even the tobit model.

6 Conclusions and future research

This paper focused on location decisions of foreign investors within one country, using econometric models with discrete dependent variables using a tax report based dataset of Hungarian firms. The rapid appearance of foreign-owned manufacturing sites offered a great opportunity: studying the geographic properties of a large number of new firms entering a region

[18] Importantly, the negative binomial model yields more efficient test statistics and prevents us from drawing overly optimistic conclusions (see Cameron and Trivedi (1998)).

previously closed to foreigners. Some conclusions may be drawn regarding theory and its empirical support as well as the application of methodologies.

First, a possible way was shown to link input-output linkage based NEG theory and a tax report based dataset - building on variables that had been generated out of firm level sales figures. Results that proved to be robust through discrete choice and count data specifications suggest that there is indeed an agglomeration effect for companies in play and input-output linkages work their way through supplier and market access providing a key reason for co-location. The importance of industrial clustering has been robustly shown and some support of agglomeration externalities was found as well. Access to markets throughout the country seemed to be a persistently important determinant of location choice. This provides some empirical support to NEG models with input output linkages.

In a quest to find out if misspecification of the econometric model is what causes unexpected results, several techniques, which have been previously used in some studies earlier, were applied. Specifications of conditional logit, nested logit, Poisson and negative binomial models were tested to compare results and investigate robustness. Although many specification test suggested that the actual model is misspecified, most coefficients kept their respective sign throughout specifications, and similar log likelihoods (or McFadden's pseudo R^2 measures, where available) suggested that most specifications are by and large equally supported by data.

Thus, we shall conclude that unexpected behaviour of some access variables should rather stem from the exclusion of important forces. Competition may be studied more directly, allowing access variables to pick up less of a market crowding effects. Furthermore results provide additional motives to study not only entry decisions of firms but to look at their behaviour once they have settled. For example, an analysis of the possibly divergent behavior of horizontal and vertical FDI may shed a light on unexpected signs.

Finally, some policy conclusion may be drawn - with caution. First, most of the industries do have a strong tendency to settle where other similar firms have already settled. Spending money on incentives to have them established elsewhere may be inefficient, and instead labour migration should be made easier, for example via development of temporary housing conditions. Further, subsidies to large firms may be efficient as long as they lure in similar firms. Second, input-output linkages are important. Thus, improving the relationship between suppliers and multinationals is key to fostering more investment. With a recent experience of loosing multinationals to non-EU Eastern Europe and China, this may be ever more important. Third, other explanatory variables that was found to be a significant are telephone and road network confirming widely held view on importance of local infrastructure.[19]

Acknowledgements

Part of this paper was written when I was visiting the Magyar Nemzeti Bank, I am grateful for their hospitality. I am also indebted to the CEU-Labor Project where major data cleaning was carried. For comments and suggestions I thank László Halpern, Gianmarco Ottaviano, István Kónya, Fabrice Defever and Álmos Telegdy as well as seminar participants at Central European University, Magyar Nemzeti Bank, Warsaw University, HWWA Hamburg and Sieps Stockholm .

References

Amiti, M. and Javorcik, B. S. (2003), Trade costs and location of foreign firms in china. World Bank.

Amiti, M. and Pissarides, C. A. (2001), Trade and industrial location with heterogeneous labor, Technical report, CEP/LSE.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G. and Robert-Nicoud, F. (2003), *Public Policy and Spatial Economics*, MIT Press.

Barrios, S., Strobl, E. and Gorg, H. (2003), Multinationals' location choice, agglomeration economies and public incentives, Discussion Paper 17, CORE.

Barry, F., Strobl, E. and Gorg, H. (2003), 'Foreign direct investment, agglomeration and demonstration effects: An empirical investigation', *Weltwirtschaftliches Archiv* (139), 583–600.

Barta, G. (2003), Developments in the geography of hungarian manufacturing. in Munkaerõpiaci Tükör.

Basile, R. (2004), 'The locationaacquisition versus greenfield investments: the location of foreign manufacturers in italy', *Regional Science and Urban Economics* **34**(1), 3–25.

Brown, J. D., Earle, J. S. and Telegdy, Á. (2004), Does privatization raise productivity?, Discussion paper 25, H.A.S Institure of Economics.

Cameron, C. A. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge University Press.

[19]However, one must bear in mind that several general equilibrium NEG models would show how construction of motorways may have an adverse effects in the long run. See Baldwin et al. (2003) for theory and Puga (2002) for some empirical support.

Carlton, D. W. (1983), 'The location and employment choices of new firms', *Review of Economics and Statistics* **65**(440-449).

Cieslik, A. (2003), Location determinants of multinational firms within poland. Warsaw University.

Crozet, M., Mayer, T. and Mucchielli, J.-L. (2004), 'How do firms agglomerate? a study of fdi in france', *Regional Science and Urban Economics* **34**(1), 27–54.

Disdier, A.-C. and Mayer, T. (2004), 'How different is eastern europe? structure and determinants of location choices by french firms in eastern and western europe', *Journal of Comparative Economics* **32**(2), 280–296.

Dixit, A. and Stiglitz, J. E. (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review* **67**, 297–308.

Fazekas, K. (2003), Effects of foreign direct investment on the performance of local labor markets - the case of hungary. Budapest Working Papers on the Labour Market, No 03/03.

Figueiredo, O., Guimaraes, P. and Woodward, D. (2002*a*), 'Home-field advantage: location decisions of portuguese entrepreneurs', *Journal of Urban Economics* **52**(2), 341–361.

Figueiredo, O., Guimaraes, P. and Woodward, D. (2002*b*), Modeling industrial location decision in us counties. University of South California.

Figueiredo, O., Guimaraes, P. and Woodward, D. (2004), 'A tractable approach to the firm location decision problem', *Review of Economic and Statistics* .

Fujita, M., Krugman, P. and Venables, A. J. (1999), *The Spatial Economy: Cities, Regions and International Trade*, MIT Press, Cambridge.

Harris, C. (1954), 'The market as a factor in the localization of industry in the united states', *Annals of the Association of American Geographers* **64**(315-348.).

Hausman, J. and McFadden, D. (1984), 'A specification test for the multinomial logit model', *Econometrica* **52**, 1219–1240.

Head, K. and Mayer, T. (2004), The empirics of agglomeration and trade, *in* V. Henderson and J. F. Thisse, eds, 'Handbook of Regional and Urban Economics, Vol 4.', number 15, North-Holland, Amsterdam, dp 59.

Head, K. and Mayer, T. (2005), 'Market potential and the location of japanese firms in the european union', *Review of Economics and Statistics* .

Head, K. and Ries, J. (2001), 'Overseas investments and firms exports', *Review of International Economics* **9**(1), 108–122.

Heiss, F. (2002), 'Structural choice analysis with nested logit', *Stata Journal* **2**(3), 227Ű252.

Holl, A. (2004), 'Manufacturing location and impacts of road transport infrastructure:empirical evidence from spain', *Regional Science and Urban Economics* **34**(3), 341–363.

Hummels, D. (2000), Toward a geography of trade costs. Univesrity of Chicago.

Krugman, P. R. (1991), 'Increasing returns and economic geography', *Journal of Political Economy* **99**, 483–499.

Krugman, P. and Venables, A. J. (1995), 'Globalization and the inequality of nations', *Quarterly Journal of Economics* (110), 857–880.

KSH (2001), Ágazati kapcsolatok mérlege 1998 (input-output tables), Technical report, Központi Statisztikai Hivatal.

Leamer, E. and Stolper, M. (2000), The economic geography of the internet age. UCLA.

Maddala, G. S. (1983), *Limited Dependent and Qualitiative Variables in Econometrics*, Cambridge University Press.

Marshall, A. (1920), *Principles of Economics*, Macmillan Press.

McFadden, D. (1974), *Conditional Logit Analysis of Qualititative Choice Behaviour*, Academic Press, New York, NY.

Midelfart-Knarvik, K. H., Overman, H. G. and Venables, A. J. (2000), Comparative advantage and economic geography: estimating the determinants of industrial location in the eu. LSE.

Ottaviano, G., Tabuchi, T. and Thisse, J.-F. (2003), 'Agglomeration and trade revisited', *International Economic Review* .

Puga, D. (2002), 'European regional policies in light of recent location theories', *Journal of Economic Geography* **2**(4), 372–406.

Redding, S. and Venables, A. J. (2004), 'Economic geography and international inequality', *Journal of International Economics* **62**(1), 53–82.

Smith, D. F. and Florida, R. (1994), 'Agglomeration and industrial location: An econometric analysis of japanese-affiliated manufacturing establishments in automotive-related industries', *jue* **36**, 23–41.

Telegdy, Á. (2004), Apeh dataset cleaning steps, Labor project, Central European University.

Train, K. (1986), *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*, MIT Press, Cambridge, MA.

Train, K. (2003), *Discrete Choice methods with Simulation*, Cambridge University Press, Cambridge.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel data*, MIT Press.

7 Appendix

7.1 Firms versus plants

A key issue is the exact nature of firm location. In effect, plant level data would be necessary to representing the actual production site. However, only firm level data are available instead. As a result, we may have data on a firm headquarter, rather than its production plant distorting our results a great deal. To check this, two exercises were carried out .

First, the National Corporate Register was consulted to see how large foreign manufacturers such as Siemens, Philips or IBM were incorporated in Hungary. Apparently, these multinationals established separate entities for many of their operations. Siemens AG, a German electronics good manufacturer established a dozen firms up to 2003 including Siemens kft, responsible for all retail activities, Siemens "Finance" (Financial Services), or Siemens "Telefongyár" (Telecom). IBM has its main production plant as part of IBM "Data Storage Systems" in Székesfehérvár (Fejér county), while consulting business is carried out via IBM "Üzleti Tanácsadó" registered in Budapest downtown.

The best example for separation of plants by industries may be the Dutch giant, Philips. It has invested in various firms including Philips "Components" (machinery) in Győr (Győr-Sopron-Moson county), Philips "Industries Hungary" (electronics) in Székesfehérvár (Fejér county), Philips "Monitor Industries" in Szombathely (Vas county) and Philips "Hungary Sales" in Budapest. A similar structure may be perceived by many other major multinational companies including Audi producer "Porsche Inter Auto", or Electrolux, whose production plant is situated somewhere in the countryside as one firm, while another one in Budapest is responsible for sales or foreign trade.

One should expect that the most problematic bias would come from an over-representation of the capital city given that many firms that entered Hungary, first established a HQ in Budapest. Thus, in a second effort, industry-level aggregates from two sources were compared: The APEH complete firm level corporate dataset and plant level employer data of the Labour Market Surveys. It showed that the share of Budapest by industries is just a few percentage points higher in the firm level data. This also supports the assumption that the application of firm level data should be of no great concern in our practice.

7.2 Corrections to the data

There has been serious effort invested in cleaning the data and several corrections were made to the original APEH dataset by the Magyar Nemzeti Bank, the CEU Labor Project[20] and the author. There has been three important steps. First, Longitudinal links for foreign firms were improved using data provided by Hungarian statistics office KSH on corporate entry and exit. CEU Labor Project looked for other longitudinal links in which the firms did not simply appear under a new id number, but actually split up into several firms or were formed via a merger. These allowed to keep track most but not all of firms under transformation. Second, The ownership structure of new firms was repaired in many cases to make sure that foreign ownership reflected the most likely case. Information from balance sheets and adjacent years' values were used.

Third, sales data for all firms were checked to avoid typing errors. For many firms, sales data were missing. Further problems I found and/or learned from others working with the same or similar datasets included: (1) 0 is imputed instead of actual figures for sales, (2) thousands written instead of millions, (3) one digit is left out making sales figure be 1/10 of actual data, (4) sales and export sales figures swapped. Overall, I made modifications reaching almost 2% of the total dataset. In some cases, sales could be estimated by using other balance sheet figures, and in others, the simple average of sales data at $(t-1)$ and $(t+1)$ was used.

As a final note, remember that for some discrete choice datasets, one has to worry about classification error i.e. measurement error in the left hand side variable. Having only a list of firm location decisions, the actual place may be mistyped or simply poorly gathered. This is not the case with the APEH dataset, since tax reports are submitted by the company to the regional Tax Authority office, and there are one per office per county (except for Budapest where there are three.). As a result, there should be very little error in the choice of location variable.

[20] For a basic description, see Brown et al. (2004) and for details see Telegdy (2004)

Table 1 New foreign manufacturing firms per Hungarian counties (1993-2002)

Counties	Number of new firms*	Inhabitants ('000)	New firms/ '000 poeple	New firms	New firms	Relative change in time
		1993-2002		1993-1997	1998-2002	(1998-2002) v. (1993-1997)
Szabolcs-Szatmár-Bereg	102	583	0.17	48	54	-53.8
Borsod-Abaúj-Zemplén	139	738	0.19	73	66	-32.1
Békés	79	393	0.20	45	34	-10.4
Hajdú-Bihar	118	550	0.21	62	56	-32.0
Jász-Nagykun-Szolnok	98	413	0.24	62	36	29.5
Somogy	104	334	0.31	73	31	92.7
Tolna	87	247	0.35	50	37	-7.6
Nógrád	80	218	0.37	43	37	-26.5
Heves	124	324	0.38	72	52	-4.3
Fejér	168	428	0.39	111	57	52.0
Csongrád	184	426	0.43	107	77	-3.8
Bács-Kiskun	236	542	0.44	146	90	19.5
Pest	504	1123	0.45	278	226	-19.7
Veszprém	183	368	0.50	112	71	15.0
Zala	164	297	0.55	95	69	-5.1
Baranya	235	402	0.58	147	88	24.3
Komárom-Esztergom	197	316	0.62	123	74	23.5
Vas	189	267	0.71	107	82	-12.3
Győr-Moson-Sopron	346	440	0.79	213	133	17.4
Budapest city	2013	1708	1.18	1167	846	-4.8

Source: KSH, APEH Corporate dataset, author's calculations

Table 2 New foreign manufacturing firms by industries

(NACE code) Industries	All FDI	Greenfield	World Bank unit price in USD/KG*
(17) Textile	327	280	11,6
(18 & 19) Cloths, leather	452	397	31,5
(20 & 21) Paper and wood products	475	414	5,8
(22) Press	648	510	22
(23 & 24) Refinery and chemicals	208	156	18
(25) Plastic rubber	383	319	12
(26) Other non-metalic	283	229	8
(27) Metal -basic	68	54	6
(28 Metal -fabricated	725	602	31
(29 Machinery	632	525	27
(30 Office equipment	57	51	140
(31 Electric machines	208	179	45
(32 & 33) Audio-video, PC, etc. instruments	429	347	140
(34 & 35) Motor vehicles	153	133	41
(36) Furniture, etc.	302	261	10
Total manufacturing (ex-food)	5350	4457	-

*Source: World Bank, APEH Corporate dataset, author's calculations. *Original data in ISIC, unit prices were transformed to NACE categories and aggregated by the author*

Table 3 Summary statistics

Variable	Description	Source	Mean	Std. Dev.
IPC	income per capita	KSH	87.234	27.646
Pop	population size	KSH	505.85	339.57
IPloc	own industry local output	APEH, „AKM" of KSH	649437	2637335
Ipnat	own industry national access	APEH, „AKM" of KSH	231341	501525
SAloc	local supplier access	APEH, „AKM" of KSH	1050664	2810798
MAloc	local market access	APEH, „AKM" of KSH	1879780	5910528
SAnat	national supplier access	APEH, „AKM" of KSH	354441	574277
MAnat	national market access	APEH, „AKM" of KSH	621667	1041976
BAloc	local business access	APEH, „AKM" of KSH	72692050	30586220
Tel_size	Size of telepohone network	KSH	123244	158637
Density	population density: inhabitants/area	KSH	.24328	.6913
Road_size	Size of highway network	KSH	1526.8	563.77
Edu_size	number of college students	KSH	9803	7332.54
dSouth	Distance of Southern export border	HAS-Institute of Economics	254120	117478
dWest	Distance of Western export border	HAS-Institute of Economics	233186	100015
dAirport	Distance of Airport	HAS-Institute of Economics	136976	68498
Wage	local wage	Minsitry of Labor „LMS"	31204.02	14371.81
Wage_ind	local, own industry wage	Minsitry of Labor „LMS"	30362.51	16232.25
Wage_bc	local blue-collar wage	Minsitry of Labor „LMS"	25585.4	12834.84
Wage_off	local office wage	Minsitry of Labor „LMS"	38096.37	22391.55
Wage_man	local manager wage	Minsitry of Labor „LMS"	80120.27	66476.41
D_{lr}	Road distance between cities	HAS-Institute of Economics	190.54	103.01

KSH: Hungarian Central Statistics Office, „AKM": Input-output tables, „LMS": Annual Labour Market Survey, APEH: Hungarian Tax Authority's corporate database. NB All variables in estimations are taken in logs.

Table 4 Location choice with conditional logit

Specification	CL(1)	CL(2)	CL(3)	CL(4)	CL(5)
Fixed effects	No	No	No	No	No
Ln (income per capita)	1.02***	1.10***	0.57***	0.78***	0.12
	(0.16)	(0.16)	(0.19)	(0.17)	(0.20)
Ln (population size)	0.16*	-0.13	0.38	0.24	0.42***
	(0.09)	(0.11)	(0.32)	(0.31)	(0.13)
Ln (local corporate access)	0.14***	0.11***	0.08*		
	(0.02)	(0.03)	(0.03)		
Ln (local supplier access)				0.40***	0.38***
				(0.04)	(0.04)
Ln (local market access)				-0.17***	-0.22***
				(0.04)	(0.04)
Ln (national corporate access)	0.31***	0.31***	0.13**		
	(0.03)	(0.03)	(0.06)		
Ln (national supplier access)				-0.52***	-0.56***
				(0.12)	(0.12)
Ln (national market access)				0.66***	0.53***
				(0.11)	(0.11)
ln (own region business serv. access)	0.34***	0.11	0.09	0.08	0.07
	(0.05)	(0.07)	(0.07)	(0.07)	(0.07)
ln (own region wage)	-1.53***	-1.42***	-1.19***	-1.04	-0.77*
	(0.39)	(0.43)	(0.44)	(0.43)	(0.44)
Ln (students enrolled in higher education)			0.02		
			(0.28)		
Ln (Size of highway network)		0.14***	0.008	0.14***	0.03
		(0.05)	(0.05)	(0.05)	(0.05)
Ln (Size of telepohone network)		0.40***	0.30***	0.34***	0.23***
		(0.08)	(0.08)	(0.08)	(0.08)
Ln (population density: inhabitants/area)		0.26***		0.30***	0.16**
		(0.06)		(0.06)	(0.06)
Ln (Weighted distance of export borders)			-0.40***		-0.47***
			(0.07)		(0.07)
LR chi^2	4887	4933	4959	5089	5124
Log likelihood	-10907	-10885	-10872	-10807	-10789
McFadden's pseudo R^2	0.1830	0.1847	0.1857	0.1906	0.1919
Number of observations	4557	4557	4557	4557	4557

Standard errors in parentheses. Significance at 1%, 5% and 10% is denoted by ***, **, and *, respectively.

Table 5 Location choice with all access variables with conditional logit

Specification	CL(6)	CL(7)	CL(8)	CL(9)
Fixed effects	NO	NO	No	No
Ln (income per capita)	0.91***	0.03	0.26	0.13
	(0.17)	(0.22)	(0.18)	(0.21)
Ln (population size)	0.20**	-0.19	0.51***	0.07
	(0.09)	(0.34)	(0.14)	(0.39)
Ln(own industry local output)	0.22***	0.20***	0.21***	0.23***
	(0.01)	(0.01)	(0.02)	(0.02)
Ln (own industry national access	-0.08	-0.19***	-0.12*	-0.03
	(0.06)	(0.06)	(0.06)	(0.07)
Ln(local supplier access)	0.10**	0.03	0.03	0.03
	(0.04)	(0.04)	(0.04)	(0.05)
Ln(local market access)	-0.17***	-0.17***	-0.43***	-0.23***
	(0.03)	(0.03)	(0.13)	(0.04)
Ln (national supplier access	-0.35***	-0.88***	-0.23***	-0.41***
	(0.12)	(0.15)	(0.03)	(0.14)
Ln (national market access)	0.64***	0.41***	0.50***	0.47***
	(0.01)	(0.12)	(0.11)	(0.12)
Ln (local business access)	0.33***	-0.006	0.04	0.01
	(0.05)	(0.07)	(0.07)	(0.08)
Ln (local wage)	-0.82**	-0.07		
	(0.39)	(0.45)		
Ln (local, own industry wage)			-0.46**	
			(0.19)	
Ln (local blue-collar wage)			0.51***	0.33**
			(0.19)	(0.16)
Ln (local office wage)				-0.06*
				(0.03)
Ln (local manager wage)				-0.11
				(0.13)
Ln (number of college students)		0.51		0.50
		(0.28)		(0.35)
Ln (Size of highway network)		0.25**	0.05	0.06
		(0.11)	(0.05)	(0.06)
Ln (population density: inhabitants/area)		0.09	0.16**	0.13**
		(0.07)	(0.06)	(0.07)
Ln (Size of telephone network)		0.18**	0.26***	0.25***
		(0.09)	(0.08)	(0.09)
Ln Weighted distance of export borders)		-0.62***	-0.52***	-0.55***
		(0.13)	(0.07)	(0.09)
Ln Distance of Airport		-0.62***		
		(0.13)		
Ln Distance of Western export border		-0.39***		
		(0.06)		
Ln Distance of Southern export border		-0.22***		
		(0.05)		
LR chi^2	5184	5331	5159	4567
Log likelihood	-10759	-10686	-10462	-7880
McFadden's pseudo R^2	0.1941	0.1997	0.1978	0.2247
Number of observations	4557	4557	4557	2860

Standard errors in parentheses. Significance at 1%, 5% and 10% is denoted by ***, **, and *, respectively.
*FORAQ: Denotes an increase in foreign ownership. This may be due to privatization or foreign acquisition of a private domestically owned firm.

Table 6 Hausman tests of IIA

	Hausman test $\chi 2$ test (p-value)	Generalised Hausman test $\chi 2$ test (p-value)
7 NUTS2 regions		
All versus no Region1	130.78 (0.000)	147.34 (0.000)
All versus no Region2	12.34 (0.579)	76.03 (0.000)
All versus no Region3	28.23 (0.013)	41.18 (0.000)
All versus no Region4	n.a.	39.83 (0.000)
All versus no Region5	n.a.	19.05 (0.163)
All versus no Region6	154.32 (0.000)	49.39 (0.000)
All versus no Region7	50.66 (0.000)	37.86 (0.000)
3 large regions: West, East, Central		
All versus no West	116.49 (0.000)	96.34 (0.000)
All versus no Central	n.a.	60.80 (0.000)
All versus no East	109.48 (0.000)	106.36 (0.000)

Note: N.a.: model fitted on these data fails to meet the asymptotic assumptions of the Hausman test

Table 7 Location choice with conditional logit and fixed effects

Specification	CLFE(1)	CLFE(2)	CLFE(3)	CLFE(4)	CLFE(5)
Fixed effects	County	county	county	Region	Region
Ln (Total income)		-0.48	-0.08		
		(0.36)	(0.31)		
Ln (income per capita)	-0.51			-0.07	-0.30
	(0.36)			(0.23)	(0.21)
Ln (population size)				0.41***	-0.51
				(.15)	(0.36)
Ln(own industry local output)	0.20***	0.20***	0.21***	0.21***	0.21***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
Ln (own industry national access	-0.24***	-0.24***	-0.26***	-0.13**	-0.18***
	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)
Ln(local supplier access)	0.12**	0.12**	0.10*	0.10**	0.09*
	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
Ln(local market access)	-0.07	-0.99***	-1.02***	-0.57***	-0.64***
	(0.04)	(0.16)	(0.17)	(0.13)	(0.15)
Ln (national supplier access	-1.01***	-0.07	-0.07	-0.08*	-0.11**
	(0.16)	(0.04)	(0.05)	(0.04)	(0.04)
Ln (national market access)	0.45***	0.46***	0.40***	0.71***	0.57***
	(0.13)	(0.13)	(0.13)	(0.11)	(0.12)
Ln (local business access)	-0.13	-0.15	-0.14	0.10	-0.04
	(0.09)	(0.10)	(0.10)	(0.07)	(0.08)
Ln (local wage)	0.87	0.95		-0.20	
	(0.56)	(0.59)		(0.49)	
Ln (local, own industry wage)			-0.36*		-0.40**
			(0.19)		(0.19)
Ln (local blue-collar wage)			0.37*		0.40**
			(0.19)		(0.19)
Ln (number of college students)					0.95***
					(0.31)
Ln (Size of highway network)		0.63***	0.75***		-0.11
		(0.20)	(0.20)		(0.08)
Ln (population density: inhabitants/area)		1.14***	1.18***	0.27***	0.02
		(0.23)	(0.24)	(0.07)	(0.10)
Ln (Size of telephone network)		0.04	-0.03		0.15*
		(0.10)	(0.10)		(0.09)
Ln Weighted distance of export borders)					-0.59***
					(0.14)
LR chi^2	5439	5439	5307	5301	5212
Log likelihood	-10632	-10632	-10388	-10701	-10436
McFadden's pseudo R^2	0.204	0.2037	0.2035	0.1985	0.1998
Number of observations	4557	4557	4557	4557	4557
Lnedu					0.95***
					(0.31)

Standard errors in parentheses. Significance at 1%, 5% and 10% is denoted by ***, **, and *, respectively.

Table 8 Location choice with nested logit

Specification	NL(1)	NL(2)	NL(3)	NL(4)	NL(5)
Top level alternatives	3	3	3	7	7
Ln (income per capita)	0.67***	0.74***	0.68**	0.10	-0.30
	(0.18)	(0.26)	(0.27)	(0.35)	(0.35)
Ln (population size)	0.32***	0.87***	0.09	1.34***	-0.53
	(0.13)	(0.18)	(0.46)	(0.37)	(0.58)
Ln (local corporate access)	0.18***				
	(0.04)				
Ln (national corporate access)	0.24***				
	(0.04)				
Ln(own industry local output)		0.31***	0.31***	0.29***	0.28***
		(0.02)	(0.03)	(0.02)	(0.02)
Ln (own industry national access		-0.11	-0.08	-0.18**	-0.17**
		(0.07)	(0.07)	(0.08)	(0.07)
Ln(local supplier access)		0.16***	0.16***	0.07	0.09*
		(0.06)	(0.06)	(0.05)	(0.05)
Ln(local market access)		-0.16***	-0.17***	-0.07	-0.05
		(0.04)	(0.04)	(0.05)	(0.04)
Ln (national supplier access		-0.72***	-0.57***	-0.86***	-0.74***
		(0.15)	(0.15)	(0.16)	(0.16)
Ln (national market access)		0.87***	0.91***	0.91***	1.03***
		(0.13)	(0.13)	(0.13)	(0.13)
Ln (local business access)	0.29***	0.29***	-0.04	0.23**	-0.02
	(0.05)	(0.07)	(0.10)	(0.09)	(0.11)
Ln (local wage)	-1.34***	-0.83	-0.18	0.48	0.91
	(0.45)	(0.56)	(0.65)	(0.69)	(0.69)
Ln (number of college students)			0.54		1.38***
			(0.43)		(0.52)
Ln (Size of highway network)			0.04		0.23**
			(0.08)		(0.11)
Ln (population density: inhabitants/area)			0.12		-0.06
			(0.10)		(0.12)
Ln (Size of telepohone network)			0.40***		0.21*
			(0.12)		(0.11)
IV1	0.96	1.43	1.37	1.43	1.46
IV2	1.10	1.77	1.45	2.41	1.84
IV3	0.86	1.45	1.77	2.77	2.28
IV4				2.62	2.11
IV5				2.16	1.51
IV6				1.64	1.16
IV7				2.18	1.71
LR test of IVs=1	76.5	93.0	59.1	147.2	103.8
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
LR chi^2	4964	6434	6461	6488	6505
Log likelihood	-10869	-12809	-12796	-12783	-12774
Number of observations	4557	4557	4557	4557	4557

Random utility model consistent nested logit. Standard errors in parentheses. Significance at 1%, 5% and 10% is denoted by ***, **, and *, respectively.

Table 9 Location choice with count data regressions

Specification	CNT(1)	CNT(2)	CNT(3)	CNT(4)	CNT(5)	CNT(6)
Model	Poisson	Poisson	Poisson	Poisson	Negative binomial	Negative binomial
FE	No	No	County	Area, time	No	Area, time
Ln (income per capita)	1.62***	0.65***	-0.10	0.39***	0.92***	0.53***
	(0.11)	(0.15)	(0.26)	(0.12)	(0.19)	(0.16)
Ln (population size)	0.82***					
	(0.07)					
Ln(own industry local output)	0.23***	0.24***	0.25***	0.25***	0.26***	0.26***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Ln (own industry national access)	-0.02*	-0.03**	-0.04***	-0.03**	-0.12***	-0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Ln(local supplier access)	-0.09***	-0.12***	-0.07***	-0.14***	-0.17***	-0.18***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Ln(local market access)	-0.06***	-0.08***	0.02	-0.01	-0.05*	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Ln (national supplier access)	0.02	0.07***	0.04	0.12***	0.25***	0.29***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.04)	(0.04)
Ln (national market access)	0.17***	0.08***	0.01	0.13***	0.05*	0.08***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)
Ln (local business access)	0.006	-0.10**	-0.13***	0.47***	-0.12*	0.43***
	(0.04)	(0.04)	(0.03)	(0.02)	(0.06)	(0.03)
Ln (local wage)	-0.86***					
	(0.08)					
Ln (local, own industry wage)		-0.68***	-0.58***	-0.81***	-0.76***	-0.86***
		(0.06)	(0.06)	(0.07)	(0.08)	(0.09)
Ln (local blue-collar wage)						
Ln (number of college students)		0.73***			0.63***	
		(0.07)			(0.09)	
Ln (Size of highway network)		0.05			0.02	
		(0.04)			(0.06)	
Ln (population density: inhabitants/area)		0.11***			0.13*	
		(0.05)			(0.06)	
Ln (Size of telepohone network)		0.11*			0.18**	
		(0.06)			(0.09)	
Ln Weighted distance of export borders)		-0.48***			-0.40***	
		(0.05)			(0.07)	
Ln Distance of Airport						
Ln Distance of Western export border						
Ln Distance of Southern export border						
LR χ^2	6936	6661	6830	6837	1748	1822
Log likelihood	-4912	-4579	-4494	-4491	-4163	-4125
McFadden's pseudo R^2	0.4138	0.4210	0.4318	0.4322	0.1735	0.1809
Over-dispersion α^+					0.36	0.33
LR (α=0), χ01 (p-value)					833 (0.00)	730(0.00)
Number of observations	3000	2737	2737	2737	2737	2737

Standard errors in parentheses. Significance at 1%, 5% and 10% is denoted by ***, **, and *, respectively.
$^+$ χ01: is a one-sided χ^2 test of the over-dispersion parameter, α.