Politecnico di Torino

Polytechnic of Turin

Class LM-31 Degree: MANAGEMENT ENGINEERING
A.y. 2021/2022
Graduation session March 2022

Localization and urbanization economies in the Italian manufacturing sector

Application of the study of Castellani and Lavoratori on Italian companies

Supervisor:
Luigi Benfratello
Anna D'Ambrosio

Candidate:
Michele Olivieri

Abstract

Italy is a country with a great history in the manufacturing sector, famous worldwide for the production of high-quality products. This thesis aims to perform, using an econometric approach, an analysis of productivity for a sample of Italian manufacturing firms, investigating in particular whether it is influenced by economies of agglomeration in order to understand whether heterogeneous or homogeneous clusters of firms can create an advantage or disadvantage in terms of productivity.

The study estimates a Cobb-Douglas production function using the balance sheets of a sample of 78,157 firms observed in the years 2000-2014.

The effects of location and urbanization economies are estimated using various estimators (fixed, random, and multilevel mixed effects) and at the end a comparison is made with the work of Castellani and Lavoratori (2021) based on UK manufacturing firms.

Index

This page intentionally left blank

Introduction

The words “economy” and “economics” are two of the most used words in our life, we heard them every day sometimes more than once. There are newspapers, tv programs, films dedicated to them, a lot of students chose to take a degree on these subjects, but however someone could make a confusion.

- Economy: it is a bundle of activities, that aid in determining how scarce resources are allocated. Usually an economy indicates a region, a particular area or country, concerning production, distribution, consumption, and exchange of goods and services, and supply of money.
- Economics: it regards a social science, to be more specific it is concerned with how an economy and its participants function and behave. It studies the whole life cycle of the goods and services: how they are produced, distributed throughout the economy, and consumed by businesses and individuals. To better understand the difference, in this case the purpose is not to understand how scarce resources are allocated, but instead: how human being behave when there is scarcity of resources.

Basically, economics is the study of an economy and one of the key areas of focus of economics is the understanding of the efficiency surrounding production and the exchange of goods as a result of incentives and policies that are designed to maximize efficiency.

Following these definitions, within an economy, individuals can trade and exchange goods and services on many markets, and finally these individuals are aggregated, and various analysis are done. This is because each economy has its own distinguishing characteristics, although they all share some basic features, they are based on a unique set of conditions and assumptions. Usually, the aggregation of individuals within an economy is finalized in understanding the production growth, unemployment and inflation change.
<u>But what is the best level of aggregation?</u> This is a question with no wrong answer, from one side as always, the larger the boundaries are, the higher the number of individuals and so there are lot of information to make inferences and statistics; but from the other side, wide boundaries mean that many local relations and dynamics are neglected or do not appear, at the same time an excessive finer level of analysis could lead to no results if there are not enough individuals.

Idea of the work

1. What is a determinant for a firm that decide to enter in a certain industry?
2. Is there evidence that firms take advantage in locating near other competitors?
3. Are there Italian areas in which the ground is more fruitful, and it is easier to have a sustainable business?

These are some questions that managers and owners of each firms meet when they want to start a new activity or when they are looking for a competitive advantage, in fact the answers are a matter of strategy and market analysis.
The previous questions can be merged into one and can be rephrased in economic terms, that is the starting point of this thesis:

➔ Is firm productivity affected by agglomeration externalities?

In fact, it will be analysed if the presence of more firms in certain Italian areas allows for a company to run its business easier, just to clarify the idea, it is looking for a situation similar to Silicon Valley in the USA, that is a cluster for IT start-up, or also like the Chinese clothing manufacturers that have seen a strong growth in manufacturing industries on the south-east coast.

The approach followed is not based on a market study and there is not any business plan; instead, it is carried out an analysis on the business data of all the Italian firms in the past years.
The focus is just on the Italian manufacturing firms, as done also by Lavoratori and Castellani in their study "Too close for comfort? Microgeography of agglomeration economies in the United Kingdom" of the 2021, from which this thesis take inspiration.

The paper of Lavoratori and Castellani contributes to the literature in three main ways:

- Provide additional evidence on the role of agglomeration economies on firm productivity, moving toward a microgeographical approach.
- Exploiting a property of mixed-effect models, they also contribute to the debate on the factors that moderate the benefits that firms achieve from locating in highly agglomerated areas.
- Lastly, they aim at contributing to the recent call for subnational and sub-regional analyses to overcome country boundaries and investigate location phenomena at extremely fine-grained geographical scales to capture within-country heterogeneity and zoom in "to a much smaller scale to get a true picture of locational advantage".

Summarizing all the previous discussion, in this study it is analysed the whole Italian manufacturing companies in the years between 2000 to the 2014, it is applied a method that aim to find these evidences listed below:

- If localization or urbanization economies affect positively the productivity of firms.
- At which spatial level of analysis it is easier to find that localization and urbanization are related to the productivity.
- If localization and urbanization coexist.
- The results are homogeneous or if there are other factors that affects these externalities and so the results are heterogeneous.

The main characteristic of the study of Lavoratori and Castellani with respect to this one is shown in the table.

Table 1

Profile of the analysis	Lavoratori & Castellani	Thesis
Country of interest	United Kingdom	Italy
Number of firm analyzed	4.927 firms The initial sample was about 10.000 enterprises, but were chosen only the firms without subsidiaries. The sample is with no missing values	78.157 firms
Period	From 2008 to 2016	From 2000 to 2014
Geographical level, dimension of the aggregation layers	1. City-wide – postcode area. 2. Within-city level – a square of 3 x 3 km with a focal company as centroid. 3. Neighbourhoods around a firm – a square of 1 x 1 km with a focal company as centroid.	1. City-wide – CAP. 2. Neighbourhoods around a firm – a circle with a focal company as centroid and a radius of 1 km
Agglomeration economies studied	Localization and Urbanization In order to better detect their presence, it was retained the whole sample of firms in the analysis of agglomeration measures.	Localization and Urbanization

The main advantage of this thesis is that the sample of firms is higher than the one of Lavoratori and Castellani, and furthermore an additional contribution that be given:

- To find some difference/evidence of the Italian situation with respect to the UK reported by Lavoratori and Castellani.

Production Function

The direction that in this study is taken is the one to apply the regression on a production function, and in turn the productivity is estimated through a model like the following:

$$Y_{it} = A_{it} K_{it}^{\beta_K} L_{it}^{\beta_L} M_{it}^{\beta_M}$$

Equation 1

A Cobb-Douglas, one of the most used models to estimate a production function. The production function describes the technology of a firm, specifically the relationship between the quantity used of the various factors of production or inputs and the level of the product or output. The input of interest are Capital, Labour and Raw material, the output instead is the Production. $\beta_K, \beta_L, \beta_M$ are parameters that determine the overall efficiency of production and the responsiveness of production to changes in input quantities and they are the parameters to be estimated.

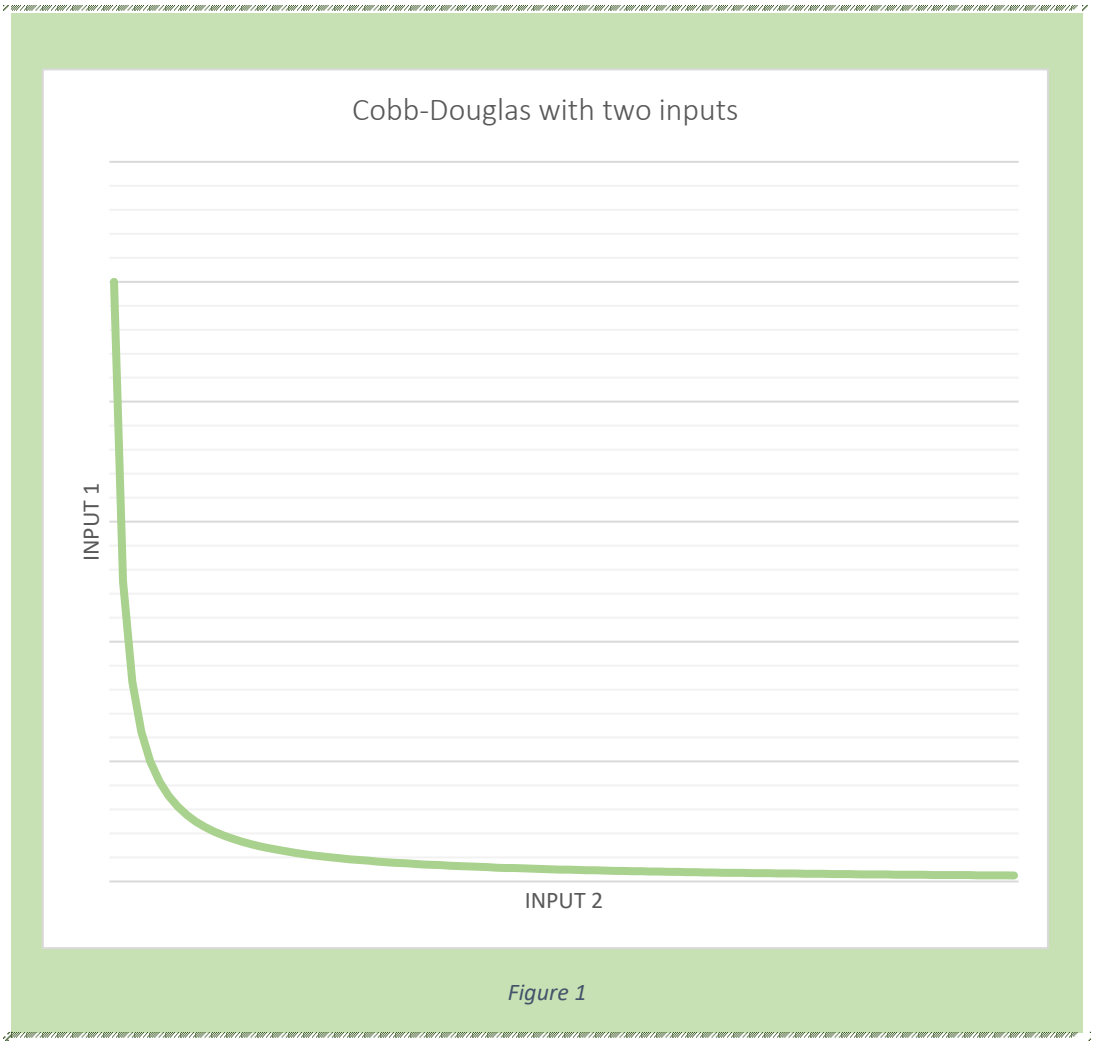


Figure 1

In the Figure 1 is presented an indifference curve typical of Cobb-Douglas with only two inputs (the two β depicted on the graph are put to 0.5 each), since it is an isoquant, each point on the curve correspond to a different combination of the two inputs (ex. Labour and Capital) that gives the same output; but unlike a linear production function, the rate at which Labour can be substituted for Capital is not constant as you move along an isoquant. The shape of the indifference curve of Cobb-Douglas depends on the values of the coefficient $\beta_K, \beta_L, \beta_M$ that have a particular meaning:

- $\sum \beta_i = 1$ constant return to scale
- $\sum \beta_i < 1$ decreasing return to scale
- $\sum \beta_i > 1$ increasing return to scale

If we consider only two input and if the two β are different from each other, the curve will take a shape closer to the input's axis with the higher β. The constant A is a multiplicative one and can be considered an indicator of the degree of efficiency in the use of all factors of production. It is therefore an efficiency parameter that indicates the level of technology. Although Y_{it}, M_{it}, K_{it} and L_{it} are all observed by the econometrician (although usually in value terms rather than in quantities), A_{it} is unobservable to the researcher.

A desirable property of the Cobb-Douglas that will be used is to be log-linear.

$$\ln(Y_{it}) = \ln(A_{it}) + \beta_K \ln(K_{it}) + \beta_L \ln(L_{it}) + \beta_M \ln(M_{it})$$

Equation 2

To better read Equation 2, the log normal are substituted by a lower-case letter.

$$y_{it} = \beta_0 + \beta_K \mathrm{k_{it}} + \beta_L l_{it} + \beta_M m_{it} + \varepsilon_{it}$$

Equation 3

With:

$$\ln(A_{it}) = \beta_0 + \varepsilon_{it}$$

Equation 4

Where β_0 measures the mean efficiency level across firms and over time; ε_{it} is the time- and producer-specific deviation from that mean, which can then be further decomposed into an observable (or at least predictable) and unobservable component; this results in the following Equation 5.

$$y_{it} = \beta_0 + \beta_K \mathrm{k_{it}} + \beta_L l_{it} + \beta_M m_{it} + v_{it} + u_{it}^q$$

Equation 5

Where $\omega_{it} = \beta_0 + v_{it}$ is the observable (by the firm and not by the econometrician) component of $\ln(A_{it})$ and represents firm-level productivity while u_{it}^q is an i.i.d. component, representing unexpected deviations from the mean due to measurement error, unexpected delays, or other external circumstances.

The final objective will be to estimate the productivity, and so:

$$\widehat{\omega_{it}} = \widehat{\beta_0} + \widehat{v_{it}} = y_{it} - \widehat{\beta_K} k_{it} - \widehat{\beta_L} l_{it} - \widehat{\beta_M} m_{it}$$

Equation 6

and productivity in levels can be obtained as the exponential of $\widehat{\Omega_{it}} = e^{\widehat{\omega_{it}}}$. Making the various substitution it is possible to highlight that it simply become a ratio of output over input:

$$\widehat{\Omega}_{it} = \frac{Y_{it}}{K^{\widehat{\beta_K}} L^{\widehat{\beta_L}} M^{\widehat{\beta_M}}}$$

Equation 7

The estimation of the parameters will be done using Equation 6.

Common methodologies

In order to implement agglomeration economies analysis, an estimate of the productivity is needed. Theoretically it is possible to think to productivity as it is showed in equation 6, but in so doing there are many unknowns parameters that need to be calculated: all the $\widehat{\beta_\iota}$ in the formula. The literature provides a huge list of methods with pros and cons, to estimate productivity, a summary of how these methods work is given below. It starts from one of the best methods used in the world of regression when a linear function is treated is OLS (Ordinary Least Squares).

OLS

The key idea is that these coefficients can be estimated by minimizing the sum of squared prediction mistakes.

$$\sum_{1}^{n}(y_{it} - b_0 - b_1 x_{ki})^2$$

Equation 8

The b_i that minimize the deviation are the ordinary least squares estimators of β_i that the equation was looking for, and they are indicated as $\widehat{\beta_\iota}$.

The residual is so computed in this way.

$$\widehat{\varepsilon_\iota} = y_i - \widehat{y_\iota}$$

Equation 9

Generation of bias

Although equation 6 can be estimated using OLS, this method requires that the inputs in the production function (the dependent variables) are exogenous or, in other words, determined independently from the firm's efficiency level.

Variables correlated with the error term are called endogenous variables, while variables uncorrelated with the error term are called exogenous variables.

If x and ε_i are correlated the OLS estimator is inconsistent, and this often is due to omitted variable bias. Omitted variable bias is an error generated if the regressor is correlated with a variable that is omitted from the analysis, and that determines, in part, the dependent variable.

And so, inputs in the production function are not independently chosen (OLS is inconsistent), but rather determined by the characteristics of the firm, including its efficiency. This endogeneity of inputs or simultaneity bias is defined as the correlation between the level of inputs chosen and unobserved productivity shocks, firms that have a large positive productivity shock may respond by using more inputs. Since OLS cannot give its contribution, three additional methods can be considered in order to calculate $\widehat{\beta_\iota}$ and lastly productivity.

- ➔ A fixed effect estimation allows to overcome this problem (with discrete results) by assuming that the productivity ω_i is plant specific but time invariant (it involve many assumption).
- ➔ An instrumental variable (IV) estimator achieves consistency by instrumenting the explanatory variables with regressors that are correlated with the inputs but uncorrelated with the error term.

➔ Control function estimator, that is a more recent approach in which unobserved firm productivity is proxied by a function of observed firm characteristics that reflect a firm's reaction to productivity changes.

Further biases not strictly related to OLS method but to consider are the following:

- Selection bias/Endogeneity of attrition: it is generated when a failure in take explicitly into account firm exit decision is done. The bias emerges because the firms' decisions on the allocation of inputs in a particular period are made conditional on its survival. If firms have some knowledge about their productivity level ω_{it} prior to their exit, this will generate correlation between ε_{it} and the fixed input capital, conditional on being in the data set. This correlation has its origin in the fact that firms with a higher capital supply will likely be able to survive with lower ω_{it} relative to firms with a lower capital stock.
  The selection bias will generate a negative correlation between ε_{it} and k_{it} causing the capital coefficient to be biased downwards, and as a result, ignoring the exit rule of the firm will result in firm-level productivity estimates that are biased upwards.
- Omitted price bias: it arises in presence of imperfect competition, when firms can set different prices for the same good. A difference is created between the firm level of prices and the industry level of prices and this can lead to misleading consideration if the regression is based on firm sales. The problem is that typically the data on hand are firm sales $Y_{it} * P_{it}$ (where Y_{it} is firm total output in quantities and P_{it} is the firm level of prices) and firm level of prices is not available to the researchers, so it is not possible to find the variation of the physical output.
  In the absence of information on firm level prices, industry level price indices are usually applied to deflate firm level sales and input expenditures in traditional production function estimates.
  ➔ If firm level price variation is correlated with input choice, this will result in biased input coefficients.
  Failure to account for firm level deviations of industry level prices can result in sizeable biases in estimated productivity. Suppose for example that cost savings realized by a more efficient producer are passed through to consumers as lower output prices. If firm level output is proxied by the deflated value of sales, this will lead to an under-estimation of that particular firm's output. Formally, if the price the firm charges is lower than the industry level price index $\overline{P_{it}}$, this will result in lower output for given inputs and hence in an under-estimation of productivity. Similarly, if firm charge higher prices compared to the industry average, productivity will be over-estimated, because higher output prices will be partly translated into higher output for a given amount of inputs.
  The same asymmetry can be found on the prices of inputs, and also in this case it is feasible to think that the prices are firm specific. Failure to take firm level deviations of industry level input price indices into account will generally introduce a bias in estimated productivity that is opposite to that introduced by omitting firm level output price differences. If the firm is able to negotiate lower prices for a given input, the use of industry level prices rather than firm-level input prices will lead to an under-estimation of its input use, causing productivity to be biased upwards.
- Multi-product firm: is present when firms produce more than one kind of product. Consistent estimation requires to know the product mix, the weight of each product on the total output on the inputs, as well as on the prices. Without this information the coefficients calculated assumes identical production techniques and final demand (through the use of common output price deflators) across products manufactured by a single firm.

Instrumental Variables (IV)

IV is a way to obtain a consistent estimator of the coefficients in the production function, it is capable to solve the simultaneity bias that affect OLS regressors.

The working principle is based on the fact that part of the variability of the endogenous variable x_i that is uncorrelated with the error term ε_i is gleaned from one or more additional variables (in this example labelled as Z), called instruments. Instruments must be an additional variable that drive/affect the endogenous variable of the model.

The assumptions that must satisfy the instrumental variables are:

1. $corr(Z_i, x_i) \neq 0$, instruments need to be correlated with the endogenous inputs.
2. $corr(Z_i, \varepsilon_i) = 0$, the instruments cannot be correlated with the error term (and hence with productivity since $\varepsilon_{it} = v_{it} + u_{it}^q$ and v_{it} is a component of ω_{it}, see equation 4, 5 and 6).
3. The instruments Z_i cannot enter the production function directly.

To understand the step with which IV is applied we start considering this productivity function (here is shown IV with 2SLS solution, in some cases it is applied the IV method with GMM).

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Equation 10

So, the IV is applied on the endogenous independent variable.

$$x_i = \pi_0 + \pi_1 Z_i + \sigma_i$$

Equation 11

Where π_0 is the intercept, π_1 is the slope and σ_i is the error term and at the same time the problematic component of x_i correlated with ε_i.

OLS is applied on this last equation; the coefficients are estimated and are generated predicted values of $\widehat{x_i}$.

$$\widehat{x_i} = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

Equation 12

Putting all together in the initial equation, the new production function become:

$$y_i = b_0 + b_1 \widehat{x_i} + \sigma_i + \varepsilon_i$$

Equation 13

In general, IV method is suitable on a conceptual standpoint, but a great problem when applied to production function is that it is very hard to find a valid instrumental variable that fit well and do not generate other problems.

Olley-Pakes Estimation Algorithm (1996)

They were the first to propose a complete method based on a control function approach to estimate parameters in a production function. Their key idea is to exploit firm investment levels as a proxy variable for unobserved productivity shocks (the original OP contains a correction for potentially endogenous firm entry and exit).

They prove their estimates of productivity to be consistent under those assumption mentioned above:

- $i_{it} = i_t(k_{it}, \omega_{it})$ is the investment policy function, invertible in ω_{it}. Moreover, i_{it} is monotonically increasing in ω_{it}, and so only non-negative value of i_{it} can be used in the analysis.
- There is only one unobserved state variable at the firm level, its productivity, which is also assumed to evolve as a first-order Markov process $\omega_{it+1} = E[\omega_{it+1}|\omega_{it}] + \xi_{it+1}$, where ξ_{it+1} represent the productivity shock, assumed to be uncorrelated with productivity and capital, and correlated with labour and material.

- The state variables k evolve according to the investment policy function i_{it}, which is decided at time $t-1$.
- If industry-wide price indices are used to deflate inputs and output in value terms to proxy for their respective quantities, it is implicitly assumed that all firms in the industry face common input and output prices.
- The free variables l_{it} and m_{it} are nondynamic, in the sense that their choice at t does not impact future profits and are chosen at time t after the firm realizes productivity shock.

At the start of each period t, each incumbent firm take the decision to exit or to continue to run their business.

- ✓ Run: it chooses the level of l_{it} and m_{it} (variable input) and investment i_{it}.
- ✗ Exit: the firm receives a sell-off value, and it never re-enters.

The firm is assumed to maximize the expected discounted value of net cash flows and investment and exit decisions will depend on the firm's perceptions about the distribution of future market structure, given the information currently available. Both the lower bound to productivity (cut-off value below which the firm exit) and the investment decision are determined as part of a Markov perfect Nash equilibrium and will hence depend on all parameters determining equilibrium behaviour.

Capital k_{it} is a state variable, only affected by current and past levels of ω_{it}. Investment i_{it} can be derived:

$$i_{it} = k_{it+1} - (1-\delta)k_{it} \quad => \quad i_{it} = i_t(k_{it}, \omega_{it})$$

Equation 14

Note that from the capital rule, lagged investment should be used to invert out productivity, Olley and Pakes experiment with both current and lagged investment in their empirical application but current investment generates correlation between capital and productivity.

Since investment is strictly increasing in ω_{it}, it can be inverted.

$$\omega_{it} = i_t^{-1}(k_{it}, i_{it}) = h_t(k_{it}, i_{it})$$

Equation 15

Equation 15 is substituted in equation 5 in this way.

$$y_{it} = \beta_0 + \beta_K \mathrm{k_{it}} + \beta_L l_{it} + \beta_M m_{it} + h_t(k_{it}, i_{it}) + u_{it}^q$$

Equation 16

Now it is defined the following function.

$$\varphi(i_{it}, k_{it}) = \beta_0 + \beta_K \mathrm{k_{it}} + h_t(k_{it}, i_{it})$$

Equation 17

Equation 16 is defined proceeds in two steps, the first one in which OLS is applied on equation 18.

$$y_{it} = \beta_L l_{it} + \beta_M m_{it} + \varphi(i_{it}, k_{it}) + u_{it}^q$$

Equation 18

The output of the first step is the estimate of the coefficients of labour $\widehat{\beta_L}$ and material $\widehat{\beta_M}$ that are the variable factors. Just as a conceptual point, $\varphi(i_{it}, k_{it})$ is approximated with a higher-order polynomial.

The second step instead restart from equation restart from equation 16 written in different way, where χ_{it+1} means conditional on firm survival (the firm continue to operate if ω_{it+1} is higher than the threshold value of productivity.

$$y_{it+1} - \beta_L l_{it+1} - \beta_M m_{it+1} = \beta_0 + \beta_K \mathrm{k_{it+1}} + E[\omega_{it+1} | \omega_{it}, \chi_{it+1}] + \xi_{it+1} + u^q_{it+1}$$

Equation 19

From the law of motion for the productivity shocks:

$$y_{it+1} - \beta_L l_{it+1} - \beta_M m_{it+1} = \beta_0 + \beta_K \mathrm{k_{it+1}} + g(P_{it}, \varphi_{it} - \beta_K \mathrm{k_{it}}) + \xi_{it+1} + u^q_{it+1}$$

Equation 20

And P_{it} is the probability of survival of firm i in the next period, and as in the first step it is approximated with a higher-order polynomial. Substituting also the coefficients of labour and material find previously, it is possible to apply non-linear least squares and to find the coefficient of the capital.

Levinsohn-Petrin Estimation Algorithm (2003)

The monotonicity condition of OP requires that investment is strictly increasing in productivity. Because this implies that only observations with positive investment can be used when estimating equations 18 and 20, this can result in a significant loss in efficiency, depending on the data at hand. Moreover, if firms report zero investment in a significant number of cases, this enters doubt on the validity of the monotonicity condition. Hence, Levinsohn and Petrin (2003) propose a similar control function approach in which they use intermediate inputs rather than investment as a proxy. Because firms typically report positive use of materials and energy in each year, it is possible to retain most observations, which also implies that the monotonicity condition is more likely to hold.

The assumption on which LP is based are:

- Firms observe their productivity shock and adjust their optimal level of intermediate inputs (materials) according to the demand function $m(\omega_{it}, k_{it})$.
- $m_{it} = m_t(\omega_{it}, k_{it})$ is the intermediate input function, invertible in ω_{it}. Moreover, m_{it} is monotonically increasing in ω_{it}.
- The state variables k evolve according to the investment policy function i_{it}, which is decided at time $t-1$.
- The free variables l_{it} and m_{it} are nondynamic, in the sense that their choice at t does not impact future profits and are chosen at time t after the firm realizes productivity shock.

From these assumptions it is possible to work (similarly as before) inverting the proxy relationship, in fact it is possible to write the followings.

$$\omega_{it} = m_t^{-1}(m_{it}, k_{it}) = s_t(m_{it}, k_{it})$$

Equation 21

And also:

$$y_{it} = \beta_0 + \beta_K \mathrm{k_{it}} + \beta_L l_{it} + \beta_M m_{it} + s_t(m_{it}, k_{it}) + u^q_{it}$$

Equation 22

They are defined:

$$\varphi(i_{it}, k_{it}) = \beta_0 + \beta_K \mathrm{k_{it}} + \beta_M m_{it} + s_t(m_{it}, k_{it})$$

Equation 23

$$y_{it} = \beta_L l_{it} + \varphi(m_{it}, k_{it}) + u_{it}^{q}$$

Equation 24

Equation 24 is initially solved exactly as in the case of OP, by approximating $\varphi(m_{it}, k_{it})$ with a higher-order polynomial, in the second step equation 23 is regressed.

It should be noted that the coefficient on the proxy variable, is now only recovered in the second stage of the estimation algorithm. The second difference between the approach used by OP and LP is in the correction for the selection bias. Although OP allow for both an unbalanced panel as well as the incorporation of the survival probability in the second stage of the estimation algorithm, LP do not incorporate the survival probability in the second stage. Estimation of a value-added production function is fully analogous to the approach used by OP and summarized above.

Considerations of OP and LP

- Timing of input choices: the methodologies of OP and LP assume that there is at least one input that is costless to adjust and that will respond to new information immediately.
- For the labour coefficient to be identified in the first stage of the estimation algorithm, it is required that there exists some variation in the data, independent of investment (or materials for LP). If this is not the case, it can be shown that the labour coefficient will be perfectly collinear with $\varphi(*, k_{it})$ in the first stage estimation and hence will not be identified, this problem can be solved for OP with some additional assumptions.

Agglomeration Economies

The concept of agglomeration economies refers to the economic benefits that come when firms locate near each other and create spatial clusters.

The literature traditionally emphasises three sources of agglomeration economies: linkages between intermediate and final goods suppliers, labour market interactions, and knowledge spillovers.

- Input-output linkages occur because savings on transaction costs means firms benefit from locating close to their suppliers and customers.
- Larger labour markets may, for example, allow for a finer division of labour or provide greater incentives for workers to invest in skills.
- Finally, knowledge or human capital spillovers arise when spatially concentrated firms or workers are more easily able to learn from one another than if they were spread out over space.

Now that the sources are well described, it is possible to make a classification of the agglomeration economies, there are two major categories: Urbanization economies and Localization economies.

Table 2

Agglomeration economies	Urbanization economies	Localization economies
What does it mean?	Firms in a number of different industries receive benefits from population and infrastructure clusters.	Firms in the same industry get benefits from being located close together.
Benefits?	This is based on the idea that the most important sources of knowledge spillovers are external to the industry in which the firm operate, and the knowledge arises more between industries. The presence of many specialized clusters and the frequent interaction among multidisciplinary individuals can determine urbanization economies and promote technological innovation. This diversified knowledge can be better achieved over a larger geographical space.	The major benefits of localization include: I. The ability to draw from the same skilled group of workers, known as labour pooling II. Quicker spread of ideas among firms within the same industry (knowledge spillovers), and this positively affect the innovation process and firm performance. Proximity can facilitate interactions and communication, coordination and monitoring, exchange of information and knowledge, as well as lead trust across economic parties, crucial factors in the relations with clients and suppliers.
Example?	A great example of this is a shopping mall. Although the stores in the mall may be unrelated, locating close together gives them the opportunity to use the same infrastructure: building, parking lots, and other common areas. Another urbanization benefit in this example is that stores have the opportunity to market and sell to customers who go to the mall to visit another store.	As an example, access to skilled labour specific to the auto industry, common suppliers, and the potential for knowledge spillovers were powerful factors in turning the city into an auto manufacturing hub.

Besides the benefits highlighted above, agglomeration of economic activities may result in negative externalities, mainly in the form of congestion costs, which increase the costs of production factors, and

competition effects, which may crowd out weaker firms and discourage industry leaders to locate in highly agglomerated locations to minimize knowledge leakages.

➔ If external benefits are greater than the added costs, there would be geographic clustering. If the opposite were the case, firms would disperse to places with lower costs.

What determines the level of benefits-costs? Technological change, globalization, government policy and a whole host of other factors change these costs and benefits and hence the nature of this trade-off, with fundamental implications for the economic geography of the state. Of course, the response to these changes is not instantaneous, instead playing out over long periods as people and organisations slowly adjust to the different forces at work.

While the extent and the role played by external agglomeration economies is a well-established fact, what is the appropriate geographical unit of analysis to detect the effects of agglomeration externalities on firms' productivity and at which level of geographical unit these externalities operate are still unclear.

On this subject, various studies have been conducted, and the majority of the results argue that "specialization externalities operate at a finer level than diversification externalities". Here are listed two important studies.

- ❖ USA analysis at ZIP code level (Rosenthal and Strange 2003) – Localization economies (measured by the employment in their own industry) are more important than urbanization externalities (employment in other industries) and that the former rapidly decline with the increase of distance. Instead, urbanization economies present a trade-off between the benefits of being located close to high-populated areas and the related congestion costs. Their results suggest that agglomerations need to be studied at a more granular level, than in previous studies.
- ❖ UK analysis at FUA and at the two ZIP codes level (Lavoratori 2018) – It was investigated the effects of external agglomeration factors on the labour productivity in 2015 at different level of relatively fine geographical disaggregation: the functional urban area level, the municipality level using the postcode area and the sub-municipality level, using postcode district and sector. In order to do this, the study carries out a multilevel empirical analysis on 5.627 manufacturing firms in the United Kingdom. Findings show that the functional urban area seems not to be an appropriate level to detect agglomeration effects, which instead are significant at finer levels of geographical aggregation. While the diversification plays a role at the municipality level (postcode area), specialization externalities operate at a finer level, within the municipality (postcode district and sector) in a closer neighbourhood to the firm.

Localization and urbanization externalities do not necessarily "compete," they may "coexist" in the same geographical areas, because the combination of these economies may contribute to firm growth, but in different ways.

Application

Estimation model

Panel data provide information on individual behaviour, both across individuals and over time, so the data have both cross-sectional and time-series dimensions, it become challenging to model the dataset in the right way.

In general, every time it is needed to describe how the dependent variable is influenced by the effect of the independent variables a mathematical model is used. And the model defined for any given situation depends on what the values of the independent variables mean.

To answer to the question wrote in the introduction "*Is firm productivity affected by agglomeration externalities?*", it was decided to apply two regression steps.

1. The first is to apply the Levinsohn & Petrin control function in order to estimate the coefficients of the production function, and subsequently using them to predict the productivity values for all the observations.

   Some important points are that:

   - differently from what said above in the description of LP method, it was decided that the variable material is used only as proxy variable and not as an input determining the production (free variable). This choice is justifiable since in general, for the LP estimator, labour and materials are both chosen simultaneously, a natural assumption could be that they are allocated in similar ways, and consequently they depend on the same variables capital $k_{it}$and productivity ω_{it}. Because it is not possible to simultaneously estimate a non-parametric function of ω_{it} and k_{it} together with the coefficient on the labour variable, which is also a function of those same variables, the labour coefficient will not be identified in the first stage. Hence collinearity between the labour variable and the non-parametric function in the first stage can cause the labour coefficient to be unidentified.
   - all the assumption of LP method are treated as verified (also the one of invertibility of the material), and all the problems related to selection bias, omitted price bias (variables are deflated with some industry specific indices) and multi-product firms are neglected during the estimation.

2. The second step is the one to consider the agglomeration economies. It is done starting from the productivity estimated earlier through LP, it becomes the dependent variable. While as independent variables they are used various indices about localization, urbanization economies and some firm characteristics. The regression now could be applied, but with which method? What model best fit the data?

   In general, there are three statistical models that can be applied, and the right one depends upon the data; these models are based on different assumption and chose the wrong model can lead to wrong conclusion. The first two model listed below are Fixed and Random effect.

Table 3

Variable/Effect	Description	Example
Fixed	It is typically referred to variables that is assumed to be measured without error. In this case the values of this independent variable represent the entire population of values.	A variable that identifies a smoker can just take two values and there is no uncertainty. Since everybody falls into one of those two categories, there are no other categories to worry about. A drug study might use 0 mg, 5 mg, and 10 mg of an experimental drug, the population is restricted to these 3 values only.

Random	It is assumed that these values are drawn from a larger population of values and thus will represent them. It is possible to think about the values of random variables as representing a random sample of all possible values or instances of that variable. It is expected to generalize the results obtained with a random variable to all other possible instances of that value. When it is treated a random variable, it does not care about the specific level values, since they will be generalized across the various levels.	A study finalized in determine the variation in acceleration among models of car, testing the idea that variation among models is relatively small compared with the variation caused by differences in the smoothness of gear changes, or the reaction speeds of the drivers. In this case it is more appropriate to select a sample of car models to be representative of the wider population of models which exist.

When the independent variables are characterized with both Fixed and Random effect, in statistical term it is called Mixed model and it is the third method to regress. Mixed models are especially useful when working with a within-subjects design because it works around the ANOVA assumption that data points are independent of one another. In a within subjects design, one participant provides multiple data points and those data will correlate with one another because they come from the same participant. Therefore, using a mixed model allows to systematically account for item-level variability (within subjects) and subject-level variability (within groups).

Mixed model is used also by Castellani and Lavoratori in their work and it is possible to say that it can be applied in the same way also on this database. In fact, the whole data are well characterized by three dimension or three level: time (15 years of observations), geography (all the observations are well established in some region/town) and lastly the firms (observations).

So, to be clearer this third method that could fit the data is a multilevel mixed model, and it make sense to use this regression when the data have an hierarchical relationship, where the levels at the bottom are nested into the higher, see the figure below.

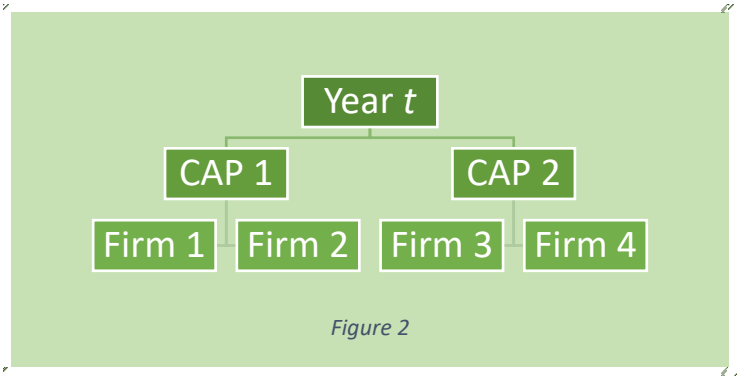


Figure 2

Obviously, there are variables such as age that can be well described by a fixed effect, their variability does not depend on different firms, and the effect due to the change in this variable is the same whatever the firm; fixed effects estimate separate levels with no relationship assumed between the levels (only variability within the layers). And there are the variables of the layers that instead are assumed to be random, in this way a variability between group is introduced in the model; each level can be thought of as a random variable from an underlying process or distribution.

Mixed model is in line with the requirements of this study, in fact it allows to simultaneously model firm and location variables, controlling for the spatial dependence due to the nested structure of the data and correcting standard errors measurements. A multilevel analysis allows to study the variance of the outcome at each level, measuring the unobserved group-level heterogeneity, but maintaining the firm as the unit of analysis.

Dataset

The first important step for this study was the one to retrieve the data on which to make a regression analysis, or more in general to operates. And the input data were provided from the online platform AIDA, that contains all the income statements (multi-step format) and balance sheets of all the Italian businesses for a time lap of 15 years, during the years 2000 – 2014.

The importance and value of information within economics is huge. It mitigates risk and uncertainty, and it makes it possible to take better choices that will report higher yields. The less risk and uncertainty there is, the higher the utility taken from the information in the decisions will be valued.

This database is very detailed and in its initial form, it was endowed with a total of 293 variables, and 17.224.170 observations (roughly 39 GB of information):

- ✓ In these observations the consolidated financial statements (the ones that present the assets, liabilities, equity, income, expenses and cash flows of a parent firm and its subsidiaries as those of a single economic entity) are not taken into account.
- ✓ Each observation is a full set of information for one firm in one year, since we are analysing the time lap of 15 years from 2000 to 2014, we have 15 observations for each company.
- ✓ The consequent number of firms was $\frac{17.224.170}{15} = 1.148.278$
- ✓ The 239 variables can be easier viewed as composed of 4 sections:
  - 34 of the variables are the information to uniquely identify the firm such as name, address, postal code, CCIAA number, latitude and longitude, fiscal code, ATECO's code and others.
  - 63 of the variables are the information related to the income statement.
  - 136 of the variables are the information related to the balance sheet.
  - 60 of the variables are the indexes computed on the statements.
- ✓ The only kinds of businesses that are not present in the database are: sole proprietorships and the freelancers.
- ✓ The whole study is made with a focus on manufacturing activities, identifiable through the ATECO code; a company belongs to this group if its code starts with 10, 33 or a number in this range.

Of course, for the final analysis there is no need to use all these variables and some observations are useless, but the best way to understand how to move and which decisions we should take with which results, it is of critical importance to study the dataset and to put down some numbers before to drop anything.

Some of the most interesting information are summarized in the graphs below:

- The region with the highest number of presences in the data is the Lombardia with 243.935 companies, followed by Lazio with 161.994. Lombardia alone has a share of 21,24% and together with Lazio they reach 35,35% presence of all the companies in Italy.

- The region with the lowest number of presences in the data is the Valle D'Aosta with only 2053 businesses.
- A consideration can be done on the businesses belonging to the manufacturer sector, since our study will be done on them; on average they are the 14,45% of all businesses but they are not equally distributed on the regions. In relative terms the region with the highest frequency is Marche with about 22% while one with the lowest share is Lazio with 6.4%.

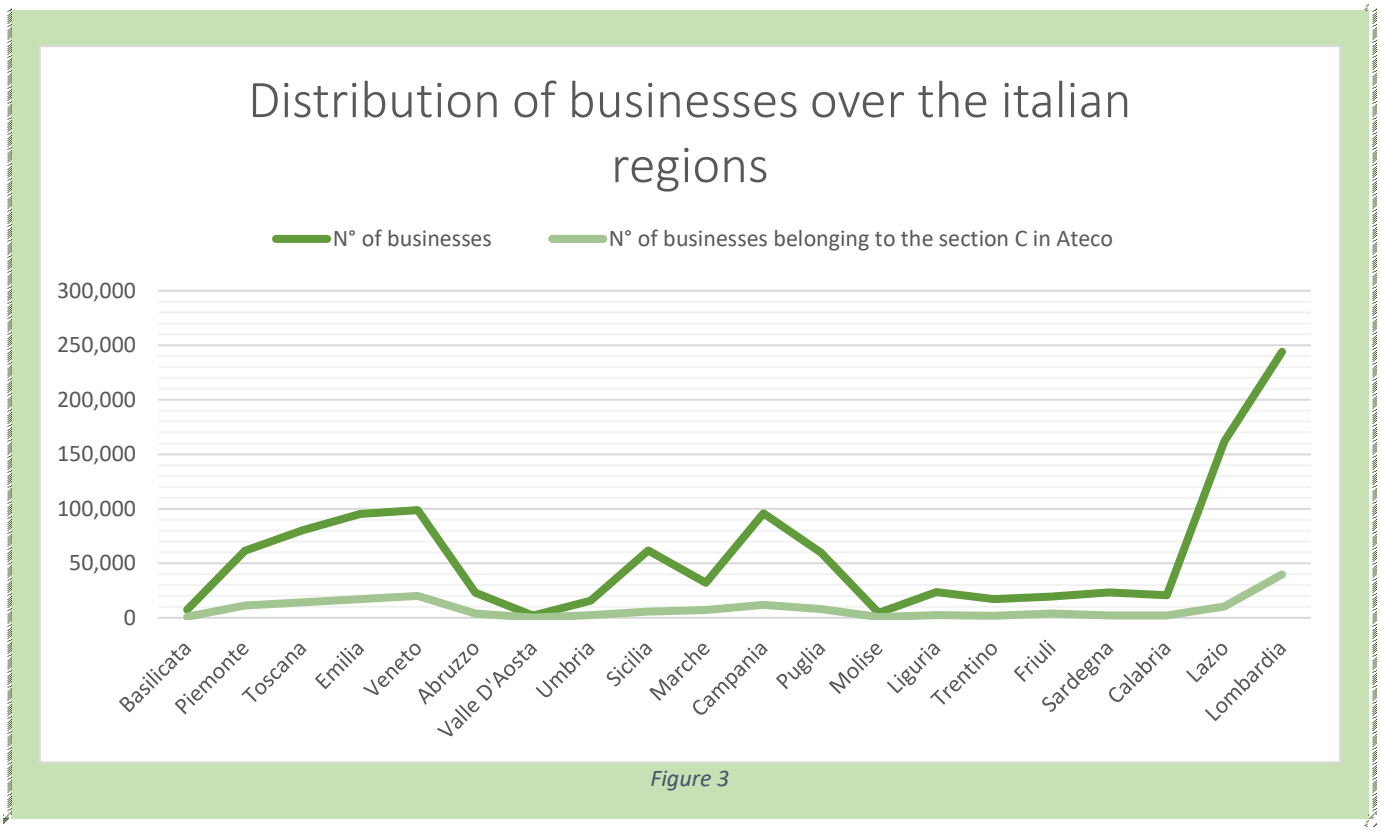


Figure 3

Insight on manufacturing

In its initial state, the dataset will give us a number of entities $n = 1.148.278$ and a number of periods $T = 15$, that is enough to classify these data as "Panel" or "Longitudinal" (Panel data means that each entity is observed at two or more time periods). A further classification can be done on the structure of panel data:

- Balanced Panel: when the observations have no missed data, the variables are observed for each entity and each time period.
- Unbalanced Panel: when there is at least one missing observation for at least one year for at least one entity; this is the typical situation when the amount of information is high and in this case is driven by variation in data coverage and firm survival.

The dataset of AIDA can be classified as unbalanced for many reasons but the most important is that the platform give us at maximum 10 years of observations for businesses over 15 years requested.

Understanding the degree of completeness of the dataset is very important, to this aim the first operation done was to cut the entities that does not belong to the object of the study and so the ATECO code C:

$$drop\ if\ real(ateco) < 100000 \mid real(ateco) \geq 340000$$

This operation gives a new number of $n = 165.942\ businesses$, this number is showed also in the Figure 3 above (distribution per region) under the light green curve.

Starting from this smaller set there are many information that could be extrapolated and between all, the more important are the following.

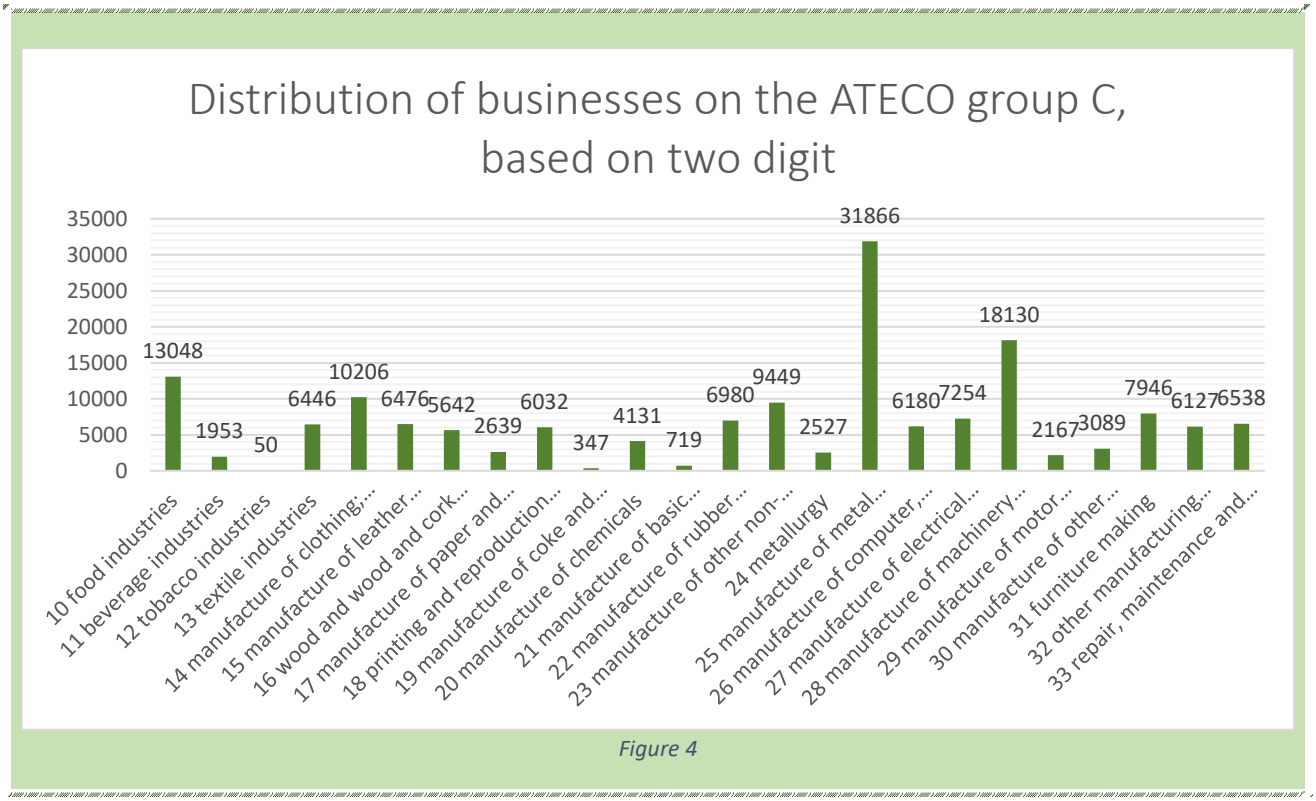
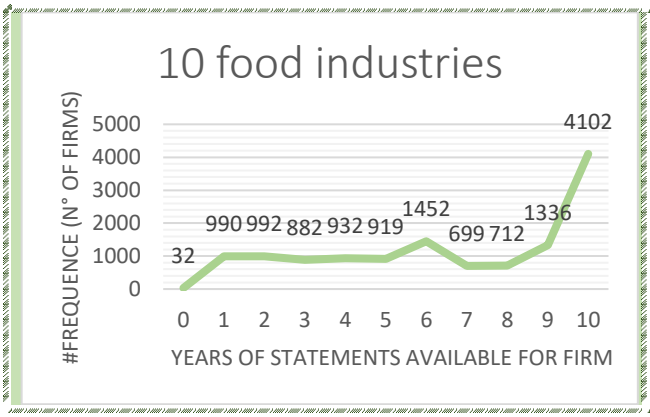


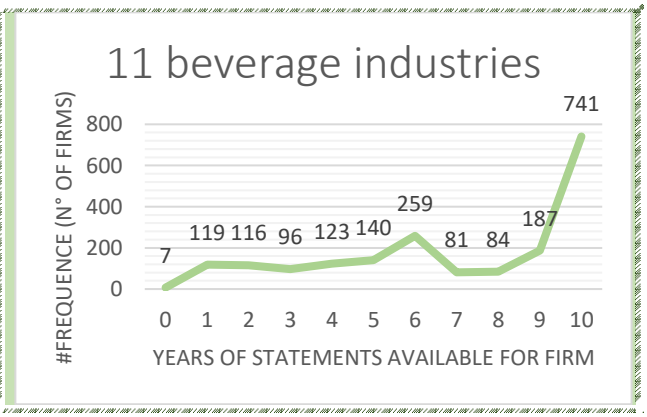
Figure 4

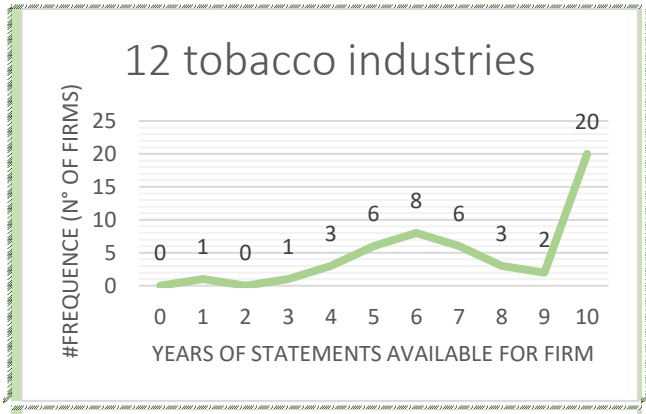
This is the distribution of all the remaining firm across all the ATECO code of group C and so from 10 to 33. The least populated is the group 12 “Industry of the tobacco” while the most populated is 25 “manufacture of metal products (excluding machinery and equipment)”.

The graphs below instead describe the completeness of the financial statements, the firms are grouped by ATECO 2 digit.

Table 4



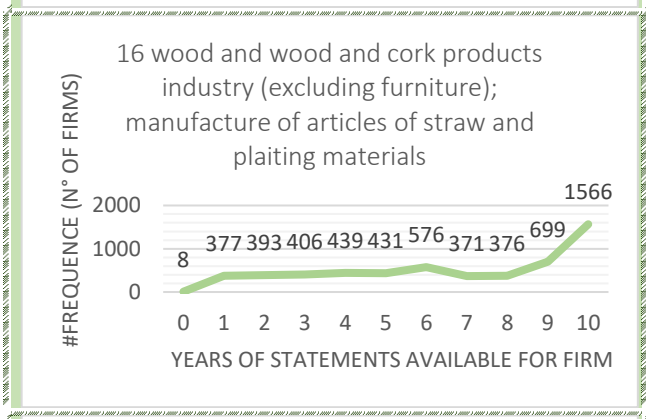


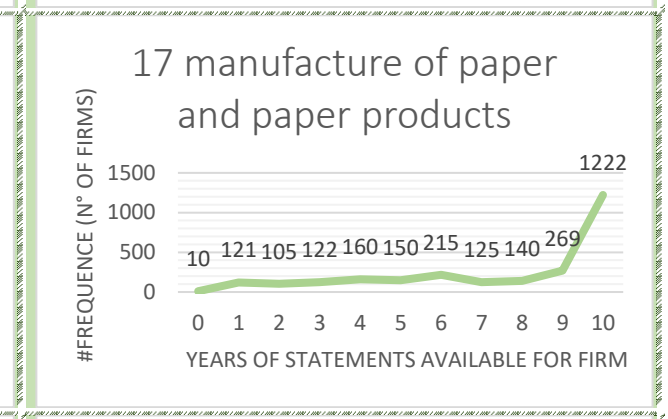

12 tobacco industries
#FREQUENCE (N° OF FIRMS)
25
20
15
10
5
0
0 1 0 1 3 6 8 6 3 2 20
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

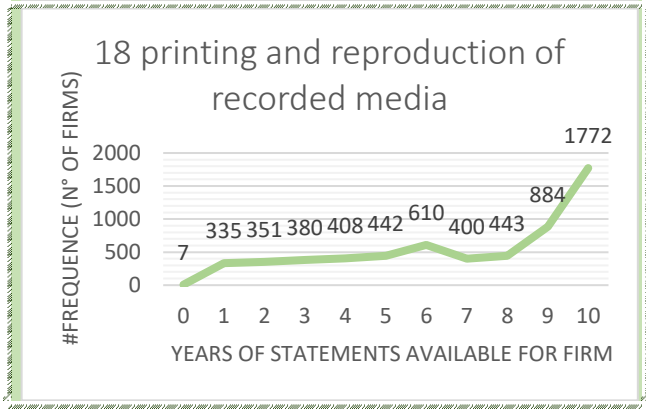

13 textile industries
#FREQUENCE (N° OF FIRMS)
3000
2000
1000
0
18 282 306 350 369 364 544 366 356 722 2769
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

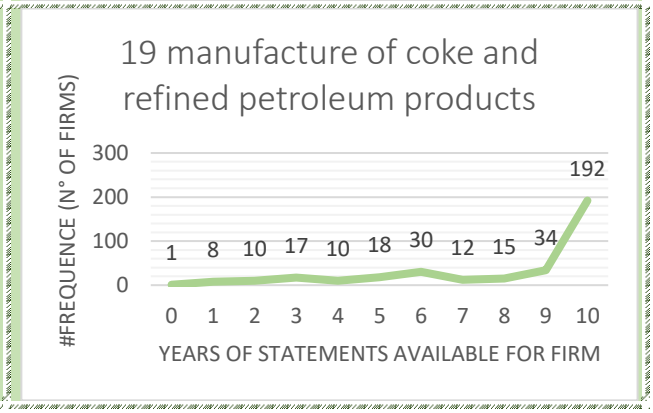

14 manufacture of clothing; manufacture of leather and fur articles
#FREQUENCE (N° OF FIRMS)
3000
2000
1000
0
29 712 755 752 759 853 1180 682 594 1131 2759
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

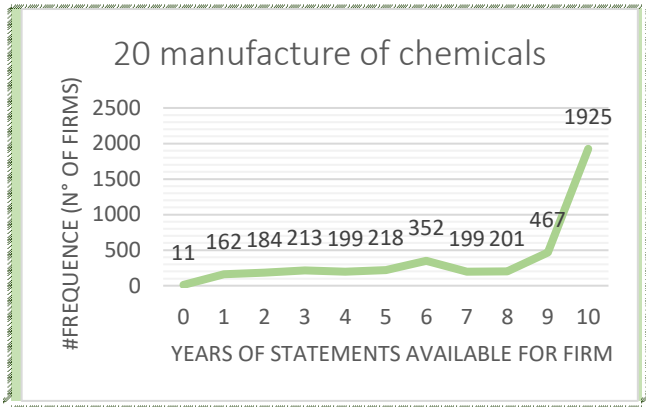

15 manufacture of leather goods and the like
#FREQUENCE (N° OF FIRMS)
3000
2000
1000
0
22 430 491 418 468 494 597 376 386 668 2126
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

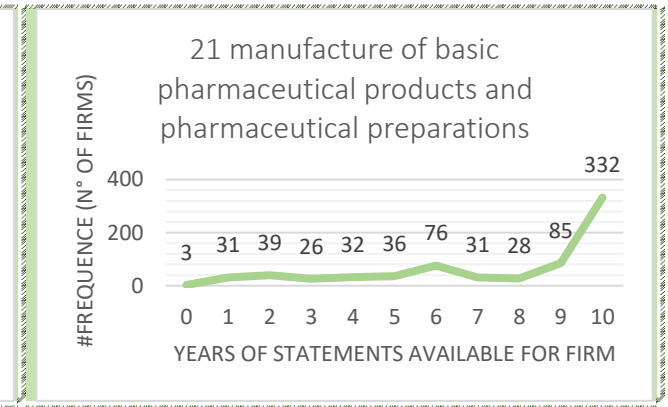

16 wood and wood and cork products industry (excluding furniture); manufacture of articles of straw and plaiting materials
#FREQUENCE (N° OF FIRMS)
2000
1000
0
8 377 393 406 439 431 576 371 376 699 1566
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

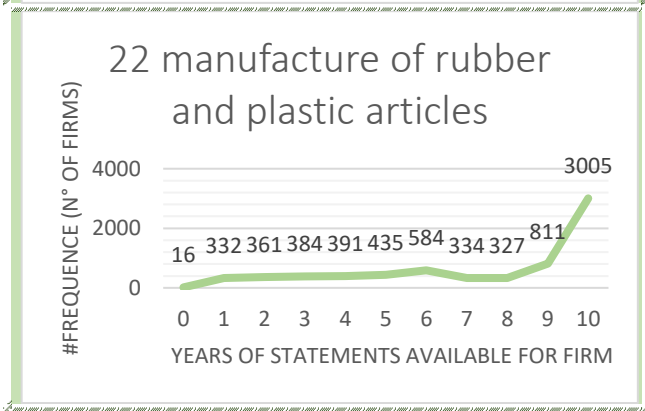

17 manufacture of paper and paper products
#FREQUENCE (N° OF FIRMS)
1500
1000
500
0
10 121 105 122 160 150 215 125 140 269 1222
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

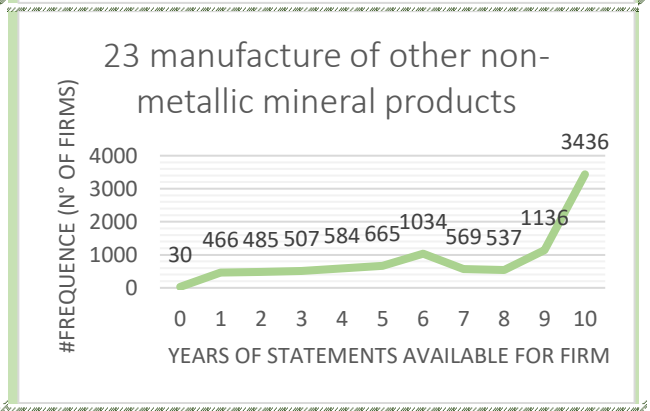

18 printing and reproduction of recorded media
#FREQUENCE (N° OF FIRMS)
2000
1500
1000
500
0
7 335 351 380 408 442 610 400 443 884 1772
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

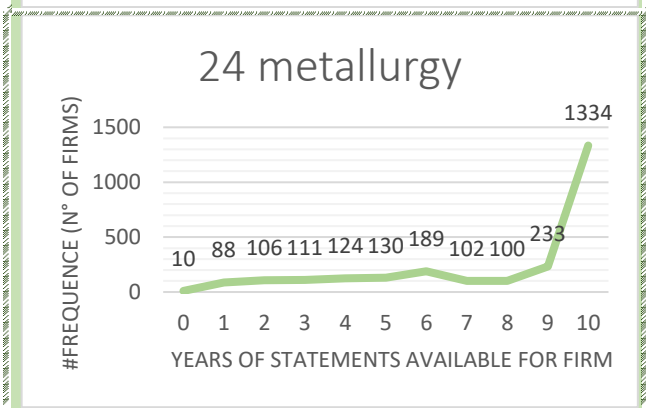

19 manufacture of coke and refined petroleum products
#FREQUENCE (N° OF FIRMS)
300
200
100
0
1 8 10 17 10 18 30 12 15 34 192
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

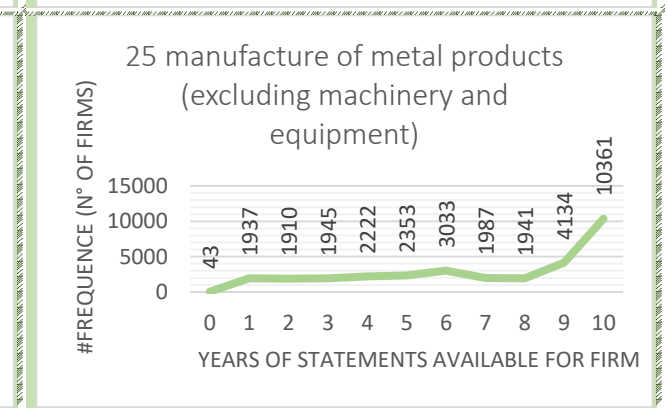

20 manufacture of chemicals
#FREQUENCE (N° OF FIRMS)
2500
2000
1500
1000
500
0
11 162 184 213 199 218 352 199 201 467 1925
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

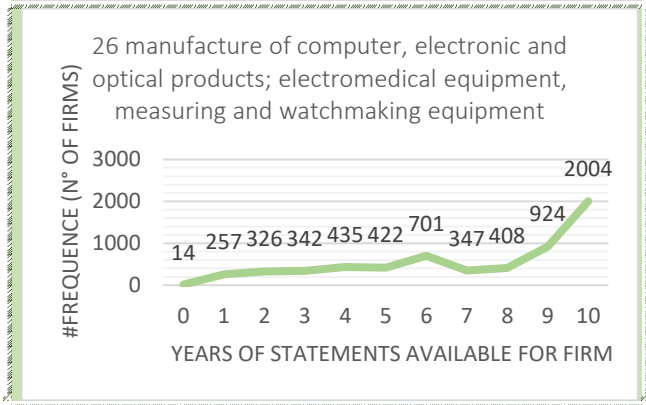

21 manufacture of basic pharmaceutical products and pharmaceutical preparations
#FREQUENCE (N° OF FIRMS)
400
200
0
3 31 39 26 32 36 76 31 28 85 332
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

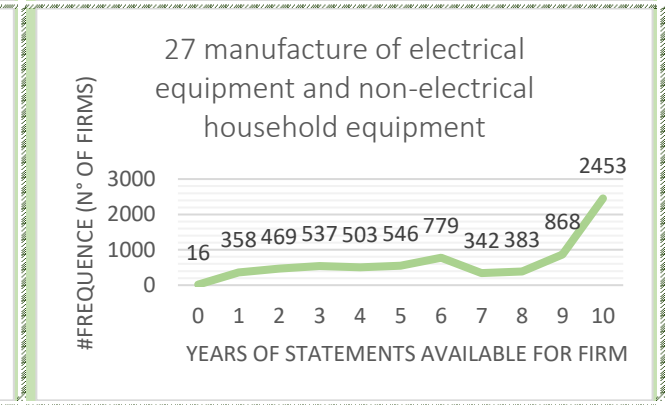

22 manufacture of rubber and plastic articles
#FREQUENCE (N° OF FIRMS)
4000
2000
0
16 332 361 384 391 435 584 334 327 811 3005
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM

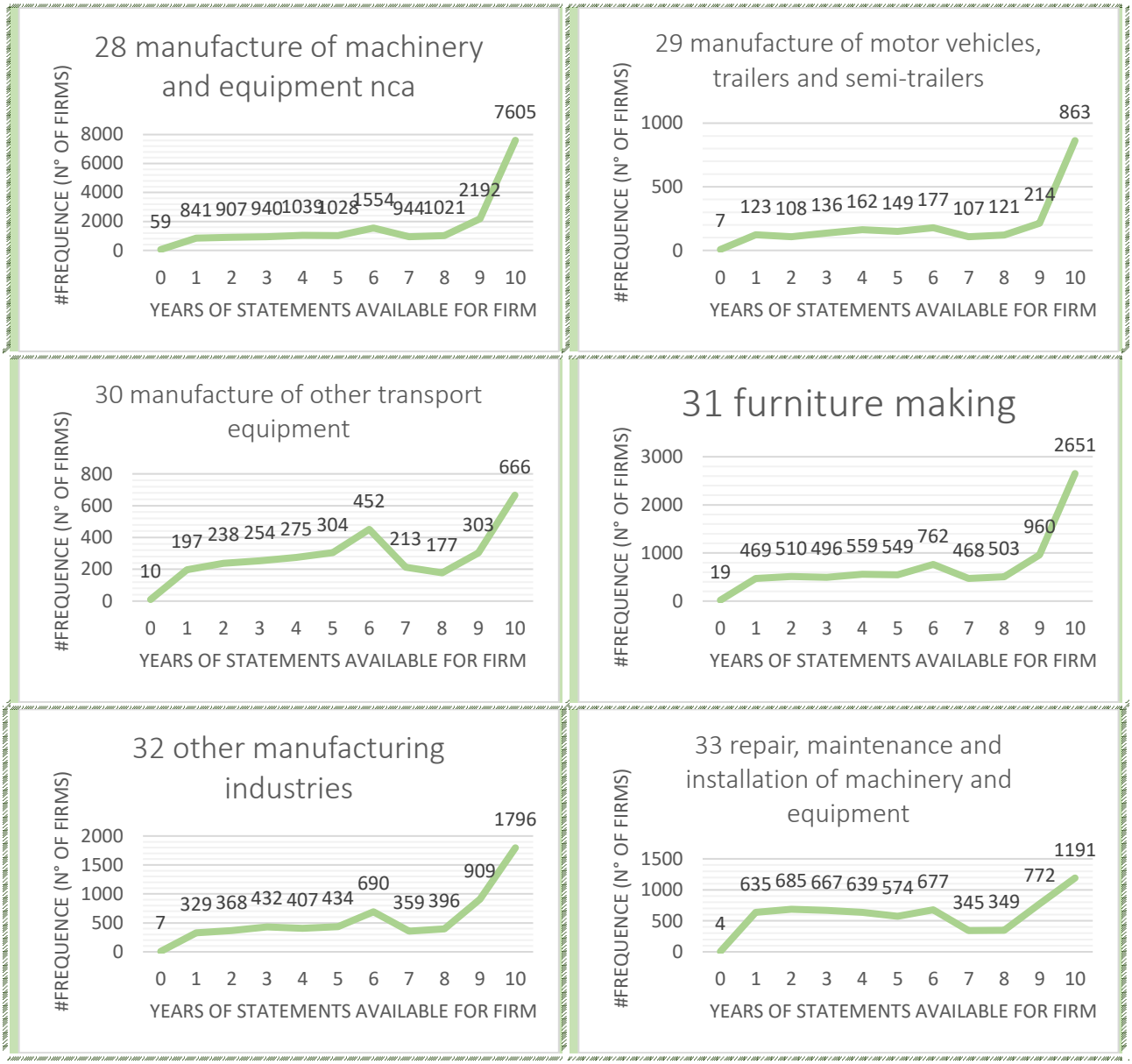

23 manufacture of other non-metallic mineral products
#FREQUENCE (N° OF FIRMS)
4000
3000
2000
1000
0
30 466 485 507 584 665 1034 569 537 1136 3436
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM


24 metallurgy
#FREQUENCE (N° OF FIRMS)
1500
1000
500
0
10 88 106 111 124 130 189 102 100 233 1334
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM


25 manufacture of metal products (excluding machinery and equipment)
#FREQUENCE (N° OF FIRMS)
15000
10000
5000
0
43 1937 1910 1945 2222 2353 3033 1987 1941 4134 10361
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM


26 manufacture of computer, electronic and optical products; electromedical equipment, measuring and watchmaking equipment
#FREQUENCE (N° OF FIRMS)
3000
2000
1000
0
14 257 326 342 435 422 701 347 408 924 2004
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM


27 manufacture of electrical equipment and non-electrical household equipment
#FREQUENCE (N° OF FIRMS)
3000
2000
1000
0
16 358 469 537 503 546 779 342 383 868 2453
0 1 2 3 4 5 6 7 8 9 10
YEARS OF STATEMENTS AVAILABLE FOR FIRM



These graphs are of a crucial importance since they give a first clear understanding of the degree of completeness of the data. The horizontal axis presents the numbers from 0 to 10, that means: "considering that there are 15 years of analysis and that AIDA give at most 10 years of observations for firm, how many years of observation does the company i have?" and consequently, the vertical axis shows the absolute frequencies.

a) For sure all the entities with zero statements will be deleted, they don't give any relevant information.
b) In this graphs it is not shown if the entities with 1-2-….-10 financial statements have their statements in consecutive time interval or if across the years there are void information (i.e. a firm with only 2 financial statements could have the 2000 and 2001 -> consecutive; but could also have 2005 and 2010 -> not consecutive). However, this is a point that will be considered better in the following chapter but for sure the entities that provide few information or possible misleading information will be deleted (firms that present the financial statements in only two years are not robust data especially if they are not consecutive).

In few cases the presence of an entity with a number of financial statements different from 10 can be explained with an entry or exit from the market. There is not information in the whole DB that help in understanding if there is an exit, but there is a variable with the establishment date and so that could tell if the business was born after the 2000.

Data cleaning

Among all the data there are some firms/businesses that cannot be considered in the analysis, because in correspondence of them, specific information is missing. Accordingly, on the number of businesses of interest previously found a reduction is expected.

Till this moment:

$$n = 165.942\ businesses$$

Useless observations

In a detail, since the variables of interest of the regression are Y, M, L, K, it is crucial that their values must be present in the years observed for all the firms, or at least for some years. For example, the financial statements of a firm with missing information for all the 15 years of interest for the variable Y is not useful and will not be considered. The same is applied to M, L, K. The Stata commands that allow to delete the financial statements of this useless businesses are the following.

First of all, for the 4 variables of interest the missing values are substituted with a zero.

$$replace\ Total\_value\_prod = 0\ if\ \ Total\_value\_prod ==.$$

$$replace\ Value\_added = 0\ if\ \ Value\_added ==.$$

$$replace\ Tangible\_fixed\_asset = 0\ if\ \ Tangible\_fixed\_asset ==.$$

$$replace\ Cost\_of\_employees = 0\ if\ \ Cost\_of\_employees ==.$$

Equations 25

Where $Total\_value\_prod$ is the Y of the Cobb-Douglas, $Total\_value\_prod - Value\_added$ is the M of material, for the capital K is considered the $Tangible\_fixed\_asset$, and in the end for the labour variable l is used $Cost\_of\_employees$.

After these operations are done, there are many firms with few observations (1, 2, 3) over the 15 years and sometimes these few years of information are not consecutive. What problem does it lead? The financial statements with only 2-3 years of data over 15 of analysis (and sometimes not on following years) are useless for the analysis or the void year can be due to systematic error or systematic misreporting and so a bias is introduced.

So, a further action of cleaning of the dataset is applied to remove these observations that give a small contribution. All the firms without 7 consecutive years are dropped, the queries are the following in Equation 26 and are repeated for all the 4 variables of interest *.

$$gen\ presence = 1\ if\ *== 0$$

$$replace\ presence = 0\ if\ presence! = 1$$

$$egen\ sequence\_of\_7 = count(presence) \qquad if\ presence == 1\ \&\ presence[\_n-1] = \\ = 1\ \&\ presence[\_n-2] == 1\ \&\ presence[\_n-3] == 1\ \&\ presence[\_n-4] = \\ = 1\ \&\ presence[\_n-5] == 1\ \&\ presence[\_n-6] == 1\ \ ,by\ (\ Num\_CCIAA\ )$$

$$egen\ filter\_7 = count(sequence\_of\_7), by\ (Num\_CCIAA)$$

$$drop\ if\ filter\_7 == 0$$

Equations 26

The results where the * is substituted with each of the listed variable:

- $*= Total\_value\_prod$

$$(1.169.700\ observations\ deleted)$$

$$\frac{1.169.700}{15} = 77.980\ businesses$$

- $*= Cost\_of\_employees$

$$(125.970\ observations\ deleted)$$

$$\frac{125.970}{15} = 8.398\ \ businesses$$

- $*= Tangible\_fixed\_asset$

$$(19.785\ observations\ deleted)$$

$$\frac{19.785}{15} = 1.319\ businesses$$

- $*= Value\_added$

$$(0\ observations\ deleted)$$

$$\frac{0}{15} = 0\ businesses$$

The remaining firms are:

$$n = 165.942\ - 77.980 - 8.398 - 1.319 = \ 78.245\ businesses$$

Due to further problem in the day-to-day work, other 88 businesses are dropped. The firms on which the estimates are done become:

$$n = 78.157\ businesses$$

Normalization over 12 months

The financial statements are documents used by investors, market analysts, and creditors to evaluate a company's financial health and earnings potential. The three major financial statement reports are the balance sheet, income statement, and statement of cash flows.

These documents summarize the accounting data at the end of an administrative period that generally coincides with the calendar year from January 1 to December 31. The ordinary duration of twelve months may, in the presence of specific conditions, change, since the administrative period may be longer or shorter than the ordinary one.

In the light of the above, beyond the canonicals financial statements of 12 months, the following types of statements can be obtained.

- With a financial year that does not coincide with the calendar year (so-called straddling financial year).

A frequent closing date is March 31, the statements belonging to this category are always of 12 months.

- With a financial year of less than twelve months.
  A typical case is represented by the first financial year in which the company was established. Other cases may occur in conjunction with extraordinary operations, such as transformation or merger, or following an extraordinary meeting which brings forward the closure of the current financial year, setting a new deadline for future ones.
- With a financial year exceeding twelve months (multi-year).
  Contrary to the previous hypothesis, the multi-year financial statements are allowed when the date of incorporation is a few months before the statutory closing date.

The dataset is made of many Italian businesses, and so often there are the situations described above. What is the problem of having financial statement with a variable duration? The equation of the regression has a temporal dimension, but knowing the form of the dataset, it fits better to consider the time in the form of years. And so, with the situation described earlier it could be possible to find some observations in which the data refer to ½ year or 1 and ½ year. It does not make sense to compare them with the majority of the observations that last 1 year: they become outliers. Obviously, this is a problem that affect only the data of the income statement that are the result of the accounting period, while it is irrelevant for the data of the balance sheet (like K) that instead is like a snapshot in a certain period of time.

To adjust these variables the easier possible approximation is the one to linearize their amounts on 12 months.

$$gen\ propor\_time = 12/real(month\_competence)$$

Equation 27

$$replace\ income\_stat = income\_stat * propor\_time$$

Equation 28

Where $income\_stat$ is substituted by all the variables of interest of the income statement.

Latitude Longitude and Matrix of distances

With regard to the agglomeration economies, as said above in the previous comma, the intention of this study is to understand which and how many businesses belong to the surroundings of the others, in order to compute the localization and urbanization indices.

There are various ways to reach the intendent result, but all of them involves the use of a software. Here are applied two of the possible methods: the first is simply through a formula on STATA (using the CAP as geographical identifier), the other instead is through a dedicated software that works on latitude and longitude. With the second one it is possible to make an analysis at a more disaggregated level, in fact it is created a matrix of distances, in which all the businesses that are less far than 1 km are listed.

Both the alternatives are developed, but as usual the DB in its current state is not perfect, the following correction are made.

- The CAP was missing for some firms.

$$tab\ Town\ if\ CAP == ""$$

Equation 29

The results in Table 5.

Table 5

Town	#frequence
Fie allo Sciliar/Voels am Schlern	1
Jesi	1
Milano	2
Priolo Gargallo	1
Sarentino/Sarntal	1

However, for these observations, the information of Latitude and Longitude are presents, so it was made a web search on Google Maps and the CAPs were recovered.

- Completely wrong Latitude and Longitude data.
  As shown in the figure below if all the firms' latitude and longitude data are plotted on some specific software, their distribution is not only over the Italy boundaries, but there is a point also below Portugal (on the left side of the Figure 5, at the bottom).

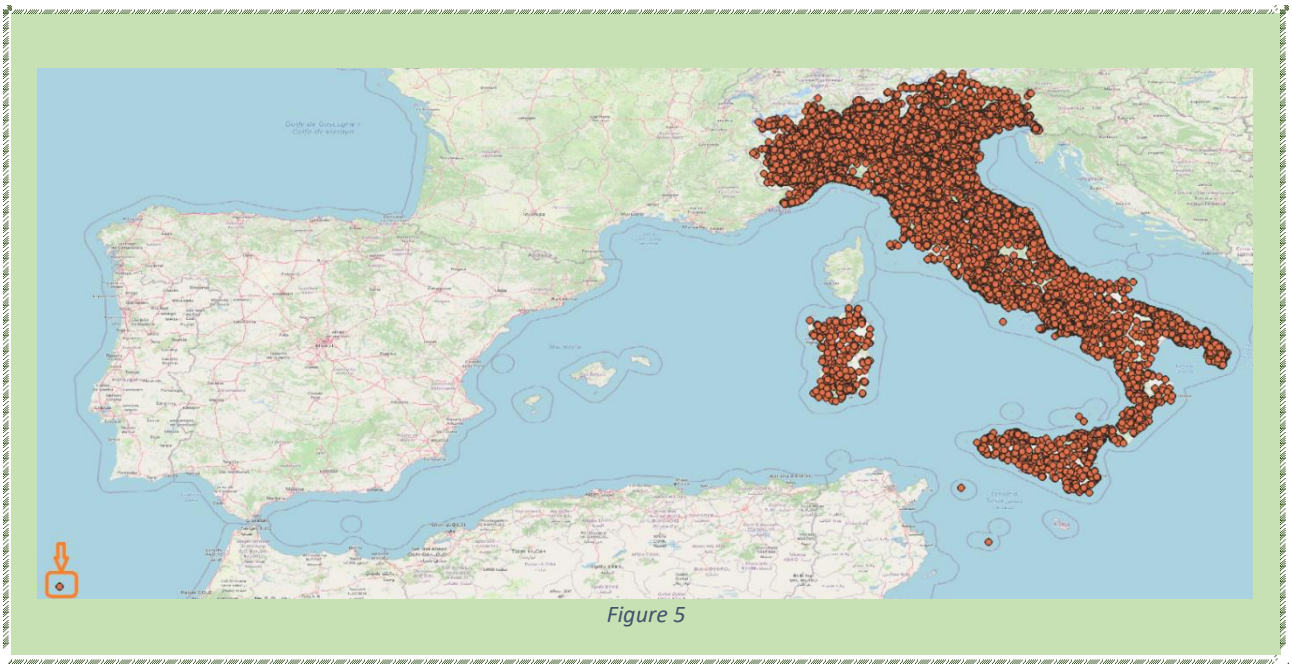


Figure 5

This situation needs a further analysis, that can be done on STATA. The investigation is easily conducted comparing the longitude of that point with the longitude of all the other point in the Italy boundaries; in fact, all the towns in Italy have longitude values that are higher than 0, while instead the incriminated point below Portugal is characterized by a negative longitude. So, to identify the points it is possible to make a query in which it is printed the ID of the firms or, as in this case, the real town of points.

$$tab\ Town\ if\ Longitude < 0$$

Equation 30

Table 6

Town	#frequence	Town	#frequence
Aielli	3	Mongrassano	1
Altomonte	2	Osidda	1

Amendolara	1	Pietrapaola	1
Arborea	4	Roccascalegna	2
Atessa	1	San Giovanni Lipioni	1
Balvano	2	San Marco Argentano	6
Banzi	1	Sant'Angelo Le Fratte	2
Canzano	3	Santa Maria del Cedro	3
Crosia	2	Santa Sofia d'Epiro	1
Gerocarne	1	Senise	2
Gessopalena	1	Ulassai	1

The sum of all these frequencies gives a total of 42 $businesses$ that presented the same wrong coordinates of Latitude and Longitude.

Similar as before, the information of Latitude and Longitude are adjusted one by one, using the address/CAP or the company name on Google Maps.

- No corrections are done in the case in which Latitude and Longitude are missing but the CAP is present.

$$tab\ Num\_CCIAA\ if\ Latitude == .\ \&\ Longiude ==.$$

Equation 31

The resulting output is:

$$\frac{12.690}{15} = 846\ businesses$$

These observations are not dropped just for one reason, that on the first method on STATA it is used the CAP to treat agglomeration economies, and these $846\ businesses$ are equipped with it.

- To better approach one of the two geographical models, it is useful to identify if a company is a later entrant or an incumbent (some businesses are founded after the 2000 and before the 2014).
  It is very easy to generate such a variable because the DB is provided with the establishment year, but it is a string and it is not written in a uniform way.
  This concept is reinforced submitting the following query.

$$gen\ lenght\_Establishment = strlen(establishment\_year)$$

Equation 32

And if the new variable is plotted, the results are shown below.

$$tab\ lenght\_Establishment$$

Equation 33

Table 7

Lenght_Establishment	#frequence
4	69
7	2
10	78.157

This help in understanding that the establishment date is not uniform and a query with the condition on that variable generate bias and errors. So, the next step is to extrapolate the year from all the observations.

$$gen\ yearEst = real(substr(establishment\_year, 7,4)\ \ if\ lenght\_establishment == 10$$

Equation 34

$$replace\ yearEst = real(substr(establishment\_year, 4,4)\ \ if\ lenght\_establishment == 7$$

Equation 35

$$replace\ yearEst = real(substr(establishment\_year, 1,4)\ \ if\ lenght\_establishment == 4$$

Equation 36

The new variable $yearEst$ could be enough, but also a dummy variable could be created, with a value of 1 if the company exist in that specific year and 0 otherwise.

$$gen\ noborn = 0\ \ if\ yearEst > Year$$

Equation 37

$$replace\ noborn = 1\ \ if\ noborn ==.$$

Equation 38

Model and variables creation

Data deflation & Price indexes collection – (Istat data)

Since the variables of the production function are expressed in money rather than in quantities, and since the time horizon of interest is of 15 years, the information in the actual state is not comparable over the years. The approach to do this is by deflating the data, removing the growth due to the inflation.

In a market economy, prices for goods and services can always change. Some prices fall while others rise. Inflation occurs if there is a broad increase in the prices of goods and services, not just of individual items; it means that there is a reduction of the market power, for €1 today it is possible to buy less than yesterday. In other words, inflation reduces the value of the currency over time.

There are various factors that can drive prices or inflation in an economy. Typically, inflation results from an increase in:

- ✓ Production costs: such as raw materials and wages. The demand for goods is unchanged while the supply of goods declines due to the higher costs of production. But as it happens with many costs, the increase of the production costs inflates the prices of the finished goods that are paid ultimately from the customers.
    - Raw materials: cause inflation since they are inputs of the production. Any rise in the commodity prices such as oil metals and copper, lead companies that use these materials to make their products to increase the prices of their goods.
    - Wages: when the economy is performing well, and the unemployment rate is low, shortages in labour or workers can occur. Hence in this case the increase in the production costs is due to the firms that increase wages to attract the best candidates. And so as before this increase in costs is suffered by the final consumers because the companies in turn rise the prices.
- ✓ Demand: as the law of the demand and supply teaches, when there is a rise in the demand for a good across an economy, the demand curve shifts upward and the intersection with the supply curve

correspond to an higher price than before; and if the supply is limited in quantity all the consumers will be willing to pay a very high price to buy the product. Sustained demand can reflect in the economy and raise costs for other goods. The increase in the demand often depends on the consumer confidence: low unemployment, rising wages, everything that lead consumers to a more spending. Economic expansion has a direct impact on the level of consumer spending in an economy, which can lead to a high demand for products and services.

The removal operation of the inflation in the prices is called deflation and is performed by dividing the monetary time series by a price index. Since inflation is a significant component of the of apparent growth in any series expressed in money, by deflation it is possible to uncover the real growth, if any.

As explained before there could be various causes of inflation and so each product or service can suffer a different amount of inflation over the year. In our Equation 6, there are four variables y_{it}, l_{it}, k_{it}, m_{it} that comes from the financial statements and so they will be characterized by inflation, but each one in different way.

Producer price index for industrial products

A company belonging to ATECO 3101 takes care of "manufacture of office and store furniture" while a company with ATECO 1394 is interested in "manufacture of twine, cordage, rope and netting", so on the whole sample of companies of interest, with ATECO between 10 – 33, each ATECO correspond to the manufacture of a different products. A detailed analysis implies to consider a price index for each ATECO code, since the production of different products does not mean same inflation prices.

The indexes collected are the one available on the Istat web site (national institute of statistics), in the section called "Industry producer prices" there is a data set divided for ATECO code. The main steps that allow to create the dataset for deflating y_{it} are listed below.

I. Starting from the fact that ATECO has 6 digits, if it is created a data indexes collection in which all the ATECO codes have their dedicated index for the 15 years, it would be a very detailed work but with many codes that share same index or with a minimal difference. Furthermore, on the Istat web site the maximum level of detail was 4 digits, so the choice fell on 4-digit ATECO codes.
II. Make a clear understanding about all the ATECO codes present in the database. The STATA commands are:

$$generate\ ateco4\ =\ int(\ real(Ateco)/100)$$

Equation 39

And after:

$$tab\ ateco4$$

Equation 40

The results tell for each ATECO with 4 digit how many firms belong to it:

Table 8

Ateco4	#frequence	Ateco4	#frequence	Ateco4	#frequence	Ateco4	#frequence
1000	102	1623	1.137	2360	231	2733	446
1010	253	1624	323	2361	632	2740	425
1011	250	1629	385	2362	20	2750	16
1012	41	1700	23	2363	468	2751	254
1013	451	1710	23	2364	12	2752	71
1020	160	1711	9	2365	16	2790	1.186
1030	234	1712	165	2369	40	2800	215

1031	6	1720	141	2370	1.319	2810	43
1032	21	1721	575	2390	2	2811	114
1039	329	1722	84	2391	121	2812	66
1040	101	1723	360	2399	301	2813	373
1041	264	1724	9	2400	35	2814	425
1042	7	1729	92	2410	166	2815	279
1050	3	1800	21	2420	196	2820	662
1051	759	1810	98	2430	89	2821	130
1052	46	1811	42	2431	15	2822	614
1060	17	1812	1.859	2432	31	2823	59
1061	274	1813	516	2433	78	2824	7
1062	1	1814	205	2434	90	2825	613
1070	107	1820	41	2440	56	2829	1.810
1071	592	1900	1	2441	43	2830	597
1072	185	1910	7	2442	118	2840	800
1073	270	1920	197	2443	20	2841	117
1080	6	2000	90	2444	20	2849	237
1081	8	2010	131	2445	36	2890	127
1082	125	2011	57	2446	1	2891	122
1083	250	2012	64	2450	168	2892	338
1084	75	2013	36	2451	115	2893	630
1085	59	2014	39	2452	18	2894	495
1086	56	2015	100	2453	128	2895	89
1089	104	2016	232	2454	100	2896	183
1090	54	2017	19	2500	253	2899	523
1091	152	2020	29	2510	286	2900	5
1092	18	2030	415	2511	2.950	2910	135
1100	16	2040	18	2512	1.287	2920	302
1101	157	2041	162	2520	12	2930	19
1102	523	2042	342	2521	119	2931	81
1104	2	2050	18	2529	136	2932	521
1105	21	2051	25	2530	27	3000	3
1106	3	2052	70	2540	41	3010	135
1107	128	2053	36	2550	899	3011	180
1200	18	2059	399	2560	41	3012	290
1300	73	2060	31	2561	1.256	3020	58
1310	602	2100	71	2562	4.796	3030	77
1320	996	2110	90	2570	12	3090	5
1330	552	2120	229	2571	30	3091	105
1390	17	2200	22	2572	140	3092	133
1391	87	2210	65	2573	1.266	3099	16
1392	513	2211	81	2590	142	3100	1.716
1393	30	2219	387	2591	20	3101	591
1394	47	2220	1.834	2592	74	3102	126
1395	57	2221	316	2593	284	3103	131
1396	246	2222	437	2594	154	3109	1.112
1399	149	2223	228	2599	1.515	3210	3

1400	13	2229	479	2600	133	3211	3
1410	1.291	2300	42	2610	13	3212	892
1411	102	2310	132	2611	429	3213	97
1412	74	2311	16	2612	46	3220	44
1413	876	2312	318	2620	602	3230	136
1414	253	2313	23	2630	428	3240	146
1419	361	2314	11	2640	97	3250	662
1420	75	2319	132	2650	1	3290	6
1430	222	2320	56	2651	520	3291	60
1431	172	2330	2	2652	43	3299	756
1439	410	2331	240	2660	367	3300	1
1500	3	2332	164	2670	101	3310	1
1510	1	2340	25	2680	24	3311	24
1511	692	2341	136	2700	180	3312	630
1512	494	2342	58	2710	35	3313	115
1520	1.604	2343	3	2711	425	3314	41
1600	25	2344	16	2712	168	3315	131
1610	490	2349	2	2720	37	3316	10
1620	9	2350	7	2730	1	3317	32
1621	126	2351	34	2731	1	3319	7
1622	9	2352	53	2732	135	3320	1.088

As always it is remarked that the sample is not changed and so:

$$\sum \#frequences = n = 78.157\ businesses$$

Equation 41

III. Knowing all the ATECO codes on the dataset, it is possible to collect the indexes. As an example, the Istat web site provide the information below. Monthly indexes from January 2000 to December 2015.

Table 9

1310	Jan-2000	Feb-2000	Mar-2000	Apr-2000	May-2000	Jun-2000	Jul-2000	Aug-2000	Sep-2000	Oct-2000	Nov-2000	Dec-2000
Total	93.7	93.8	94.3	94.9	95.4	95.3	95.3	95.4	96.6	96.9	97.2	97.3

1310	Jan-2001	Feb-2001	Mar-2001	Apr-2001	May-2001	Jun-2001	Jul-2001	Aug-2001	Sep-2001	Oct-2001	Nov-2001	Dec-2001
Total	97.6	98.1	98.3	98.3	98.3	98.2	98.2	98.6	98.2	97.9	97.5	97.8

1310	Jan-2002	Feb-2002	Mar-2002	Apr-2002	May-2002	Jun-2002	Jul-2002	Aug-2002	Sep-2002	Oct-2002	Nov-2002	Dec-2002
Total	97.2	97.3	97.1	97.1	97.1	96.8	96.9	97	95.6	95.5	95.6	95.4

1310	Jan-2003	Feb-2003	Mar-2003	Apr-2003	May-2003	Jun-2003	Jul-2003	Aug-2003	Sep-2003	Oct-2003	Nov-2003	Dec-2003
Total	95.4	95.7	94.9	95.7	95.7	95.8	95.6	96.5	95.7	96.2	96.2	95.6

1310	Jan-2004	Feb-2004	Mar-2004	Apr-2004	May-2004	Jun-2004	Jul-2004	Aug-2004	Sep-2004	Oct-2004	Nov-2004	Dec-2004
Total	96	95.6	95.4	95.3	95.7	95.7	95.9	96	95.8	95.7	95.2	94.9

1310	Jan-2005	Feb-2005	Mar-2005	Apr-2005	May-2005	Jun-2005	Jul-2005	Aug-2005	Sep-2005	Oct-2005	Nov-2005	Dec-2005
Total	94.6	94.4	94.3	94.4	94.8	95.1	95	95.1	95.1	95.3	95.3	95.4

1310	Jan-2006	Feb-2006	Mar-2006	Apr-2006	May-2006	Jun-2006	Jul-2006	Aug-2006	Sep-2006	Oct-2006	Nov-2006	Dec-2006
Total	95.8	95.9	95.6	95.8	95.4	95.5	95.7	95.7	96	96	95.8	95.8

1310	Jan-2007	Feb-2007	Mar-2007	Apr-2007	May-2007	Jun-2007	Jul-2007	Aug-2007	Sep-2007	Oct-2007	Nov-2007	Dec-2007
Total	95.4	94.8	95.4	96.7	96.2	96.8	96.2	95.7	97.1	97	98.1	96.7

1310	Jan-2008	Feb-2008	Mar-2008	Apr-2008	May-2008	Jun-2008	Jul-2008	Aug-2008	Sep-2008	Oct-2008	Nov-2008	Dec-2008
Total	97.4	98	98	98	98.3	98.3	97.9	98	98.6	98.6	100.2	100.1

1310	Jan-2009	Feb-2009	Mar-2009	Apr-2009	May-2009	Jun-2009	Jul-2009	Aug-2009	Sep-2009	Oct-2009	Nov-2009	Dec-2009
Total	98.8	99.2	97.8	99.8	98.9	99.1	98.8	98.5	97.7	97.9	98	97.4

1310	Jan-2010	Feb-2010	Mar-2010	Apr-2010	May-2010	Jun-2010	Jul-2010	Aug-2010	Sep-2010	Oct-2010	Nov-2010	Dec-2010
Total	97.7	98.7	99	98.9	99.2	99.5	99.7	100.8	101	101.4	101.8	102.5

1310	Jan-2011	Feb-2011	Mar-2011	Apr-2011	May-2011	Jun-2011	Jul-2011	Aug-2011	Sep-2011	Oct-2011	Nov-2011	Dec-2011
Total	104	105.9	108.1	109.4	110.3	110.6	111.1	111.4	111.7	112.7	112.5	112.6

1310	Jan-2012	Feb-2012	Mar-2012	Apr-2012	May-2012	Jun-2012	Jul-2012	Aug-2012	Sep-2012	Oct-2012	Nov-2012	Dec-2012
Total	112.9	112.4	112.7	112.6	113	113.1	113	112.6	112.5	111.7	111.2	110.5

1310	Jan-2013	Feb-2013	Mar-2013	Apr-2013	May-2013	Jun-2013	Jul-2013	Aug-2013	Sep-2013	Oct-2013	Nov-2013	Dec-2013
Total	111.1	111.1	111.4	111.4	111.9	111.9	111.5	112.4	112.5	112.5	112.3	112.4

1310	Jan-2014	Feb-2014	Mar-2014	Apr-2014	May-2014	Jun-2014	Jul-2014	Aug-2014	Sep-2014	Oct-2014	Nov-2014	Dec-2014
Total	112.9	113.9	113.7	113.4	113.5	113.4	113.3	113.4	113.4	113	113.4	112.6

How are these indexes computed? They are called base 2010, and it means that 2010 has an index of 100, it is the year of reference, while all the other months presents an index number such that if it is higher than 100 that year was with more inflation that the 2010 and vice versa. The farer it is from 100 the higher the inflation/deflation was. The formula is shown below.

$$index_{it} = \frac{Level\ of\ prices\ of\ the\ production\ in\ the\ year\ t\ for\ ATECO\ i}{Level\ of\ prices\ of\ the\ production\ in\ the\ year\ 2010\ for\ ATECO\ i} * 100$$

Equation 42

IV. The monthly indexes are used, and its average become the index of the corresponding year.

Table 10

ATECO	Year	Total
1310	2000	95.51
1310	2001	98.08
1310	2002	96.55
1310	2003	95.75
1310	2004	95.60
1310	2005	94.90
1310	2006	95.75
1310	2007	96.34
1310	2008	98.45
1310	2009	98.49
1310	2010	100.02
1310	2011	110.03
1310	2012	112.35
1310	2013	111.87
1310	2014	113.33

V. Point III and IV are repeated for all the 4-digit ATECO found at point II.

Labour and wages

Regarding the variable labour l_{it} of equation 6, it can sound strange to adjust for inflation since it represents the number of employees. This is clearly true, but in the great majority of the businesses the information of employees was missing for many years. So, there were two alternatives:

- ✓ Estimate the variable employee for the years in which it is missing. How? Using the variable "total employee costs" (that is rarely missing), and the variable "employee". Two conditions apply in order to estimate employee:
  - The variable employee is available for at least one year over the 15 of the horizons.
  - There is at least one year over the 15 in which the two variables "total employee costs" and "employee" are both available.

  If one company does not support these conditions cannot be considered, the variable "employee" will never be estimated.

  The steps for the estimation are the following.

  - Compute an "average cost for employee" that change across the various firms but is the same for the 15 years of the analysis.
  - Estimate the "employee" number simply by making the ratio of the "total employee costs" (that is different in the various years) over "average cost for employee" (that is fixed for every firms).
- ✓ Using directly the "total employee costs", as it is done in this thesis.

In both cases a monetary variable is interested and so the "total employee costs" must be deflated.
As before the source of the information is the Istat web site, and the indexes that better fit the work are hourly contractual wage index, that is available with a detail of 3 digits ATECO; but unfortunately, it is not

possible to implement it, since for these indexes the data collected start from the 2005, while the dataset has from 2000.
The Istat site has also other indexes available that differ for grouping criteria or other small feature but that refers always to wages, so the one selected is labour cost index per Ula (Ula means "unità di lavoro annuale" annual work units); there is only one collection, it does not take into account the ATECO codes, but this is not a problem, the approximation is very small.

Table 11

ATECO	Year	Labor cost index
1012	2000	72.25
1012	2001	74.18
1012	2002	76.35
1012	2003	78.75
1012	2004	82.08
1012	2005	84.50
1012	2006	87.63
1012	2007	89.93
1012	2008	93.90
1012	2009	96.10
1012	2010	100.00
1012	2011	102.93
1012	2012	105.50
1012	2013	108.18
1012	2014	110.13

Capital goods & Intermediate goods

The third and fourth variables are the capital k_{it} and the raw/intermediate materials m_{it} both in log form.

As a short review K_{it} means those goods that a company acquires for multi-year use, as they contribute to the business for a period longer than the financial year. Included in this specific category is a building or even a piece of machinery, such as a computer or other accessories useful for the development of the business itself; while M_{it} stands for economic goods that can only be used in a production cycle to produce other goods, for many businesses in the ATECO groups 10-33 (all manufacturers) these materials are the same.
These variables according to the descriptions do not seem to be highly correlated with the ATECO code, and so since the analysis regard only manufacturer businesses a unique general index as for wages and labour is safely acceptable.

Table 12

ATECO	Year	Capital goods	Intermediate goods
1012	2000	89.62	84.58
1012	2001	90.94	85.66
1012	2002	91.52	85.49
1012	2003	91.71	85.93
1012	2004	92.66	89.19
1012	2005	93.80	91.23

1012	2006	95.53	94.95
1012	2007	98.02	98.97
1012	2008	100.00	102.06
1012	2009	99.87	96.63
1012	2010	100.00	99.99
1012	2011	101.56	104.96
1012	2012	102.30	105.55
1012	2013	102.56	104.91
1012	2014	103.00	104.33

Approximations

The previous description about the procedures on the Istat indexes describe an ideal situation and so does not take into account all the missing information in the tables downloaded by the web site. The missing information leads to problems, or to incomplete work; and that is the point, in this paragraph it is going to describe the decision taken to overcome the stalemate.
Summarizing, at least two situations in which a decision should be taken are found, and the decision then is repeated all the time the same problem comes out.

1. Monthly absences of the index, this problem was the more frequent.
   There are some indexes for which the data was missing for the twelve months of certain years.

Table 13

1310	Jan-2000	Feb-2000	Mar-2000	Apr-2000	May-2000	Jun-2000	Jul-2000	Aug-2000	Sep-2000	Oct-2000	Nov-2000	Dec-2000
Total	.	.	.	.	.	.	.	.	.	.	.	.

   Problem: the corresponding year for that ATECO code is with no index.
   Decision: in this case it was meaningless to assign the same index of the following/previous year, while to make the average of the previous and the following years was not always possible because the majority of the times it was missing the 2000 and 2001 at the same time, and the 2000/2014 does not have a predecessor/successor. So, these possibilities to internally correct the DB were abandoned.
   To solve this problem, the practise applied was to download the DB of the same period of time, but with the ATECO code hierarchically superior. In the final solution there are many indexes computed from the DB of reference is of 3-digit ATECO code.

2. ATECO codes present on the DB but absent on the Istat web site.
   This was a strange problem since the codes in the Table 14 are recorded as ATECO codes in the DB, but they do not correspond to anyone real ATECO code. After a first look it seems that they are rounded to the 2/3 digit.
   There are two ways to list all these codes: create a full indexes collection just with the real ATECO codes, to join it with the DB and after to print the firms without match; or the second alternative is to print all the ATECO codes at the beginning and to check one by one with the real list.
   The query for the first alternative is the following:

$$tab\ ateco4\ if\ \_merge == no\_match$$

   The result is in the table.

Table 14

ATECO	#frequence	ATECO	#frequence	ATECO	#frequence	ATECO	#frequence
1000	102	1620	9	2340	25	2710	35
1010	253	1700	23	2350	7	2730	1

1030	234	1710	23	2360	231	2750	16
1040	101	1720	141	2390	2	2800	215
1050	3	1800	21	2400	35	2810	43
1060	17	1810	98	2430	89	2820	662
1070	107	1900	1	2440	56	2840	800
1080	6	2000	90	2450	168	2890	127
1090	54	2010	131	2500	253	2900	5
1100	16	2040	18	2510	286	2930	19
1300	73	2050	18	2520	12	3000	3
1390	17	2100	71	2560	41	3010	135
1400	13	2200	22	2570	12	3090	5
1410	1291	2210	65	2590	142	3100	1716
1430	222	2220	1834	2600	133	3210	3
1500	3	2300	42	2610	13	3290	6
1510	1	2310	132	2650	1	3300	1
1600	25	2330	2	2700	180	3310	1

For a total of:

$$\sum \#frequences = 10.758\ businesses$$

Equation 43

Problem: there are no index available for these codes in the table.
Decision: probably this rounded code is due to the accountant or for following conversion of all the whole ATECO codes in the version of the 2007, so it does not seem to be a dangerous approximation to solve as before. For each code in the table are taken the indexes of the ATECO hierarchically superior.

Distances

As anticipated, there are two methods to treat with the geographical analysis, and here it is provided a description of how they are reached. It is computed the number of neighbours that in the following chapter will be used to determine the indexes of localization and urbanization economies.

1. The first method is implemented simply by making a command on Stata. They are created eight variables called n_firms* whose value is equal to: "*the number of similar firms in the same year with the same CAP*".
   These eight variables are especially necessary to make some comparisons as their difference is very subtle, and as said in previous chapters the best aggregation level is always a question mark. In detail, they have been constructed as follows.
   The first four are the output of:
   1) $bysort\ ateco4\ CAP\ Anno: egen\ n\_firms4\_NA = count(Num\_CCIAA)$
   2) $bysort\ ateco3\ CAP\ Anno: egen\ n\_firms3\_NA = count(Num\_CCIAA)$
   3) $bysort\ ateco2\ CAP\ Anno: egen\ n\_firms2\_NA = count(Num\_CCIAA)$
   4) $bysort\ CAP\ Anno: egen\ n\_firmstot\_NA = count(Num\_CCIAA)$

Equations 44

Their difference is in the aggregation level of the ATECO code, it is used ATECO code at 4, 3 and 2 digits for the first three equation, while in the fourth it is calculated the number of neighbour firms independently on the ATECO code.
The second four, instead are calculated at the same way but with a if condition that discriminate if a firm is entered in the market.

5) $bysort\ ateco4\ CAP\ Anno: egen\ n\_firms4\_A = count(Num\_CCIAA)\ if\ noborn == 1$
6) $bysort\ ateco3\ CAP\ Anno: egen\ n\_firms3\_A = count(Num\_CCIAA)\ if\ noborn == 1$
7) $bysort\ ateco2\ CAP\ Anno: egen\ n\_firms2\_A = count(Num\_CCIAA)\ if\ noborn == 1$
8) $bysort\ CAP\ Anno: egen\ n\_firmstot\_A = count(Num\_CCIAA)\ if\ noborn == 1$

Equations 45

So, if a firm is established in the 2008, it is not taken into account in the counts until that year.

2. The second method is characterized not only by the use of STATA, since they are involved also ArcMap and Excel. ArcMap is a suite of ArcGIS that allow to manipulate geospatial data coming from a dataset. In this case it is exploited its capabilities to compute pairwise distances between all firms. The steps needed to apply this method are the following:
    a. Starting from three variables of the DB, $Latitude$, $Longitude$, and Num_CCIAA; they are exported for each firm, from a .dta version to an excel file.
    b. The excel file is imported in ArcMap and the data of each firm are converted into points. These points are featured with $Latitude$ and $Longitude$ and are printed on the map replicating their original position in Italy.
    c. On ArcMap it is used the function of proximity called "*Point Distance*".

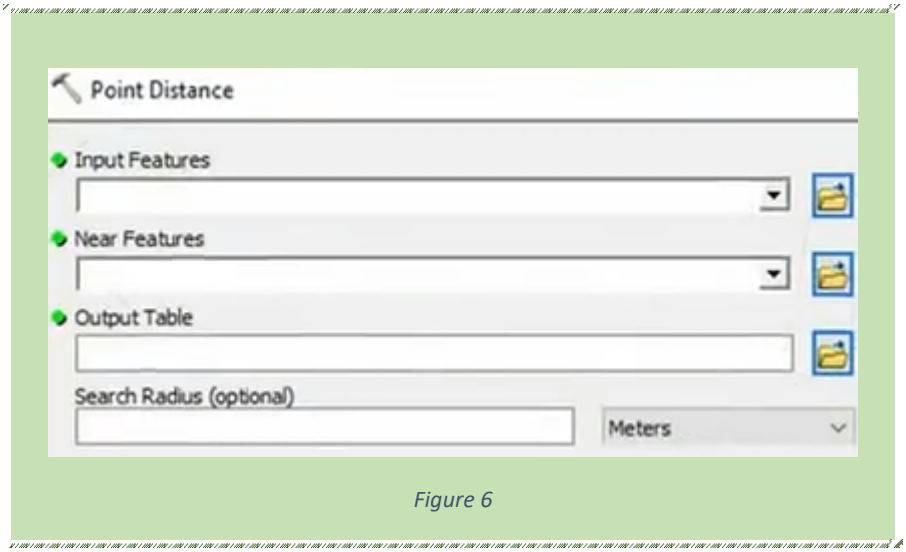


Figure 6

This function allows to compute the distances between two datasets: the "*Input Features*" and the "*Near Feature*", within a specified "*Search Radius*". For the aim of this thesis, the Input Features and the Near Feature are filled with the same dataset, the one coming from the Excel file, and the Search Radius is set to 1 km.
As an example, suppose there are only 3 firms in the dataset (each one identified with a Num_CCIAA): $CCIAA\_1$, $CCIAA\_2$, $CCIAA\_3$, and since it is put at the same time as Input and Near Feature, ArcMap allow to compute the distances in this way:

Table 15

Input Feature	Distance	Near Feature
CCIAA_1	0.84	CCIAA_2
CCIAA_1	0.29	CCIAA_3
CCIAA_2	0.84	CCIAA_1
CCIAA_2	0.06	CCIAA_3
CCIAA_3	0.29	CCIAA_1
CCIAA_3	0.06	CCIAA_2

d. The output of the Point Distance is a list of about 4 million rows and three columns (Input Feature, Near Feature and distance among them), Input and Near Feature are valued with the Num_CCIAA and it is exported into an excel file.

e. The excel file is imported in Stata. To reach the final aim it is needed to make some manipulation and so various intermediate database are created.
In fact, the output of ArcMap keep the number of observations, but additional variables are needed: ATECO codes, years and the establishment year, for both *Input Feature* and *Near Feature* (from now on identified as $Firm\_i$ and $Firm\_j$).

Year	Firm_i	Ateco4_i	Ateco3_i	Ateco2_i	Establishment_i
Distance	Firm_j	Ateco4_j	Ateco3_j	Ateco2_j	Establishment_j

These table can be obtained simply through various merges and drops starting from the complete database.

f. As in the previous method, 8 variables are created, with the same logic, the difference is that the geographic base is $Latitude$ and $Longitude$ and the area in which the other firms are considered neighbour is 1 km.
The commands used are the following.

1) Creation of the support counter:
$$gen\ counter4 = 1\ if\ Ateco4\_i ==\ Ateco4\_j\ \&\ establishment\_i \leq\ year\ \&\ establishment\_j \leq\ year$$
And after:
$$egen\ neighbourateco4\_A = sum(counter4), by(Firmi\ year)$$

2) Creation of the support counter:
$$gen\ counter3 = 1\ if\ Ateco3\_i ==\ Ateco3\_j\ \&\ establishment\_i \leq\ year\ \&\ establishment\_j \leq\ year$$
And:
$$egen\ neighbourateco3\_A = sum(counter3), by(Firmi\ year)$$

3) Creation of the support counter:
$$gen\ counter2 = 1\ if\ Ateco2\_i ==\ Ateco2\_j\ \&\ establishment\_i \leq\ year\ \&\ establishment\_j \leq\ year$$
And:
$$egen\ neighbourateco2\_A = sum(counter2), by(Firmi\ year)$$

4) Creation of the support counter:
$$gen\ counter = 1\ if\ establishment\_i \leq\ year\ \&\ establishment\_j \leq\ year$$
And:
$$egen\ neighbourtot\_A = sum(counter), by(Firmi\ year)$$

5) Creation of the support counter:
$$gen\ counter4 = 1\ if\ Ateco4\_i ==\ Ateco4\_j$$

And after:

$$egen\ neighbourateco4\_NA = sum(counter4), by(Firmi\ year)$$

6) Creation of the support counter:

$$gen\ counter3 = 1\ if\ Ateco3\_i == Ateco3\_j$$

And:

$$egen\ neighbourateco3\_NA = sum(counter3), by(Firmi\ year)$$

7) Creation of the support counter:

$$gen\ counter2 = 1\ if\ Ateco2\_i == Ateco2\_j$$

And:

$$egen\ neighbourateco2\_NA = sum(counter2), by(Firmi\ year)$$

8) Creation of the support counter:

$$gen\ counter = 1$$

And:

$$egen\ neighbourtot\_NA = sum(counter), by(Firmi\ year)$$

The first four only consider companies that have already been established in previous years.

Variable used

Production function

For the productivity function the variable are created with these queries.

$$y = \frac{Total\_value\_prod}{Producer\ price\ index\ for\ industrial\ products}$$

$$m = \frac{Total\_value\_prod - Value\_added}{Intermediate\ goods\ price\ index}$$

$$l = \frac{Cost\_of\_employees}{Labour\ and\ wages\ price\ index}$$

$$k = \frac{Tangible\_fixed\_asset}{Capital\ goods\ price\ index}$$

They are simply deflated by the corresponding industry level indices.

External agglomeration variables

Starting from the 16 variables created previously to detect geographical relationships (number of similar firms in the same area), for each of them are created the indices that will be used in the regressions analysis. These new variables are computed in the following way.

- Localization for the firm i is considered as the number of neighbours in the same industry s (ATECO 2, 3 or 4 digits) and in the same location l:

  $$Localization_{islt} = \ln(n\_firms\%_{islt})$$

  Or:

  $$Localization_{islt} = \ln(neighbourateco\%_{islt} + 1)$$

  If this variable is valued with 1 means that there are no neighbour, the firm i is the only one in the area l.

- Urbanization, instead for the business i is computed as the number of firms in different industry than s in the same are l in which the firm i is located.

$$Urbanization_{ilt} = \ln(n\_firmstot\%_{lt} - n\_firms\%_{islt} + 1)$$

Or:

$$Urbanization_{ilt} = \ln(neighbourtot\%_{ilt} - neighbourateco\%_{islt} + 1)$$

- Industrial diversity: it is a measure of industrial diversity faced by firm i operating in industry s in the area l. In order to compute it, the Herfindahl–Hirschman Index (common measure of market concentration and is used to determine market competitiveness) is calculated, in which way?
  Using excel it is computed for each area these fractions:

$$\left(\frac{n\_firms2\_NA_{ls_i}}{n\_firmstot\_NA_l}\right)^2 = \left(\frac{number\ of\ firm\ in\ area\ l\ with\ ATECO\ s_i}{total\ number\ of\ firm\ in\ l}\right)^2$$

  They are a sort of squared market shares based on the number of firms in the sector s_i in the area l. The further step is the one to sum for all the s_i in the area l. This variable is common for each firm i in the area l.
  So, it is possible to summarize with the following equation:

$$HH_{isl} = \sum_s \left(\frac{n\_firms2\_NA_{ls}}{n\_firmstot\_NA_l}\right)^2$$

$$Industrial\_diversity_{ils} = \frac{1}{HH_{isl}}$$

  This $Industrial\_diversity_{ils}$ is computed only considering the neighbour with ATECO code at two digits.

Differently from Castellani and Lavoratori, in this thesis these indexes are firm specific, since each firm has a different number of neighbours and so they change across each i; while in the paper from which this thesis take inspiration create location specific indexes based on postcode area; a further difference is that in their work in the computation of number of neighbours they consider the full sample of firms, not only the manufacturing firms.

Results

The results of the regressions are described in the following pages. As already explained previously, the regression will be divided in two steps, the first in which it is estimated the productivity and the second instead where it is calculated the impact of the agglomeration economies.

Productivity

The estimate of the production function is done through the $levpet$ formula: the total value of the production is set as dependent variable and it is specified that this variable is not expressing the value added (it set as a default), but it represents revenues.

$$levpet\ y, free(l)\ proxy\ (m)\ capital\ (k)\ revenue$$

Equation 46

The free variable is validated with the labour (in the sense that theoretically is chosen by the firm after that the shock in the productivity appears), as a remark, from the financial statement it is taken the total cost of workers. The material cost variable could be declared as free, but to reduce collinearity problems it is preferred to use it only as proxy. Finally, the state variable (the one you assume may change more slowly over time and especially before the productivity shock reveals itself) is the total fixed asset.

Table 16

Levinsohn-Petrin productivity estimator Dependent variable represents revenue.				Number of obs= 727814 Number of groups= 78157 Obs per group: min= 15 avg= 15.0 max= 15		
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
l	.1915741	.0014254	134.40	0.000	.1887805	.1943678
k	.0120811	.0011763	10.27	0.000	.0097756	.0143866
m	.7989092	.0029236	273.26	0.000	.793179	.8046394
Wald test of constant returns to scale: Chi2 = 0.96 (p = 0.3269).						

The resulting coefficients are showed in the table 16 and they seem to be solid, in line with what anyone can imagine, in fact the Wald test allow to accept the null hypothesis of constant returns to scale.

Now that the coefficients are ready, it is possible to estimate the productivity. It is done through this equation:

$$predict\ productivity, omega$$

Equation 47

Table 17

Variable	Obs	Mean	Std. Dev.	Min	Max
y	736,262	9.924527	1.54912	-4.721619	19.27435
l	731,209	8.242349	1.541555	-4.658711	16.2846
k	734,580	7.895665	2.045224	-4.62791	17.25675
m	738,336	9.564434	1.605117	-4.453572	19.12398
$productivity$	727,814	1.899841	2.565719	4.17e-06	2111.003

Each observation in the DB is now characterized with this new variable $productivity$, and the main statistics about the variables are available in the table 17 above. The negative minimum values of the variables listed in the fifth column point out the presence of outliers in the data (maybe other cleaning operations were needed), because it makes no sense such non positive values, hopefully thanks to a post estimation check, they belong to very few firms with respect to the whole sample.

Agglomeration economies

When the productivity values are ready, the estimates about the agglomeration economies can be done, as described above all the three model are applied: fixed effect, random effect and multilevel mixed model. In the chapter "Distances" and "Variable used" is listed how the variables of neighbours and relative localization and urbanization are created, for a total of 12 different kind of aggregation, they are all implemented, and the results are listed.

Application 1 (Localization and Urbanization based on Latitude and Longitude, establishment not considered)
The first regression is based on $Localization_{islt}$ and $Urbanization_{islt}$ variables created in this way:

- Localization and Urbanization are created with the output of ArcMap, so two firms are considered neighbour if their latitude and longitude give a distance lower than one km.
- Despite one firm is established for example in the 2004, it is considered as a neighbour also in the previous years. These indexes are constant for a firm over the 15 years of observations.
- A replication of the indices is created based on the different ATECO at 2, 3 and 4 digits.
- The variable $Year$ is used only in the multilevel mixed model (the only one that allow the presence of more layers).

The outputs of Stata are available in the exhibits at the end of the thesis, but in sum their directions are schematized in this table.

The fixed effect is not applicable since it calculates the variation within the firms and the independent variables are all constants over the 15 years for each firm.

Multilevel mixed effect and random effect give the same intuition in terms of direction of the coefficients, localization contribute in a positive way to the productivity, the contrary can be said for urbanization, age and industrial diversity. Localization is present and with a large confidence interval it can be said that is different from zero for ATECO 4 and ATECO 3, by widening the aggregation and considering similar sectors as identical (ATECO 2), its P-value tends to increase to 0.2%. Urbanization is always negative, but its P-value in those six regressions is on average 1.5%, so if is presents it has negative impact on the productivity. This last point can be led to the fact that is considered neighbour everyone in the range of 1 km, so maybe there are not many different sectors considered in such around (in fact the P-value increase passing from ATECO 4 to ATECO 2, so considering less sectors). Industrial diversity is almost always with a high P-value but however its direction is negative.

In the mixed model β_0 become zero due to the presence of the variable $Year$.

Table 18

No Entry	ATECO 2			ATECO 3			ATECO 4		
Variables and their direction	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model
Age	OMITTED	-	-	OMITTED	-	-	OMITTED	-	-
Localization		+	+		++	++		+++	+++

Urbanization		-	-		-	-		-	-
Industrial diversity		---	---		--	--		-	-
β_0		+	-		+	-		+	-
Year			+			+			+

Application 2 (Localization and Urbanization based on Latitude and Longitude, establishment considered)
The second regression is based on $Localization_{islt}$ and $Urbanization_{islt}$ variables created in this way:

- Localization and Urbanization are created with the output of ArcMap, so two firms are considered neighbour if their latitude and longitude give a distance lower than one km.
- If one firm is established for example in the 2004, it is not considered as a neighbour in the previous years. These allows the calculation of the coefficients for fixed effect model.
- A replication of the indices is created based on the different ATECO at 2, 3 and 4 digits.
- The variable $Year$ is used only in the multilevel mixed model (the only one that allow the presence of more layers).

The considerations on the results of random effect and multilevel mixed effect are the same as before, and their values are very close.

Concerning fixed effect, the results are poor since localization and urbanization have a P-value that fluctuates between 31% and 85%. However, in mean terms their impact is positive for localization and negative for urbanization, and this can confirm what said for Application 1.

The variable Industrial diversity does not provide any results since once again it is constant for each firm over the 15 years of observation.

Table 19

Entry	ATECO 2			ATECO 3			ATECO 4		
Variables and their direction	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model
Age	+	-	-	+	-	-	+	-	-
Localization	+	+	+	++	++	++	+++	+++	+++
Urbanization	-	-	-	-	-	-	-	-	-
Industrial diversity		---	---		--	--		-	-
β_0	+	+	-	+	+	-	+	+	-
Year			+			+			+

Application 3 (Localization and Urbanization based on CAP, establishment not considered)
This regression is based on $Localization_{islt}$ and $Urbanization_{islt}$ variables created in this way:

- Localization and Urbanization are created using the Italian postal code, so two firms are considered neighbour if their CAP is the same.
- Despite one firm is established for example in the 2004, it is considered as a neighbour also in the previous years. These indexes are constant for a firm over the 15 years of observations.
- A replication of the indices is created based on the different ATECO at 2, 3 and 4 digits.
- The variable $Year$ is used only in the multilevel mixed model (the only one that allow the presence of more layers).

The outputs of Stata are available in the exhibits at the end of the thesis, but in sum their directions are schematized in the table below.

As in the first application, fixed effect gives no results, the variables are all constant for each individual firms over the 15 years.

This time, also the multilevel mixed model when the aggregation variables based on the ATECO 4 are used, gives no results. During its execution Stata tries many times to find the optimum of the regression function, but in every iterations it is not reachable, a "backed up" message is printed out. This means that the function probably is convex.

Turning now to the regressions that provide some results, the possible considerations are very similar to the two previous applications.
The coefficient of localization contributes in a positive way to the productivity, vice versa for urbanization and age, these results were obtained also in application 1 and 2. Contrary than application 1, urbanization is negative with no doubt, its P-value is 0 for all the five regressions; and localization coefficient is 0 with a probability that ranges from 0 to 0.4%.
One of the main differences of the agglomeration indexes considered here is that the firms in one specific CAP are surely higher than the neighbours in one km (indexes calculated with latitude and longitude). So, this application 3 (the same is valid for the next application 4) can be seen as a regression in which the boundaries that define the groups of neighbours are expanded; each firm now meets more neighbours and more sectors.

This boundary enlargement produces an increase in absolute value of urbanization coefficient and a decrease of the impact of localization betas. This means that the benefits are reducing and maybe this is due to congestion costs and a higher competition created by the higher number of competitors or the higher costs of input resources.

Industrial diversity contributes positively to productivity enlarging the boundaries, but it is almost always with a high P-value.

Table 20

No Entry	ATECO 2			ATECO 3			ATECO 4		
Variables and their direction	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model
Age	OMITTED	-	-	OMITTED	-	-	OMITTED	-	BACKED UP
Localization		+	+		+	+		++	
Urbanization		-	-		-	-		-	
Industrial diversity		+	++		+	+		+	
β_0		+	-		+	-		+	
Year			+			+			

Application 4 (Localization and Urbanization based on CAP, establishment considered)
This regression is based on $Localization_{islt}$ and $Urbanization_{islt}$ variables created in this way:

- Localization and Urbanization are created using the Italian postal code, so two firms are considered neighbour if their CAP is the same.
- If one firm is established for example in the 2004, it is not considered as a neighbour in the previous years. These allows the calculation of the coefficients for fixed effect model.
- A replication of the indices is created based on the different ATECO at 2, 3 and 4 digits.

- The variable $Year$ is used only in the multilevel mixed model (the only one that allow the presence of more layers).

The outputs of Stata are available in the exhibits at the end of the thesis, but in sum their directions are schematized in the table below.

In this application just one regression is not available, the one based on ATECO 4 with the multilevel mixed model, as before Stata find some difficulties in reach the optimum.

Since the variables are not constant now, fixed effect can provide some results, but they are not robust, in fact they have an incredibly large standard error that led to a P-values that reach the 98% for urbanization and the 76% for localization variables. However, if the analysis is limited to the direction in mean terms, localization have a positive impact as in all the other regression.

With respect to application 2, when fixed effect is applied, it is found an increase of both the coefficients of localization and urbanization, this is in contrast with the results of random effect and multilevel mixed model.

In this application the considerations about random effect and multilevel mixed model are the same of the application 3, so a decrease of the coefficients of urbanization and localization. Localization is still positive but it is closer to the zero, while urbanization boost its negative impact. This variation is due to the enlargement of the boundaries for the consideration of the neighbours.

Table 21

Entry	ATECO 2			ATECO 3			ATECO 4		
Variables and their direction	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model	Fixed Effect	Random Effect	Multilevel Mixed Model
Age	+	-	-	+	-	-	+	-	BACKED UP
Localization	+	+	+	++	++	+	++	+++	
Urbanization	++	-	-	-	-	-	+	-	
Industrial diversity		+	+		+	+		+	
β_0	+	+	-	+	+	-	+	+	
Year			+			+			

Exploiting source of heterogeneity in firm benefits from localization economies

The common and more important result achieved so far is regarding the behaviour of localization, that is always positive, and its impact is higher if the geographical boundaries are tight, the ATECO code used is with 4 digits and finally establishment is considered.

A further investigation could be the one to exploit the potentialities of the multilevel mixed-effect model to understand if there is an heterogeneous response to such localization effect and if it change across the various firms.

The deepening is made choosing one of the previous combination of variables, and so the one that best fit the data is: Application 2 with ATECO 4 digit.

Starting from these variables, a new set of coefficients are estimated. Obviously, this time the formula of the multilevel mixed model is written with some differences, the levels are defined as CAP and firm, the second nested in the first; and just about the firm-level ($IDfirm$) it is allowed the randomness in the slope for the localization coefficient.

$$mixed\ productivity\ age\ industrial\_div\ year\ Localization4\ Urbanization4 \\ ||\ CAP:\ ||\ IDfirm:Localization4$$

Equation 48

Table 22

Variables	Coeff.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Age	-0.002565	0.000221	-11.63	0.000	-0.003	-0.00213
Localization	0.0190511	0.007083	2.69	0.007	0.005169	0.032933
Urbanization	-0.006451	0.002768	-2.33	0.02	-0.01188	-0.00103
Industrial diversity	-0.00107	0.001192	-0.9	0.37	-0.00341	0.001267
β_0	-21.58756	2.038475	-10.59	0.000	-25.5829	-17.5922
Year	0.0117316	0.001016	11.55	0.000	0.009741	0.013723
	Estimate	Std. Err.	[95% Conf. Interval]		Note:	
Random-effects Paramets Firm: Localization	0.4200994	0.0053966	0.4096543	0.4308108	Localization – ATECO 4 + Latitude and Longitude	

The results highlight a significant standard deviation of roughly 0.42 with relatively low standard error, suggesting that a component of heterogeneity does exist, some firms may benefit more than others from localization. This scenario can be due to several possible factors, including those related to firm characteristics, as well as specificities of sectors and locations. However, based only on the results of table 22, there are no intuition on how and which firms are more or less affected.

So, to add some contributions on this heterogeneity, it is possible to proceed in three ways:

- Splitting the analysis into subsamples by firm/industry/location factors.
- Introducing interaction effects between the localization variable and the variables capturing firm, industry and location characteristics.

Due to the high number of interactions, these approaches are not efficient.

➔ The solution is to predict firm-specific random parameters for localization and to model them.

In fact, based on the previous regression of table 22, the coefficients of the random slope have been predicted: for each firm and for each year. The query used on Stata is:

$$predict\ b*, reffects$$

Equation 49

Three coefficients for each level are generated, but the one in which this study is interested is $b2$ referring to the random slope of the localization analysed across the firm-level, the outliers are removed and at the end it is plotted.

The kernel density distribution of $b2$ for the year 2007 is showed in the figure 7, someone can think that there is something of strange since the mean values are below the zero instead of 0.019 as reported in table 22, but this is absolutely under control, it is the effect of removing outliers and also the boundaries of the picture has been shrunk. However, the presence of some firms below the zero means that there is heterogeneity in the effect of localization.

$$histogram\ b2\ if\ b2 < 2\ \&\ year == 2007\ \&\ b2 > -2\ \&\ b2! = 0, kdensity$$

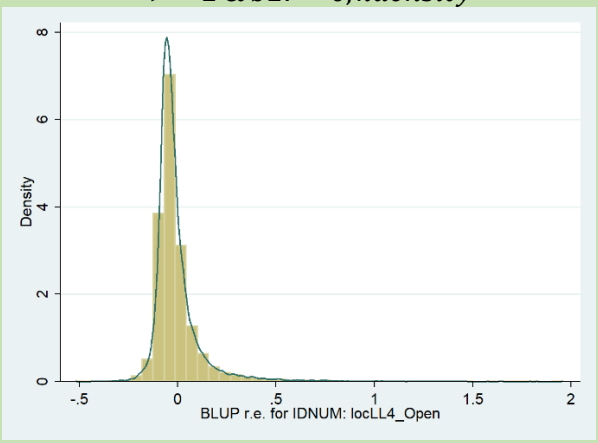


Figure 7

As mentioned earlier, there are many characteristics that can determine whether the location economy positively or negatively influences a firm's productivity, and in this study we looked at one of them: age. It was decided to investigate how the slope of the coefficient of localization economies differently vary across the firms based on their age. Two main classes are created: old firms if the age is higher than 15 years and young in the other case.

The outliers are removed, the results are in the table below.

$$ttest\ b2\ if\ year == 2007\ \&\ b2 < 10, by(old)$$

Equation 50

Table 23

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Young	39,862	0.001155	0.000531	0.105968	0.000115	0.002195
Old	38,289	-0.00283	0.000438	0.085763	-0.00369	-0.00197
Combined (Just one outlier)	78,151	-0.0008	0.000346	0.096618	-0.00148	-0.00012
Diff = Young - Old		0.003986	0.000691		0.002631	0.005341

The interpretation of this results is that young firms on average benefit more than the older. Although these values could be better investigated with further analysis, one possible interpretation is that in the early years young firms get a lot of help from the economies of scale present in their surroundings, while older now efficient firms prefer less competition.

Conclusions

The four applications implemented and discussed previously are pretty much in line with each other, and a clear picture of the influence of the agglomeration economies can be taken. Fixed effect cannot be considered a consistent estimator, the data on hand are not well fit with it.

The final result of the Italian study tells that Localization has a positive influence on the productivity of the firms and it decreases when the geographical boundaries get bigger. Another grouping principle that makes localization decreasing is when more similar sectors are considered equals and so passing from ATECO 4 to ATECO 2. The considerations about Urbanization are slightly different, since its impact is always negative and passing from the count of the neighbours in 1 km to the neighbours in the CAP its impact increases.

Why Localization affects positively?

Since having one more neighbour that belong to the same sector allows to create economies of scale in the infrastructure, on the resources and on the creation of innovative ideas. This effect is higher the nearer the firms are, so when the area of analysis is one km, firms draw all the possible benefits.

Many similar firms localized in the same area attract suppliers, allow for the partition of some investments, and many other advantages.

While when the geographic boundaries are expanded, localization benefits decrease, but a different reasoning is needed in this case: in a larger geographical area the common advantages among the neighbours in the same sectors are still present but in lower entity (the common benefits that are present among firms in one km are higher).

Finally, it was discovered that younger firms benefit more from localization than the older.

Why Urbanization affects negatively?

Regarding urbanization it can be said that in both the geographical approaches, the areas are never very large, since when the latitude and longitude data are used, a radius of one km is the aggregation way, while when the CAP is treated it is considered an area larger than one km, but always relatively small. In fact, in Italy the bigger cities have more than one CAPs, and in this analysis two different CAPs are considered two different areas, so probably the real benefit of urbanization never comes out, maybe urbanization needs a larger area to be productive. However, the results are in contrast with this last sentence, since enlarging the geographic boundaries from 1 km to the CAP urbanization become lower, probably many congestion costs arises and there are no benefits that the firm can take from firms in different manufacturing sectors.

Remember that in urbanization in this study they are counted all the neighbours in the manufacturing sector (ATECO C) but with a different ATECO code, so at the end there is no such a diversity between the various sectors of manufacturing and so many benefits of urbanization remain hidden. In the Castellani and Lavoratori work, in the urbanization index the neighbours that enter into the sum are also the ones belonging to sector different from the manufacturing one.

Localization and Urbanization coexists?

Nearly always the results suggests that they are both present simultaneously. But overall they result in positive externalities only when the one km neighbourhood is considered, localization is higher in absolute value than urbanization.

Castellani and Lavoratori

As described in the introduction Castellani and Lavoratori apply three different geographical aggregation manners for the computation of the neighbours, “Mod. 1” the one with the largest boundaries using the Postal Code, “Mod. 2” and “Mod. 3” with a square of 3x3 and 1x1 km with each firm as focal point. Out of

the sum is London since has more Postal Code, and it is implemented only in the Mod. 4. Differently than this thesis they used also two variables to discriminate against the firm size, based on the number of employees.

They start with Mod. 1 and after progressively add the indices of Mod. 2, Mod. 3 and Mod. 4. The regression is carried out with mixed effect and fixed effect (only for Mod. 1-3).

- In Mod. 1 Urbanization and Localization have both a positive impact on productivity, while for Mod. 2 and Mod. 3 Urbanization is an important factor at a higher level of geographical aggregation (Postal Code), while Localization plays a crucial role at a finer level of geographical disaggregation, in a closer neighbourhood of the firm. Urbanization decreases and becomes negative at the finest level of aggregation, while Localization increases positively at the finest level and tend to become zero at the higher level, passing through Mod. 1 till Mod. 3.
  - ➔ Firms may benefit from being located in a more specialized neighbourhood, within a diversified city. In other words, the local environment may be characterized by a combination of urbanization and specialization agglomeration economies, supporting the idea that these forces may “coexist” in the same area, but at different geographical scales.
- It is worth mentioning that these findings are robust to a FE estimation.
- Lavoratori and Castellani analysis is in line with Andersson et al. (2019)'s results in the case of Sweden.
- As in this study, they found an heterogeneous behaviour of how localization economies affect the productivity.

So, the main difference with this thesis is in the direction of the coefficients of Urbanization. The explanation to this result is due to the fact that in the work of Castellani and Lavoratori it is taken into account also the benefits deriving from the firms that do not belong to the manufacturing sector. Since in the generation of the localization and urbanization index they sum all the firms, that comes from different sectors than manufacturing, also the benefits that each firm can receive from them consequently is different. All this heterogeneity of sectors leads to positive Urbanization coefficients when the boundary considered expands.

References

Introduction

https://keydifferences.com/difference-between-economics-and-economy.html

https://arts-sciences.buffalo.edu/economics/about/what-is-economics.html

https://www.pnas.org/content/103/10/3510

https://study.com/academy/lesson/how-economic-systems-influence-social-structure.html

K. Lavoratori and D. Castellani, Too close for comfort? Micro-geography of agglomeration economies in the United Kingdom, Wiley, 2021

Production

https://www.investopedia.com/articles/04/022504.asp

https://policonomics.com/lp-information-economics1-economics-of-information/#:~:text=The%20importance%20and%20value%20of,that%20will%20report%20higher%20yields.&text=The%20study%20of%20economics%20of,agency%20theory%20or%20contract%20theory.

https://www.sciencedirect.com/topics/engineering/cobb-douglas-production-function

D. Besanko, Microeconomics, Wiley, 2020

P. Rossi, Omitted Variables, Instruments and Fixed Effects, Structural Econometrics Conference, July 2013

I. Van Beveren, TOTAL FACTOR PRODUCTIVITY ESTIMATION: A PRACTICAL REVIEW, Journal of Economic Surveys, 2010

J. Levinsohn and A. Petrin, Estimating Production Functions Using Inputs to Control for Unobservables, Review of Economic Studies, 2003

J. H. Stock and M. W. Watson, Introduction to econometrics, Pearson, 2014

Y. Lee and A. Stoyanovy and N. Zubanovz, - Olley and Pakes-style production function estimatoers with firm fixed effect, 2017 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2899248

G. Rovigliatti V. Mollisi, Theory and practice of total-factor productivity estimation: The control function approach using Stata, The Stata Journal, 2018

Teaching material, Politecnico di Torino

Agglomeration

https://www.parisschoolofeconomics.eu/IMG/pdf/Overman3-PSE-MEEDM.pdf

https://study.com/academy/lesson/what-is-agglomeration-in-economics-definition-process-theory-effects.html

https://link.springer.com/article/10.1007/s00168-019-00957-4#:~:text=In%20theory%2C%20if%20external%20benefits,to%20places%20with%20lower%20costs

K. Lavoratori and D. Castellani, Too close for comfort? Micro-geography of agglomeration economies in the United Kingdom, Wiley, 2021

T. McKillop and D. Coyle and Ed Glaeser and J. Kestenbaum and J. O'Neill, The Case for Agglomeration Economies, Manchester Independent Economic Review

Mixed effect

https://www.statstest.com/mixed-effects-model/

http://web.pdx.edu/~newsomj/mlrclass/ho_randfixd.pdf

https://ademos.people.uic.edu/Chapter17.html

https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13434#:~:text=Formally%2C%20the%20assumptions%20of%20a,effects%20versus%20covariates%20(exogeneity)%2C

https://www.youtube.com/watch?v=FCcVPsq8VcA

https://www.youtube.com/watch?v=QeCJ9ON0WDc&list=PLqRWVl_MGhBlkEU2rhPz8Wg-PJ09sWm0C&index=3

https://www.youtube.com/watch?v=QeCJ9ON0WDc&list=PLqRWVl_MGhBlkEU2rhPz8Wg-PJ09sWm0C&index=3

https://bookdown.org/steve_midway/DAR/random-effects.html

K. Lavoratori and D. Castellani, Too close for comfort? Micro-geography of agglomeration economies in the United Kingdom, Wiley, 2021

Reg

https://www.youtube.com/watch?v=fIzuQ6wBxHI

https://www.stata.com/features/overview/multilevel-mixed-effects-models/

https://www.stata.com/features/overview/linear-fixed-and-random-effects-models/

https://lost-stats.github.io/Model_Estimation/OLS/fixed_effects_in_linear_regression.html

https://journals.sagepub.com/doi/pdf/10.1177/1536867X0400400202

https://www.princeton.edu/~otorres/Panel101.pdf

https://www.youtube.com/watch?v=KALxDwwqX1A

https://bookdown.org/steve_midway/DAR/random-effects.html

https://www.youtube.com/watch?v=rUWT_EWV6QI

Exhibit

Exhibit 1 – Random Effect

Random-effects GLS regression Group variable: IDNUM				Number of obs= 727814 Number of groups= 78152 Obs per group: min= 1 avg= 9.3 max= 10		
xtreg productivity eta locLL4_noOpen urbLL4_noOpen industrial_div, re						
productivity R-sq: within = 0.0000 between = 0.0018 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0022178	.0002679	-8.28	0.000	-.0027429	-.0016928
$Localization_{islt}$	.0279885	.0057504	4.87	0.000	.016718	.039259
$Urbanization_{ilt}$	-.009559	.0035054	-2.73	0.006	-.0164295	-.0026885
$Industrial\_diversity_{ils}$	-.0013291	.0014906	-0.89	0.373	-.0042507	.0015924
_cons	1.965008	.0120678	162.83	0.000	1.941356	1.988661
sigma_u	.69679162					
sigma_e	2.4973377					
rho	.07222599	(fraction of variance due to u_i)				
					Wald chi2(4)=102.09 Prob > chi2=0.0000	
xtreg productivity eta locLL3_noOpen urbLL3_noOpen industrial_div, re						
productivity R-sq: within = 0.0000 between = 0.0017 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0022155	.0002679	-8.27	0.000	-.0027406	-.0016903
$Localization_{islt}$	.0193951	.0052292	3.71	0.000	.0091461	.0296441
$Urbanization_{ilt}$	-.0088859	.003644	-2.44	0.015	-.0160279	-.0017439
$Industrial\_diversity_{ils}$	-.0019419	.0014858	-1.31	0.191	-.0048541	.0009703
_cons	1.967584	.0119725	164.34	0.000	1.944118	1.991049
sigma_u	.69691465					
sigma_e	2.4973377					
rho	.07224965	(fraction of variance due to u_i)				
					Wald chi2(4)=91.98 Prob > chi2=0.0000	
xtreg productivity eta locLL2_noOpen urbLL2_noOpen industrial_div, re						
productivity R-sq: within = 0.0000 between = 0.0016	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	

overall = 0.0002						
age	-.0022329	.0002679	-8.33	0.000	-.002758	-.0017078
$Localization_{islt}$	.0139896	.0045879	3.05	0.002	.0049975	.0229816
$Urbanization_{ilt}$	-.0092815	.0039806	-2.33	0.020	-.0170834	-.0014796
$Industrial\_diversity_{ils}$	-.0021564	.0015057	-1.43	0.152	-.0051075	.0007948
$\_cons$	1.967217	.0118208	166.42	0.000	1.944049	1.990385
$sigma\_u$	.69697497					
$sigma\_e$	2.4973377					
rho	.07226126	(fraction of variance due to u_i)				
					Wald chi2(4)=87.36	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locLL4\_Open\ urbLL4\_Open\ industrial\_div, re$						
$productivity$ R-sq: within = 0.0000 between = 0.0018 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.002216	.0002681	-8.26	0.000	-.0027415	-.0016904
$Localization_{islt}$	.0286921	.0057921	4.95	0.000	.0173398	.0400444
$Urbanization_{ilt}$	-.0082268	.0034392	-2.39	0.017	-.0149674	-.0014861
$Industrial\_diversity_{ils}$	-.0015301	.0014844	-1.03	0.303	-.0044395	.0013792
$\_cons$	1.96271	.0119591	164.12	0.000	1.93927	1.986149
$sigma\_u$	.69679256					
$sigma\_e$	2.4973393					
rho	.07222608	(fraction of variance due to u_i)				
					Wald chi2(4)=102.36	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locLL3\_Open\ urbLL3\_Open\ industrial\_div, re$						
$productivity$ R-sq: within = 0.0000 between = 0.0017 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0022144	.0002682	-8.26	0.000	-.00274	-.0016887
$Localization_{islt}$	.0199502	.0052632	3.79	0.000	.0096345	.030266
$Urbanization_{ilt}$	-.0076183	.0035785	-2.13	0.033	-.0146321	-.0006045
$Industrial\_diversity_{ils}$	-.0021419	.0014796	-1.45	0.148	-.0050418	.000758
$\_cons$	1.965468	.0118687	165.60	0.000	1.942205	1.98873
$sigma\_u$	.69691659					
$sigma\_e$	2.4973395					
rho	.07224993	(fraction of variance due to u_i)				
					Wald chi2(4)=92.08	
					Prob > chi2=0.0000	

xtreg productivity eta locLL2_Open urbLL2_Open industrial_div, re

productivity R-sq: within = 0.0000 between = 0.0016 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0022331	.0002682	-8.33	0.000	-.0027587	-.0017075
$Localization_{islt}$	.0142516	.0046085	3.09	0.002	.0052191	.0232842
$Urbanization_{ilt}$	-.007931	.003918	-2.02	0.043	-.0156101	-.0002519
$Industrial\_diversity_{ils}$	-.0023943	.0014986	-1.60	0.110	-.0053316	.0005429
_cons	1.965382	.0117328	167.51	0.000	1.942386	1.988378
sigma_u	.69697617					
sigma_e	2.4973402					
rho	.07226135	(fraction of variance due to u_i)				

Wald chi2(4)=87.18
Prob > chi2=0.0000

xtreg productivity eta locCAP4_noOpen urbCAP4_noOpen industrial_div, re

productivity R-sq: within = 0.0000 between = 0.0020 overall = 0.0003	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0021924	.0002679	-8.18	0.000	-.0027174	-.0016674
$Localization_{islt}$	.0167787	.0042519	3.95	0.000	.0084452	.0251123
$Urbanization_{ilt}$	-.0249214	.0040813	-6.11	0.000	-.0329206	-.0169222
$Industrial\_diversity_{ils}$	.0022538	.0016565	1.36	0.174	-.0009928	.0055005
_cons	2.008319	.0137178	146.40	0.000	1.981432	2.035205
sigma_u	.69665502					
sigma_e	2.4973377					
rho	.07219972	(fraction of variance due to u_i)				

Wald chi2(4)=115.84
Prob > chi2=0.0000

xtreg productivity eta locCAP3_noOpen urbCAP3_noOpen industrial_div, re

productivity R-sq: within = 0.0000 between = 0.0019 overall = 0.0003	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0021894	.0002679	-8.17	0.000	-.0027145	-.0016643
$Localization_{islt}$	.0115623	.0040382	2.86	0.004	.0036476	.0194771
$Urbanization_{ilt}$	-.0236411	.0042633	-5.55	0.000	-.031997	-.0152852
$Industrial\_diversity_{ils}$	.0015881	.0016526	0.96	0.337	-.001651	.0048272
_cons	2.008051	.0136247	147.38	0.000	1.981347	2.034755

$sigma\_u$	.69674312					
$sigma\_e$	2.4973377					
rho	.07221666	(fraction of variance due to u_i)				
					Wald chi2(4)=108.37	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locCAP2\_noOpen\ urbCAP2\_noOpen\ industrial\_div, re$						
$productivity$ R-sq: within = 0.0000 between = 0.0019 overall = 0.0003	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0022096	.0002678	-8.25	0.000	-.0027345	-.0016846
$Localization_{islt}$	.0110527	.0038133	2.90	0.004	.0035788	.0185266
$Urbanization_{ilt}$	-.025726	.0046573	-5.52	0.000	-.0348542	-.0165978
$Industrial\_diversity_{ils}$	.0019845	.0016909	1.17	0.241	-.0013297	.0052987
$\_cons$	2.004744	.0133353	150.33	0.000	1.978607	2.030881
$sigma\_u$	.6967473					
$sigma\_e$	2.4973377					
rho	.07221747	(fraction of variance due to u_i)				
					Wald chi2(4)=108.61	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locCAP4\_Open\ urbCAP4\_Open\ industrial\_div, re$						
$productivity$ R-sq: within = 0.0000 between = 0.0020 overall = 0.0003	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0021838	.0002679	-8.15	0.000	-.0027089	-.0016587
$Localization_{islt}$	.0165785	.0042607	3.89	0.000	.0082277	.0249292
$Urbanization_{ilt}$	-.0229734	.0040466	-5.68	0.000	-.0309045	-.0150422
$Industrial\_diversity_{ils}$	.0018213	.0016513	1.10	0.270	-.0014152	.0050577
$\_cons$	2.003614	.0136278	147.02	0.000	1.976904	2.030324
$sigma\_u$	.69668057					
$sigma\_e$	2.4973405					
rho	.07220448	(fraction of variance due to u_i)				
					Wald chi2(4)=111.22	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locCAP3\_Open\ urbCAP3\_Open\ industrial\_div, re$						
$productivity$ R-sq: within = 0.0000 between = 0.0018	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	

overall = 0.0003						
age	-.0021816	.000268	-8.14	0.000	-.0027068	-.0016563
$Localization_{islt}$	.0113545	.0040441	2.81	0.005	.0034282	.0192809
$Urbanization_{ilt}$	-.0216991	.0042279	-5.13	0.000	-.0299857	-.0134125
$Industrial\_diversity_{ils}$	.0011562	.0016474	0.70	0.483	-.0020728	.0043851
$\_cons$	2.00351	.0135355	148.02	0.000	1.976981	2.030039
$sigma\_u$	.69676844					
$sigma\_e$	2.4973404					
rho	.07222138	(fraction of variance due to u_i)				
					Wald chi2(4)=103.88	
					Prob > chi2=0.0000	

$xtreg\ productivity\ eta\ locCAP2\_Open\ urbCAP2\_Open\ industrial\_div, re$

$productivity$ R-sq: within = 0.0000 between = 0.0018 overall = 0.0002	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.002201	.0002679	-8.22	0.000	-.0027261	-.001676
$Localization_{islt}$	.0108318	.0038154	2.84	0.005	.0033538	.0183098
$Urbanization_{ilt}$	-.0236593	.0046221	-5.12	0.000	-.0327185	-.0146002
$Industrial\_diversity_{ils}$	.0015041	.0016853	0.89	0.372	-.001799	.0048073
$\_cons$	2.000396	.0132533	150.94	0.000	1.97442	2.026373
$sigma\_u$	.69677068					
$sigma\_e$	2.4973411					
rho	.07222178	(fraction of variance due to u_i)				
					Wald chi2(4)=103.88	
					Prob > chi2=0.0000	

Exhibit 2 – Fixed effect

Fixed-effects (within) regression Group variable: IDNUM				Number of obs= Number of groups= Obs per group: min= avg= max=	727814 78152 1 9.3 10	
xtreg productivity eta locLL4_Open urbLL4_Open, fe						
productivity R-sq: within = 0.0000 between = 0.0018 overall = 0.0002 corr(u_i, Xb) = -0.1152	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.0061142	.0010995	5.56	0.000	.0039592	.0082692
$Localization_{islt}$	.0262349	.0422387	0.62	0.535	-.0565516	.1090214
$Urbanization_{ilt}$	-.0168751	.0160835	-1.05	0.294	-.0483982	.0146481
_cons	1.814881	.036455	49.78	0.000	1.743431	1.886332
sigma_u	1.0863617					
sigma_e	2.4973393					
rho	.15912092	(fraction of variance due to u_i)				
					F(3,649659) = 10.64 Prob > F = 0.0000	
F test that all u_i=0: F(78151, 649659) = 1.51			Prob > F = 0.0000			

xtreg productivity eta locLL3_Open urbLL3_Open, fe						
productivity R-sq: within = 0.0000 between = 0.0010 overall = 0.0001 corr(u_i, Xb) = -0.1162	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.0061149	.0010999	5.56	0.000	.0039592	.0082707
$Localization_{islt}$	.020345	.0380217	0.54	0.593	-.0541763	.0948663
$Urbanization_{ilt}$	-.0173755	.0171153	-1.02	0.310	-.0509209	.01617
_cons	1.814743	.036377	49.89	0.000	1.743446	1.886041
sigma_u	1.0864384					
sigma_e	2.4973395					
rho	.1591398	(fraction of variance due to u_i)				
					F(3,649659) = 10.61 Prob > F = 0.0000	
F test that all u_i=0: F(78151, 649659) = 1.51			Prob > F = 0.0000			

xtreg productivity eta locLL2_Open urbLL2_Open, fe					
productivity R-sq: within = 0.0000 between = 0.0012	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]

overall = 0.0002 corr(u_i, Xb) = -0.1175						
age	.0061173	.0011008	5.56	0.000	.0039597	.0082749
$Localization_{islt}$	.0063255	.0335373	0.19	0.850	-.0594064	.0720574
$Urbanization_{ilt}$	-.0142506	.0198456	-0.72	0.473	-.0531473	.0246461
$\_cons$	1.811332	.0359772	50.35	0.000	1.740818	1.881846
$sigma\_u$	1.0865459					
$sigma\_e$	2.4973402					
rho	.15916621	(fraction of variance due to u_i)				
						F(3,649659) = 10.48 Prob > F = 0.0000
F test that all u_i=0: F(78151, 649659) = 1.51			Prob > F = 0.0000			

$xtreg\ productivity\ eta\ locCAP4\_Open\ urbCAP4\_Open, fe$						
$productivity$ R-sq: within = 0.0000 between = 0.0009 overall = 0.0001 corr(u_i, Xb) = -0.1188	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.0057802	.001107	5.22	0.000	.0036105	.0079499
$Localization_{islt}$	.0416073	.0629032	0.66	0.508	-.081681	.1648957
$Urbanization_{ilt}$	.0008021	.0350212	0.02	0.982	-.0678383	.0694425
$\_cons$	1.753895	.1200046	14.62	0.000	1.51869	1.9891
$sigma\_u$	1.0867767					
$sigma\_e$	2.4973405					
rho	.15922304	(fraction of variance due to u_i)				
						F(3,649659) = 10.43 Prob > F = 0.0000
F test that all u_i=0: F(78151, 649659) = 1.51			Prob > F = 0.0000			

$xtreg\ productivity\ eta\ locCAP3\_Open\ urbCAP3\_Open, fe$						
$productivity$ R-sq: within = 0.0000 between = 0.0010 overall = 0.0001 corr(u_i, Xb) = -0.1205	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.0057728	.0011079	5.21	0.000	.0036014	.0079442
$Localization_{islt}$	.0416439	.0587049	0.71	0.478	-.0734159	.1567037
$Urbanization_{ilt}$	-.0034117	.0370178	-0.09	0.927	-.0759653	.0691419
$\_cons$	1.757094	.1196295	14.69	0.000	1.522624	1.991564
$sigma\_u$	1.0869452					
$sigma\_e$	2.4973404					
rho	.15926456	(fraction of variance due to u_i)				
						F(3,649659) = 10.45

Prob > F = 0.0000

F test that all u_i=0: F(78151, 649659) = 1.51 Prob > F = 0.0000

xtreg productivity eta locCAP2_Open urbCAP2_Open, fe

productivity R-sq: within = 0.0000 between = 0.0014 overall = 0.0002 corr(u_i, Xb) = -0.1185	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.0058044	.0011088	5.23	0.000	.0036311	.0079777
$Localization_{islt}$	.0176728	.0582157	0.30	0.761	-.0964282	.1317737
$Urbanization_{ilt}$	.0025251	.0429915	0.06	0.953	-.0817369	.0867871
_cons	1.750105	.1185383	14.76	0.000	1.517774	1.982436
sigma_u	1.0866139					
sigma_e	2.4973411					
rho	.15918287	(fraction of variance due to u_i)				

F(3,649659) = 10.33
Prob > F = 0.0000

F test that all u_i=0: F(78151, 649659) = 1.51 Prob > F = 0.0000

Exhibit 3 – Multilevel mixed model

Mixed-effects ML regression Group variable:	N° of Groups	Observations per Group Min	Avg	Max	Number of obs= 727814
CAP	3995	6	182.2	6,617	
IDFIRM	78152	1	9.3	10	

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL4\_noOpen\ urbLL4\_noOpen\ ||\ dvd2014\_h07: R.anno, covariance(identity)\ ||\ IDNUM:$

$productivity$ Log likelihood = -1716667.1 Wald chi2(5)=241.78 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028165	0.000251	-11.24	0.000	-0.00331	-0.00233
$Industrial\_diversity$	-0.0012794	0.001376	-0.93	0.352	-0.00398	0.001418
$Year$	0.0115816	0.001077	10.75	0.000	0.00947	0.013693
$Localization_{islt}$	0.0278586	0.005267	5.29	0.000	0.017536	0.038181
$Urbanization_{ilt}$	-0.00894	0.003225	-2.77	0.006	-0.01526	-0.00262
$\_cons$	-21.27608	2.161824	-9.84	0.000	-25.5132	-17.039
$CAP: Identity\ var\ (R.anno)$	0.00242	0.00149			0.00072	0.00812
$IDNUM: Identity\ var(\_cons)$	0.27703	0.00560			0.26627	0.28823
$var(Residual)$	6.31125	0.01132			6.28910	6.33347

LR test vs. linear model: chi2(2) = 3349.12 Prob > chi2 = 0.0000

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL3\_noOpen\ urbLL3\_noOpen\ ||\ dvd2014\_h07: R.anno, covariance(identity)\ ||\ IDNUM:$

$productivity$ Log likelihood = -1716673 Wald chi2(5)=229.85 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028137	0.000251	-11.23	0.000	-0.0033	-0.00232
$Industrial\_diversity$	-0.0018846	0.001372	-1.37	0.17	-0.00457	0.000805
$Year$	0.0115726	0.001077	10.74	0.000	0.009461	0.013684
$Localization_{islt}$	0.0193154	0.004788	4.03	0.000	0.009931	0.0287
$Urbanization_{ilt}$	-0.00827	0.003349	-2.47	0.013	-0.01484	-0.00171
$\_cons$	-21.25554	2.162193	-9.83	0.000	-25.4934	-17.0177
$CAP: Identity\ var\ (R.anno)$	0.00245	0.00148			0.00075	0.00798
$IDNUM: Identity\ var(\_cons)$	0.27718	0.00560			0.26641	0.28838
$var(Residual)$	6.31122	0.01132			6.28907	6.33344

LR test vs. linear model: chi2(2) = 3352.38 Prob > chi2 = 0.0000

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL2\_noOpen\ urbLL2\_noOpen\ ||\ dvd2014\_h07: R.anno, covariance(identity)\ ||\ IDNUM:$

$productivity$ Log likelihood = -1716675.5 Wald chi2(5)=224.83 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028309	0.000251	-11.3	0.000	-0.00332	-0.00234
$Industrial\_diversity$	-0.0020787	0.001391	-1.49	0.135	-0.0048	0.000647
$Year$	0.0115736	0.001078	10.74	0.000	0.009461	0.013686
$Localization_{islt}$	0.0142214	0.004201	3.39	0.001	0.005988	0.022455
$Urbanization_{ilt}$	-0.00883	0.00365	-2.42	0.016	-0.01598	-0.00168
$\_cons$	-21.25783	2.163039	-9.83	0.000	-25.4973	-17.0184
$CAP: Identity$ $var\ (R.anno)$	0.00253	0.00146			0.00082	0.00781
$IDNUM: Identity$ $var(\_cons)$	0.27722	0.00560			0.26645	0.28842
$var(Residual)$	6.31115	0.01132			6.28901	6.33337
LR test vs. linear model: chi2(2) = 3353.46		Prob > chi2 = 0.0000				

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL4\_Open\ urbLL4\_Open \|\|$ $dvd2014\_h07: R.anno, covariance(identity) \|\| IDNUM:$						
$productivity$ Log likelihood = -1716666.7 Wald chi2(5)=242.60 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028114	0.000251	-11.22	0.000	-0.0033	-0.00232
$Industrial\_diversity$	-0.0012522	0.001371	-0.91	0.361	-0.00394	0.001436
$Year$	0.0116141	0.001078	10.78	0.000	0.009502	0.013726
$Localization_{islt}$	0.0284359	0.005314	5.35	0.000	0.018021	0.038851
$Urbanization_{ilt}$	-0.00913	0.003178	-2.87	0.004	-0.01536	-0.00290
$\_cons$	-21.34152	2.162477	-9.87	0.000	-25.5799	-17.1032
$CAP: Identity$ $var\ (R.anno)$	0.00241	0.00012			0.00219	0.00266
$IDNUM: Identity$ $var(\_cons)$	0.27703	0.00560			0.26627	0.28823
$var(Residual)$	6.31125	0.01123			6.28927	6.33331
LR test vs. linear model: chi2(2) = 3359.15		Prob > chi2 = 0.0000				

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL3\_Open\ urbLL3\_Open \|\|$ $dvd2014\_h07: R.anno, covariance(identity) \|\| IDNUM:$						
$productivity$ Log likelihood = -1716672.7 Wald chi2(5)=230.41 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028091	0.000251	-11.21	0.000	-0.0033	-0.00232
$Industrial\_diversity$	-0.0018546	0.001368	-1.36	0.175	-0.00454	0.000826
$Year$	0.0116009	0.001078	10.76	0.000	0.009488	0.013713
$Localization_{islt}$	0.0197234	0.004828	4.09	0.000	0.010261	0.029185
$Urbanization_{ilt}$	-0.00850	0.003303	-2.57	0.01	-0.01497	-0.00203

$\_cons$	-21.31235	2.162847	-9.85	0.000	-25.5515	-17.0733
$CAP: Identity$ $var\ (R.anno)$	0.00245	0.00149			0.00074	0.00810
$IDNUM: Identity$ $var(\_cons)$	0.27718	0.00560			0.26642	0.28838
$var(Residual)$	6.31121	0.01132			6.28907	6.33344
LR test vs. linear model: chi2(2) = 3352.44		Prob > chi2 = 0.0000				

$mixed\ productivity\ eta\ industrial\_div\ anno\ locLL2\_Open\ urbLL2\_Open\ ||$
$dvd2014\_h07: R.anno, covariance(identity)\ ||\ IDNUM:$

$productivity$ Log likelihood = -1716675.4 Wald chi2(5)=225.01 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-0.0028274	0.000251	-11.28	0.000	-0.00332	-0.00234
$Industrial\_diversity$	-0.0020633	0.001385	-1.49	0.136	-0.00478	0.000652
$Year$	0.0115964	0.001078	10.75	0.000	0.009483	0.01371
$Localization_{islt}$	0.0143577	0.004227	3.4	0.001	0.006072	0.022643
$Urbanization_{ilt}$	-0.00898	0.003607	-2.49	0.013	-0.01605	-0.00191
$\_cons$	-21.30356	2.163744	-9.85	0.000	-25.5444	-17.0627
$CAP: Identity$ $var\ (R.anno)$	0.00253	0.00145			0.00082	0.00781
$IDNUM: Identity$ $var(\_cons)$	0.27722	0.00560			0.26646	0.28842
$var(Residual)$	6.31115	0.01132			6.28901	6.33337
LR test vs. linear model: chi2(2) = 3353.51		Prob > chi2 = 0.0000				

$mixed\ productivity\ eta\ industrial\_div\ anno\ locCAP3\_noOpen\ urbCAP3\_noOpen\ ||$
$dvd2014\_h07: R.anno, covariance(identity)\ ||\ IDNUM:$

$productivity$ Log likelihood = -1716664.1 Wald chi2(5)=248.06 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-0.002786	0.000251	-11.12	0.000	-0.00328	-0.0023
$Industrial\_diversity$	0.0016516	0.001527	1.08	0.28	-0.00134	0.004645
$Year$	0.011515	0.001076	10.7	0.000	0.009407	0.013623
$Localization_{islt}$	0.0112998	0.003691	3.06	0.002	0.004065	0.018534
$Urbanization_{ilt}$	-0.02317	0.003935	-5.89	0.000	-0.03088	-0.01546
$\_cons$	-21.09984	2.15887	-9.77	0.000	-25.3312	-16.8685
$CAP: Identity$ $var\ (R.anno)$	0.00215	0.00150			0.00055	0.00842
$IDNUM: Identity$ $var(\_cons)$	0.27692	0.00560			0.26616	0.28812
$var(Residual)$	6.31154	0.01132			6.28939	6.33377
LR test vs. linear model: chi2(2) = 3345.24		Prob > chi2 = 0.0000				

mixed productivity eta industrial_div anno locCAP2_noOpen urbCAP2_noOpen || dvd2014_h07: R.anno, covariance(identity) || IDNUM:						
productivity Log likelihood = -1716663.7 Wald chi2(5)=248.85 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028058	0.00025	-11.2	0.000	-0.0033	-0.00232
Industrial_diversity	0.0020794	0.001563	1.33	0.183	-0.00098	0.005142
Year	0.0115224	0.001076	10.71	0.000	0.009414	0.013631
$Localization_{islt}$	0.0110265	0.003485	3.16	0.002	0.004197	0.017856
$Urbanization_{ilt}$	-0.02539	0.004287	-5.92	0.000	-0.03380	-0.01699
_cons	-21.11798	2.15894	-9.78	0.000	-25.3494	-16.8865
CAP: Identity var (R.anno)	0.00216	0.00149			0.00056	0.00835
IDNUM: Identity var(_cons)	0.27689	0.00560			0.26613	0.28809
var(Residual)	6.31155	0.01132			6.28940	6.33378
LR test vs. linear model: chi2(2) = 3344.43		Prob > chi2 = 0.0000				

mixed productivity eta industrial_div anno locCAP3_Open urbCAP3_Open || dvd2014_h07: R.anno, covariance(identity) || IDNUM:						
productivity Log likelihood = -1716665 Wald chi2(5)=246.22 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.002784	0.000251	-11.11	0.000	-0.00328	-0.00229
Industrial_diversity	0.0014722	0.001523	0.97	0.334	-0.00151	0.004457
Year	0.011663	0.001076	10.84	0.000	0.009555	0.013772
$Localization_{islt}$	0.0111764	0.003699	3.02	0.003	0.003927	0.018426
$Urbanization_{ilt}$	-0.02238	0.003906	-5.73	0.000	-0.03004	-0.01473
_cons	-21.39892	2.158695	-9.91	0.000	-25.6299	-17.168
CAP: Identity var (R.anno)	0.00212	0.00151			0.00053	0.00857
IDNUM: Identity var(_cons)	0.27694	0.00560			0.26618	0.28814
var(Residual)	6.31156	0.01132			6.28941	6.33380
LR test vs. linear model: chi2(2) = 3345.68		Prob > chi2 = 0.0000				

mixed productivity eta industrial_div anno locCAP2_Open urbCAP2_Open || dvd2014_h07: R.anno, covariance(identity) || IDNUM:						
productivity Log likelihood = -1716664.7 Wald chi2(5)=246.89 Prob > chi2=0.0000	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-0.0028035	0.00025	-11.19	0.000	-0.00329	-0.00231
Industrial_diversity	0.0018816	0.001558	1.21	0.227	-0.00117	0.004936
Year	0.0116725	0.001076	10.85	0.000	0.009564	0.013781

$Localization_{islt}$	0.0108957	0.003489	3.12	0.002	0.004058	0.017733
$Urbanization_{ilt}$	-0.02456	0.00426	-5.76	0.000	-0.03290	-0.01621
$\_cons$	-21.42103	2.15919	-9.92	0.000	-25.653	-17.1891
$CAP: Identity$ $var\ (R.anno)$	0.00217	0.00149			0.00056	0.00835
$IDNUM: Identity$ $var(\_cons)$	0.27691	0.00560			0.26614	0.28810
$var(Residual)$	6.31154	0.01132			6.28939	6.33377
LR test vs. linear model: chi2(2) = 3344.83		Prob > chi2 = 0.0000				